Testing Clustered Equal Predictive Ability with Unknown Clusters

Oğuzhan Akgün[1], Alain Pirotte[2], Giovanni Urga[3], and Zhenlin Yang[4]

[1] *LEDi, University of Burgundy, France*

[2] *CRED, Paris-Panthéon-Assas University, France*

[3] *Bayes Business School (formerly Cass), London, United Kingdom*

[4] *School of Economics, Singapore Management University, Singapore*

August 14, 2024

Abstract

This paper proposes new tests for the hypothesis of clustered equal predictive ability (C-EPA) in panels in the case of unknown clusters. The unknown clusters are estimated via *kmeans* which are then used to perform the test. To address the problem with testing a hypothesis selected from data, the selective inference approach as well as a sample splitting solution are adopted. The proposed framework allows comparing the forecast performance of agents or predictive models. The asymptotic properties of the tests are studied. Monte Carlo results reveal excellent finite sample properties, with only negligible size distortions and very high power even under weak deviations from the C-EPA null. Last, the empirical relevance of our tests is illustrated over a large panel data set of exchange rate forecasts.

Keywords: Forecast Evaluation; Hypothesis Testing; *kmeans*; Selective Inference.

JEL classification: C12, C23.

Email addresses: oguzhan.akgun@u-bourgogne.fr (Corresponding author, O. Akgun), alain.pirotte@u-paris2.fr (A. Pirotte), g.urga@city.ac.uk (G. Urga), zlyang@smu.edu.sg (Z. Yang).

1. Introduction

Despite the large and ever-growing time series literature,[1] testing equal predictive ability (EPA) using panel data has attracted attention among econometricians only recently. To the best of our knowledge, the only contributions are those of Akgun et al. (2024, APUY hereafter) and Qu et al. (2024, QTZ hereafter). Both papers focus on two EPA hypotheses: the overall EPA (O-EPA) hypothesis and the clustered EPA (C-EPA) hypothesis, where *overall* refers to the equivalence of two forecasts for a given loss function on average over all time periods and all panel units, whereas *clustered* means that two competing forecasts are equivalent for G clusters of panel units on average over all time periods.

The main contribution of this paper is twofold. First, we develop tests for the C-EPA hypothesis considered by APUY and QTZ in the case of unknown clusters. These unknown clusters are estimated from data which are then used to test the C-EPA hypothesis. To address the problem with testing a hypothesis selected from data, we establish a selective inference framework and a sample splitting solution based on *kmeans* estimates of the cluster membership and cluster centers. Second, we propose conditional panel EPA tests, thus extending to panels the framework by Giacomini and White (2006) in a time series context, and by Qu et al. (2023) in cross-sections.

Developing tests on the cluster centers which successfully control the Type I error rate following the estimation of the unknown clusters constitutes the main theoretical contribution of this paper. Several well-known clustering methods exist, such as hierarchical and *kmeans* clustering.[2] In this paper, we focus on the *kmeans* clustering approach which is arguably the most commonly used method in econometrics (see Lin and Ng, 2012; Bonhomme and Manresa, 2015; Sarafidis and Weber, 2015; Bonhomme et al., 2022, among others). If the predictive abilities of two forecasters differ so that, while they are equally good (or bad) within clusters, they differ between clusters, *kmeans* can detect these clusters under general conditions. That is, the *kmeans* estimator of the cluster centers is consistent if the clusters are well separated. However, under the C-EPA hypothesis, this assumption does not hold, which implies that there is only one cluster in the population because all clusters are mean zero. Hence, the problem of C-EPA testing post-clustering falls into the category which is called "double dipping" (see, for instance Kriegeskorte et al., 2009) to describe the use of the same data to generate a hypothesis and test it. Recent literature in statistics and econometrics recognizes the consequence of

[1] See Giacomini (2011), Clark and McCracken (2013), and Rossi (2021) for reviews of the early and more recent contributions to the area.

[2] See Ikotun et al. (2023) for a recent review of the *kmeans* clustering algorithms.

double dipping in clustering and shows that the Wald tests based on estimated clusters are *extremely anti-conservative*, if there are no heterogeneous clusters in the underlying population distribution (Gao et al., 2024; Chen and Witten, 2023; Patton and Weller, 2023).

A straightforward way to deal with the problem of double dipping is sample splitting. In a cross-sectional setting, Gao et al. (2024) show that sample splitting does not provide a valid way to test null hypotheses on cluster centers. However, the time series dimension of a panel provides a solution to this, as in Patton and Weller (2023) where a split-sample test statistic is proposed to test the homogeneity of the mean of a panel process among clusters chosen by *kmeans*. Despite its theoretical and computational simplicity, sample splitting has its drawbacks. First of all, split-sample test statistics rely on the selection of two sub-samples. One sample is used for the estimation of the clusters and another for inference on the centers of these estimated clusters. However, this selection can be arbitrary in practice and there is no guidance in the literature on how to split the sample.[3] Second, the fact that the inference is based on a number of observations smaller than the whole sample may cause the associated test statistics to have high size distortions or low power. Third, the validity of the split-sample method is not guaranteed for dependent data (Kuchibhotla et al., 2022). Lunde (2019) shows that the split-sample approach is valid under weak-dependence conditions but their framework covers variable selection in a regression model and it is not necessarily valid for clustering. Patton and Weller (2023) propose a solution to the case where general time series dependence of $q \geq 1$ lags are allowed but impose independence beyond q lags. They show that, in this case, sample splitting continues to be valid if q periods between the two sub-samples are discarded. This validates the use of the split-sample tests but makes the drawbacks even more pronounced.

In this paper, we develop an alternative selective inference framework which uses the full sample of observations for both estimating the unknown clusters and making inference on their centers. Our framework is motivated by recent papers by Gao et al. (2024) and Chen and Witten (2023), who propose a selective p-value to test the equality of two cluster means post-clustering which are, however, based on strong assumptions on the data generating process such as normality, homoskedasticity and independent observations which we relax in this paper. Furthermore, their testing procedure concerns with the equality of only two cluster means.

We develop a selective p-value procedure for testing the equality of two cluster centers for heteroskedastic, dependent and potentially non-Gaussian panel data. Then we apply a p-value combina-

[3]For an attempt to answering this question in a related but different context, see Hansen and Timmermann (2012).

tion method using the p-values of $G-1$ pairwise homogeneity tests together with that of an O-EPA test. To deal with the dependencies in the time series dimension, a heteroskedasticity and autocorrelation robust variance estimator is employed following Sun (2013, 2014). This estimator is then applied to the cross-sectional averages of the loss differentials which in turn provides a test statistic robust to arbitrary form and strength of cross-sectional dependence (CD) (see Driscoll and Kraay, 1998).

We derive the limiting theory of the proposed test statistics. In particular, we show that the tests are correctly sized, and consistent under general alternatives even in the presence of arbitrary weak time series correlation and strong CD. In order to establish the asymptotic power of the tests, we prove that the *kmeans* estimator of the cluster centers remain consistent under strong CD contrary to the weak dependence assumptions in Bonhomme and Manresa (2015) and Patton and Weller (2023) which is, to the best of our knowledge, a result which has not previously appeared in the literature.

The small sample properties of the proposed tests are assessed via an extensive Monte Carlo simulation, and are compared with a set of split-sample test statistics. The results show that our test statistics have optimal properties even in samples which can be considered very small in potential applications. In particular, our tests have negligible size distortions in very small samples and have high power even under weak deviations from the C-EPA null.

The empirical relevance of our proposed methodology is illustrated in an application on model comparison. Using the data set on the exchange rate forecasts compiled by Spreng and Urga (2023, SU hereafter), different time series models are compared. The results show that there are exchange rate clusters for which the predictive ability of different models differ significantly.

The remainder of the paper is organized as follows. Section 2 presents the null and the alternative hypotheses of interest, and the *kmeans* estimator of the unknown clusters. Section 3 introduces the test statistics and presents their asymptotic properties. Section 4 presents essential Monte Carlo results. An empirical illustration is reported in Section 5. Section 6 concludes. Appendices A-C contain the proofs of the theoretical results and additional simulation evidence.

Notation. Random variables are denoted by upper-case letters and their realizations by the corresponding lower-case letters. For instance, $w_{NT}(\cdot,\cdot)$ denotes a particular realization of the test statistic $W_{NT}(\cdot,\cdot)$. Further, $\|\cdot\|$ denotes Euclidean norm, $\mathbf{1}\{\cdot\}$ is indicator function, $\text{diag}(\cdot)$ forms a diagonal matrix by given elements, $[\ \cdot\ ]$ returns an integer by rounding, $\otimes$ denotes Kronecker product, $\xrightarrow{p}$ convergence in probability, $\xrightarrow{d}$ convergence in distribution, and $(T,N)\to\infty$ the joint passage to infinity of T and N.

2. Setup

In this section, we introduce the testing framework, the C-EPA null and alternative hypotheses, and the assumptions; we then introduce our proposed conditional C-EPA test for predetermined clusters; finally, we present the *kmeans* estimator and the associated algorithm.

2.1. Testing framework, hypotheses and assumptions

Let $\widehat{Y}_{a,it}$ be the τ-steps ahead, $\tau \geq 1$, forecast of forecasters $a = 1, 2$ for the target variable Y_{it} made at time $t-\tau$, $t = 1, 2, \ldots, T$, for unit $i = 1, 2, \ldots, N$. Here, a represents a forecasting agent such as IMF, OECD, or a forecasting model. Let $L(\cdot, \cdot)$ be a generic loss function. This can be a quadratic loss, an absolute loss or a loss function which is not necessarily in the forecast error form. Define the loss differentials of the two forecasts as $\Delta L_{it} = L(\widehat{Y}_{1,it}, Y_{it}) - L(\widehat{Y}_{2,it}, Y_{it})$ which are assumed to be specified on a complete probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

The null and the alternative hypotheses. The null hypothesis of interest is the generalized C-EPA hypothesis. We call the null of this paper the "*generalized* C-EPA hypothesis" because it allows conditioning variables contrary to the null hypotheses considered in recent papers by APUY and QTZ. This null hypothesis is stated as

$$\mathcal{H}_0 : \lim_{N\to\infty} \frac{1}{n_g} \sum_{i=1}^{N} \mathrm{E}(\Delta L_{it} \mid \mathcal{G}_{t-\tau}) \mathbf{1}\{g_i = g\} = 0, \text{ a.s., for all } g = 1, 2, \ldots, G \tag{1}$$

where $\mathcal{G}_t \subseteq \mathcal{A}$ is a conditioning set (see the description below) and $n_g = \sum_{i=1}^{N} \mathbf{1}\{g_i = g\}$ with $g_i \in \{1, 2, \ldots, G\}$ being the cluster membership variable stating the cluster which ith unit belongs to. Typically, the clusters are chosen to be mutually exclusive and exhaustive. That is, for any $g \neq g'$, $\sum_{i=1}^{N} \mathbf{1}\{g_i = g\} \times \mathbf{1}\{g_i = g'\} = 0$ and $\sum_{g=1}^{G} \sum_{i=1}^{N} \mathbf{1}\{g_i = g\} = N$. Moreover, n_g is assumed to be diverging to infinity with N (see Assumption 2 below). The alternative hypothesis is

$$\mathcal{H}_1 : \lim_{N\to\infty} \frac{1}{n_g} \sum_{i=1}^{N} \mathrm{E}(\Delta L_{it} \mid \mathcal{G}_{t-\tau}) \mathbf{1}\{g_i = g\} \neq 0, \text{ a.s., for at least one } g = 1, 2, \ldots, G. \tag{2}$$

In the formulation of the hypotheses, it is implicitly assumed that the conditional expectation of interest is time invariant almost surely. With a more complicated notation and without much gain of insight, we could also focus on the averages of these expectations over time.

Two cases covered in the null hypothesis (1) and the corresponding alternative hypothesis are important. The first null is the unconditional C-EPA hypothesis which is obtained when $\mathcal{G}_t = \{\emptyset, \Omega\}$.

For predetermined clusters, the tests for this null hypothesis were developed by APUY and QTZ under different assumptions on the autocorrelation and CD properties of the loss differentials. The second null hypothesis that we consider is the conditional C-EPA hypothesis. Two sub-cases of the conditional null are particularly useful. First, let $\mathcal{W}_t = \sigma(\{W_{is}\}_{i=1}^N, s \leq t)$, the σ-field generated by the present and the past of the measurable-$\mathcal{A}$ random variables $W_{it} = (Y_{it}, X_{it}')'$ with X_{it} being a vector of external predictors used to make the predictions $\widehat{Y}_{a,it}$. Then an interesting null hypothesis of the form (1) is obtained when $\mathcal{G}_t = \mathcal{W}_t$. Second, a researcher may be interested in the conditional EPA with respect to the realization of a vector of measurable-$\mathcal{A}$ common factors F_t. In this case, we let $\mathcal{F}_t = \sigma(F_s, s \leq t)$ and $\mathcal{G}_t = \mathcal{F}_t$. Some common factors can be particularly useful to model via dummy variables indicating, for example, the global financial crises, the COVID-19 period, etc. Through a careful choice of these dummies, our framework makes it possible to focus on local differences in the predictive abilities of the two forecasters.

The null hypothesis $\mathcal{H}_0$ implies that $\lim_{N\to\infty} n_g^{-1} \sum_{i=1}^N \mathrm{E}(\widetilde{H}_{i,t-\tau}\Delta L_{it})\mathbf{1}\{g_i = g\} = 0$, for any measurable-$\mathcal{A}$ vector of random variables $\widetilde{H}_{it}$ (Giacomini and White, 2006). Let H_{it} be such a $K \times 1$ vector, called a "testing function" by Giacomini and White (2006) and $Z_{it} = H_{i,t-\tau}\Delta L_{it}$. We define $\beta_i^0 = \mathrm{E}(Z_{it})$ which is assumed to be time invariant, and $V_{it} = Z_{it} - \beta_i^0$. Moreover, $\theta_{g,n_g}^0(\gamma) = n_g^{-1} \sum_{i=1}^N \beta_i^0 \mathbf{1}\{g_i = g\}$, $\theta_{g,n_g}^0(\gamma) \in \Theta_g^K \subseteq \mathbb{R}^K$, for all g, where $\gamma := \gamma_{N,G} = (g_1, \ldots, g_N)' \in \Gamma_{N,G} \subseteq \mathbb{R}^N$ is the vector of membership variables with G distinct elements. Here, Θ_g^K is the parameter space of the gth cluster center and $\Gamma_{N,G}$ denotes the set of membership vectors associated to all vectors γ involving G distinct elements. Then testing $\mathcal{H}_0$ is equivalent to testing

$$\mathcal{H}_0' : \lim_{N\to\infty} \theta_{g,n_g}^0(\gamma) = 0, \text{ for all } g = 1, 2, \ldots, G. \tag{3}$$

Let $\bar{V}_{g,n_g,t}(\gamma) = n_g^{-1} \sum_{i=1}^N V_{it}\mathbf{1}\{g_i = g\}$ be the cross-sectional mean of the innovations of cluster g, $\bar{V}_{N,t}(\gamma) = [\bar{V}_{1,n_1,t}'(\gamma), \ldots, \bar{V}_{G,n_G,t}'(\gamma)]'$ with a given $\gamma \in \Gamma_{N,G}$ where $G \geq 1$. We also collect the true cluster means in $\theta^0(\gamma) = (\theta_{1,n_1}^{0\prime}(\gamma), \ldots, \theta_{G,n_G}^{0\prime}(\gamma))' \in \Theta^{GK} \subseteq \mathbb{R}^{GK}$ where $\Theta^{GK} = \Theta_1^K \times \cdots \times \Theta_G^K$. The following assumptions will be referred to throughout the paper.

Assumption 1. *(a) Θ^{GK} is a compact set of $\mathbb{R}^{GK}$, (b) $\mathrm{E}\|V_{it}\|^4 \leq C$, (c) $T^{-1}\sum_{t,s=1}^T \mathrm{E}\|V_{it}V_{is}'\| \leq C$.*

Assumption 2. *Let $\gamma \in \Gamma_{N,G}$ with $G \geq 2$. Then as $N \to \infty$, $n_g/N \to \kappa_g \in (0,1)$ for each $g = 1, \ldots, G$.*

Assumption 3. *$\Omega(\gamma)^{-1/2}\mathcal{N}^{1-\epsilon}T^{-1/2}\sum_{t=1}^T \bar{V}_{N,t}(\gamma) \xrightarrow{d} \mathbb{N}(0, I_{GK})$ for some $\epsilon \in [1/2, 1]$ as $(T, N) \to \infty$, for each $\gamma \in \Gamma_{N,G}$, with $\Omega(\gamma) = \sum_{s=-\infty}^{\infty} \mathcal{N}^{1-\epsilon}\mathrm{E}[\bar{V}_{N,t}(\gamma)\bar{V}_{N,t-s}'(\gamma)]\mathcal{N}^{1-\epsilon}$ being positive definite and*

$\mathcal{N} = \text{diag}(n_1, \ldots, n_G) \otimes I_K$.

Assumptions 1(a) and 1(b) are standard conditions which ensure that the cluster centers are well defined and all moments up to the fourth of the innovation process V_{it} exist. Assumption 1(c) limits the time series dependence. We do not place any restriction on the CD characteristics of the panel and allow for both strong and weak CD.

Assumption 2 controls the asymptotic number of units per cluster for a given γ. It is standard in the econometrics literature of clustering (see, for instance Assumption 2(a) of Bonhomme and Manresa (2015) and Assumption A1(vii) of Su et al. (2016)). It states that each cluster has a non-negligible contribution. This assumption can be relaxed at the expense of more complicated notation.

The CLT in Assumption 3 holds under standard mixing conditions. Let $V_{k,it}$, $k = 1, \ldots, K$, be the kth element of V_{it}. The sufficient conditions for it to hold are, for all $i = 1, \ldots, N$, V_{it} is weakly stationary with $\Omega_i = \sum_{j=-\infty}^{\infty} \text{E}[V_{it} V_{i,t-j}']$ being positive definite and either (i) $\text{E}(|V_{k,i1}|)^{\zeta} < \infty$ for all $k = 1, \ldots, K$ and for $\zeta \geq 2$, (ii) V_{it} is φ-mixing with $\sum_{l=1}^{\infty} \varphi_l^{1-1/\zeta} < \infty$, or (i) $\text{E}(|V_{k,i1}|)^{\zeta} < \infty$ for all $k = 1, \ldots, K$ and for $\zeta > 2$, (ii) V_{it} is α-mixing with $\sum_{l=1}^{\infty} \alpha_l^{1-2/\zeta} < \infty$ (Phillips and Durlauf, 1986). Then the same conditions hold for $\mathcal{N}^{1-\epsilon} \bar{V}_{N,t}(\gamma)$ because the mixing properties are hereditary. The scalar $\epsilon \in [1/2, 1]$ defined in Assumption 3 measures the degree of CD in V_{it}. The case of $\epsilon = 1$ corresponds to the case of strong CD of the loss differentials and if $\epsilon = [1/2, 1)$, they are weakly cross-sectionally dependent. Weak CD includes the case of independence over $i = 1, \ldots, N$. We refer to Chudik et al. (2011) for examples of panel models satisfying different cases of CD and Bailey et al. (2016) for the estimation of the parameter ϵ when $\epsilon = (1/2, 1]$.

By Assumption 3, $\Omega(\gamma)$ is positive definite which is a well-known condition for the validity of EPA testing using Diebold and Mariano (1995) type tests (West, 1996). This assumption means that the forecasts are made by either non-nested models or they satisfy the conditions of Giacomini and White (2006) for nested models. In particular, if two models are nested, they need to be made using rolling window or fixed estimation sample forecasting schemes. An expanding window scheme is ruled out in the case of nested model comparisons (see McCracken, 2020; Zhu and Timmermann, 2020, for counter arguments for the validity of fixed estimation sample scheme). For general nested model comparisons, we refer to the recent paper by Clark and McCracken (2015) and the references therein.

Let $\hat{\theta}_{g,n_g,T}(\gamma) = \bar{Z}_{g,n_g,T}(\gamma) = (n_g T)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} Z_{it} \mathbf{1}\{g_i = g\}$ be the sample mean of cluster g, and $\hat{\theta}_{NT}(\gamma) = (\hat{\theta}_{1,n_1,T}'(\gamma), \ldots, \hat{\theta}_{G,n_G,T}'(\gamma))'$. The following standard result will be used as the basis for robust EPA testing in this paper.

Lemma 1. Under Assumptions 1-3, the following results hold as $(T, N) \to \infty$:

(a) $\hat{\theta}_{NT}(\gamma) = \theta^0(\gamma) + o_p(1)$,

(b) $\Omega(\gamma)^{-1/2}\mathcal{N}^{1-\epsilon}T^{1/2}(\hat{\theta}_{NT}(\gamma) - \theta^0(\gamma)) \xrightarrow{d} \mathbb{N}(0, I_{GK})$.

Part (a) is a law of large numbers which shows that Assumptions 1-3 are sufficient for the consistency of the sample means for the cluster centers defined by a given γ. Part (b) is the corresponding central limit theorem. The result concerns with the case of a fixed $\gamma \in \Gamma_{N,G}$ and does not necessarily hold with estimated cluster memberships. Below, we will make use of this result in a conditional framework to obtain the asymptotic properties of our proposed tests with estimated clusters.

2.2. Generalized clustered EPA tests with predetermined clusters

The tests for the unconditional C-EPA hypothesis have been developed by APUY and QTZ for predetermined clusters. When $\mathcal{G}_t = \{\emptyset, \Omega\}$, the C-EPA null reduces to $\lim_{N\to\infty} n_g^{-1} \sum_{i=1}^{N} \mathrm{E}(\Delta L_{it})\mathbf{1}\{g_i = g\} = 0$ for all $g = 1, 2, \ldots, G$ and we obtain the unconditional C-EPA hypothesis. APUY suggested several test statistics under different assumptions on the dependence structure of the loss differentials. Here, we generalize their methodology to the case of $\mathcal{G}_t \neq \{\emptyset, \Omega\}$ together with a small sample adjustment to their test statistics. Let γ be a vector of predetermined cluster membership variables implying $n_g \geq 1$ for all $g = 1, 2, \ldots, G$. The case of $n_g = 0$ for some g is trivial and leads only to the reduction of the number of groups. The hypothesis in (3) can be tested using

$$W_{NT}(H_{i,t-\tau}, \gamma) = \frac{B - GK + 1}{GKB} T\hat{\theta}'_{NT}(\gamma)\widehat{\Omega}^{-1}_{NT}(\gamma)\hat{\theta}_{NT}(\gamma), \tag{4}$$

where $\widehat{\Omega}_{NT}(\gamma)$ is an orthonormal series (OS) estimator of $\Omega(\gamma)$ which we introduce below in (5), and B is the number of orthonormal basis functions used in its estimation. The first factor in (4), $(B - GK + 1)/GKB$, is a small-sample correction which will allow us to use the connection between Hotelling's T^2 distribution and the F distribution together with an appropriate variance estimator. APUY and QTZ suggest kernel type estimators as in Newey and West (1987) and Andrews (1991). APUY show that their test statistics are asymptotically distributed as a χ^2 variate. For better small sample properties, we propose using the following OS estimator:

$$\begin{aligned} \widehat{\Omega}_{NT}(\gamma) &= \frac{1}{B}\sum_{j=1}^{B} \widehat{\Lambda}_j(\gamma)\widehat{\Lambda}'_j(\gamma), \\ \widehat{\Lambda}_j(\gamma) &= \sqrt{\frac{2}{T}}\sum_{t=1}^{T} \left[\bar{Z}_{N,t}(\gamma) - \hat{\theta}_{NT}(\gamma)\right] \cos\left[\pi j \left(\frac{t - 1/2}{T}\right)\right], \end{aligned} \tag{5}$$

with $\bar{Z}_{N,t}(\gamma) = [\bar{Z}'_{1,n_1,t}(\gamma), \ldots, \bar{Z}'_{G,n_G,t}(\gamma)]'$, $\bar{Z}_{g,n_g,t}(\gamma) = n_g^{-1} \sum_{i=1}^{N} Z_{it} \mathbf{1}\{g_i = g\}$.

The general class of OS estimators of a long-run variance (LRV) was first proposed by Phillips (2005). Different OS were then used to construct estimators by Müller (2007), Sun (2011, 2013, 2014), among others. Under the results of Lemma 1 and following Sun (2013), it is easy to show that $\mathcal{N}^{1-\epsilon}\widehat{\Omega}_{NT}(\gamma)\mathcal{N}^{1-\epsilon} - \Omega(\gamma) = o_p(1)$ (see Lemma A.3 in Appendix A). This implies that the variance estimator is consistent if $\epsilon = 1$ and it is proportional to the true value if $\epsilon \in [1/2, 1)$. It follows that $T\hat{\theta}'_{NT}(\gamma)\widehat{\Omega}^{-1}_{NT}(\gamma)\hat{\theta}_{NT}(\gamma) \xrightarrow{d} \mathbb{T}^2_{GK,B}$ for B fixed, where $\mathbb{T}^2_{v_1,v_2}$ denotes the Hotelling's T^2 distribution with v_1 and v_2 being its degrees of freedoms. The connection between Hotelling's T^2 distribution and the F distribution then implies that $W_{NT}(H_{i,t-\tau}, \gamma) \xrightarrow{d} \mathbb{F}_{GK,B-GK+1}$ under the null, where $\mathbb{F}_{v_1,v_2}$ denotes the F distribution with numerator and denominator degrees of freedom of v_1 and v_2, respectively. When $B \to \infty$, a generalization of the usual results of APUY hold such that $T\hat{\theta}'_{NT}(\gamma)\widehat{\Omega}^{-1}_{NT}(\gamma)\hat{\theta}_{NT}(\gamma) \xrightarrow{d} \chi^2_{GK}$. The results of Sun (2013) show that when B is not too large, using the $\mathbb{F}_{GK,B-GK+1}$ critical values instead of (scaled) χ^2_{GK} critical values results in better size properties. With some abuse of notation, let $p_{NT}[w_{NT}(H_{i,t-\tau}, \gamma)] = \mathbb{P}_{\mathcal{H}_0}[\mathbb{F}_{GK,B-GK+1} \geq w_{NT}(H_{i,t-\tau}, \gamma)]$ be the p-value associated with $w_{NT}(H_{i,t-\tau}, \gamma)$. Here, as in the rest of the paper, we do not show the dependency of $p_{NT}[\,\cdot\,]$ to the reference distribution for the sake of simplicity. Moreover, we simply write $\mathbb{P}_{\mathcal{H}_0}[\,\cdot\,]$ to mean the null hypothesis of interest even if different statistics may test different nulls. These will be clear from the context as we establish the asymptotic distribution of each test, and the respective p-values are defined by this asymptotic distribution under the respective null. Using this p-value, a level-α test rejects the null hypothesis if $p_{NT}[w_{NT}(H_{i,t-\tau}, \gamma)] \leq \alpha$ where $\alpha \in (0, 1)$ is the predetermined Type I error rate.

2.3. *kmeans* estimators of cluster membership and centers

If there is no *a priori* information allowing us to specify the cluster membership variables $g_i \in \{1, 2, \ldots, G\}$, testing the C-EPA hypothesis becomes non-trivial. However, under suitable assumptions, it is possible to estimate them from data. Namely, it is required that $\beta_i^0 = \mathrm{E}(Z_{it})$ is homogeneous within clusters but heterogeneous between them which we formalize in the following assumption.

Assumption 4. *$\beta_i^0 = \sum_{g=1}^{G^0} \theta_g^0(\gamma^0)\mathbf{1}\{g_i^0 = g\}$ where $\theta_g^0(\gamma^0)$ with $\|\theta_g^0(\gamma^0)\| > 0$ is the true cluster center of the gth cluster, $\gamma^0 = (g_1^0, \ldots, g_N^0)' \in \Gamma_{N,G^0} \subseteq \mathbb{R}^N$, and $\theta^0(\gamma^0) = (\theta_1^{0\prime}(\gamma^0), \ldots, \theta_G^{0\prime}(\gamma^0))' \in \Theta^{G^0K} \subseteq \mathbb{R}^{G^0K}$ be the true cluster centers.*

For a given G, the *kmeans* estimators of the cluster membership and centers are defined as the

solution to the following optimization problem:

$$(\hat{\theta}_{NT}, \hat{\gamma}_{NT}) = \underset{(\theta,\gamma)\in\Theta^{GK}\times\Gamma_{N,G}}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} \|Z_{it} - \theta_{g_i}\|^2 . \tag{6}$$

This optimization problem is usually solved by an iterative algorithm such as Lloyd (1982) or Hartigan (1975). Below, we provide an iterative algorithm which is a generalization of that of Lloyd's. To see how this iteration is implemented, we define the estimator of the cluster membership variables g_i for any given $\theta = (\theta_1', \ldots, \theta_G')'$ and panel unit $i \in \{1, \ldots, N\}$ as

$$\widehat{g}_i(Z, \theta) = \underset{g\in\{1,\ldots,G\}}{\arg\min} \sum_{t=1}^{T} \|Z_{it} - \theta_g\|^2 . \tag{7}$$

Using this estimator, the *kmeans* estimator of the cluster centers θ^0 can then be written as

$$\hat{\theta}_{NT}(\hat{\gamma}_{NT}) = \underset{\theta\in\Theta^{GK}}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} \left\|Z_{it} - \theta_{\widehat{g}_i(Z,\theta)}\right\|^2 , \tag{8}$$

which shows that the *kmeans* estimator of the center of cluster g is the sample mean of that cluster: $\hat{\theta}_{g,\hat{n}_g,T}(\hat{\gamma}_{NT}) = \bar{Z}_{g,\hat{n}_g,T}(\hat{\gamma}_{NT}) = (\hat{n}_g T)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} Z_{it}\mathbf{1}\{\widehat{g}_i = g\}$, where $\hat{n}_g = \sum_{i=1}^{N} \mathbf{1}\{\widehat{g}_i = g\}$. Following these lines, using the $NT \times K$ matrix $Z = (Z_1', \ldots, Z_N')'$ where $Z_i = (Z_{i1}', \ldots, Z_{iT}')'$, the *kmeans* estimates of the cluster membership variables and the cluster centers can be calculated using the following algorithm.

Algorithm 1: Iterative algorithm for panel data

Input: Data matrix Z, number of clusters G.

Output: Cluster membership variables $g_i^{(m)}$, $i = 1, \ldots, N$, calculated at each iteration $m = 1, 2, \ldots$

1. Initialize the $GK \times 1$ vector of cluster centers $\theta^{(0)}$ by randomly sampling G observations from the time averages $\bar{Z}_{i,T} = T^{-1}\sum_{t=1}^{T} Z_{it}$, $i = 1, \ldots, N$, without replacement.
2. Compute $g_i^{(0)}(Z, \theta^{(0)}) = \arg\min_{g\in\{1,\ldots,G\}} \sum_{t=1}^{T} \left\|Z_{it} - \theta_g^{(0)}\right\|^2$ for all $i = 1, \ldots, N$.
3. Initialize $m = 0$.
   a. Update centers: $\theta_g^{(m+1)} = (n_g^{(m)} T)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} Z_{it}\mathbf{1}\{g_i^{(m)} = g\}$, $g = 1, \ldots, G$, where $n_g^{(m)} = \sum_{i=1}^{N} \mathbf{1}\{g_i^{(m)} = g\}$.
   b. Update assignment: $g_i^{(m+1)}(Z, \theta^{(m+1)}) = \arg\min_{g\in\{1,\ldots,G\}} \sum_{t=1}^{T} \left\|Z_{it} - \theta_g^{(m+1)}\right\|^2$, $i = 1, \ldots, N$.
   c. Stop if $g_i^{(m+1)} = g_i^{(m)}$ for all $i = 1, \ldots, N$ and save $M = m + 1$, the number of iterations. Otherwise, Set $m = m + 1$ and go to Step a..

Algorithm 1 is a panel data generalization of Lloyd's *kmeans* algorithm as used by Chen and Witten (2023). A similar iterative algorithm is used in panel data econometrics by Bonhomme and Manresa (2015). The difference lies in the choice of the initial cluster centers $\theta^{(0)}$ given in the first step. While Bonhomme and Manresa (2015) are agnostic about how to choose these values, our conditional testing framework is crucially dependent on the choice. Although other initialization methods can be used, this would potentially change the derivations in Appendix B. To establish the asymptotic properties of the *kmeans* estimator, we make following assumptions.

Assumption 5. *Let $G^0 \geq 2$. Then for all $g, g' \in \{1, \ldots, G^0\}$, $g \neq g'$, there exists $C_{g,g'} > 0$ such that $\|\theta_g^0(\gamma^0) - \theta_{g'}^0(\gamma^0)\|^2 \geq C_{g,g'}$.*

Assumption 6. *There exist constants $a_1 > 0$ and $b_1 > 0$ such that, for all $i = 1, \ldots, N$, V_{it} is α-mixing with mixing coefficients $\alpha[t] \leq e^{-a_1 t^{b_1}}$. Moreover, there exist constants $a_2 > 0$ and $b_2 > 0$ such that $\mathbb{P}(|V_{k,it}| > C) \leq e^{1-(C/a_2)^{b_2}}$ for all k, i, t and $C > 0$.*

Assumption 5 formalizes the situation where $\mathcal{H}_0$ fails because there are clusters in the population which differ in terms of their expectations. It simply states that the true cluster centers are well separated. Although it implies that the C-EPA null hypothesis fails, this cluster separation assumption is not necessary (but sufficient) for our tests to have power. As documented in the next section, even if $G^0 = 1$, that is there is only one cluster in the population, our proposed tests have power if the population overall mean is different from zero.

Assumption 6 places additional constraints on dependence properties and tail probabilities of the process $\{V_{it}\}_t$ over Assumption 1. These conditions are imposed for the consistent estimation of cluster membership and the asymptotic equivalence of the cluster center estimators based on *kmeans* to the one based on true clusters.

We have the following result which summarizes the asymptotic properties of the *kmeans* estimator.

Lemma 2. Under Assumptions 1-5 and if $G = G^0$, as $(T, N) \to \infty$,

(a) $\hat{\theta}_{NT}(\hat{\gamma}_{NT}) = \theta^0(\gamma^0) + o_p(1)$.

If Assumption 6 also holds, for all $\xi > 0$,

(b) $\mathbb{P}(\sup_{i \in \{1,\ldots,N\}} |\widehat{g}_i - g_i^0| > 0) = o(1) + o(NT^{-\xi})$,

(c) $\hat{\theta}_{NT}(\hat{\gamma}_{NT}) = \hat{\theta}_{NT}(\gamma^0) + o_p(T^{-\xi})$,

(d) if also $N/T^{\xi} \to 0$, $\Omega(\gamma^0)^{-1/2}\mathcal{N}^{1-\epsilon}T^{1/2}(\hat{\theta}_{NT}(\hat{\gamma}_{NT}) - \theta^0(\gamma^0)) \xrightarrow{d} \mathbb{N}(0, I_{GK})$.

Based on this result, a naive attempt to test the null hypothesis of C-EPA would be to estimate the unknown clusters using the *kmeans* estimator and then to use these estimates to construct a Wald test statistic. Let $W_{NT}(H_{i,t-\tau}, \hat{\gamma}_{NT})$ be the test statistic of the form (4) calculated using the *kmeans* estimates obtained using the above algorithm. Consider the test which rejects the associated null if $p_{NT}[w(H_{i,t-\tau}, \hat{\gamma}_{NT})] \leq \alpha$ for $\alpha \in (0, 1)$. The problem with this approach is that the clusters are estimated from the data which are then used to test the null hypothesis of C-EPA. It is now well known in the literature that testing the null hypothesis of homogeneity (that is no clusters exist), following a clustering method such as *kmeans* or hierarchical clustering leads to extremely anti-conservative test statistics, as the results of Gao et al. (2024), Patton and Weller (2023), Chen and Witten (2023) show. As explained in Section 3 below, the null hypothesis of these studies is a sub-hypothesis of the null in our paper, hence, the naive tests of EPA suffer from the same problem. We demonstrate the consequences of this naive approach in Appendix C.1.

3. Clustered EPA Tests with Unknown Clusters

In this section, we develop the valid tests of the C-EPA null hypothesis with clusters estimated by *kmeans*. As is shown in the previous section, using the estimated clusters for testing in a naive manner results in over rejection of the null hypothesis. Here, first the selective inference approach will be employed to control for the Type I error rate by conditioning on the estimated clusters. Then a more straightforward sample splitting solution will be considered.

To begin, we first break down the C-EPA hypothesis into its sub-hypotheses of homogeneity and O-EPA. Namely, the implication (3) of the null hypothesis of interest (1) can be written as $\mathcal{H}_0' : \mathcal{H}_0^{homo} \bigcap \mathcal{H}_0^{oepa}$, with

$$\mathcal{H}_0^{homo} : \bigcap_{g \in \{2,\ldots,G\}} \left[\lim_{N\to\infty} \theta^0_{1,n_1}(\gamma) = \lim_{N\to\infty} \theta^0_{g,n_g}(\gamma) \right], \tag{9}$$

being the homogeneity hypothesis and

$$\mathcal{H}_0^{oepa} : \lim_{N\to\infty} \frac{1}{N} \sum_{g=1}^{G} n_g \theta^0_{g,n_g}(\gamma) = 0, \tag{10}$$

the O-EPA hypothesis, as named by APUY. Both $\mathcal{H}_0^{homo}$ and $\mathcal{H}_0^{oepa}$ are of particular empirical relevance. The tests of the unconditional O-EPA hypothesis are studied by APUY under different

assumptions on the dependence structure of the loss differentials. The empirical importance of testing the homogeneity hypothesis $\mathcal{H}_0^{homo}$ goes beyond EPA testing (see, in particular, the applications of Patton and Weller, 2023, and the discussion therein).

In Section 3.1, we first develop a selective inference framework to test the homogeneity of a pair of clusters selected by *kmeans*. Then we propose a p-value combination test of $\mathcal{H}_0^{homo}$. In Section 3.2, an O-EPA test and the main test statistic of $\mathcal{H}_0$ are presented. Section 3.3 contains a split-sample test based on Patton and Weller (2023) as an alternative method. Finally, in Section 3.4 we cover the case of an unknown number of clusters and develop a method for its estimation.

3.1. Testing the null of homogeneity

The homogeneity null $\mathcal{H}_0^{homo}$ is the intersection of $G-1$ pairwise homogeneity hypotheses. For each of these pairwise homogeneity nulls, we define the following test statistic:

$$D_{g,NT}(H_{i,t-\tau}) = \sqrt{T}\left\|\widehat{\sigma}_{1,g,NT}^{-1/2}(\hat{\gamma}_{NT})[\hat{\theta}_{1,\hat{n}_1,T}(\hat{\gamma}_{NT}) - \hat{\theta}_{g,\hat{n}_g,T}(\hat{\gamma}_{NT})]\right\|,$$

where $\widehat{\sigma}_{1,g,NT}^2(\hat{\gamma}_{NT}) = \widehat{\omega}_{1,1,NT}(\hat{\gamma}_{NT}) + \widehat{\omega}_{g,g,NT}(\hat{\gamma}_{NT}) - 2\widehat{\omega}_{1,g,NT}(\hat{\gamma}_{NT})$ with $\widehat{\omega}_{g,g',NT}(\hat{\gamma}_{NT})$ being the $\{g,g'\}$th $K\times K$ block of $\widehat{\Omega}_{NT}(\hat{\gamma}_{NT})$. For notational simplicity, we ignore showing the dependence of the test statistic on $\hat{\gamma}_{NT}$ as it does not risk confusion, i.e. $D_{g,NT}(H_{i,t-\tau}) := D_{g,NT}(H_{i,t-\tau},\hat{\gamma}_{NT})$. It is easily seen that, this test statistic is the square root of a Wald test of the equality of two estimated cluster centers. Hence, under appropriate conditions, $D_{g,NT}(H_{i,t-\tau},\gamma) \xrightarrow{d} \chi_K$ as $T\to\infty$, where χ_K is a random variable distributed as a χ variate with K degrees of freedom. However, as discussed in the previous section, the associated critical values lose their validity when used with estimated clusters. We define the following asymptotic selective Type I error rate which will be the basis for valid C-EPA testing with unknown clusters.

Definition 1. Let $g,g'\in\{1,\dots,G\}$ be two cluster indexes such that $g\neq g'$ and define the following null hypothesis:

$$\mathcal{H}_0^{g,g'} : \lim_{N\to\infty}\theta^0_{g,n_g}(\hat{\gamma}_{NT}) = \lim_{N\to\infty}\theta^0_{g',n_{g'}}(\hat{\gamma}_{NT}).$$

Then a test of $\mathcal{H}_0^{g,g'}$ controls the selective Type I error rate asymptotically as $(T,N)\to\infty$ if, under $\mathcal{H}_0^{g,g'}$,

$$\lim_{(T,N)\to\infty}\mathbb{P}_{\mathcal{H}_0}\left[\text{Reject } \mathcal{H}_0^{g,g'} \text{at level } \alpha \;\middle|\; \bigcap_{i=1}^{N}\{\widehat{g}_i(Z,\theta)=\widehat{g}_i(z,\theta)\}\right] \leq \alpha,\ \forall\alpha\in(0,1), \tag{11}$$

where $\widehat{g}_i(Z,\theta)$, $i=1,\dots,N$ is the output of Algorithm 1 and $\widehat{g}_i(z,\theta)$ is its sample counterpart associ-

ated with the realization z of Z.

The definition states that a valid test of the pairwise equality hypothesis $\mathcal{H}_0^{g,g'}$ is the one that controls the selective Type I error rate given the clusters estimated by the *kmeans* algorithm. More specifically, the conditioning event in (11) implies that $\mathcal{H}_0^{g,g'}$ should be rejected if the probability of obtaining a test statistic as large as the one in hand does not exceed α among all realizations of Z which result in the same clustering as the one obtained using the realization z. As stated by Chen and Witten (2023), characterising this condition is not trivial but we can instead condition on the clusters estimated at all $m = 1, \ldots, M$ steps of the algorithm. Hence, following Gao et al. (2024) and Chen and Witten (2023), we define the asymptotic p-value

$$
\begin{aligned}
&p_{g,\infty}(d_{g,NT}(H_{i,t-\tau})) \\
&\quad = \lim_{(T,N)\to\infty} P_{\mathcal{H}_0}\Bigg[D_{g,NT}(H_{i,t-\tau}) \geq d_{g,NT}(H_{i,t-\tau}) \;\Bigg|\; \bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{ g_i^{(m)}(Z,\theta) = g_i^{(m)}(z,\theta) \right\}, \\
&\qquad \Pi_g Z = \Pi_g z, \operatorname{dir}(\widehat{\sigma}^{1/2}_{1,g,NT}(\hat{\gamma}_{NT}) Z' \hat{\nu}_g) = \operatorname{dir}(\widehat{\sigma}^{1/2}_{1,g,NT}(\hat{\gamma}_{NT}) z' \hat{\nu}_g) \Bigg],
\end{aligned}
\tag{12}
$$

for $g \in \{2, \ldots, G\}$ where $\Pi_g = I_{NT} - \hat{\nu}_g \hat{\nu}_g' / \|\hat{\nu}_g\|^2$, $\hat{\nu}_g = (\hat{\nu}_{g,1}', \ldots, \hat{\nu}_{g,N}')'$, $\hat{\nu}_{g,i} = \iota_T \hat{\delta}_{g,i}$, ι_T being a T-vector of ones, $\hat{\delta}_{g,i} = \mathbf{1}\{\widehat{g}_i = 1\}/\hat{n}_1 - \mathbf{1}\{\widehat{g}_i = g\}/\hat{n}_g$. Notice that $Z'\hat{\nu}_g = \hat{\theta}_{1,\hat{n}_1,T}(\hat{\gamma}_{NT}) - \hat{\theta}_{g,\hat{n}_g,T}(\hat{\gamma}_{NT})$ and $\|\hat{\nu}_g\|^2 = (\hat{n}_1 T)^{-1} + (\hat{n}_g T)^{-1}$.

The first condition in (12) is the most crucial to the selective inference framework. It states that the cluster to which each panel unit i assigned in every iteration m of the *kmeans* algorithm using the realization z, namely $g_i^{(m)}(z)$, corresponds to the cluster obtained using Z, that is $g_i^{(m)}(Z)$. In other words, as required by Definition 1, we focus on the realization of the random matrix Z resulting in the same clustering as the one results from the application of the *kmeans* algorithm applied to the particular realization z in hand. The next two conditions allow us to remove the nuisance parameters $\Pi_g Z$ and $\operatorname{dir}(\widehat{\sigma}^{1/2}_{1,g,NT}(\hat{\gamma}_{NT}) Z' \hat{\nu}_g)$. These are standard conditions in selective inference literature (see Fithian et al., 2014; Gao et al., 2024; Chen and Witten, 2023).

The p-value $p_{g,\infty}(d_{g,NT}(H_{i,t-\tau}))$ is based on the selective inference methodology of Chen and Witten (2023) but it generalizes it in several ways. First of all, here, we have double indexed random variables Z_{it}, $i = 1, \ldots, N$, $t = 1, \ldots, T$. Second, their study does not allow for dependencies between Z_{it} and Z_{js}, for either $i \neq j$ and $t \neq s$, but only across different variables of the same observation, i.e. between $Z_{k,it}$ and $Z_{k',it}$, the kth and the k'th elements of Z_{it}. Whereas, we allow for arbitrary autocorrelation and CD as well as dependencies between different elements of Z_{it}. Third, their method

depends crucially on the normality of the data generating process, whereas we make use of the CLT in Lemma 1 by exploiting the time series dimension of the data. The following lemma shows how to calculate a p-value in observed samples following this definition.

Lemma 3. Let $g \in \{2, \ldots, G\}$ with $G \geq 2$ given, and $B \to \infty$ as $(T, N) \to \infty$ such that $B/T \to 0$. Under Assumptions 1-3 and $\mathcal{H}_0^{1,g}$, a p-value following the asymptotic principle (12) can be calculated using

$$p_{g,NT}(d_{g,NT}(H_{i,t-\tau})) = 1 - F_{\chi_K}\left[d_{g,NT}(H_{i,t-\tau}); \mathcal{S}\right], \tag{13}$$

where $F_{\chi_K}(\,\cdot\,; \mathcal{S})$ denotes the cumulative distribution function of a χ_K random variable truncated to the set $\mathcal{S}$ with

$$\mathcal{S} = \left\{ \phi \in \mathbb{R} : \bigcap_{m=0}^{M} \bigcap_{i=1}^{N} g_i^{(m)}(z(\phi), \theta) = g_i^{(m)}(z, \theta) \right\}, \tag{14}$$

and

$$z(\phi) = z - \frac{\|z'\hat{\nu}_g\|}{\|\hat{\nu}_g\|^2}\hat{\nu}_g[\mathrm{dir}(z'\hat{\nu}_g)]' + \phi\frac{\hat{\nu}_g}{\sqrt{T}\|\hat{\nu}_g\|^2}\frac{\|z'\hat{\nu}_g\|}{\|\widehat{\sigma}_{1,g,NT}^{-1/2}(\hat{\gamma}_{NT})z'\hat{\nu}_g\|}[\mathrm{dir}(z'\hat{\nu}_g)]'. \tag{15}$$

The equation in (15) defines a perturbation $z(\phi)$ of the original data matrix z. Depending on ϕ, $z(\phi)$ is a version of z such that the two clusters g and 1 are either pushed towards each other or pulled further apart. If $\phi = d_{g,NT}(H_{i,t-\tau})$ then $z(\phi) = z$. If $\phi > d_{g,NT}(H_{i,t-\tau})$ then the two clusters are pulled apart. If instead $\phi < d_{g,NT}(H_{i,t-\tau})$, the two clusters are pushed towards each other and in the extreme case of $\phi = 0$, their centers correspond to each other. Hence, the variable ϕ measures the degree of perturbation (see, Figure 2 of Chen and Witten, 2023). We document the steps of the calculation of this p-value in Appendix B through a characterization of the truncation set $\mathcal{S}$. The following result establishes the asymptotic validity of $p_{g,NT}(d_{g,NT}(H_{i,t-\tau}))$ for the pairwise null hypothesis $\mathcal{H}_0^{1,g}$ defined in Definition 1.

Proposition 1. Let $g \in \{2, \ldots, G\}$ with $G \geq 2$ given, and $B \to \infty$ as $(T, N) \to \infty$ such that $B/T \to 0$.

(a) Under Assumptions 1-3, and $\mathcal{H}_0^{1,g}$,

$$\lim_{(T,N)\to\infty} \mathbb{P}[p_{g,NT}(D_{g,NT}(H_{i,t-\tau})) \leq \alpha] = \alpha, \ \forall \alpha \in (0, 1).$$

(b) Suppose now that $G = G^0 \geq 2$ and $N/T^{\xi} \to 0$. Under Assumptions 1-6, and if $\mathcal{H}_0^{1,g}$ fails,

$$\lim_{(T,N)\to\infty} \mathbb{P}[p_{g,NT}(D_{g,NT}(H_{i,t-\tau})) \leq \alpha] = 1, \ \forall \alpha \in (0, 1).$$

Part (a) of the proposition states that the random variable $p_{g,NT}(D_{g,NT}(H_{i,t-\tau}))$ satisfies the definition of a p-variable of Vovk and Wang (2020) asymptotically, under the null of pairwise cluster equality. The p-value $p_{g,NT}(d_{g,NT}(H_{i,t-\tau}))$ is a realization of this p-variable. Following the common practice, hereafter we refer to both of these quantities as p-values. In Part (b), it is shown that $D_{g,NT}(H_{i,t-\tau})$ is consistent whenever $\mathcal{H}_0^{1,g}$ fails. Here, it is required that the number of clusters G is correctly chosen to be equal to G^0. We relax this assumption in Section 3.4 by proposing an information criterion to estimate G^0.

Now, using the fact that each $G-1$ pairwise cluster equality p-values $p_{g,NT}(D_{g,NT}(H_{i,t-\tau}))$ are asymptotically uniform under their corresponding null, we can define a p-value combination test statistic for the homogeneity null (9) as follows:

$$W_{NT}^{homo}(H_{i,t-\tau}) = \frac{1}{G-1}\left(\sum_{g=2}^{G} p_{g,NT}(D_{g,NT}(H_{i,t-\tau}))^{-r}\right)^{1/r}, \; r \in (1,\infty), \tag{16}$$

The test statistic combines $G-1$ p-values following the methodology developed by Vovk and Wang (2020). Another recent application of this methodology to EPA testing, in particular to multivariate predictive ability comparison, is proposed by SU. Alternative p-value combination methods such as Fisher (1932) are not necessarily suitable to the problem in hand because the p-values $p_{g,NT}(D_{g,NT}(H_{i,t-\tau}))$ use overlapping samples through their dependence on Cluster 1 by construction. Proposition 2 of SU shows that a *pseudo* p-value associated with the test statistic W_{NT}^{homo} can be calculated as

$$p_{NT}(w_{NT}^{homo}(H_{i,t-\tau})) = \min\left(\frac{r}{r-1}\frac{1}{w_{NT}^{homo}(H_{i,t-\tau})}, 1\right). \tag{17}$$

As discussed by SU, the variable $p_{NT}(W_{NT}^{homo}(H_{i,t-\tau}))$ can be interpreted as a p-value although it is not distributed uniformly. However, it satisfies the desired properties of a p-value. These properties are summarized in the following theorem.

Theorem 1. Let $G \geq 2$ be given, and $B \to \infty$ as $(T,N) \to \infty$ such that $B/T \to 0$.

(a) Under Assumptions 1-3, and $\mathcal{H}_0^{homo}$,

$$\limsup_{(T,N)\to\infty} p_{NT}(W_{NT}^{homo}(H_{i,t-\tau})) \leq \alpha, \; \forall \alpha \in (0,1).$$

(b) Suppose now that $G = G^0 \geq 2$ and $N/T^{\xi} \to 0$. Under Assumptions 1-6, and if $\mathcal{H}_0^{homo}$ fails,

$$\lim_{(T,N)\to\infty} \mathbb{P}[p_{NT}(W_{NT}^{homo}(H_{i,t-\tau})) \leq \alpha] = 1, \; \forall \alpha \in (0,1).$$

Although non-crucial for the development of our C-EPA test statistic with unknown clusters, the test statistic W_{NT}^{homo} is of particular empirical importance as it is a strong alternative to the split-sample homogeneity test proposed by Patton and Weller (2023). Part (a) of the theorem shows that the test statistic controls for the Type I error rate asymptotically whereas Part (b) shows that it is consistent if at least one of the pairwise equality null hypothesis $\mathcal{H}_0^{1,g}$ fails.

3.2. The overall EPA test and the main result

The second sub-hypothesis of the C-EPA hypothesis (1), namely the O-EPA hypothesis $\mathcal{H}_0^{oepa}$ states that the two forecasts are equally good on average given past information. To test this sub-hypothesis, consider the test statistic

$$W_{NT}^{oepa}(H_{i,t-\tau}) = \frac{B-K+1}{KB} T \bar{Z}'_{o,NT} \widehat{\Omega}_{o,NT}^{-1} \bar{Z}_{o,NT}, \tag{18}$$

where $\bar{Z}_{o,NT} = T^{-1}\sum_{t=1}^{T} \bar{Z}_{N,t}$, $\bar{Z}_{N,t} = N^{-1}\sum_{i=1}^{N} Z_{it}$, and $\widehat{\Omega}_{o,NT}$ is given by

$$\begin{aligned} \widehat{\Omega}_{o,NT} &= \frac{1}{B}\sum_{j=1}^{B} \widehat{\Lambda}_{o,j}\widehat{\Lambda}'_{o,j}, \\ \widehat{\Lambda}_{o,j} &= \sqrt{\frac{2}{T}} \sum_{t=1}^{T} [\bar{Z}_{N,t} - \bar{Z}_{o,NT}] \cos\left[\pi j \left(\frac{t-1/2}{T}\right)\right]. \end{aligned} \tag{19}$$

The asymptotic properties of this test statistic are summarized in the following proposition.

Proposition 2. Suppose that Assumptions 1 and 3 hold with $\gamma = (1, \ldots, 1)$, that is $G = 1$. Then, for B fixed as $(T, N) \to \infty$, the following results hold.

(a) Under $\mathcal{H}_0^{oepa}$, $W_{NT}^{oepa}(H_{i,t-\tau}) \xrightarrow{d} \mathbb{F}_{K,B-K+1}$.

(b) Suppose that $\mathcal{H}_0^{oepa}$ fails. Then, for any $C > 0$, $\mathbb{P}[W_{NT}^{oepa}(H_{i,t-\tau}) > C] \to 1$.

The test rejects the null of O-EPA if $p_{NT}(w_{NT}^{oepa}(H_{i,t-\tau})) = \mathbb{P}_{\mathcal{H}_0}\left[\mathbb{F}_{K,B-K+1} \geq w_{NT}^{oepa}(H_{i,t-\tau})\right] \leq \alpha$ where $\alpha \in (0,1)$ is the predetermined Type I error rate. When $B = T$ and $K = 1$, the test statistic becomes a Wald-type statistic which is robust to arbitrary CD but does not control for autocorrelation. It becomes then a special case of the $S_{NT}^{(3)}$ test of APUY where the bandwidth parameter of the kernel function is chosen to ignore potential autocorrelation.

We now turn to our main test statistic for the C-EPA null $\mathcal{H}_0$. As in the previous section, we propose the following p-value combination statistic which uses the p-values associated with the $G-1$

pairwise homogeneity tests and the O-EPA test:

$$W_{NT}^{SI}(H_{i,t-\tau}) = \frac{1}{G}\left(p_{NT}(W_{NT}^{oepa}(H_{i,t-\tau}))^{-r} + \sum_{g\in\{2,\ldots,G\}} p_{g,NT}(D_{g,NT}(H_{i,t-\tau}))^{-r}\right)^{1/r}, \quad (20)$$

for any $r \in (1,\infty)$. Again, for notational simplicity, we ignore showing the dependence of the test statistic on $\hat{\gamma}_{NT}$, i.e. $W_{NT}^{SI}(H_{i,t-\tau}) := W_{NT}^{SI}(H_{i,t-\tau}, \hat{\gamma}_{NT})$. The associated *pseudo* p-value is given by

$$p_{NT}(w_{NT}^{SI}(H_{i,t-\tau})) = \min\left(\frac{r}{r-1}\frac{1}{w_{NT}^{SI}(H_{i,t-\tau})}, 1\right). \quad (21)$$

As in the case of (17), $p_{NT}(W_{NT}^{SI}(H_{i,t-\tau}))$ is not a p-value because it is not necessarily uniform under the associated null. However, it can be interpreted as one as the following main result of the paper summarizes its desired asymptotic properties.

Theorem 2. Let $G \geq 2$ be given, and $B \to \infty$ as $(T,N) \to \infty$ such that $B/T \to 0$.

(a) Under Assumptions 1-3, and $\mathcal{H}_0$,

$$\limsup_{(T,N)\to\infty} p_{NT}(W_{NT}^{SI}(H_{i,t-\tau})) \leq \alpha, \ \forall \alpha \in (0,1).$$

(b) Suppose now that $G = G^0 \geq 2$ and $N/T^{\xi} \to 0$. Under Assumptions 1-6, and if either $\mathcal{H}_0^{homo}$ or $\mathcal{H}_0^{oepa}$ fails, then,

$$\lim_{(T,N)\to\infty} \mathbb{P}[p_{NT}(W_{NT}^{SI}(H_{i,t-\tau})) \leq \alpha] = 1, \ \forall \alpha \in (0,1).$$

The asymptotic result shows that the proposed selective inference test successfully controls the Type I error rate and it is consistent as its power approaches one when either $\mathcal{H}_0^{homo}$ or $\mathcal{H}_0^{oepa}$ fails. The finite sample properties of the test statistic are investigated in Section 4 where the simulation results confirm these theoretical expectations.

3.3. Split-sample test statistic

In the previous subsection, the selective inference approach was adopted to condition on the estimated cluster memberships. An alternative and more straightforward method is sample splitting in the time dimension. The current section develops a testing procedure similar to the homogeneity tests developed by Patton and Weller (2023).

Let $\mathcal{R}$ and $\mathcal{P}$ be two mutually exclusive but not necessarily exhaustive subsets of $\mathcal{T} = \{1,\ldots,T\}$ given by $\mathcal{R} = \{1,2,\ldots,S-q+1\}$ and $\mathcal{P} = \{S+1, S+2,\ldots,T\}$ where $q \geq 1$. We denote $R :=$

$S-q+1=\text{card}(\mathcal{R})$ and $P=\text{card}(\mathcal{P})$. Let $\hat{\gamma}_{NR}=(\widehat{g}_{1,NR},\ldots,\widehat{g}_{N,NR})'$ be the vector of the estimated cluster membership variables obtained from the *kmeans* estimator given in (6) using the sample of N cross-sectional units and the subsample $\mathcal{R}$. We define $\hat{\theta}_{NP}(\hat{\gamma}_{NR})=[\hat{\theta}'_{1,\hat{n}_1,P}(\hat{\gamma}_{NR}),\ldots,\hat{\theta}'_{G,\hat{n}_G,P}(\hat{\gamma}_{NR})]$, $\hat{\theta}_{g,\hat{n}_g,P}(\hat{\gamma}_{NR})=P^{-1}\sum_{t=S+1}^{T}\bar{Z}_{g,\hat{n}_g,t}(\hat{\gamma}_{NR})$, $\bar{Z}_{g,\hat{n}_g,t}(\hat{\gamma}_{NR})=\hat{n}_g^{-1}\sum_{i=1}^{N}Z_{it}\mathbf{1}\{\widehat{g}_{i,NR}=g\}$. A split-sample test statistic for $\mathcal{H}_0$ is

$$W_{NT}^{SS}(H_{i,t-\tau},\hat{\gamma}_{NR})=\frac{B-GK+1}{GKB}P\hat{\theta}'_{NP}(\hat{\gamma}_{NR})\widehat{\Omega}_{NP}^{-1}(\hat{\gamma}_{NR})\hat{\theta}_{NP}(\hat{\gamma}_{NR}), \tag{22}$$

$$\begin{aligned}\widehat{\Omega}_{NP}(\hat{\gamma}_{NR})&=\frac{1}{B}\sum_{j=1}^{B}\widehat{\Lambda}_j(\hat{\gamma}_{NR})\widehat{\Lambda}'_j(\hat{\gamma}_{NR}),\\ \widehat{\Lambda}_j(\hat{\gamma}_{NR})&=\sqrt{\frac{2}{P}}\sum_{t=S+1}^{T}\left[\bar{Z}_{N,t}(\hat{\gamma}_{NR})-\hat{\theta}_{NP}(\hat{\gamma}_{NR})\right]\cos\left[\pi j\left(\frac{t-1/2}{P}\right)\right].\end{aligned} \tag{23}$$

The asymptotic properties of the split-sample test crucially depends on the following assumption.

Assumption 7. *V_{it} is independent of all measurable-$\mathcal{G}_{t-q}$ random variables for some $q\geq 1$ and for all $t=1,\ldots,T$, $i=1,\ldots,N$.*

According to Assumption 7, time series dependence in the process $\{V_{it}\}_t$ is limited such that V_{it} is independent of V_{js} whenever $|t-s|\geq q$ for all i and j. This assumption is somewhat restrictive as it rules out many mixing processes for $\{V_{it}\}_t$. We can now state the following result which is similar to Theorem 6 of Patton and Weller (2023) with the differences we discuss in the remarks below.

Theorem 3. Suppose that Assumptions 1-3 and 7 hold. Then, for B fixed, $R,P\to\infty$ as $(T,N)\to\infty$, the following results hold.

(a) Under $\mathcal{H}_0$, $W_{NT}^{SS}(H_{i,t-\tau},\hat{\gamma}_{NR})\xrightarrow{d}\mathbb{F}_{GK,B-GK+1}$.

(b) Suppose now that $G=G^0\geq 2$. Under Assumptions 1-5 and 7, and if $\mathcal{H}_0$ fails, then, for any $C>0$, $\mathbb{P}[W_{NT}^{SS}(H_{i,t-\tau},\hat{\gamma}_{NR})>C]\to 1$.

The result above leads us to the following remarks. First, the split-sample test statistics rely on the selection of the two sub-samples $\mathcal{R}$ and $\mathcal{P}$ which can be arbitrary in practice. Furthermore, the fact that the inference is based on P observations may cause the associated test statistics to have low power. However, we note that the selective inference approach has extra conditioning due to the nuisance parameters discussed above. Hence, the comparative power of the split-sample statistics is an empirical question which we investigate with simulations. Second, here, we apply a small sample

correction contrary to the asymptotic tests of Patton and Weller (2023). Third, our framework allows for strong CD which is ruled out by the authors. Finally, their testing procedure focuses only on homogeneity of the panel whereas we test if each cluster has zero mean.

3.4. Estimating the number of clusters under the alternative

When the researcher wishes to learn the number of clusters under the alternative from data, the sample in hand can be used to obtain an estimate of it. For this purpose, Patton and Weller (2023) suggest to use a multiple testing procedure based on the Bonferroni correction. An adaptation of their proposal would be calculating the p-value associated to the test statistic (20) or (22) for $G = 2, \ldots, G_{max}$ and applying the usual Bonferroni correction to these p-values. The test rejects $\mathcal{H}_0$ if the Bonferroni p-value does not exceed the predetermined Type I error rate. As an alternative, we propose an information criterion (IC) to estimate the number of clusters. Consider the following IC:

$$IC_{NT}(G) = \log\left[\det\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{V}_{it}(G)\widehat{V}_{it}'(G)\right)\right] + (GK+N)\frac{\varsigma\log(NT)}{NT},$$

where $\widehat{V}_{it}(G) = Z_{it} - \hat{\theta}_{G,\widehat{g}_i}$ with $\hat{\theta}_{G,\widehat{g}_i}$ being the solution to (6) with G clusters, and ς is a tuning constant. In our simulations we found that $\varsigma = 3$ works well, as was also suggested by the results of Lumsdaine et al. (2023). The IC estimate of the number of clusters is given by

$$\widehat{G}_{NT} = \underset{G\in\{2,\ldots,G_{max}\}}{\arg\min}\ IC_{NT}(G). \tag{24}$$

For the split-sample test, this IC can be adapted by using only the $\mathcal{R}$ portion of the data. This IC is denoted as $IC_{NR}(G)$. Penalty functions other than the one used here can also be employed (see, for instance Bai and Ng, 2002, for different penalties for determination of the number of factors in factor models). Our IC is an adaptation of the one used by Lumsdaine et al. (2023) to our multivariate framework. It is easy to see that the $\widehat{G}_{NT}$ is consistent for $G^0 \geq 2$ under Assumptions 1-5 if N and T diverge at the same rate. We explore the finite sample properties of this estimator in Section 4 with numerical experiments.

The main advantage of using an information criterion instead of a Bonferroni p-value is its computational efficiency. Although the extra computational burden is negligible in the case of split-sample test statistics, it is quite important for the selective inference tests. This is because the computation of the conditioning set $\mathcal{S}$ is time consuming, and contrary to the Bonferroni p-value, an information criterion requires only the *kmeans* estimates for different values of G and not $\mathcal{S}$.

4. Monte Carlo Study

4.1. Design

To investigate the small sample properties of the proposed test statistics, we generate samples from the following models under different parameter constraints:

$$\text{DGP1: } \Delta L_{it} = d + \mu_i + \lambda_{1i} F_{1,t-1} + \lambda_{2i} F_{2,t-1} + U_{1,it}$$

$$\text{DGP2: } \Delta L_{it} = d + \mu_i + \lambda_{1i} F_{1,t-1} + U_{1,it}$$

$$\text{DGP3: } \Delta L_{it} = \mu_i + \lambda_{1i} F_{1,t-1} + \lambda_{2i} F_{2,t-1} + U_{1,it}$$

$$\text{DGP4: } \Delta L_{it} = d + \mu_i + \lambda_{1i} F_{1,t-1} + U_{2,it}$$

$$\text{DGP5: } \Delta L_{it} = d + \mu_i + \lambda_{1i} F_{1,t-1} + \lambda_{2i} F_{2,t-1} + U_{2,it}$$

where $F_{1t}, F_{2t} \sim iidN(0,1)$. For all experiments, the loadings of F_{2t} are drawn once as $\lambda_{2i} \sim iidN(0, 0.2)$ for all i so they are fixed. We also set their variance to be exactly 0.2 for each experiment. The error terms are generated as $U_{1,it} \sim iidN(0,1)$ and $U_{2,it} = U_{1,it} + 0.5U_{1,i,t-1}$. The parameter d controls the distance between the clusters. We consider values $d \in \{0, 0.125, 0.25, 0.375, 0.5\}$. The number of true clusters is $G = 1$ for size experiments and $G = 2$ otherwise. When $G = 2$, we have $g_i = 1$ for $i = 1, \ldots, N/2$ and $g_i = 2$ for $i = N/2 + 1, \ldots, N$. The number of replications is 2000. For the main experiment we have $N \in \{50, 100\}$ and $T \in \{20, 50, 100, 200\}$. In a relatively limited set of robustness checks, we set $N = 50$, $T \in \{50, 100\}$ and $d = 0.25$. We evaluate the performance of the tests by setting either $H_{i,t-1} = 1$ or $H_{i,t-1} = F_{1,t-1}$. These correspond to unconditional and conditional C-EPA tests, respectively. The parameter constraints particular to these cases are reported in Table 1.

Table 1: Constraints in Monte Carlo Design

Test type	Size ($d = 0$)	Power ($d \neq 0$)
Unc. tests ($H_{i,t-1} = 1$)	$\mu_i = 0$ $\lambda_{1i} = 0$	$\mu_i = -d$ for $i = 1, \ldots, N/2$, $\mu_i = d$ otherwise $\lambda_{1i} = 0$
Con. tests ($H_{i,t-1} = F_{1,t-1}$)	$\mu_i = 0$ $\lambda_{1i} = 0$	$\mu_i = -d$ for $i = 1, \ldots, N/2$, $\mu_i = d$ otherwise $\lambda_{1i} = -d$ for $i = 1, \ldots, N/2$, $\lambda_{1i} = d$ otherwise

DGP1 is considered the main experiment of the Monte Carlo study and the others to be robustness checks, as dictated by the lengthy computations. DGP1 corresponds to the case where we have

cross-sectionally dependent but serially uncorrelated loss differentials. DGP2 imposes cross-sectional independence by setting $\lambda_{2i} = 0$ for all i. As seen in Table 1 $\lambda_{1i} \neq 0$ only for conditional tests but they condition on $F_{1,t-1}$ hence, the independence holds conditionally. DGP3 is a particular case of DGP1 where the O-EPA hypothesis is satisfied even when the C-EPA hypothesis fails. When $d = 0$, DGP3 is identical to DGP1 so we omit the associated results in Table 3. DGP4 and DGP5 are used to check the effect of autocorrelation in loss differentials.

Our main interest is in the test statistics W_{NT}^{SI}, W_{NT}^{SS} computed with $G = 2$ as well as the same statistics computed using the IC-based estimate, $\widehat{G}$, of the number of clusters. For the latter we use $G_{max} = 5$. As a benchmark, we also report the unfeasible test W_{NT} which is computed using $g_i = 1$ for $i = 1, \ldots, N/2$ and $g_i = 2$ for $i = N/2+1, \ldots, N$ so it is based on the true clusters under the alternative. We omit the results of the naive tests as they are clearly invalid in the light of Figure 2. Split-sample tests are computed by splitting the sample into two equally sized parts as $\mathcal{R} = \{1, 2, \ldots, T/2 - q + 1\}$ and $\mathcal{P} = \{T/2 + 1, T/2 + 2, \ldots, T\}$. In DGP1, DGP2 and DGP3, we set $B = T$, $q = 1$. In the other DGPs with autocorrelation, we set $q = 2$, and $B = 2 \max(\lfloor (GK + 4)/2 \rfloor, 1)$ which provides tests with the best size properties in the simulation study of Sun (2013).

4.2. Results

Main Results. The main results on the empirical size of the test statistics are reported in Table 2. The first observation which we draw is that all tests have excellent size properties in small samples under DGP1. In particular, all unconditional tests are correctly sized with the exception of very small size distortions for the split-sample tests when T is small. We note that the Monte Carlo results of APUY as well as Qu et al. (2024) show that the Wald tests of the C-EPA display important size distortions. Our results on the comparable test statistic with known clusters show that the small sample correction using the OS LRV estimator together with the p-values calculated from the F distribution works perfectly. Furthermore, the uncertainty due to the estimation of the clusters does not change the conclusions.

The findings are only marginally different for the conditional tests. One particular observation here is that, in certain cases, W_{NT}^{SI} and W_{NT}^{SI} [IC] tests are slightly undersized whereas W_{NT}^{SS} and W_{NT}^{SS} [IC] are oversized as before. The size of both classes of tests reaches to the nominal value with increases in T. We conclude by noting that these small deviations may be due to the limited number of replications dictated by the computational capacity. Further research may shed light on this issue.

Table 2: Empirical Size of the Test Statistics–Main Results

N	T	W_{NT} [Known]	W_{NT}^{SI}	W_{NT}^{SI} [IC]	W_{NT}^{SS}	W_{NT}^{SS} [IC]
Unconditional Tests ($H_{i,t-1} = 1$)						
50	20	0.06	0.06	0.07	0.07	0.07
50	50	0.05	0.06	0.06	0.05	0.05
50	100	0.05	0.04	0.05	0.06	0.06
50	200	0.05	0.04	0.05	0.05	0.05
100	20	0.07	0.06	0.05	0.07	0.07
100	50	0.05	0.04	0.05	0.05	0.05
100	100	0.06	0.06	0.05	0.05	0.05
100	200	0.05	0.04	0.04	0.05	0.05
Conditional Tests ($H_{i,t-1} = F_{1,t-1}$)						
50	20	0.05	0.05	0.05	0.06	0.06
50	50	0.05	0.04	0.04	0.05	0.05
50	100	0.05	0.04	0.04	0.05	0.05
50	200	0.05	0.04	0.05	0.05	0.05
100	20	0.06	0.04	0.04	0.06	0.07
100	50	0.05	0.04	0.05	0.05	0.05
100	100	0.06	0.05	0.04	0.05	0.05
100	200	0.05	0.04	0.04	0.05	0.05

The power properties of the test statistics are reported in Figure 1. The overall conclusion is that all test statistics are consistent. The first and the second columns of the figure let us analyze the power loss due to the estimation of the number of clusters. We can conclude that this effect of estimation uncertainty is negligible. One particularly important comparison is the power of the conditional selective inference and conditional split-sample tests. We see that the selective inference tests are more powerful than their split-sample counterparts for small T.

Alternative DGPs and additional results. The results on the empirical power and size of the test statistics for DGP2 to DGP5 are reported in Table 3. First, let us focus on the size and power of the unconditional tests for different DPGs. In DGP2, which differs from the main experiment in that it does not contain CD, we show that our tests display only minor size distortions. They reach the power of 100% in the sample sizes and the deviations from the null that we consider.

We remind that DPG3 is identical to DGP1 when $d = 0$. So we discuss only its power properties. A particularly important finding arises in this case as for both unconditional and conditional tests the power of W_{NT}^{SI} and W_{NT}^{SI} [IC] is low when the O-EPA fails, although it increases with the sample size. In this DGP, when the O-EPA hypothesis fails, split-sample tests seem to be the reliable choice.

With DGP4 and DGP5, we analyze the effect of autocorrelation in the loss differentials on the small sample properties of the test statistics. In this case, both conditional and unconditional W_{NT}^{SI} and W_{NT}^{SI} [IC] tests are slightly oversized whereas the W_{NT}^{SS} and W_{NT}^{SS} [IC] show no particular size distortions. Both sets of tests have their power increased with sample size, as expected. We see in particular that, in the case of autocorrelation, selective inference based tests have superior power to the split-sample tests, although this superiority might be due to the slight size distortions of W_{NT}^{SI} and W_{NT}^{SI} [IC] in this particular case.

Some supporting Monte Carlo evidence is reported in Tables 5-7. Measures on the performance of the *kmeans* estimator and the proposed IC are presented in Table 5 where we set $N = 50$ and $T \in \{20, 50, 100, 200\}$, as above. In particular, for each DGP and $d \neq 0$, we report the recovery probability (abbreviated as Rec.) which is the percentage of Monte Carlo replications which result in the true clusters, and the Rand Index (abbreviated as Rand. Ind.) which is a measure of the similarity between the estimated clusters and the true clusters. This index varies between 0 and 1 with the latter signifying perfect recovery of the true clusters. Finally, this table also reports the average of the estimated number of clusters by the IC over 2000 replications. The results can be summarized as follows. First, as expected the performance of the *kmeans* estimator increases with T

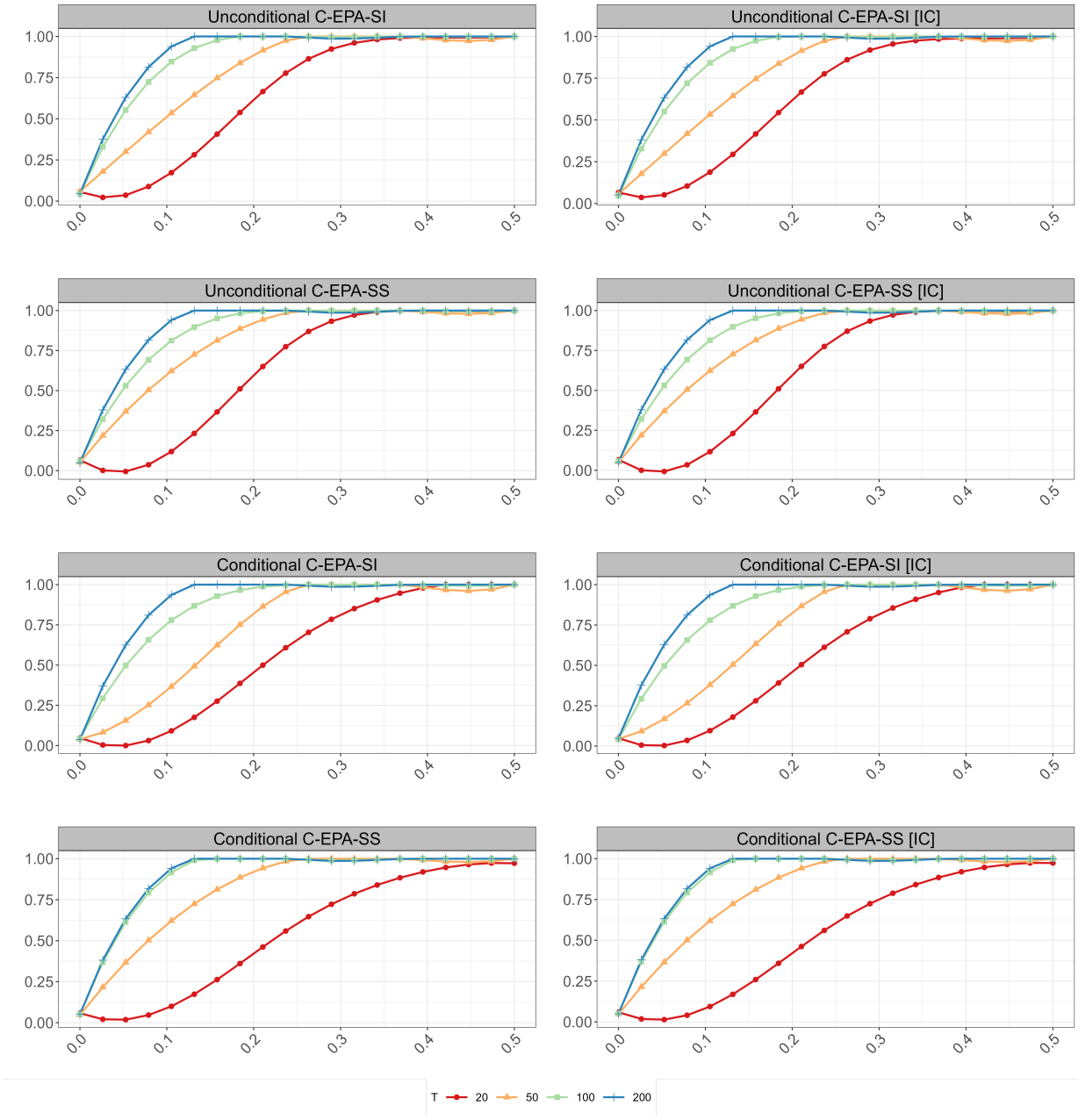

Figure 1: Empirical Power of the Test Statistics–DGP1

Table 3: Empirical Size and Power of the Test Statistics–Alternative DGPs

T	d	DGP	W_{NT} [Known]	W_{NT}^{SI}	W_{NT}^{SI} [IC]	W_{NT}^{SS}	W_{NT}^{SS} [IC]
Unconditional Tests ($H_{i,t-1} = 1$)							
50	0	2	0.05	0.06	0.06	0.07	0.07
100	0	2	0.05	0.05	0.05	0.05	0.05
50	0	4	0.06	0.10	0.09	0.05	0.05
100	0	4	0.04	0.09	0.09	0.05	0.05
50	0	5	0.05	0.09	0.08	0.05	0.05
100	0	5	0.05	0.09	0.09	0.06	0.06
50	0.25	2	1.00	1.00	1.00	1.00	1.00
100	0.25	2	1.00	1.00	1.00	1.00	1.00
50	0.25	3	1.00	0.54	0.51	1.00	1.00
100	0.25	3	1.00	0.87	0.83	1.00	1.00
50	0.25	4	1.00	1.00	1.00	0.99	0.96
100	0.25	4	1.00	1.00	1.00	1.00	1.00
50	0.25	5	1.00	0.80	0.72	0.80	0.71
100	0.25	5	1.00	0.98	0.96	0.99	0.95
Conditional Tests ($H_{i,t-1} = F_{1,t-1}$)							
50	0	2	0.04	0.04	0.05	0.05	0.05
100	0	2	0.05	0.05	0.04	0.05	0.06
50	0	4	0.05	0.06	0.06	0.04	0.04
100	0	4	0.05	0.07	0.06	0.05	0.05
50	0	5	0.05	0.06	0.06	0.05	0.05
100	0	5	0.04	0.06	0.05	0.05	0.05
50	0.25	2	1.00	1.00	1.00	1.00	1.00
100	0.25	2	1.00	1.00	1.00	1.00	1.00
50	0.25	3	1.00	0.32	0.31	0.96	0.96
100	0.25	3	1.00	0.83	0.83	1.00	1.00
50	0.25	4	1.00	1.00	1.00	0.97	0.97
100	0.25	4	1.00	1.00	1.00	1.00	1.00
50	0.25	5	0.96	0.64	0.64	0.68	0.68
100	0.25	5	1.00	0.96	0.97	0.96	0.96

as well as the distance between the true clusters measured by d. Second, especially in the alternative DGPs that we consider, both the *kmeans* estimator and the IC perform better in the conditional setting. The low performance is most apparent in DGP4 with $T = 50$ in unconditional case where the true clusters are never correctly recovered by *kmeans* and the IC overestimates the true number of clusters with average estimated number being 2.86. However, in general, the estimators have very good small sample properties.

Finally, Tables 6 and 7 focus on the performance of the components of the selective inference tests, i.e. W_{NT}^{oepa} and W_{NT}^{homo}. Table 6 reports the performance of these tests under DGP1. First, we see that both tests have excellent size control even in smallest samples. The biggest size distortion is observed when the number of clusters is estimated for unconditional homogeneity testing W_{NT}^{homo} which results in a size of 9% when $N = 50$ and $T = 20$. However, this seems to be an exceptional case as no other setting results in a size superior to 6%. Second, we see that the power of the O-EPA test approaches 1 very rapidly with d. Even when $N = 50$ and $T = 20$ it has a power of 87% when $d = 0.25$ for the unconditional test. Although it has in general lower power than W_{NT}^{oepa}, this observation applies to W_{NT}^{homo} as well.

Among the results of the alternative DGPs reported in Table 7, two findings stand out. First, in the DGPs with autocorrelation, W_{NT}^{homo} tests display some size distortions which is not the case for W_{NT}^{oepa}. Hence, the size distortions observed in Table 3 are due to the fact that the selective inference component of the W_{NT}^{SI} test uses the χ distribution instead of the F distribution. Second, the low power of W_{NT}^{SI} in DGP3 discussed above is entirely due to the fact that the O-EPA hypothesis holds, as in this case W_{NT}^{homo} still has high power.

5. Empirical Illustration

In this section, we present an empirical illustration of the usage of our test statistics for the comparison of predictive models. For this purpose, we use the data set constructed by SU on a large set of daily exchange rates. The complete data set contains 84 exchange rates with the longest time series spanning between January 4, 2011 and April 1, 2021. These contain 39 currencies against the USD, 23 currencies against the EUR, and 22 currencies against the GBP. The time series for the currencies against the USD are observed for a total number of 2558 days, against the GBP for 2590 days, and against the EUR for 2622 days.

For our illustration, we use the out-of-sample forecasts generated by 4 different time series models.

These are given by

$$\begin{aligned}
\text{RW: } & Y_{it} = Y_{i,t-1} + U_{it} \\
\text{AR1: } & Y_{it} = \beta_{i,1} Y_{i,t-1} + U_{it} \\
\text{AR2: } & Y_{it} = \beta_{i,1} Y_{i,t-1} + \beta_{i,2} Y_{i,t-2} + U_{it} \\
\text{TVP: } & Y_{it} = \psi_{it} Y_{i,t-1} + U_{it}, \quad \psi_{it} = \rho_i \psi_{i,t-1} + \varepsilon_{it}
\end{aligned}$$

where Y_{it} is the first difference of the natural logarithm of exchange rate i at day t. Using the Matlab codes made available by the authors,[4] we generate the one-step ahead forecasts with a rolling estimation window of 750 days. This results in 1806, 1870 and 1838 time series observations for the exchange rates against the USD, EUR and GBP, respectively. Our focus is on the balanced portion of the data set after leaving out the COVID-19 period, i.e. before 1st of January 2020. The final forecast errors data set contains $T = 1517$ observations of each of the $N = 84$ exchange rates.

For each model above, we compute the quadratic loss differentials as $\Delta L_{it} = (\widehat{Y}_{1,it} - Y_{it})^2 - (\widehat{Y}_{2,it} - Y_{it})^2$ where $\widehat{Y}_{2,it}$ denoting the RW forecasts. We compare each of the 3 other models with RW using the test statistics W_{NT} [Known], W_{NT} [Naive], W_{NT}^{SI}, and W_{NT}^{SS} for both the unconditional and the conditional EPA hypotheses. The tests with predetermined clusters, noted W_{NT} [Known], use the base exchange rate, that is $G = 3$ for USD, EUR, and GBP. The conditional tests are evaluated using $H_{i,t-1} = \Delta L_{i,t-1}$. For the *kmeans* estimates, the number of random initial observations and the maximum number of iterations in the *kmeans* algorithm are chosen as 10. We set $G = 3$ for W_{NT} [Naive], and for the tests taking into account the estimation of the clusters, the number is chosen by IC with $G_{max} = 5$. The split-sample statistics are calculated using $\mathcal{R} = \{1, 2, \ldots, T/2 - q + 1\}$ and $\mathcal{P} = \{T/2 + 1, T/2 + 2, \ldots, T\}$ with $q = 2$. All tests are robust to autocorrelation with $B = 2\max(\lfloor (GK + 4)/2 \rfloor, 1)$ as in the Monte Carlo simulations. Following SU, we set $r = 20$ for the calculation of the p-value combination tests.

The results are reported in Table 4. The first observation is that all tests conclude that each of AR1, AR2 and TVP models have significantly different predictive power compared to the RW model at 5% level. In fact, unreported results show that, for any exchange rate series that we observe, the RW model produces a higher root mean squared error. Therefore, we conclude that all alternative models have significantly better forecast performance than the RW.

As expected, in each comparison, naive tests result in very small p-values. In all cases except one,

[4]The data set and the Matlab codes are available in https://doi.org/10.1080/07350015.2022.2067545.

Table 4: p-values of Test Statistics Comparing Exchange Rate Forecasts

Model Compared	W_{NT}	W_{NT} [Naive]	W_{NT}^{SI} [IC]	W_{NT}^{SS} [IC]
Unconditional Tests ($H_{i,t-1} = 1$)				
AR1	0.020	0.012	0.025^a	0.025
AR2	0.020	0.007	0.025^a	0.025
TVP	0.010	0.048	0.045	0.030
Conditional Tests ($H_{i,t-1} = \Delta L_{i,t-1}$)				
AR1	0.015	0.000	0.021^c	0.033^a
AR2	0.014	0.000	0.023^b	0.031^b
TVP	0.006	0.002	0.044	0.029^a

Note: ^{a}, ^{b} and ^{c} signify $\widehat{G}_{NT} = 3$, $\widehat{G}_{NT} = 4$ and $\widehat{G}_{NT} = 5$, respectively. Otherwise IC chooses $\widehat{G}_{NT} = 2$.

namely unconditional TVP comparison, selective inference tests give larger p-values than the naive tests. This is similar for split-sample test statistics. One important observation is on the estimated number of clusters: we see that using the full sample (selective inference tests) and the first half of the sample (split-sample tests) result in different estimations, the exceptions being unconditional TVP comparison ($\widehat{G}_{NT} = 2$), and conditional AR2 comparison ($\widehat{G}_{NT} = 4$). We finally note that, in unreported results of W_{NT}^{oepa} and W_{NT}^{homo} tests, we find that the inferiority of the RW forecasts is due to the overall differences instead of heterogeneity, except in one case for which the homogeneity hypothesis is rejected using the W_{NT}^{homo} test with a p-value of 0.027. This is the conditional AR2 comparison which results in the estimated exchange rate clusters given in Table 8.

6. Conclusion

This paper developed a statistical framework for testing a linear hypothesis on the cluster centers of a panel process after having estimated these clusters using the *kmeans* estimator. This statistical framework was then applied to conditional C-EPA testing in order to compare the forecast performance of agents or predictive models. In particular, we developed two distinct strategies to deal with the problem of what is sometimes called "double dipping" in recent statistical literature. Our proposed method is a conditional testing procedure based on recent developments in the area of selective inference. The main idea behind the methodology is to compute a p-value for the C-EPA hypothesis which can be thought as the percentage of rejections of a true null among all realizations of the panel process which result in the same clustering obtained using *kmeans* with the realization in hand. The

second strategy resulted in a set of more straightforward split-sample tests. The two methodologies were then compared theoretically as well as in Monte Carlo experiments.

Our simulation results show that both testing strategies work very well in small samples. They are correctly sized even in very small samples and they have power against viable alternative hypotheses. In particular, selective inference tests perform very well and together with their theoretical and practical advantages, they stand out as the preferred methodology.

Finally, to illustrate the empirical validity of our tests, we compared several time series models with random walk forecasts in terms of their predictive ability, using a large data set of 84 exchange rates. The results showed that all alternative models are have superior predictive power over the random walk, and there is some evidence on exchange rate clusters for which the predictive ability of alternative models differs significantly.

Appendices

A. Proofs

A.1. Proof of Lemma 1

To prove Part (a), we will show that each $K\times 1$ component of $\hat{\theta}_{NT}(\gamma)$ satisfies $\hat{\theta}_{g,n_g,T}(\gamma)=\theta^0_{g,n_g}(\gamma)+o_p(1)$. By definition, $Z_{it}=\beta^0_i+V_{it}$ and $\mathrm{E}(V_{it})=0$. Since $\hat{\theta}_{g,n_g,T}(\gamma)=(n_gT)^{-1}\sum_{i=1}^N\sum_{t=1}^T Z_{it}\mathbf{1}\{g_i=g\}$, we have

$$\hat{\theta}_{g,n_g,T}(\gamma)-\theta^0_{g,n_g}(\gamma)=\frac{1}{n_gT}\sum_{i=1}^N\sum_{t=1}^T V_{it}\mathbf{1}\{g_i=g\}, \tag{25}$$

which gives $\mathrm{E}(\hat{\theta}_{g,n_g,T}(\gamma)-\theta^0_{g,n_g}(\gamma))=0$. Turning to the variance, we have

$$\begin{aligned}
&\left\|\mathrm{E}[(\hat{\theta}_{g,n_g,T}(\gamma)-\theta^0_{g,n_g}(\gamma))(\hat{\theta}_{g,n_g,T}(\gamma)-\theta^0_{g,n_g}(\gamma))']\right\| \\
&\qquad=\left\|\frac{1}{(n_gT)^2}\sum_{i,j=1}^N\sum_{t,s=1}^T\mathrm{E}(V_{it}V_{js}')\mathbf{1}\{g_i=g\}\mathbf{1}\{g_j=g\}\right\| \\
&\qquad\le\frac{1}{n_g^2T}\sum_{i,j=1}^N\left(\frac{1}{T}\sum_{t,s=1}^T\mathrm{E}\|V_{it}V_{js}'\|\right)\mathbf{1}\{g_i=g\}\mathbf{1}\{g_j=g\} \\
&\qquad\le\frac{1}{n_g^2T}\sum_{i,j=1}^N\left(\frac{1}{T}\sum_{t,s=1}^T\mathrm{E}\|V_{it}V_{js}'\|\right)=O\left(\frac{1}{\kappa_g^2T}\right),
\end{aligned} \tag{26}$$

by Assumptions 1 and 2 which concludes Part (a). For Part (b), we write

$$\Omega(\gamma)^{-1/2}\mathcal{N}^{1-\epsilon}T^{1/2}(\hat{\theta}_{NT}(\gamma)-\theta^0(\gamma))=\Omega(\gamma)^{-1/2}\mathcal{N}^{1-\epsilon}T^{-1/2}\sum_{t=1}^{T}\bar{V}_{N,t}(\gamma),$$

and the result follows from Assumption 3.

A.2. Proof of Lemma 2

Let $\widehat{\mathcal{Q}}(\theta,\gamma)=(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|Z_{it}-\theta_{g_i}\|^2$, be the objective function of the *kmeans* estimator divided by NT, and $\widetilde{\mathcal{Q}}(\theta,\gamma)=N^{-1}\sum_{i=1}^{N}\|\theta^0_{g_i^0}-\theta_{g_i}\|^2+(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|V_{it}\|^2$, the auxiliary objective function. We also define the Hausdorff distance between $\hat{\theta}_{NT}(\gamma)$ and $\theta^0(\gamma)$ as

$$\begin{aligned}d_H(\hat{\theta}_{NT}(\gamma),\theta^0(\gamma))=\max\Bigg[&\max_{g\in\{1,\ldots,G\}}\left(\min_{g'\in\{1,\ldots,G\}}\left\|\hat{\theta}_{g',n_{g'},T}(\gamma)-\theta^0_{g,n_g}(\gamma)\right\|^2\right),\\&\max_{g'\in\{1,\ldots,G\}}\left(\min_{g\in\{1,\ldots,G\}}\left\|\hat{\theta}_{g',n_{g'},T}(\gamma)-\theta^0_{g,n_g}(\gamma)\right\|^2\right)\Bigg].\end{aligned}$$

Our proof is based on the proof of Theorem 1 and Proposition S.4 of Bonhomme and Manresa (2015) but it generalizes their results for the multivariate case with potentially strong CD. Part (a) of Lemma 2 is proved by the following lemma.

Lemma A.1. Under the assumptions of Lemma 2, we have

(a) $\widehat{\mathcal{Q}}(\theta,\gamma)-\widetilde{\mathcal{Q}}(\theta,\gamma)=o_p(1)$,

(b) $\widetilde{\mathcal{Q}}(\hat{\theta}_{NT}(\hat{\gamma}_{NT}),\hat{\gamma}_{NT})-\widetilde{\mathcal{Q}}(\theta^0,\gamma^0)=o_p(1)$.

Proof. To prove (a), we write

$$\begin{aligned}\left|\widehat{\mathcal{Q}}(\theta,\gamma)-\widetilde{\mathcal{Q}}(\theta,\gamma)\right|&=\left|\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}'(\theta^0_{g_i^0}-\theta_{g_i})\right|\\&\leq 2\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\theta^0_{g_i^0}-\theta_{g_i}\right\|\left\|\frac{1}{T}\sum_{t=1}^{T}V_{it}\right\|\right)=o_p(1),\end{aligned}$$

which follows directly from Assumption 1(a)-(c). To show (b), we first note that $\widetilde{\mathcal{Q}}(\theta,\gamma)$ is uniquely

minimized at true values. To see this, it suffices to write

$$\begin{aligned}
\widetilde{\mathcal{Q}}(\theta,\gamma)-\widetilde{\mathcal{Q}}(\theta^0,\gamma^0) &= \frac{1}{N}\sum_{i=1}^{N}\left\|\theta^0_{g_i^0}-\theta_{g_i}\right\|^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{g'=1}^{G}\mathbf{1}\{g_i^0=g\}\mathbf{1}\{g_i=g'\}\left\|\theta^0_g(\gamma^0)-\theta_{g'}(\gamma)\right\|^2 \\
&\geq \sum_{g=1}^{G}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{g_i^0=g\}\min_{g'\in\{1,\ldots,G\}}\left\|\theta^0_g(\gamma^0)-\theta_{g'}(\gamma)\right\|^2 \\
&= \sum_{g=1}^{G}\frac{n_g^0}{N}\min_{g'\in\{1,\ldots,G\}}\left\|\theta^0_g(\gamma^0)-\theta_{g'}(\gamma)\right\|^2,
\end{aligned} \tag{27}$$

where $n_g^0/N \to \kappa_g^0 \in (0,1)$ by Assumption 2. Note that, by definition, *kmeans* estimator satisfies $\widehat{\mathcal{Q}}(\hat{\theta}_{NT}(\hat{\gamma}_{NT}),\hat{\gamma}_{NT}) \leq \widehat{\mathcal{Q}}(\theta,\gamma)$. Combining this with (a), we find $\widetilde{\mathcal{Q}}(\hat{\theta}_{NT}(\hat{\gamma}_{NT}),\hat{\gamma}_{NT})+o_p(1) \leq \widetilde{\mathcal{Q}}(\theta,\gamma)+o_p(1)$. Hence, by (27), we have $\widetilde{\mathcal{Q}}(\hat{\theta}_{NT}(\hat{\gamma}_{NT}),\hat{\gamma}_{NT})-\widetilde{\mathcal{Q}}(\theta^0,\gamma^0)=o_p(1)$ which ends the proof. □

For Part (a), we will show the consistency of the *kmeans* estimator of the cluster centers with respect to the Hausdorff distance, as in Proposition S.4 of Bonhomme and Manresa (2015). Namely, we will show that $d_H(\hat{\theta}_{NT}(\hat{\gamma}_{NT}),\theta^0(\gamma^0))=o_p(1)$. Define the permutation $\upsilon:\{1,\ldots,G\}\to\{1,\ldots,G\}$ as $\upsilon(g)=\arg\min_{g'\in\{1,\ldots,G\}}\|\theta^0_g(\gamma^0)-\hat{\theta}_{g',NT}(\hat{\gamma}_{NT})\|^2$. Following steps similar to those in (27), it is easy to show that $\|\theta^0_g(\gamma^0)-\hat{\theta}_{g',NT}(\hat{\gamma}_{NT})\|^2$ is bounded away from zero. It follows that $\upsilon(g)\neq\upsilon(g')$ for all $g'\neq g$, with probability approaching to one. Thus, for all $g'\in\{1,\ldots,G\}$, $\min_{g'\in\{1,\ldots,G\}}\|\theta^0_g(\gamma^0)-\hat{\theta}_{g',NT}(\hat{\gamma}_{NT})\|^2 \leq \|\theta^0_{\upsilon^{-1}(g')}(\gamma^0)-\hat{\theta}_{g',NT}(\hat{\gamma}_{NT})\|^2 = \min_{\tilde{g}\in\{1,\ldots,G\}}\|\theta^0_{\upsilon^{-1}(g')}(\gamma^0)-\hat{\theta}_{\tilde{g},NT}(\hat{\gamma}_{NT})\|^2 = o_p(1)$ where the last equality follows from (27) and Lemma A.1(b). This in turn implies that

$$\max_{g\in\{1,\ldots,G\}}\left(\min_{g'\in\{1,\ldots,G\}}\|\theta^0_g-\theta_{g'}\|^2\right)=o_p(1).$$

Combining this with the definition of the Hausdorff distance, we find $d_H(\hat{\theta}_{NT}(\hat{\gamma}_{NT}),\theta^0(\gamma^0))=o_p(1)$ which shows that there exists a permutation $\upsilon(g)$ such that $\|\theta^0_{\upsilon(g)}(\gamma^0)-\hat{\theta}_{g,NT}(\hat{\gamma}_{NT})\|^2=o_p(1)$ which ends the proof of Part (a).

For Part (b), we define Θ_η as the set of parameters $\theta\in\Theta^{GK}$ that satisfy $\|\theta-\theta^0(\gamma^0)\|^2<\eta$ for $\eta>0$. We state the following result which is similar to Lemma B.4 of Bonhomme and Manresa (2015).

Lemma A.2. For $\eta>0$ small enough, we have, for all $\xi>0$ and as $(T,N)\to\infty$,

$$\sup_{\theta\in\Theta_\eta}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i(Z,\theta)\neq g_i^0\}=o_p(T^{-\xi}).$$

Proof. As in the proof of Lemma B.4 of Bonhomme and Manresa (2015), we first note that, by

the definition of $\widehat{g}_i(Z,\theta)$ in (7), $\mathbf{1}\{\widehat{g}_i(Z,\theta)=g\} \leq \mathbf{1}\left\{\sum_{t=1}^T \|Z_{it}-\theta_g\|^2 \leq \sum_{t=1}^T \|Z_{it}-\theta_{g_i^0}\|^2\right\}$. Notice also that we can write $N^{-1}\sum_{i=1}^N \mathbf{1}\{\widehat{g}_i(Z,\theta) \neq g_i^0\} = \sum_{g=1}^G N^{-1}\sum_{i=1}^N \mathbf{1}\{g_i^0 \neq g\}\mathbf{1}\{\widehat{g}_i(Z,\theta)=g\}$. Combining these two gives $N^{-1}\sum_{i=1}^N \mathbf{1}\{\widehat{g}_i(Z,\theta) \neq g_i^0\} \leq \sum_{g=1}^G N^{-1}\sum_{i=1}^N Q_{ig}(\theta)$ where $Q_{ig}(\theta) = \mathbf{1}\{g_i^0 \neq g\}\mathbf{1}\left\{\sum_{t=1}^T \|Z_{it}-\theta_g\|^2 \leq \sum_{t=1}^T \|Z_{it}-\theta_{g_i^0}\|^2\right\}$. We will bound $Q_{ig}(\theta)$. By the fact that $Z_{it} = \theta_{g_i^0}^0 + V_{it}$, we have

$$\begin{aligned}
Q_{ig}(\theta) &= \mathbf{1}\{g_i^0 \neq g\}\mathbf{1}\left\{\sum_{t=1}^T\sum_{k=1}^K \left[2V_{k,it}(\theta_{k,g_i^0}-\theta_{k,g}) + (\theta_{k,g_i^0}^0-\theta_{k,g})^2 - (\theta_{k,g_i^0}^0-\theta_{k,g_i^0})^2\right] \leq 0\right\} \\
&\leq \max_{g'\neq g} \mathbf{1}\left\{\sum_{t=1}^T\sum_{k=1}^K \left[2V_{k,it}(\theta_{k,g'}-\theta_{k,g}) + (\theta_{k,g'}^0(\gamma^0)-\theta_{k,g})^2 - (\theta_{k,g'}^0(\gamma^0)-\theta_{k,g'})^2\right] \leq 0\right\}.
\end{aligned}$$

Define

$$\begin{aligned}
A_T = \Bigg|&\sum_{t=1}^T\sum_{k=1}^K \left[2V_{k,it}(\theta_{k,g'}-\theta_{k,g}) + (\theta_{k,g'}^0(\gamma^0)-\theta_{k,g})^2 - (\theta_{k,g'}^0(\gamma^0)-\theta_{k,g'})^2\right] \\
&- \sum_{t=1}^T\sum_{k=1}^K \left[2V_{k,it}(\theta_{k,g'}^0(\gamma^0)-\theta_{k,g}^0(\gamma^0)) + (\theta_{k,g'}^0(\gamma^0)-\theta_{k,g}^0(\gamma^0))^2\right]\Bigg|.
\end{aligned}$$

Rearranging and using the triangular inequality, we find,

$$A_T \leq |A_{1T}| + |A_{2T}| + |A_{3T}| + |A_{4T}|,$$

where

$$\begin{aligned}
A_{1T} &= 2\sum_{t=1}^T\sum_{k=1}^K V_{k,it}(\theta_{k,g'} - \theta_{k,g'}^0(\gamma^0)), \\
A_{2T} &= 2\sum_{t=1}^T\sum_{k=1}^K V_{k,it}(\theta_{k,g}^0(\gamma^0) - \theta_{k,g}), \\
A_{3T} &= T\sum_{k=1}^K (\theta_{k,g'}^0(\gamma^0) - \theta_{k,g'})^2,
\end{aligned}$$

and

$$\begin{aligned}
A_{4T} &= T\sum_{k=1}^K \left[(\theta_{k,g'}^0(\gamma^0)-\theta_{k,g})^2 - (\theta_{k,g'}^0(\gamma^0)-\theta_{k,g}^0(\gamma^0))^2\right] \\
&= T\sum_{k=1}^K \left[\theta_{k,g}^2 - \theta_{k,g}^0(\gamma^0)^2 - 2\theta_{k,g'}^0(\gamma^0)(\theta_{k,g}-\theta_{k,g}^0(\gamma^0))\right] \\
&= T\sum_{k=1}^K \left[\theta_{k,g}^2 - \theta_{k,g}^0(\gamma^0)^2\right] - 2T\sum_{k=1}^K \left[\theta_{k,g'}^0(\gamma^0)(\theta_{k,g}-\theta_{k,g}^0(\gamma^0))\right].
\end{aligned}$$

For $\theta \in \Theta_\eta$, we find that

$$A_T \leq TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} V_{k,it}^2\right)^{1/2} + TC_2\eta + TC_3\sqrt{\eta},$$

with C_1, C_2, C_3 being constants independent of η and T which follows from the definition of Θ_η. We find

$$\begin{aligned} Q_{ig}(\theta) &\leq \max_{g' \neq g} \mathbf{1}\left\{\sum_{t=1}^{T}\sum_{k=1}^{K}\left[2V_{k,it}(\theta_{k,g'}^0(\gamma^0) - \theta_{k,g}^0(\gamma^0)) + (\theta_{k,g'}^0(\gamma^0) - \theta_{k,g}^0(\gamma^0))^2\right]\right. \\ &\qquad \left. \leq TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} V_{k,it}^2\right)^{1/2} + TC_2\eta + TC_3\sqrt{\eta}\right\}. \end{aligned}$$

The right-hand side does not depend on θ, hence, $\sup_{\theta\in\Theta_\eta} Q_{ig}(\theta) \leq \widetilde{Q}_{ig}$ with

$$\begin{aligned} \widetilde{Q}_{ig} &\leq \max_{g' \neq g} \mathbf{1}\left\{\sum_{t=1}^{T}\sum_{k=1}^{K} 2V_{k,it}(\theta_{k,g'}^0(\gamma^0) - \theta_{k,g}^0(\gamma^0))\right. \\ &\qquad \left. \leq -T\sum_{k=1}^{K}(\theta_{k,g'}^0(\gamma^0) - \theta_{k,g}^0(\gamma^0))^2 + TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} V_{k,it}^2\right)^{1/2} + TC_2\eta + TC_3\sqrt{\eta}\right\}. \end{aligned}$$

This gives $\sup_{\theta\in\Theta_\eta} N^{-1}\sum_{i=1}^{N} \mathbf{1}\{\widehat{g}_i(Z,\theta) \neq g_i^0\} \leq N^{-1}\sum_{i=1}^{N}\sum_{g=1}^{G}\widetilde{Q}_{ig}$. Now we have

$$\begin{aligned} \mathbb{P}(\widetilde{Q}_{ig} = 1) &\leq \sum_{g' \neq g} \mathbb{P}\left(\sum_{t=1}^{T}\sum_{k=1}^{K} 2V_{k,it}(\theta_{k,g'}^0(\gamma^0) - \theta_{k,g}^0(\gamma^0))\right. \\ &\qquad \left. \leq -T\sum_{k=1}^{K}(\theta_{k,g'}^0(\gamma^0) - \theta_{k,g}^0(\gamma^0))^2 + TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} V_{k,it}^2\right)^{1/2} + TC_2\eta + TC_3\sqrt{\eta}\right) \\ &\leq \sum_{g' \neq g}\left[\mathbb{P}\left(\sum_{t=1}^{T}\sum_{k=1}^{K} 2V_{k,it}(\theta_{k,g'}^0(\gamma^0) - \theta_{k,g}^0(\gamma^0)) \leq -TC_{g,g'} + TC_1\sqrt{\eta}\sqrt{C} + TC_2\eta + TC_3\sqrt{\eta}\right)\right. \\ &\qquad \left. + \mathbb{P}\left(\sum_{k=1}^{K}(\theta_{k,g'}^0(\gamma^0) - \theta_{k,g}^0(\gamma^0))^2 < C_{g,g'}\right) + \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} V_{k,it}^2 > C\right)\right]. \end{aligned}$$

By Assumption 5, the second term above is null, and by Lemma B.5 of Bonhomme and Manresa (2015) and under Assumption 6, $\mathbb{P}\left(T^{-1}\sum_{t=1}^{T}\sum_{k=1}^{K} V_{k,it}^2 > C\right) = o(T^{-\xi})$, for all $\xi > 0$. Furthermore, by choosing η suitably, we find

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} 2V_{k,it}(\theta_{k,g'}^0(\gamma^0) - \theta_{k,g}^0(\gamma^0)) \leq -C_{g,g'} + C_1\sqrt{\eta}\sqrt{C} + C_2\eta + C_3\sqrt{\eta}\right) \\ &\qquad \leq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} V_{k,it}(\theta_{k,g'}^0(\gamma^0) - \theta_{k,g}^0(\gamma^0)) \leq -\frac{C_{g,g'}}{2}\right) = o(T^{-\xi}) \end{aligned}$$

where we obtain the last equality by applying Lemma B.5 of Bonhomme and Manresa (2015) with $z_t = V_{k,it}(\theta^0_{k,g'}(\gamma^0) - \theta^0_{k,g}(\gamma^0))$ and $z = C_{g,g'}/2$. This in turn implies that $N^{-1}\sum_{i=1}^N \sum_{g=1}^G \mathbb{P}(\widetilde{Q}_{ig} = 1) = o(T^{-\xi})$. Finally we note that, for all $\xi > 0$ and $\tilde{\xi} > 0$,

$$\begin{aligned}\mathbb{P}\left(\sup_{\theta\in\Theta_\eta} \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{\widehat{g}_i(Z,\theta) \neq g_i^0\} > \tilde{\xi}T^{-\xi}\right) &\leq \mathbb{P}\left(N^{-1}\sum_{i=1}^N\sum_{g=1}^G \widetilde{Q}_{ig} > \tilde{\xi}T^{-\xi}\right) \\ &\leq \frac{\mathrm{E}\left(N^{-1}\sum_{i=1}^N\sum_{g=1}^G \widetilde{Q}_{ig}\right)}{\tilde{\xi}T^{-\xi}} = o(1),\end{aligned}$$

which ends the proof. □

We now prove the last three parts of Lemma 2. For Part (b) we refer to the proof of Bonhomme and Manresa (2015) which is identical to our case. Part (c) also follows similar lines to the proof of Theorem 2 and Corollary 1 of Bonhomme and Manresa (2015) with a difference that we have in (28). First, like in the proof of Part (a), let $\mathcal{Q}^*(\theta) = (NT)^{-1}\sum_{i=1}^N\sum_{t=1}^T \|Z_{it} - \theta_{\widehat{g}_i}\|^2$, be the concentrated version of $\widehat{\mathcal{Q}}(\theta,\gamma)$, and $\mathcal{Q}^\dagger(\theta) = (NT)^{-1}\sum_{i=1}^N \|Z_{it} - \theta_{g_i^0}\|^2$. By choosing η small enough, Lemma A.2 leads to $\sup_{\theta\in\Theta_\eta}|\mathcal{Q}^*(\theta) - \mathcal{Q}^\dagger(\theta)| = o_p(T^{-\xi})$ for all $\xi > 0$. Furthermore, by consistency of $\hat{\theta}_{NT}(\hat{\gamma}_{NT})$ and $\hat{\theta}_{NT}(\gamma^0)$, as $(T,N)\to\infty$, $\mathcal{Q}^*(\hat{\theta}_{NT}(\hat{\gamma}_{NT})) - \mathcal{Q}^\dagger(\hat{\theta}_{NT}(\hat{\gamma}_{NT})) = o_p(T^{-\xi})$ and $\mathcal{Q}^*(\hat{\theta}_{NT}(\gamma^0)) - \mathcal{Q}^\dagger(\hat{\theta}_{NT}(\gamma^0)) = o_p(T^{-\xi})$ which in turn gives $\mathcal{Q}^\dagger(\hat{\theta}_{NT}(\hat{\gamma}_{NT})) - \mathcal{Q}^\dagger(\hat{\theta}_{NT}(\gamma^0)) = o_p(T^{-\xi})$. Now, as in (27),

$$\begin{aligned}\mathcal{Q}^\dagger(\hat{\theta}_{NT}(\hat{\gamma}_{NT})) - \mathcal{Q}^\dagger(\hat{\theta}_{NT}(\gamma^0)) &= \frac{1}{N}\sum_{i=1}^N \left\|\hat{\theta}_{\widehat{g}_i} - \hat{\theta}_{g_i}\right\|^2 \\ &= \frac{1}{N}\sum_{i=1}^N\sum_{g=1}^G\sum_{g'=1}^G \mathbf{1}\{\widehat{g}_i = g\}\mathbf{1}\{g_i = g'\}\left\|\hat{\theta}_{\widehat{g}_i} - \hat{\theta}_{g_i}\right\|^2 \\ &\geq \sum_{g=1}^G \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{\widehat{g}_i = g\}\mathbf{1}\{g_i = g'\} \min_{g'\in\{1,\ldots,G\}}\left\|\hat{\theta}_{\widehat{g}_i} - \hat{\theta}_{g_i}\right\|^2 \\ &= \sum_{g=1}^G \frac{\hat{n}_g}{N}\min_{g'\in\{1,\ldots,G\}}\left\|\hat{\theta}_{\widehat{g}_i} - \hat{\theta}_{g_i}\right\|^2,\end{aligned} \tag{28}$$

where $\hat{n}_g/N \to \kappa_g^0 \in (0,1)$ by Assumption 2. We thus obtain $\hat{\theta}_{\widehat{g}_i} - \hat{\theta}_{g_i} = o_p(T^{-\xi})$ which ends the proof of Part (c). Part (d) then follows from the consistency of the estimator and Assumption 3.

A.3. Proof of Lemma 3

Let $\gamma \in \Gamma_{N,G}$, $G \geq 2$, and ν_g the associated $NT \times 1$ vector. For convenience, we remind that, $\nu_g = (\nu'_{g,1},\ldots,\nu'_{g,N})'$, $\nu_{g,i} = \iota_T\delta_{g,i}$, ι_T being a T-vector of ones, $\delta_{g,i} = \mathbf{1}\{g_i = 1\}/n_1 - \mathbf{1}\{g_i = g\}/n_g$.

As in the main text, we also have $\Pi_g = I_{NT} - \nu_g \nu_g'/\|\nu_g\|^2$. Following lemmas will be referred to in the proof of our result.

Lemma A.3. Suppose that Assumption 1 holds. Then as $B \to \infty$ as $(T, N) \to \infty$ such that $B/T \to 0$, $\mathcal{N}^{1-\epsilon}\widehat{\Omega}_{NT}(\gamma)\mathcal{N}^{1-\epsilon} - \Omega(\gamma) = o_p(1)$.

Proof. Let $\Omega_{NT}(\gamma) = \mathrm{V}[T^{-1/2}\sum_{t=1}^{T} \bar{V}_{N,t}(\gamma)]$. We have, $\Omega_{NT}(\gamma) = T^{-1}\sum_{t,s=1}^{T} \mathrm{E}[\bar{V}_{N,t}(\gamma)\bar{V}_{N,s}'(\gamma)]$. Under the assumptions and standard arguments $\mathcal{N}^{1-\epsilon}\Omega_{NT}(\gamma)\mathcal{N}^{1-\epsilon} - \Omega(\gamma) = o_p(1)$ as $(T, N) \to \infty$. Following Sun (2013), we have $\widehat{\Omega}_{NT}(\gamma) - \Omega_{NT}(\gamma) = o_p(1)$ from which the desired result follows. □

Lemma A.4. Suppose that Assumption 1, and $\mathcal{H}_0^{1,g} : \lim_{N\to\infty} \theta_{1,n_1}^0(\gamma) = \lim_{N\to\infty} \theta_g^0(\gamma)$ hold. Then, $D_{g,NT}(H_{i,t-\tau}, \gamma) \xrightarrow{d} \chi_K$ for all $g \in \{2, \ldots, G\}$, as $B \to \infty$, $(T, N) \to \infty$ such that $B/T \to 0$.

Proof. We write

$$D_{g,NT}^2(H_{i,t-\tau}) = T[\hat{\theta}_{1,n_1,T}(\gamma) - \hat{\theta}_{g,n_g,T}(\gamma)]'\widehat{\sigma}_{1,g,NT}^{-1}(\gamma)[\hat{\theta}_{1,n_1,T}(\gamma) - \hat{\theta}_{g,n_g,T}(\gamma)],$$

which is a standard Wald-type statistic for the test of the difference of the two cluster centers. Therefore, for the result to hold, it suffices to show that $D_{g,NT}^2(H_{i,t-\tau}) \xrightarrow{d} \chi_K^2$ under the assumptions. Let $R_{1,g}$ be the $2K \times GK$ selection matrix with its first and gth columns being $(\iota_K', 0_K')'$ and $(0_K', -\iota_K')'$, respectively, and zeros otherwise, where 0_K is a K-vector of zeros. Using Lemma 1, we find $\Sigma(\gamma)^{-1/2}T^{1/2}R_{1,g}\mathcal{N}^{1-\epsilon}[\hat{\theta}_{NT}(\gamma) - \theta^0(\gamma)] \xrightarrow{d} \mathbb{N}(0, I_K)$ where $\Sigma(\gamma) = R_{1,g}\Omega(\gamma)R_{1,g}'$. Under $\mathcal{H}_0^{1,g}$, this in turn gives that

$$T[n_1^{1-\epsilon}\hat{\theta}_{1,n_1,T}(\gamma) - n_g^{1-\epsilon}\hat{\theta}_{g,n_g,T}(\gamma)]'\sigma_{1,g}^{-1}(\gamma)[n_1^{1-\epsilon}\hat{\theta}_{1,n_1,T}(\gamma) - n_g^{1-\epsilon}\hat{\theta}_{g,n_g,T}(\gamma)] \xrightarrow{d} \chi_K^2,$$

where $\sigma_{1,g}^2(\gamma) = \omega_{1,1}(\gamma) + \omega_{g,g}(\gamma) - 2\omega_{1,g}(\gamma)$ with $\omega_{g,g'}(\gamma)$ begin the $\{g, g'\}$th $K \times K$ block of $\Omega(\gamma)$. But by Lemma A.3, we have $(n_g n_{g'})^{1-\epsilon}\widehat{\omega}_{g,g',NT}(\gamma) - \omega_{g,g'}(\gamma) = o_p(1)$ from which the result follows. □

Lemma A.5. Suppose that Assumptions 1-3, and $\mathcal{H}_0^{1,g} : \lim_{N\to\infty} \theta_{1,n_1}^0(\gamma) = \lim_{N\to\infty} \theta_g^0(\gamma)$ hold. Then, as $(T, N) \to \infty$, $\Pi_g Z$, $D_{g,NT}(H_{t-\tau}, \gamma)$ and $\mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g)$ are asymptotically pairwise independent.

Proof. Notice first that we can write $D_{g,NT}(H_{t-\tau}, \gamma) = \|\sqrt{T}\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g\|$ and under Assumptions 1-3 $\sqrt{T}\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g \xrightarrow{d} \mathbb{N}(0, \|\nu_g\|^2 I_K)$ if $\mathcal{H}_0^{1,g}$ holds, by Lemma 1. It follows that $D_{g,NT}(H_{t-\tau}, \gamma)$ is asymptotically independent of $\mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g)$ as the length and the direction of a standard normal random vector are independent of each other.

To show that $D_{g,NT}(H_{t-\tau},\gamma)$ is asymptotically independent of $\Pi_g Z$, we first note that $\Pi_g \nu_g = 0$. This implies by the properties of the matrix normal distribution that $Z'\nu_g$ is independent of $\Pi_g Z$ from which the desired result follow immediately. □

Our proof of Lemma 3 follows lines similar to the proof of Theorem 1 of Gao et al. (2024) and Proposition 1 of Chen and Witten (2023). We first write

$$\begin{aligned} Z &= \Pi_g Z + (I_{NT} - \Pi_g)Z = \Pi_g Z + \frac{\nu_g \nu_g' Z \widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma)}{\|\nu_g\|^2} \\ &= \Pi_g Z + \frac{\|\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g\|}{\|\nu_g\|^2}\nu_g[\mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g)]'\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma) \\ &= \Pi_g Z + D_{g,NT}(H_{t-\tau},\gamma)\frac{\nu_g}{\sqrt{T}\|\nu_g\|^2}[\mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g)]'\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma). \end{aligned} \tag{29}$$

By replacing this equation into (12), we find

$$\begin{aligned} &p_{g,\infty}(d_{g,NT}(H_{i,t-\tau},\gamma)) \\ &\quad= \lim_{(T,N)\to\infty} \mathbb{P}_{\mathcal{H}_0}\Bigg[D_{g,NT}(H_{i,t-\tau},\gamma) \geq d_{g,NT}(H_{i,t-\tau},\gamma) \,\Bigg| \\ &\quad\bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{g_i^{(m)}\left(\Pi_g Z + D_{g,NT}(H_{t-\tau},\gamma)\frac{\nu_g}{\sqrt{T}\|\nu_g\|^2}[\mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g)]'\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma)\right) = g_i^{(m)}(z)\right\}, \\ &\quad\Pi_g Z = \Pi_g z, \mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma)Z'\nu_g) = \mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma)z'\nu_g)\Bigg] \\ &\quad= \lim_{(T,N)\to\infty} \mathbb{P}_{\mathcal{H}_0}\Bigg[D_{g,NT}(H_{i,t-\tau},\gamma) \geq d_{g,NT}(H_{i,t-\tau},\gamma) \,\Bigg| \\ &\quad\bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{g_i^{(m)}\left(\Pi_g z + D_{g,NT}(H_{t-\tau},\gamma)\frac{\nu_g}{\sqrt{T}\|\nu_g\|^2}[\mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)z'\nu_g)]'\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma)\right) = g_i^{(m)}(z)\right\}, \\ &\quad\Pi_g Z = \Pi_g z, \mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma)Z'\nu_g) = \mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma)z'\nu_g)\Bigg], \end{aligned}$$

where we used the two conditions $\Pi_g Z = \Pi_g z$ and $\mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g) = \mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)z'\nu_g)$ to obtain the second equality. By Lemma A.5, this implies

$$\begin{aligned} &p_{g,\infty}(d_{g,NT}(H_{i,t-\tau},\gamma)) \\ &\quad= \lim_{(T,N)\to\infty} \mathbb{P}_{H_0}\Bigg[D_{g,NT}(H_{i,t-\tau},\gamma) \geq d_{g,NT}(H_{i,t-\tau},\gamma) \,\Bigg| \\ &\quad\bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{g_i^{(m)}\left(\Pi_g z + D_{g,NT}(H_{t-\tau},\gamma)\frac{\nu_g}{\sqrt{T}\|\nu_g\|^2}[\mathrm{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)z'\nu_g)]'\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma)\right) = g_i^{(m)}(z)\right\}\Bigg]. \end{aligned}$$

Next, by plugging in the definition of Π_g into the first term of (29), we have,

$$\begin{aligned} z(\phi) &:= z - \frac{\|z'\nu_g\|}{\|\nu_g\|^2}\nu_g[\text{dir}(z'\nu_g)]' + D_{g,NT}(H_{i,t-\tau},\gamma)\frac{\nu_g}{\sqrt{T}\|\nu_g\|^2}[\text{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)z'\nu_g)]'\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma) \\ &= z - \frac{\|z'\nu_g\|}{\|\nu_g\|^2}\nu_g[\text{dir}(z'\nu_g)]' + D_{g,NT}(H_{i,t-\tau},\gamma)\frac{\nu_g}{\sqrt{T}\|\nu_g\|^2}\frac{\|z'\nu_g\|}{\|\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)z'\nu_g\|}[\text{dir}(z'\nu_g)]' \\ &= z - \frac{\|z'\nu_g\|}{\|\nu_g\|^2}\nu_g[\text{dir}(z'\nu_g)]' + \phi\frac{\nu_g}{\sqrt{T}\|\nu_g\|^2}\frac{\|z'\nu_g\|}{\|\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)z'\nu_g\|}[\text{dir}(z'\nu_g)]'. \end{aligned} \tag{30}$$

with $\phi \sim \chi_K$ which follows from Lemma A.4 under $\mathcal{H}_0$. This in turn gives

$$p_{g,\infty}(d_{g,NT}(H_{i,t-\tau},\gamma)) = \mathbb{P}_{\mathcal{H}_0}\left[\phi \geq d_{g,NT}(H_{i,t-1},\gamma) \;\middle|\; \bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{g_i^{(m)}(z(\phi)) = g_i^{(m)}(z)\right\}\right],$$

which shows that $p_{g,\infty}(d_{g,NT}(H_{i,t-1},\gamma))$ can be calculated as the survival function of a χ_K variable truncated to the set $\mathcal{S} = \left\{\phi \in \mathbb{R} : \bigcap_{m=0}^{M}\bigcap_{i=1}^{N} g_i^{(m)}(z(\phi)) = g_i^{(m)}(z)\right\}$, that is, $p_{g,NT}(d_{g,NT}(H_{i,t-\tau})) = 1 - F_{\chi_K}\left[d_{g,NT}(H_{i,t-\tau});\mathcal{S}\right]$. This completes the proof.

A.4. Proof of Proposition 1

Lemma A.6. Suppose that Assumptions 1-5, and $\mathcal{H}_1^{1,g} : \lim_{N\to\infty}\theta_{1,n_1}^0(\gamma^0) \neq \lim_{N\to\infty}\theta_{g,n_g}^0(\gamma^0)$ hold. Then, $D_{g,NT}(H_{i,t-\tau})$ diverges as $B\to\infty$, $(T,N)\to\infty$ such that $B/T\to 0$.

Proof. We first write

$$\begin{aligned} T^{-1/2}D_{g,NT}(H_{i,t-\tau}) &= \left\|\widehat{\sigma}_{1,g,NT}^{-1/2}(\hat{\gamma}_{NT})[\hat{\theta}_{1,\hat{n}_1,T}(\hat{\gamma}_{NT}) - \hat{\theta}_{g,\hat{n}_g,T}(\hat{\gamma}_{NT})]\right\| \\ &\xrightarrow{p} \left\|\sigma_{1,g}^{-1/2}(\gamma^0)[\theta_{1,n_1}^0(\gamma^0) - \theta_{g,n_g}^0(\gamma^0)]\right\| > 0, \end{aligned}$$

by Lemma 1(a), Lemma 2 and Assumption 3 from which $\sigma_{1,g}$ is positive definite. Then the result follows from the fact that under $\mathcal{H}_1^{1,g}$, $\|\theta_{1,n_1}^0(\gamma^0) - \theta_{g,n_g}^0(\gamma^0)\| > 0$. □

To prove the first part of the proposition, we write

$$\begin{aligned} &\limsup_{(T,N)\to\infty} \mathbb{P}\Bigg[p_{g,NT}(D_{g,NT}(H_{i,t-\tau},\gamma)) \leq \alpha \;\Bigg| \\ &\quad \bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{g_i^{(m)}\left(\Pi_g Z + D_{g,NT}(H_{t-\tau},\gamma)\frac{\nu_g}{\sqrt{T}\|\nu_g\|^2}\{\text{dir}(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g)\}'\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)\right) = g_i^{(m)}(z)\right\}, \\ &\quad \Pi_g Z = \Pi_g z, \text{dir}\left(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)Z'\nu_g\right) = \text{dir}\left(\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma)z'\nu_g\right)\Bigg] \\ &= \limsup_{(T,N)\to\infty} \mathbb{P}\left[p_{g,NT}(D_{g,NT}(H_{i,t-\tau},\gamma)) \leq \alpha \;\middle|\; \bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{g_i^{(m)}\left(z[D_{g,NT}(H_{i,t-\tau},\gamma)]\right) = g_i^{(m)}(z)\right\}\right] \end{aligned}$$

$$= \limsup_{(T,N)\to\infty} \mathbb{P}\left[1 - F_{\chi_K}\left[D_{g,NT}(H_{i,t-\tau},\gamma);\mathcal{S}\right] \leq \alpha \;\middle|\; \bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{g_i^{(m)}\left(z[D_{g,NT}(H_{i,t-\tau},\gamma)]\right) = g_i^{(m)}(z)\right\}\right],$$

which follows similar lines to the ones above and the definition of $F_{\chi_K}\left[D_{g,NT}(H_{i,t-\tau},\gamma);\mathcal{S}\right]$ as the cumulative distribution function of a χ_K variate truncated to the set $\mathcal{S}$. It remains to show that

$$\limsup_{(T,N)\to\infty} \mathbb{P}\left[1 - F_{\chi_K}\left[D_{g,NT}(H_{i,t-\tau},\gamma);\mathcal{S}\right] \leq \alpha \;\middle|\; \bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{g_i^{(m)}\left(z[D_{g,NT}(H_{i,t-\tau},\gamma)]\right) = g_i^{(m)}(z)\right\}\right] = \alpha.$$

To show this, we note that, under $\mathcal{H}_0$, the conditional distribution function of $D_{g,NT}(H_{i,t-\tau},\gamma)$ given $\bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{g_i^{(m)}\left(z[D_{g,NT}(H_{i,t-\tau},\gamma)]\right) = g_i^{(m)}(z)\right\}$ is $F_{\chi_K}(\cdot,\mathcal{S})$.

$$\begin{aligned}
&\limsup_{(T,N)\to\infty} \mathbb{P}\left[p_{g,NT}(D_{g,NT}(H_{i,t-\tau},\gamma)) \leq \alpha \;\middle|\; \bigcap_{i=1}^{N}\left\{g_i^{(M)}(Z) = g_i^{(M)}(z)\right\}\right] \\
&\quad= \lim_{(T,N)\to\infty} \mathrm{E}\left[\mathbf{1}\left\{p_{g,NT}(D_{g,NT}(H_{i,t-\tau},\gamma)) \leq \alpha\right\} \;\middle|\; \bigcap_{i=1}^{N}\left\{g_i^{(M)}(Z) = g_i^{(M)}(z)\right\}\right] \\
&\quad= \lim_{(T,N)\to\infty} \mathrm{E}\left[\mathrm{E}\left(\mathbf{1}\left\{p_{g,NT}(D_{g,NT}(H_{i,t-\tau},\gamma)) \leq \alpha\right\} \;\middle|\; \bigcap_{m=0}^{M}\bigcap_{i=1}^{N}\left\{g_i^{(m)}(Z) = g_i^{(m)}(z)\right\}, \Pi_g Z = \Pi_g z,\right.\right. \\
&\qquad \left.\left.\operatorname{dir}\left(\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma)Z'\nu_g\right) = \operatorname{dir}\left(\widehat{\sigma}_{1,g,NT}^{1/2}(\gamma)z'\nu_g\right)\right) \;\middle|\; \bigcap_{i=1}^{N}\left\{g_i^{(M)}(Z) = g_i^{(M)}(z)\right\}\right] \\
&\quad= \lim_{(T,N)\to\infty} \mathrm{E}\left[\alpha \;\middle|\; \bigcap_{i=1}^{N}\left\{g_i^{(M)}(Z) = g_i^{(M)}(z)\right\}\right] = \alpha,
\end{aligned}$$

which concludes the proof of Part (a).

Part (b) follows directly from Lemma A.6 which implies that $D_{g,NT}(H_{i,t-\tau}) \to \infty$ under the alternative hypothesis, hence, for any $\alpha \in (0,1)$

$$\lim_{(T,N)\to\infty} \mathbb{P}[p_{g,NT}(D_{g,NT}(H_{i,t-\tau})) \leq \alpha] = 1,$$

and noting that under Lemma (d)(b), the conditioning event holds with probability 1.

A.5. Proof of Theorem 1

Lemma A.7. Let $P_{NT} = (P_{1,NT},\ldots,P_{n,NT})'$ a random n-vector such that $P_{NT} \xrightarrow{d} P$ as $(T,N) \to \infty$ where P is an n-vector of p-values. Let $h(x_1,\ldots,x_n) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^{-r}\right)^{1/r}$, $r \in (1,\infty)$. Then $h(P_{NT}) \xrightarrow{d} h(P)$.

Proof. This follows from the Continuous Mapping Theorem by the fact that $h(\,\cdot\,)$ is continuous. □

Lemma A.8. Let $P_{NT} = (P_{1,NT}, \ldots, P_{n,NT})'$ a random n-vector such that $P_{NT} \xrightarrow{d} P$ as $(T, N) \to \infty$ where P is an n-vector of p-values. Let $\mathcal{C}_\alpha = \{(p_1, \ldots, p_n) \in [0,1]^n : f(p_1, \ldots, p_n) \leq \alpha\}$, for all $\alpha \in (0,1)$, with $f(p_1, \ldots, p_n) = \min\left[\frac{r}{r-1} n \left(\sum_{i=1}^n p_i^{-r}\right)^{-1/r}, 1\right]$, $r \in (1, \infty)$. Then $\lim_{(T,N)\to\infty} \mathbb{P}(P_{NT} \in \mathcal{C}_\alpha) \leq \mathbb{P}(P \in \mathcal{C}_\alpha)$.

Proof. This follows from the Portmanteau Theorem by the fact that the sets $\mathcal{C}_\alpha$ are closed (see, Section 3.4 of Gasparin et al., 2024). □

From Theorem 1 of SU, we have

$$\mathbb{P}\left[\left(\frac{1}{G-1}\sum_{g=2}^{G} P_g^{-r}\right)^{-1/r} \leq \alpha(G-1)^{(1-r)/r}\frac{r-1}{r}\right] \leq \alpha,$$

where, the random variables P_g are defined by $p_{g,NT}(D_{g,NT}(H_{i,t-\tau})) \xrightarrow{d} P_g \sim U[0,1]$ as $(T, N) \to \infty$ for all $g \in \{2, \ldots, G\}$ which holds by Proposition 1(a). For part (a), as in the proof of Proposition 2 of SU, we write the above inequality as

$$\mathbb{P}\left[\frac{r}{r-1}(G-1)\left(\sum_{g=2}^{G} P_g^{-r}\right)^{-1/r} \leq \alpha\right] = \mathbb{P}\left(\frac{r}{r-1}\frac{1}{W^{homo}(H_{i,t-\tau})} \leq \alpha\right) \leq \alpha,$$

where, $W^{homo}(H_{i,t-\tau})$ is defined by $W_{NT}^{homo}(H_{i,t-\tau}) \xrightarrow{d} W^{homo}(H_{i,t-\tau})$ as $(T, N) \to \infty$. Now, by applying Lemma A.8, we find

$$\lim_{(T,N)\to\infty} \mathbb{P}\left(\frac{r}{r-1}\frac{1}{W_{NT}^{homo}} \leq \alpha\right) = \mathbb{P}\left(\frac{r}{r-1}\frac{1}{W^{homo}(H_{i,t-\tau})} \leq \alpha\right) \leq \alpha,$$

which ends the proof of part (a).

Part (b) now follows from Proposition 1(a) under which at least for one $g \in \{1, \ldots, G\}$ the p-value satisfies $p_{g,NT}(D_{g,NT}(H_{i,t-\tau})) \to 1$.

A.6. Proof of Proposition 2

Part (a) follows directly from Theorem 3.1 of Sun (2013) under our Assumptions 1 and 3 by setting $\gamma = (1, \ldots, 1)'$. Part (b) follows from Section 4.1 of Sun (2011) under the same assumptions.

A.7. Proof of Theorem 2

Part (a) follows the same lines as the proof of Theorem 1 and noting that the p-value associated to the O-EPA test statistic is asymptotically uniform by Proposition 2. Similarly, Part (b) follows

from the fact that under the alternative hypothesis, either at least for one $g \in \{1, \ldots, G\}$ the p-value satisfies $p_{g,NT}(D_{g,NT}(H_{i,t-\tau})) \to 1$ and the conditioning event holds with probability 1 by Lemma 2(b), or the O-EPA test statistic diverges.

A.8. Proof of Theorem 3

The proof begins algebraically similar to the proof of Lemma 1 except that we will establish a CLT conditional on $\mathcal{G}_{\mathcal{R}} = \sigma(\{Z_{it}\}_{i=1}^{N}, t \in \mathcal{R})$. First, we will show that each $K \times 1$ sub-vector of $\hat{\theta}_{NP}(\hat{\gamma}_{NR})$ satisfies $\hat{\theta}_{g,\hat{n}_g,P}(\hat{\gamma}_{NR}) = \theta^0_{g,n_g}(\hat{\gamma}_{NR}) + o_p(1)$. We have,

$$\begin{aligned}
&\mathrm{E}(\hat{\theta}_{g,\hat{n}_g,P}(\hat{\gamma}_{NR}) - \theta^0_{g,n_g}(\hat{\gamma}_{NR}) \mid \mathcal{G}_{\mathcal{R}}) \\
&\qquad = \mathrm{E}\left(\frac{1}{\hat{n}_g P} \sum_{i=1}^{N} \sum_{t=S+1}^{T} V_{it} \{\hat{g}_{i,NR} = g\} \;\middle|\; \mathcal{G}_{\mathcal{R}} \right) \\
&\qquad = \frac{1}{\hat{n}_g P} \sum_{i=1}^{N} \sum_{t=S+1}^{T} \mathrm{E}(V_{it} \mid \mathcal{G}_{\mathcal{R}}) \{\hat{g}_{i,NR} = g\} \\
&\qquad = 0,
\end{aligned} \tag{31}$$

by Assumption 7. For the variance, we find,

$$\begin{aligned}
&\left\| \mathrm{E}\left[(\hat{\theta}_{g,\hat{n}_g,P}(\hat{\gamma}_{NR}) \; - \; \theta^0_{g,n_g}(\hat{\gamma}_{NR}) (\hat{\theta}_{g,\hat{n}_g,P}(\hat{\gamma}_{NR}) - \theta^0_{g,n_g}(\hat{\gamma}_{NR})' \;\middle|\; \mathcal{G}_{\mathcal{R}} \right] \right\| \\
&\qquad = \left\| \mathrm{E}\left[\frac{1}{(\hat{n}_g P)^2} \sum_{i,j=1}^{N} \sum_{t,s=S+1}^{T} V_{it} V_{js}' \mathbf{1}\{\hat{g}_{i,NR} = g\} \mathbf{1}\{\hat{g}_{j,NR} = g\} \;\middle|\; \mathcal{G}_{\mathcal{R}} \right] \right\| \\
&\qquad \leq \frac{1}{\hat{n}_g^2 P} \sum_{i,j=1}^{N} \left\| \frac{1}{P} \sum_{t,s=S+1}^{T} \mathrm{E}\left(V_{it} V_{js}' \mid \mathcal{G}_{\mathcal{R}} \right) \right\| \mathbf{1}\{\hat{g}_{i,NR} = g\} \mathbf{1}\{\hat{g}_{j,NR} = g\} \\
&\qquad \leq \frac{1}{\hat{n}_g^2 P} \sum_{i,j=1}^{N} \left\| \frac{1}{P} \sum_{t,s=S+1}^{T} \mathrm{E}\left(V_{it} V_{js}' \mid \mathcal{G}_{\mathcal{R}} \right) \right\| = O_p\left(\frac{1}{\kappa_g^2 P} \right),
\end{aligned} \tag{32}$$

by Assumptions 1 and 2 from which it follows that $\hat{\theta}_{g,\hat{n}_g,P}(\hat{\gamma}_{NR}) = \theta^0_{g,n_g}(\hat{\gamma}_{NR}) + o_p(1)$. Now, by Assumption 3, conditional on $\mathcal{G}_{\mathcal{R}}$ and under $\mathcal{H}_0$, as $P, R \to \infty$, $(T, N) \to \infty$ we have

$$\begin{aligned}
&\Omega_{NP}(\hat{\gamma}_{NR})^{-1/2} \hat{\mathcal{N}}^{1-\epsilon} P^{1/2} (\hat{\theta}_{g,\hat{n}_g,P}(\hat{\gamma}_{NR}) - \theta^0_{g,n_g}(\hat{\gamma}_{NR})) \\
&\qquad = \Omega_{NP}(\hat{\gamma}_{NR})^{-1/2} \hat{\mathcal{N}}^{1-\epsilon} P^{-1/2} \sum_{t=S+1}^{T} \bar{V}_{N,t}(\hat{\gamma}_{NR}) \xrightarrow{d} \mathbb{N}(0, I_K),
\end{aligned}$$

with $\Omega_{NP}(\hat{\gamma}_{NR}) = P^{-1} \sum_{t,s=S+1}^{T} \hat{\mathcal{N}}^{1-\epsilon} \mathrm{E}[\bar{V}_{N,t}(\hat{\gamma}_{NR}) \bar{V}_{N,s}'(\hat{\gamma}_{NR})] \hat{\mathcal{N}}^{1-\epsilon}$ where $\hat{\mathcal{N}} = \mathrm{diag}(\hat{n}_1, \ldots, \hat{n}_G) \otimes I_K$. Part (a) then follows from Theorem 1 of Sun (2013) noting that $\mathcal{N}^{1-\epsilon} \widehat{\Omega}_{NP}(\hat{\gamma}_{NR}) \mathcal{N}^{1-\epsilon} - \Omega(\hat{\gamma}_{NR}) =$

$o_p(1)$, conditional on $\mathcal{G}_{\mathcal{R}}$.

For Part (b), we first write

$$\begin{aligned}\hat{\theta}_{NP}(\hat{\gamma}_{NR}) - \theta^0(\gamma^0) &= [\hat{\theta}_{NP}(\hat{\gamma}_{NR}) - \hat{\theta}_{NR}(\hat{\gamma}_{NR})] + [\hat{\theta}_{NR}(\hat{\gamma}_{NR}) - \theta^0(\gamma^0)] \\ &= [\hat{\theta}_{NP}(\hat{\gamma}_{NR}) - \hat{\theta}_{NR}(\hat{\gamma}_{NR})] + o_p(1),\end{aligned}$$

as $(R, N) \xrightarrow{p} \infty$, which follows from Lemma 2(a). We will show that the first term is also $o_p(1)$. To see this we focus on the $K \times 1$ sub-vectors of the term:

$$\begin{aligned}\hat{\theta}_{g,\hat{n}_g,P}(\hat{\gamma}_{NR}) - \hat{\theta}_{g,\hat{n}_g,R}(\hat{\gamma}_{NR}) &= \frac{1}{\hat{n}_g P}\sum_{i=1}^{N}\sum_{t=S+1}^{T} Z_{it}\mathbf{1}\{\hat{g}_{i,NR} = g\} - \frac{1}{\hat{n}_g R}\sum_{i=1}^{N}\sum_{t=1}^{R} Z_{it}\mathbf{1}\{\hat{g}_{i,NR} = g\} \\ &= \frac{1}{\hat{n}_g P}\sum_{i=1}^{N}\sum_{t=S+1}^{T} V_{it}\mathbf{1}\{\hat{g}_{i,NR} = g\} - \frac{1}{\hat{n}_g R}\sum_{i=1}^{N}\sum_{t=1}^{R} V_{it}\mathbf{1}\{\hat{g}_{i,NR} = g\} \\ &= \frac{1}{P}\sum_{t=S+1}^{T}\bar{V}_{g,\hat{n}_g,t} - \frac{1}{R}\sum_{t=1}^{R}\bar{V}_{g,\hat{n}_g,t} \\ &= O_p\left(\frac{\kappa_g^{\epsilon-1}}{N^{1-\epsilon}\sqrt{P}}\right) + O_p\left(\frac{\kappa_g^{\epsilon-1}}{N^{1-\epsilon}\sqrt{R}}\right) = o_p(1).\end{aligned}$$

This in turn gives that

$$\hat{\theta}'_{NP}(\hat{\gamma}_{NR})\mathcal{N}^{\epsilon-1}\widehat{\Omega}^{-1}_{NP}(\hat{\gamma}_{NR})\mathcal{N}^{\epsilon-1}\hat{\theta}_{NP}(\hat{\gamma}_{NR}) \xrightarrow{p} \theta^{0\prime}(\gamma^0)\Omega^{-1}(\gamma^0)\theta^0(\gamma^0) > 0,$$

by Assumptions 3 and 4 from which it follows that $\widehat{\Omega}_{NP}(\hat{\gamma}_{NR}) \to \infty$ which completes the proof.

