Journal of Accounting Research

DOI: 10.1111/1475-679X.12219
Journal of Accounting Research
Vol. 56 No. 4 September 2018
Printed in U.S.A.

CHICAGO BOOTH

Inference with Dependent Data in Accounting and Finance Applications

TIMOTHY CONLEY,[*] SILVIA GONÇALVES,[†]
AND CHRISTIAN HANSEN[‡]

Received 8 November 2017; accepted 17 April 2018

ABSTRACT

We review developments in conducting inference for model parameters in the presence of intertemporal and cross-sectional dependence with an emphasis on panel data applications. We review the use of heteroskedasticity and autocorrelation consistent (HAC) standard error estimators, which include the standard clustered and multiway clustered estimators, and discuss alternative sample-splitting inference procedures, such as the Fama–Macbeth procedure, within this context. We outline pros and cons of the different procedures. We then illustrate the properties of the discussed procedures within a simulation experiment designed to mimic the type of firm-level panel data that might be encountered in accounting and finance applications. Our conclusion, based on theoretical properties and simulation performance, is that sample-splitting procedures with suitably chosen splits are the most likely to deliver robust inferential statements with approximately correct coverage

[*] University of Western Ontario; [†] McGill; [‡] University of Chicago Booth School of Business.

Accepted by Christian Leuz. We would like to thank Rodrigo Verdi for providing us with the data we used to construct our simulation experiment and JAR editors for helpful comments. We thank Aldo Sandoval Hernandez for providing excellent research assistance. This material is based on work supported by the National Science Foundation under Grant No. 1558636, the University of Chicago Booth School of Business, and the Social Sciences and Humanities Research Council of Canada. An online appendix to this paper can be downloaded at http://research.chicagobooth.edu/arc/journal-of-accounting-research/online-supplements.

© University of Chicago on behalf of the Accounting Research Center, 2018

properties in the types of large, heterogeneous panels many researchers are likely to face.

JEL codes: C12; C23

Keywords: hypothesis testing; confidence intervals; robust standard error estimation; spatial dependence; bootstrap; fixed-effects

1. *Introduction*

Empirical research in accounting and finance often uses panel data on firms, sectors, or regions over time. It is routine for such data to be interrelated. In particular, unobservable factors ("shocks") are typically important in determining outcomes and seem likely to be related across observations. Time series shocks affecting an individual firm or geographic region are often taken to be serially correlated. Similarly, shocks at a single point in time affecting different observations may be correlated with each other. As examples, supply shocks may jointly impact all firms in industries with similar technology, and shocks to interest rate expectations may jointly impact firms with similar exposure to interest rate risk. Moreover, such shocks are likely to be correlated across firms at different time periods. For example, firms with similar investment opportunities will tend to make similar choices and experience correlated shocks. Furthermore, with multiperiod investments, such firms will routinely exhibit correlation across nearby time periods, not just contemporaneously.[1]

It is well known that researchers need to account for the presence of dependent unobservables when conducting statistical inference for model parameters. For example, a 5% level test formed from a *t*-statistic with standard error that is estimated assuming independence across observations can have size—the probability of rejecting a true null hypothesis—very far from 5% when the data are in fact dependent. This potential for distortions to inferential statements from failing to properly account for dependence has long been recognized in the time series literature. In the empirical economics literature, Bertrand, Duflo, and Mullainathan [2004] highlighted this point in the context of panel data with cross-sectional independence and intertemporal correlation; and, casual empiricism based on papers appearing after Bertrand, Duflo, and Mullainathan [2004] suggests that applied researchers dealing with panel data are now acutely aware of the potential for distortions from failing to account for dependence when conducting inference. Indeed, the vast majority of applied work with panel data in accounting, finance, and economics uses inference procedures that are robust to some form of correlation across observations.

[1] We highlight that this intuitive structure leads to correlation across different firms in different time periods and would render the common practice of using two-way clustering by firm and time inappropriate, as this two-way clustering structure imposes that different firms in different time periods are uncorrelated.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The importance of adequately accounting for dependence, both intertemporal and cross sectional, has led to the development of a variety of statistical procedures that aim to deliver valid inferential statements about parameters of interest when data may be dependent and heterogeneous. These methods include the use of clustered standard error estimators, sample-splitting procedures such as the Fama–Macbeth procedure, and bootstrap procedures. While the menu of available methods offers researchers many high-quality options, the methods are not equivalent and involve substantive choices. For example, when using clustered standard errors, the obvious decision that must be made is at what level to form clusters. There are also less obvious choices such as what critical values to use and what fixed effects structure to maintain that have important impacts on the quality of inference.

The goal of this review is to offer a heuristic overview of leading inferential approaches with dependent data and a practical guide to some of the tradeoffs between methods and choices that must be made. In addition to reviewing different broad classes of methods, we talk about practical issues that are common to all approaches and often ignored in the literature. Two particularly important issues are the choice of group structure, for example, on what level(s) to cluster for one- or multiway clustering, and how the choice of the group structure interacts with fixed effects structures.

The practical recommendations we make are based on a simulation study that was designed to allow an evaluation of alternative procedures in a typical accounting or finance application. Importantly, our simulation model is not based on a stylized model with simple dependence structure. Rather, we base the simulation on the empirical analysis in Balakrishnan, Core, and Verdi [2014] and use a data-generating process (DGP) that is heterogeneous, allows for dependence along multiple dimensions, and is designed to approximate the correlation structure that is present in the data. Our simulation results should therefore be empirically relevant for accounting and corporate finance applications.

1.1 MAIN RESULTS

We evaluate alternative inference procedures in the context of a simple panel-data regression estimated by ordinary least squares (OLS) under strict exogeneity of regressors so there are no finite sample bias concerns. To summarize the main points from our simulation, we conclude that sample-splitting strategies (e.g., Fama–MacBeth) with a small number (e.g., 5–10) of groups that each consist of many observations are likely to yield the most reliable inferential statements in many accounting and finance applications. The use of a small number of large groups allows one to accommodate very rich dependence structures and the use of sample-splitting allows one to accommodate very heterogeneous data. Having many observations per group is also important for sample-splitting estimators as they require estimation of model parameters within each group, and these group-level estimates may be very unstable if the groups have few observations.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Our recommendation to use sample-splitting approaches, as opposed to using clustered standard errors, and a small number of groups may be surprising to some readers. Using a small number of large groups results in inconsistent estimates of standard errors and differs from the folk recommendation, motivated by approximations that rely on consistent standard estimation, to use at least a modest number (e.g., 30–40 or more) of groups. Our recommendation is in line with a key contribution of the recent theoretical literature, outlined below, that highlights that consistent estimation of standard errors is *not* necessary to obtain high-quality inference and that inferential performance tends to be much more robust when a small number of groups is used. The use of sample-splitting strategies versus clustering is motivated by the fact that inference using clustered standard errors with few, large clusters relies on homogeneity conditions across clusters that seem unlikely to be satisfied in firm-level panels. In contrast, sample-splitting strategies do not require these homogeneity conditions.[2]

The simulation evidence also suggests that it is important that fixed effects estimated with a small number of observations do not cross group boundaries as the estimation of fixed effects using observations from multiple clusters results in mechanical correlation across the clusters. Rather, fixed effects should ideally be nested within any clustering structure. A nice by-product of sample-splitting strategies is that they mechanically keep any fixed effects structure nested within the structure used to split the sample.

1.2 OUTLINE

In the remainder of this introductory section, we provide a high-level overview of inferential approaches and a literature review. We outline the basic problem of inference with dependent data in a simplified setting with heuristic derivations that allows us to convey the key theoretical insights without dwelling on technical details in section 2. In section 3, we provide a discussion of leading inferential approaches. Our discussion is informal and meant to allow us to provide some intuition into the functioning of the procedures along with providing a venue to outline their pros and cons. Finally, we present our simulation results along with detailed discussion of the performance of the different procedures and key practical takeaways in section 4. While presenting the simulation results in section 4, we also comment on many practical implementation details for the different methods. The main results, summarized above, are elaborated on in a concluding section.

[2] The discussion becomes more nuanced in scenarios where finite sample bias is a key concern. Key examples where finite sample bias may be prominent are instrumental variables estimation and estimation in fixed effects panels including lagged dependent variables. In such cases, the performance of the point estimator resulting from sample-splitting deteriorates due to the presence of bias and there is a tradeoff between bias of the point estimator and robustness of inference to heterogeneity. We briefly comment further on these issues in section 3.3.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1.3 LITERATURE REVIEW AND OVERVIEW

We consider four basic approaches to estimation and inference in this review. First, we consider traditional large sample approximations that use the full sample to estimate the parameters of interest and use heteroskedasticity and autocorrelation consistent (HAC) estimators, which include one- and multiway clustering as special cases, to estimate the standard errors associated with these parameters. These traditional approximations maintain the classical assumption that the standard error estimators are extremely accurate (consistent). Second, we turn to recent alternative approximations that modify the traditional approach by considering estimators of standard errors that are not assumed to be consistent. Approximations that do not rely on consistent standard error estimators but instead directly account for uncertainty in standard error estimates have the potential to improve inference in applications with firm-level panel data where it is difficult to estimate standard errors. We then turn to a third approach, which adapts the sample-splitting approach of Fama and MacBeth [1973]. Finally, we examine the performance of bootstrap-based inference.

Probably the most popular method for performing inference allowing for dependent and heterogeneous observations in practice is to use so-called robust standard error estimators coupled with conventional test statistics. These methods were popularized in economic applications by White [1980], which presented a standard error estimator robust to heteroskedasticity under the assumption of independence across observations; see also Eicker [1967] and Huber [1967]. These methods have then been adapted to handle a variety of different dependence structures. See, for example, Levine [1983], White and Domowitz [1984], Newey and West [1987], and Andrews [1991] for time series dependence; Liang and Zeger [1986], Arellano [1987], and Bertrand, Duflo, and Mullainathan [2004] for panel data with dependence within the panel unit of observation but independence across units of observation[3]; Cameron, Gelbach, and Miller [2011] for clustering along multiple dimensions; and Conley [1999] and Kelejian and Prucha [2007] for cross-sectional dependence. We refer to the general class of robust standard error estimators as HAC estimators for simplicity.

An alternative for estimating the standard error of an estimator is to adopt a sample-splitting approach. This method splits the data into a set of G subsamples and then estimates the model of interest within each subsample. Thus, one obtains a set of G estimates of any parameter of interest, one from each subsample. These G estimates are then treated as G independent observations. The final point estimate for a parameter is obtained simply as the average of its G within-subsample estimates, and its standard

[3] This structure corresponds to the so-called "clustered" standard error estimators where the panel unit of observation denotes a cluster and observations within-cluster are allowed to be essentially arbitrarily correlated.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

error corresponds to the usual standard error of a sample mean estimated from G independent observations. A special case of this method is the approach taken in Fama and MacBeth [1973]. We will refer to these and related procedures as "sample-splitting" or Fama–MacBeth approaches (FM hereafter) in honor of the fundamental contribution and insight of Fama and MacBeth [1973]. We wish to highlight that our usage of this term is far more general than that often associated with the procedure of Fama and MacBeth [1973], which involves estimating a model within each time period and then proceeding with inference using the resulting time series of estimated coefficients. We are rather using FM to denote any procedure that subsets the data and estimates an appropriate model within subsets. For example, our ultimate recommendation will be to perform FM-style inference using a small number of subsets, which are large enough to accommodate any relevant fixed effects structure and capture rich forms of intertemporal and cross-sectional dependence. We refer to the procedure resulting from this recommendation as FM, though implementing this recommendation will generally involve forming subsets consisting of multiple time periods and cross-sectional units, which differs in detail from the original implementation in Fama and MacBeth [1973].

A common feature of early HAC papers is that they rely on an approximation that leverages being able to consistently estimate the sampling variance matrix of the parameter estimator. Inference relies on an argument that essentially starts by approximating the behavior of an inference procedure, for example, a hypothesis test, when the true sampling variance of the estimator of the parameter of interest is known. The next step is to show that, due to consistency, the estimated sampling variance will be arbitrarily close to the true sampling variance in a large enough sample. Typically, consistency of standard error estimators allowing for dependence will require strong conditions. For example, multiway clustering estimators are only consistent if the smallest number of groups along any of the dimensions that clustering occurs is large, which may be hard to satisfy in many common uses in accounting and finance where clustering may be by firm and time and the time dimension is relatively short.

While simple, basing inference on approximations that take the estimated standard errors to be arbitrarily close to the true standard errors ignores the fact that standard errors themselves are estimated quantities that are not exactly equal to the true standard errors. In finite samples, estimated standard errors exhibit sampling variability themselves and potentially suffer from bias. The variability in standard errors is especially important in dependent data settings where accounting for dependence leads to additional sampling variability that may be large even in big data sets. Failing to account for this variability in estimating standard errors then potentially results in poor performance of standard inference procedures as evidenced, for example, in Kiefer and Vogelsang [2002, 2005]. Within the finance and accounting literature, Petersen [2009] and Gow,

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Ormazabal, and Taylor [2010] provide excellent reviews focused on comparing different variants of HAC and FM approaches within this classical setting.

To address the drawbacks of relying on consistent standard error estimators, the recent theoretical literature on inference for dependent data in econometrics and statistics has turned to providing inference procedures that explicitly account for sampling uncertainty in standard error estimators. These approaches are focused on using standard test statistics (e.g., Student's *t*-statistics) but providing alternative reference distributions[4] to account for the sampling variation in the estimated standard errors. These alternative reference distributions tend to have critical values that are larger than those for conventional reference distributions. Effectively, these larger critical values account for the additional uncertainty due to sampling variation in the standard error estimator.

A popular approach for accounting for uncertainty in standard error estimation within the HAC framework was pioneered in Kiefer and Vogelsang [2002, 2005] and Vogelsang [2003] for data with time series dependence. Kiefer and Vogelsang [2002, 2005] and Vogelsang [2003] provide a limiting distribution for commonly used test statistics with an approximation that treats standard error estimators as inconsistent in the sense that they are approximately unbiased but have variance that does not vanish even in very large samples. This approach thus explicitly approximates the finite-sample situation where the sampling variability in estimating the standard error is not vanishingly small, which seems to capture many real-world scenarios. This work has been extended to panel data with independent units of observation in Hansen [2007], to cross-sectionally dependent data in Bester et al. [2016], and to panels with intertemporal and cross-sectional correlation in Vogelsang [2012] and Bester, Conley, and Hansen [2011].

Similar results that do not rely on consistent standard error estimation are provided for sample-splitting procedures in Ibragimov and Müller [2010, 2016] and Canay, Romano, and Shaikh [2017]. The primary benefit of FM methods relative to inference based on HAC estimators in settings where noise in estimating standard errors is important is that good performance of HAC-based inference has been shown only under conditions that strongly restrict heterogeneity across observations and that are unnecessary for FM methods.[5] This robustness to heterogeneity is an important advantage in our example given the substantial heterogeneity across observations in the firm-level panel data that we use in our simulation.

We wish to emphasize that a key insight from this theoretical work is that it is *not* necessary to have consistent estimators of standard errors to conduct inference. Rather two conditions are sufficient. First, an inferential

[4] The reference distribution is the distribution that characterizes the behavior of the statistic under the null hypothesis and is used, for example, for obtaining critical values that control the size of the test at the desired level.

[5] The costs of this robustness to heterogeneity are discussed in section 3.3.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

procedure needs to adequately accommodate the dependence that is actually in the data. Accommodating rich dependence structures, such as those one might anticipate in firm-level panel data, in estimating standard errors will tend to result in highly variable standard error estimators. A second ingredient is then making use of an approximation for a statistic's sampling distribution that accounts for this variability in the standard error estimator. Importantly, the reference distributions for such approximations will not, in general, be usual distributions such as the standard normal. Much of the remainder of the paper is spent on elaborating on these two points.

The aforementioned approaches rely on analytically approximating the behavior of a statistic to obtain inferential statements. The bootstrap offers an alternative that replaces analytic approximation with simulation. The main idea underlying the bootstrap is to obtain a sampling distribution for a statistic of interest by approximating the distribution of the data itself. Typically, simulation methods are used to construct many samples of bootstrap data that each have approximately the same distribution as the original data. The empirical distribution of simulated statistics corresponding to these bootstrap data is then used to approximate the sampling distribution of the statistic computed from the original data. Thus, the bootstrap avoids the need for asymptotic approximations for statistics and may capture variability in the standard error component of commonly used statistics.

Generally, bootstrap inference works well when the bootstrap data have approximately the same dependence structure that exists in the true data. Replicating this structure is complicated in data exhibiting dependence over time and across space. We consider two different mechanisms for generating bootstrap data that have been proposed in this setting. We apply the overlapping blocks bootstrap (e.g. Künsch [1989] and Liu and Singh [1992]) by treating the panel as a vector time series and treating each cross-section as an observation from this time series. This structure allows for very general cross-sectional correlation but may suffer due to the relatively short time series in many panel data sets. We also consider the cluster wild bootstrap (e.g. Cameron, Gelbach, and Miller [2008] and Mackinnon and Webb [2016]) for several different definitions of clusters. This approach captures dependence within clusters but imposes approximate independence across clusters.

2. *Basic Statistical Problem*

Typical statistical methods involve averages across sample observations and the variability in these averages needs to be accounted for in order to make good inferences. As a simple example that captures the main ideas, consider the problem of estimating a population mean, μ, given data that consist of observations y_{it} that are noisy measures of μ from many firms, N, and time periods T. The statistical model is then simply

$$y_{it} = \mu + \eta_{it}, \tag{1}$$

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where the η_{it} are mean zero but may be correlated across firms and/or time periods. A natural estimator of μ is the sample average of the y_{it}, $\bar{y}$:

$$\bar{y} = \frac{1}{NT}\sum_{i,t} y_{it} = \mu + \frac{1}{NT}\sum_{i,t} \eta_{it}. \tag{2}$$

The variance of $\bar{y}$, V_{NT}, is equal to the variance of $\frac{1}{NT}\sum_{i,t}\eta_{it}$, where

$$V_{NT} = \frac{1}{(NT)^2}\left[\sum_{i,t}\mathrm{Var}(\eta_{it}) + \sum_{i,t}\sum_{(j,s)\neq(i,t)}\mathrm{Cov}(\eta_{it}, \eta_{js})\right] \tag{3}$$

by a standard calculation. Thus, V_{NT} depends in general on not just the variances of the η_{it} but also on all the covariances between them. Note that the first summation in (3), which accounts for the contribution of the variances to the sampling variability of the sample mean, depends on NT elements while the second summation, which provides the contribution of covariances to the sampling variability of the sample mean, is a sum of $(NT)^2 - NT$ terms. Even when these covariances are each small, their sum can still account for a large portion of the variance of the sample mean estimator, V_{NT}, due to the large number of covariances in (3).

We focus on applications where it is reasonable to assume the estimator of the parameter of interest, $\bar{y}$ in this example, is consistent and $\sqrt{NT}$-asymptotically normal.[6] Such an approximation will typically rely on a notion of weak dependence between observations. By weakly dependent, we mean that correlations between observations decay relatively quickly as one considers observations farther apart in time or in the cross section. This decay will allow the use of usual asymptotic approximations, which require correlations across observations to shrink fast enough with a notion of economic or time distance, so that $\frac{1}{NT}\sum_{i,t}\sum_{(j,s)\neq(i,t)}\mathrm{Cov}(\eta_{it}, \eta_{js}) \to c$ with $|c| < \infty$. Under weak dependence, the sampling distribution of $\bar{y}$ will then be approximately $N(\mu, V_0/(NT))$, where $V_{NT} \to V_0$.

Weak dependence may not be appropriate for all applications, of course; for example, weak dependence will not hold in data with trends or random walk–like behavior. We feel that dealing with such structures would take us too far afield and add little to understanding the key issues regarding inference with dependent data. We note that weak dependence is a reasonably general structure and that often in applications weak dependence will only need to hold after sensible data transformations are used to remove trends from data. Further, most linear panel models include firm and time fixed effects, so weak dependence needs to hold only after removing an arbitrary common trend and any time-invariant firm-level heterogeneity. Finally, one could allow many types of strong dependence without substantively

[6] Note that by $\sqrt{NT}$-asymptotically normal, we mean that $\sqrt{NT}(\bar{y} - \mu)$ has a limiting distribution that is normal.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

changing the main features of the discussion beyond adding technical complication.

Maintaining the assumption of weak dependence, we now turn to inference. For simplicity, we consider the case where interest is in testing a simple null hypothesis of the form $H_0 : \mu = \mu_0$ for some number μ_0. Under this hypothesis and assuming that we know V_0, we can form a t-statistic, t_1, as

$$t_1 = \frac{\sqrt{NT}(\bar{y} - \mu_0)}{\sqrt{V_0}}.$$

Using the asymptotic normality of $\bar{y}$, we can approximate the distribution of the test statistic using $t_1 \overset{\text{approx}}{\sim} N(0, 1)$ under the null hypothesis. V_0 is, of course, unknown in practice, so it must be estimated from the data to make this approximation useful. Estimating V_0 or obtaining some other way to benchmark sampling uncertainty under dependence poses complications. Providing insight into these complications and outlining solutions is the focus of the remainder of this review.

3. Inference Approaches

In the following, we outline four key basic approaches to estimation and inference, including a heuristic discussion of their theoretical underpinnings and mathematical motivation. We highlight practical distinctions in section 4.

3.1 INFERENCE BASED ON CONSISTENT HAC ESTIMATION

Typical large sample inference scales the statistic t_1 using an estimator of V_0 that is assumed to be consistent, call it $\hat{V}_c$. Plugging this estimator in when forming the t-statistic for testing the null hypothesis that $\mu = \mu_0$ yields

$$t = \frac{\sqrt{NT}(\bar{y} - \mu_0)}{\sqrt{\hat{V}_c}} \overset{\text{approx}}{\sim} \frac{\sqrt{V_0}N(0, 1)}{\sqrt{\hat{V}_c}}. \tag{4}$$

Assuming $\hat{V}_c$ is a consistent estimator of V_0, we then have that the ratio $\frac{\sqrt{V_0}}{\sqrt{\hat{V}_c}}$ may be taken to be arbitrarily close to one with high probability in large enough samples. Thus, we have that the t-statistic approximately follows a standard normal distribution in this setting. In practice, this approximation will only work well when $\hat{V}_c$ is an extremely accurate estimator of V_0 so that approximating the ratio $\frac{\sqrt{V_0}}{\sqrt{\hat{V}_c}}$ by one has little impact on the behavior of the t-statistic. Such an approximation essentially relies on having an estimator of V_0 that has little bias and little sampling variability in finite samples.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In terms of estimating V_0, we focus on estimators that can be written using residuals $\hat{\eta}_{it} = y_{it} - \bar{y}$ as

$$\hat{V}_{NT} = \frac{1}{NT} \sum_{i,t} \sum_{j,s} W(it,\, js)\hat{\eta}_{it}\hat{\eta}_{js}. \tag{5}$$

The term $W(it, js)$ in (5) is a weight on the pair of residuals from observations *it* and *js*. We require that this weight is one when both indexes are the same, $W(it, it) = 1$. For distinct observations *it* and *js*, nonzero weights allow for a contribution of the corresponding covariance term, that is, $\mathrm{Cov}(\eta_{it}, \eta_{js})$, to the estimator. This type of estimator is commonly referred to as an HAC estimator.

The weights used in an HAC estimator play a crucial rule in controlling the types of dependence that are allowed to influence the estimator of V_0. If all weights are set equal to zero when $it \neq js$, the HAC estimator reduces to the standard heteroskedasticity-consistent variance estimator of White [1980]. One can allow for cross-sectional dependence by choosing weights such that $W(it, js) \neq 0$ for $i \neq j$ and may produce weights that are appropriate for weak dependence in the cross-section by choosing $W(it, js)$ to depend on "economic distance" between firms, with smaller values of $W(it, js)$ for more distant firms. Similarly, weak dependence in the time series can be accounted for by allowing $W(it, js)$ to depend on the intertemporal distance between observations, $|t - s|$.

In this review, we focus on the special case of cluster covariance estimators. Conventional cluster covariance estimators correspond to (5) using the weighting $W(it, js) = 1$ if observations *it* and *js* belong to the same specified group of observations. Typically, these groups are formed by partitioning the data along a particular dimension or set of dimensions. As examples, a researcher may choose to cluster by firm, which corresponds to setting $W(it, js) = 1$ whenever observation *it* and observation *js* belong to the same firm; or, a researcher may choose to cluster by state which corresponds to setting $W(it, js) = 1$ whenever observation *it* and observation *js* belong to the same state. Clustered covariance estimators have a number of appealing features. They are positive semidefinite by construction and are easy to compute. They also allow for very general correlation within cluster but ignore, in terms of forming the estimator for V_0, any correlations across groups. Cluster covariance estimators are also easily explained; a researcher simply needs to specify the definitions of groups, for example, states, to convey that estimated standard errors are robust to very general correlation between observations belonging to each state but rely on assuming that correlations between observations from different states are ignorable. These estimators have long been employed in applications with a large number of groups (clusters) that are assumed independent; see, for example, Liang and Zeger [1986] and Arellano [1987].

HAC estimators can also be used with multiple metrics used to define the W weights. For example, $W(it, js)$ could be set equal to one if t and

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

s are close in time and if either firms i and j have similar technology or have suppliers located in similar geographic regions. Multiway clustering, for example, Cameron, Gelbach, and Miller [2011], is a popular special case of HAC estimation accounting for multiple indices. The most popular version is two-way clustering, which relies on using two different clustering partitions. For example, one set of N clusters could be individual firms and another set could define T groups consisting of all observations with a common time period. Two-way clustering with these partition sets $W(it, js) = 1$ if $i = j$ or $t = s$ and equal to zero otherwise. Like (one-way) clustering, multiway clustering estimators are easy to compute and explain. To explain the type of dependence one is robust to, a researcher again simply needs to specify the definition of groups, for example, two-way clustering by state and time, to convey that estimated standard errors are robust to very general correlation between observations belonging to the same state or belonging to the same time period but rely on assuming that correlations between observations that share neither state nor time period are ignorable.

Multiway clustering estimators are intuitive but do suffer from some drawbacks. Unlike one-way clustering, multiway clustering may produce negative variance estimates. In practice, this defect tends to be resolved by adding an ad hoc positive number to the variance in cases where a negative variance is estimated as in Cameron, Gelbach, and Miller [2011]. It is also important to remember that consistency of the multiway clustering estimator relies on very stringent conditions that require, at a minimum, that the smallest number of groups along any of the dimensions along which one is clustering is very large. In applications like Balakrishnan, Core, and Verdi [2014], where the number of time periods per firm ranges from 2 to 17, regarding such estimators as consistent is problematic. Moreover, it will be implausible for there to be appreciable correlations across some sets of firms at a point in time, appreciable correlation within each firm over time, but no correlations between these same firms at distinct but close points in time in many applications with firm-level panels.

Inference using approximation (4) and an HAC estimator for $\hat{V}_c$ will work well in scenarios where dependence is weak enough that only a modest number of the $W(it, js)$ weights need to be nonzero, relative to the sample size, to capture the contribution of covariances to V_0. A classic example where this seems like a sensible approximation is in a very short T panel on a large number N of independently sampled individuals. In this setting, maintaining that $W(it, js) = 0$ for all $i \neq j$ seems reasonable, which leaves relatively few nonzero weights. However, there are also many applications where correlations across observations are high enough that it is difficult to estimate V_0 because of the contributions of all the covariances to V_0 in (3). We suspect that this setting encompasses many applications in accounting and finance with observational data on firm-level or aggregate-level data that are likely correlated both in space and time. Estimators of V_0 in such examples are better viewed as noisy and perhaps biased estimators of V_0.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In such applications, the standard normal distribution can be a very bad approximation to the sampling distribution of *t*-statistics.

3.1.1 Illustrating the Bias–Variance Tradeoff in HAC Estimation: Moving Average Example. As a heuristic illustration of what goes into establishing consistency of an HAC estimator, consider the simple case of a single time series of T observations from a stationary first-order moving average with $\mathrm{Cov}(\eta_t, \eta_{t-1}) = \theta$ and $\mathrm{Var}(\eta_t) = \sigma^2$. Note that we would have $V_0 = \sigma^2 + 2\theta$ in this case, and we would have an exact variance of $\sqrt{T}(\bar{y} - \mu_0)$ equal to $V_T = \sigma^2 + 2\theta\frac{T-1}{T}$. Further suppose that standard errors will be estimated using a clustering estimator that groups together blocks of adjacent time series observations, and suppose that the $\{\eta_t\}_{t=1}^{T}$ were actually observed. Note that the following discussion could be made rigorous and allow for estimation of the residuals; see Bester, Conley, and Hansen [2011].

Suppose that one wished to use $G_T > 1$ clusters. Ignoring integer problems, forming G_T groups would correspond to each group consisting of $n_T = T/G_T$ consecutive observations with group 1 containing observations $1, \ldots, n_T$, group 2 containing observations $n_T + 1, \ldots, 2n_T$, and so forth. In this case, we would have

$$\widehat{V}_{G_T} = \frac{1}{T}\sum_{g=1}^{G_T}\sum_{s=(g-1)n_T+1}^{gn_T}\sum_{t=(g-1)n_T+1}^{gn_T}\eta_s\eta_t.$$

We would then have $\mathrm{E}[\widehat{V}_{G_T}] = \frac{1}{T}G_T(n_T\sigma^2 + 2(n_T - 1)\theta) = \sigma^2 + 2\theta\frac{T-G_T}{T}$ from which we obtain the bias of the estimator

$$\mathrm{Bias}\left[\widehat{V}_{G_T}\right] = \left\{\sigma^2 + 2\theta\frac{T-G_T}{T}\right\} - \left\{\sigma^2 + 2\theta\frac{T-1}{T}\right\} = -2\theta\frac{G_T - 1}{T}. \quad (6)$$

Letting $\omega_{st} = \eta_s\eta_t - \mathrm{E}[\eta_s\eta_t]$, we can also write the variance of $\widehat{V}_{G_T}$ as

$$\mathrm{Var}\left[\widehat{V}_{G_T}\right] = \frac{1}{T^2}\sum_{g=1}^{G_T}\sum_{q=(g-1)n_T+1}^{gn_T}\sum_{r=(g-1)n_T+1}^{gn_T}\sum_{h=1}^{G_T}\sum_{s=(h-1)n_T+1}^{hn_T}\sum_{t=(h-1)n_T+1}^{hn_T}\mathrm{E}[\omega_{qr}\omega_{st}]. \quad (7)$$

Importantly, assuming all expectations are bounded, the summation in (7) will consist of the order of $G_T n_T^2$ nonzero terms even in the best possible case where all observations were independent (and recall that they are not in this example). Taking this optimistic rate gives

$$\mathrm{Var}\left[\widehat{V}_{G_T}\right] = C\frac{G_T n_T^2}{T^2} = C\frac{n_T}{T} \quad (8)$$

for some constant C that would depend on higher order moments of the η_t. Expressions (6) and (8) then give us a crude set of conditions, which can be used to discuss the bias–variance tradeoff in this simple example.

First, suppose that one wished to use T clusters, which would correspond to setting $W(t, t) = 1$ and $W(t, s) = 0$ for all $s \neq t$. In this case, the HAC

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

estimator would correspond to White [1980], which does not allow for dependence. Looking at expression (6), we can see that bias of the estimator is $2\theta\frac{T-1}{T}$, which does not approach 0 even for large T. On the variance side, (8) gives $\text{Var}[\widehat{V}_{G_T}] = C/T \to 0$; so the estimator would be inconsistent. The inconsistency in this case is because the clustering structure, which does not allow any correlation in this example, is not rich enough to capture the true dependence in the data; so the estimator is arbitrarily close to the wrong value with high probability in large enough samples.

Next, we can see that using any sequence of clustering structures with $\frac{n_T}{T} \to 0$ will be sufficient to produce an estimator with arbitrarily small variance in a large enough sample based on the heuristic bound in (8). However, as illustrated in the preceding example, having variance go to zero is insufficient to produce a good estimator as we would also like bias to be small. Producing small bias relies on having $\frac{G_T}{T} \to 0$. Having variance go to zero requires that we consider a sequence of clustering structures where we would use larger numbers of clusters when faced with larger sample sizes such that $G_T \to \infty$ and, thus, $\frac{n_T}{T} = \frac{1}{G_T} \to 0$. Note that the combination of the bias and variance conditions, $\frac{G_T}{T} \to 0$ and $\frac{1}{G_T} \to 0$, requires that the number of groups increase slowly relative to the total sample size. Asymptotically, we would have $\widehat{V}_{G_T} \xrightarrow{p} V_0$ and the standard normal distribution would provide a good approximation to the sampling distribution of the t-statistic under these rate conditions.

It is this heuristic analysis that underlies the traditional use of HAC estimators. The basic argument, of course, carries through to much more general dependence structures and other, non–cluster-based HAC estimators. The approach is useful in many circumstances but is somewhat unsatisfactory. We can see from the previous discussion that, for example, in the clustering case, any sequence of grouping choices where the number of clusters eventually grows large would satisfy the asymptotic conditions.

More insight can be gained by noting that the number of groups chosen corresponds to a choice of how much dependence to account for in estimating V_0, which translates directly into a bias–variance tradeoff for the V_0 estimator. We can see from equation (6) that using a large number of groups allows for very little dependence and may thus produce strongly biased estimates of V_0 even in large samples, while keeping variance low by having few covariance terms. Choosing a smaller number of groups reduces bias by including more covariance terms in the estimator of V_0 at the cost of producing an estimator with more variance, as can be seen from equation (8). Through the choice of weighting function, all HAC estimators have a similar tradeoff with estimators that have most weights close to or equal to zero tending to have high bias and low variance and estimators that set relatively few weights close to or equal to zero tending to have low bias but high variance.

There are many papers that address the choice of weighting structure to trade off these forces in the econometrics literature within the framework

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

illustrated that start from the premise that $\widehat{V}_{G_T} \xrightarrow{p} V_0$. Early work in this area focuses on properties of the HAC estimator itself, for example, Andrews [1991] and Newey and West [1994]. More recent work has focused on making this choice in a way that trades off Type I and Type II errors in testing hypotheses about parameters of interest, for example, Sun, Phillips, and Jin [2008] and Wilhelm [2015].

REMARK 1. ***A Comment on Mechanical Bias When the Number of Clusters Is Small.*** *There are actually two sources of bias inherent in HAC, and therefore clustered standard error, estimation. The first source of bias, which is what we have focused on, results from using an estimator for standard errors that fails to account for the actual dependence in the data. In our simple MA(1) example, this bias is captured in expression* (6). *The other source of bias is mechanical and results from estimation error in residuals. To see this bias, consider again the case where we are interested in estimating the sampling variance of a sample mean using only one cluster. Define estimated residuals* $\hat{\eta}_t = y_t - \bar{y}$. *With* $G_T = 1$ *and using estimated residuals, we would have* $\widehat{V}_{G_T} = \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\widehat{\eta}_s\widehat{\eta}_t = \frac{1}{T}(\sum_{t=1}^{T}\widehat{\eta}_t)^2 = 0$ *because* $\sum_{t=1}^{T}\widehat{\eta}_t = \sum_{t=1}^{T}(y_t - \bar{y}) = 0$. *This mechanical bias is especially pronounced when small numbers of groups are used in clustered variance estimators that may seem disquieting and may be pointed to as a reason to avoid using a small number of clusters. However, while technical, it is also relatively easy to appropriately adjust for this mechanical bias in many cases. Indeed, the familiar "degrees of freedom" adjustments made when estimating sample variances, for example, normalizing by* $N - 1$ *rather than* N *when estimating the usual sample variance, is an example of such an adjustment in a simple setting. More generally, the approaches discussed in sections 3.2–3.4 account for and remove the effect of this mechanical bias. As a practical matter, it is important to recognize that this mechanical bias exists and thus to use a procedure that addresses it; but the existence of this bias is not a reason to be wary of using a small number of groups in clustered variance estimation.*

3.2 INFERENCE BASED ON HIGH VARIANCE HAC ESTIMATION

Looking at the finite sample bias and variance results given in (6) and (8) in section 3.1, we can see a potential issue with an approximation that relies on the HAC estimator, $\widehat{V}_{G_T}$, being arbitrarily close to the true sampling variance, V_0, with high probability. Such an approximation implicitly relies on the estimator having both ignorable bias and ignorable variance in a researcher's given finite sample. However, we can see from (6) and (8) that there will be both bias and variance in any finite sample. It is also clear that keeping bias small leads to the use of a small number of groups, which will result in having a large variance in a given finite sample. Again, we note that the simple heuristic discussion provided in section 3.1 carries over to much more general and realistic settings where, under richer dependence structures, there is even more pressure to use a small number of groups (or a large number of nonnegligible weights in the more general HAC setting)

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to manage bias, which will be associated with large finite-sample variance of the estimator of V_0.

This tension provides the motivation for a different style of approximation, developed in Kiefer and Vogelsang [2002, 2005] and Vogelsang [2003]. These approximations give up on the notion of obtaining a consistent estimator of V_0 and instead focus on HAC estimators that allow for a large amount of dependence, for example, by using a small number of clusters in the case of clustered standard errors, and thus have relatively small bias but potentially large variance. In this case, the asymptotic approximation makes use of an inconsistent standard error estimator. Unlike the example given in section 3.1, the inconsistency in this scenario does not arise from using an estimator that concentrates quickly around the wrong value but rather arises from using an estimator that keeps bias small but has a nonvanishing variance even in large samples. The cleverness of this approach is that the approximation then accounts for a feature that is always present in finite samples: The estimated standard error itself has sampling variation.

Slightly more formally, this approach leverages the insight that it is not necessary to have a consistent estimator of V_0 in order to get a good approximation of the sampling distribution of t-statistics. The chief difficulty in inference is the presence of the unknown scale V_0, and it turns out that inference can be done using an estimator of V_0 with the right scale even when this estimator is quite noisy and properly viewed as inconsistent. Specifically, suppose an estimator $\sqrt{\widehat{V}_{NT}}$ has a distribution that is approximately $\sqrt{V_0}$ times a random variable W, where W does not depend on V_0 or other unknown nuisance parameters; that is, the estimator satisfies an appropriate sense of "unbiasedness."[7] Under this condition, the distribution of the t-statistic is

$$t = \frac{\sqrt{NT}(\bar{y} - \mu_0)}{\sqrt{\widehat{V}_{NT}}} \overset{\text{approx}}{\sim} \frac{\sqrt{V_0}N(0,1)}{\sqrt{V_0}W} = \frac{N(0,1)}{W}. \tag{9}$$

If the ratio $\frac{N(0,1)}{W}$ has a tractable, nuisance-parameter–free distribution, one can do inference using a standard t-statistic formed using an HAC estimator for $\widehat{V}_{NT}$ but with critical values from the distribution of $\frac{N(0,1)}{W}$, which will in general differ from the usual standard normal critical values. The chief complication that arises in this general setting is that the distribution of $\frac{N(0,1)}{W}$ will rarely correspond to a known distribution and appropriate critical values must be obtained by simulation; see, for example, Kiefer and Vogelsang [2002, 2005], Vogelsang [2003], and Bester et al. [2016]. Results in the aforementioned papers, Sun, Phillips, and Jin [2008], and Bester, Conley, and Hansen [2011] show that using the approximation in (9) outperforms inference based on the standard normal approximation across

[7] Note that W will depend on the mechanical bias discussed in section 3.1. By explicitly accounting for W, the approximation (9) will thus directly account for this mechanical bias.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

a wide range of DGPs and settings.[8] This approach is readily extended to handle tests of multiple hypotheses if desired, as shown in the papers referenced previously.

While simulating the distribution in (9) is generally straightforward, Bester, Conley, and Hansen [2011] show that the special structure of (one-way) clustered standard error estimators allows this distribution to be characterized analytically when (i) there are a large number of observations per group, (ii) groups have approximately equal numbers of observations, and (iii) variances of observables are approximately homogeneous across groups. This analytic characterization bypasses the need for simulation. Specifically, Bester, Conley, and Hansen [2011] show that

$$t = \frac{\sqrt{NT}(\bar{y} - \mu_0)}{\sqrt{\widehat{V}_{NT}}} \overset{\text{approx}}{\sim} \frac{\sqrt{V_0}N(0,1)}{\sqrt{V_0}W} = \frac{N(0,1)}{W} = \sqrt{\frac{G}{G-1}}t_{G-1}, \quad (10)$$

where t_{G-1} denotes a Student *t*-distribution with $G - 1$ degrees of freedom and G is the number of clusters used in forming $\widehat{V}_{NT}$. Note that the expression in (10) is equivalent to stating that a rescaled version of the *t*-statistic, $\sqrt{\frac{G-1}{G}}t$, approximately follows the usual Student *t*-distribution with $G - 1$ degrees of freedom. Further note that rescaling the *t*-statistic in this way is equivalent to rescaling the variance estimator as $\frac{G}{G-1}\widehat{V}_{NT}$. This rescaling of the variance is exactly what offsets the mechanical bias discussed in Remark 1.[9] There seems to be little reason not to use these critical values in all cases where clustered standard errors are used. When the number of groups is small, the use of *t*, rather than standard normal, critical values helps account for the additional uncertainty in inference due to noisily estimated standard errors, and the t critical values converge to the usual critical values when the number of groups is large. Of course, when conditions (i)–(iii) stated above are not satisfied, the approximation (10) may not hold and one may wish to use different inferential methods, such as those discussed in section 3.3.

