Testing for Clustering Under Switching

Igor Custodio João*

* Vrije Universiteit Amsterdam and Tinbergen Institute

November 2023

Abstract

We refine the test for clustering of Patton and Weller (2022) to allow for cluster switching. In a multivariate panel setting, clustering on time-averages produces consistent estimators of means and group assignments. Once switching is introduced, we lose the consistency. In fact, under switching the time-averaged k-means clustering converges to equal, indistinguishable means. This causes the test for a single cluster to lose power under the alternative of multiple clusters. Power can be regained by clustering the N times T observations independently and carefully subsampling the time dimension. When applied to the empirical setting of Bonhomme and Manresa (2015) of an autoregression of democracy in a panel of countries, we are able to detect clusters in the data under noisier conditions than the original test.

1 Introduction

We develop a test for clustering based on Patton and Weller (2022) that has power under the alternative hypothesis that units of a panel switch clusters over time. Clustering identifies patterns of heterogeneity in the data and can be used to find an interpretable group structure or reduce the dimensionality of the parameter space, for example. A wide range of methods have been proposed, some designed to identify particular patterns. We focus here on clustering of means in multivariate panel data. That is, each observation is drawn from one of several distributions with different means, playing the role of their true clusters. Variations of the popular k-means algorithm can then be used to identify and estimate the cluster means given the true number of clusters. Selecting this number, however, is not a trivial task. The most used methods rely on calculating an array of measures of goodness-of-fit for each possible number of clusters considered, and condensing the information contained in each measure to turn it into a decision. See Tibshirani et al. (2001) for a comparison of several popular techniques. But the case of a single cluster is at the periphery of this literature, however important it may be. Indeed, several goodness-of-fit measures such as the Silhouette are not defined in this setting. Patton and Weller (2022) elegantly bridge this gap by proposing a test for the null hypothesis that there is only a single cluster in the data. In their setting, each unit of the panel belongs to a single cluster, and, in combination with a clustering method, they are able to consistently estimate the cluster means. Relying on the central limit theorem of Pollard (1982) to guarantee the normality of the estimated means, they can finally derive an F test for the difference in cluster means. Importantly, they do not allow for the panel units to switch clusters over time.

This paper concerns the case where panel units can belong to different clusters over time. This is not a threat for the validity of the Patton and Weller (2022) test (hereafter P&W), as under the null hypothesis of a single cluster switching does not make any difference. Its power, however, can be seriously hindered, as exemplified by our simulation study. We modify their test to allow for cluster switching by clustering each observation independently, as a flattened cross-section, and proposing a suitable subsampling scheme. In doing so we mostly ignore the

unit dimension of the panel. This means that our test, and the clustering method underneath, allow for arbitrary, non-parametric switching. We show that under the alternative hypothesis allowing for switching, the P&W test relies on estimates of the cluster means that are biased towards each other. In contrast, our test produces estimates that are biased away from each other under the alternative. This difference inflates our test statistic under the alternative hypothesis only, while maintaining the validity under the null.

We illustrate the power gain of our test by revisiting the application of Bonhomme and Manresa (2015). They use a variation of the k-means algorithm in the context of a regression, using clustering to identify groups in a panel of countries with common parameters. Their application is of special interest for us because, although they do not allow for countries to switch clusters, their regression includes grouped time fixed effects that are interpreted as transitional clusters. We can model them alternatively as cluster switching.

Our paper is not the first to model cluster switching in multivariate panel data. See Catania (2021) for a score-driven mixture model with time-varying weights. Munro and Ng (2022) model a group-structure in survey responses through latent Dirichlet analysis, allowing for time-varying group membership. Custodio João, Lucas, Schaumburg, and Schwaab (2022) use a score-driven hidden Markov model to account for cluster switching in the context of banks' business models. Custodio João, Schaumburg, Lucas, and Schwaab (2022) use a similarly non-parametric clustering model to allow for cluster switching.

Closely related is the literature on regime switching in Markov models. For example, Cho and White (2007) propose a test for regime switching with the null hypothesis of a single regime against the alternative of two regimes in panel data. Although our aim is similar, their setting does not allow for different units being subject to different regimes (or clusters) concurrently. Lumsdaine, Okui, and Wang (2023) introduce structural breaks in a panel data model with grouped heterogeneity. This includes cluster switching at the break points. The main difference to ours is that we allow for unconstrained switching that need not be synchronized across the units of the panel.

This paper is organized as follows. The following section briefly presents the

P&W test. We develop our modified test in Section 2. In Section 3 we present a simulation study to expose the cases where there is a difference in the power of these two tests. Section 4 offers some insights on why there is a difference in power in a simplified setting. Section 5 revisits the application of Bonhomme and Manresa (2015) to compare the tests. Finally, Section 6 concludes.

1.1 Testing for clustering under switching

P&W proposed a statistical test for the null hypothesis that there is no clustering in the data. In their setting a vector Y_{it} of dimension d is observed for $i = 1, \ldots, N$ units over $t = 1, \ldots, T$ periods, and comes from $Y_{it} = m_i + \varepsilon_{it}$, where ε_{it} are i.i.d. and have zero mean and covariance Σ_i. Then, the null hypothesis can be expressed as $m_i = \mu^* \; \forall \; i$.

As we develop on their test, it is important to be detailed on how it was originally constructed. They employ a sample splitting scheme, dividing the T periods in a group $\mathcal{R}$ of size R, and a group $\mathcal{P}$ of size P. As there is no time-dependecy in their DGP, it does not matter how the sample is split. They suggest an equal split between the first and second halves. It is assumed that units can belong to one of G groups. The assignment of unit i is denoted by $\gamma_i \in \{1, \ldots, G\}$, and γ denotes the vector of stacked γ_i. The mean vector of group g is denoted by μ_g and the stacked vectors are denoted by μ.

The procedure begins by estimating means and cluster assignments jointly on the $\mathcal{R}$ sample using a k-means procedure. The estimators are denotes $\hat{\mu}_{NR}$ and $\hat{\gamma}_{NR}$. We can write the estimator as:

$$(\hat{\mu}_{NR}, \hat{\gamma}_{NR}) = \arg\min_{\mu,\gamma} \frac{1}{NR} \sum_{i=1}^{N} \sum_{t \in \mathcal{R}} \sum_{g=1}^{G} ||Y_{it} - \mu_g||^2 \mathbb{1}\{\gamma_i = g\}$$

Next the means on the $\mathcal{P}$ sample are calculate based on the estimated $\hat{\gamma}_{NR}$. This is simply an averaging:

$$\tilde{\mu}_{NP}(\hat{\gamma}_{NR}) = \arg\min_{\mu} \frac{1}{NP} \sum_{i=1}^{N} \sum_{t \in \mathcal{P}} \sum_{g=1}^{G} ||Y_{it} - \mu_g||^2 \mathbb{1}\{\hat{\gamma}_{i,NR} = g\} \tag{1}$$

Cluster variances ($d \times d$) are estimated by:

$$\hat{\Omega}_{g,NPR} = \frac{1}{NP} \sum_{t \in \mathcal{P}} \sum_{i=1}^{N} (Y_{it} - \bar{Y}_{it})(Y_{it} - \bar{Y}_{it})' \hat{\pi}_{g,NR}^{-2} \mathbb{1}\{\hat{\gamma}_{i,NR} = g\}$$

where $\hat{\pi}_{g,NR} = N^{-1} \sum_{i=1}^{N} \mathbb{1}\{\hat{\gamma}_{i,NR} = g\}$ are estimates of the clusters sizes. Also define

$$\hat{\Omega}_{NPR} = \text{diag}\{\hat{\Omega}_{1,NPR}, \ldots, \hat{\Omega}_{G,NPR}\}$$

The null hypothesis of the test can be stated as $H_0 : \mu_g^* = \mu_{g'}^* \ \forall \ g$, where μ_g^* denotes the true mean of group g. It can be written as $H_0 : A_{d,G}\mu^* = 0$ by defining

$$\underset{d(G-1) \times dG}{A_{d,G}} = [(\iota_{G-1} \otimes I_d), -I_{d(G-1)}] \tag{2}$$

where ι_n is an $n \times 1$ vector of ones and I_n is the $n \times n$ identity matrix.

Lastly, define the test statistic

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

Then, Theorem 1 of Patton and Weller (2022) state that

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

They show that the test is correctly sized and has power in a range of relevant

2 Testing for clustering under switching

Once we introduce switching in the DGP, however, the power of the PW test decreases steadily. Here we define the switching as a probability that the mean m_{it} is different from the next-period mean $m_{i,t+1}$, for unit i. In short, the switching rate p is

$$p = \mathbb{P}(m_{i,t+1} \neq m_{it}). \tag{3}$$

The precise definition of switching used does not matter for our main result in Theorem 1.

The power loss is intuitive. If we ignore the possibility of switching, two well-separated clusters will blur together as we average the positions over time of units that have been part of both of them. To see this more clearly let us focus on a single unit first. As time goes by it will be more likely to have switched clusters at some point, but the estimator of P&W would still classify it to a unique cluster. When averaging its observations over time in (1), we would be averaging over draws from different clusters.

To produce a test with power under the alternative hypothesis of cluster switching we propose to modify the original test slightly in order to accommodate switching in the estimation of the clusters' parameters and membership. We do this by clustering at the cross-sectional level instead of at the panel level, allowing for a unit to be freely assigned to different clusters over time.

We start by modifying the assumptions of P&W by adding time-variation in the cluster assignments and on the cluster sizes.

Assumption 1. (a) The data comes from $Y_{it} = m_{it} + \varepsilon_{it}$ where $\varepsilon_{it} \sim$ 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, $m_i \in \mathcal{M} \subset \mathbb{R}^d$, Σ_i is positive definite, and $\mathbb{E}\left[||\varepsilon_{it}||^4\right] \leq \bar{\kappa} < \infty \forall i$, (b) $\varepsilon_{it} \perp\!\!\!\perp \varepsilon_{js} \forall\, i \neq j$ and $\forall s$, (c) $N, P, R \to \infty$.

Assumption 2. $m_{it} = \mu^* \;\forall\; i, t$

Assumption 2'. For known $G \geq 2$,

(a) $m_{it} \in \{\mu_1^*, \ldots, \mu_G^*\} \;\forall\; i, t$,

(b) $||\mu_g^* - \mu_{g'}^*|| > c > 0 \;\forall\; g \neq g'$, and

(c) $\lim_{N\to\infty} N_{g,t}/NT \equiv \pi_g \geq \underline{\pi} > 0$ for $g = 1, \ldots, G$, $t = 1, \ldots, T$,

where $N_{g,t} \equiv \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1}\{\gamma_{i,t}^* = g\}$, and $\gamma_{i,t}^* \in \{1, \ldots, G\}$ indicates to which cluster unit i belongs at time t.

Assumption 2 represents the null hypothesis of a single cluster and 2' represents the alternative. Note that we assume that the limiting proportions π_g and the true cluster means μ_g^* do not vary over time.

Our clustering procedure on sample $\mathcal{R}$ now accounts for switches:

$$(\hat{\mu}_{NR}, \hat{\gamma}_{NR}) = \arg\min_{\mu,\gamma} \frac{1}{NR} \sum_{i=1}^{N} \sum_{t\in\mathcal{R}} \sum_{g=1}^{G} ||Y_{it} - \mu_g||^2 \mathbb{1}\{\gamma_{it} = g\} \tag{4}$$

where $\hat{\gamma}_{NR}$ is the vector of stacked assignments over N and $\mathcal{R}$ of size NR. This is equivalent to a k-means procedure where each observation in the panel is treated as an independent unit in a cross-section of size NR. The subsequent estimators of the cluster mean and variances are also modified similarly:

$$\tilde{\mu}_{NP}(\hat{\gamma}_{NR}) = \arg\min_{\mu} \frac{1}{NP} \sum_{i=1}^{N} \sum_{t\in\mathcal{P}} \sum_{g=1}^{G} ||Y_{it} - \mu_g||^2 \mathbb{1}\{\hat{\gamma}_{it,NR} = g\} \tag{5}$$

and

$$\hat{\Omega}_{g,NPR} = \frac{1}{NP} \sum_{t\in\mathcal{P}} \sum_{i=1}^{N} (Y_{it} - \bar{Y}_{it})(Y_{it} - \bar{Y}_{it})' \hat{\pi}_{g,NR}^{-2} \mathbb{1}\{\hat{\gamma}_{it,NR} = g\}$$

where $\hat{\pi}_{g,NR} = (NR)^{-1} \sum_{i=1}^{N} \sum_{t\in\mathcal{R}} \mathbb{1}\{\hat{\gamma}_{it,NR} = g\}$ are estimates of the clusters sizes. As before, define

$$\hat{\Omega}_{NPR} = \text{diag}\{\hat{\Omega}_{1,NPR}, \ldots, \hat{\Omega}_{G,NPR}\}.$$

Although P&W put no restrictions on how to split the sample into $\mathcal{R}$ and $\mathcal{P}$, we cannot choose freely. The possible switching imposes time-dependence that affects the quality of the estimates of the means depending on how we split the sample. We suggest an interspersed scheme where odd time indices go into one sample, and even indices into the other. In this way the estimated cluster assignments at some $t \in \mathcal{R}$ are used to predict assignments at $t + 1$, which is itself in $\mathcal{P}$. We can write $\hat{\gamma}_{it,NR} = \hat{\gamma}_{i,t+1,NR}$ for all $t \in \mathcal{R}$ to be consistent with (5).

The null hypothesis can be represented as before, with the same matrix $A_{d,G}$ in (2). We can now state our version of the theorem.

Theorem 1. Let $\hat{\gamma}_{NR}$ be the estimated time-varying group assignments based on sample $\mathcal{R}$. Let $\tilde{\mu}_{NP}(\hat{\gamma}_{NR})$ be the estimated group means from sample $\mathcal{P}$ using group assignments $\hat{\gamma}_{NR}$. Define the test statistic for the differences in the estimated

means as

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

$$\text{where } \underset{(dG\times dG)}{\hat{\Omega}_{NPR}} = \text{diag}\left\{\hat{\Omega}_{1,NPR},\ldots,\hat{\Omega}_{G,NPR}\right\}$$

$$\text{and } \underset{(d\times d)}{\hat{\Omega}_{g,NPR}} = \frac{1}{NP}\sum_{t\in\mathcal{P}}\sum_{i=1}^{N}\left(Y_{it}-\bar{Y}_{it}\right)\left(Y_{it}-\bar{Y}_{it}\right)'\hat{\pi}_{g,NR}^{-2}\mathbb{1}\{\hat{\gamma}_{it,NR}=g\}$$

$$\text{and } \hat{\pi}_{g,NR} \equiv \frac{1}{NR}\sum_{i=1}^{N}\sum_{t\in\mathcal{R}}\mathbb{1}\{\hat{\gamma}_{it,NR}=g\}$$

(a) Under Assumptions 1 and 2,

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

(b) Under Assumptions 1 and 2',

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

The proof of this theorem follows closely that in P&W and is presented in the Appendix, with the differences being concentrated in the case under the alternative hypothesis. As part of it we also show that the k-means procedure in (4) necessarily produces distinct means when $G=2$ in Lemma 1.

3 Simulation study

To illustrate the power gain of accounting for cluster switching we present a simulation study in Figure 1. We generate data from either one of two normally-distributed clusters with means (m,m) and $(-m,-m)$, and identity covariance matrix, and several values of N and T. m varies from zero (the null hypothesis) to 2. We introduce cluster switching as in (3), with a probability of switching varying from zero to 20%. The rejection rates from the P&W test are plotted in blue, and from our test in red, considering a 5% confidence level.