B. Calculation of the Truncation Set $\mathcal{S}$

We start by writing the truncation set $\mathcal{S}$ as follows:

$$\mathcal{S} = \left\{\phi \in \mathbb{R} : \bigcap_{i=1}^{N} g_i^{(0)}(z(\phi)) = g_i^{(0)}(z)\right\} \bigcap \left\{\phi \in \mathbb{R} : \bigcap_{m=1}^{M}\bigcap_{i=1}^{N} g_i^{(m)}(z(\phi)) = g_i^{(m)}(z)\right\}.$$

As stated by Chen and Witten (2023), according to Step 2 of Algorithm 1, the equality in the first term holds if and only if the initial cluster center which is closest to z_{it} in total over t, coincides with the initial cluster center which is closest to $[z(\phi)]_{it}$ in total over t, for all $i = 1, \ldots, N$. This is similar for the equality in second term except that the cluster centers are determined by the assignments of the previous step in the iteration. Proposition 2 of Chen and Witten (2023) can then be generalized

as follows:

$$\mathcal{S} = \left(\bigcap_{i=1}^{N} \bigcap_{g=1}^{G} \left\{ \phi : \sum_{t=1}^{T} \left\| [z(\phi)]_{it} - \theta^{(0)}_{g_i^{(0)}(z)} [z(\phi)] \right\|^2 \leq \sum_{t=1}^{T} \left\| [z(\phi)]_{it} - \theta^{(0)}_{g} [z(\phi)] \right\|^2 \right\} \right) \bigcap$$
$$\left(\bigcap_{m=1}^{M} \bigcap_{i=1}^{N} \bigcap_{g=1}^{G} \left\{ \phi : \sum_{t=1}^{T} \left\| [z(\phi)]_{it} - \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} w_j^{(m-1)} \left(g_i^{(m)}(z) \right) [z(\phi)]_{jt} \right\|^2 \right. \right. \tag{33}$$
$$\left. \left. \leq \sum_{t=1}^{T} \left\| [z(\phi)]_{it} - \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} w_j^{(m-1)}(g) [z(\phi)]_{jt} \right\|^2 \right\} \right),$$

where $w_i^{(m)}(g) = \mathbf{1}\left\{ g_i^{(m)}(z) = g \right\} / \sum_{j=1}^{N} \mathbf{1}\left\{ g_j^{(m)}(z) = g \right\}$. By (30), we see that,

$$[z(\phi)]_{it} = z_{it} - \hat{\delta}_i \frac{\|z'\nu_g\|}{\|\nu_g\|^2} \mathrm{dir}(z'\nu_g) + \left(\frac{\|z'\nu_g\|}{\|\widehat{\sigma}^{-1/2}_{1,g,NT}(\gamma) z'\nu_g\|} \frac{\hat{\delta}_i}{\sqrt{T}\|\nu_g\|^2} \phi \right) \mathrm{dir}(z'\nu_g). \tag{34}$$

Straightforward calculations which are similar to the proofs of Lemmas 14 and 15 of Chen and Witten (2023) give

$$\left\| [z(\phi)]_{it} - \frac{1}{T} \sum_{t=1}^{T} [z(\phi)]_{jt} \right\|^2 = a_{ij}\phi^2 + b_{ijt}\phi + c_{ijt},$$

where

$$a_{ij} = \left(\frac{\|z'\nu_g\|}{\|\widehat{\sigma}^{-1/2}_{1,g,NT}(\gamma) z'\nu_g\|} \right)^2 \left(\frac{\hat{\delta}_i - \hat{\delta}_j}{\sqrt{T}\|\nu_g\|^2} \right)^2,$$
$$b_{ijt} = 2 \left(\frac{\|z'\nu_g\|}{\|\widehat{\sigma}^{-1/2}_{1,g,NT}(\gamma) z'\nu_g\|} \right) \left(\frac{\hat{\delta}_i - \hat{\delta}_j}{\sqrt{T}\|\nu_g\|^2} \langle z_{it} - \bar{z}_{j,T}, \mathrm{dir}(z'\nu_g) \rangle - \frac{(\hat{\delta}_i - \hat{\delta}_j)^2}{\sqrt{T}\|\nu_g\|^4} \|z'\nu_g\| \right),$$
$$c_{ijt} = \left\| z_{it} - \bar{z}_{j,T} - (\hat{\delta}_i - \hat{\delta}_j) \frac{z'\nu_g}{\|\nu_g\|^2} \right\|^2,$$

and

$$\left\| [z(\phi)]_{it} - \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} w_j^{(m-1)}(g) [z(\phi)]_{jt} \right\|^2 = \tilde{a}_{ij}\phi^2 + \tilde{b}_{ijt}\phi + \tilde{c}_{ijt},$$

where

$$\tilde{a}_{ij} = \left(\frac{\|z'\nu_g\|}{\|\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma) z'\nu_g\|} \right)^2 \left(\frac{\hat{\delta}_i - \sum_{j=1}^N w_j^{(m-1)}(g)\hat{\delta}_j}{\sqrt{T}\|\nu_g\|^2} \right)^2,$$

$$\tilde{b}_{ijt} = 2 \left(\frac{\|z'\nu_g\|}{\|\widehat{\sigma}_{1,g,NT}^{-1/2}(\gamma) z'\nu_g\|} \right) \times \left\{ \frac{\hat{\delta}_i - \sum_{j=1}^N w_j^{(m-1)}(g)\hat{\delta}_j}{\sqrt{T}\|\nu_g\|^2} \left\langle z_{it} - \frac{1}{T}\sum_{t=1}^T \sum_{j=1}^N w_j^{(m-1)}(g) z_{jt}, \mathrm{dir}(z'\nu_g) \right\rangle - \frac{(\hat{\delta}_i - \sum_{j=1}^N w_j^{(m-1)}(g)\hat{\delta}_j)^2}{\sqrt{T}\|\nu_g\|^4} \|z'\nu_g\| \right\},$$

$$\tilde{c}_{ijt} = \left\| z_{it} - \frac{1}{T}\sum_{t=1}^T \sum_{j=1}^N w_j^{(m-1)}(g) z_{jt} - \left(\hat{\delta}_i - \sum_{j=1}^N w_j^{(m-1)}(g)\hat{\delta}_j \right) \frac{z'\nu_g}{\|\nu_g\|^2} \right\|^2.$$

These in turn show that the truncation set $\mathcal{S}$ can be analytically calculated as the inequalities defined in the two components of (33) are all quadratic in ϕ.

C. Additional Results

C.1. Consequences of naive testing

To illustrate the consequences of C-EPA inference following *kmeans*, we generate 2000 samples from the following DGP: $\Delta L_{it} = \lambda_i F_{t-1} + U_{it}$ where $U_{it} \sim iidN(0,1)$, $F_t \sim iidN(0,1)$, $\lambda_i \sim iidN(0, 0.2)$ with $N = 50$, $T = 20$. We then calculate $p_{NT}(W_{NT}(1,\gamma))$ for each 2000 samples, that is, the p-values associated with the unconditional C-EPA test statistic based on predetermined clusters. For this test, we assume $G = 2$ is predetermined and we set $g_i = 1$ for $i = 1, \ldots, 25$ and $g_i = 2$ for $i = 26, \ldots, 50$. In addition, we compute the naive p-values $p_{NT}(W_{NT}(1, \hat{\gamma}_{NT}))$ with $G = 2$. The histograms of these p-values are given in Figure 2.

The results show that the p-value of the test (4) is approximately uniform under the null of C-EPA whereas the test statistic using the *kmeans* estimates is extremely anti-conservative. Hence, the consequences of double dipping to test the homogeneity of the cluster means are equally present in the case of C-EPA testing post-clustering.

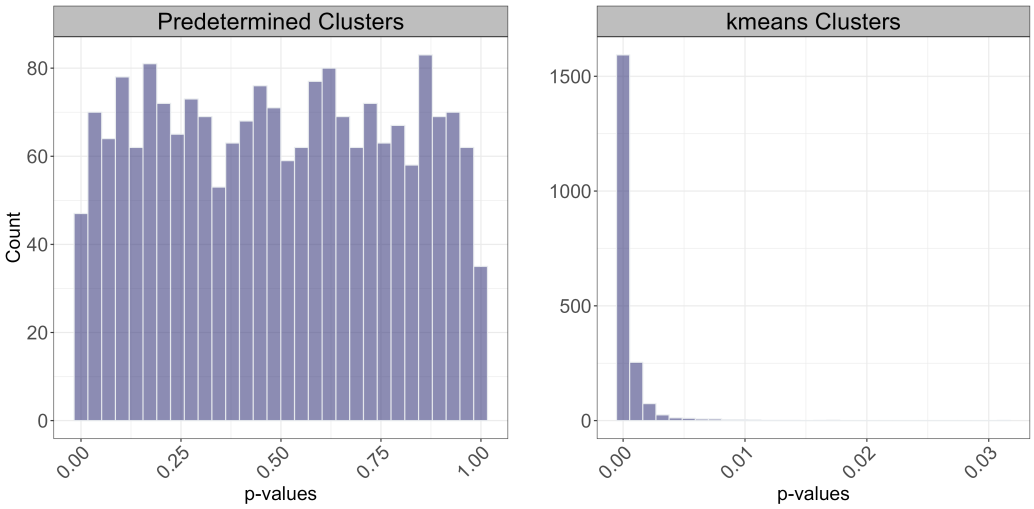

Figure 2: Histograms of the p-values $p_{NT}(W_{NT}(1,\gamma))$ and $p_{NT}(W_{NT}(1,\hat{\gamma}_{NT}))$

C.2. Details on Monte Carlo simulations and the empirical illustration

Table 5: Performance of the *kmeans* Estimator and the Proposed Information Criterion–$G^0 = 2$

T	d	DGP	Rec.	Rand. Ind.	$\widehat{G}_{NT}$	T	d	DGP	Rec.	Rand. Ind.	$\widehat{G}_{NT}$
Unconditional Tests ($H_{i,t-1} = 1$)						Conditional Tests ($H_{i,t-1} = F_{1,t-1}$)					
Main Results											
20	0.125	1	0.00	0.59	2.00	20	0.125	1	0.00	0.62	2.00
50	0.125	1	0.00	0.68	2.01	50	0.125	1	0.00	0.78	2.00
100	0.125	1	0.00	0.79	2.02	100	0.125	1	0.13	0.91	2.00
200	0.125	1	0.10	0.91	2.05	200	0.125	1	0.62	0.98	2.00
20	0.25	1	0.00	0.75	2.10	20	0.25	1	0.06	0.85	2.02
50	0.25	1	0.09	0.91	2.15	50	0.25	1	0.61	0.98	2.00
100	0.25	1	0.64	0.98	2.11	100	0.25	1	0.97	1.00	2.00
200	0.25	1	0.97	1.00	2.05	200	0.25	1	1.00	1.00	2.00
20	0.375	1	0.06	0.89	2.27	20	0.375	1	0.48	0.96	2.00
50	0.375	1	0.74	0.99	2.14	50	0.375	1	0.97	1.00	2.00
100	0.375	1	0.99	1.00	2.06	100	0.375	1	1.00	1.00	2.00
200	0.375	1	1.00	1.00	2.03	200	0.375	1	1.00	1.00	2.00
20	0.5	1	0.42	0.96	2.31	20	0.5	1	0.86	0.99	2.00
50	0.5	1	0.98	1.00	2.12	50	0.5	1	1.00	1.00	2.00
100	0.5	1	1.00	1.00	2.05	100	0.5	1	1.00	1.00	2.00
200	0.5	1	1.00	1.00	2.04	200	0.5	1	1.00	1.00	2.00
Alternative DGPs											
50	0.25	2	0.13	0.92	2.24	50	0.25	2	0.66	0.98	2.00
100	0.25	2	0.71	0.99	2.12	100	0.25	2	0.98	1.00	2.00
50	0.25	3	0.11	0.91	2.13	50	0.25	3	0.61	0.98	2.00
100	0.25	3	0.65	0.98	2.09	100	0.25	3	0.97	1.00	2.00
50	0.25	4	0.00	0.79	2.86	50	0.25	4	0.23	0.93	2.05
100	0.25	4	0.08	0.90	2.94	100	0.25	4	0.79	0.99	2.02
50	0.25	5	0.00	0.78	2.70	50	0.25	5	0.20	0.92	2.02
100	0.25	5	0.07	0.90	2.80	100	0.25	5	0.76	0.99	2.01

Table 6: Empirical Size and Power of the Homogeneity and Overall EPA Tests–Main Results

N	T	d	DGP	W_{NT}^{oepa}	W_{NT}^{homo}	$W_{NT}^{homo}[IC]$	N	T	d	DGP	W_{NT}^{oepa}	W_{NT}^{homo}	$W_{NT}^{homo}[IC]$
Unconditional Tests							Conditional Tests						
50	20	0	1	0.05	0.06	0.09	50	20	0	1	0.05	0.04	0.05
100	20	0	1	0.05	0.06	0.06	100	20	0	1	0.05	0.04	0.03
50	50	0	1	0.05	0.06	0.05	50	50	0	1	0.05	0.04	0.04
100	50	0	1	0.05	0.05	0.05	100	50	0	1	0.05	0.05	0.05
50	100	0	1	0.06	0.04	0.05	50	100	0	1	0.05	0.05	0.05
100	100	0	1	0.06	0.05	0.05	100	100	0	1	0.06	0.04	0.04
50	200	0	1	0.05	0.05	0.06	50	200	0	1	0.05	0.05	0.05
100	200	0	1	0.05	0.05	0.05	100	200	0	1	0.05	0.04	0.05
50	20	0.125	1	0.32	0.09	0.10	50	20	0.125	1	0.22	0.05	0.05
50	50	0.125	1	0.70	0.14	0.15	50	50	0.125	1	0.58	0.06	0.07
50	100	0.125	1	0.95	0.19	0.19	50	100	0.125	1	0.90	0.15	0.16
50	200	0.125	1	1.00	0.60	0.62	50	200	0.125	1	1.00	0.67	0.66
50	20	0.25	1	0.87	0.24	0.25	50	20	0.25	1	0.77	0.07	0.08
50	50	0.25	1	1.00	0.62	0.59	50	50	0.25	1	0.99	0.34	0.36
50	100	0.25	1	1.00	0.90	0.87	50	100	0.25	1	1.00	0.84	0.83
50	200	0.25	1	1.00	0.98	0.96	50	200	0.25	1	1.00	0.98	0.98
50	20	0.375	1	0.99	0.50	0.43	50	20	0.375	1	0.98	0.16	0.17
50	50	0.375	1	1.00	0.92	0.90	50	50	0.375	1	1.00	0.75	0.77
50	100	0.375	1	1.00	0.99	0.98	50	100	0.375	1	1.00	0.96	0.96
50	200	0.375	1	1.00	1.00	0.99	50	200	0.375	1	1.00	0.99	0.99
50	20	0.5	1	1.00	0.80	0.72	50	20	0.5	1	1.00	0.36	0.36
50	50	0.5	1	1.00	0.99	0.95	50	50	0.5	1	1.00	0.92	0.92
50	100	0.5	1	1.00	1.00	0.99	50	100	0.5	1	1.00	0.99	0.99
50	200	0.5	1	1.00	1.00	0.99	50	200	0.5	1	1.00	1.00	0.99

Table 7: Empirical Size and Power of the Homogeneity and Overall EPA Tests–Alternative DGPs

N	T	d	DGP	W_{NT}^{oepa}	W_{NT}^{homo}	$W_{NT}^{homo}[IC]$	N	T	d	DGP	W_{NT}^{oepa}	W_{NT}^{homo}	$W_{NT}^{homo}[IC]$
Unconditional Tests							Conditional Tests						
50	50	0	2	0.05	0.06	0.06	50	50	0	2	0.05	0.05	0.06
50	100	0	2	0.06	0.05	0.06	50	100	0	2	0.05	0.05	0.05
50	50	0	4	0.05	0.12	0.10	50	50	0	4	0.05	0.08	0.08
50	100	0	4	0.05	0.10	0.11	50	100	0	4	0.05	0.09	0.08
50	50	0	5	0.05	0.10	0.09	50	50	0	5	0.04	0.06	0.07
50	100	0	5	0.05	0.10	0.10	50	100	0	5	0.05	0.07	0.07
50	50	0.25	2	1.00	0.63	0.58	50	50	0.25	2	1.00	0.38	0.37
50	100	0.25	2	1.00	0.93	0.88	50	100	0.25	2	1.00	0.87	0.87
50	50	0.25	3	0.06	0.60	0.59	50	50	0.25	3	0.05	0.38	0.37
50	100	0.25	3	0.05	0.89	0.85	50	100	0.25	3	0.05	0.86	0.86
50	50	0.25	4	1.00	0.35	0.27	50	50	0.25	4	1.00	0.22	0.19
50	100	0.25	4	1.00	0.59	0.39	50	100	0.25	4	1.00	0.65	0.64
50	50	0.25	5	0.89	0.34	0.27	50	50	0.25	5	0.78	0.20	0.20
50	100	0.25	5	0.99	0.57	0.40	50	100	0.25	5	0.98	0.62	0.61

Table 8: Estimated Exchange Rate Clusters–Conditional Test, Full Sample (AR1 vs. RW)

$g=1$			$g=2$	$g=3$	$g=4$	$g=5$
$\hat{\theta}_{1,\hat{n}_1,T}=(-0.12,0.01)'$			$\hat{\theta}_{2,\hat{n}_2,T}=(-0.3,0.21)'$	$\hat{\theta}_{3,\hat{n}_3,T}=(-0.32,0.3)'$	$\hat{\theta}_{4,\hat{n}_4,T}=(-0.2,0.08)'$	$\hat{\theta}_{5,\hat{n}_5,T}=(-0.36,0.32)'$
EUR/AUD	GBP/HKD	USD/HRK	EUR/RUB	EUR/BRL	GBP/RUB	GBP/CHF
EUR/CAD	GBP/ILS	USD/HUF	EUR/TRY	USD/BRL		GBP/TRY
EUR/CHF	GBP/INR	USD/IDR	USD/ARS			GBP/ZAR
EUR/CNY	GBP/JPY	USD/ILS	USD/RUB			
EUR/HKD	GBP/KRW	USD/INR	USD/TRY			
EUR/IDR	GBP/MYR	USD/ISK				
EUR/ILS	GBP/NOK	USD/JPY				
EUR/INR	GBP/NZD	USD/KRW				
EUR/JPY	GBP/SEK	USD/LKR				
EUR/KRW	GBP/SGD	USD/MKD				
EUR/MXN	GBP/THB	USD/MXN				
EUR/MYR	GBP/TWD	USD/MYR				
EUR/NOK	GBP/USD	USD/NOK				
EUR/NZD	USD/ALL	USD/NZD				
EUR/PHP	USD/AUD	USD/PHP				
EUR/SEK	USD/BAM	USD/PLN				
EUR/SGD	USD/BGN	USD/RON				
EUR/THB	USD/CAD	USD/SEK				
EUR/USD	USD/CHF	USD/SGD				
EUR/ZAR	USD/CNY	USD/THB				
GBP/AUD	USD/COP	USD/TWD				
GBP/CAD	USD/CZK	USD/UYU				
GBP/CNY	USD/DKK	USD/ZAR				
GBP/DKK	USD/DZD					
GBP/EUR	USD/GBP					

Acknowledgements

We thank Lucy L. Gao, Antonio Montañes, Ryo Okui, Hashem Pesaran, Esther Ruiz-Ortega and the participants of Centre for Econometric Analysis Occasional Econometrics Seminar at the Bayes Business School (London, 27 October 2023), the Annual Spatial Econometrics Association Conference (San Diego, 16-17 November 2023, and the 29$^{\text{th}}$ International Panel Data Conference (Orléans, 3-5 July 2024). The usual disclaimer applies.

References

Akgun, O., Pirotte, A., Urga, G., and Yang, Z. (2024). Equal predictive ability tests based on panel data with applications to OECD and IMF forecasts. *International Journal of Forecasting*, 40(1):202–228.

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, 59(3):817–858.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bailey, N., Kapetanios, G., and Pesaran, M. H. (2016). Exponent of cross-sectional dependence: Estimation and inference. *Journal of Applied Econometrics*, 31(6):929–960.

Bonhomme, S., Lamadon, T., and Manresa, E. (2022). Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643.

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

Chen, Y. T. and Witten, D. M. (2023). Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41.

Chudik, A., Pesaran, M. H., and Tosetti, E. (2011). Weak and strong cross-section dependence and estimation of large panels. *The Econometrics Journal*, 14(1):C45–C90.

Clark, T. and McCracken, M. (2013). Advances in forecast evaluation. *Handbook of Economic Forecasting*, 2:1107–1201.

Clark, T. E. and McCracken, M. W. (2015). Nested forecast model comparisons: a new approach to testing equal accuracy. *Journal of Econometrics*, 186(1):160–177.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Driscoll, J. C. and Kraay, A. C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics*, 80(4):549–560.

Fisher, R. (1932). *Statistical Methods for Research Workers*. Oliver & Boyd.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Gao, L. L., Bien, J., and Witten, D. (2024). Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545):332–342.

Gasparin, M., Wang, R., and Ramdas, A. (2024). Combining exchangeable p-values. *arXiv preprint arXiv:2404.03484*.

Giacomini, R. (2011). Testing conditional predictive ability. In *The Oxford Handbook of Economic Forecasting*. Oxford University Press.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Hansen, P. R. and Timmermann, A. (2012). Choice of sample split in out-of-sample forecast evaluation. Unpublished manuscript, Stanford and UCSD.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., and Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5):535–540.

Kuchibhotla, A. K., Kolassa, J. E., and Kuffner, T. A. (2022). Post-selection inference. *Annual Review of Statistics and Its Application*, 9:505–527.

Lin, C.-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1(1):42–55.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

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

Lunde, R. (2019). Sample splitting and weak assumption inference for time series. *arXiv preprint arXiv:1902.07425*.

McCracken, M. W. (2020). Diverging tests of equal predictive ability. *Econometrica*, 88(4):1753–1754.

Müller, U. K. (2007). A theory of robust long-run variance estimation. *Journal of Econometrics*, 141(2):1331–1352.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Patton, A. J. and Weller, B. M. (2023). Testing for unobserved heterogeneity via *k-means* clustering. *Journal of Business & Economic Statistics*, 41(3):737–751.

Phillips, P. C. (2005). HAC estimation by automated regression. *Econometric Theory*, 21(1):116–142.

Phillips, P. C. and Durlauf, S. N. (1986). Multiple time series regression with integrated processes. *The Review of Economic Studies*, 53(4):473–495.

Qu, R., Timmermann, A., and Zhu, Y. (2023). Comparing forecasting performance in cross-sections. *Journal of Econometrics*, 237(2):105–186.

Qu, R., Timmermann, A., and Zhu, Y. (2024). Comparing forecasting performance with panel data. *International Journal of Forecasting*, 40(3):918–941.

Rossi, B. (2021). Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them. *Journal of Economic Literature*, 59(4):1135–90.

Sarafidis, V. and Weber, N. (2015). A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics*, 77(2):274–296.

Spreng, L. and Urga, G. (2023). Combining p-values for multivariate predictive ability testing. *Journal of Business & Economic Statistics*, 41(3):765–777.

Su, L., Shi, Z., and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

Sun, Y. (2011). Robust trend inference with series variance estimator and testing-optimal smoothing parameter. *Journal of Econometrics*, 164(2):345–366.

Sun, Y. (2013). A heteroskedasticity and autocorrelation robust F test using an orthonormal series variance estimator. *The Econometrics Journal*, 16(1):1–26.

Sun, Y. (2014). Fixed-smoothing asymptotics in a two-step generalized method of moments framework. *Econometrica*, 82(6):2327–2370.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4):791–808.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, 64(5):1067–1084.

Zhu, Y. and Timmermann, A. (2020). Can two forecasts have the same conditional expected accuracy? *arXiv preprint arXiv:2006.03238*.