Operationally, the inference procedure for a single parameter from Bester, Conley, and Hansen [2011] outlined above is very similar to usual practice. For example, suppose one were interested in testing a hypothesis at the 5% level. Common practice would be to reject the hypothesis if the absolute value of the *t*-statistic for testing the hypothesis exceeded 1.96. Similarly, common practice for providing a 95% level confidence interval is to

[8] We also provide some evidence of this within our simulation example in results reported in the supplementary appendix to this paper.

[9] It is interesting that this scaling is what is implemented in Stata (Stata Corporation [2013], p. 314) for calculating clustered standard errors with most commands. The exception is for linear models where the calculation in Stata rescales by $\frac{G}{G-1}$ and then makes an additional adjustment that will be negligible in most applications. Confidence intervals reported by Stata after the use of the cluster command also use critical values from a t_{G-1} rather than a standard normal. That is, the confidence intervals returned by Stata implement the approximation in (10).

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

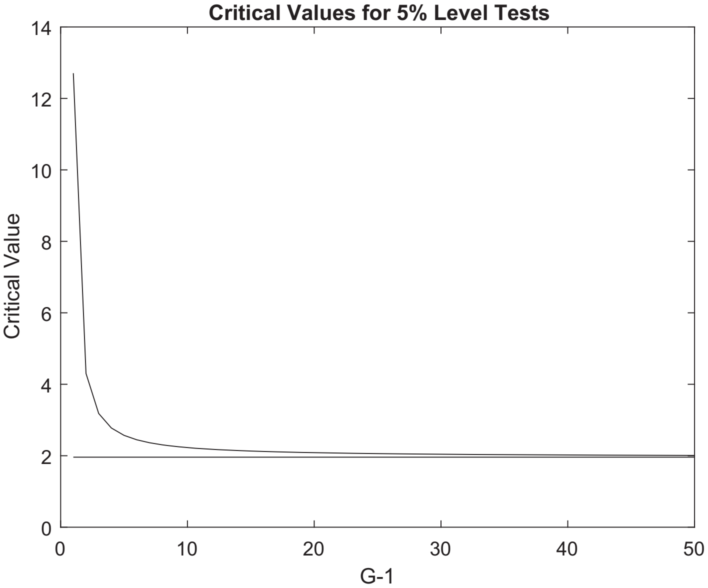

FIG. 1.—Critical values for 5% level tests for differing number of groups used in forming clustered standard errors. The number of clusters used minus one is presented on the axis. The horizontal line is the usual standard normal critical value, provided for reference.

take the point estimate plus and minus 1.96 times the estimated standard error. The procedure outlined above, which accounts for the uncertainty in estimating the standard error, simply replaces the critical value 1.96 with the appropriate 97.5% quantile from a t-distribution with $G - 1$ degrees of freedom. For example, if one used five groups in estimating standard errors, one would reject a hypothesis at the 5% level only if the t-statistic were larger than 2.78 in absolute value and would form a 95% level confidence interval as the point estimate of the parameter of interest plus and minus 2.78 times the estimated standard error.

For small numbers of groups, the critical values from a t-distribution differ sharply from those that are conventionally employed. For reference, critical values for two-sided 5% level tests or 95% level confidence intervals for different numbers of groups are presented in figure 1. The horizontal line in the figure denotes the usual critical value from the standard normal approximation. It is clear from figure 1 that critical values from a t-distribution differ substantively from standard normal critical values for small numbers of groups but converge rapidly as the number of groups increases.

The pattern of critical values illustrates a tradeoff between robustness and the precision with which conclusions can be drawn from dependent data. To be robust to potential strong dependence, a small number of groups is necessary to keep bias small. The cost of keeping bias small is an increase in

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

uncertainty in the standard error estimator itself, which translates into an increase in critical values. The increase in critical values then naturally leads to less precise statements about what can be learned about parameters of interest in a given data set. To the authors' knowledge, there is no reliable data-dependent way to choose the number of clusters at present. However, a heuristic argument suggests that using between 5 and 10 groups—provided these groups have similar sizes and the within-group variance of observables is roughly homogeneous across groups—is likely to be desirable in many applications. From the pattern of critical values in figure 1, we can see that most of the gains to precision of inferential statements in terms of decreased critical values are made by increasing the number of groups to this region. At the same time, using these small numbers of groups should provide robustness to reasonably rich dependence structures. The simulation study in section 4 provides further evidence regarding this choice.

The more refined approximations outlined in this section are relatively easy to implement, either leveraging the analytic results available within the clustering framework or via simulation in the more general setting, and offer considerable improvements relative to approximations that rely on treating standard error estimators as extremely well estimated. However, the results also rely on some crucial assumptions as mentioned above and further elaborated below.

First, the weights used in forming the HAC estimator must have been chosen in a way that results in capturing all important sources of dependence in the data so that the nonmechanical bias in the estimator is small. This condition is true of any approach that aims to deliver valid inference but is worth remembering. In the clustering case, this condition will tend to be more plausible when one uses relatively few groups with large numbers of observations formed by keeping observations that seem likely to be related together.

Some intuition for the dependence condition in the clustering case is that the theoretical results rely on averages within clusters being approximately Gaussian and approximately independent. The approximate Gaussianity of group averages relies on having a reasonably large number of observations within each cluster, and the approximate independence relies on there being only a weak correlation across clusters. Assuming that clusters are formed respecting the dimensions along which dependence truly lies, correlation across clusters can only occur near cluster boundaries. If clusters are large, most observations within a group will be in the interior of the cluster and therefore, as a consequence of weak dependence, approximately independent of observations in other groups. With large clusters, group boundaries are small relative to their interior, and thus the amount of neglected correlation resulting from spillovers across group boundaries will be correspondingly small relative to the correlation captured by the clustered standard error estimator, which accounts for all the correlations among the observations interior to the clusters. Of course, the stronger the correlations across observations, the larger the spillovers across cluster

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

boundaries will be. Keeping these spillovers small relative to the total correlation that is captured by the clustered standard error estimator will then require larger group sizes, and hence smaller numbers of groups, in more strongly dependent cases.

Second, the results depend on the distribution in (9) not depending on any unknown parameters. To achieve this, the formal analysis depends on strong conditions that restrict heterogeneity across observations. In the clustering context, a homogeneity condition across clusters is required to hold. This homogeneity condition requires that, in a linear model, the design matrix needs to be similar across clusters. With a single right-hand-side variable and equally sized groups, this homogeneity would require that, in the limit, the variance of the right-hand-side variable within each cluster be the same across all clusters. This restriction is strong and may be implausible. It will likely be violated when clusters are formed from very different numbers of observations. For example, clusters based on industry classification or countries could easily have very disparate numbers of firms. This condition could be especially unpalatable in some situations in which clusters are formed based on measures of "economic distance." For example, if one thought there were shocks that affected all small firms differentially from how they affected large firms, it would make sense to form clusters by grouping firms on the basis of their size. Of course, one might anticipate that the variability in observables within the firms in the "small firm" cluster would be quite different from firms in the "large firm" cluster, at odds with the homogeneity restriction.

3.3 INFERENCE WITH SAMPLE-SPLITTING

An idea that is closely related to clustering is to base inference on a procedure that splits the sample into the groups that would be defined by the clusters and then estimates the model of interest within each group. A point estimate of the parameter of interest can then be obtained as the average of the within-group estimates, and variation across these estimates can be used to estimate the sampling variance of this estimator. This approach was popularized in Fama and MacBeth [1973] and was rigorously justified under a wide range of structures for the case of inference about scalar hypotheses allowing for a small number of groups in Ibragimov and Müller [2010, 2016]. Canay, Romano, and Shaikh [2017] build on this idea and consider permutation inference using the group-level parameter estimates. Canay, Romano, and Shaikh [2017] may be used to test joint hypotheses and may have advantages relative to Ibragimov and Müller [2010, 2016] at the cost of needing to do permutation inference.

In the following, we illustrate the basic FM procedure in the simple location model used in the previous sections. We refer readers to Canay, Romano, and Shaikh [2017] for discussion of the permutation inference approach. Suppose the data are partitioned into G groups with each observation belonging to one and only one group. We first partition the data into the corresponding G subsets and estimate the sample mean within each

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

subsample. Using $\bar{y}_g$ to denote the sample mean obtained in group g, this process results in a collection of group-specific estimators $\{\bar{y}_g\}_{g=1}^G$. Under weak regularity conditions, each group-specific estimator will be approximately normally distributed; and they will be approximately independent following the same reasoning outlined for cluster estimators in section 3.2 as long as the number of observations within each group is large and the groups were chosen to appropriately capture the dependence in the data.

We can then treat the G group-level estimates, $\{\bar{y}_g\}_{g=1}^G$, as though they were observations from an independent Gaussian location model and use these G "pseudo-observations" to estimate the parameters of the Gaussian distribution and do inference. Following this logic, we can define a point estimator for the overall mean, $\widehat{\mu}$, as

$$\hat{\mu} = \frac{1}{G}\sum_{g=1}^{G} \bar{y}_g.$$

Inference also proceeds in a straightforward manner using the usual estimator of the sampling variance of a sample mean estimated from G observations:

$$S = \left(\frac{1}{G-1} \sum_{g=1}^{G} (\bar{y}_g - \hat{\mu})^2 \right) / G.$$

Ibragimov and Müller [2010] show that inference for μ can proceed under general conditions using the usual t-statistic

$$t = \frac{\hat{\mu} - \mu_0}{S^{1/2}} \tag{11}$$

along with critical values from a t_{G-1} distribution. Importantly, the conditions employed in Ibragimov and Müller [2010] allow quite general heterogeneity across groups. Ibragimov and Müller [2016] and Canay, Romano, and Shaikh [2017] extend these results and similarly allow for general across-group heterogeneity.

The conditions for how the sample should be split for this approach to work well are analogous to those for choosing clusters in section 3.2. First, just as with clustering, the sample splits need to have been chosen in a way that all important sources of correlation have been captured so the bias in the standard error estimator is relatively small. Subsamples must be large enough for their subsample averages to be approximately Gaussian and so that most observations are "interior." Note that, as in section 3.2, it is not required that the subsamples are literally independent; it is only required that the missed sources of correlation are small relative to the overall variance of the estimator (which we often proxy for with the subsample size).

The main benefit of sample-splitting procedures relative to clustering or other HAC-based procedures is that their validity in the case where a small number of groups is used does not rely on strong homogeneity conditions. They remain valid when groups are not approximately equal

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

sized and when observables across groups are heterogeneous. This robustness makes sample-splitting approaches especially appealing in complex settings such as firm-level panel data where heterogeneity across observations and among groups seems likely ex ante.

For the simple FM approach discussed above, this robustness comes at a cost. The reason that the approach discussed in section 3.2 fails under heterogeneity is that the distribution of the t-statistic in (9) depends on unknown objects that are related to the exact heterogeneity in the data. Ibragimov and Müller [2016] show that the structure induced by sample-splitting and aggregating in the manner employed in the FM procedure allows one to prove that the critical values from a t_{G-1} are upper bounds on the critical values from the distribution of the statistic in (11) regardless of the actual nature of heterogeneity.[10] This result means that the procedure outlined above using critical values from a t_{G-1} distribution is valid very generally in the sense that tests will have size that is no worse than the desired level—for example, a 5% level test will reject true null hypotheses *no more than* 5% of the time—and confidence intervals will have at least the desired level of coverage—for example, a 95% confidence interval will cover the true parameter *at least* 95% of the time. Importantly, this conservativeness is different from the potential conservativeness resulting from allowing for too much dependence shared by all procedures that aim to be robust to dependence and is the result of using an upper bound on the critical value that one would obtain from the true distribution of the statistic in (11) under the actual heterogeneity in the data.

The permutation inference approach of Canay, Romano, and Shaikh [2017] is designed to remove the conservativeness from the simple FM procedure by simulating the distribution of (11) using the information in the data. Theoretically, Canay, Romano, and Shaikh [2017] promises to exactly control size and so is not conservative under across-group heterogeneity. The simulation required in the Canay, Romano, and Shaikh [2017] approach presents a small cost. However, the main limitation of this procedure is that, because it relies on a permutation distribution produced from G pseudo-observations, one needs a sufficient number of groups to have any power. For example, one needs at least six groups to have any power for 5% level tests. Thus, it is not well suited to applications where the number of observations per group needs to be quite large due to substantial correlations across observations, resulting in very few groups.

REMARK 2. ***Clustering Versus Sample-Splitting.*** *There is a tight connection between sample-splitting and clustering approaches for inference. In terms of large sample performance, there is relatively little reason to prefer clustering with a small number of groups relative to a sample-splitting approach with the same groups. Inference*

[10] The result actually only holds in the tails of the distribution but covers inference at levels 5% and less, which suffices for most applications.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

from the sample-splitting approach remains valid under weaker technical conditions than are used in establishing validity of inference based on clustered standard errors, and the conservativeness of the sample-splitting approach can be removed by applying Canay, Romano, and Shaikh [2017].

The discussion is more nuanced when one examines finite-sample performance. The chief potential for problems with sample-splitting approaches comes from the fact that only a fraction of the observations are used to estimate each of the group-specific parameters that are then aggregated to form the final estimator of the parameter of interest. In a situation in which each of the underlying estimators is well behaved and approximately unbiased, the sample-splitting simply induces larger variability in the estimators within each subsample that is then averaged out when producing the final estimator. In this case, there is no obvious benefit to using the full sample to obtain point estimates, and the robustness to heterogeneity available through the sample-splitting approaches makes them extremely attractive.

This discussion changes when the point estimates obtained within subsamples are not well behaved due to, for example, being highly variable or having substantial finite sample bias. If the subsamples are small relative to the number of parameters in a model, then estimates within the subsamples may vary wildly, which may substantially degrade the performance of the ultimate estimator. For example, one might want to form groups in a firm-level panel by using SIC codes or some other notion of industrial classification to capture the presence of shocks within industries that affect the firms within-industry differently. The presence of some industries with very few firms may then lead to unstable estimates of parameters of interest. This behavior is illustrated in the simulation example in section 4.

Scenarios with finite sample biases can also result in poorly behaved subsample estimators. Leading examples are estimation of instrumental variables models and estimation of panel data models with fixed effects and predetermined variables such as lagged dependent variables. When there are finite sample biases, the bias in the within-group estimated parameters will correspond to the number of observations per subsample. In general, the bias in each subgroup estimator will be much larger than the bias that would result when using the full sample and will not average out in forming the overall point estimator. For example, consider instrumental variables estimators with a full sample of 500 observations cut into 10 groups of 50 observations each. Intuitively, each subsample estimator would be expected to have bias 10 times the magnitude of the bias in the full sample estimator, and the final point estimator resulting from averaging these 10 biased estimates would also be expected to have bias 10 times the magnitude of the bias of the full sample estimator. The presence of this bias can then result in worse performance than inference based on the full sample estimator with HAC standard errors even in the presence of heterogeneity as illustrated in Bester, Conley, and Hansen [2011]. This bias problem can be reduced by using fewer, larger groups, which provides further motivation for using a small number of large groups beyond the benefit of allowing for substantial dependence among observations.

REMARK 3. ***Aggregate Variables.*** *Having a variable of interest that only varies at some more aggregate level is a fairly common occurrence. As a simple illustration,*

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

suppose the model of interest is

$$y_{ic} = \alpha + \beta x_c + \varepsilon_{ic}, \tag{12}$$

where y_{ic} is an outcome of interest for firm i in country c, x_c is a policy variable that varies only at the country level, and α and β are the model parameters with β being the parameter of interest. If one believed that observations on firms from the same country were dependent but that observations on firms from different countries were independent, one could directly estimate and do inference for β by estimating (12) *by least squares and estimating standard errors by clustering at the country level. Of course, if one split the data by country and attempted to estimate* (12) *using only the within-country subsets of the data, there would be no variation in x_c leaving one unable to identify β.*

Despite this problem, one may still use sample-splitting to estimate parameters on variables that only have variation on some aggregate level. The easiest approach would be to form subsamples that are large enough to have variation in the variable of interest within subsample. For example, one could group collections of geographically similar countries together in the country-level example above. In this case, one could directly apply a sample-splitting procedure using these larger subsamples and would also obtain additional robustness to dependence that spills across country borders but is related to geographic location. A different intuitive, but generally invalid, strategy would be to just form groups based on some other variable that maintains within-group variation in the aggregate variable. For example, one could form groups by industry rather than country assuming industries have presence in multiple countries. Grouping by industry would, however, result in large spillovers across groups under the assumption that firms from the same country are correlated as every group would involve observations from many of the same countries.

It is worth noting that the fundamental problem of having only aggregate-level variation in a variable of interest is that often there is little variation in the data for learning about the effects of such variables. Valid inference will therefore often be imprecise and often rely on imposing strong restrictions. In the context of sample-splitting, one would like within-subsample estimates to be well behaved, as discussed in Remark 2, which may be difficult to achieve in some practical situations. For example, if one has a single cross-section of U.S. data and state-level policy variables, one may be tempted to consider groups formed from, say, the nine U.S. census divisions in order to have a modest number of groups. However, each of these groups will consist of only between three and nine states; and it may seem implausible to maintain that this little state-level variation is sufficient to generate well-behaved within-subsample estimates. In this example, one may prefer to use a much coarser grouping scheme, say splitting into only two groups, and the resulting large critical values if one wishes to use the sample-splitting approach.

3.4 BOOTSTRAP

The bootstrap has been a standard tool of inference since Efron [1979]. The basic idea of the bootstrap is to use simulation to approximate the finite sample behavior of statistics. Bootstrap inference proceeds by

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

generating many "bootstrap data sets," where each of these bootstrap data sets ideally has approximately the same distribution as the original data. One may then compute a statistic of interest within each generated bootstrap data set and use the resulting distribution of the statistic to approximate the sampling distribution of the statistic computed from the original data. If the mechanism used to generate the bootstrap data exactly coincided with the true data-generating mechanism, this bootstrap distribution would be exactly the *finite sample* distribution of the statistic of interest. More generally, the bootstrap distribution will provide a high-quality approximation to the sampling distribution of the statistic when the mechanism used to generate the bootstrap data approximates the true data-generation mechanism and may provide a useful approximation even when the bootstrap data are far from perfect at replicating the true DGP.

Bootstrap inference is potentially valuable for *t*-statistics in our firm-level panel data for at least two reasons.[11] The method inherently accounts for sampling variation in standard errors because the components of test statistics, for example, the point estimate for the regression coefficient and the associated standard error used in forming a *t*-statistic, are recomputed on every simulated data set. Accounting for this sampling variation in standard errors is especially important in dependent data settings such as our motivating firm-level panel data model as has been discussed in sections 3.2 and 3.3. Bootstrapping test statistics is thus similar to the approaches for obtaining reference distributions accounting for standard error sampling variation via high variance HAC estimators, clustering approaches with a small number of large clusters, or sample-splitting FM approaches. Moreover, there is also potential for bootstrap approximations to be more accurate for *t*-statistics than the usual large-sample approximations. Theoretical work has demonstrated that when bootstrap simulations closely approximate the true DGP, bootstrap reference distributions can provide more accurate approximations to sampling distributions than conventional large sample approximations; see Gotze and Kunsch [1996], Hall and Horowitz [1996], Andrews [2002], and Inoue and Shintani [2006]. It is a challenge, of course, to approximate the dependence structure across firms and time in a firm-level panel. However, we may still see benefits in practice from using bootstrap inference even if only part of the correlation structure in the true DGP is captured by our simulated data.

To illustrate the basic approach, let μ_0 denote the true value of a parameter of interest and $\hat{\mu}$ be its estimator using the actual data. Let s denote an estimator of the standard error of $\hat{\mu}$. In a regression example, μ_0 could be the true coefficient on a variable of interest, $\hat{\mu}$ the OLS estimator of this coefficient, and s the estimated standard error of this coefficient based on one-way clustering with a given group structure. We then wish to use

[11] One could also use the bootstrap to generate standard error estimates. We do not pursue this approach as it tends to perform poorly relative to bootstrapping test statistics both theoretically and in practice when analytic standard errors are easy to compute.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the bootstrap to approximate the sampling distribution of the t-statistic, $t = \frac{\hat{\mu}-\mu_0}{s}$.

Our first step is to draw $b = 1, \ldots, B$ "bootstrap samples" that have the same dimensions as the actual data set. We draw these bootstrap samples using a computer simulation from what we will call a bootstrap DGP. There are a variety of options for this bootstrap DGP, which we will discuss below, all of which are easy to simulate and attempt to mimic the true DGP that generated the actual data set. Within each bootstrap sample, we compute $\hat{\mu}_b$ and s_b by applying the same estimator of μ and standard error estimator as used in the actual data to the bootstrap data. In our regression example, the $\hat{\mu}_b$ and s_b are the OLS estimator of the coefficient of interest and the associated clustered standard error computed using the bootstrap data set b.