Both tests display nominal rejection rates under the null hypothesis. If there is no switching and m takes small positive values, the P&W test diligently rejects the null as expected. In contrast, our test needs a greater separation between

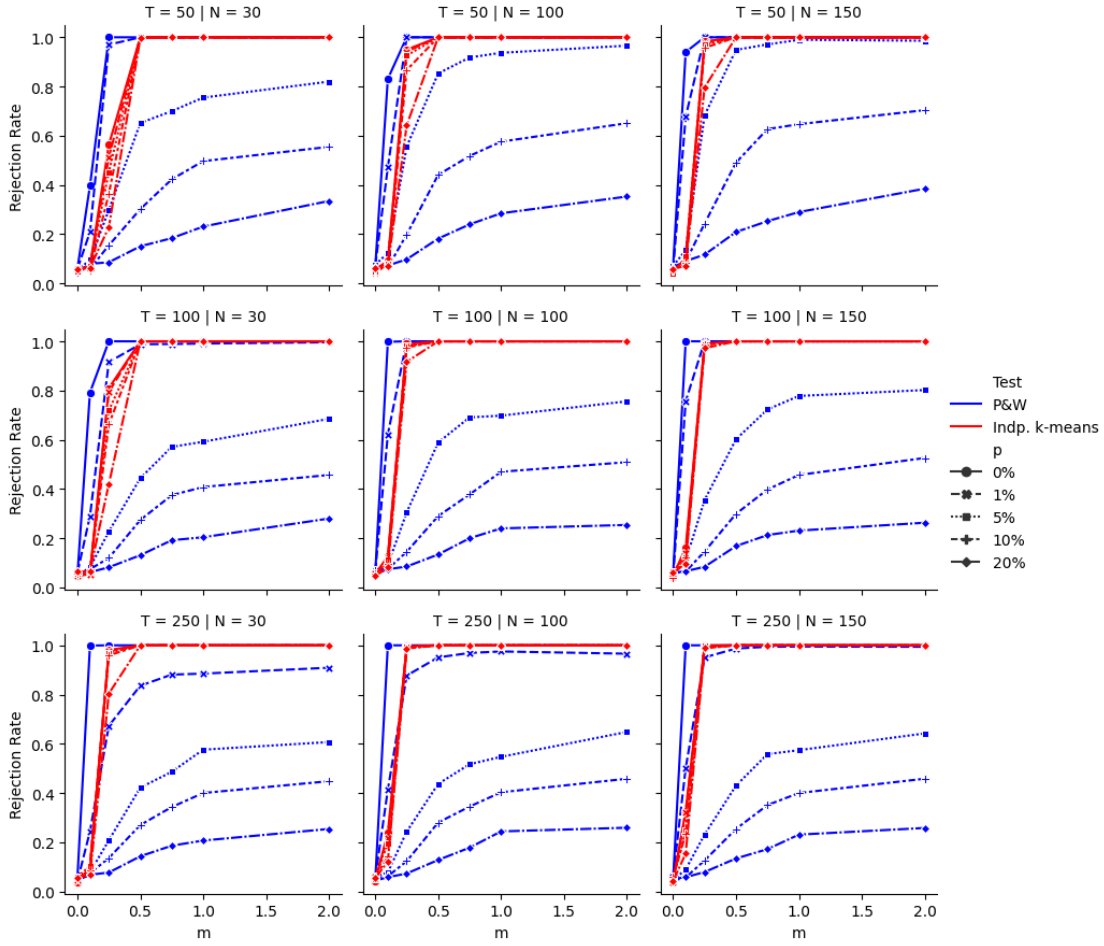

Figure 1: Simulation of rejection rates of the Patton and Weller (2022) test (blue) and our modified version (red) under different values of the switching rate. Points are generated from either one of two clusters centered at (m, m) and $(-m, -m)$, both with identity covariance matrix, and can change clusters with probability p. The plot shows that the original test is severely underpowered under switching rates of at least 5% in large samples. It is also noteworthy that a large T does not correct the problem, as there are more switches in consequence.

the clusters to achieve a high power. Still under no switching, a larger N and T uniformly increase power.

Once switching is introduced, power drops in all cases. This is drastically visible in the P&W test, where even at large values of m we get low power in most

cases. A large N still uniformly increases power, but T does not anymore. In fact, a large T hinders the P&W test as it gets us closer to the limiting Markov chain where the cluster assignments are completely mixed, meaning that one half of the sample cannot predict the assignment of the other half reliably anymore.

The effect of a larger p in our test is limited. It causes a drop in power that is only noticeable in settings with smaller sample sizes and less separated clusters. In all settings, at $m \geq 0.5$ the power of our test is close to 1 regardless of p. Here there's no distinction between N and T. As we cluster each point (i, t) independently, as a single cross-section, an increase in either one leads to an increase in power.

4 A few insights on the power of the test

We focus on the case of two clusters in one dimension in the DGP to compare the estimated cluster means through the estimator in P&W and ours. We show that under switching both are biased, but the bias works in opposite ways that make the estimates look more separated when allowing for switching, and less so when not.

One can think of the loss of power in the P&W test as a blurring of the clusters induced by the misclassification that does not vanish asymptotically under switching. Instead of averaging over true members of the cluster, we have to average over some pairs (i, t) that are not in the correct cluster. This brings the mean of these two clusters together the more misclassification there is.

In the modified test the opposite can happen. The 'tails' of a cluster that spread closer to another cluster are classified in the latter. Symmetrically the tails of the latter cluster are classified in the former. This causes a concentration of probability mass closer to the mean compared to the true cluster's distribution, and moves the estimated mean away from the other clusters.

Taken together, these mean that the difference in estimated means is larger when using the modified test than the standard under switching, inflating the F statistic and making a rejection of the null more likely and increasing the power of the test.

Starting with the P&W test, recall the definition of the estimator of the mean $\hat{\mu}_{NR}$ and assume without loss of generality that the true mean of cluster 1 is zero:

$$(\hat{\mu}_{NR}, \hat{\gamma}_{NR}) = \arg\min_{\mu,\gamma} \frac{1}{NR} \sum_{i=1}^{N} \sum_{t\in\mathcal{R}} \sum_{g=1}^{G} ||Y_{it} - \mu_g||^2 \mathbb{1}\{\gamma_i = g\}$$

The solution given the assignments is also the average:

$$\hat{\mu}_{g,NR}(\hat{\gamma}_{NR}) = \frac{1}{\hat{N}_{g,R}} \sum_{i=1}^{N} \sum_{t\in\mathcal{R}} \mathbb{1}\{\hat{\gamma}_{i,NR} = g\} Y_{it}$$

And we can write the classifier as

$$\hat{\gamma}_{NR}(\hat{\mu}_{NR}) = \arg\min_{\gamma} \sum_{i=1}^{N} \sum_{t\in\mathcal{R}} \sum_{g=1}^{G} ||Y_{it} - \hat{\mu}_{g,NR}||^2 \mathbb{1}\{\gamma_i = g\}$$

$$\hat{\gamma}_{i,NR}(\hat{\mu}_{NR}) = \arg\min_{\gamma} \sum_{g=1}^{G} \sum_{t\in\mathcal{R}} ||Y_{it} - \hat{\mu}_{g,NR}||^2 \mathbb{1}\{\gamma_i = g\}$$

The usual k-means procedure alternates between estimating the means given the classes, and estimating the classes given the means, until convergence. Assume as the best case that we start the procedure at the true values of the means, $\hat{\mu}^0_{g,NR} = \mu_g$ where the superscript indicates the iteration. If the assignments then do not produce the same means $\hat{\mu}^1_{g,NR} = \mu_g$ the procedure cannot achieve consistency. We will look for the estimated means that produce such stable assignments. First, we write down the classifier $\hat{\gamma}(\hat{\mu})$ omitting the subscripts NR:

$$\hat{\gamma}^0_i(\hat{\mu}^0) = \arg\min_{\gamma} \sum_{g=1}^{2} \sum_{t\in\mathcal{R}} (Y_{it} - \hat{\mu}^0_g)^2 \mathbb{1}\{\gamma_i = g\}$$

$$\hat{\gamma}^0_i(\hat{\mu}^0) = \begin{cases} 1 \text{ if } R^{-1}\sum_{t\in\mathcal{R}} Y_{it} \leq \hat{\mu}^0_2/2 \\ 2 \text{ otherwise} \end{cases}$$

That is, unit i is assigned to cluster 1 if its average over time is closer to the center of cluster 1. The next-iteration estimated means will be

$$\hat{\mu}^1_g = \left(R \sum_{i=1}^{N} \mathbb{1}\{\hat{\gamma}^0_i = g\} \right)^{-1} \sum_{t\in\mathcal{R}} \sum_{i=1}^{N} Y_{it} \mathbb{1}\{\hat{\gamma}^0_i = g\}$$

At the limit of N this is

$$\lim_{N\to\infty} \hat{\mu}_g^1 = \mathbb{E}_i \left(\frac{1}{R} \sum_{t\in\mathcal{R}} x_t \middle| \frac{1}{R} \sum_{t\in\mathcal{R}} x_t \leq \frac{\hat{\mu}_2^0}{2} \right)$$

Next we can take the limit over R. Note that the Markov chain of true assignments converges to a proportion of half of the t in one cluster and the other half in the other cluster for $p > 0$.

$$\lim_{R\to\infty} \lim_{N\to\infty} \hat{\mu}_g^1 = \mathbb{E}_i \left(\mathbb{E}_t(x) \middle| \mathbb{E}_t(x) \leq \frac{\hat{\mu}_2^0}{2} \right) = \frac{\mu_2}{2}$$

That is, at the limit the estimator of the means converges to the midpoint of the true means, where the clusters are indistinguishable. Note that $\hat{\gamma}_i^0(\hat{\mu}^0)$ is still a solution to the classification step and so the k-means procedure can stop.

