Taylor & Francis
Taylor & Francis Group

Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Testing for Unobserved Heterogeneity via *k-means* Clustering

Andrew J. Patton & Brian M. Weller

To cite this article: Andrew J. Patton & Brian M. Weller (2023) Testing for Unobserved Heterogeneity via *k-means* Clustering, Journal of Business & Economic Statistics, 41:3, 737-751, DOI: 10.1080/07350015.2022.2061983

To link to this article: https://doi.org/10.1080/07350015.2022.2061983

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2023, VOL. 41, NO. 3, 737–751
https://doi.org/10.1080/07350015.2022.2061983

Taylor & Francis
Taylor & Francis Group

Testing for Unobserved Heterogeneity via *k-means* Clustering

Andrew J. Patton[a] and Brian M. Weller[b]

[a]Duke University, Durham, NC; [b]Amazon.com, Seattle, WA

ABSTRACT

Clustering methods such as *k-means* have found widespread use in a variety of applications. This article proposes a split-sample testing procedure to determine whether a null hypothesis of a single cluster, indicating homogeneity of the data, can be rejected in favor of multiple clusters. The test is simple to implement, valid under mild conditions (including nonnormality, and heterogeneity of the data in aspects beyond those in the clustering analysis), and applicable in a range of contexts (including clustering when the time series dimension is small, or clustering on parameters other than the mean). We verify that the test has good size control in finite samples, and we illustrate the test in applications to clustering vehicle manufacturers and U.S. mutual funds.

ARTICLE HISTORY
Received December 2020
Accepted March 2022

KEYWORDS
Classification methods; Machine learning; Model selection; Overfitting

1. Introduction

Clustering methods provide researchers with a means of imposing some structure on a set of data under analysis. They are a middle ground between imposing strict homogeneity and allowing complete heterogeneity across the variables under analysis, enabling the researcher to group variables into clusters and impose homogeneity only within a cluster. Such methods have proven useful in a wide variety of applications including medical research (e.g., Eisen et al. 1998; Liu et al. 2008), economics (e.g., Francis, Owyang, and Savascin 2017; Patton and Weller 2019), and computer science (e.g., Ray and Turi 1999; Steinbach, Karypis, and Kumar 2000).

A key input to cluster analysis is the number of clusters to employ, and several methods for making this choice have been proposed in the literature. Perhaps most widely known is the "gap" statistic of Tibshirani, Walther, and Hastie (2003), which looks at the reduction in a measure of within-cluster heterogeneity as a function of the number of clusters. Other approaches include those based on information criteria (e.g., Fraley and Raftery 2002; Sugar and James 2003; Bonhomme and Manresa 2015) and those based on cross-validation methods (e.g., Tibshirani and Walther 2005; Wang 2010; Fu and Perry 2017).

In many applications there is economic interest in the null hypothesis of a *single* cluster, that is, that the variables under analysis are homogeneous, or, more generally, homogeneous in the attribute(s) under analysis. A rejection of this hypothesis in favor of a model with multiple clusters represents evidence of heterogeneity, a conclusion that can have important implications. For example, a rejection could indicate that a medical treatment is effective only for some subpopulations; that investments with equal risk may have different expected returns; or that different geographic regions respond differently to economic shocks. Existing methods for selecting the number of clusters do not allow for a formal probabilistic statement about the empirical evidence for or against a model with a single cluster. For that, we need a hypothesis test.

We combine results from panel econometrics (e.g., Hansen 2007; Bonhomme and Manresa 2015) with methods from out-of-sample forecast performance for model comparison (e.g., Diebold and Mariano 1995; West 1996) to present general methods for testing the null hypothesis of a single cluster imposing only mild regularity conditions on the data. We do so in the context of a panel of data containing N variables, each with T repeated observations, where the length of each dependent variable is d. Our testing approach exploits a standard assumption made in cluster analyses: cluster assignments are stable across repeated observations (e.g., time). This assumption is often used to tune "bandwidth" parameters in applications such as those considered here, and it enables us to estimate the cluster assignments on one sample (e.g., the first $T/2$ observations, or all odd-numbered observations) and then test the significance of the differences across clusters in a separate sample. Our asymptotic theory is developed for $N, T \rightarrow \infty$, although we can also accommodate any fixed $T \geq 2$.

The split-sample approach proposed here is simple to implement and we show that it allows one to conduct inference under much weaker assumptions than existing methods. For example, Liu et al. (2008) consider a high-dimensional setting $(d \gg N)$, and no repeated observations $(T = 1)$. Their approach takes a Gaussian distribution as the null hypothesis, which makes obtaining critical values for a test straightforward, however, the assumption of Gaussianity is much stronger than the null of homogeneous means, and in many economic applications Gaussianity is not plausible. Maitra, Melnykov, and Lahiri (2012)

CONTACT Andrew J. Patton ✉ andrew.patton@duke.edu Duke University, Durham, NC.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

consider a bootstrap test for multiple clusters, replacing the assumption of Gaussianity with an assumption that the data are identically distributed after some known transformation. Our assumption that we have at least one repeated observation allows us to weaken these assumptions considerably: we impose *no* distributional assumptions on the data beyond standard regularity conditions and do not require homogeneity of the data beyond that implied by the clustering analysis.

The remainder of our article is structured as follows. In Section 2 we present the main theoretical results, along with extensions to consider clustering on general estimated parameters (rather than means); tests when one of the clusters is "small"; and tests when the time series sample size is small. Section 3 presents simulation results on the finite-sample performance of the proposed methods, and Section 4 applies these tests to clustering vehicle manufacturers and U.S. mutual funds. Section 5 concludes. The Appendix contains all proofs, and the supplementary materials contain additional theoretical and simulation results.

2. Testing for Multiple Clusters

Below we present our main result on testing for multiple clusters, followed by results related to the choice of the number of clusters to consider under the alternative, and some empirically useful extensions of our main results.

2.1. *Main Result*

We observe T realizations of a collection of N variables, $\mathbf{Y}_{it}$ for $i = 1, 2, \ldots, N$, and $t = 1, 2, \ldots, T$, where $\dim(\mathbf{Y}_{it}) = d$. In all cases we consider a split of the full sample of T observations into two mutually exclusive, though not necessarily exhaustive, subsamples $\mathcal{R}$ and $\mathcal{P}$, where $\text{card}(\mathcal{R}) = R$ and $\text{card}(\mathcal{P}) = P$. Define $\mathcal{F}_R$ as the information set $\sigma\left(\{\mathbf{Y}_{it}\}_{i=1}^{N}, t \in \mathcal{R}\right)$, and let $\|\cdot\|$ denote the Euclidean norm.

Assumption 1. (a) The data come from $\mathbf{Y}_{it} = \mathbf{m}_i + \boldsymbol{\varepsilon}_{it}$, where $\boldsymbol{\varepsilon}_{it} \sim \text{iid}\, F_i\,(\mathbf{0}, \Sigma_i)$, where F_i is some distribution with mean zero and covariance matrix Σ_i, for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and where, for all $i, \mathbf{m}_i \in \mathcal{M} \subset \mathbb{R}^d$, Σ_i is strictly positive definite, and $E\left[\|\boldsymbol{\varepsilon}_{it}\|^4\right] \leq \bar{\kappa} < \infty \;\forall\, i$, (b) $\boldsymbol{\varepsilon}_{it} \perp\!\!\!\perp \boldsymbol{\varepsilon}_{js} \forall\, i \neq j$ and $\forall\, s$, (c) $N, P, R \to \infty$.

Assumption 1(a) allows the data to have arbitrary heterogeneity in variances and higher-order moments, subject to the existence of fourth-order moments. Importantly, it does not impose normality, as in Liu et al. (2008), nor does it require the observations to be a known transformation away from homogeneity, as in Maitra, Melnykov, and Lahiri (2012). For our baseline result, Assumption 1(b) imposes that the data are independent across time and cross-sections, and Assumption 1(c) imposes that the sizes of the cross-section and each of the subsamples diverge; later in the article we relax each of these conditions.

Assumption 2. $\mathbf{m}_i = \boldsymbol{\mu}^* \;\forall\, i$.

Assumption 2′. For known $G \geq 2$, (a) $\mathbf{m}_i \in \left\{\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_G^*\right\} \;\forall\, i$, (b) $\left\|\boldsymbol{\mu}_g^* - \boldsymbol{\mu}_{g'}^*\right\| > c > 0 \;\forall\, g \neq g'$, and (c) $\lim_{N\to\infty} N_g/N \equiv \pi_g \geq \underline{\pi} > 0$ for $g = 1, \ldots, G$, where $N_g \equiv \sum_{i=1}^{N} \mathbf{1}\left\{\gamma_i^* = g\right\}$, and $\gamma_i^* \in \{1, \ldots, G\}$ indicates to which cluster variable i belongs.

Assumption 2 defines the homogeneous case we study under the null hypothesis. Assumption 2′ covers the alternative hypothesis: (a) imposes that each variable belongs to one of the G clusters, indicated by the group membership vector $\boldsymbol{\gamma}$, (b) imposes that the cluster means are "well separated," which rules out any repeated values in the set $\left\{\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_G^*\right\}$, and (c) imposes that each cluster contains a nontrivial fraction of the variables.

We stack the mean vectors for the G clusters into a single $dG \times 1$ vector $\boldsymbol{\mu} \equiv \left[\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_G'\right]'$. Define the full-sample estimator:

$$\left(\hat{\boldsymbol{\mu}}_{NT}, \hat{\boldsymbol{\gamma}}_{NT}\right) = \tag{1}$$

$$\underset{(\boldsymbol{\mu},\boldsymbol{\gamma})\in\mathcal{M}^{dG}\times\Gamma_{N,G}}{\arg\min} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G}\left\|\mathbf{Y}_{it} - \boldsymbol{\mu}_g\right\|^2 \mathbf{1}\left\{\gamma_i = g\right\}$$

The set $\Gamma_{N,G}$ is the subset of all possible allocations of N variables to G groups that satisfies $\min_g \; N_g/N \geq \underline{\pi} > 0$, that is, it only allows for "nonnegligible" clusters.

Next define the estimator of the location parameters for a given value of $\boldsymbol{\gamma}$:

$$\tilde{\boldsymbol{\mu}}_{NT}\left(\boldsymbol{\gamma}\right) = \underset{\boldsymbol{\mu}\in\mathcal{M}^{dG}}{\arg\min} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G}\left\|\mathbf{Y}_{it} - \boldsymbol{\mu}_g\right\|^2 \mathbf{1}\left\{\gamma_i = g\right\} \tag{2}$$

We will look at a joint test that $\boldsymbol{\mu}_g^* = \boldsymbol{\mu}_{g'}^*$ for all $g \neq g'$, a total of $d\,(G-1)$ restrictions. To do so we will use the matrix:

$$\underset{(d(G-1)\times dG)}{A_{d,G}} = \left[\left(\boldsymbol{\iota}_{G-1} \otimes I_d\right), -I_{d(G-1)}\right] \tag{3}$$

where $\boldsymbol{\iota}_n$ is a $n \times 1$ vector of ones, I_n is the $n \times n$ identity matrix, and $\otimes$ is the Kronecker product. This allows us to state the null as

$$H_0 : A_{d,G}\boldsymbol{\mu}^* = 0 \Leftrightarrow H_0 : \boldsymbol{\mu}_g^* = \boldsymbol{\mu}_{g'}^* \;\forall\, g \neq g' \tag{4}$$

Theorem 1. Let $\hat{\boldsymbol{\gamma}}_{NR}$ be the estimated group assignments based on sample $\mathcal{R}$, and let $\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)$ be the estimated group means from sample $\mathcal{P}$ using group assignments $\hat{\boldsymbol{\gamma}}_{NR}$. Define the test statistic for the differences in the estimated means as

$$\begin{aligned} F_{NPR} &= NP\tilde{\boldsymbol{\mu}}_{NP}'\left(\hat{\boldsymbol{\gamma}}_{NR}\right) \\ &\quad A_{d,G}'\left(A_{d,G}\hat{\Omega}_{NPR}A_{d,G}'\right)^{-1} A_{d,G}\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right) \\ \text{where} \quad \underset{(dG\times dG)}{\hat{\Omega}_{NPR}^2} &= \text{diag}\left\{\hat{\Omega}_{1,NPR}, \ldots, \hat{\Omega}_{G,NPR}\right\} \\ \text{and} \quad \underset{(d\times d)}{\hat{\Omega}_{g,NPR}} &= \frac{1}{NP}\sum_{t\in\mathcal{P}}\sum_{i=1}^{N}\left(\mathbf{Y}_{it} - \bar{\mathbf{Y}}_{iP}\right)\left(\mathbf{Y}_{it} - \bar{\mathbf{Y}}_{iP}\right)' \\ &\quad \hat{\pi}_{g,NR}^{-2}\mathbf{1}\left\{\hat{\gamma}_{i,NR} = g\right\} \\ \hat{\pi}_{g,NR} &\equiv \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\hat{\gamma}_{i,NR} = g\right\}, \text{ for } g = 1, \ldots, G. \end{aligned}$$

(a) Under Assumptions 1 and 2,

$$F_{NPR} \xrightarrow{d} \chi^2_{d(G-1)}, \text{ as } N, P, R \to \infty.$$

(b) Under Assumptions 1 and 2′,

$$F_{NPR} \xrightarrow{p} \infty, \text{ as } N, P, R \to \infty.$$

The proof is presented in the Appendix. This theorem shows that if the means of the variables are homogeneous (i.e., Assumption 2 is satisfied) then the test statistic has a χ^2 limit distribution, while if the variables are heterogeneous (Assumption 2′ is satisfied) then the test statistic diverges, and so this test has power to detect multiple groups.