We then compute a bootstrap t-statistic as $t_b = \frac{\hat{\mu}_b-\hat{\mu}}{s_b}$. Within the bootstrap, the true value of the parameter of interest is (either exactly or approximately) $\hat{\mu}$ under any of the bootstrap DGPs we consider. Therefore, t_b is a draw from the (approximate) distribution of the t-statistic when the null hypothesis is true and data are generated by the bootstrap DGP. The collection $\{t_b\}_{b=1}^{B}$, thus, represents a series of draws from the (approximate) distribution of the test statistic when the null is true, and we can use percentiles of the $\{t_b\}_{b=1}^{B}$ to estimate critical values under the bootstrap DGP. [12] When the bootstrap DGP mimics the true DGP, these bootstrap critical values will be good approximations for the critical values for the true sampling distribution of t. Importantly, these bootstrap critical values will account for the sampling properties of *both* the estimator of the parameter of interest $\hat{\mu}$ and its estimated standard error s because both are recomputed in each new bootstrap sample. We will use the notation F for the true DGP and $\hat{F}$ for a bootstrap DGP.

The key choice in implementing any bootstrap is what to use for the bootstrap DGP, $\hat{F}$, that is, how to generate the bootstrap samples. In order for the bootstrap to improve upon analytic large-sample approximations, the bootstrap needs to generate data that mimic the dependence structure in the true DGP, F, that is relevant for the statistic of interest. If $\hat{F}$ captures all the dependence in F, the bootstrap will work with any statistic of interest. Of course, generating data in a manner that perfectly captures all sources of dependence is a tall order with the complicated dependence structure found in firm-level panels. Fortunately, how much dependence $\hat{F}$ needs to capture depends on the test statistic and perfect replication of dependence structure is unnecessary for many statistics. The demands on $\hat{F}$ can in principle be reduced by choosing a t-statistic with a well-scaled denominator (high variance HAC estimator) just as in equation (10). If the HAC estimator used for the denominator captures the dependence, it will

[12] For example, let $\hat{cv}_\alpha$ be defined as the $1-\alpha$ percentile of $\{|t_b|\}_{b=1}^{B}$. We can perform an α level test by rejecting the null hypothesis whenever $|t| > \hat{cv}_\alpha$.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

effectively "cancel out" the influence of dependence on the distribution of the t-statistic. In fact, Gonçalves and Vogelsang [2011] show that with a perfect choice of HAC estimator for the denominator even using an $\hat{F}$ that generates independent and identically distributed (iid) data, and thus fails to capture any type of dependence, is adequate to approximate the t-statistic's sampling distribution in a time series setting. We anticipate the most relevant case in practice with firm-level panels is that the estimators used in forming the denominators in t-statistics will be imperfect, and it will thus be important for the bootstrap DGP to capture at least some of the dependence in F. In the following two subsections, we outline two different potential panel data bootstrap DGPs.

3.4.1 Time Block Bootstrap. One approach to bootstrapping in a panel data setting is to note that panel data may be viewed as multivariate time series. With time series data, the goal is to produce an approximation $\hat{F}$ that mimics the relevant serial dependence in F.[13] We focus on a popular nonparametric method that generates bootstrap samples by drawing sets of consecutive observations called blocks from the original time series data; see, for example, Carlstein [1986] and Künsch [1989]. To describe this idea, suppose the data consisted of a time series Z_t for $t = 1, 2, \ldots, T$. Blocks could be constructed from the original time series using pairs of consecutive observations, that is, $\{Z_1, Z_2\}, \{Z_2, Z_3\}, \ldots \{Z_{T-1}, Z_T\}$. The jargon for this set of blocks is that they are overlapping and the block size is two. An overlapping blocks bootstrap with block size of two then generates a simulated time series from $\hat{F}$ by taking independent draws with replacement from this set of pairs and stringing them together.[14] Note that one could consider different block sizes by defining blocks over longer sequences of adjacent observations in the obvious way and correspondingly define overlapping blocks bootstraps with different block sizes.[15]

Resampling from blocks of observations results in an $\hat{F}$ in which the bootstrap data exhibit some time series correlation and thus may be able to approximate an F that generates serially correlated data. Specifically, the correlation structure of the actual data is maintained within each block by construction. It follows that using longer blocks of observations allows the

[13] Early work on the bootstrap, Efron [1979], assumed iid data and used iid draws from the empirical distribution of the sample—a uniform distribution with N points of support each equal to the N observed data points—as $\hat{F}$. This is equivalent to resampling with replacement from the original sample of observations. As first remarked by Singh [1981], this resampling scheme will not generally work for dependent data because iid draws from an empirical distribution cannot approximate the dependence present in the true DGP.

[14] When T is even, $T/2$ draws from the set of pairs will be strung together to form the simulated time series. When T is odd a simulated series one period longer than T is simulated and the last point discarded. In STATA the command `bsample` can be used to generate these series via drawing from clusters defined as $\{Z_1, Z_2\}, \{Z_2, Z_3\}$, etc.

[15] The use of overlapping blocks is not crucial. For instance, Carlstein [1986] suggested resampling nonoverlapping blocks of consecutive observations. The use of overlapping blocks is generally advocated due to the larger number of blocks available to be resampled.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

bootstrap data to exhibit more complicated dependence structures, which may improve the bootstrap's ability to approximate the dependence structure in the actual data. The cost of using longer blocks is that, by construction, one is left with fewer blocks from which to sample, which leads to a deterioration in the performance of the bootstrap. In practice, one aims to choose a block size that is long enough to capture the important correlations in the true DGP but is small enough to leave an adequate number of blocks from which to sample.

Block bootstraps can be used with balanced panel data by simply viewing the panel data as a vector time series. To implement a panel version of the overlapping blocks bootstrap, one can construct bootstrap samples as above with Z_t consisting of all firms' data at time t. Gonçalves [2011] presents a formal treatment of this overlapping blocks bootstrap in the context of linear panel data models with cross-sectional and serial dependence of unknown form.[16] The combination of unbalanced panel data where some firms are observed for many fewer time periods than others and common fixed effects structures complicates implementation of an overlapping blocks bootstrap. These complications will likely lead one implementing the overlapping blocks bootstrap in practice to first restrict any unbalanced panel to a balanced subset. Assuming the panel is unbalanced because observations are missing at random, which is a maintained assumption in most commonly applied panel data methods, restricting attention to a balanced subset does not introduce selection bias but clearly potentially leads to efficiency losses as some observations are being discarded. We explore these issues in the simulation example in section 4.

REMARK 4. ***Block Choice and Variance Heterogeneity for Block Bootstrap.*** *The choice of block size is highly related to the choice of groups or sample splits discussed in sections 3.2 and 3.3. Groups discussed in sections 3.2 and 3.3 that include all firms in sets of consecutive time periods are equivalent to nonoverlapping blocks. In order to mimic a true DGP that generates highly serially correlated time series, blocks need to be long (include many time periods) in order for the simulated data to match this dependence pattern. This requirement is analogous to needing large-enough groups so that correlations are mostly within-group and not across-group when using a cluster HAC estimator or FM approach. However, the length of blocks is inherently more limited in the block bootstrap approach because of the need to have enough different blocks for the simulation to give reasonable approximations of draws from the true DGP. Using overlapping blocks rather than nonoverlapping blocks helps but does not remove this limitation. As a result, this type of bootstrap inference will likely perform poorly in applications that exhibit strong serial correlations where it would*

[16] It is common to use time to define blocks due to the widespread presence of serial dependence and difficulty modeling cross-section dependence in panels. However, in principle any partitioning of the data into mutually exclusive and exhaustive groups could be used to define nonoverlapping blocks, which can be independently sampled as a bootstrap DGP; see Cameron and Miller (2015) and MacKinnon and Webb (2017).

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

be desirable to use a very small number, such as two or three, groups in cluster HAC or FM approaches.

The panel block bootstrap is not subject to concerns about cross-sectional heterogeneity because entire cross sections are contained in blocks as the data are split into blocks only along the time dimension. However, it is important for the true DGP for the time series to be stable in a sense over time. For example, block bootstraps would not work well for nonstationary DGPs with growing variances over time. Thus, the block bootstrap shares a drawback with inference approaches based on clustered standard errors estimated using a small number of groups with many observations per group in that neither will work well in scenarios with substantial variance heterogeneity across groups/blocks.

3.4.2 Residual "Wild" Bootstrap. The second bootstrap DGP we consider is better suited to unbalanced panels than block bootstraps. Working with a balanced subset of firm-level panels will be adequate for some applications, but many researchers will want to work with a full, unbalanced panel with all available firms. This motivates our investigation of a method that works well with unbalanced panels called a residual bootstrap DGP, which exploits a regression model and its residuals to generate bootstrap simulations. We focus on a special case of a residual bootstrap called a wild bootstrap, for example, Härdle and Mammen [1993] and Mammen [1993].

Consider the linear regression model

$$y_i = x_i'\beta + \epsilon_i,$$

where x_i and β are $k \times 1$ vectors. The wild bootstrap uses an estimate of β, call it $\hat{\beta}$, and residuals $\hat{\epsilon}_i$ to generate a bootstrap sample that exploits this regression model.[17] For each value of x_i in the actual data, simulated outcomes are generated as

$$y_i^* = x_i'\hat{\beta} + \epsilon_i^*,$$

where ϵ_i^* is generated so that by construction ϵ_i^* has mean zero, variance $\hat{\epsilon_i}^2$, and third moment $\hat{\epsilon_i}^3$. There are many mechanisms available for constructing such an ϵ_i^*. In the simulations in the following section, we use a suggestion from Mammen [1993] that generates a random variable w_i with mean zero, variance one, and third moment one via combining two iid $N(0,1)$ draws, $v_{1,i}$ and $v_{2,i}$, as $w_i = v_{1,i}/\sqrt{2} + (v_{2,i}^2 - 1)/2$. One then generates a draw of ϵ_i^* as

$$\epsilon_i^* = \hat{\epsilon}_i * w_i. \tag{13}$$

We use a simple adaptation of the wild bootstrap described above to allow for dependence within groups/clusters. We use the full vector of residuals

[17] We will use OLS point estimates for $\hat{\beta}$. However, we note that the estimate of β could be constrained to impose the null hypothesis, which could improve performance of the bootstrap, as noted by Djogbenou, MacKinnon, and Nielsen [2017].

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

for all observations within a given cluster in place of $\hat{\epsilon}_i$. That is, we form a single weight for each cluster as above, call it w_g, and then construct the bootstrap residual for each observation i in cluster g as $w_g\hat{\epsilon}_i$. Because each observation in the same cluster shares the same weight, the within-cluster dependence structure is maintained. Note that, because the w_g are generated independently, this bootstrap produces bootstrap data that are independent across groups. This modification of the wild bootstrap is called a cluster wild bootstrap; see Cameron, Gelbach, and Miller [2008] and Djogbenou, MacKinnon, and Nielsen [2017]. Because it holds the observed values of the regressors fixed, the wild bootstrap is well suited to unbalanced firm-level panels, where it implicitly holds fixed the timing of the observations for each firm.

REMARK 5. ***Block Choice and Heterogeneity for Cluster Wild Bootstrap.*** *The choice of cluster structure for the cluster wild bootstrap is essentially identical to the choice of groups or sample splits discussed in sections 3.2 and 3.3. Clusters need to be constructed such that correlations are mostly within-group not across-group. This may result in a researcher having a very limited number of clusters, which, in general, may lead to this method performing poorly. It is possible for cluster wild bootstraps to work well when a homogeneity condition holds across groups. In recent work, Canay, Santos, and Shaikh [2018] show that the cluster wild bootstrap delivers valid inference with a small number of clusters using a strong homogeneity condition and particular choice of weights. Although the homogeneity condition of Canay, Santos, and Shaikh [2018] is weaker than that used in obtaining the few clusters HAC approximation in* (9)*, this version of the cluster wild bootstrap will also struggle with across-group heterogeneity in firm-level panels where complex dependence motivates the use of a few, large clusters.*

4. Simulation Performance

In this section, we present evidence on the performance of several procedures for performing inference in dependent data using simulated data. The simulated data are calibrated to capture the types of variables and dependence that one might encounter in typical accounting or corporate finance data. Because the data we use are simulated, we know the true values of the parameters in the simulation DGP. Thus, we can evaluate the performance of various procedures for conducting inference about these parameters.

We base our simulation on data from Balakrishnan, Core, and Verdi [2014], who investigate how firms' financing and investments are related to reporting quality and how reporting quality is influenced by financing capacity. We focus on a specification in which they look at how a firm's investment is influenced by its collateral value and how this effect varies with its reporting quality. The hypothesized mechanism is that increases in reporting quality reduce information asymmetries, which lower financing frictions. These lower financing frictions then reduce the sensitivity of

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

investment to fluctuations in collateral values. The regression we use is motivated by the Balakrishnan, Core, and Verdi [2014] hypothesis that a change in a firm's collateral value will have a lower impact on its investment when its reporting quality is higher. Under this hypothesis, the effect of a change in a firm's real estate assets is anticipated to be lower for firms with higher reporting quality, corresponding to a negative coefficient on an interaction term between a measure of reporting quality and a measure of collateral value.

It is routine in finance and accounting applications to use panel data like those in Balakrishnan, Core, and Verdi [2014] that contain diverse firms. With such data, making cross-firm comparisons credible often requires using a set of firm characteristics as conditioning information like those used in Balakrishnan, Core, and Verdi [2014]. Which of these firm characteristics is the key variable and which are viewed as conditioning information will vary across applications. Clearly, the variable of interest in other studies may differ from the key variables in Balakrishnan, Core, and Verdi [2014], so it is important to consider how inference methods perform for regression coefficients on a variety of predictors. Therefore, we examine inference performance across all coefficients rather than focus solely on the coefficient of interest from Balakrishnan, Core, and Verdi [2014].

Our simulation data start from a baseline firm-level panel that replicates the final column of table 2 in Balakrishnan, Core, and Verdi [2014], which gives results from the regression of capital expenditure scaled by lagged assets (y_{it}) on a vector of nine variables, x_{it}. These explanatory variables are constructed from the following firm characteristics. *RE_VALUE* is the market value of the firm's real estate assets as of year t scaled by the lagged book value of assets. *STATE_INDEX* measures the growth in real estate prices in the firm's state from 1993 until year t. *FRQ* is one of the Balakrishnan, Core, and Verdi [2014] measures of reporting quality in year $t-1$. *CASH FLOW* is the year t cash flow from operations scaled by the lagged book value of assets. *Q* is the market value of assets in year $t-1$ divided by their book value. *LN_MVE* is the log of market value of equity in year $t-1$. *LN_AGE* is the log of the number of years a firm has a record in Compustat as of year $t-1$. *LEVERAGE* is the sum of short- and long-term debt divided by the book value of assets at year $t-1$. The product of *FRQ* and *RE_VALUE* (*FRQ* × *RE_VALUE*) is also included and is the key regressor for Balakrishnan, Core, and Verdi [2014].

We use data on 21,290 observations spread across 2,159 firms with the number of observations per firm ranging from 3 to 17. With these data, we estimate a standard additive fixed effects model with firm and time effects

$$y_{it} = x_{it}'\beta + \alpha_i + \delta_t + \varepsilon_{it},$$

and then estimate a model for the ε_{it} that allows for a rich, realistic spatial-temporal covariance structure that accommodates correlation not only between firms within a given time period and between time periods within a given firm but also between different firms in different time periods. For

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

example, the model allows that the regression error to firm i in year t is correlated with the shock for some other firm(s) j in year $s \neq t$. We provide specific details about the model for the ε_{it} in an additional supplementary appendix.

We then generate $m = 1, \ldots, 1{,}000$ simulated data sets by using the values of x_{it}, the firm identifiers, and the time identifiers from the actual data coupled with the associated point estimates from the data of their parameters, $\widehat{\beta}$, $\{\widehat{\alpha}_i\}_{i=1}^{n}$, and $\{\widehat{\delta}_t\}_{t=1}^{T}$ to form the linear index $x_{it}'\widehat{\beta} + \widehat{\alpha}_i + \widehat{\delta}_t$.[18] We then generate $y_{it}^{m} = x_{it}'\widehat{\beta} + \widehat{\alpha}_i + \widehat{\delta}_t + \varepsilon_{it}^{m}$, where the ε_{it}^{m} are drawn from a model for ε_{it} estimated using the data. We thus know the true values of the parameters on the covariates x_{it} in the simulated data are $\widehat{\beta}$.

Before turning to the results, we wish to remind the reader that the point of the exercise is to illustrate the performance of different inferential procedures in a variety of settings, not to comment specifically on the particular analysis of Balakrishnan, Core, and Verdi [2014] from which we took the data. Researchers will often have data whose composition and dependence structure is analogous to our data from Balakrishnan, Core, and Verdi [2014], but details will differ across studies so a procedure that performs well across a variety of different stochastic settings will be valuable. We examine how procedures fare in this regard by looking at their behavior across all the different variables used in the original study, which captures reasonable variation in the properties of the underlying data. There are some procedures that perform reasonably well across the different variables and others that do not. Focusing on or drawing attention to the one or two columns that happened to have been of interest in the original paper would present a misleadingly favorable impression of the performance of many of the procedures. We caution the reader against focusing on any small number of columns when trying to extract generalizable information from the results.

We now turn to evaluating the performance of various procedures for conducting inference. We first report performance in terms of size of tests as the reason we are interested in accounting for dependence in making inferential statements is to have these statements accurately reflect the uncertainty with which parameters are estimated. We report results for size of tests based on different inferential procedures in sections 4.1– 4.5. We then turn to point estimation properties (mean-squared error) and power in section 4.6 as, conditional on having reliable assessments of uncertainty,

[18] Our simulation conditions on the set of observed regressors from Balakrishnan, Core, and Verdi [2014] and thus corresponds to a fixed design. The published formal theory for clustered standard errors with a small number of groups in Hansen [2007] and Bester, Conley, and Hansen [2011] that provides convergence to simple, standard limiting distributions does not apply in the fixed design case. These formal results could be extended to the case of a fixed design under a suitable modification of the design homogeneity condition and appropriate modification of technical conditions.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

we would like to use procedures that are as informative as possible about the underlying model parameters.

4.1 CLUSTERING PROCEDURES

We start by considering the performance of various one- and two-way clustering procedures. To perform the evaluation, we first estimate models of the form

$$y_{it} = x_{it}'\beta + FE_{i,t} + \varepsilon_{it} \tag{14}$$

using the simulated data for different fixed effects structures $FE_{i,t}$ discussed later. We then estimate standard errors using a variety of one-way clustering schemes (with G clusters) and two-way clustering schemes (with G_1 clusters in the first dimension and G_2 clusters in the second dimension). When using one-way clustered standard errors, we remove the mechanical bias by rescaling as discussed in section 3.2 and then use critical values from a t_{G-1} distribution as justified in Bester, Conley, and Hansen [2011]. We follow an ad hoc rule-of-thumb when basing inference on two-way clustered standard errors and use critical values from a $t_{\min\{G_1,G_2\}-1}$ distribution.[19]

The key choice in doing inference based on clustered standard errors is how to form the groups of observations that define the clusters. Recall that a central condition underlying the validity of inference based on clustered standard errors is that the contribution of covariances between observations from different clusters is negligible relative to the overall variance of the estimator. Heuristically, this condition requires that clusters are broad enough so that unobservables for most observations within a given cluster are essentially uncorrelated with unobservables of observations from other clusters. Justification of this condition will typically become more plausible as the number of observations within each cluster increases. As examples, take three of the schemes we consider in the simulation experiment: clustering by state, two-way clustering by state and time, and clustering by eight-year time blocks. Note that two-way clustering by state and time allows for more general correlation structures than clustering by state while clustering by eight-year time blocks is neither more nor less general than either other strategy.

When we cluster by state, all observations on all firms in the same state across all time periods are included in the same cluster. This grouping accommodates very general types of within-firm intertemporal correlation in

[19] We also make an ad hoc adjustment to the two-way clustered standard errors to counteract the mechanical bias. Computationally, two-way clustered standard errors based on groups g_1 and g_2 with, respectively, G_1 and G_2 clusters can be computed as $\hat{V}_{g_1} + \hat{V}_{g_2} - \hat{V}_{g_1 \times g_2}$, where $\hat{V}_g$ is the one-way clustered variance estimator based on clusters in g. For our simulation, we rescale each term in this expression and use $\hat{V}_{two\text{-}way} = \frac{G_1}{G_1-1}\hat{V}_{g_1} + \frac{G_2}{G_2-1}\hat{V}_{g_2} - \frac{G_1 G_2}{G_1 G_2-1}\hat{V}_{g_1 \times g_2}$. This rescaling essentially corresponds to what one would obtain if using Stata to compute each term in the formula for two-way clustering with default scaling. Another sensible alternative would be to use $\frac{\min G_1, G_2}{\min G_1, G_2-1}(\hat{V}_{g_1} + \hat{V}_{g_2} - \hat{V}_{g_1 \times g_2})$. It may be useful to further explore finite sample adjustments for use with two-way clustering.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

unobservables as well as correlations in unobservables across firms within the same state both within the same time period and across time periods. Such correlation could be induced, for example, by firms in the same state facing similar time-varying legal environments due to differences in state laws. This structure seems quite general and indeed allows substantial correlation within and across firms, but suppose that there are also industry-specific shocks that affect different industries differently and perhaps interact with the time-varying legal environment. Such shocks would induce correlation in unobservables across all firms in a given industry. If industries were largely concentrated within state boundaries, then clustering by state would allow for such shocks; but if firms from the same industry are spread across many different states, the assumption that spillovers in correlations across cluster boundaries are negligible would likely be violated and lead to a failure of standard errors clustered by state being able to adequately capture the impact of dependence in the data on uncertainty about parameter estimates.

