Testing Clustered Equal Predictive Ability with Unknown Clusters

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

[1] *Université Bourgogne Europe, LEDi UR 7467, 21000 Dijon, 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*

July 19, 2025

Abstract

We develop new tests of clustered equal predictive ability (C-EPA) in panels where the clusters are unknown and estimated by a Panel Kmeans algorithm. This algorithm differs from the standard Kmeans algorithm by employing the time series variation of the panel rather than relying merely on time averages of observations. To address the challenge of testing hypotheses that depend on data-driven cluster estimates, we adopt a selective conditional inference framework. Specifically, we derive a Wald-type test statistic for pairwise equality and show that the limiting distribution of its square root conditional on the estimated cluster structure is that of a truncated χ random variable. We characterize the associated truncation set as a polyhedron in the data space. As a test of the C-EPA hypothesis, we propose a p-value combination method which aggregates the evidence against the pairwise equality and overall EPA null hypotheses. In addition, we prove that using an information criterion to select the unknown number of clusters under the alternative hypothesis prior to testing does not require further conditioning to obtain a valid test. Monte Carlo simulations confirm the excellent finite sample performance of the proposed tests. An empirical application to forecasting exchange rates using traditional time series models as well as machine learning methods illustrates the practical importance of our procedure.

Keywords: Big Data; Forecast Evaluation; