Testing for Clustering Under Switching

Igor Custodio João[a]

[a]Vrije Universiteit Amsterdam and Tinbergen Institute

9 October 2023

Duke Financial Econometrics Lunch Group

The setting

Consider a panel of N units, observed over T periods across d dimensions with individual means:

$$Y_{it} = m_i + \varepsilon_{it} \ , \ \varepsilon_{it} \sim F(0, \Sigma_i).$$

Now assume that each individual belong to one of G groups with group-specific means:

$$m_i \in \{\mu_1^*, \dots, \mu_G^*\}.$$

We can use k-means clustering to recover the means and group structure.

Patton and Weller (2022), P&W hereafter, develop a test for clustering with $H_0 : G = 1$

The setting

Now we allow for cluster switching. Add a subscript t on m_{it}

$$Y_{it} = m_{it} + \varepsilon_{it} \ , \ \varepsilon_{it} \sim F(0, \Sigma_i).$$

$$m_{it} \in \{\mu_1^*, \dots, \mu_G^*\}.$$

Individuals can switch cluster, so their means can change over time:

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

We say $\gamma_{it} = g$ if $m_{it} = \mu_g$.

What is this about?

- I refine the test for clustering of P&W to allow for cluster-switching.
- This improves power in settings with frequent switching.
- Some insights are provided on why power increases.
- I present an illustration based on the well-know application of Bonhomme and Manresa (2015).

Objective of the test

Most clustering procedure use a criterion to determine the number of cluster.

These criteria are often undefined for $G = 1$ (e.g. the Silhouette).

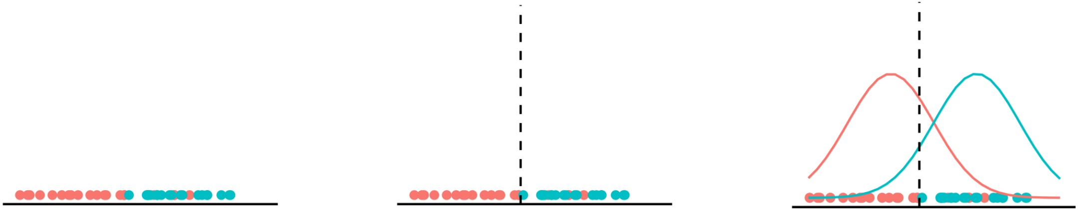
We want a test for $G = 1$, i.e. the null hypothesis that $m_{it} = \mu^* \ \forall \ i, t$.

Intuition of the test

k-means will divide the data into k groups no matter what.

Their centers are asymptotically normal means.

An F-test of equal means can be constructed.



What if there's switching

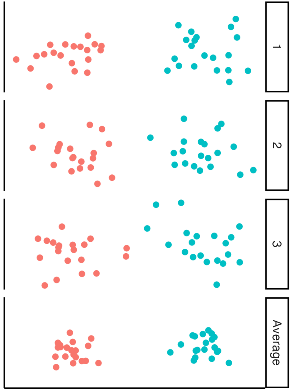
P&W is still valid, but it loses power!

Clustering in settings with a lot of switching cannot be detected.

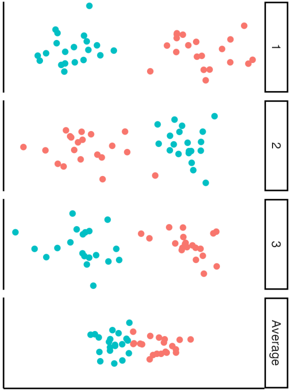
Their test works on average distance of Y_{it} to the cluster centers over t.
The estimated means given cluster assignments γ are:

$$\hat{\mu}(\gamma) = \arg\min_{\mu} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{g=1}^{G} ||Y_{it} - \mu_g||^2 \mathbb{1}\{\gamma_i = g\}$$

What if there’s switching



$p = 0$



$p = 1$

Under switching that is unaccounted for, it is harder to distinguish the cluster means.

Intuition of the solution

Simply cluster every point (i, t) as an independent observation.

Sample splitting

The original test employs an arbitrary sample splitting approach.

We cluster on sample $\mathcal{R}$, and estimate the means on sample $\mathcal{P}$.

$$\{1, 2, \ldots, T\} = \mathcal{R} \cup \mathcal{P}$$

To account for switching, let $\mathcal{R}$ be odd time indices, and $\mathcal{P}$ be the even indices.

$$\mathcal{R} = \{1, 3, 5, \ldots, T-1\}$$

$$\mathcal{P} = \{2, 4, 6, \ldots, T\}$$

Overview of the testing procedure

In P&W:

1. Apply k-means on sample $\mathcal{R}$ yielding assignments $\hat{\gamma}_i$
2. Calculate cluster means on sample $\mathcal{P}$ yielding $\tilde{\mu}_{NP}(\hat{\gamma})$
3. Calculate the test statistic F_{NPR} based on $\hat{\gamma}_i$, $\tilde{\mu}_{NP}$, and the $\mathcal{P}$ sample.
4. Under $H_0 : \mu_1 = \mu_2 = \ldots = \mu_G$, $F_{NPR} \xrightarrow{d} \chi^2_{d(G-1)}$

Here:

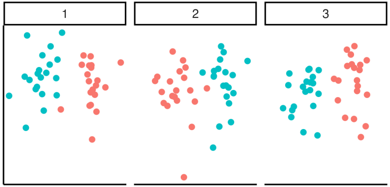
- Different sample splitting.
- Time-varying assignments γ_{it}.

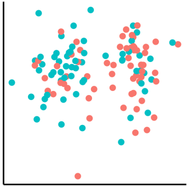
Clustering

As P&W, we use k-means clustering but let assignments vary over time:

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

This is akin to clustering as if there was no time dimension.





Clustering

Then, cluster means are estimated on the $\mathcal{P}$ sample:

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

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

Because P&W doesn't have time-varying assignments, means are calculated from a mix of observations in and out of the cluster.

The test statistic: building blocks

Estimator of the cluster-specific means:

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

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

and $\overline{Y}_{i,g}$ are cluster-specific individual means.

The null hypothesis is denoted $H_0 : \mu_g^* = \mu_{g'}^* \; \forall \; g \neq g' \iff A_{d,G}\mu^* = 0$ for a suitably defined matrix $A_{d,G}$.

The test statistic

Theorem

Define the test statistic for the differences in the estimated means as

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

(a) Under H_0,

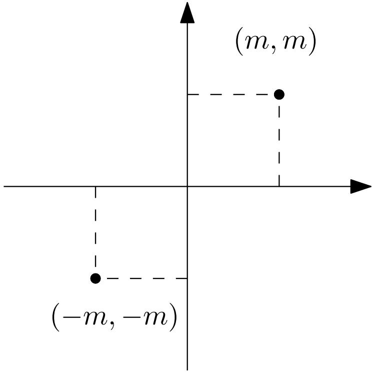
$$F_{NPR} \xrightarrow{d} \chi^s_{d(G-1)}, \text{ as } N, P, R \to \infty$$

(b) Under H_1,

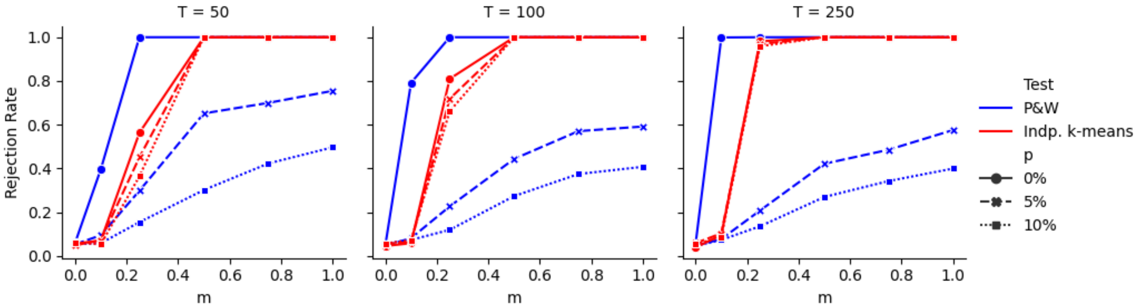
$$F_{NPR} \xrightarrow{p} \infty, \text{ as } N, P, R \to \infty$$