Now suppose we use two-way clustering by state and time. In this case, all observations on all firms in the same state across all time periods are included in the same cluster, and all observations in the same time period are also included. We can thus handle very general types of within-firm intertemporal correlation as well as correlations across firms within the state both within the same time period and across time periods as before, but we are also allowing any firm in the same time period to be correlated to any other firm in the same period. Allowing for contemporaneous correlation among all firms allows us to accommodate industry-specific shocks that affect different industries differently (as well as more general macroeconomic shocks), as long as these shocks are not correlated over time. Such industry shocks would be problematic for one-way clustering by state. However, if industry shocks were intertemporally correlated, they would produce correlation between firms in the same industry across different time periods. Once again, this correlation would generally lead to a failure of two-way clustering by state and time to capture the impact of dependence in the data on uncertainty about parameter estimates unless industries were concentrated within states where intertemporal dependence is allowed.

Finally, consider clustering by eight-year time blocks. In this case, all observations within the first eight years of the sample are included in one cluster, all observations in the next eight years are included in another cluster, and so on. Relative to the two previous strategies, this clustering scheme imposes stronger restrictions on dependence within firms in the same state but much weaker restrictions on cross-sectional correlation. This strategy will allow not only for correlation between firms in the same time period but also for correlation between different firms in different time periods within the eight-year block. Allowing for this additional dependence between different firms in different time periods comes at the cost of needing further restrictions on intertemporal correlation. To see this, note that

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

in the first two strategies, dependence between unobservables between say firm i at time t and firm i at time s are allowed regardless of t and s, while this dependence would be neglected in the last strategy whenever t and s do not belong to the same eight-year block. Neglecting this intertemporal correlation will have relatively little impact when dependence is weak as would be implied, for example, by a first-order autoregressive process with small slope coefficient.

Whether it is more palatable to restrict intertemporal correlations as in the clustering by eight-year time blocks strategy or restrict the spatiotemporal correlations as in two-way clustering by state and time is not obvious and depends on which source of correlation is stronger. It is important to note though that spatial correlations tend to accumulate much faster than temporal correlations due to generally larger cross-sectional sample sizes and that temporal correlations are easier to model and remove via prewhitening methods such as taking first-differences or quasi-differences in the time series.[20] For these reasons, we believe that many researchers may wish to consider structures that allow quite general spatiotemporal dependence as in the clustering by eight-year time block strategy at the cost of needing stronger restrictions on intertemporal correlations.

A further consideration in choosing a cluster structure is that as the number of clusters decreases, clustered standard error estimators of any variety become more variable. To accommodate high variability standard error estimators, the critical values used in conducting inference need to be appropriately set and will generally differ from the usual standard normal cutoffs, as discussed in section 3.2. For one-way clustered estimators, a sensible guide that can be theoretically justified is to use critical values from a t-distribution with degrees of freedom equal to the number of clusters minus one; see, for example, Bester, Conley, and Hansen [2011]. For multiway clustering, a somewhat ad hoc rule-of-thumb motivated by the previous statement is to use critical values from a t-distribution with degrees of freedom equal to the minimum of the number of clusters along each dimension minus one.[21] Looking at our three examples, we have 49 states

[20] For quasi-differencing, one would first need to estimate a low-order autoregressive model. A simple estimation procedure would be to assume that every individual time series follows the same autoregressive model and estimate the coefficients from pooled OLS using residuals from the baseline model, (14) in our simulation example.

[21] As Stata is commonly employed among applied researchers, the authors wish to stress that, to their knowledge, there is no official Stata code at present for implementing multiway clustering, though there are user-provided packages. We also remind the reader that the theory with two-way clustering with a small number of groups is not developed. We suggest caution when trying to use black-boxed multiway clustering unless the number of clusters in all dimensions is large. We also note that one might wish to adopt the bias-correction and degrees of freedom adjustments for clustering estimators suggested in Imbens and Kolesar [2012], which build on ideas from Bell and McCaffrey [2002]. As there are already many moving parts and estimators being considered in this review, we have chosen not to consider these adjustments.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and 17 years represented in the data. For clustering by state, we thus have 49 clusters, two-way clustering by state and year has 49 clusters in the state dimension and 17 in the time dimension, and clustering by eight-year time block has two clusters.[22] Looking at 5% level tests, we would thus use critical values of 2.01, 2.12, and 12.71, respectively, for standard errors clustered by state, two-way clustered by state and time, and clustered by eight-year time block, respectively. Note that with small numbers of clusters, these critical values differ sharply from the usual standard normal critical values and use of the standard normal values, which do not account for estimation error in the standard error estimator itself, could lead to a dramatic overstatement in the precision with which parameters can be determined.

4.1.1 Fixed Effects Choices. A second choice that needs to be made is how to structure the fixed effects. This choice not only should reflect a researcher's beliefs about important sources of unobserved heterogeneity but also needs to be chosen carefully with the type of clustering structure that will be employed in mind. The latter consideration often seems to be ignored in practice. Recall that the within transformation for removing fixed effects induces dependence between observations that belong to the unit defined by the fixed effect. If a fixed effect is defined that crosses cluster boundaries, removing this effect by estimating the unobserved component or equivalently subtracting the group-level mean defined by the fixed effect category will lead to correlation of observations across clusters even if they were originally independent. For example, estimating industry fixed effects would induce correlations across clusters defined by state if industries are present in multiple states and would induce correlations across clusters defined by time blocks.

To further illustrate this point, recall that the error terms for firm i at time t and firm i at time s after removing a firm fixed effect are defined as $\varepsilon_{it} - \bar{\varepsilon}_i$ and $\varepsilon_{is} - \bar{\varepsilon}_i$, where $\bar{\varepsilon}_i = \frac{1}{T}\sum_{\tau=1}^{T} \varepsilon_{i\tau}$. Note that $\varepsilon_{it} - \bar{\varepsilon}_i$ and $\varepsilon_{is} - \bar{\varepsilon}_i$ are correlated generally for all t and s due to the presence of $\bar{\varepsilon}_i$. If one used the eight-year time blocks clustering strategy discussed above, we would then have correlation across clusters by construction. In this case, to eliminate the potential for induced correlation, one could instead define a new fixed effects structure by including a set of firm by eight-year time block fixed effects.[23] This structure nests a model with just firm fixed effects and produces a within transformation that does not spill across groups. Note that this structure also removes any unobserved components at the firm by eight-year time block level, which may also serve to lessen the unobserved dependence in the data and, in cases where one is concerned that such

[22] We form one cluster from the first nine years and the second from the last eight years.

[23] We use "X by Y fixed effects" to denote including a full set of fixed effects for groups formed by taking all possible combinations of X and Y or equivalently for including fixed effects for each category of X, fixed effects for each category of Y, and the full set of interactions between these effects.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

error components may be related to observed explanatory variables, lessen concerns about endogeneity of observables. Finally, note that the problem of inducing correlation across clusters due to fixed effects crossing cluster boundaries is lessened as the number of observations used to estimate the fixed effect becomes large as the induced correlation generally shrinks proportionally with the number of observations used to estimate the fixed effect. Therefore, this issue is much less likely to be a concern for time fixed effects than for firm, industry, or cross-sectional group fixed effects in many firm-level panels due to relatively short time spans versus much larger cross-sectional dimensions.

In our simulation, we consider one-way clustered standard errors estimated with clustering by (1) firm, (2) state, (3) one-digit SIC code, (4) two-digit SIC code, (5) size category, (6) eight-year time block, (7) six-year time block, (8) four-year time block, and (9) two-year time block. We also consider two-way clustering based on (10) firm and year, (11) state and year, (12) one-digit SIC code and year, and (13) two-digit SIC code and year.[24] We chose these different structures to encompass clustering strategies that are common in empirical practice and to illustrate some less common strategies that we thought seemed ex ante plausible for capturing complex dependence structures. This set of clustering approaches is certainly not exhaustive but we think offers a rich enough set to provide a useful exploration of different possibilities in this example.

The fixed effects structures we consider are motivated by the cluster structures defined above. In particular, we consider fixed effects structures that respect cluster boundaries in addition to using just the conventional firm and time fixed effects. For each of the one- and two-way clustering schemes, we consider the use of additive firm and year fixed effects. With clustering by state and two-way clustering by state and year, we also consider a specification with a full set of state by year fixed effects in addition to firm fixed effects. With clustering by one-digit SIC code and two-way clustering by one-digit SIC code and year, we consider two additional fixed effects specifications: one with a full set of one-digit SIC code by year fixed effects in addition to firm fixed effects and the other with a full set of two-digit SIC code by year fixed effects in addition to firm fixed effects. With clustering by two-digit SIC code and two-way clustering by two-digit SIC code and year, we consider an additional specification with a full set of two-digit SIC code by year fixed effects in addition to firm fixed effects. With clustering by size category, we also consider a specification with a full set of size category by year fixed effects along with firm fixed effects. In addition to the additive firm and year effect specification, we also consider a specification with two-year time block by firm fixed effects as well as year fixed effects, a

[24] Two-way clustered covariance matrix estimators are not guaranteed to be positive semidefinite. We use the ad hoc adjustment suggested by Cameron, Gelbach, and Miller [2011] to ensure that covariance matrices estimated by two-way clustering are always positive semidefinite.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

specification with four-year time block by firm fixed effects as well as year fixed effects, a specification with six-year time block by firm fixed effects as well as year fixed effects, and a specification with eight-year time block by firm fixed effects as well as year fixed effects when clustering by two-year time block, four-year time block, six-year time block, and eight-year time block, respectively.

Table 1 contains sizes for 5% level tests based on clustered standard error estimators obtained in the simulated data. Panel A of the table gives results using one-way clustering, and panel B gives results using two-way clustering. The first column of the table indicates the level of clustering, the second column lists the number of clusters minus one, and the third column gives the fixed effects structure. The remaining columns give the size of *t*-tests for the coefficient associated with each of the firm and time varying variables in the data about which one might wish to perform inference. In order to highlight specifications where rejection proportions are close to 5% uniformly across coefficients, we use bold font for rows in which the maximum rejection frequency is 10%—that is, rows in which the maximum distortion of the 5% test is 5 percentage points.

Perhaps the most striking feature in table 1 is the tendency for tests to have poor performance in uniformly controlling size across the full set of covariates. Of all the strategies considered, only two are successful at producing tests in this simulation that keep distortions at smaller than 0.05 uniformly across variables. These are clustering by eight-year time block with firm by eight-year time block fixed effects and year fixed effects and clustering by two-year time block with firm by two-year time block fixed effects and year fixed effects. Note that both of these use fixed effects that respect cluster boundaries. If one wishes to restrict attention to procedures that seem to control size in the sense of keeping the size of 5% level tests near 5% or smaller, only the approach that clusters by two-year time block and takes out firm by two-year time block fixed effects and year fixed effects is successful. Having a procedure that controls size uniformly across the columns in this table is important as, in practice, a researcher generally does not know the stochastic properties of variables of interest ex ante and would thus like to have a procedure that performs well regardless of the identity of the underlying covariate of interest.

It is also interesting to note that none of the two-way clustering approaches perform uniformly well. The difficulty in obtaining good performance for inference based on two-way clustering in finite samples is also highlighted in Villacorta [2015], who focuses on theoretical difficulties in obtaining good behavior in samples where either the cross-sectional or time series dimension is not very large. We also remind the reader of the undesirable feature of usual uses of two-way clustering, which allows observations in unit i at time t to be correlated to observations in unit j at time t but rules out correlation between observations in unit i at time t and unit j at time s for $s \neq t$. This implied correlation structure seems inappropriate for many real-world settings.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1

Size of 5% Level Tests Based on Clustered Standard Errors from Simulation Experiment

Clusters	G−1	Fixed Effects	*RE_VALUE*	*STATE_ INDEX*	*FRQ*	*FRQ* RE_VALUE*	*CASH FLOW*	*Q*	*LN_MVE*	*LEVE- RAGE*	*LN_AGE*
Panel A: One-way clustering											
Firm	2,158	Firm, year	0.12	0.30	0.41	0.22	0.09	0.17	0.27	0.06	0.35
State	48	Firm, year	0.13	0.30	0.42	0.25	0.13	0.20	0.28	0.08	0.39
One-digit SIC	8	Firm, year	0.14	0.27	0.37	0.21	0.15	0.18	0.26	0.09	0.32
Two-digit SIC	53	Firm, year	0.15	0.29	0.38	0.21	0.14	0.20	0.27	0.08	0.34
Size category	4	Firm, year	0.10	0.19	0.27	0.14	0.11	0.13	0.10	0.06	0.17
Eight-year time block	1	Firm, year	0.08	0.06	0.04	0.06	0.06	0.08	0.08	0.05	0.11
Six-year time block	2	Firm, year	0.09	0.09	0.05	0.09	0.06	0.10	0.10	0.07	0.21
Four-year time block	3	Firm, year	0.11	0.10	0.07	0.07	0.06	0.12	0.10	0.09	0.24
Two-year time block	7	Firm, year	0.15	0.20	0.07	0.10	0.07	0.14	0.19	0.10	0.29
State	48	Firm, state × year	0.10	0.10	0.41	0.22	0.12	0.18	0.27	0.07	0.34
One-digit SIC	8	Firm, one-digit SIC × year	0.14	0.24	0.38	0.20	0.13	0.18	0.25	0.08	0.32
Two-digit SIC	53	Firm, one-digit SIC × year	0.14	0.14	0.36	0.18	0.13	0.17	0.24	0.07	0.31
Two-digit SIC	53	Firm, two-digit SIC × year	0.13	0.13	0.37	0.17	0.13	0.17	0.25	0.07	0.29
Size category	4	Firm, size category × year	0.07	0.20	0.29	0.13	0.11	0.11	0.07	0.06	0.14
Eight-year time block	**1**	**Firm × eight-year time block, year**	**0.04**	**0.05**	**0.02**	**0.04**	**0.04**	**0.04**	**0.03**	**0.05**	**0.10**
Six-year time block	2	Firm × six-year time block, year	0.05	0.06	0.05	0.05	0.03	0.09	0.03	0.06	0.22
Four-year time block	3	Firm × four-year time block, year	0.02	0.04	0.02	0.03	0.02	0.05	0.02	0.03	0.13
Two-year time block	**7**	**Firm × two-year time block, year**	**0.01**	**0.03**	**0.01**	**0.02**	**0.01**	**0.02**	**0.01**	**0.02**	**0.04**

(*Continued*)

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1—*Continued*

Clusters	G−1	Fixed Effects	*RE_VALUE*	*STATE_ INDEX*	*FRQ*	*FRQ* RE_VALUE*	*CASH FLOW*	*Q*	*LN_MVE*	*LEVE- RAGE*	*LN_AGE*
Panel B: Two-way clustering											
Firm, year	16	Firm, year	0.08	0.17	0.07	0.09	0.05	0.11	0.18	0.05	0.18
State, year	16	Firm, year	0.10	0.21	0.08	0.11	0.06	0.11	0.21	0.07	0.22
One-digit SIC, year	8	Firm, year	0.11	0.19	0.08	0.10	0.05	0.09	0.16	0.09	0.20
Two-digit SIC, year	16	Firm, year	0.11	0.18	0.07	0.10	0.08	0.11	0.19	0.08	0.20
Firm, year	16	Firm, state × year	0.07	0.04	0.07	0.09	0.05	0.09	0.18	0.05	0.18
State, year	16	Firm, state × year	0.08	0.11	0.08	0.09	0.06	0.10	0.20	0.07	0.21
Firm, year	16	Firm, one-digit SIC × year	0.08	0.15	0.06	0.08	0.05	0.10	0.18	0.05	0.18
One-digit SIC, year	8	Firm, one-digit SIC × year	0.11	0.16	0.07	0.10	0.05	0.08	0.16	0.10	0.21
Two-digit SIC, year	16	Firm, one-digit SIC × year	0.11	0.11	0.05	0.09	0.04	0.07	0.15	0.09	0.20
Firm, year	16	Firm, two-digit SIC × year	0.07	0.07	0.04	0.07	0.05	0.09	0.16	0.04	0.17
Two-digit SIC, year	16	Firm, two-digit SIC × year	0.09	0.10	0.05	0.09	0.06	0.10	0.17	0.06	0.18

Size of 5% level tests obtained from simulation study; 1,000 simulation replications were performed. The simulation standard error for a 5% level test is 0.0069. Panel A shows results based on one-way clustering, and panel B shows results based on two-way clustering. The column "Clusters" gives the level at which clustering occurs, and the column "Fixed Effects" gives the levels of fixed effects that are included in the estimated model. The remaining columns give the names of the firm and time-varying variables in the data whose effects we may be interested in inferring. Critical values are obtained from a t-distribution with $G-1$ degrees of freedom where G is the number of clusters used in forming one-way clustered standard errors or the smaller of the number of clusters along each dimension for two-way clustering. $G-1$ is provided in the column "G−1" for reference. Bold rows indicate that the largest size distortion in that row is 0.05 or less. Further details are provided in the main text, and details about the simulation design are given in a supplementary appendix.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The poor performance of the majority of clustering approaches in this setting can be explained by first noting that the data on which we base the simulation design shows evidence of complicated cross-sectional and intertemporal correlation. The cross-sectional correlation is not perfectly captured by state, industry, or any other single factor on which it would be easy to cluster. There is also intertemporal correlation not just within-firm but across firms in different time periods. Clusters that accommodate this rich structure are necessarily large and, due to the complexity of the cross-sectional patterns, involve forming clusters that keep all cross-sectional observations together and split only on the time dimension, which results at best in a small number of clusters.

The use of a small number of clusters then brings about a second complication. Current approaches to showing good theoretical properties of inference based on clustered standard errors rely either on having a large number of approximately independent clusters or on having a small number of clusters with many observations and a form of within-cluster homogeneity.[25] Specifically, with a small number of clusters and letting X_g denote the elements of the design matrix corresponding to cluster g after partialling out any nuisance variables such as fixed effects, we need $X_g'X_g \approx X_h'X_h$ for all g and h. It is important to note that having $X_g'X_g \approx X_h'X_h$ for all g and h will be hard to satisfy when there are large differences in the numbers of observations across clusters. This type of homogeneity also seems very unlikely in settings where there is substantial heterogeneity among observations and clusters are formed by grouping similar observations together. Carter, Schnepel, and Steigerwald [2013] provide further discussion of these issues and offer a measure of cluster heterogeneity that is meant to capture relevant deviations from the condition $X_g'X_g \approx X_h'X_h$ for all g and h. We note that concerns about cluster heterogeneity would tend to favor forming clusters from time blocks or similar partitions where the number of observations and composition of firms within-cluster can be kept relatively controlled.

A simple, ad hoc device to assess the degree of design heterogeneity across clusters is to regress the squared values of the covariates onto a complete set of cluster dummy variables. One could then look at an F-statistic for testing the null that the squared values of the covariates are not predicted by cluster identity. That is, for each right-hand-side variable of interest x_j, one could estimate the model

$$x_{j,it}^2 = d_{it}'\gamma + v_{j,it},$$

[25] With a large number of clusters, one does not need homogeneity as the heterogeneity "averages out." However, it is not clear what a large number is in practice and the number will depend on the actual extent of heterogeneity with more heterogeneity requiring correspondingly more clusters. See discussion in Carter, Schnepel, and Steigerwald [2013] and Mackinnon and Webb [2016].

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where d_{it} is a complete set of cluster membership dummies and test that $\gamma = 0$. Rejecting $\gamma = 0$ then suggests that cluster identity predicts the value of the squared observable, which indicates that the observable is not homogeneous across clusters. While only valid under restrictive and somewhat unrealistic conditions, such an exercise is still informative about the degree of design heterogeneity in the data.

We have implemented this diagnostic using the covariates from the Balakrishnan, Core, and Verdi [2014] data. Looking at one covariate at a time, the *p*-value associated with testing the null hypothesis of variance homogeneity—that cluster identity does not predict the squared covariate—is smaller than 0.02 across all variables, cluster structures, and fixed effects structures considered with only five exceptions; and in every case, the *p*-value associated with this null is smaller than 0.02 for at least eight of the nine potential variables.[26] Overall, the evidence seems to contradict the hypothesis of cluster homogeneity and suggests that one should be hesitant to trust inference based on clustered standard errors with a small number of groups in these data.

It is also worth noting that if one excludes consideration of the age variable (*LN_AGE*), there are a few additional approaches that control size reasonably. In particular, both rows relating to clustering by eight-year time block do well. The other strategies based on clustering by time blocks are also reasonably effective as long as fixed effects are removed in a way that does not cross cluster boundaries. We note that the age variable is particularly problematic as it is clearly trending within-firm, and one may reasonably wish to exclude inference on coefficients of trending variables, which is known to be problematic. Noting this also helps explain the success of the approach that partials out firm by two-year time block fixed effects as this approach will essentially remove any systematic firm-specific trends from the variables in the model that vary at both the firm level and over time.

4.2 BOOTSTRAP CRITICAL VALUES

Size of tests based on bootstrapped critical values, as described in section 3.4, are provided in tables 2 and 3. The results in table 2 are based on generating bootstrap samples via the cluster wild bootstrap, and the results in table 3 generate bootstrap samples via the overlapping blocks bootstrap.

Table 2 reports size using *t*-statistic critical values obtained via the cluster wild bootstrap following Cameron, Gelbach, and Miller [2008] using the Mammen [1993] weights described in section 3.4. In each simulation replication, we form 2,000 bootstrap samples from which bootstrap *t*-statistics

[26] The exceptions are for *LEVERAGE* with eight-year time block clusters and either fixed effects strategy, *LEVERAGE* with six-year time block clusters and only firm and year fixed effects, *FRQ* × *RE_VALUE* with six-year time block clusters and firm by six-year time block plus year fixed effects, and *LEVERAGE* with four-year time block clusters and firm and year fixed effects. The full set of results is available in the supplementary appendix to this paper.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2

Size of 5% Level Tests Based on Clustered Standard Errors with Wild Bootstrap Critical Values from Simulation Experiment

Clusters	Fixed Effects	*RE_VALUE*	*STATE_ INDEX*	*FRQ*	*FRQ* RE_VALUE*	*CASH FLOW*	*Q*	*LN_MVE*	*LEVERAGE*	*LN_AGE*
Firm	Firm, year	0.14	0.32	0.44	0.25	0.11	0.20	0.29	0.08	0.38
State	Firm, year	0.14	0.31	0.44	0.26	0.15	0.23	0.29	0.09	0.40
One-digit SIC	Firm, year	0.21	0.32	0.41	0.28	0.18	0.24	0.31	0.14	0.34
Two-digit SIC	Firm, year	0.16	0.31	0.40	0.23	0.15	0.20	0.29	0.09	0.35
Size category	Firm, year	0.18	0.29	0.37	0.23	0.18	0.19	0.16	0.12	0.23
Eight-year time block	Firm, year	0.20	0.15	0.10	0.17	0.15	0.24	0.27	0.18	0.24
Six-year time block	Firm, year	0.20	0.18	0.11	0.20	0.12	0.21	0.23	0.20	0.31
Four-year time block	Firm, year	0.18	0.18	0.11	0.14	0.13	0.20	0.20	0.17	0.29
Two-year time block	Firm, year	0.18	0.23	0.08	0.13	0.09	0.17	0.24	0.15	0.27
State	Firm, state × year	0.13	0.16	0.43	0.27	0.14	0.21	0.29	0.10	0.37
One-digit SIC	Firm, one-digit SIC × year	0.21	0.28	0.42	0.25	0.18	0.23	0.31	0.13	0.35
Two-digit SIC	Firm, one-digit SIC × year	0.21	0.19	0.39	0.25	0.18	0.23	0.29	0.13	0.35
Two-digit SIC	Firm, two-digit SIC × year	0.15	0.17	0.39	0.20	0.15	0.18	0.28	0.08	0.32
Size category	Firm, size category × year	0.14	0.30	0.39	0.22	0.19	0.19	0.13	0.12	0.22
Eight-year time block	Firm × eight-year time block, year	0.10	0.10	0.06	0.09	0.06	0.09	0.06	0.08	0.20
Six-year time block	Firm × six-year time block, year	0.13	0.15	0.12	0.13	0.10	0.19	0.11	0.15	0.24
Four-year time block	Firm × four-year time block, year	0.09	0.13	0.09	0.09	0.09	0.13	0.07	0.11	0.17
Two-year time block	Firm × two-year time block, year	0.07	0.17	0.08	0.08	0.08	0.10	0.07	0.10	0.09

Size of 5% level tests obtained from simulation study; 1,000 simulation replications were performed. The simulation standard error for a 5% level test is 0.0069. The column "Clusters" gives the level at which clustering occurs, and the column "Fixed Effects" gives the levels of fixed effects that are included in the estimated model. The remaining columns give the names of the firm and time-varying variables in the data whose effects we may be interested in inferring. Critical values are obtained by applying the cluster wild bootstrap of Cameron, Gelbach, and Miller (2008) using the clusters defined in the column "Clusters." Note that in no row are size distortions uniformly 0.05 or smaller. Further details are provided in the main text, and details about the simulation design are given in a supplementary appendix.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3

Size of 5% Level Tests Based on Clustered Standard Errors with Moving Block Bootstrap Critical Values from Simulation Experiment

Clusters	Fixed Effects	*RE_VALUE*	*STATE_ INDEX*	*FRQ*	*FRQ* RE_VALUE*	*CASH FLOW*	*Q*	*LN_MVE*	*LEVERAGE*	*LN_AGE*
Firm	Firm, year	0.25	0.29	0.21	0.21	0.23	0.26	0.26	0.23	0.26
State	Firm, year	0.24	0.27	0.20	0.20	0.22	0.24	0.25	0.22	0.24
One-digit SIC	Firm, year	0.21	0.24	0.15	0.16	0.16	0.19	0.21	0.19	0.21
Two-digit SIC	Firm, year	0.24	0.28	0.19	0.20	0.19	0.24	0.25	0.22	0.24
Size category	Firm, year	0.20	0.22	0.17	0.18	0.17	0.19	0.19	0.19	0.19
Eight-year time block	**Firm, year**	**0.07**	**0.08**	**0.05**	**0.05**	**0.07**	**0.06**	**0.06**	**0.05**	**0.08**
Six-year time block	Firm, year	0.06	0.07	0.06	0.06	0.06	0.05	0.06	0.06	0.11
Four-year time block	Firm, year	0.08	0.07	0.06	0.05	0.07	0.05	0.07	0.06	0.15
Two-year time block	Firm, year	0.09	0.09	0.05	0.05	0.09	0.07	0.08	0.07	0.14
State	Firm, state × year	0.22	0.20	0.20	0.21	0.20	0.21	0.24	0.20	0.24
One-digit SIC	Firm, one-digit SIC × year	0.21	0.24	0.15	0.17	0.15	0.18	0.22	0.19	0.19
Two-digit SIC	Firm, two-digit SIC × year	0.23	0.24	0.18	0.18	0.17	0.23	0.25	0.22	0.22
Size category	Firm, size category × year	0.19	0.23	0.18	0.18	0.17	0.21	0.17	0.17	0.16
Eight-year time block	**Firm × eight-year time block, year**	**0.04**	**0.04**	**0.04**	**0.04**	**0.05**	**0.04**	**0.03**	**0.04**	**0.05**
Six-year time block	**firm × six-year time block, year**	**0.02**	**0.02**	**0.05**	**0.03**	**0.04**	**0.04**	**0.03**	**0.03**	**0.07**
Four-year time block	**Firm × four-year time block, year**	**0.01**	**0.02**	**0.04**	**0.03**	**0.03**	**0.03**	**0.02**	**0.02**	**0.01**
Two-year time block	**Firm × two-year time block, year**	**0.00**	**0.00**	**0.02**	**0.02**	**0.02**	**0.02**	**0.00**	**0.01**	**0.00**

Size of 5% level tests obtained from simulation study; 1,000 simulation replications were performed. The simulation standard error for a 5% level test is 0.0069. The column "Clusters" gives the level at which clustering occurs for computing standard errors, and the column "Fixed Effects" gives the levels of fixed effects that are included in the estimated model. The remaining columns give the names of the firm and time-varying variables in the data whose effects we may be interested in inferring. Critical values are obtained by applying the time series moving block bootstrap with a block size of three to simulate the distribution of *t*-statistics formed from models with fixed effects as in column "Fixed Effects" and standard errors clustered according to column "Clusters." Bold rows indicate that the largest size distortion in that row is 0.05 or less. Further details are provided in the main text, and details about the simulation design are given in a supplementary appendix.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

are obtained. We then use these 2,000 bootstrap *t*-statistics to estimate the critical value to use in constructing 5% level tests. The first column of table 2 provides the type of clustering used in estimating standard errors in the data and in the bootstrap samples. This column also describes the independence structure induced in the bootstrap data by generation of the wild bootstrap weights. The second column in table 2 reports the fixed effects structure maintained in estimating the model. The remaining columns of table 2 provide rejection proportions for *t*-tests performed using the critical values estimated in the bootstrap simulations.

Overall, the performance of the cluster wild bootstrap is on par or inferior to the clustering approaches described in table 1. The procedure does not control size uniformly across regressors for any specification. Even if one excludes consideration of *LN_AGE*, only one scheme controls size to be at most 10% for a 5% level test: eight-year time block clustering with fixed effects for year and eight-year time block by firm. Given the evidence for substantial across-group heterogeneity for most grouping structures discussed in section 4.1, this performance is in line with results in Mackinnon and Webb [2016] and Canay, Santos, and Shaikh [2018] that suggest the cluster wild bootstrap performs poorly with heterogeneous groups.

Table 3 presents analogous results for the overlapping blocks bootstrap with a block size of three. Due to the difficulty in implementing the overlapping blocks bootstrap in unbalanced panels with complex fixed effects structures, we only consider the performance of the overlapping block bootstraps on a balanced subset of our calibrated firm-level panel where the panel is balanced by dropping all firms that are observed for less than the entire time span of 17 periods. This subset contains approximately 10% of the original firms and 17% of all firm-year observations. We then use only this balanced subset in each simulation replication and apply the overlapping blocks bootstrap to generate 2,000 bootstrap samples. We use these bootstrap samples to bootstrap *t*-statistics from which we estimate critical values. The first two columns in table 3, respectively, report the group structure, which was used to estimate standard errors used in constructing *t*-statistics and the fixed effects structure. The remaining columns report size of 5% level tests.

The overlapping blocks bootstrap results are more promising than those for the cluster wild bootstrap in our simulation. Size is controlled in cases where clustering is by time blocks with fixed effects that do not cross group boundaries and when clustering is by eight-year time block with just firm and year fixed effects included. Results reported in the supplemental appendix suggest that this good performance relative to the cluster wild bootstrap is not driven solely by using the balanced panel subset of firms. We obtain similar results to those in table 2 when the cluster wild bootstrap is applied using the same balanced panel that we use with the overlapping

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

blocks bootstrap.[27] It is worth noting rejection probabilities tend to be less than 5% in cases where size is controlled using critical values obtained by the overlapping blocks bootstrap. We return to this issue in section 4.6.

4.3 SAMPLE-SPLITTING PROCEDURES

In the present context, an FM-style approach proceeds by positing a model with common parameters of exactly the form given in (14). One then obtains estimates of the common parameters of interest in β by following a multistep procedure. First, one estimates a model with group-specific parameters of the form

$$y_{it} = x_{it}'\beta_g + FE_{i,t,g} + \varepsilon_{it}, \tag{15}$$

where $g = 1, \ldots, G$ denotes a group of observations used to estimate $\{\widehat{\beta}_g\}_{g=1}^{G}$. Of course, estimates of the parameters of (15) can be obtained by splitting the data into groups and estimating the parameters of the model separately group by group. One then obtains a point estimate of the common parameter of interest, β, by taking the sample mean of the $\widehat{\beta}_g$:

$$\widehat{\beta} = \frac{1}{G}\sum_{g=1}^{G}\widehat{\beta}_g.$$

Similarly, an estimator of the sampling variation of $\widehat{\beta}$ is given by the usual estimator of the sampling variance of a sample mean estimated from G observations:

$$S = \left(\frac{1}{G-1}\sum_{g=1}^{G}(\widehat{\beta}_g - \widehat{\beta})(\widehat{\beta}_g - \widehat{\beta})'\right)/G.$$

Ibragimov and Müller [2010] show that inference for $\beta(j)$, where $\beta(j)$ is a scalar element of β, about the null hypothesis $H_0 : \beta(j) = \beta_0(j)$ can proceed using the usual t-statistic

$$t_j = \frac{\widehat{\beta(j)} - \beta_0(j)}{S_{j,j}^{1/2}},$$

where $S_{j,j}$ is the jth diagonal element of S along with critical values from a t_{G-1} distribution under reasonably general conditions, though the inference will be conservative when there is heterogeneity across groups. These results are extended to more general problems of testing one-dimensional hypotheses in Ibragimov and Müller [2016]. Canay, Romano, and Shaikh [2017] uses $\widehat{\beta}$ and S in conjunction with a permutation inference procedure to develop a valid inference procedure

[27] We also report additional results for the overlapping blocks bootstrap with block size of two in the supplemental appendix. These results are quite similar to those obtained with the block length of three.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

that allows for general joint hypothesis testing and is not (asymptotically) conservative.

These approaches rely on a similar set of conditions. The key requirement is that the $\widehat{\beta}_g$ are approximately independent across groups, which is motivated by exactly the same considerations as the condition underlying the use of clustered standard errors that covariances between observations across clusters provide an asymptotically negligible contribution to the overall variance of an estimator. The intuition for ensuring this type of independence is identical to the intuition for choosing clusters; and as such, we do not restate that discussion here but refer the reader to the discussion of this point in the preceding section.

A second condition underlying the validity of FM-style procedures is that there are many weakly dependent observations within each group. In practice, this requirement suggests that one will wish to use a small number of groups that each consist of many observations. Note that this differs somewhat from the use of clustered standard errors, which would allow for a large number of small clusters under the assumption that observations in these clusters are approximately independent. The important benefit provided by sample-splitting is that the homogeneity assumption required for validity of clustered standard errors with a small number of clusters may be dropped. Dropping this homogeneity requirement broadens the set of applications under which sample-splitting estimators will provide reliable inference and allows, for example, for settings where there are quite different numbers of observations per group or where there is substantial variability in observables across groups as found in the actual Balakrishnan, Core, and Verdi [2014] data and discussed above.

We also note that the choice of group structure interacts with the fixed effects structure maintained just as the choice of grouping structure does when clustered standard errors are used. Specifically, the key assumption of approximate independence of $\widehat{\beta}_g$ across groups suggests that fixed effects should not spill across group boundaries. Keeping the fixed effects from spilling over can readily be accomplished by splitting the data into the desired clusters and then estimating the desired model. For example, if one wished to allow firm-specific effects and thought an appropriate grouping scheme would form groups by taking groups as two-year time blocks, estimation could proceed by first splitting the data into two-year time blocks and then running separate regressions that include firm-specific effects within each two-year time block. Note that this is equivalent to including a full set of firm by two-year time block fixed effects in the original model as $FE_{i,t,g}$.

We illustrate the inferential properties of FM-style procedures by applying them in our simulated data. We assume a baseline model given by

$$y_{it} = x_{it}'\beta + \alpha_i + \delta_t + \varepsilon_{it}, \tag{16}$$

where α_i and δ_t are, respectively, firm and time unobserved heterogeneity, which is allowed to depend arbitrarily on observed variables and β are parameters of interest as in the preceding section. As with the application

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of clustered standard errors, we consider a variety of grouping (sample-splitting) schemes. Specifically, we consider groups formed by (1) state, (2) one-digit SIC code, (3) firm size categories, (4) eight-year time blocks, (5) six-year time blocks, (6) four-year time blocks, and (7) two-year time blocks. When using groups based on state, one-digit SIC code, and two-year time blocks, we also consider the permutation inference procedure of Canay, Romano, and Shaikh [2017].[28] Note that doing FM in this model with different group structures effectively corresponds to different sets of fixed effects.[29]

We provide simulation results based on FM in table 4. The results shown in table 4 are strikingly different from those in tables 1–3 in that most of the considered schemes appear to control size relatively well for tests regarding the coefficient on all variables. In particular, all procedures based on splitting in the time series control size across all coefficients. This good performance is likely due to the fact that the procedures that group on time blocks allow for very flexible patterns of cross-sectional/spatial correlation and weak spatio-temporal dependence. The presence of firm by time block fixed effects serves to effectively remove any low or moderate frequency time-varying firm-specific effects that might lead to stronger intertemporal dependence. Finally, good inferential properties are retained despite the strong heterogeneity across groups briefly described in the previous section due to the robustness of sample-splitting procedures to heterogeneity in both observables and errors demonstrated in Ibragimov and Müller [2010, 2016] and Canay, Romano, and Shaikh [2017].

Before concluding this section, it is important to note that the apparent size control of FM grouping on state or one-digit SIC code is a fortuitous coincidence in that neither grouping by state nor grouping by one-digit SIC code adequately captures the dependence in the data. Evidence for this is provided in the modest size distortions in the permutation inference procedure of Canay, Romano, and Shaikh [2017], which is correctly sized, not conservative, in the presence of heterogeneity across groups. However, with the FM procedure, the impact on specifying groups that fail to capture the correlation structure is essentially offset by the conservativeness of the FM procedure due to heterogeneity in the groups. This behavior is likely peculiar to this specific example, and one should not infer that these two

[28] Note that the baseline procedure in Canay, Romano, and Shaikh [2017] is based on a permutation distribution that will have 2^G points of support and thus a minimum p-value of $1/2^{G-1}$. We thus need at least six groups to test at the 5% level if we wish the result of the test to be informed by the data.

[29] Specifically, grouping by state, by one-digit SIC code, by firm size category, by two-year time block, by four-year time block, by six-year time block, and by eight-year time block are effectively including firm effects and a full set of state by year effects, firm effects and a full set of one-digit SIC code by year effects, firm effects and a full set of size category by year effects, year effects and a full set of firm by two-year time block effects, year effects and a full set of firm by four-year time block effects, year effects and a full set of firm by six-year time block effects, and year effects and a full set of firm by eight-year time block effects, respectively.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4

Size of 5% Level Tests Based on Sample-Splitting from Simulation Experiment

Groups	G−1	Effective Fixed Effects	*RE_VALUE*	*STATE_ INDEX*	*FRQ*	*FRQ* RE_VALUE*	*CASH FLOW*	*Q*	*LN_MVE*	*LEVERAGE*	*LN_AGE*
Panel A: Fama–MacBeth											
State	**48**	**Firm, state × year**	**0.04**	**0.05**	**0.08**	**0.02**	**0.02**	**0.03**	**0.06**	**0.05**	**0.07**
One-digit SIC	**8**	**Firm, one-digit SIC × year**	**0.02**	**0.03**	**0.07**	**0.03**	**0.01**	**0.02**	**0.02**	**0.02**	**0.03**
Size category	4	Firm, size category × year	0.06	0.19	0.29	0.15	0.06	0.12	0.07	0.06	0.14
Eight-year time block	**1**	**Firm × eight-year time block, year**	**0.05**	**0.01**	**0.03**	**0.04**	**0.04**	**0.04**	**0.04**	**0.05**	**0.02**
Six-year time block	**2**	**Firm × six-year time block, year**	**0.05**	**0.02**	**0.04**	**0.05**	**0.05**	**0.06**	**0.03**	**0.06**	**0.02**
Four-year time block	**3**	**Firm × four-year time block, year**	**0.04**	**0.01**	**0.03**	**0.04**	**0.03**	**0.06**	**0.03**	**0.05**	**0.02**
Two-year time block	**7**	**Firm × two-year time block, year**	**0.04**	**0.02**	**0.04**	**0.05**	**0.05**	**0.05**	**0.03**	**0.06**	**0.03**
Panel B: Canay, Romano, Shaikh											
State	-	Firm, state × year	0.07	0.06	0.11	0.04	0.06	0.06	0.06	0.06	0.10
One-digit SIC	-	Firm, one-digit SIC × year	0.08	0.07	0.11	0.09	0.06	0.06	0.07	0.05	0.08
Two-year time block	**-**	**firm × two-year time block, year**	**0.04**	**0.05**	**0.04**	**0.05**	**0.06**	**0.05**	**0.05**	**0.06**	**0.03**

Size of 5% level tests obtained from simulation study; 1,000 simulation replications were performed. The simulation standard error for a 5% level test is 0.0069. Panel A shows results based on FM using critical values from a *t*-distribution with $G-1$ degrees of freedom, where G is the number of groups. $G-1$ is provided in the column "$G-1$" for reference. Panel B shows results based on the permutation inference procedure of Canay, Romano, and Shaikh (2014). The column "Groups" gives the level at which sample splitting occurs, and the column "Effective Fixed Effects" gives the levels of fixed effects that are included in the estimated model. The remaining columns give the names of the firm and time-varying variables in the data whose effects we may be interested in inferring. Bold rows indicate that the largest size distortion in that row is 0.05 or less. Further details are provided in the main text, and details about the simulation design are given in a supplementary appendix.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

forces will offset in different settings with groups specified in a way that does not adequately account for sources of unobserved residual dependence.

4.4 SENSITIVITY ANALYSIS

In the preceding sections, we have tried to provide intuition into the key features underlying group choice for use with clustered standard error estimators, FM estimators, and the bootstrap. The key in all cases is that groups are chosen in such a way as to capture important directions of correlation among observations so that dependence between observations that belong to different groups makes only a small contribution to the total sampling variation of the estimator of interest. In practice, we believe that this requirement suggests that one will typically wish to use a small number of quite large groups when doing inference to minimize the potential for neglected sources of correlation and spillovers across group boundaries.

While the advice to use a small number of groups that consist of many observations and capture important directions of dependence is sensible, it is hard to know how to choose such groups in practice. All of the grouping strategies considered in our simulation experiment, with the possible exception of grouping by state, rely on constructing a small number of groups made up of many observations. We also see that inferential performance, as measured by size of tests, depends nontrivially on the grouping structure one assumes when conducting inference. This dependence means that the choice of grouping structure can have important practical implications.

One simple possibility to group selection for inference is to not choose just one structure but rather to report inferential results based on several different sensible group structures. This type of sensitivity analysis is analogous to the usual procedure in empirical research of trying many different specifications to assess the sensitivity of conclusions about the parameter of interest to model specification. One might also worry about the strength of a conclusion that depends sensitively on which group structure is used when conducting inference. One could formalize this to a requirement that a finding not be deemed significant unless it is significant across each of a variety of grouping structures. Such a procedure would equate to using the union of confidence intervals across different schemes as a confidence region for the parameter of interest and would thus be conservative by construction under the assumption that at least one of the structures considered adequately captured the dependence in the data.

We consider such a procedure in the row labeled “Sensitivity” in table 5. To obtain these results, we consider inference based on the conventional clustered standard error estimator but consider clustering based on firm, state, one-digit SIC code, size category, or four-year time block. With clustering based on four-year time block, we include a full set of firm by four-year time block fixed effects and year fixed effects. In all other cases, we use just firm fixed effects and year fixed effects. We then reject a null hypothesis only if the hypothesis would be rejected based on each of the five

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5

Size of 5% Level Tests Based on Sensitivity Analysis and Ad Hoc Group Selection

Procedure	*RE_VALUE*	*STATE_INDEX*	*FRQ*	*FRQ** *RE_VALUE*	*CASH FLOW*	*Q*	*LN_MVE*	*LEVERAGE*	*LN_AGE*
Sensitivity analysis	**0.00**	**0.01**	**0.01**	**0.00**	**0.00**	**0.01**	**0.00**	**0.00**	**0.02**
Ad hoc group selection	0.04	0.02	0.05	0.04	0.05	0.06	0.03	0.06	0.02

Size of 5% level tests obtained from simulation study; 1,000 simulation replications were performed. The simulation standard error for a 5% level test is 0.0069. "Sensitivity analysis" gives size of tests based on estimating clustered standard errors under several different clustering structures and rejecting only if the hypothesis is rejected across all structures. "Ad hoc group selection" gives size of tests based on an ad hoc data-dependent group selection procedure outlined in the main text. The remaining columns give the names of the firm and time-varying variables in the data whose effects we may be interested in inferring. Bold rows indicate that the largest size distortion in that row is 0.05 or less. Further details are provided in the main text, and details about the simulation design are given in a supplementary appendix.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

procedures. In the simulation, this procedure does indeed control size in the sense of producing tests with rejection rates less than the nominal level of 5% across all variables considered. However, the test is quite conservative as expected, with size uniformly substantially below the nominal level.

4.5 AD HOC GROUP SELECTION

Rather than conducting sensitivity analysis or simply choosing some a priori plausible structure, one may wish to use the data to try to infer an appropriate clustering structure. Traditionally, clustering and grouped data inference procedures were advocated in settings where assuming independence across cross-sectional sampling units and thus forming clusters at the unit of observation seems natural; see Liang and Zeger [1986] and Arellano [1987]. Fama and MacBeth [1973] were also chiefly interested in returns regressions where the assumption of a lack of intertemporal correlation is natural and thus forming clusters by grouping by time provides a natural grouping with roughly independent groups. In these settings, the choice of grouping structure is thus a nonissue and little attention was paid to the choice of grouping strategy. The more recent literature such as Vogelsang [2012], Ibragimov and Müller [2010], Bester, Conley, and Hansen [2011], Ibragimov and Müller [2016], and Canay, Romano, and Shaikh [2017] has advocated the use of grouped data-based inference strategies in more general dependent data settings, but little work has been done on providing serious, data-dependent choice of group structure in these more general settings. One exception is Ibragimov and Müller [2016], which provides a testing procedure that can be used to test the adequacy of one grouping structure with a larger number of clusters relative to another with a smaller number of clusters. While useful, this procedure does require maintaining the hypothesis that the grouping structure with a small number of clusters is appropriate and allows only for the comparison of this maintained structure against one alternative with a large number of groups. Exploring data-dependent group choice is definitely an area that calls for further additional research.

As a concrete suggestion, we consider an ad hoc data-dependent procedure for choosing a grouping strategy. We first note that spatial dependence among observations in many contexts, such as firm-level data, may operate along many dimensions simultaneously. Thinking about all the potential directions along which spatial correlation can operate and forming groups that split the data along cross-sectional dimensions that appropriately capture this dependence is thus challenging. It is, however, much easier to think about and model time series dependence. We thus choose to form groups that split only along the time series dimension. This treatment mirrors the original motivation for the procedure in Fama and MacBeth [1973] and also underlies the approaches for inference in spatially and temporally correlated panels considered in Driscoll and Kraay [1998] and Vogelsang [2012]. Unlike Fama and MacBeth [1973], whose focus on returns motivates an assumption of independence across time, we do not consider

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

forming groups from individual cross sections as intertemporal independence seems implausible in many accounting and finance contexts. We also do not wish to use large T approximations as in Driscoll and Kraay [1998] and Vogelsang [2012]. Rather, we wish to use the intertemporal dependence structure in the data to guide our choice of groups.

For our ad hoc group selection procedure, we first estimate the linear model (16) by OLS and take the residuals from this regression. We then use these residuals to form the score for each covariate by multiplying the within-transformed covariates to these estimated residuals.[30] With nine right-hand-side variables in the model, this operation gives us nine score vectors, one for each variable. We then collapse each of these vectors to a time series by taking the within-time-period means of each vector, which results in a $T \times 1$ vector of within-time-period means for each covariate. We treat these $T \times 1$ vectors as independent time series, and estimate the first three autocorrelations for each series. We then use these autocorrelation estimates to benchmark the strength of the time series correlation by looking at 5% level tests for the autocorrelations being equal to 0 assuming iid sampling of the time series.[31] We then take the maximum lag length at which the hypothesis is rejected across all variables. If this maximum is one or the hypothesis is never rejected, we group by two-year time blocks. If the maximum is two, we group by four-year time blocks; and we group by six-year time blocks if the maximum is three.[32] We use FM-style inference as design heterogeneity seems likely in this application.

Results based on this data-dependent procedure are reported in the row “ad hoc group selection” in table 5. As we limit ourselves to a small number of grouping schemes, which all allow for quite general cross-sectional/spatial correlation and some weak time series dependence, it is not surprising that the resulting inference seems to do a good job controlling size across all coefficients. Such an approach would struggle in cases with stronger time series dependence, though a setting with strong temporal correlation and general spatial correlation would be challenging for any procedure. Finally, we note that the ability to form groups by splitting only in the time series and keeping all cross-sectional observations in the same time period together is predicated on having a long enough time series that one may form at least a few clusters by splitting only on the time series. If one is faced with a short panel, forming groups by splitting along the cross-section seems to be necessary to generate enough groups to produce useful inferential statements.

[30] With time and firm fixed effects, the within-transformed x_{it} is $\tilde{x}_{it} = x_{it} - \bar{x}_i - \bar{x}_t + \bar{x}$, where $\bar{x}_i$ is the within firm mean of x, $\bar{x}_t$ is the within-time-period mean of x, and $\bar{x}$ is the overall mean of x.

[31] We ignore any concerns about incidental parameters bias, neglected dependence, and other finite-sample problems as the procedure is ad hoc and is only meant to get a ballpark idea of the strength of time series correlation.

[32] This rule results in grouping by two-, four-, and six-year time blocks in 254, 294, and 452 simulation replications, respectively.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.6 POWER AND POINT ESTIMATION PROPERTIES

In sections 4.1–4.5, we have considered the performance of a variety of grouped data-based estimation and inference strategies for learning about model parameters in a firm-level panel with performance measured by size of 5% level tests. While size of tests or coverage of confidence intervals is a primary concern when thinking about statistical inference, it is also important to consider power and efficiency. In this section, we report results on efficiency of point estimators underlying the procedures examined in the previous sections and report results on power of tests focusing on only the procedures that controlled size in the simulation.

The use of different fixed effects structures means that the point estimator of the regression coefficients underlying the results in table 1 differs across some of the entries, and the point estimator obtained from applying an FM-type procedure is not equivalent to the OLS estimator of the parameters β in (15) even when the fixed effects structures across the procedures align. The point estimators underlying the overlapping blocks bootstrap results also differ from those obtained in the full sample as we use only the balanced subset of firms with data available for all periods when implementing the overlapping blocks bootstrap. It is thus instructive to consider how the use of these different estimators impacts point estimation properties. In table 6, we report root mean squared error (RMSE) of the OLS point estimator of each parameter obtained under each different fixed effects structure used in table 1, of the FM point estimator of each parameter under each sample-splitting scheme used in table 2, and of the OLS point estimator of each parameter obtained under each different fixed effects structure using only the subset of the data used in the overlapping blocks bootstrap. We note that all of the estimators are unbiased in this example because the covariates are strictly exogenous, so RMSE is dominated by the estimators' variance.

The first thing to note from the RMSE results is that there are a few point estimators that appear to be uniformly dominated. The FM-style procedure grouping by state or one-digit SIC code performs very poorly relative to the other estimators considered. This poor performance seems to be driven by using very heterogeneous groups, some of which are very small, in both cases. The presence of very small groups leads to group-specific estimators with high sampling variability, which significantly degrades the estimation performance of the overall procedure; see the discussion in Remark 2 in section 3.3. We also see that the performance of estimators using the largest balanced subset of the data presented in panel B of the table is dominated by OLS using the full sample or any of the FM estimators excepting the two previously mentioned. This result is unsurprising as a substantial fraction of the data is discarded in forming the underlying balanced subset.

Excluding the few clearly dominated procedures, the RMSE results are not clear cut. We can see that, within this simulation design, there is no single procedure that performs best across all the different variables. The

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6

Simulation Root Mean Squared Error

	RE_VALUE	*STATE_INDEX*	*FRQ*	*FRQ∗ RE_VALUE*	*CASH FLOW*	*Q*	*LN_MVE*	*LEVERAGE*	*LN_AGE*
Panel A: OLS on full sample									
Fixed Effects Structure									
Firm, year	0.351	0.348	0.135	0.158	0.046	0.072	0.098	0.320	0.437
Firm, state × year	0.338	0.870	0.134	0.158	0.046	0.070	0.099	0.327	0.433
Firm, one-digit SIC × year	0.346	0.317	0.134	0.153	0.045	0.071	0.096	0.318	0.422
Firm, two-digit SIC × year	0.328	0.249	0.129	0.148	0.045	0.068	0.097	0.320	0.405
Firm, size category × year	0.322	0.341	0.146	0.157	0.045	0.067	0.069	0.313	0.318
Firm × eight-year time block, year	0.383	0.272	0.155	0.158	0.048	0.063	0.085	0.343	0.558
Firm × six-year time block, year	0.483	0.397	0.161	0.167	0.048	0.071	0.088	0.381	0.668
Firm × four-year time block, year	0.439	0.347	0.149	0.148	0.051	0.069	0.089	0.386	0.761
Firm × two-year time block, year	0.607	0.453	0.133	0.171	0.055	0.084	0.100	0.517	1.308
Panel B: OLS on balanced subsample									
Fixed Effects Structure									
Firm, year	0.571	0.425	0.180	0.217	0.129	0.125	0.180	0.734	0.734
Firm, state × year	0.591	2.126	0.181	0.232	0.134	0.127	0.186	0.778	0.812
Firm, one-digit SIC × year	0.575	0.418	0.180	0.217	0.126	0.124	0.180	0.730	0.747
Firm, two-digit SIC × year	0.605	0.413	0.180	0.227	0.124	0.124	0.188	0.741	0.773
Firm, size category × year	0.558	0.424	0.185	0.216	0.127	0.125	0.152	0.703	0.620
Firm × eight-year time block, year	0.715	0.471	0.190	0.237	0.127	0.129	0.175	0.778	0.974
Firm × six-year time block, year	0.816	0.556	0.200	0.249	0.126	0.135	0.195	0.851	1.206
Firm × four-year time block, year	0.838	0.557	0.185	0.252	0.130	0.135	0.193	0.893	1.561
Firm × two-year time block, year	1.065	0.762	0.206	0.311	0.155	0.164	0.215	1.159	2.789

(*Continued*)

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6—*Continued*

	RE_VALUE	*STATE_INDEX*	*FRQ*	*FRQ∗ RE_VALUE*	*CASH FLOW*	*Q*	*LN_MVE*	*LEVERAGE*	*LN_AGE*
Panel C: Sample-splitting on full sample									
Groups									
State	1.493	5.996	0.268	1.355	0.285	0.327	0.226	0.808	1.099
One-digit SIC	2.198	0.986	0.250	0.735	0.309	0.236	0.318	1.783	1.246
Size category	0.326	0.334	0.144	0.167	0.057	0.067	0.070	0.313	0.317
Eight-year time block	0.419	0.912	0.130	0.161	0.049	0.061	0.083	0.335	0.639
Six-year time block	0.464	0.781	0.110	0.146	0.047	0.064	0.084	0.370	0.822
Four-year time block	0.513	1.084	0.114	0.151	0.054	0.066	0.101	0.411	0.808
Two-year time block	0.749	1.734	0.118	0.176	0.065	0.078	0.119	0.570	1.256
Ad hoc group selection	0.568	1.159	0.114	0.159	0.054	0.069	0.097	0.437	0.932

Root mean squared error obtained from simulation study; 1,000 simulation replications were performed. Panel A shows results based on the standard OLS regression of the linear fixed effects model using the fixed effects structure specified in the first column of the table. Panel B shows results based on the standard OLS regression of the linear fixed effects model using the fixed effects structure specified in the first column of the table using only the data from the largest balanced subset of the panel. The results in panel B are for the point estimator corresponding to the moving block bootstrap, which uses this subsample. Panel C shows results based on FM with splits specified in the first column of the table. The row labeled "ad hoc group selection" in panel C is based on the adaptive procedure outlined in section 3.3 of the text. The remaining columns give the names of the firm and time-varying variables in the data whose effects we may be interested in inferring. Further details are provided in the main text, and details about the simulation design are given in a supplementary appendix.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 7

Power for Coefficient on FRQ∗RE_VALUE from Simulation

			k								
Clusters	$G-1$	Fixed Effects	−4	−3	−2	−1	0	1	2	3	4
Panel A: Clustered standard errors with *t*-critical value											
Eight-year time block	1	Firm × eight-year time block, year	0.26	0.18	0.12	0.07	0.04	0.07	0.13	0.20	0.26
Two-year time block	7	Firm × two-year time block, year	0.68	0.43	0.18	0.04	0.02	0.06	0.18	0.45	0.68
Panel B: Moving blocks boostrap											
Eight-year time block	–	Firm, year	0.20	0.15	0.10	0.06	0.05	0.07	0.10	0.15	0.21
Eight-year time block	–	Firm × eight-year time block, year	0.14	0.10	0.07	0.04	0.04	0.05	0.07	0.11	0.13
Six-year time block	–	Firm × six-year time block, year	0.24	0.14	0.08	0.04	0.03	0.06	0.10	0.17	0.26
Four-year time block	–	Firm × four-year time block, year	0.34	0.22	0.13	0.05	0.03	0.05	0.14	0.23	0.34
Two-year time block	–	Firm × two-year time block, year	0.27	0.16	0.07	0.03	0.02	0.04	0.09	0.18	0.30
Panel C: Sample-splitting											
State	48	Firm, state × year	0.08	0.06	0.03	0.02	0.02	0.03	0.04	0.06	0.09
One-digit SIC	8	Firm, one-digit SIC × year	0.27	0.19	0.10	0.04	0.03	0.05	0.11	0.19	0.28
Eight-year time block	1	Firm × eight-year time block, year	0.26	0.19	0.13	0.07	0.04	0.08	0.14	0.20	0.26
Six-year time block	2	Firm × six-year time block, year	0.63	0.44	0.25	0.10	0.05	0.12	0.27	0.46	0.65
Four-year time block	3	Firm × four-year time block, year	0.77	0.56	0.31	0.11	0.04	0.11	0.34	0.58	0.80
Two-year time block	7	Firm × two-year time block, year	0.85	0.65	0.36	0.13	0.04	0.12	0.37	0.67	0.86
Panel D: Canay, Romano, and Shaikh											
Two-year time block	–	Firm × two-year time block, year	0.83	0.63	0.35	0.13	0.05	0.13	0.38	0.65	0.85
Panel E: Sensitivity analysis and ad hoc group selection											
Sensitivity analysis	–	–	0.73	0.47	0.19	0.03	0.00	0.04	0.22	0.48	0.74
Ad hoc group selection	–	–	0.70	0.50	0.28	0.09	0.04	0.12	0.30	0.52	0.71

Power of 5% level tests obtained from simulation study. Power is against the alternative that the parameter value is equal to $\beta_{0,FRQ*RE\_VALUE} + ks_{FRQ*RE\_VALUE}$, where k is given in the column labels, $\beta_{0,FRQ*RE\_VALUE}$ denotes the true parameter value, and s_{FRQ*RE_VALUE} is the standard error of the OLS estimator of $\beta_{0,FRQ*RE\_VALUE}$ from a model using only firm and time fixed effects obtained from the simulation; 1,000 simulation replications were performed. Panel A shows results based on one-way clustering with critical values from a *t*-distribution with $G-1$ degrees of freedom, where G is the number of clusters used in forming the one-way clustered standard errors. Panel B shows results based on using critical values from the moving blocks bootstrap. Panel C presents results based on FM with critical values from a *t*-distribution with $G-1$ degrees of freedom where G is the number of groups. The $G-1$ used to obtain critical values for clustering or FM is provided in the column "$G-1$" for reference. Panel D shows results based on the permutation inference procedure of Canay, Romano, and Shaikh (2014). Panel E presents results from the sensitivity analysis procedure and the ad hoc procedure for selecting groups. The column "Clusters" gives the level at which clustering or sample-splitting occurs, and the column "Fixed Effects" gives the levels of fixed effects that are included in the estimated model. Further details are provided in the main text, and details about the simulation design are given in a supplementary appendix.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

lack of a uniformly dominating procedure is not surprising. None of the considered estimators are theoretically efficient, and each of the observed explanatory variables exhibits different time series and spatial dependence properties. We do see that using the most aggressive fixed effects strategy, including a full set of firm by two-year time block fixed effects and a full set of year fixed effects as in the rows "firm × 2-year time block, year" in panel A and "2-year time block" in panel B, tends to be more variable than the less aggressive strategies. Again, this is unsurprising given how much of the variation in the explanatory variables is absorbed by the inclusion of these fixed effects. What is perhaps surprising is how competitive the point estimators seem to be in this case despite the inclusion of this rich set of fixed effects. One suspects the good performance is driven by the fact that, with the exception of *LN_AGE*, none of the variables exhibit strong firm-specific trends that drive their variability and are essentially eliminated by the fixed effects. Finally, we see that the ad hoc group selection strategy performs reasonably well in producing a procedure with competitive sampling variation.

In table 7, we report power for 5% level tests using the 16 procedures that did a reasonable job controlling size in the simulation, in the sense of having size of 5% tests uniformly smaller than 10% across the covariates in our example. These procedures are (i) standard errors clustered by eight-year time block with firm by eight-year time block and year fixed effects with *t*-critical value, (ii) standard errors clustered by two-year time block with firm by two-year time block and year fixed effects with *t*-critical value, (iii) standard errors clustered by eight-year time block with firm and year fixed effects with overlapping blocks bootstrap critical value, (iv) standard errors clustered by eight-year time block with firm by eight-year time block and year fixed effects with overlapping blocks bootstrap critical value, (v) standard errors clustered by six-year time block with firm by six-year time block and year fixed effects with overlapping blocks bootstrap critical value, (vi) standard errors clustered by four-year time block with firm by four-year time block and year fixed effects with overlapping blocks bootstrap critical value, (vii) standard errors clustered by two-year time block with firm by two-year time block and year fixed effects with overlapping blocks bootstrap critical value, (viii) FM by state, (ix) FM by one-digit SIC code, (x) FM by eight-year time block, (xi) FM by six-year time block, (xii) FM by four-year time block, (xiii) FM by two-year time block, (xiv) Canay–Romano–Shaikh by two-year time block, (xv) sensitivity analysis, and (xvi) ad hoc group selection.