Next we turn to the modified statistic in Theorem 1. As we treat every observation (i, t) independently, we can use just one index i running from 1 to $M = NR$. Again we start at $\hat{\mu}_g^0 = \mu_g$. Then the estimator for the mean in the b-th step of the k-means procedure is:

$$\hat{\mu}_g^b(\hat{\gamma}) = \left(\sum_{i=1}^{M} \mathbb{1}\{\hat{\gamma}_i = g\} \right)^{-1} \sum_{i=1}^{M} \mathbb{1}\{\hat{\gamma}_i = g\} Y_i$$

and the initial classifier:

$$\hat{\gamma}_i^0(\hat{\mu}^0) = \arg\min_{\gamma} \sum_{g=1}^{2} (Y_i - \hat{\mu}_g^0)^2 \mathbb{1}\{\gamma = g\} = \begin{cases} 1 & \text{if } Y_i \leq \hat{\mu}_2^0/2 \\ 2 & \text{otherwise} \end{cases}$$

The next-iteration means will be

$$\hat{\mu}_g^1 = \left(\sum_{i=1}^{M} \mathbb{1}\{\hat{\gamma}_i^0 = g\} \right)^{-1} \sum_{i=1}^{M} Y_i \mathbb{1}\{\hat{\gamma}_i^0 = g\}$$

which, at the limit of M and for $g = 1$ initially, becomes

$$\lim_{M\to\infty} \hat{\mu}_1^1 = \mathbb{E}_f \left(x \middle| x \leq \frac{\hat{\mu}_2^0}{2} \right) = \frac{1}{\int_{x\in\mathbb{R}} f(x) \mathbb{1}\{x \leq \hat{\mu}_2^0/2\} \, \mathrm{d}x} \int_{x\in\mathbb{R}} x f(x) \mathbb{1}\{x \leq \hat{\mu}_2^0/2\} \, \mathrm{d}x$$

where f is the mixture distribution composed of the two clusters with equal weights (with distributions f_1 and f_2). We can decompose it in equal parts for each cluster.

$$\begin{aligned}\lim_{M\to\infty} \hat{\mu}_1^1 &= \frac{1}{2\int_{x\in\mathbb{R}} f(x)\mathbb{1}\{x \leq \hat{\mu}_2^0/2\}\,\mathrm{d}x} \int_{x\leq\hat{\mu}_2^0/2} x(f_1(x)+f_2(x))\,\mathrm{d}x \\ &= \frac{1}{2(1/2)} \int_{x\leq\hat{\mu}_2^0/2} x(f_1(x)+f_2(x))\,\mathrm{d}x\end{aligned}$$

The condition for this estimator to be lower, and therefore farther from cluster 2, is

$$\begin{aligned} &\lim_{M\to\infty} \hat{\mu}_1^1 < \mu_1 = \mathbb{E}_{f_1}(x) \\ \iff &\int_{x<\hat{\mu}_2^0/2} x(f_1(x)+f_2(x))\,\mathrm{d}x < \int_{x\in\mathbb{R}} x f_1(x)\,\mathrm{d}x \\ &\int_{x\leq\hat{\mu}_2^0/2} x(f_1(x)+f_2(x))\,\mathrm{d}x < \int_{x\leq\mu_2^0/2} x f_1(x)\,\mathrm{d}x + \int_{x>\mu_2^0/2} x f_1(x)\,\mathrm{d}x \\ &\int_{x\leq\hat{\mu}_2^0/2} x f_2(x)\,\mathrm{d}x < \int_{x>\mu_2^0/2} x f_1(x)\,\mathrm{d}x \end{aligned}$$

Note that $f_2(x+\mu_2) = f_1(x)$ as they are only different up to a shift of the mean.

$$\begin{aligned} \int_{x\leq\hat{\mu}_2^0/2} x f_2(x)\,\mathrm{d}x &< \int_{x>\mu_2^0/2} x f_2(\mu_2+x)\,\mathrm{d}x \\ \int_{z\geq\hat{\mu}_2^0/2} (\mu_2 - z) f_2(\mu_2+z)\,\mathrm{d}x &< \int_{x>\mu_2^0/2} x f_2(x+\mu_2)\,\mathrm{d}x \\ \int_{z\geq\hat{\mu}_2^0/2} (\mu_2 - 2z) f_2(\mu_2+z)\,\mathrm{d}x &< 0 \end{aligned}$$

The condition is satisfied as $f_2(x) > 0 \quad \forall \quad x$ and $(\mu_2 - 2z) < 0 \quad \forall \quad z > \hat{\mu}_2^0/2$. Symmetrically, it can be shown that $\lim_{M\to\infty} \hat{\mu}_2^1 > \mu_2$. Still due to the symmetry the assignments $\hat{\gamma}^1$ will be the same and the k-means procedure can stop; both means move in opposite directions by the same magnitude between iterations. The estimated means are further away from each other than the true ones, inflating the F statistic and inducing a higher power in the test.

5 Application

To showcase the effect of accounting for cluster switching in a concrete setting we revisit the application of Bonhomme and Manresa (2015). They build on Acemoglu et al. (2008) to estimate a model for democracy in a yearly panel of countries. They add time fixed effects which are group specific, and estimate the groups jointly with the estimation of the regression coefficients. Their model reads as

$$democracy_{it} = \theta_1 democracy_{i,t-1} + \theta_2 \log GDPpc_{i,t-1} + \alpha_{g_i,t} + \nu_{it} \tag{6}$$

where $\alpha_{g_i,t}$ are fixed effects specific of time t and group g_i of country i. Group assignments are fixed in time. *democracy* is the measure published by Freedom House and $\log GDPpc$ is the percent change of log GDP per capita.

Their estimation procedure alternates between estimating the group assignments and estimating the parameters θ and α until convergence by minimizing the sum of squared errors of the equation in (6). The number of groups is not estimated but chosen as the least amount of groups that produce parameter values similar to those produced using more groups. This technique is common in cluster analysis, see Tibshirani et al. (2001).

Bonhomme and Manresa (2015) find 4 clusters in their application. They are characterized by grouped residuals which are plotted in the right-most panel of Figure 6. The average level of the *democracy* variable of each group follows a similar pattern. They interpret these groups as one high-democracy group, one low-democracy group, and two transition groups: one early transition and one late transition.

Could the same transition pattern be described by two clusters, one with a high means and the other with a low mean, and a switching rate? Denote the partial residuals by

$$\hat{\nu}'_{it} = democracy_{it} - \hat{\theta}_1 democracy_{i,t-1} - \hat{\theta}_2 \log GDPpc_{i,t-1}.$$

We apply ours and P&W's test to $\hat{\nu}'_{it}$, considering the alternative hypothesis of two clusters. Can we detect clustering if we ignore the switching?

Our simulation study indicates that in settings with clear clustering both tests are able to reliably reject the null. If the clustering structure is clouded by means that are too close together, the P&W test seems to have a higher power, and in settings where switching is the problem our test seems to be a better choice. Bonhomme and Manresa (2015) use data sampled in 5-year intervals and the partial residuals in their setting are indeed clearly clustered (see the right-most panel of Figure 7) such that both tests reject the null hypothesis. We muddy the waters by starting from annual data instead. The annual *democracy* variable is largely driven by discrete jumps stemming from, for example, sudden regime change. This is reflected in the jumpy nature of the residuals (see, for example, the left-most panel of Figure 7). In this setting, can we still reject the null hypothesis?

We design two sets of experiments based on the methods of Bonhomme and Manresa (2015) and the data of Acemoglu et al. (2008). In the first one we use annual data on *democracy* and $logGDPpc$ from 1975 to 2000 and calculate moving averages using zero to 10 lags. For each value of the length of this window we estimate (6) using four groups. As the window expands, the partial residuals $\hat{\nu}'_{it}$ of the regression become smoother. In the second experiment we use annual data from 1970 to 2000, and sample the data at increasing frequencies from 1 to 5 years, such that at the 5-year frequency we coincide with the setting in Bonhomme and Manresa (2015). Again, as we increase the frequency, pooling the jumps in *democracy*, the residuals become smother. In both settings we take the partial residuals produced as given and compare the p-values produced by either test.

The p-values plotted in Figures 2 (first experiment) and 5 (second experiment) show that as we approach the smoother settings both tests are able to reject the null hypothesis, but the independent k-means version also does it in settings where the clustering structure is only apparent when allowing for cluster switching. See, for example, the middle panel of Figure 3. It shows the grouped fixed effects resulting from estimating (6) over annual data averaged in moving windows of 5 lags. Although one of the resulting groups has a stable mean around zero, the other three are volatile, sometimes diverging from zero, sometimes merging back. These patterns might not be visible when looking for two clusters with no switching as there is no stable cluster with a high mean. Once we allow for switching we can

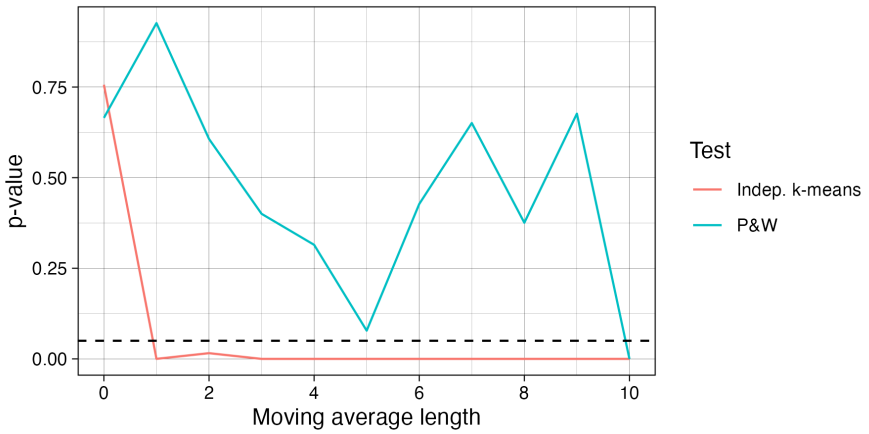

Figure 2: p-values in the first experiment according to the length of the moving average window. The dashed line marks the 5% level.

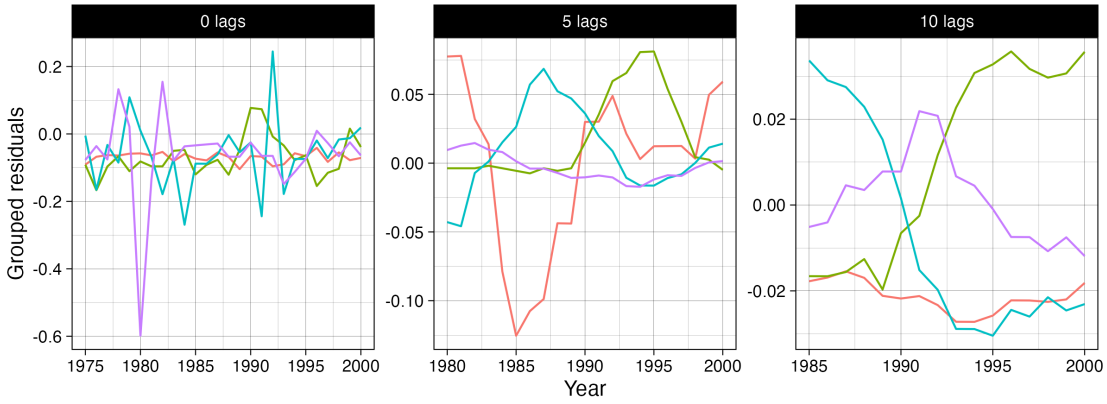

Figure 3: Grouped fixed effects in the first experiment for three values of the length of the moving average window. As we increase the window length, the groups become more clearly separated.

form a high-mean cluster with variable composition whose mean is significantly different from the low-mean cluster.

These results indicate that ignoring the possibility of cluster-switching might lead us to ignore a clustering structure that is in fact present in the data.

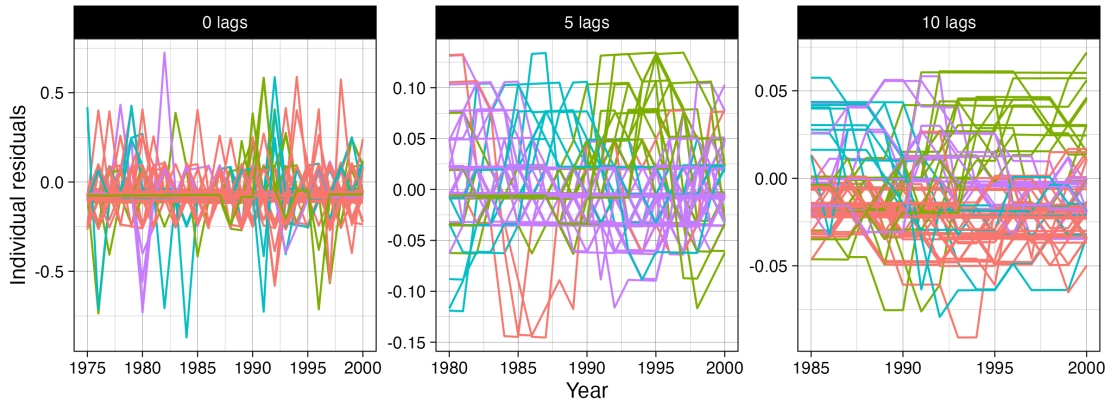

Figure 4: Individual partial residuals in the first experiment for three values of the length of the moving average window. As we increase the window length, the residuals become smoother and the clustering more apparent.

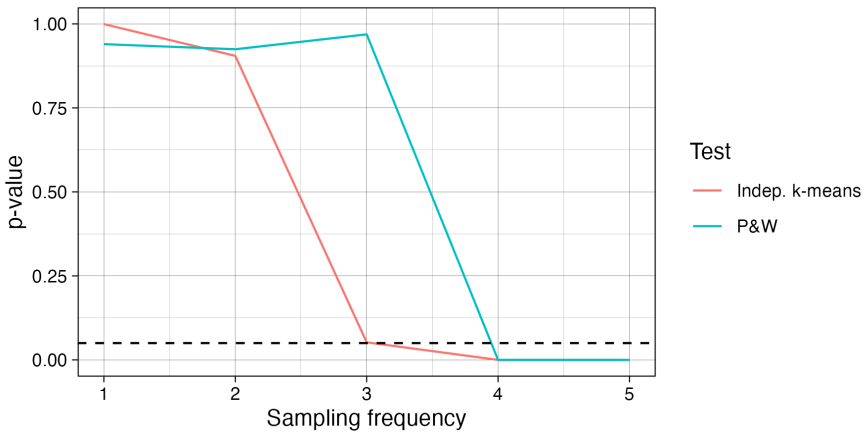

Figure 5: p-values in the second experiment according to the sampling frequency. The dashed line marks the 5% level. The right-most point coincides with the application in Bonhomme and Manresa (2015).

6 Conclusion

In this paper we have argued that a clustering structure might be masked by switching, and when testing for clustering as in Patton and Weller (2022) we might want to allow for this possibility. We have suggested a variation of their test that can handle variable cluster compositions and is still valid under the same

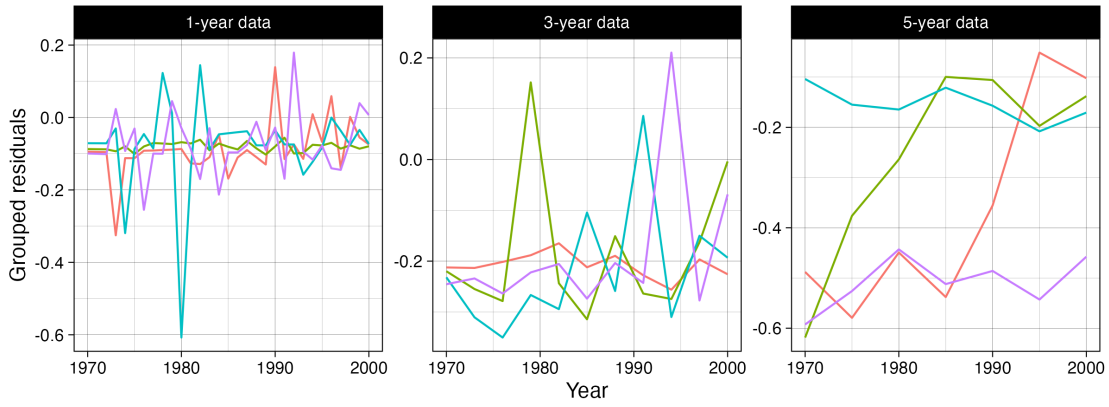

Figure 6: Grouped fixed effects in the second experiment for three values of the sampling frequency. As we increase the frequency, the groups become more clearly separated. The right-most panel coincides with the application in Bonhomme and Manresa (2015).

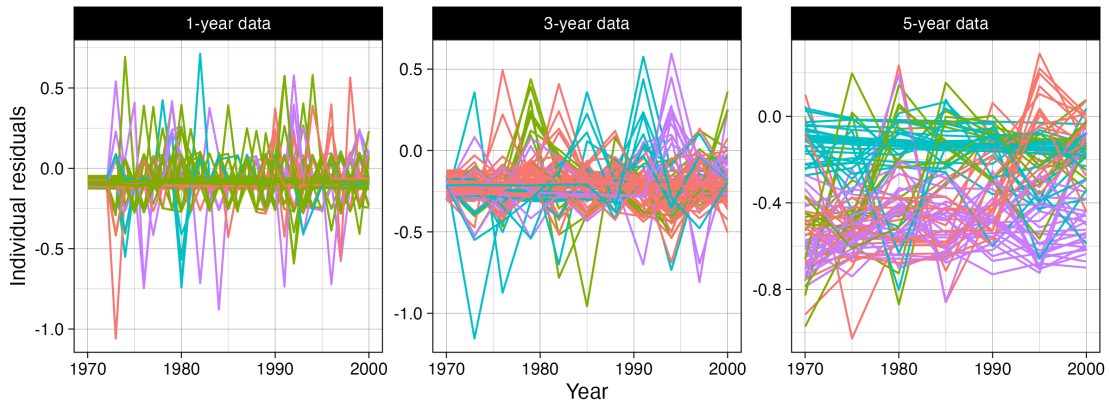

Figure 7: Individual partial residuals in the second experiment for three values of the sampling frequency. As we increase the sampling frequency, the residuals become smoother and the clustering more apparent. The right-most panel coincides with the application in Bonhomme and Manresa (2015).

null hypothesis. This new test has power under the alternative hypothesis that have different means, but units switch between them, which is indistinguishable from a single cluster for the test in P&W. We have shown, in a simplified setting, that the power of the test comes from the bias in the estimation of the cluster means, while the P&W test presents a bias that hinders its power. Revisiting the application of Bonhomme and Manresa (2015) we have shown that our test can

detect a clustering structure in settings where the clustering is not clear and there is potentially strong switching.

References

Acemoglu, D., S. Johnson, J. A. Robinson, and P. Yared (2008). Income and democracy. *American Economic Review 98*(3), 808–42.

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*(3), 1147–1184.

Catania, L. (2021). Dynamic Adaptive Mixture Models with an Application to Volatility and Risk. *Journal of Financial Econometrics 19*(4), 531–564.

Cho, J. S. and H. White (2007). Testing for regime switching. *Econometrica 75*(6), 1671–1720.

Custodio João, I., A. Lucas, J. Schaumburg, and B. Schwaab (2022). Dynamic clustering of multivariate panel data. *Journal of Econometrics*.

Custodio João, I., J. Schaumburg, A. Lucas, and B. Schwaab (2022). Dynamic Nonparametric Clustering of Multivariate Panel Data. *Journal of Financial Econometrics*.

Lumsdaine, R. L., R. Okui, and W. Wang (2023). Estimation of panel group structure models with structural breaks in group memberships and coefficients. *Journal of Econometrics 233*(1), 45–65.

Munro, E. and S. Ng (2022). Latent dirichlet analysis of categorical survey responses. *Journal of Business & Economic Statistics 40*(1), 256–271.

Patton, A. J. and B. M. Weller (2022). Testing for unobserved heterogeneity via k-means clustering. *Journal of Business & Economic Statistics 41*(3), 737–751.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics 9*(1), 135–140.

Pollard, D. (1982). A Central Limit Theorem for k-Means Clustering. *The Annals of Probability 10*(4), 919 – 926.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 411–423.

7 Appendix

Proof of Theorem 1. (a) We first find the limiting distribution of $\sqrt{NP}\tilde{\mu}_{NP}(\hat{\gamma}_{NR})$ conditional on the information set $\mathcal{F}_R = \sigma\left(\{Y_{it}\}_{i=1}^N, t \in \mathcal{R}\right)$.

Denote $\hat{N}_{g,R} \equiv \sum_{i=1}^N \sum_{t\in\mathcal{R}} \mathbb{1}\{\hat{\gamma}_{it,NR} = g\}$ the total (estimated) size of group g on sample $\mathcal{R}$ and $\hat{\pi}_{g,R} = \hat{N}_{g,R}/NR$ the estimated relative size. Note that the minimization in (5) results in a group average, so[1]

$$\begin{aligned}\tilde{\mu}_{g,NP} &= \frac{R}{P\hat{N}_{g,R}} \sum_{i=1}^N \sum_{t\in\mathcal{P}} \mathbb{1}\{\hat{\gamma}_{it,NR} = g\} Y_{it} \\ &= \frac{1}{NP} \sum_{i=1}^N \sum_{t\in\mathcal{P}} Y_{it} \hat{\pi}_{g,R}^{-1} \mathbb{1}\{\hat{\gamma}_{it,NR} = g\}\end{aligned}$$

for $g = 1, \ldots, G$. Thus,[2]

$$\begin{aligned}\sqrt{NP}\left(\tilde{\mu}_{g,NP}(\hat{\gamma}_{NR}) - \mu_g^*\right) &= \frac{1}{\sqrt{NP}} \sum_{i=1}^N \sum_{t\in\mathcal{P}} (\mu_g^* + \varepsilon_{it}) \hat{\pi}_{g,R}^{-1} \mathbb{1}\{\hat{\gamma}_{it,NR} = g\} - \sqrt{NP}\mu_g^* \\ &= \frac{1}{\sqrt{NP}} \sum_{i=1}^N \sum_{t\in\mathcal{P}} \hat{U}_{itg,NR} \varepsilon_{it} \\ &\quad + \frac{\mu_g^* NR}{\hat{N}_{g,R}\sqrt{NP}} \sum_{i=1}^N \sum_{t\in\mathcal{P}} \mathbb{1}\{\hat{\gamma}_{it,NR} = g\} - \sqrt{NP}\mu_g^* \\ &= \frac{1}{\sqrt{NP}} \sum_{i=1}^N \sum_{t\in\mathcal{P}} \hat{U}_{itg,NR} \varepsilon_{it}\end{aligned}$$

with $\hat{U}_{itg,NR} \equiv \hat{\pi}_{gR}^{-1} \mathbb{1}\{\hat{\gamma}_{it,NR} = g\}$. Note that it is bounded due to the assumption of a minimum cluster size in Assumption 2 and 2'(c). Define

$$\bar{\Omega}_{g,NR} \equiv \text{Var}\left[\frac{1}{\sqrt{NP}} \sum_{i=1}^N \sum_{t\in\mathcal{P}} \hat{U}_{itg,NR} \varepsilon_{it} \mid \mathcal{F}_R\right] = \frac{1}{NP} \sum_{i=1}^N \sum_{t\in\mathcal{P}} \hat{U}_{itg,NR}^2 \Sigma_i$$

where $\Sigma_i = \text{Var}[\varepsilon_{it}]$ and the second line holds as ε_{it} is uncorrelated in the time series and cross-section. We then obtain the asymptotic distribution of $\tilde{\mu}_{g,NP}(\hat{\gamma}_{NR})$:

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

[1] We should assume additional that $\hat{N}_{g,R} = \hat{N}_{g,P}$ due to our sampling scheme and how we translate assignments from $\mathcal{R}$ to $\mathcal{P}$. Then the notation could be simplified.

[2] Again we should assume that the samples are symmetric and predict $t+1$.

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

Next we show that $\text{Cov}[\tilde{\mu}_{g,NP}(\hat{\gamma}_{NR}), \tilde{\mu}_{g',NP}(\hat{\gamma}_{NR})] = 0$ for $g \neq g'$. Consider elements k and k' in groups g and g' of the vector $(\tilde{\mu}_{g,NP}(\hat{\gamma}_{NR}) - \mu^*_g)$. Their covariance is:

$$\begin{aligned}
&\mathbb{E}\left[(\tilde{\mu}_{gk,NP}(\hat{\gamma}_{NR}) - \mu^*_{gk})(\tilde{\mu}_{g'k',NP}(\hat{\gamma}_{NR}) - \mu^*_{g'k'}) \mid \mathcal{F}_R\right] \\
&= \frac{1}{N^2P^2}\mathbb{E}\left[\left(\sum_{i=1}^{N}\sum_{t\in\mathcal{P}} \pi^{-1}_{g,R}\mathbb{1}\{\hat{\gamma}_{it,NR} = g\}\varepsilon_{itk}\right)\left(\sum_{j=1}^{N}\sum_{s\in\mathcal{P}} \pi^{-1}_{g',R}\mathbb{1}\{\hat{\gamma}_{jt,NR} = g'\}\varepsilon_{jsk'}\right)\middle|\mathcal{F}_R\right] \\
&= 0
\end{aligned}$$

As $\mathbb{1}\{\hat{\gamma}_{it,NR} = g\}\mathbb{1}\{\hat{\gamma}_{it,NR} = g'\} = 0$ and $\mathbb{E}(\varepsilon_{itk}\varepsilon_{jsk'}) = 0\ \forall (i,t) \neq (j,s)$. Thus we obtain the limiting distribution for the entire vector $\tilde{\mu}_{NP}(\hat{\gamma}_{NR})$:

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

where $\bar{\Omega}^{-1/2}_{NR}$ is block-diagonal with $\left(\bar{\Omega}^{-1/2}_{1,NR}, \ldots, \bar{\Omega}^{-1/2}_{G,NR}\right)$ along the diagonal. Consider the following estimator of $\bar{\Omega}_{g,NR}$:

$$\begin{aligned}
\hat{\Omega}_{g,NPR} &= \frac{1}{NP}\sum_{t\in\mathcal{P}}\sum_{i=1}^{N} \hat{U}^2_{igt,NR}(Y_{it} - \bar{Y}_i)(Y_{it} - \bar{Y}_i)' \\
&= \frac{1}{NP}\sum_{t\in\mathcal{P}}\sum_{i=1}^{N} \hat{U}^2_{igt,NR}\hat{\varepsilon}_{it}\hat{\varepsilon}'_{it}
\end{aligned}$$

which can be shown to be consistent for all g. So $\hat{\Omega}_{NPR} - \bar{\Omega}_{NR} \xrightarrow{p} 0$ and

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

Under the null hypothesis of one cluster we have $A_{dG}\mu^* = \mathbf{0}_{d(G-1)}$, and finally:

$$F_{NPR} = NP\tilde{\mu}'_{NP}(\hat{\gamma}_{NR})A'_{d,G}\left(A_{d,G}\hat{\Omega}_{NPR}A'_{d,G}\right)^{-1}A_{d,G}\mu_{NP}(\hat{\gamma}_{NR}) \xrightarrow{d} \chi^2_{d(G-1)}$$

(b) Note that $\tilde{\mu}_{NP}(\hat{\gamma}_{NR}) - \mu^* = (\hat{\mu}_{NR} - \mu^*) + (\tilde{\mu}_{NP}(\hat{\gamma}_{NR}) - \hat{\mu}_{NR})$. The second term

is

$$
\begin{aligned}
&\tilde{\mu}_{g,NP}(\hat{\gamma}_{NR}) - \hat{\mu}_{g,NR} \\
&= \frac{1}{NP}\sum_{i=1}^{N}\sum_{t\in\mathcal{P}} Y_{it}\hat{\pi}_{g,R}^{-1}\mathbb{1}\{\hat{\gamma}_{it,NR}=g\} \\
&\quad - \frac{1}{NR}\sum_{i=1}^{N}\sum_{t\in\mathcal{R}} Y_{it}\hat{\pi}_{g,R}^{-1}\mathbb{1}\{\hat{\gamma}_{it,NR}=g\} \\
&= \frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_{g,R}^{-1}\left(\frac{1}{P}\sum_{t\in\mathcal{P}}(m_{it}+\varepsilon_{it})\mathbb{1}\{\hat{\gamma}_{it,NR}=g\} - \frac{1}{R}\sum_{t\in\mathcal{R}}(m_{it}+\varepsilon_{it})\mathbb{1}\{\hat{\gamma}_{it,NR}=g\}\right) \\
&= o_p(1) + \frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_{g,R}^{-1}\left(\frac{1}{P}\sum_{t\in\mathcal{P}} m_{it}\mathbb{1}\{\hat{\gamma}_{it,NR}=g\} - \frac{1}{R}\sum_{t\in\mathcal{R}} m_{it}\mathbb{1}\{\hat{\gamma}_{it,NR}=g\}\right) \\
&= o_p(1) + \pi_{g,R}^{-1}\mu_g^*\left(\frac{\hat{N}_{g,P}}{NP} - \frac{\hat{N}_{g,R}}{NR}\right) \to 0, \text{ as } N,P,R\to\infty
\end{aligned}
$$

where the last line is due to Assumption 2'. This holds for $g = 1,\ldots,G$. Pollard (1981) shows that k-means estimator of the means $\hat{\mu}_{NR}$ converges to the minimum of the objective function $\mu^\flat$. For $G \geq 3$ it is assumed that this is a set of k distinct points, such that $\mu^{\flat\prime}A'_{d,G} > 0$. For $G = 2$, Lemma 1 guarantees that the solution will consist of distinct points. Thus $\tilde{\mu}_{NP}(\gamma_{NR}) \xrightarrow{p} \mu^\flat$ as $N,P,R\to\infty$. This implies that

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

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

Lemma 1. The solution to the clustering problem in (4) converges in probability to two distinct means when $G = 2$.

Proof of Lemma 1. State the problem as

$$
(\hat{\mu},\hat{\gamma}) = \arg\min_{\mu,\gamma} W(\mu) = \arg\min_{\mu,\gamma}\frac{1}{M}\sum_{i=1}^{M}\sum_{g=1}^{G}||Y_i-\mu_g||^2\mathbb{1}\{\gamma_i=g\}
$$

The solution given the assignments is the average:

$$\hat{\mu}_g(\hat{\gamma}) = \frac{1}{\sum_{i=1}^{M} \mathbb{1}\{\hat{\gamma}_i = g\}} \sum_{i=1}^{M} \mathbb{1}\{\hat{\gamma}_i = g\} Y_{it}$$

And the classification rule given the estimated means:

$$\hat{\gamma}(\hat{\mu}) = \arg\min_{\gamma} \sum_{i=1}^{M} \sum_{g=1}^{G} ||Y_i - \hat{\mu}_g||^2 \mathbb{1}\{\gamma_i = g\}$$

Pollard (1981) shows that the solution converges to the minimizer of the within-cluster distance W. We show that the minimum $W(\mu)$ when μ contains two coinciding centers occurs when these centers lie in a convex combination $\mu^\sharp$ of the true centers. Then we show that $\mu^\sharp$ is not a minimizer of W. Let the centers be indexed by j and k. The average within cluster distance is:

$$\begin{aligned}
W(\hat{\mu}) &= \frac{1}{M} \sum_{i=1}^{M} \sum_{g=j,k} ||Y_i - \hat{\mu}_g||^2 \mathbb{1}\{\hat{\gamma}_i = g\} \\
\operatorname{plim} W(\hat{\mu}) &= \mathbb{E}\left(||Y_i - \hat{\mu}_{\hat{\gamma}_i}||^2\right) \\
&= \mathbb{E}\left(||\mu_{\gamma_i^*} - \hat{\mu}_{\hat{\gamma}_i}||^2 + 2\varepsilon_i'(\mu_{\gamma_i^*} - \hat{\mu}_{\hat{\gamma}_i}) + \varepsilon_i'\varepsilon_i\right) \\
&= \mathbb{E}\left(||\mu_{\gamma_i^*} - \hat{\mu}_{\hat{\gamma}_i}||^2\right) + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k} \\
&= \mathbb{E}\left(||\mu_{\gamma_i^*} - \hat{\mu}_{\gamma_i^*}||^2\right) \mathbb{P}\left(\hat{\gamma}_i = \gamma_i^*\right) + \mathbb{E}\left(||\mu_{\gamma_i^*} - \hat{\mu}_{\gamma_i'}||^2\right) \mathbb{P}\left(\hat{\gamma}_i \neq \gamma_i^*\right) \\
&\quad + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k}
\end{aligned}$$