Simulation setting

- 2 clusters in the DGP, on 2 dimensions.
- Normally distributed with identity covariance matrix.
- Centered at (m, m) and $(-m, -m)$ with m varying from 0 to 1.
- Probability of switching $p \in \{0,\ 5\%,\ 10\%\}$.
- Compare the test above with P&W.

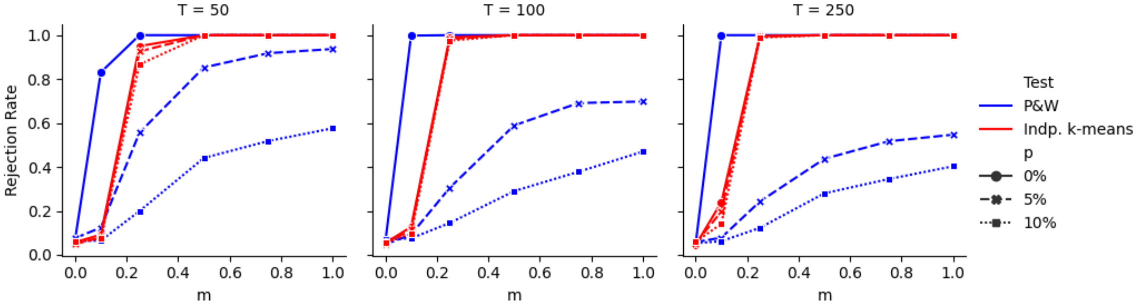


Power results, $N = 30$



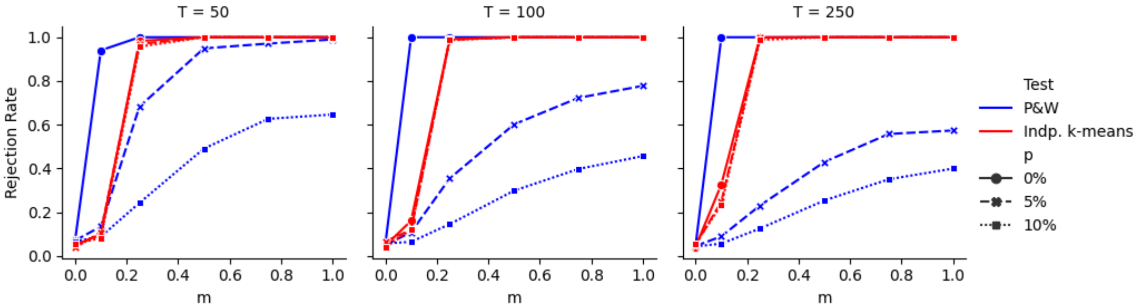
With switching, larger T increases the misclassification rate and reduces the power of the P&W test.

Power results, $N = 100$



In almost all settings, 5% is enough to create a difference in power.

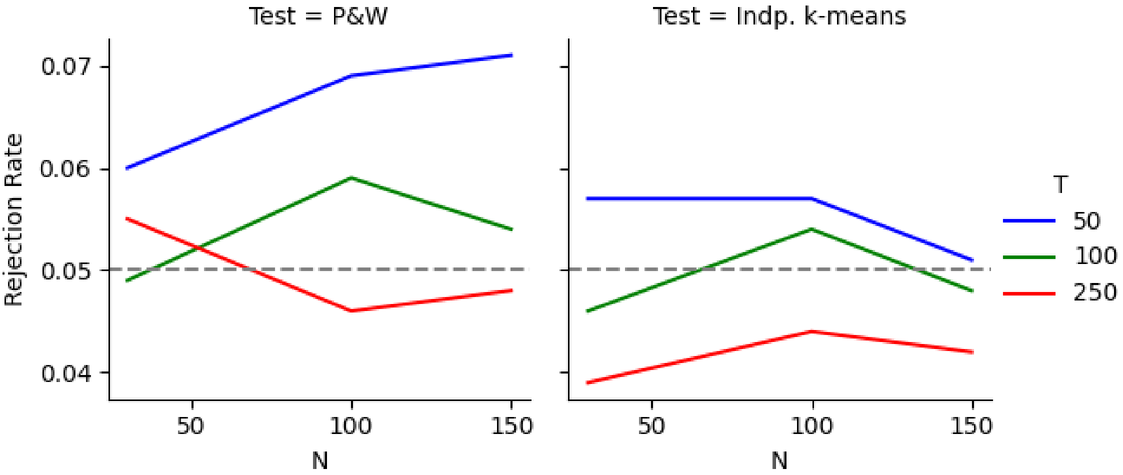
Power results, $N = 150$



Power increases when clusters are more separated.

But even at $m = 1$ power can be low in P&W, around 60%.

Size results ($m = 0$)



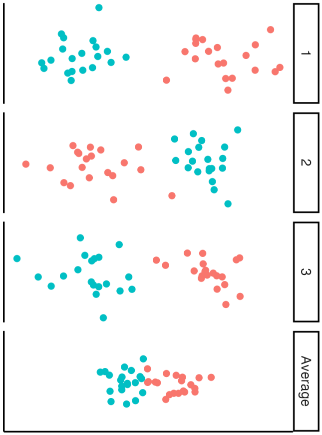
Power: Intuition

When $p = 0$, clustering on time averages consistently estimates the true means.

With switching, this is not possible. Both methods systematically misclassify.

- Pooled k-means in P&W mixes the clusters and produces means closer together than they should be.
- Independent k-means misclassify outliers in different clusters and produces means farther apart than they should be.

Intuition: P&W case

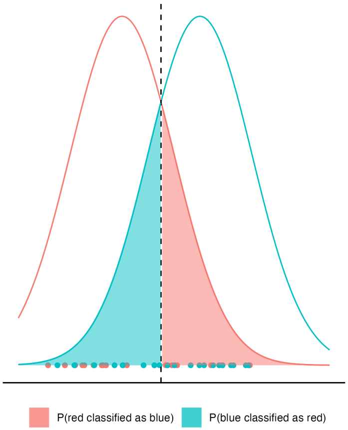


Each unit *i* can only belong to one cluster.

Averaging over time includes realizations from both distributions.

The mean gets closer to the global mean with higher *p*.

Intuition: independent case



As this can be seen as a large cross-section, there's always a non-zero probability of misclassification.

Misclassification happens on the tails of the cluster distributions.

These misclassified points shift the cluster means away from each other.

Closer look: P&W case

Setting: 2 clusters, 1 dimension.

The k-means procedure alternates between:

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

and

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

Closer look: P&W case

Suppose that I start with the correct estimate of the means. Let $\mu_1 < \mu_2$. First assignment step:

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

$$= \begin{cases} 1 & \text{if } R^{-1} \sum_{t \in \mathcal{R}} Y_{it} \leq (\hat{\mu}_1^0 + \hat{\mu}_2^0)/2 \\ 2 & \text{otherwise} \end{cases}$$

Closer look: P&W case

Recalculating the mean:

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

At the limit of N and R, clusters are mixed.

$$\lim_{N \to \infty} \hat{\mu}_1^1 = \mathbb{E}_i \left(\frac{1}{R} \sum_{t \in \mathcal{R}} y_{it} \middle| \frac{1}{R} \sum_{t \in \mathcal{R}} y_{it} \leq \frac{\hat{\mu}_1^0 + \hat{\mu}_2^0}{2} \right)$$

$$\lim_{R \to \infty} \lim_{N \to \infty} \hat{\mu}_1^1 = \mathbb{E}_i \left(\mathbb{E}_t(y_{it}) \middle| \mathbb{E}_t(y_{it}) \leq \frac{\hat{\mu}_1^0 + \hat{\mu}_2^0}{2} \right) = \frac{\mu_1 + \mu_2}{2}$$

And likewise for $\hat{\mu}_2$. The centers approach their average becoming indistinguishable.

Closer look: independent case

Same setting as before, but the subscript t is irrelevant. So let's count from 1 to $M := NR$

Again, start from the correct means.

$$\hat{\gamma}_i^0(\hat{\mu}^0) = \arg\min_{\gamma} \sum_{g=1}^{2} (Y_i - \hat{\mu}_g^0)^2 \mathbb{1}\{\gamma = g\} = \begin{cases} 1 & \text{if} \quad Y_i \leq (\hat{\mu}_1^0 + \hat{\mu}_2^0)/2 \\ 2 & \text{otherwise} \end{cases}$$

The next-iteration means will be

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

Closer look: independent case

At the limit of M:

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

where f is the mixture distribution with equal weights. We can decompose it in f_1 and f_2 and write:

$$\lim_{M\to\infty} \hat{\mu}_1^1 = \int_{x \leq \hat{\mu}_2^0/2} x(f_1(x) + f_2(x)) \, \mathrm{d}x$$

Then we can show that

$$\lim_{M\to\infty} \hat{\mu}_1^1 < \mu_1 \quad \text{and} \quad \lim_{M\to\infty} \hat{\mu}_2^1 > \mu_2$$

And hence the estimated means are farther apart.

Closer look: independent case

$$\lim_{M\to\infty} \hat{\mu}_1^1 < \mu_1 = \mathbb{E}_{f_1}(x)$$

$$\int_{x \le \hat{\mu}_2^0/2} x f_2(x)\,\mathrm{d}x < \int_{x > \mu_2^0/2} x f_1(x)\,\mathrm{d}x$$

$$\int_{z \ge \hat{\mu}_2^0/2} (\mu_2 - z) f_2(\mu_2 + z)\,\mathrm{d}x < \int_{x > \mu_2^0/2} x f_2(x + \mu_2)\,\mathrm{d}x$$

$$\int_{z \ge \hat{\mu}_2^0/2} (\mu_2 - 2z) f_2(\mu_2 + z)\,\mathrm{d}x < 0$$

The condition is satisfied as $f_2(x) > 0 \ \forall \ x$ and $(\mu_2 - 2z) < 0 \ \forall \ z > \hat{\mu}_2^0/2$.

Application

I revisit the application of Bonhomme and Manresa (2015).

They build on Acemoglu et al. (2008) and their data to estimate a model for democracy

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

where $\alpha_{g_i,t}$ are group fixed effects.

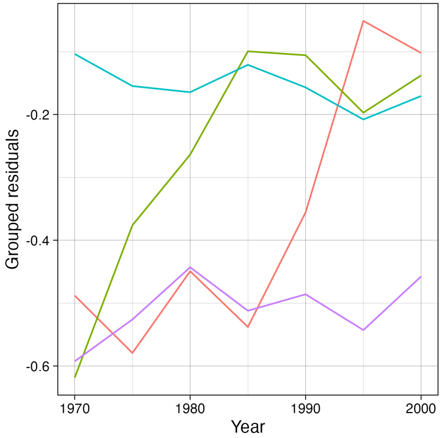
Application

They employ an iterative procedure to estimate the parameters and group assignments.

$$g_i^{(s)} = \underset{g \in \{1,\ldots,G\}}{\arg\min} \sum_{t=1}^{T} (y_{it} - x_{it}'\theta^{(s)} - \alpha_{g,t}^{(s)})^2$$

$$(\theta^{(s+1)}, \alpha^{(s+1)}) = \underset{\theta,\alpha}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it}'\theta - \alpha_{g_i^{(s+1)},t})^2$$

Application: motivation



They find 4 clusters of fixed effects.

2 of them are characterized by moving up over time.

Looks like switching between two clusters.

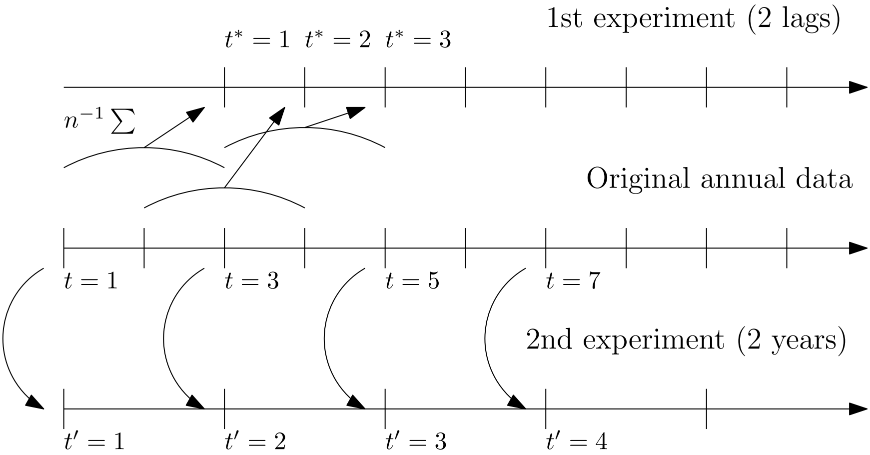
I estimate their model and test for clustering on the individual residuals on a variety of settings.

Application: data

Two exercises:

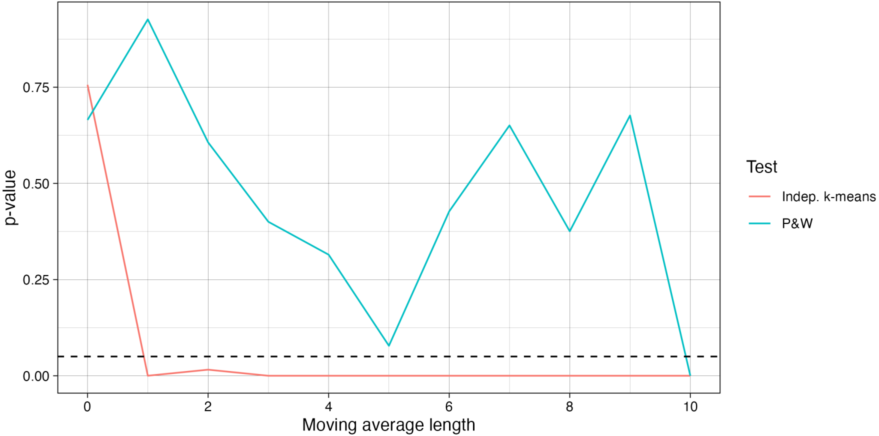
1. Annual data from 1975 to 2000.
   I calculate overlapping moving averages using 0 to 10 year lags.
   As the moving average window expand, clusters become clearer.
2. Annual data from 1970 to 2000.
   I sample the data at intervals of 1 to 5 years.
   At 5 years, we are in the empirical setting of Bonhomme and Manresa (2015).

Application: data



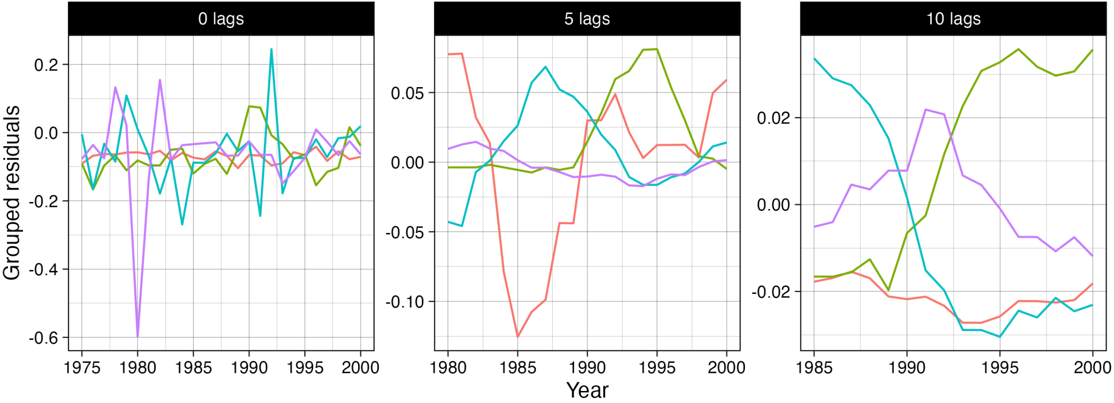
Application: 1st exercise

As the residuals become smoother, p-values drop for both tests.



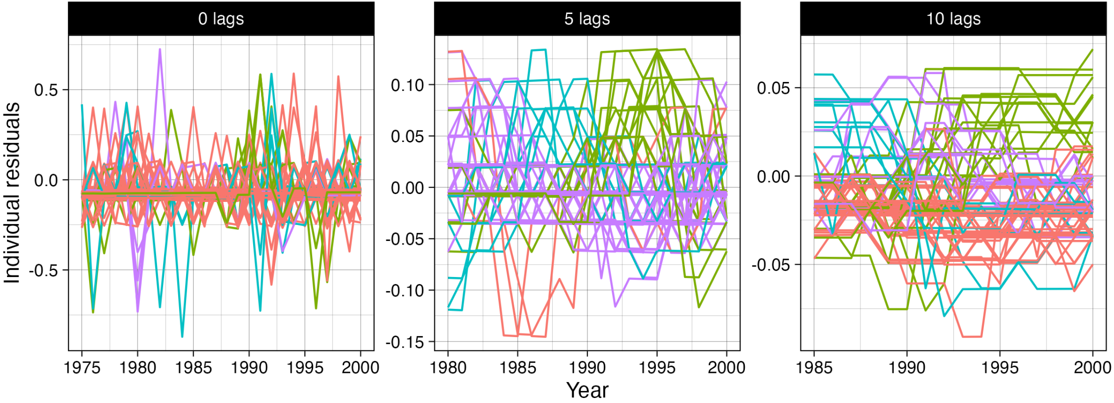
Application: 1st exercise

The grouped fixed effect terms $\alpha_{g,t}$ become smoother:

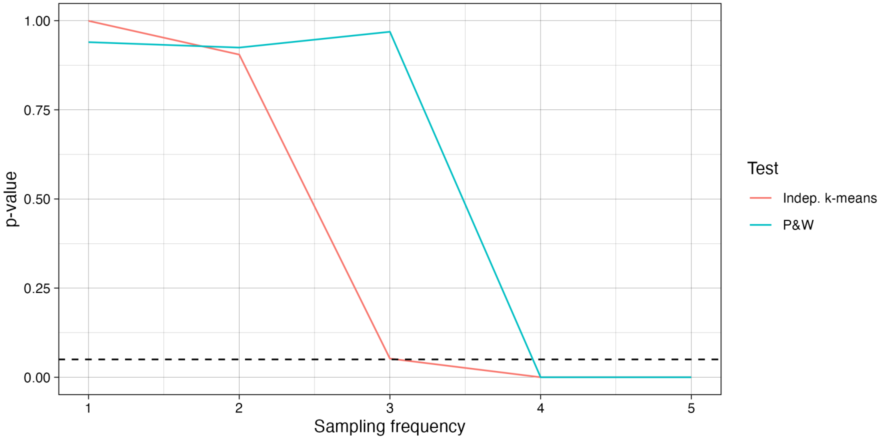


Application: 1st exercise

The individual residuals too:

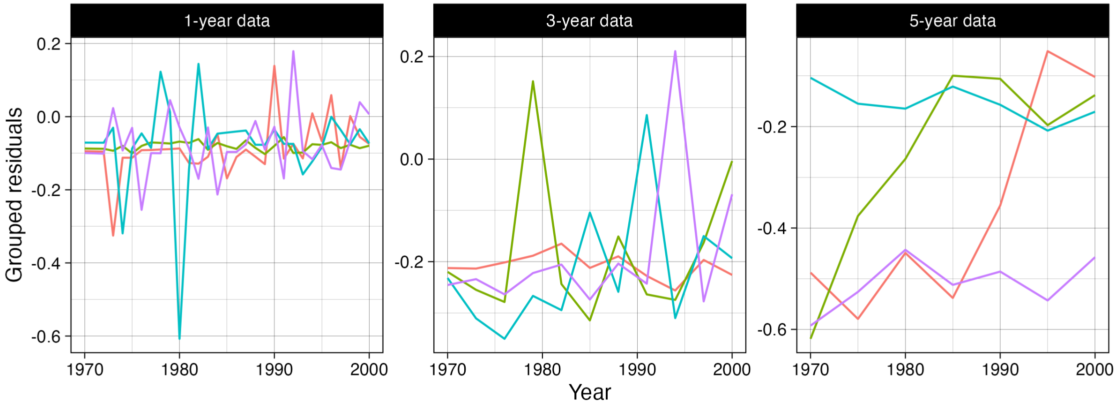


Application: 2nd exercise



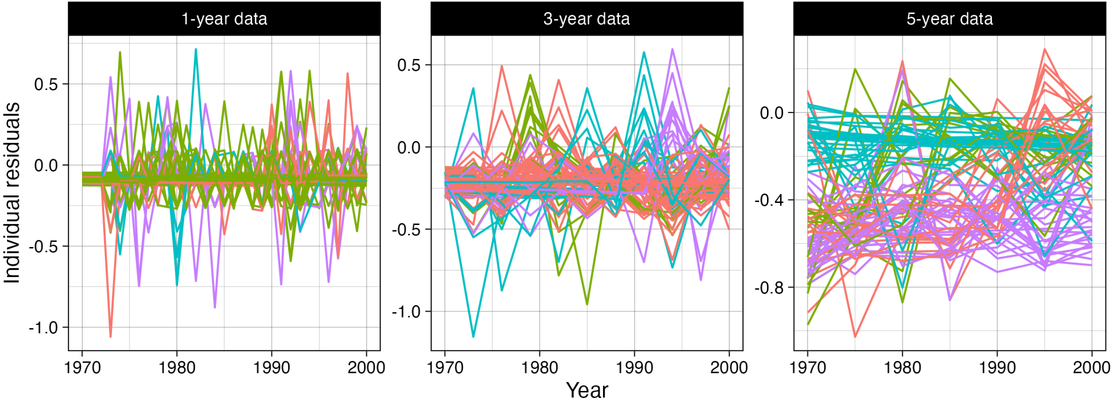
Application: 2nd exercise

The grouped fixed effect terms $\alpha_{g,t}$ become smoother:



Application: 2nd exercise

The individual residuals too:



Conclusion

- In settings with cluster switching, P&W can be underpowered.
- We can improve the power by clustering independently.
- Size is still controlled, power increases with large T and switching probability p.
- Power comes from the asymptotic bias of the means being on opposite directions.
- In an empirical setting this is relevant when clusters are not so neatly salient.

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.

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.