Importantly, the limit distribution of the test statistic is not affected by the problem of estimated cluster assignments. Cluster assignments are unidentified under the null hypothesis, and obtaining results on the behavior of the estimated cluster assignments in such a case is difficult. Indeed, even when the clusters are well separated (i.e., under the alternative hypothesis), estimation error in cluster assignments is difficult to treat, see Pollard (1981, 1982) and Bonhomme and Manresa (2015). Without distribution theory for the estimated cluster assignments it is difficult to quantify the over-fitting problem that arises when estimating a multi-cluster model on homogeneous data, and simply ignoring the over-fitting problem leads to tests with poor size control: in the simulation study described in Section 3 we find rejection rates as high as 100% for a nominal 0.05 level test. Our test overcomes the overfitting problem via a simple split-sample approach.

Theorem 1 can be generalized to accommodate various departures from the assumptions given above. Time series dependence can be accommodated by employing results from Hansen (2007). The main change required when allowing for time series dependence is that the formation of subsamples ($\mathcal{R}$ and $\mathcal{P}$) now requires some structure. We suggest using simply the first and second halves of the time series. It is also possible to allow for general time series and cross-sectional dependence, drawing on results in Bonhomme and Manresa (2015) adapted to our application. The supplementary materials contain details and formal results for these two extensions.

2.2. Choice of G Under the Alternative

The test above requires a choice of the number of clusters under the alternative, and in practice the value chosen may be incorrect. Below we consider the behavior of the test when the chosen value is too large or too small. The theory for behavior of the test statistic under the null is unaffected by this problem, of course, as under the null the true number of clusters is one and Theorem 1(a) applies. To simplify exposition, we assume that $d \equiv \dim(\mathbf{Y}_{it}) = 1$ in this section.

First, consider the case that the model under the alternative ($\tilde{G}$) has more clusters than are needed (G). In this case the model considered under the alternative is "too big," but importantly it nests the correct model. We show below that the test remains consistent in this case, although in finite samples it may have lower power than the case where the correct value for the number of clusters is chosen. Consider an assumption based on the optimal $\tilde{G}$-cluster model:

Assumption 3′. Assume $\tilde{G} > G > 1$, and (a) $p\lim_{N,R\to\infty} \hat{\boldsymbol{\mu}}_{NR}$ exists, and is denoted $\boldsymbol{\mu}^{\bigstar}$. (b) $\min_g \lim_{N\to\infty} \tilde{N}_g/N \geq \underline{\pi} > 0$, where $\tilde{N}_g \equiv \sum_{i=1}^{N} \mathbf{1}\left\{\gamma_i^{\bigstar} = g\right\}$, and $\gamma_i^{\bigstar} \in \left\{1, \ldots, \tilde{G}\right\}$ indicates to which cluster variable i is assigned.

The lemma below shows that the optimal $\tilde{G}$-cluster parameter vector is the true G-cluster parameter vector, $\boldsymbol{\mu}^*$, with one or more of its elements repeated.

Lemma 1. Assume that the DGP satisfies Assumptions 1 and 2′, but the researcher estimates a $\tilde{G} > G$ cluster model. Let $\boldsymbol{\mu}^{\bigstar} = \left[\boldsymbol{\mu}^{*\prime}, \boldsymbol{\varphi}^{*\prime}\right]'$, where $\boldsymbol{\varphi}^*$ is a $(\tilde{G} - G)$ vector with elements drawn with replacement from $\boldsymbol{\mu}^*$, and let $\boldsymbol{\gamma}^{\bigstar}$ be such that $\gamma_i^{\bigstar} = \gamma_j^{\bigstar} \Rightarrow \gamma_i^* = \gamma_j^* \; \forall\, i, j$. Then $\left(\boldsymbol{\mu}^{\bigstar}, \boldsymbol{\gamma}^{\bigstar}\right)$ is a solution to the $\tilde{G}$-cluster model as $N, T \to \infty$.

The proof is presented in the supplementary materials. Lemma 1 reveals that under Assumption 2′ the vector $\boldsymbol{\mu}^{\bigstar}$ is "weakly well separated," in that $\left|\mu_g^{\bigstar} - \mu_{g'}^{\bigstar}\right| > c > 0$ for at least $(G-1)$ pairs $(g, g') \in \left\{1, \ldots, \tilde{G}\right\}^2$. The presence of repeated values in $\boldsymbol{\mu}^{\bigstar}$ means that some pair-wise differences will be zero.

Next consider the case that the model under the alternative ($\tilde{G}$) has fewer clusters than are needed (G). Choosing $\tilde{G}$ to be too small will generally mean that the estimated cluster means are not consistent for their true values, however, our concern is only whether the null of a single cluster will be rejected. Assumption 3″(b) states that the population values of the cluster means are, like the true cluster means, "well separated." We show in Lemma 2 that if $d = 1$ then $c^{\bigstar} > c$ and well-separatedness is ensured, while for $d > 1$ it is possible to find cases where $c^{\bigstar} < c$, and so in such cases one must simply assume the true cluster means are sufficiently well separated that the misspecified cluster means are also well separated.[1] Thus, if G is not known ex ante, then if $d = 1$ choosing $\tilde{G}$ "small" is preferable, while if $d > 1$, the above arguments suggest that using a larger value of G is preferable.

Assumption 3″. Assume $G > \tilde{G} > 1$, and (a) $p\lim_{N,R\to\infty} \hat{\boldsymbol{\mu}}_{NR}$ exists and is denoted $\boldsymbol{\mu}^{\bigstar}$. (b) $\left|\mu_g^{\bigstar} - \mu_{g'}^{\bigstar}\right| > c^{\bigstar} > 0 \; \forall\, g \neq g'$, and (c) $\min_g \lim_{N\to\infty} \tilde{N}_g/N \geq \underline{\pi} > 0$, where $\tilde{N}_g \equiv \sum_{i=1}^{N} \mathbf{1}\left\{\gamma_i^{\bigstar} = g\right\}$, and $\gamma_i^{\bigstar} \in \left\{1, \ldots, \tilde{G}\right\}$ indicates to which cluster variable i is assigned.

Lemma 2. For $d = 1$, Assumption 2′(b) implies Assumption 3″(b), while for $d > 1$, this implication need not hold.

[1] Note that $c^{\star}$ depends on a variety of features of the problem: the true number of clusters (G), the number of clusters used in the alternative $(\tilde{G})$, the separation of the true cluster means (c), the dimension of the data (d), and, for $d > 1$, the within-variable correlation of the data (Σ_i), which can vary across $i = 1, \ldots, n$.

The following theorem contains results when the true number of clusters under the alternative is larger or smaller than that chosen by the researcher.

Theorem 2. Let $\tilde{G}$ denote the number of groups considered by the researcher and let $\hat{\boldsymbol{\gamma}}_{NR}$ be the estimated group assignments based on sample $\mathcal{R}$, and let $\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)$ be the estimated group means from sample $\mathcal{P}$ using group assignments $\hat{\boldsymbol{\gamma}}_{NR}$. Define the test statistic for the differences in the estimated means as

$$F_{NPR} = NP\tilde{\boldsymbol{\mu}}'_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)A'_{\tilde{G}}\left(A_{\tilde{G}}\hat{\Omega}_{NPR}A'_{\tilde{G}}\right)^{-1} A_{\tilde{G}}\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)$$

$$\text{where}\quad \underset{\left(\tilde{G}\times\tilde{G}\right)}{\hat{\Omega}_{NPR}} = \text{diag}\left\{\hat{\omega}^2_{1,NPR},\ldots,\hat{\omega}^2_{\tilde{G},NPR}\right\}$$

$$\text{and}\quad \underset{(1\times 1)}{\hat{\omega}^2_{g,NPR}} = \frac{1}{NP}\sum_{t\in\mathcal{P}}\sum_{i=1}^{N}\left(Y_{it}-\bar{Y}_{iP}\right)^2\hat{\pi}^{-2}_{g,R}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=g\right\}$$

$$\hat{\pi}_{g,R} \equiv \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=g\right\},\ \text{for } g=1,\ldots,\tilde{G}.$$

(a) Under Assumptions 1 and 2′,

$$F_{NPR}\xrightarrow{d}\chi^2_{\tilde{G}-1},\ \text{as } N,P\to\infty.$$

(b) Under Assumptions 1, 2′, and 3′ , or (c) 1, 2′, and 3″

$$F_{NPR}\xrightarrow{p}\infty,\ \text{as } N,P,R\to\infty.$$

The proof is presented in the supplementary materials. Theorem 2(b) shows that the test has unit asymptotic power under the alternative, even when $\tilde{G} > G$. In finite samples, power may be lower than if the correct number of clusters was used, as the critical values from a $\chi^2_{\tilde{G}}$ distribution are increasing in G. Theorem 2(c) confirms that if the cluster model with too few clusters is well separated, then we obtain the expected result for the test statistic under the alternative. We investigate the finite-sample impact of choosing an incorrect value of $\tilde{G}$ in Section 3.

With the results above we can consider a simple multiple testing procedure that applies when the researcher does not know the correct value for G under the alternative, and wants to consider a range of possible values. For example, the researcher implements the test for $\tilde{G} = 2,\ldots,\bar{G}$, a total of $\bar{G}-1$ tests. The p-values from each of these tests, denoted $p_{\tilde{G}}$, can be combined via a Bonferroni adjustment: define the joint p-value as

$$p_{\text{Bonf}} = \min\left\{1,\left(\bar{G}-1\right)\times\min_{\tilde{G}\in\{2,\ldots,\bar{G}\}}p_{\tilde{G}}\right\} \tag{5}$$

then reject the null that $G = 1$ in favor of $G \in \left\{2,\ldots,\bar{G}\right\}$ if $p_{\text{Bonf}} < \alpha$, where α is the desired level for the test. The choice of $\bar{G}$ is important in this approach: if it is chosen it too large, the Bonferroni adjustment will make the test conservative, while if it is chosen too small then, for $d > 1$, the estimated cluster parameters may be closer together than the true cluster parameters (as discussed above) and thus harder to distinguish via a test. We investigate this in our simulation study in Section 3 and, foreshadowing those results, we find that for larger sample sizes setting $\bar{G} = 5$ and combining the resulting four individual tests using a Bonferroni correction works well, while for smaller sample sizes setting $\bar{G} = 2$ and using just a single test avoids a loss of power.[2]

2.3. Extensions

2.3.1. Dealing with "Small" Clusters

Our interest is in the joint restriction that $\mu^*_g = \mu^*_{g'}$ for all $g \neq g'$, a total of $(G-1)$ restrictions. To allow for the presence of "small" clusters, we will test an implication of this null, namely that $\mu^*_g = \mu^*_{g'}$ for all $g \neq g'$ s.t. $\pi_g, \pi_{g'} \geq \underline{\pi}$. We adjust Assumption 2(c) to require only that at least two clusters are "large." We simplify the exposition by assuming that $d \equiv \dim\left(\mathbf{Y}_{it}\right) = 1$, but the results generalize naturally to the case that $d > 1$.

Assumption 2′(c^S). $\sum_{g=1}^{G}\mathbf{1}\left\{\pi_g \geq \underline{\pi}\right\} \geq 2$, where $\underline{\pi} > 0$, $\pi_g \equiv \lim_{N\to\infty} N_g/N \geq \underline{\pi} > 0$, $N_g \equiv \sum_{i=1}^{N}\mathbf{1}\left\{\gamma^*_i = g\right\}$, and $\gamma^*_i \in \{1,\ldots,G\}$ indicates to which cluster variable i belongs.

To implement this test, order the clusters so that $\hat{\pi}_{1,NR} \geq \hat{\pi}_{2,NR} \geq \cdots \geq \hat{\pi}_{G,NR}$, and define $\hat{G}_{NR}$ as the number of "large" estimated clusters, that is, the number of clusters that satisfy the condition $\hat{\pi}_{g,NR} \geq \underline{\pi}$. For $2 \leq G' \leq G$, define the matrix

$$\underset{((G'-1)\times G)}{B_{G',G}} = \left[\boldsymbol{\iota}_{G'-1}, -I_{(G'-1)}, \mathbf{0}_{(G'-1,G-G')}\right]. \tag{6}$$

This is the matrix comprised of the first $\left(G'-1\right)$ rows of $A_{1,G}$ defined in Equation (3). This allows us to obtain an implication of the null for the $\hat{G}_{NR}$ "large" clusters:

$$H^S_0 : B_{\hat{G}_{NR},G}\boldsymbol{\mu}^* = 0. \tag{7}$$

Note that below we characterize the asymptotic distribution of the p-value of the test statistic rather than the test statistic itself. The limiting distribution of the latter depends on the value for $\hat{G}_{NR}$, which in turn depends on $\mathcal{F}_R \equiv \sigma\left(\{\mathbf{Y}_{it}\}^N_{i=1}, t\in\mathcal{R}\right)$. Our proof technique relies on the limiting distribution being independent of $\mathcal{F}_R$; we achieve this below by transforming the test statistic to a p-value.

Theorem 3. Let $\hat{\boldsymbol{\gamma}}_{NR}$ be the estimated group assignments based on sample $\mathcal{R}$, and let $\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)$ be the estimated group means from sample $\mathcal{P}$ using group assignments $\hat{\boldsymbol{\gamma}}_{NR}$. Let $\Upsilon\left(\cdot;q\right)$ denote the CDF of a χ^2 variable with q degrees of freedom, and define the p-value for the differences in the estimated means as:

[2] An alternative approach for choosing a value of G to use in the alternative hypothesis is via cross-validation. In this approach, the sample is split into three subsamples: the first for estimation of cluster assignments across a range of values of G; the second for choosing the optimal value $\hat{G}^*$; the third for testing $G = 1$ versus $G = \hat{G}^*$. Such an approach avoids the need for a Bonferroni adjustment, which can cost power, but involves using splitting the data across three subsamples rather than two, which can also cost power. We leave a detailed investigation of such an approach for future research.

$$\text{Pval}_{NPR} = 1 - \Upsilon\left(F_{NPR}; \hat{G}_{NR} - 1\right)$$

$$\begin{aligned}
\text{where}\quad F_{NPR} &= NP\tilde{\boldsymbol{\mu}}'_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)B'_{\hat{G}_{NR},G}\left(B_{\hat{G}_{NR},G}\hat{\Omega}_{NPR}B'_{\hat{G}_{NR},G}\right)^{-1} \\
&\quad B_{\hat{G}_{NR},G}\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right) \\
\underset{(dG\times dG)}{\hat{\Omega}^2_{NPR}} &= \text{diag}\left\{\hat{\Omega}_{1,NPR},\ldots,\hat{\Omega}_{G,NPR}\right\} \\
\underset{(d\times d)}{\hat{\Omega}_{g,NPR}} &= \frac{1}{NP}\sum_{t\in\mathcal{P}}\sum_{i=1}^{N}\left(\mathbf{Y}_{it}-\bar{\mathbf{Y}}_{iP}\right)\left(\mathbf{Y}_{it}-\bar{\mathbf{Y}}_{iP}\right)' \\
&\quad \hat{\pi}^{-2}_{g,R}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=g\right\} \\
\hat{\pi}_{g,R} &\equiv \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=g\right\},\ \text{for } g=1,\ldots,G.
\end{aligned}$$

(a) Under Assumptions 1 and 2′,

$$\text{Pval}_{NPR}\xrightarrow{d}\text{Unif}\,(0,1)\,,\ \text{as } N,P,R\to\infty.$$

(b) Under Assumptions 1 and 2′(a),(b),(c^S),

$$\text{Pval}_{NPR}\xrightarrow{p}0,\ \text{as } N,P,R\to\infty.$$

2.3.2. Diverging N and finite T

We consider here the case that the number of repeated observations (T, in our notation) is small relative to the number of variables, N. Our split-sample approach to overcome the overfitting problem requires only $T\geq 2$, not $T\to\infty$. We consider the finite T case by modifying Assumption 1 as follows. We again simplify exposition by assuming that $d\equiv\dim(\mathbf{Y}_{it})=1$ and $G=2$, but the results generalize naturally to the case that $d>1$ and/or $G>2$.

Assumption 1′. (a) The data come from $Y_{it}=m_i+\varepsilon_{it}$, for $i=1,\ldots,N$, and $t=1,\ldots,T\geq 2$, where $m_i\in\left[\underline{m},\bar{m}\right]\subset\mathbb{R}$ and $V\left[\varepsilon_{it}\right]\equiv\sigma_i^2\in\left[\underline{\sigma}^2,\bar{\sigma}^2\right]\subset\mathbb{R}_{++}\ \forall\, i, E\left[\varepsilon_{it}\right]=0$ and $E\left[\left|\varepsilon_{it}\right|^{4+\delta}\right]\leq\bar{\kappa}<\infty\forall i$ for some $\delta>0$, (b) $\varepsilon_{it}\perp\!\!\!\perp\varepsilon_{jt}\forall\, i\neq j$, and $\varepsilon_{it}\perp\!\!\!\perp\varepsilon_{js}\forall\, i,j$ for $(t,s)\in\{\mathcal{R},\mathcal{P}\}$, and (c) $N\to\infty$, and $R,P\geq 1$.

Assumption 1′(a) allows for cross-sectional heteroscedasticity, and heterogeneity more generally, in the distribution of residuals, subject to them being mean zero and having finite $4+\delta$ moments. Assumption 1′(b) imposes cross-sectional independence, and time series independence across the $\mathcal{R}$ and $\mathcal{P}$ subsamples. Within each of the $\mathcal{R}$ and $\mathcal{P}$ subsamples, time series dependence is not constrained. Assumption 1′(c) requires the cross-sectional dimension to diverge, and each subsample to have at least one observation.

Theorem 4. Let $\hat{\boldsymbol{\gamma}}_{NR}$ be the estimated group assignments based on sample $\mathcal{R}$, and let $\tilde{\mu}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)$ be the estimated group means from sample $\mathcal{P}$ using group assignments $\hat{\boldsymbol{\gamma}}_{NR}$. Define the t-statistic for the differences in the estimated means as:

$$\text{tstat}_{NPR}=\frac{\sqrt{NP}\left(\tilde{\mu}_{1,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)-\tilde{\mu}_{2,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\right)}{\hat{\omega}_{NPR}}$$

$$\begin{aligned}
\text{where}\quad \hat{\omega}^2_{NPR} &\equiv \frac{1}{NP}\sum_{i=1}^{N}\boldsymbol{\iota}'_P\hat{\boldsymbol{\varepsilon}}_i\hat{\boldsymbol{\varepsilon}}'_i\boldsymbol{\iota}_P\left(\hat{\pi}^{-2}_{1,NR}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=1\right\}\right. \\
&\quad \left.+\,\hat{\pi}^{-2}_{2,NR}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=2\right\}\right) \\
\underset{(P\times 1)}{\hat{\boldsymbol{\varepsilon}}_i} &= \mathbf{Y}_i-\tilde{\mu}_{\hat{\gamma}_{i,NR},NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\boldsymbol{\iota}_P \\
\text{and}\quad \hat{\pi}_{g,NR} &\equiv \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=g\right\},\ \text{for } g=1,2.
\end{aligned}$$

where $\boldsymbol{\iota}_P$ is a $(P\times 1)$ vector of ones.

(a) Under Assumption 1′ and 2,

$$\text{tstat}_{NPR}\xrightarrow{d}N\,(0,1)\,,\ \text{as } N\to\infty.$$

(b) Under Assumption 1′ and 2′,

$$\left|\text{tstat}_{NPR}\right|\xrightarrow{p}\infty\,,\ \text{as } N\to\infty.$$

This theorem expands the applicability of the testing approach proposed in this article: we now only need $T\geq 2$, rather than T "large" in an asymptotic sense. Of course, the power of the test will be greater if a larger sample size is available, but this theorem shows that even in applications with a small time series sample size, the proposed testing approach may be adopted.

2.3.3. Clustering on Estimated Parameters

Here we consider the problem of clustering on some parameter, $\boldsymbol{\beta}\in\mathcal{A}\subset\mathbf{R}^b$, estimated for each of the N variables. This allows researchers to cluster on other features of the data, such as variances, higher-order moments, regression coefficients, or other estimated parameters. For each variable i and each subsample $\mathcal{R}$ and $\mathcal{P}$, we have a $(b\times 1)$ vector of estimated parameters $\left(\hat{\boldsymbol{\beta}}_{i,R},\hat{\boldsymbol{\beta}}_{i,P}\right)$. We assume that the estimated parameters satisfy the following assumption.

Assumption 4. (a) The estimated parameters satisfy $\hat{\boldsymbol{\beta}}_{i,S}=\boldsymbol{\beta}^*_i+\boldsymbol{\varepsilon}_{i,S}$, for $i=1,\ldots,N$, and $S\in\{R,P\}$, where $\left\|\boldsymbol{\beta}^*_i\right\|$ is bounded $\forall\, i$, $V\left[\boldsymbol{\varepsilon}_{i,S}\right]\equiv\Sigma_i$ a positive definite matrix, $E\left[\boldsymbol{\varepsilon}_{i,S}\right]=0$ and $E\left[\left|\varepsilon_{i,k,S}\right|^{4+\delta}\right]\leq\bar{\kappa}<\infty\forall i$, all $k=1,\ldots,b$ for some $\delta>0$, (b) $\varepsilon_{i,S}\perp\!\!\!\perp\varepsilon_{j,S}\forall\, i\neq j, S\in\{R,P\}$, and $\varepsilon_{i,R}\perp\!\!\!\perp\varepsilon_{j,P}\forall\, i,j$, and (c) $N\to\infty$.

These assumptions mimic those in Assumption 1′, in that Assumption 4(a) allows for cross-sectional heteroscedasticity and heterogeneity in the distribution of errors in the estimated parameters, subject to them being mean zero and having finite $4+\delta$ moments. Assumption 4(b) imposes cross-sectional independence, and time series independence across the $\mathcal{R}$ and $\mathcal{P}$ subsamples. In applications with time series dependence, the latter assumption may be satisfied if the $\mathcal{R}$ and $\mathcal{P}$ subsamples are separated by a middle, unused, sample of the data.

Assumption 2$_P$. (a) The cluster parameters satisfy $\boldsymbol{\beta}^*_i=\alpha^*\ \forall\, i$.

Assumption 2′$_P$. (a) The cluster parameters satisfy $\boldsymbol{\beta}^*_i\in\left\{\boldsymbol{\alpha}^*_1,\ldots,\boldsymbol{\alpha}^*_G\right\}\ \forall\, i$, (b) $\left\|\boldsymbol{\alpha}^*_g-\boldsymbol{\alpha}^*_{g'}\right\|>c>0\ \forall\, g\neq g'$, (c) $\lim_{N\to\infty}N_g/N\equiv\pi_g\geq\underline{\pi}>0$ for $g=1,\ldots,G$, where $N_g\equiv\sum_{i=1}^{N}\mathbf{1}\left\{\gamma^*_i=g\right\}$, and $\gamma^*_i\in\{1,\ldots,G\}$ indicates to which cluster variable i belongs.

We stack the parameter vectors for the G clusters into a single $bG \times 1$ vector $\boldsymbol{\alpha} \equiv \left[\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_G'\right]'$ and define the estimator:

$$\left(\hat{\boldsymbol{\alpha}}_{NS}, \hat{\boldsymbol{\gamma}}_{NS}\right) = \underset{(\boldsymbol{\alpha},\boldsymbol{\gamma})\in\mathcal{A}^G\times\Gamma_{N,G}}{\arg\min} \frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\left(\hat{\boldsymbol{\beta}}_{i,S} - \boldsymbol{\alpha}_g\right)^2 \mathbf{1}\left\{\gamma_i = g\right\}, \text{ for } S \in \{R, P\} \tag{8}$$

as well as the estimator of the cluster parameters for a given value of the group membership vector:

$$\tilde{\boldsymbol{\alpha}}_{NP}\left(\boldsymbol{\gamma}\right) = \underset{\alpha\in\mathcal{A}^G}{\arg\min}\ \frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\left(\hat{\boldsymbol{\beta}}_{i,P} - \boldsymbol{\alpha}_g\right)^2 \mathbf{1}\left\{\gamma_i = g\right\}. \tag{9}$$

The theorem provides a test for multiple clusters based on a general estimated parameter vector.

Theorem 5. Let $\hat{\boldsymbol{\gamma}}_{NR}$ be the estimated group assignments based on sample $\mathcal{R}$, and let $\tilde{\boldsymbol{\alpha}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)$ be the estimated cluster parameters from sample $\mathcal{P}$ using group assignments $\hat{\boldsymbol{\gamma}}_{NR}$. Define the test statistic for the differences in the estimated means as

$$\begin{aligned} F_{NPR} &= N\tilde{\boldsymbol{\alpha}}_{NP}'\left(\hat{\boldsymbol{\gamma}}_{NR}\right)A_{bG}'\left(A_{bG}\hat{\Omega}_{NPR}A_{bG}'\right)^{-1} \\ &\quad A_{bG}\tilde{\boldsymbol{\alpha}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right) \\ \text{where}\quad \underset{(bG\times bG)}{\hat{\Omega}_{NPR}} &= \operatorname{diag}\left\{\hat{\Omega}_{1,NPR}, \ldots, \hat{\Omega}_{G,NPR}\right\} \\ \text{and}\quad \underset{(b\times b)}{\hat{\Omega}_{g,NPR}} &= \frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_{g,NR}^{-2}\mathbf{1}\left\{\hat{\gamma}_{i,NR} = g\right\} \\ &\quad \left(\hat{\boldsymbol{\beta}}_{i,P} - \tilde{\boldsymbol{\alpha}}_{g,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\right)^2 \\ \hat{\pi}_{g,NR} &\equiv \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\hat{\gamma}_{i,NR} = g\right\}, \text{ for } g = 1, \ldots, G. \end{aligned}$$

(a) Under Assumptions 4 and 2_P,

$$F_{NPR} \xrightarrow{d} \chi^2_{d(G-1)}, \text{ as } N \to \infty.$$

(b) Under Assumptions 4 and $2_P'$,

$$F_{NPR} \xrightarrow{p} \infty, \text{ as } N \to \infty.$$

3. Simulation Study

In this section we investigate the finite-sample behavior of the proposed tests. We first study the finite-sample size of the test, using the design:

$$\begin{aligned} \mathbf{Y}_{it} &= \mathbf{m}_i + \boldsymbol{\varepsilon}_{it},\ i = 1, \ldots, N;\ t = 1, \ldots, T \\ \boldsymbol{\varepsilon}_{it} &\sim \text{iid } F_i\left(\mathbf{0}, I_d\right). \end{aligned} \tag{10}$$

We impose $\mathbf{m}_i = \mathbf{0}\forall\, i$, thereby ensuring that the null of homogeneous means is satisfied. We consider a variety of configurations of the problem: $N \in \{30, 150, 300\}$, $T \in \{50, 250, 1000\}$, $d \in \{1, 2, 5\}$, $G \in \{2, 3, 4, 5\}$. In addition to the four individual values of G considered under the alternative, we also study the performance of a Bonferroni-corrected combination method that considers all four tests.

We take $\boldsymbol{\varepsilon}_{it}$ to be Normally distributed or heterogeneously distributed; in the latter case the distribution for each variable i is randomly selected from one of $N(0, 1)$, $\text{Exp}(2)$, $\text{Unif}(-3, 3)$, $\chi^2(4)$ or $t(5)$, standardized to have zero mean and unit variance. The heterogeneous data cannot be considered using the tests of Liu et al. (2008) and Maitra, Melnykov, and Lahiri (2012). We implement the test in Theorem 1 at the 0.05 significance level, splitting the time series evenly to form the $\mathcal{R}$ and $\mathcal{P}$ samples. We use 1000 replications.

Table 1 reports the finite-sample size results. We see that the rejection rates are generally very close to the nominal level of 0.05, for both the Normal and the heterogeneous data. It is noteworthy that the Bonferroni-based test, which sets $\bar{G} = 5$ and combines four individual tests, also has rejection rates close to the nominal level, indicating that the well-known tendency for a Bonferroni-adjusted test to be conservative does not arise here. In the supplementary materials we repeat this simulation study using a test that does *not* split the time series into $\mathcal{R}$ and $\mathcal{P}$ samples. Table SA.1 reveals that the finite-sample rejection rates for such an approach are 100% in all but one configuration (where it is instead 99%), confirming the finite-sample size problems stemming from k-means overfitting the data, and motivating our approach.

We next consider the finite-sample power of the proposed test. We fix $d \equiv \dim\left(\mathbf{Y}_{it}\right) = 1$ and we consider an alternative containing $G = 2$ clusters. The cluster means are set to $(0, \mu_2)$, with $\mu_2 \in [0, 0.5]$. The case that $\mu_2 = 0$ corresponds to the null of a single cluster, and the rejection rate at that point should equal 0.05, the size of the test. As μ_2 increases the cluster means become better separated and we expect the test to reject the null with greater frequency. Figure 1 reveals that the test has strong power to reject the null hypothesis when the sample sizes (N, T) are large, and when the distance between cluster means is large. When $(N, T) = (30, 50)$ the test fails to detect small differences between the cluster means, and unit power is only achieved when $\mu_2 = 0.5$. When $(N, T) = (150, 1000)$ even small differences are significant, and unit power is achieved at $\mu_2 = 0.1$. It is noteworthy that the power of the test is essentially identical for Normally and heterogeneously distributed data. For the remainder of the simulation results we focus on Normally distributed data; the results for heterogeneously distributed data are very similar.

Figure 2 studies the sensitivity of the test to the choice of number of clusters under the alternative. We consider two representative combinations of sample sizes $(N, T) = (30, 50)$ and $(150, 250)$, and for each sample size pair we choose a value of μ_2 such that the test has power strictly inside $(0.05, 1)$, namely $\mu_2 = 0.2$ and $\mu_2 = 0.075$, respectively. In the left panel, the true number of clusters is two, and we consider tests that allow for between two and five clusters under the alternative. Consistent with intuition, for both sample size pairs, we observe a decrease in power as the number of clusters is increased from two to five, though the decrease is small (e.g., power drops from 0.21 to 0.20 for the smaller sample size).

In the right panel of Figure 2 the true number of clusters is five (with cluster means evenly spaced between zero and either 0.2 or 0.075 depending on the sample size). Like the left panel, we find that power is nearly unaffected by the choice of G, with a slight increase in power from using smaller G. Though the

Table 1. Finite sample rejection rates.

		$N = 30$	30	30	150	150	150	600	600	600
d	*G*	$T = 50$	250	1000	50	250	1000	50	250	1000
Panel A: Normal data										
1	2	0.055	0.049	0.056	0.057	0.046	0.055	0.059	0.056	0.048
1	3	0.052	0.047	0.035	0.046	0.045	0.047	0.050	0.043	0.055
1	4	0.057	0.042	0.064	0.034	0.059	0.052	0.053	0.056	0.036
1	5	0.058	0.054	0.059	0.034	0.048	0.053	0.049	0.051	0.047
1	Bonf.	0.057	0.057	0.055	0.045	0.048	0.049	0.059	0.045	0.045
2	2	0.040	0.048	0.064	0.040	0.046	0.052	0.049	0.046	0.039
2	3	0.048	0.051	0.060	0.040	0.062	0.045	0.036	0.058	0.061
2	4	0.072	0.061	0.046	0.050	0.040	0.047	0.039	0.061	0.044
2	5	0.058	0.044	0.043	0.045	0.063	0.043	0.067	0.053	0.054
2	Bonf.	0.049	0.044	0.066	0.042	0.050	0.040	0.045	0.064	0.047
5	2	0.040	0.058	0.060	0.049	0.051	0.062	0.052	0.041	0.044
5	3	0.050	0.052	0.059	0.052	0.054	0.044	0.052	0.049	0.053
5	4	0.066	0.047	0.054	0.041	0.067	0.053	0.052	0.055	0.051
5	5	0.083	0.049	0.049	0.055	0.037	0.044	0.059	0.049	0.048
5	Bonf.	0.065	0.051	0.063	0.043	0.051	0.049	0.050	0.044	0.050
Panel B: Heterogeneous data										
1	2	0.055	0.045	0.040	0.042	0.061	0.060	0.049	0.040	0.055
1	3	0.062	0.046	0.062	0.041	0.057	0.050	0.047	0.047	0.057
1	4	0.045	0.045	0.053	0.053	0.062	0.053	0.052	0.052	0.036
1	5	0.058	0.050	0.057	0.052	0.053	0.039	0.051	0.044	0.057
1	Bonf.	0.060	0.043	0.053	0.040	0.052	0.056	0.050	0.045	0.051
2	2	0.048	0.049	0.048	0.045	0.048	0.045	0.042	0.053	0.043
2	3	0.053	0.039	0.045	0.050	0.046	0.062	0.049	0.045	0.052
2	4	0.063	0.059	0.051	0.055	0.053	0.050	0.037	0.045	0.058
2	5	0.049	0.037	0.075	0.053	0.038	0.041	0.041	0.047	0.041
2	Bonf.	0.050	0.044	0.049	0.053	0.039	0.046	0.038	0.043	0.049
5	2	0.054	0.049	0.044	0.048	0.049	0.057	0.044	0.047	0.039
5	3	0.056	0.038	0.050	0.053	0.045	0.055	0.052	0.043	0.040
5	4	0.076	0.067	0.039	0.063	0.047	0.048	0.053	0.045	0.054
5	5	0.069	0.055	0.050	0.070	0.041	0.047	0.040	0.057	0.051
5	Bonf.	0.066	0.054	0.050	0.047	0.032	0.062	0.040	0.046	0.046

NOTE: This table presents the proportion of simulations in which we reject the null of a single cluster in favor of multiple clusters, using the test proposed in Theorem 1 at a 0.05 significance level. The top panel presents results for iid Normal data; the lower panel presents results when the distribution is randomly drawn from one of $N(0,1)$, Exp (2), Unif $(-3,3)$, $\chi^2(4)$ or $t(5)$, each standardized to have zero mean and unit variance. The dimension of the variables is denoted *d*, the number of groups considered under the alternative is denoted *G*, the number of variables is denoted *N*, and the number of time series observations is denoted *T*. Rows labeled "Bonf." use tests with a Bonferroni correction to consider $G \in \{2,3,4,5\}$ under the alternative. The number of simulations is 1000.

models with $G < 5$ are misspecified, Lemma 2 shows that for $d = 1$, as in this design, the too-small models will have cluster means that are better separated than the correct model, increasing power. Importantly, when $d > 1$, Lemma 2 shows that choosing a too-small value for G can make the estimated cluster parameters harder to separate, thereby reducing the power of the test, and it is possible that in such cases the optimal G is larger than two. Broadly, though, Figure 2 suggests that the test exhibits reasonable robustness to the choice of G.

Figure 3 examines the performance of a test based on a Bonferroni adjustment to combine four tests using $G = 2, 3, 4, 5$, compared with a test that correctly chooses $G = 2$. Unsurprisingly, the Bonferroni-corrected test is conservative, at least for the smaller sample size, and exhibits lower power than the test using the correct value of the G. When the sample sizes are small, $(N, T) = (30, 50)$, the loss of power is sizeable, however, for larger sample sizes, $(N, T) = (150, 250)$, the power loss is minimal.

Next we study the performance of the test in Theorem 3, designed to accommodate small clusters. We again consider $(N, T) = (30, 50)$ and $(150, 250)$, with $d = 1$. We set the number of clusters to three, and we look at the impact of a small cluster by varying the proportion of variables in the third cluster, denoted π_3. We set $\pi_1 = \pi_2 = (1 - \pi_3)/2$, and consider $\pi_3 \in [1/100, 1/3]$, with the largest value for π_3 corresponding to all clusters having the same weight. We set the mean of the first cluster to zero in all cases, $\mu_1 = 0$, and we set the second cluster mean $\mu_2 = \mu_3/2$. To study the finite-sample size of the test, we set $\mu_3 = 0$. To study power we choose μ_3 such that the test has power strictly inside $(0.05, 1)$, namely $\mu_3 = 0.2$ for $(N, T) = (30, 50)$ and $\mu_3 = 0.075$ for $(N, T) = (150, 250)$. We use a threshold of $\underline{\pi} = 0.1$ to decide whether a cluster is "small" and thus excluded from the test. The left panel of Figure 4 shows that the test in Theorem 3 controls the size of the test. The right panel shows, as expected, that the power of the test increases as the smallest cluster grows to be closer in size to the other two clusters.

To illustrate the applicability of the test for clusters on a general estimated parameter, we next consider an application where the clusters are found using the variables' autoregressive coefficients. That is, for each variable $Y_{i,t}$ we consider the autoregression:

$$Y_{i,t} = \phi_{0,i} + \phi_{1,i} Y_{i,t-1} + \varepsilon_{i,t} \tag{11}$$

and the cluster model assumption is that the AR(1) coefficients take one of only two values

$$\phi_{1,i} = \alpha_{\gamma_i}, \; i = 1, 2. \tag{12}$$

We fix $\alpha_1 = 0.5$ and we vary the autoregressive coefficient of the second cluster, $\alpha_2 \in [0.1, 0.9]$. Figure 5 shows the results

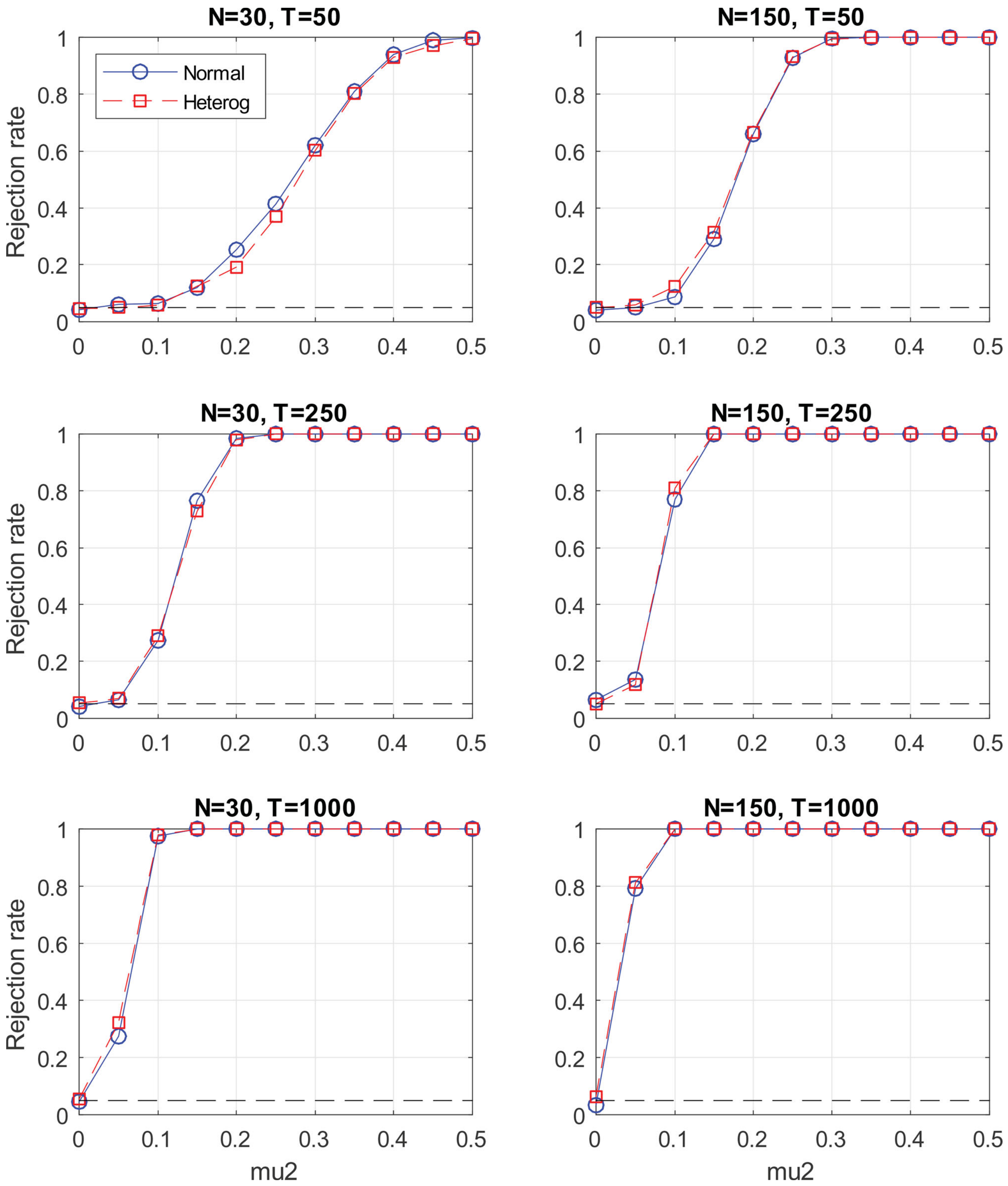

Figure 1. This figure shows test rejection frequencies as a function of μ_2, holding $\mu_1 = 0$, for six different combinations of sample sizes, and for two types of data (Normally and heterogeneously distributed). When $\mu_2 = 0$ the rejection frequency should equal 0.05, the nominal size of the test.

for two representative combinations of sample sizes, $(N, T) = (30, 50)$ and $(150, 250)$. For the smaller sample size the test is unable to reject the null of a single cluster for values of α_2 within about 0.1 of α_1; the sampling variation in the estimated AR(1) parameters is simply too large in that case. As the differences between the cluster AR(1) parameters grows, or if we use a larger sample size, the power of the test increases. For both sample sizes the finite-sample size of the test is close to the nominal value.

Finally, we investigate the performance of the test in Theorem 4, which is applicable when T is small. We consider $T \in \{2, 4, 6, 10\}$, and values of $N \in \{30, 150, 600\}$. Figure 6 shows that even when $T = 2$, the test has reasonable size control: the rejection rate for $N = 30$ is 0.07, so only slightly above

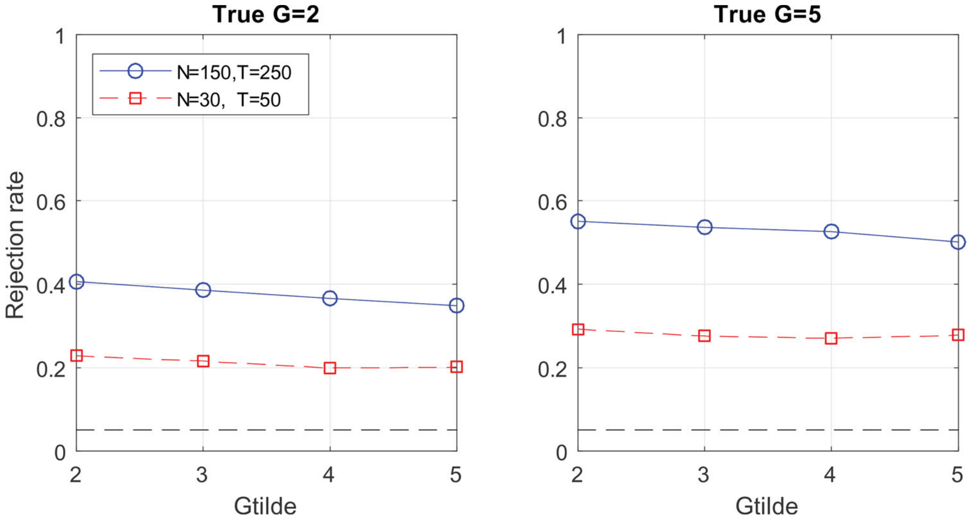

Figure 2. This figure shows test rejection frequencies as a function of $\tilde{G}$, the number of clusters considered under the alternative, for two different combinations of sample sizes. In the left panel the correct number of clusters is 2, while in the right panel it is 5. The distance between the first and last (ordered) cluster means is 0.2 when $(N, T) = (30, 50)$ and 0.075 when $(N, T) = (150, 250)$. The nominal size is 0.05.

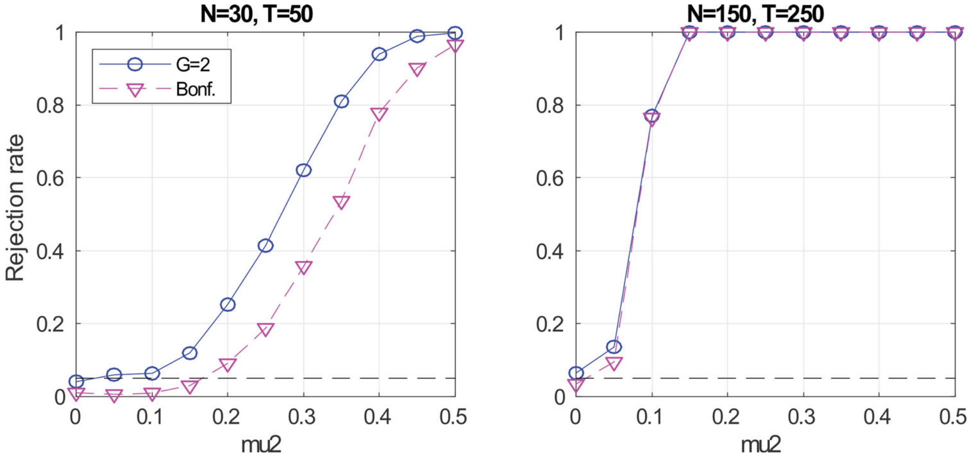

Figure 3. This figure shows test rejection frequencies as a function of μ_2, holding $\mu_1 = 0$, for two different combinations of sample sizes, and for two tests: the first uses $G = 2$ under the alternative, the second considers $G \in \{2, 3, 4, 5\}$ and uses a Bonferroni correction to control for multiple testing. When $\mu_2 = 0$ the rejection frequency should equal 0.05, the nominal size of the test.

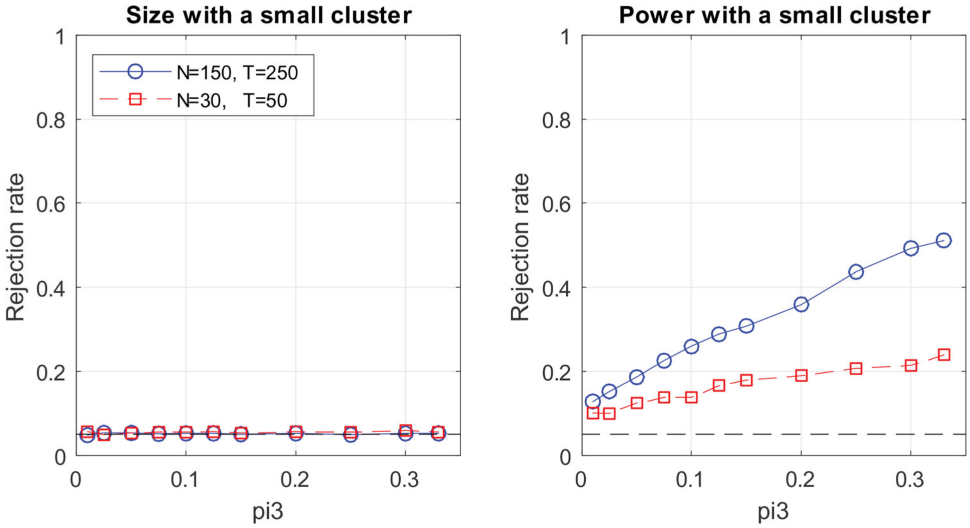

Figure 4. This figure shows test rejection frequencies as a function of π_3, the fraction of variables in the smallest cluster. The fractions of variables in the other two groups is set at $(1 - \pi_3)/2$. Two different combinations of sample sizes are considered. The distance between the first and third (ordered) cluster means is zero in the left panel, and is 0.2 when $(N, T) = (30, 50)$ and 0.075 when $(N, T) = (150, 250)$ in the right panel. The nominal size of the test is 0.05.

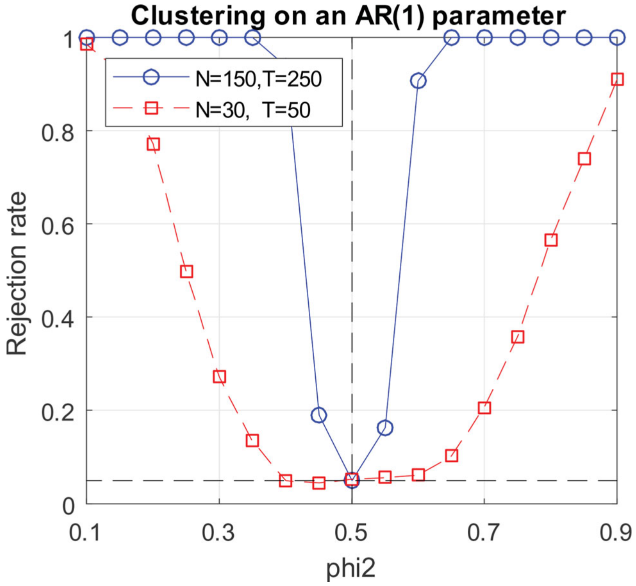

Figure 5. This figure shows test rejection frequencies as a function of ϕ_2, the AR(1) parameter of the second cluster, when the parameter for the first cluster is $\phi_1 = 0.5$. Two different combinations of sample sizes are considered. When $\phi_2 = 0.5$ the rejection frequency should equal 0.05, the nominal size of the test.

the nominal level. (The rejection rates when $N = 30$ and $T = 4$, 6, or 10 are between 0.07 and 0.08.) The power is low for the smallest value of N, but when $N = 150$ or 600 power is nontrivial. As T increases to 4, 6 and 10 we see that size control is maintained, and power increases. Naturally, a test with such small values of T has lower power than for larger values of T, for example, the results in Figure 1, however, the results in Figure 6 show that even for very small values of T, size control is maintained and nontrivial power can be achieved with a large cross-sectional sample size.

4. Empirical Applications

4.1. Vehicle Manufacturer Clusters

To illustrate our methodology in a well-known setting, we use a standard dataset, built into Matlab, on car attributes for 307 vehicle models from 30 manufacturers in seven countries during 1970–1982. Vehicle attributes include acceleration, number of cylinders, engine displacement, horsepower, miles per gallon, and weight, the last of which we log to avoid identifying outliers as separate clusters. We split our data into $\mathcal{R}$ and $\mathcal{P}$ samples of 1970–1975 and 1976–1982, respectively. Within each sample, we average vehicle attributes by manufacturer across all combinations of models and model years, and we retain only observations with nonmissing values of all attributes and manufacturers with models in both samples. Our resulting sample consists of 24 manufacturers. To make scales comparable across characteristics, we standardize each attribute within each sample by demeaning and dividing by the standard deviation across manufacturers.

We assuming $G = 2$ clusters, and use k-means on the $\mathcal{R}$ sample, with 1000 starting values initialized by k-means++ (Arthur and Vassilvitskii 2007). Table 2 summarizes the results. Panel A shows that "Group 1" (American) manufacturers typically produce vehicles with more cylinders, larger engines, greater horsepower, lower mileage, and greater weight than "Group 2" (European and Japanese) manufacturers. Note that all cross-cluster characteristic differences are larger on the $\mathcal{R}$ sample than on the $\mathcal{P}$ sample, consistent with the clustering procedure fitting both true differences among manufacturers as well as noise. The p-value from the test for multiple clusters is less than 0.001, indicating strong evidence in against the null of a single cluster. We conclude that at least two clusters are needed to describe vehicle manufacturers during this period.

In Panel B of Table 2 we present the constituents of each cluster, and a clear pattern emerges: we find that manufacturers cleave perfectly by region of origin, with Group 1 comprised completely of the American manufacturers, and Group 2 containing all the non-American manufacturers.

4.2. Mutual Fund Clusters

Performance evaluation, for example, for mutual funds or hedge funds, is one of the central concerns of empirical finance. Most performance evaluation takes the form of comparing fund returns to a benchmark return, for example, the return on a strategy or style with similar risk characteristics. A popular article in style analysis, Brown and Goetzmann (1997) pioneered the application of k-means clustering for the purpose of benchmark formation and assignment of funds to benchmarks. We use the testing approach proposed in this article to determine whether mutual fund styles are truly distinct in the data. We cluster based on risk exposures (betas) rather than returns themselves (as done in the original study) to facilitate interpretation of the results.

We use daily data from the CRSP Mutual Fund Database, see Patton and Weller (2019) for the data construction, filtering, and aggregation methodology. We use the first full year of the daily series (1999) for the $\mathcal{R}$ sample and the second year (2000) for the $\mathcal{P}$ sample, and we retain only U.S. domestic equity mutual funds that report for at least half the days in each year. The resulting sample consists of 1743 mutual funds.

We run the following regression for each fund:

$$r_{it} = \alpha_i + \sum_{k=1}^{4} \beta_{ik} f_{kt} + \sigma_i \varepsilon_{it} \tag{13}$$

As factors, f_{kt}, we use the value-weighted market (MKT), size (SMB), value (HML), and momentum (UMD) returns of the Carhart (1997) model.[3] We also estimate average abnormal returns (α_i) and idiosyncratic volatility (σ_i) for each fund but we do not cluster on these attributes.

We use k-means clustering on the $\mathcal{R}$ sample, with 1000 starting values initialized by k-means++. We follow Brown and Goetzmann (1997) and use $G = 8$ clusters. Table 3 summarizes the results of the clustering procedure. Fund clusters differ markedly in the parameters on which the clustering was done (the risk exposures, β_{ik}) and interestingly also in the other parameters of the model (α_i and σ_i). For example, annualized average abnormal returns (α_i) range between -3% and 22% across the clusters. This heterogeneity cannot be accommodated by other tests for multiple clusters.

[3] The factor data is available at *http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html*.

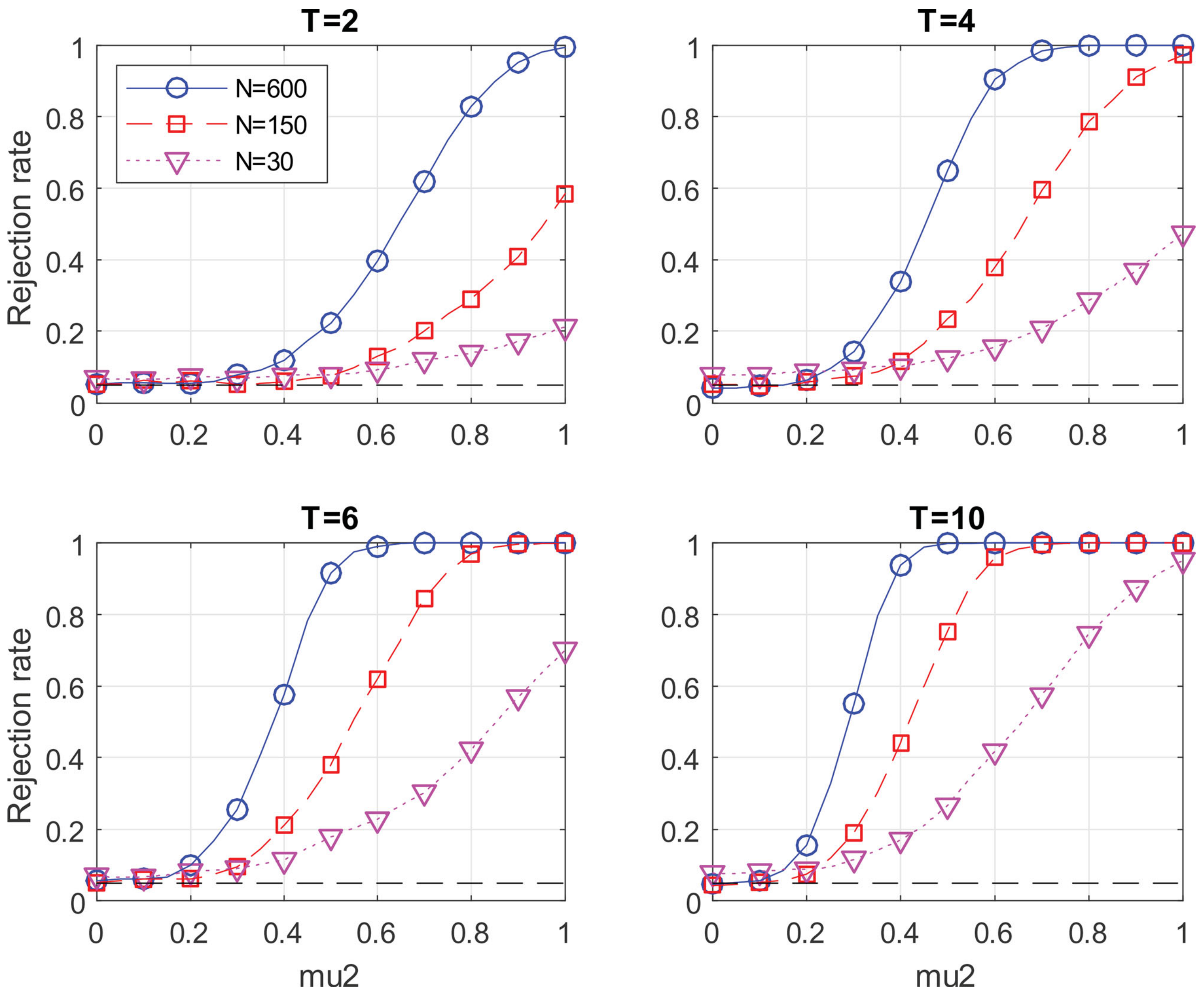

Figure 6. This figure shows test rejection frequencies as a function of μ_2, holding $\mu_1 = 0$, for four different time series sample sizes, and three different cross-sectional sample sizes. When $\mu_2 = 0$ the rejection frequency should equal 0.05, the nominal size of the test.

Table 2. Vehicle manufacturer clusters.

Panel A: Cluster properties by characteristic						
	Acceleration (s to 60mph)	Cylinders (#)	Displacement (in^3)	Horsepower (hp)	MPG (mpg)	Weight (log lbs)
	Normalized values, $\mathcal{R}$ sample					
Group 1	−0.441	0.641	0.709	0.690	−0.639	0.636
Group 2	0.326	−1.110	−1.057	−0.704	0.973	−0.949
	Normalized values, $\mathcal{P}$ sample					
Group 1	−0.125	0.507	0.628	0.450	−0.486	0.573
Group 2	0.152	−0.557	−0.650	−0.318	0.266	−0.461
	Raw values, $\mathcal{P}$ sample					
Group 1	15.76	5.82	217.25	105.02	23.21	8.04
Group 2	16.46	4.21	111.06	83.67	28.94	7.78
Panel B: Cluster assignments						
Group 1	AMC Mercury	Buick Oldsmobile	Chevrolet Plymouth	Chrysler Pontiac	Dodge	Ford
Group 2	Audi Opel Volkswagon	BMW Peugeot Volvo	Datsun Renault	Fiat Saab	Honda Subaru	Mazda Toyota

NOTE: This table presents group averages of manufacturer-level characteristics in a $G = 2$ cluster model for $\mathcal{R}$ (1970–1975) and $\mathcal{P}$ (1976–1982) samples (Panel A) and manufacturer names by group (Panel B). The "raw values" in Panel A are renormalized using the $\mathcal{P}$-sample characteristic means and standard deviations.

Table 3. Mutual fund clusters.

Group	$\hat{N}_g$	$\hat{\beta}_{\text{MKT}}$	$\hat{\beta}_{\text{SMB}}$	$\hat{\beta}_{\text{HML}}$	$\hat{\beta}_{\text{UMD}}$	$\hat{\alpha}$	$\hat{\sigma}$
	$\mathcal{R}$ sample						
1	420	0.905	−0.088	0.073	−0.022	−0.817	4.579
2	198	1.266	0.925	0.287	0.233	17.875	11.263
3	84	1.320	0.677	−0.342	0.414	21.722	15.294
4	238	1.058	0.685	0.705	−0.164	3.500	9.224
5	224	0.840	0.368	0.341	−0.026	5.463	8.469
6	210	0.959	−0.017	−0.286	0.176	0.216	7.561
7	270	1.004	−0.004	0.522	−0.194	0.768	6.791
8	99	0.029	0.047	0.075	0.006	−3.155	4.994
	$\mathcal{P}$ sample						
1	420	0.883	−0.142	0.104	−0.021	3.370	6.238
2	198	1.210	0.807	−0.041	0.080	14.896	13.849
3	84	1.219	0.428	−0.825	0.228	27.074	19.375
4	238	0.975	0.494	0.493	−0.168	12.257	11.096
5	224	0.809	0.288	0.198	−0.067	7.883	10.617
6	210	0.966	−0.061	−0.306	0.118	12.108	11.034
7	270	0.954	−0.077	0.538	−0.175	6.725	9.31
8	99	0.042	0.066	0.029	−0.027	2.398	6.769

NOTE: This table presents group averages of fund-level characteristics in a $G = 8$ group model for $\mathcal{R}$ (year 1999) and $\mathcal{P}$ (year 2000) samples. Average abnormal returns (α) and idiosyncratic volatility (σ) are annualized and reported in percent.

Unlike the two-group example in the previous section, differences between these eight groups, each with four dimensions of characteristics, are more difficult to present in tabular form. Nevertheless, the factor loadings in Table 3 reveal some clear clusters: Group 1, with a loading of near one of the market factor and relatively small loadings on the other three factors, is a "market" style cluster; Group 2, with high loadings on both the market and the size factor, is a "small capitalization" style cluster; Group 7, with high loadings on the aggregate market and value factors, is a value cluster; and Group 8, with factor loadings close to zero on all four factors, is a "market-neutral" style cluster. The p-value from the test for multiple clusters is less than 0.001, indicating strong evidence against the null of a single cluster. We conclude that mutual funds indeed have different styles.

5. Conclusion

This article proposes methods to determine whether a null hypothesis of a single cluster, indicating homogeneity of the data, can be rejected in favor of multiple clusters. The new test is simple to implement and valid under relatively mild conditions, including nonnormality, and heterogeneity of the data in aspects beyond those in the clustering analysis. We show via an extensive simulation study that the test has good finite-sample size control. We present extensions of the test for a range of applications, including clustering when the time series dimension is small, or clustering on parameters other than the mean.

Some interesting extensions remain. For example, García-Escudero and Gordaliza (1999) propose a robust version of k-means based on trimmed means, Witten and Tibshirani (2010) propose a method to optimally choose the features on which to cluster, and Ng, Jordan and Weiss (2002) propose a spectral clustering method. Another interesting extension is rethinking how the number of clusters is treated in the asymptotic theory: The results in this article apply when $\tilde{G}$ is small relative to the sample size, however, if a researcher were to consider a large number of clusters, then an asymptotic framework in which $\tilde{G}$ is modeled as diverging with the sample size may provide better finite-sample approximations. In an estimation context, such a framework is studied by Bester and Hansen (2016) with known group assignments and by Bonhomme, Lamadon, and Manresa (2021) with estimated group assignments. Recent work on high-dimensional Gaussian approximations and uniform confidence bounds on parameter vectors of diverging lengths, see Chernozhukov, Lee, and Rosen (2013), Belloni et al. (2015) and Li and Liao (2020) for example, may also be useful for such an analysis. We leave these interesting extensions for future research.

Appendix A: Proofs

Proof of Theorem 1. (a) We first find the limiting distribution of $\sqrt{NP}\tilde{\boldsymbol{\mu}}_{NP}(\hat{\gamma}_{NR})$ conditional on $\mathcal{F}_R$. Denote $\hat{N}_{g,R} \equiv \sum_{i=1}^N \mathbf{1}\{\hat{\gamma}_{g,NR} = 1\}$ and $\hat{\pi}_{g,R} = \hat{N}_{g,R}/N$, and note that

$$\tilde{\boldsymbol{\mu}}_{g,NP}(\hat{\gamma}_{NR}) = \frac{1}{\hat{N}_{g,R}} \sum_{i=1}^{N} \left(\mathbf{1}\{\hat{\gamma}_{i,NR} = g\} \frac{1}{P} \sum_{t \in \mathcal{P}} \mathbf{Y}_{i,t} \right)$$
$$= \frac{1}{NP} \sum_{i=1}^{N} \sum_{t \in \mathcal{P}} \mathbf{Y}_{i,t} \hat{\pi}_{g,R}^{-1} \mathbf{1}\{\hat{\gamma}_{i,NR} = g\}$$

for $g = 1, \ldots, G$. Thus,

$$\sqrt{NP}\left(\tilde{\boldsymbol{\mu}}_{g,NP}(\hat{\gamma}_{NR}) - \boldsymbol{\mu}_g^*\right) = \frac{1}{\sqrt{NP}} \sum_{i=1}^{N} \sum_{t \in \mathcal{P}} \left(\boldsymbol{\mu}_g^* + \boldsymbol{\varepsilon}_{it}\right)$$
$$\hat{\pi}_{g,R}^{-1} \mathbf{1}\{\hat{\gamma}_{i,NR} = g\} - \sqrt{NP}\boldsymbol{\mu}_g^*$$
$$= \frac{1}{\sqrt{NP}} \sum_{i=1}^{N} \sum_{t \in \mathcal{P}} \hat{U}_{ig,NR} \boldsymbol{\varepsilon}_{it}$$

where $\hat{U}_{ig,NR} \equiv \hat{\pi}_{g,R}^{-1} \mathbf{1}\{\hat{\gamma}_{i,NR} = g\}$. Note that this variable is bounded as $\hat{\pi}_{g,R} \geq \underline{\pi} > 0$. Conditional on $\mathcal{F}_R$, the sequence $\left\{\hat{U}_{ig,NR}\boldsymbol{\varepsilon}_{it}\right\}$ is independent and heterogeneously distributed. Define

$$\bar{\Omega}_{gNR} \equiv V\left[\frac{1}{\sqrt{NP}} \sum_{i=1}^{N} \sum_{t \in \mathcal{P}} \hat{U}_{ig,NR} \boldsymbol{\varepsilon}_{it} \middle| \mathcal{F}_R \right] = \frac{1}{N} \sum_{i=1}^{N} \hat{U}_{ig,NR}^2 \Sigma_i$$

where $\Sigma_i \equiv V[\boldsymbol{\varepsilon}_{it}]$ and the second line holds as $\boldsymbol{\varepsilon}_{it}$ is uncorrelated in the time series and cross section. Combining the Cramér-Wold device with Theorem 5.11 of White (2001), for example, we then obtain the asymptotic distribution of $\tilde{\boldsymbol{\mu}}_{g,NP}(\hat{\gamma}_{NR})$:

$$\sqrt{NP}\bar{\Omega}_{gNR}^{-1/2}\left(\tilde{\boldsymbol{\mu}}_{g,NP}(\hat{\gamma}_{NR}) - \boldsymbol{\mu}_g^*\right) \xrightarrow{d} N(0, I)$$

This holds for each $g = 1, \ldots, G$. Next we show that $\text{Cov}\Big[\tilde{\boldsymbol{\mu}}_{g,NP}(\hat{\gamma}_{NR}),$ $\tilde{\boldsymbol{\mu}}_{g',NP}(\hat{\gamma}_{NR})\Big] = 0$ for all $g \neq g'$. Define $\bar{\varepsilon}_{ikP} \equiv \frac{1}{P}\sum_{t \in \mathcal{P}} \varepsilon_{itk}$, and consider elements (k, k') of the vector $\left(\tilde{\boldsymbol{\mu}}_{g,NP}(\hat{\gamma}_{NR}) - \boldsymbol{\mu}_g^*\right)$. The covariance between any two elements (k, k') in groups $g \neq g'$ is

$$E\Big[\left(\tilde{\mu}_{gk,NP}(\hat{\gamma}_{NR}) - \mu_{gk}^*\right)\left(\tilde{\mu}_{g'k',NP}(\hat{\gamma}_{NR}) - \mu_{g'k'}^*\right)\Big| \mathcal{F}_R\Big]$$
$$= \frac{1}{N^2} E\left[\left(\sum_{i=1}^{N} \hat{\pi}_{g,R}^{-1} \mathbf{1}\{\hat{\gamma}_{i,NR} = g\} \bar{\varepsilon}_{ikP} \right) \left(\sum_{j=1}^{N} \hat{\pi}_{g',R}^{-1} \mathbf{1}\left\{\hat{\gamma}_{j,NR} = g'\right\} \bar{\varepsilon}_{jk'P} \right) \middle| \mathcal{F}_R \right]$$
$$= 0$$

since $\mathbf{1}\left\{\hat{\gamma}_{i,NR}=g\right\}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=g'\right\}=0$ for $g\neq g'$. Thus, we obtain the limiting distribution for the entire vector $\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\gamma}_{NR}\right)$:

$$\sqrt{NP}\bar{\Omega}_{NR}^{-1/2}\left(\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\gamma}_{NR}\right)-\boldsymbol{\mu}^{*}\right)\xrightarrow{d}N\left(0,I\right)$$

where $\bar{\Omega}_{NR}$ is block-diagonal, with $\left(\bar{\Omega}_{1NR},\ldots,\bar{\Omega}_{GNR}\right)$ along the diagonal. Consider the following estimator of $\bar{\Omega}_{gNR}$:

$$\begin{aligned}\hat{\Omega}_{gNPR}&=\frac{1}{NP}\sum_{t\in\mathcal{P}}\sum_{i=1}^{N}\hat{U}_{ig,NR}^{2}\left(\mathbf{Y}_{it}-\bar{\mathbf{Y}}_{i}\right)\left(\mathbf{Y}_{it}-\bar{\mathbf{Y}}_{i}\right)'\\&=\frac{1}{NP}\sum_{t\in\mathcal{P}}\sum_{i=1}^{N}\hat{U}_{ig,NR}^{2}\hat{\boldsymbol{\varepsilon}}_{it}\hat{\boldsymbol{\varepsilon}}_{it}'\end{aligned}$$

This can be shown to be consistent for $\bar{\Omega}_{gNR}$ using Kolmogorov's law of large numbers for independent heterogeneous data (e.g., White 2001, Theorem 3.7), and noting that Assumption 1(a) ensures the 2+δ moment condition on $\boldsymbol{\varepsilon}_{it}$ and the finiteness of $\Sigma_{i}\ \forall i$. This holds for all g, and so we have $\hat{\Omega}_{NPR}-\bar{\Omega}_{NR}\xrightarrow{p}0$. This implies that

$$\sqrt{NP}\hat{\Omega}_{NPR}^{-1/2}\left(\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\gamma}_{NR}\right)-\boldsymbol{\mu}^{*}\right)\xrightarrow{d}N\left(\mathbf{0},I\right)$$

Under the null hypothesis of one cluster we have $\boldsymbol{\mu}^{*}=\boldsymbol{\iota}_{G}\otimes\boldsymbol{\mu}^{\sharp}$ for some $(d\times 1)$ vector $\boldsymbol{\mu}^{\sharp}$, which implies that $A_{dG}\boldsymbol{\mu}^{*}=\mathbf{0}_{d(G-1)}$. Thus, the F -statistic obeys

$$\begin{aligned}&F_{NPR}=NP\tilde{\boldsymbol{\mu}}_{NP}'\left(\hat{\gamma}_{NR}\right)A_{d,G}'\left(A_{d,G}\hat{\Omega}_{NPR}A_{d,G}'\right)^{-1}\\&A_{d,G}\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\gamma}_{NR}\right)\xrightarrow{d}\chi_{d(G-1)}^{2}\end{aligned}$$

As the limiting distribution of the F-statistic does not depend on $\mathcal{F}_{R}$, its unconditional distribution is also $\chi_{d(G-1)}^{2}$, completing the proof.

(b) Note that $\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\gamma}_{NR}\right)-\boldsymbol{\mu}^{*}=\left(\hat{\boldsymbol{\mu}}_{NR}-\boldsymbol{\mu}^{*}\right)+\left(\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\gamma}_{NR}\right)-\hat{\boldsymbol{\mu}}_{NR}\right)$. Our Assumption 1 is sufficient for Assumption 1 of Bonhomme and Manresa (2015), and their Theorem 1 implies that the first term is $o_{p}(1)$ as $N,R\rightarrow\infty$. The second term is

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{g,NP}\left(\hat{\gamma}_{NR}\right)-\hat{\boldsymbol{\mu}}_{g,NR}&=\frac{1}{NP}\sum_{i=1}^{N}\sum_{t\in\mathcal{P}}\mathbf{Y}_{i,t}\hat{\pi}_{g,R}^{-1}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=g\right\}\\&\quad-\frac{1}{NR}\sum_{i=1}^{N}\sum_{t\in\mathcal{R}}\mathbf{Y}_{i,t}\hat{\pi}_{g,R}^{-1}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=g\right\}\\&=\frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_{g,R}^{-1}\mathbf{1}\left\{\hat{\gamma}_{i,NR}=g\right\}\left(\frac{1}{P}\sum_{t\in\mathcal{P}}\boldsymbol{\varepsilon}_{i,t}-\frac{1}{R}\sum_{t\in\mathcal{R}}\boldsymbol{\varepsilon}_{i,t}\right)\\&\leq\underline{\pi}^{-1}\left(\frac{1}{NP}\sum_{i=1}^{N}\sum_{t\in\mathcal{P}}\boldsymbol{\varepsilon}_{i,t}-\frac{1}{NR}\sum_{i=1}^{N}\sum_{t\in\mathcal{R}}\boldsymbol{\varepsilon}_{i,t}\right)\\&=\underline{\pi}^{-1}\left(O_{p}\left((NP)^{-1/2}\right)+O_{p}\left((NR)^{-1/2}\right)\right)\\&=o_{p}(1)\text{, as }N,P,R\rightarrow\infty.\end{aligned}$$

The penultimate line follows from a law of large numbers of independent heterogeneous data (e.g., White 2001, Theorem 3.7) the conditions for which are satisfied given our Assumption 1. This holds for $g=1,\ldots,G$, and thus $\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\gamma}_{NR}\right)\xrightarrow{p}\boldsymbol{\mu}^{*}$ as $N,P,R\rightarrow\infty$. This implies that

$$\begin{aligned}&\tilde{\boldsymbol{\mu}}_{NP}'\left(\hat{\boldsymbol{\gamma}}_{NR}\right)A_{d,G}'\left(A_{d,G}\hat{\Omega}_{NPR}A_{d,G}'\right)^{-1}\\&\qquad A_{d,G}\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\xrightarrow{p}\boldsymbol{\mu}^{*\prime}A_{d,G}'\left(A_{d,G}\bar{\Omega}_{NR}A_{d,G}'\right)^{-1}A_{d,G}\boldsymbol{\mu}^{*}>0\end{aligned}$$

by Assumption 2′(b) (clusters are "well separated"), the positive definiteness of $\bar{\Omega}_{NR}$, and the full row rank of $A_{d,G}$. Thus,

$$\begin{aligned}&F_{NPR}=NP\tilde{\boldsymbol{\mu}}_{NP}'\left(\hat{\boldsymbol{\gamma}}_{NR}\right)A_{d,G}'\left(A_{d,G}\hat{\Omega}_{NPR}A_{d,G}'\right)^{-1}\\&A_{d,G}\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\xrightarrow{p}\infty,\text{ as }N,P,R\rightarrow\infty\end{aligned}$$

completing the proof. □

Proof of Theorem 3. (a) This is done using the same methods as Theorem 1 for $d=1$ and $G=\hat{G}_{NR}$. We note that the reordering of the clusters (from largest to smallest) is known given $\mathcal{F}_{R}$, as is the value of $\hat{G}_{NR}$. Thus, following the steps in the proof of Theorem 1(a) we have $F_{NPR}\xrightarrow{d}\chi_{q}^{2}$ where $q=\hat{G}_{NR}-1$. This limit distribution depends on $\mathcal{F}_{R}$ via the value of $\hat{G}_{NR}$; by transforming the test statistic using its limiting CDF we obtain $Pval_{NPR}\xrightarrow{d}\text{Unif}\,(0,1)$ conditional on $\mathcal{F}_{R}$, and since the limit distribution does not depend on $\mathcal{F}_{R}$, this result also holds unconditionally.

(b) As in the proof of Theorem 1(b), note that $\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)-\boldsymbol{\mu}^{*}=\left(\hat{\boldsymbol{\mu}}_{NR}-\boldsymbol{\mu}^{*}\right)+\left(\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)-\hat{\boldsymbol{\mu}}_{NR}\right)$. Our Assumption 1 is sufficient for Assumption 1 of Bonhomme and Manresa (2015), and their Theorem 1 implies that the first term on the RHS is $o_{p}(1)$, as $N,R\rightarrow\infty$. (Note that their Theorem 1 does not require nonnegligibility of group sizes.) The first $\hat{G}_{NR}$ elements of the second term are $o_{p}(1)$ as $N,P,R\rightarrow\infty$ using the same derivation as in the proof of Theorem 1(b), noting that the condition that $\underline{\pi}>0$ holds for $g\in\left\{1,\ldots,\hat{G}_{NR}\right\}$,and we have $\hat{G}_{NR}\geq 2$ by Assumption 2′(c^{S}). This implies that

$$\begin{aligned}&\tilde{\boldsymbol{\mu}}_{NP}'\left(\hat{\boldsymbol{\gamma}}_{NR}\right)B_{\hat{G}_{NR},G}'\left(B_{\hat{G}_{NR},G}\hat{\Omega}_{NPR}B_{\hat{G}_{NR},G}'\right)^{-1}B_{\hat{G}_{NR},G}\tilde{\boldsymbol{\mu}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\\&\rightarrow\boldsymbol{\mu}^{*\prime}B_{\hat{G}_{NR},G}'\left(B_{\hat{G}_{NR},G}\hat{\Omega}_{NPR}B_{\hat{G}_{NR},G}'\right)^{-1}B_{\hat{G}_{NR},G}\boldsymbol{\mu}^{*}>0\end{aligned}$$

by Assumption 2′(b), the positive definiteness of $\bar{\Omega}_{NP}$, and the full row rank of $B_{\hat{G}_{NR},G}$. Thus, $F_{NPR}\xrightarrow{p}\infty$ and $P_{NPR}\xrightarrow{p}0$ as $N,P,R\rightarrow\infty$. □

Proof of Theorem 4. The proof of this theorem follows from that of Theorem 5, using the mean as the parameter on which to cluster, and specializing to the $d=1$ and $G=2$ case. □

Proof of Theorem 5. (a) As in the main theorem, we have

$$\tilde{\boldsymbol{\alpha}}_{g,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)=\frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_{g,R}\mathbf{1}\left\{\gamma_{i}=g\right\}\hat{\boldsymbol{\beta}}_{i,P}$$

for $g=1,\ldots,G$. Then

$$\begin{aligned}\sqrt{N}\left(\tilde{\boldsymbol{\alpha}}_{g,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)-\boldsymbol{\alpha}_{g}^{*}\right)&=\frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_{g,R}\mathbf{1}\left\{\gamma_{i}=g\right\}\left(\boldsymbol{\alpha}^{*}+\boldsymbol{\varepsilon}_{i,S}-\boldsymbol{\alpha}^{*}\right)\\&=\frac{1}{N}\sum_{i=1}^{N}U_{ig,R}\boldsymbol{\varepsilon}_{i,S}\end{aligned}$$

where $U_{ig,R}\equiv\hat{\pi}_{g,R}\mathbf{1}\left\{\gamma_{i}=g\right\}$. Conditional on $\mathcal{F}_{R}$ the sequence $\left\{U_{ig,R}\boldsymbol{\varepsilon}_{i,S}\right\}$ is *i.n.i.d.* Our Assumption 4 is sufficient for Assumptions 1, 2, and 3(b) of Hansen (2007, Theorem 1) and thus we have

$$\sqrt{N}\bar{\Omega}_{gNR}^{-1/2}\left(\tilde{\boldsymbol{\alpha}}_{g,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)-\boldsymbol{\alpha}_{g}^{*}\right)\xrightarrow{d}N\left(0,I\right)\text{, for }g=1,\ldots,G$$

$$\text{where }\ \bar{\Omega}_{gNR}\equiv V\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}U_{ig,R}\boldsymbol{\varepsilon}_{i,S}\,\middle|\,\mathcal{F}_{R}\right]=\frac{1}{N}\sum_{i=1}^{N}U_{ig,R}^{2}\Sigma_{i}$$

Stacking the parameter estimates for each group, and noting that the errors are uncorrelated across groups, we obtain

$$\sqrt{N}\bar{\Omega}_{NR}^{-1/2}\left(\tilde{\boldsymbol{\alpha}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)-\boldsymbol{\alpha}^{*}\right)\xrightarrow{d}N\left(0,I\right)$$

where $\bar{\Omega}_{NR}$ is block-diagonal, with $\left(\bar{\Omega}_{1NR},\ldots,\bar{\Omega}_{GNR}\right)$ along the diagonal. Following Hansen (2007, Theorem 1) we can consistently estimate $\bar{\Omega}_{gNR}$ using

$$\hat{\Omega}_{gNPR}=\frac{1}{N}\sum_{i=1}^{N}U_{ig,R}^{2}\left(\hat{\boldsymbol{\beta}}_{i,P}-\tilde{\boldsymbol{\alpha}}_{g,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\right)^{2}$$

which implies

$$\sqrt{N}\hat{\Omega}_{NPR}^{-1/2}\left(\tilde{\boldsymbol{\alpha}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)-\boldsymbol{\alpha}^{*}\right)\xrightarrow{d}N\left(0,I\right)$$

Under the null hypothesis we have $\boldsymbol{\alpha}^{*}=\boldsymbol{\iota}_{G}\otimes\boldsymbol{\alpha}^{\sharp}$ for some $(b\times 1)$ vector $\boldsymbol{\alpha}^{\sharp}$, implying that $A_{bG}\boldsymbol{\alpha}^{*}=0_{b(G-1)}$. Thus, the F-statistic obeys

$$\begin{aligned}F_{NPR}=N\tilde{\boldsymbol{\alpha}}_{NP}'\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\\ A_{bG}'\left(A_{bG}\hat{\Omega}_{NPR}A_{bG}'\right)^{-1}A_{bG}\tilde{\boldsymbol{\alpha}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\xrightarrow{d}\chi_{b(G-1)}^{2}\end{aligned}$$

As the limiting distribution does not depend on $\mathcal{F}_{R}$, its unconditional distribution is also $\chi_{b(G-1)}^{2}$, completing the proof.

(b) Note that $\tilde{\boldsymbol{\alpha}}_{g,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)-\boldsymbol{\alpha}^{*}=\left(\hat{\boldsymbol{\alpha}}_{g,NR}-\boldsymbol{\alpha}^{*}\right)+\left(\tilde{\boldsymbol{\alpha}}_{g,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)-\hat{\boldsymbol{\alpha}}_{g,NR}\right)$. Our Assumption 4 is sufficient for the assumptions in Bonhomme and Manresa (2015, Appendix S2.2.1), which implies that the first term on the RHS is $o_{p}\left(1\right)$, as $N\rightarrow\infty$. The second term is

$$\begin{aligned}\tilde{\boldsymbol{\alpha}}_{g,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)-\hat{\boldsymbol{\alpha}}_{g,NR}&=\frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_{g,R}\mathbf{1}\left\{\gamma_{i}=g\right\}\left(\hat{\boldsymbol{\beta}}_{i,P}-\hat{\boldsymbol{\beta}}_{i,R}\right)\\&=\frac{1}{N}\sum_{i=1}^{N}U_{ig,R}\left(\boldsymbol{\varepsilon}_{i,P}-\boldsymbol{\varepsilon}_{i,R}\right)\\&=o_{p}\left(1\right)\end{aligned}$$

under Assumption 4 and noting that $U_{ig,R}$ is bounded. Thus, $\tilde{\boldsymbol{\alpha}}_{g,NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\xrightarrow{p}\boldsymbol{\alpha}^{*}$ as $N\rightarrow\infty$. This implies that

$$\begin{aligned}\tilde{\boldsymbol{\alpha}}_{NP}'\left(\hat{\boldsymbol{\gamma}}_{NR}\right)A_{bG}'\left(A_{bG}\hat{\Omega}_{NPR}A_{bG}'\right)^{-1}\\ A_{bG}\tilde{\boldsymbol{\alpha}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\xrightarrow{p}\boldsymbol{\alpha}^{*\prime}A_{bG}'\left(A_{bG}\bar{\Omega}_{NR}A_{bG}'\right)^{-1}A_{bG}\boldsymbol{\alpha}^{*}>0\end{aligned}$$

by Assumption $2'_{P}$ (clusters are well separated), the positive definiteness of $\bar{\Omega}_{NR}$, and the full row rank of A_{bG}. Thus,

$$\begin{aligned}F_{NPR}=N\tilde{\boldsymbol{\alpha}}_{NP}'\left(\hat{\boldsymbol{\gamma}}_{NR}\right)A_{bG}'\left(A_{bG}\hat{\Omega}_{NPR}A_{bG}'\right)^{-1}\\ A_{bG}\tilde{\boldsymbol{\alpha}}_{NP}\left(\hat{\boldsymbol{\gamma}}_{NR}\right)\rightarrow\infty\text{ as }N\rightarrow\infty\end{aligned}$$

completing the proof. □

Supplementary Materials

The supplemental appendix contains additional theoretical and simulation results.

Acknowledgments

For helpful comments we thank the editor and two referees, as well as Tim Bollerslev and Jia Li.

References

Arthur, D., and Vassilvitskii, S. (2007), "K-means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07*, pp. 1027–1035, Philadelphia, PA, USA: Society for Industrial and Applied Mathematics. [746]

Belloni A., Victor Chernozhukov, V., Chetverikov, D., and Kato, K. (2015), "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186, 345–366. [748]

Bester, C. A., and Hansen, C. B. (2016), "Grouped Effects Estimators in Fixed Effects Models," *Journal of Econometrics*, 190, 197–208. [748]

Bonhomme, S., and Manresa, E. (2015), "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83, 1147–1184. [737,739,749,750]

Bonhomme, S., Lamadon, T., and Manresa, E. (2021), "Discretizing Unobserved Heterogeneity," *Econometrica*, forthcoming. [748]

Brown, S. J., and Goetzmann, W. N. (1997), "Mutual Fund Styles," *Journal of Financial Economics*, 43, 373–399. [746]

Carhart, M. M. (1997), "On Persistence in Mutual Fund Performance," *Journal of Finance*, 52, 57–82. [746]

Chernozhukov, V., Lee, S., and Rosen, A. M. (2013), "Intersection Bounds: Estimation and Inference, *Econometrica*, 81, 667–737. [748]

Diebold, F. X., and Mariano, R. S. (1995), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263. [737]

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy Science*, 95, 14863–14868. [737]

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. [737]

Francis, N., Owyang, M. T., and Savascin, Ö. (2017), "An Endogenously Clustered Factor Approach to International Business Cycles," *Journal of Applied Econometrics*, 32, 1261–1276. [737]

Fu, W., and Perry, P. O. (2017), "Estimating the Number of Clusters Using Cross-Validation," working paper, Stern School of Business, New York University. [737]

García-Escudero, L. A., and Gordaliza, A. (1999), "Robustness Properties of *k* means and Trimed *k* Means," *Journal of the American Statistical Association*, 94, 956–969. [748]

Hansen, C. B. (2007), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When *T* is Large," *Journal of Econometrics*, 141, 597–620. [737,739,749,750]

Li, J., and Liao, Z. (2020), "Uniform Nonparametric Inference for Time Series," *Journal of Econometrics*, 219, 38–51. [748]

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2008), "Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data," *Journal of the American Statistical Association*, 103, 1281–1293. [737,738,742]

Maitra, R., Melnykov, V., and Lahiri, S. N. (2012), "Bootstrapping for Significance of Compact Clusters in Multidimensional Datasets," *Journal of the American Statistical Association*, 107, 378–392. [737,738,742]

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002), "On Spectral Clustering: Analysis and an Algorithm," in *Advances in Neural Information Processing Systems*, pp. 849–856. [748]

Patton, A. J., and Weller, B. M. (2019), "What You See is not What You Get: The Costs of Trading Market Anomalies," *Journal of Financial Economics*. [737,746]

Pollard, D. (1981), Strong Consistency of *k*-Means Clustering," *Annals of Statistics*, 9, 135–140. [739]

——— (1982), "A Central Limit Theorem for *k*-means Clustering," *Annals of Probability*, 10, 919–926. [739]

Ray, S., and Turi, R. H. (1999), "Determination of Number of Clusters in k-Means Clustering and Application in Colour Image Segmentation," in *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*, pp. 137–143. [737]

Steinbach, M., Karypis, G., and Kumar, V. (2000), "A Comparison of Document Clustering Techniques," in *KDD workshop on Text mining*, 400, 525–526. [737]

Sugar, C. A., and James, G. M. (2003), "Finding the Number of Clusters in a Dataset," *Journal of the American Statistical Association*, 98, 750–763. [737]

Tibshirani, R., Walther, G., and Hastie, T. J. (2003), "Estimating the Number of Clusters in a Dataset via the Gap Statistic," *Journal of the Royal Statistical Society*, Series B, 63, 411–423. [737]

Tibshirani, R., and Walther, G. (2005), "Cluster Validation by Prediction Strength," *Journal of Computational and Graphic Statistics*, 14, 511–528. [737]

Wang, J. (2010), "Consistent Selection of the Number of Clusters via Cross-validation," *Biometrika*, 97, 893–904. [737]

West, K. D. (1996), "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067–1084. [737]

White, H. (2001), *Asymptotic Theory for Econometricians (2nd ed.)*, San Diego, CA: Academic Press. [748,749]

Witten, D. M., and Tibshirani, R. (2010), "A Framework for Feature Selection in Clustering," *Journal of the American Statistical Association*, 105, 713–726. [748]