Now assume $\hat{\mu}_j = \hat{\mu}_k = \hat{\mu}^\sharp$.

$$\begin{aligned}
\operatorname{plim} W(\hat{\mu}^\sharp) &= \mathbb{E}\left(||\mu_{\gamma_i^*} - \hat{\mu}^\sharp||^2\right) \mathbb{P}\left(\hat{\gamma}_i = \gamma_i^*\right) + \mathbb{E}\left(||\mu_{\gamma_i^*} - \hat{\mu}^\sharp||^2\right) \mathbb{P}\left(\hat{\gamma}_i \neq \gamma_i^*\right) \\
&\quad + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k} \\
&= \left(||\mu_j - \hat{\mu}^\sharp||^2\right) \mathbb{P}\left(\hat{\gamma}_i = \gamma_i^* | \gamma_i^* = j, \mu_j, \hat{\mu}^\sharp\right) \pi_j \\
&\quad + \left(||\mu_k - \hat{\mu}^\sharp||^2\right) \mathbb{P}\left(\hat{\gamma}_i = \gamma_i^* | \gamma_i^* = k, \mu_k, \hat{\mu}^\sharp\right) \pi_k \\
&\quad + \left(||\mu_j - \hat{\mu}^\sharp||^2\right) \mathbb{P}\left(\hat{\gamma}_i \neq \gamma_i^* | \gamma_i^* = j, \mu_j, \hat{\mu}^\sharp\right) \pi_j \\
&\quad + \left(||\mu_k - \hat{\mu}^\sharp||^2\right) \mathbb{P}\left(\hat{\gamma}_i \neq \gamma_i^* | \gamma_i^* = k, \mu_k, \hat{\mu}^\sharp\right) \pi_k \\
&\quad + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k}
\end{aligned}$$

When $\hat{\mu}_j = \hat{\mu}_k = \hat{\mu}^\sharp$ the assignment criterion is not defined. Let units are assigned with probability $\pi_g^\sharp$ to cluster g.

$$\begin{aligned}\operatorname{plim} W(\hat{\mu}^\sharp) = &\left(||\mu_j - \hat{\mu}^\sharp||^2\right)\pi_j^\sharp\pi_j + \left(||\mu_k - \hat{\mu}^\sharp||^2\right)\pi_k^\sharp\pi_k \\ &+ \left(||\mu_j - \hat{\mu}^\sharp||^2\right)\pi_k^\sharp\pi_j + \left(||\mu_k - \hat{\mu}^\sharp||^2\right)\pi_j^\sharp\pi_k \\ &+ \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k}\end{aligned}$$

Then the $\hat{\mu}^\sharp$ that minimizes the variation W is the weighted midpoint $\hat{\mu}^\sharp = \pi_j\mu_j + \pi_k\mu_k$. This rules out other exotic allocations of $\hat{\mu}^\sharp$. The variation in this case is:

$$\begin{aligned}\operatorname{plim} W(\hat{\mu}^\sharp) &= ||(1-\pi_j)\mu_j - \pi_k\mu_k||^2\pi_j + ||(1-\pi_k)\mu_k - \pi_j\mu_j||^2\pi_k \\ &\quad + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k} \\ &= 2\pi_k\pi_j||\mu_j - \mu_k||^2 + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k}\end{aligned}$$

Consider now the estimates $\hat{\mu}_j = \mu^\sharp + \Xi_j = \pi_j\mu_j + \pi_k\mu_k + \Xi_j$ and $\hat{\mu}_k = \mu^\sharp - \Xi_k = \pi_j\mu_j + \pi_k\mu_k - \Xi_k$.

$$\begin{aligned}\operatorname{plim} W(\hat{\mu}) &= \mathbb{E}\left(||\mu_{\gamma_i^*} - \hat{\mu}_{\gamma_i^*}||^2\right)\mathbb{P}\left(\hat{\gamma}_i = \gamma_i^*\right) + \mathbb{E}\left(||\mu_{\gamma_i^*} - \hat{\mu}_{\gamma_i'}||^2\right)\mathbb{P}\left(\hat{\gamma}_i \neq \gamma_i^*\right) \\ &\quad + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k} \\ &= ||\mu_j - \hat{\mu}_j||^2(1 - mis_j)\pi_j + ||\mu_k - \hat{\mu}_k||^2(1 - mis_k)\pi_k \\ &\quad + ||\mu_j - \hat{\mu}_k||^2 mis_j\pi_j + ||\mu_k - \hat{\mu}_j||^2 mis_k\pi_k \\ &\quad + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k} \\ &= ||\pi_k(\mu_j - \mu_k) - \Xi_j||^2(1 - mis_j)\pi_j \\ &\quad + ||\pi_j(\mu_k - \mu_j) + \Xi_k||^2(1 - mis_k)\pi_k \\ &\quad + ||\pi_k(\mu_j - \mu_k) + \Xi_k||^2 mis_j\pi_j \\ &\quad + ||\pi_j(\mu_k - \mu_j) - \Xi_j||^2 mis_k\pi_k \\ &\quad + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k}\end{aligned}$$

Let $\Xi_j = \xi_j(\mu_j - \mu_k)$ and $\Xi_k = \xi_k(\mu_j - \mu_k)$

$$
\begin{aligned}
\operatorname{plim} W(\hat{\mu}) = & ||\pi_k(\mu_j - \mu_k) - \xi_j(\mu_j - \mu_k)||^2(1 - mis_j)\pi_j \\
& + ||\pi_j(\mu_k - \mu_j) - \xi_k(\mu_k - \mu_j)||^2(1 - mis_k)\pi_k \\
& + ||\pi_k(\mu_j - \mu_k) - \xi_k(\mu_k - \mu_j)||^2 mis_j\pi_j \\
& + ||\pi_j(\mu_k - \mu_j) - \xi_j(\mu_j - \mu_k)||^2 mis_k\pi_k \\
& + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k} \\
= & ||\mu_j - \mu_k||^2(1 - mis_j)\pi_j(\pi_k - \xi_j)^2 \\
& + ||\mu_k - \mu_j||^2(1 - mis_k)\pi_k(\pi_j - \xi_k)^2 \\
& + ||\mu_j - \mu_k||^2 mis_j\pi_j(\pi_k + \xi_k)^2 \\
& + ||\mu_k - \mu_j||^2 mis_k\pi_k(\pi_j + \xi_j)^2 \\
& + \frac{\pi_j\Sigma_j + \pi_k\Sigma_k}{\pi_j + \pi_k}
\end{aligned}
$$

Where mis_j is the misclassification probability of units in cluster j.

At $\xi_k = \xi_j = 0$, let us again denote by $\pi_j^\sharp$ the probability of assigning a unit to cluster j. Then the misclassification probability is $mis_j^\sharp = (1 - \pi_j^\sharp)\pi_j$, and $mis_k^\sharp = \pi_j^\sharp(1 - \pi_j)$. Finally, we take the derivative of the probability limit of the variation and evaluate it at the threshold $\xi_k = \xi_j = 0$.

$$
\begin{aligned}
\left.\frac{\partial \operatorname{plim} W(\hat{\mu})}{\partial \xi_j}\right|_{\xi_j=0} = & ||\mu_j - \mu_k||^2\big(-2(1 - mis_j)\pi_j\pi_k \\
& - 2(1 - mis_k)\pi_k\pi_j + 2mis_j\pi_j\pi_k + 2mis_j\pi_j\pi_k\big) < 0 \\
& \iff mis_j + mis_k < 1 \iff (1 - \pi_j^\sharp)\pi_j + \pi_j^\sharp(1 - \pi_j) < 1
\end{aligned}
$$

Which is true for any combination of $\pi_j < 1$ and $\pi_j^\sharp < 1$, and similarly for the derivative with respect to ξ_k. In conclusion, $\operatorname{plim} W(\mu^\sharp)$ is not minimized at any vector of means where the means of two clusters are the same. □