We report results only for the coefficient on *FRQ*× *RE_VALUE* as it shares the broad properties of power for the other variables; results for the remaining covariates are provided in the supplementary appendix. For comparability, all tests are of the hypothesis that the population parameter is equal to $\beta_{FRQ\times RE\_VALUE,0} + ks_{FRQ\times RE\_VALUE}$, where $\beta_{FRQ\times RE\_VALUE,0}$ is the true coefficient on *FRQ*× *RE_VALUE* and $s_{FRQ\times RE\_VALUE}$ is the

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

standard deviation across simulation replications of the OLS estimator of the coefficient on *FRQ* × *RE_VALUE* from model (14) using only firm and year fixed effects. k thus measures how many standard errors from the true parameter value the hypothesized value is, and we consider $k \in \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ in table 7. We note that the results for $k = 0$ thus return the size of the test and we would like power to increase as we move k away from 0. Thus, ideally we would see power of 0.05 for $k = 0$ and power of 1 for $|k| > 0$.

The results in table 7 suggest two broad classes of inferential procedures based on power. The first class of procedures are all of the clustering procedures with overlapping blocks bootstrap critical values; clustering by eight-year time block with t-critical values; and FM by state, one-digit SIC code, and eight-year time block. These procedures tend to have power that increases very slowly as one moves the alternative farther away from the population value of the parameter. Two of these procedures are the same two based on the FM-style procedure grouping by state or one-digit SIC code that were highlighted previously as having poor point estimation properties in terms of RMSE. Their high sampling variability, of course, shows up in providing tests with relatively low power. It is also clear why clustering and FM using groups made of eight-year time blocks has low power. Grouping by eight-year time blocks corresponds to using two clusters in this study. Using only two clusters produces very noisy estimates of sampling variability that must be offset by using very large critical values to produce tests that have correct size. The use of these large critical values (12.71 in this case) then results in lower power despite the point estimators themselves not exhibiting high variability. It is less clear why using the overlapping blocks bootstrap critical values uniformly results in low power. We conjecture that this is due to having a relatively small time series to use in the bootstrap, which results in a deterioration of its performance in estimating critical values.

The other class of procedures are clustering by two-year time blocks with firm by two-year time block fixed effects and year fixed effects; FM with groups made by six-, four-, and two-year time blocks; Canay–Romano–Shaikh by two-year time block; sensitivity analysis; and ad hoc group selection. Among these procedures, there is no uniform dominance in terms of power. To illustrate, we provide power curves for each of the nine coefficients from the simulation study in figure 2 for four of these relatively higher powered procedures[33]: (a) ad hoc group selection, (b) standard errors clustered by two-year time block with firm by two-year time block and year fixed effects, (c) Canay–Romano–Shaikh by two-year time block, and (d) sensitivity analysis. Looking at just these procedures, we see that the approach of Canay, Romano, and Shaikh [2017] controls size and tends

[33] We chose these four to have one method from each of the broad classes of methods considered that have relatively high power.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

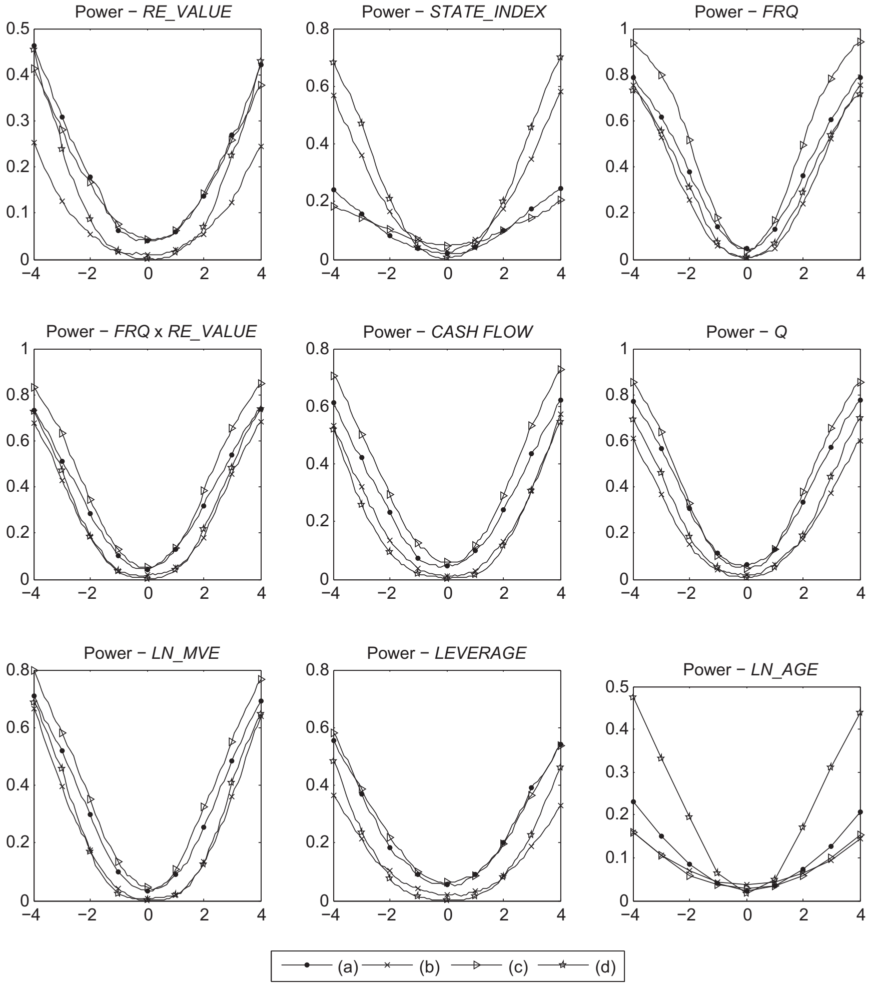

FIG. 2.—Power of 5% level tests based on simulation data. The *y*-axis gives the rejection frequency of 5% level tests of the null hypothesis that the population parameter is equal to $\beta_{j,0} + ks_j$, where k is the value on the *x*-axis, $\beta_{j,0}$ is the true coefficient on variable j, and s_j is the standard deviation across simulation replications of the OLS estimator of the coefficient on variable *j* from model (14) using only firm and year fixed effects. The procedures considered are (a) ad hoc group selection, (b) standard errors clustered by two-year time block with firm by two-year time block and year fixed effects, (c) Canay–Romano–Shaikh by two-year time block, and (d) sensitivity analysis. Further details are provided in the main text, and details about the simulation design are given in a supplementary appendix.

to have the highest power most often looking across variables. The ad hoc adaptive procedure tends to perform similarly to Canay, Romano, and Shaikh [2017] in most cases. However, there are cases where either of the other two procedures would be preferred to either the Canay, Romano, and Shaikh [2017] splitting on two-year time blocks or the adaptive procedure.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

It is interesting that the sensitivity analysis approach is substantially undersized in all cases but has power that tends to increase quite rapidly as one considers alternatives farther from the true population value, which leads to its overtaking other procedures in terms of power against more distant alternatives in a handful of cases and its not being severely outperformed in terms of power against moderate and far alternatives across the board.

The main conclusion from this section is that neither FM with sensible groups nor OLS with the full sample obviously dominates in terms of point estimation properties. This lack of dominance in point estimation suggests that the main focus in deciding on an inferential procedure should be on finding procedures that control size without losing too much power. Among the options we consider, we have some preference for the Canay, Romano, and Shaikh [2017] variant of FM due to its relatively good performance and theoretical guarantees. However, all of the sample-splitting procedures that use a small number of groups formed by splitting only along the time series seem to do well and are competitive with the Canay, Romano, and Shaikh [2017] approach. In addition, the simple and intuitive sensitivity approach is surprisingly competitive and may be appealing as deciding on a single adequate grouping structure may be difficult in practice.

5. *Concluding Comments*

In this review, we provide some intuition for recent developments that have occurred in theoretical econometrics and statistics regarding inference with dependent data, focusing on clustering, sample-splitting, and bootstrap procedures. Recent theoretical developments have focused less on obtaining consistent estimates of standard errors and more on providing inferential procedures that deliver accurate inferential statements. Several key insights arise from these theoretical developments. First, it is important to choose inferential schemes that accommodate the dependence that is present in the data. For clustering or sample-splitting approaches, accommodating the rich types of dependence that are likely present in real-world accounting and finance data will generally lead a researcher to prefer an approach that makes use of a few groups consisting of many observations, which lessens the potential for omitting important sources of correlation.

A complication that arises when a few, large groups are used for cluster estimators is that usual approximations to the behavior of test statistics do not work well in this scenario. Technically, the difficulty arises due to a high degree of sampling variability in estimating the sampling variability of an estimator of interest in this environment. Developments in statistics and econometrics allow us to overcome this difficulty by appropriate modification of critical values. Unfortunately, these modifications rely on strong homogeneity conditions, in the case of clustered standard error estimators, that seem likely to be violated in many accounting and finance applications.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Fortunately, this homogeneity is not required when sample-splitting estimators, such as FM, are used. We thus advocate, as a current rule-of-thumb, the use of sample-splitting estimators as in Fama and MacBeth [1973], Ibragimov and Müller [2010, 2016], and Canay, Romano, and Shaikh [2017] with a small number of large groups when faced with firm-level panel data where complicated dependence structures between observations may exist.

Bootstrap procedures seem to be challenging in the type of complicated panel data applications we consider. Fundamentally, it is difficult to come up with a mechanism for simulating data that capture the type of complicated dependence structure that one would expect to exist in firm-level panel data. A widely advocated approach in settings where clustered standard errors have been used is to use the cluster wild bootstrap. However, to accommodate rich dependence structures, one will likely wish to use a small number of large clusters, in which case the validity of the cluster wild bootstrap seems to rely on strong homogeneity conditions. A different option is to be relatively agnostic about cross-sectional dependence and use a time block bootstrap. This approach relies on temporal homogeneity, at least a moderate length time span, and is difficult to implement in unbalanced panels. We anticipate that many researchers in accounting and finance will want to work with panel data that are heterogeneous, unbalanced, and of only moderate time series dimension and thus recommend sample-splitting approaches relative to bootstrap methods based on the current state of the literature.

To illustrate the procedures and provide context for our discussion, we provide simulation results in a scenario designed to mimic data from Balakrishnan, Core, and Verdi [2014] that is representative of many accounting and finance applications. The simulation evidence illustrates the potential drawback of using clustered standard errors, either one-way or multiway, and the bootstrap and also demonstrate the relative robustness of sample-splitting techniques. The results are consistent with the advice that one should use a small number of large groups when faced with situations with complicated dependence coupled with the use of an FM-style approach to inference. For inference about individual regression coefficients, the simulation results are favorable to both the FM approaches with a small number of groups and the permutation inference variant suggested in Canay, Romano, and Shaikh [2017]. We do note that a modest number of groups are required for Canay, Romano, and Shaikh [2017] to have power. We also note that sensitivity analysis does surprisingly well in our simulation and may be useful in many settings. The simulation evidence also illustrates the practical difficulty of choosing grouping schemes that split along cross-sectional dimensions and appropriately capture cross-sectional/spatial dependence. This difficulty leads to our final piece of advice, which is that one might consider grouping together all cross-sectional observations and splitting only on the time dimension in cases where there is not strong a priori reason to believe there is some type of independence among cross-sectional units and a moderate length time series is available.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The big picture idea that one needs to form groups of observations for use in grouping-based inference schemes that result in dependence between observations that belong to different groups being small relative to the total sampling variation of the estimator of interest is true in general. Inference will not be approximately valid if this condition is violated. Just like discussion of identifying assumptions and model specification, discussion of why this condition on dependence is plausible for a grouping scheme used in practice in any given empirical setting should be provided.

We wish to conclude by noting that we think the advice given above is sensible and will produce reliable inference across many settings. Of course, there remains substantial work to be done. Choosing an appropriate grouping strategy is still largely an ad hoc and intuitive exercise. Providing implementable, adaptive procedures to group selection is an important topic for further research. Grouping strategies are simple and often one has some intuition about sensible strategies. However, they are crude ways to approximate covariance structures as they allow quite general forms of correlation between observations that share groups but impose, in estimation, no correlation between observations across groups. Smoother approximations as provided by appropriate generalizations of traditional HAC estimators as in Driscoll and Kraay [1998], Vogelsang [2012], and Bester et al. [2016] may offer gains in finite samples, though other issues arise in the application of these methods that could use further exploration. Finally, inference that is robust to very general forms of cross-sectional/spatial and temporal dependence may be very imprecise. There may be substantial gains available from modeling covariance structures among observations and exploiting these models to gain efficiency. Thinking about the tradeoffs between robustness and efficiency and providing tractable flexible models for the types of rich dependence structures that seem likely in observational accounting and finance data may be a useful avenue for further research.

REFERENCES

ANDREWS, D. W. K. "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica* 59 (1991): 817–58.

ANDREWS, D. W. K. "Higher Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators." *Econometrica* 70 (2002): 119–62.

ARELLANO, M. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics* 49 (1987): 431–34.

BALAKRISHNAN, K.; J. E. CORE; AND R. S. VERDI. "The Relation Between Reporting Quality and Financing and Investment: Evidence from Changes in Financing Capacity." *Journal of Accounting Research* 52 (2014): 1–36.

BELL, R. M., AND D. F. MCCAFFREY. "Computing Robust Standard Errors for Within-Groups Estimators." Mimeo RAND, 2002.

BERTRAND, M.; E. DUFLO; AND S. MULLAINATHAN "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (2004): 249–75.

BESTER, C. A.; T. G. CONLEY; AND C. B. HANSEN. "Inference for Dependent Data Using Cluster Covariance Estimators." *Journal of Econometrics* 165 (2011): 137–51.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Bester, C. A.; T. G. Conley; C. B. Hansen; and T. J. Vogelsang. “Fixed-b Asymptotics for Spatially Dependent Robust Nonparametric Covariance Matrix Estimators.” *Econometric Theory* 32 (2016): 154–86.

Cameron, C. A.; J. B. Gelbach; and D. L. Miller. “Bootstrap-Based Improvements for Inference with Clustered Errors.” *Review of Economics and Statistics* 90 (2008): 414–27.

Cameron, C. A.; J. B. Gelbach; and D. L. Miller. “Robust Inference with Multiway Clustering.” *Journal of Business and Economic Statistics* 29 (2011): 238–49.

Canay, I. A.; J. P. Romano; and A. M. Shaikh. “Randomization Tests Under an Approximate Symmetry Assumption.” *Econometrica* 85 (2017): 1013–30.

Canay, I. A.; A. Santos; and A. M. Shaikh. “The Wild Bootstrap with a ‘Small’ Number of ‘Large’ Clusters.” cemmap (centre for microdata methods and practice) Working Paper CWP27/18, 2018. Available at https://www.cemmap.ac.uk/publication/id/12915.

Carlstein, E. “The Use of Subseries Methods for Estimating the Variance of a General Statistic from a Stationary Time Series.” *Annals of Statistics* 14 (1986): 1171–79.

Carter, A. V.; K. T. Schnepel; and D. G. Steigerwald. “Asymptotic Behavior of a *t* Test Robust to Cluster Heterogeneity.” Working paper, University of California at Santa Barbara, 2013.

Conley, T. G. “GMM Estimation with Cross Sectional Dependence.” *Journal of Econometrics* 92 (1999): 1–45.

Djogbenou, A.; J. G. MacKinnon; and M. Ø. Nielsen. “Validity of Wild Bootstrap Inference with Clustered Errors.” Queen’s Economics Department Working paper No. 1383, 2017.

Driscoll, J., and A. Kraay. “Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data.” *Review of Economics and Statistics* 80 (1998): 549–60.

Efron, B. “Bootstrap Methods: Another Look at the Jackknife.” *Annals of Statistics* 7 (1979): 1–26.

Eicker, F., “Limit Theorems for Regression with Unequal and Dependent Errors.” *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1 (1967): 59–82.

Fama, E. F., and J. D. MacBeth. “Risk, Return, and Equilibrium: Empirical Tests.” *Journal of Political Economy* 81 (1973): 607–36.

Gonçalves, S. “The Moving Blocks Bootstrap for Panel Linear Regression Models with Individual Fixed Effects.” *Econometric Theory* 27 (2011): 1048–82.

Gonçalves, S., and T. J. Vogelsang. “Block Bootstrap HAC Robust Tests: The Sophistication of the Naive Bootstrap.” *Econometric Theory* 27 (2011): 745–91.

Gotze, F., and H. R. Kunsch. “Second-Order Correctness of the Blockwise Bootstrap for Stationary Observations.” *Annals of Statistics* 24 (1996): 1914–33.

Gow, I.; G. Ormazabal; and D. Taylor. “Correcting for Cross-Sectional and Time-Series Dependence in Accounting Research.” *The Accounting Review* 85 (2010): 483–512.

Hall, P., and J. L. Horowitz. “Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators.” *Econometrica* 59 (1996): 891–916.

Hansen, C. B. “Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When *T* Is Large.” *Journal of Econometrics* 141 (2007): 597–620.

Härdle, W., and E. Mammen. “Comparing Nonparametric Versus Parametric Regression Fits.” *Annals of Statistics* 21 (1993): 1926–47.

Huber, P. J. “The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions.” *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1 (1967): 221–33.

Ibragimov, R., and U. K. Müller. “*t*-Statistic Based Correlation and Heterogeneity Robust Inference.” *Journal of Business and Economic Statistics* 28 (2010): 453–68.

Ibragimov, R., and U. K. Müller. “Inference with Few Heterogeneous Clusters.” *Review of Economics and Statistics* 98 (2016): 83–96.

Imbens, G. W., and M. Kolesar. “Robust Standard Errors in Small Samples: Some Practical Advice.” Working paper, University of California, Berkeley, 2012.

Inoue, A., and M. Shintani. “Bootstrapping GMM Estimators for Time Series.” *Journal of Econometrics* 127 (2006): 531–55.

Kelejian, H., and I. Prucha. “HAC Estimation in a Spatial Framework.” *Journal of Econometrics* 140 (2007): 131–54.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

KIEFER, N. M., AND T. J. VOGELSANG. "Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size." *Econometric Theory* 18 (2002): 1350–66.

KIEFER, N. M., AND T. J. VOGELSANG. "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests." *Econometric Theory* 21 (2005): 1130–64.

KÜNSCH, H. R. "The Jackknife and the Bootstrap for General Stationary Observations." *Annals of Statistics* 17 (1989): 1217–41.

LEVINE, D. "A Remark on Serial Correlation in Maximum Likelihood." *Journal of Econometrics* 23 (1983): 337–42.

LIANG, K.-Y., AND S. ZEGER. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73 (1986): 13–22.

LIU, R. Y., AND K. SINGH. "Efficiency and Robustness in Resampling." *Annals of Statistics* 20 (1992): 370–84.

MACKINNON, J. G., AND M. D. WEBB. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Journal of Applied Econometrics* 32 (2016): 233–54.

MAMMEN, E. "Bootstrap and Wild Bootstrap for High Dimensional Linear Models." *The Annals of Statistics* 21 (1993): 255–85.

NEWEY, W. K., AND K. D. WEST. "A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (1987): 703–708.

NEWEY, W. K., AND K. D. WEST. "Automatic Lag Selection in Covariance Matrix Estimation." *Review of Economic Studies* 61 (1994): 631–54.

PETERSEN, M. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *Review of Financial Studies* 22 (2009): 435–80.

SINGH, K. "On the Asymptotic Accuracy of Efron's Bootstrap." *Annals of Statistics* 9 (1981): 1187–95.

STATA CORPORATION. *Stata User's Guide Release 13*, College Station, TX: Stata Press, 2013.

SUN, Y.; P. C. B. PHILLIPS; AND S. JIN. "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing." *Econometrica* 76 (2008): 175–94.

VILLACORTA, L. "Robust Standard Errors to Spatial and Time Dependence When *N* and *T* Are Not Large." Working paper, Central Bank of Chile, 2015.

VOGELSANG, T. "Heteroskedasticity, Autocorrelation, and Spatial Correlation Robust Inference in Linear Panel Models with Fixed-Effects." *Journal of Econometrics* 166 (2012): 303–19.

VOGELSANG, T. J. "Testing in GMM Models Without Truncation," in *Advances in Econometrics Volume 17, Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*, edited by T. B. Fomby and R. C. Hill. New York: Elsevier Science, 2003: 199–233.

WHITE, H. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (1980): 817–38.

WHITE, H., AND I. DOMOWITZ. "Nonlinear Regression with Dependent Observations." *Econometrica* 52 (1984): 143–61.

WILHELM, D. "Optimal Bandwidth Selection for Robust Generalized Methods of Moments Estimation." *Econometric Theory* 31 (2015): 1054–77.

1475679x, 2018, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12219 by Western University, Wiley Online Library on [20/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License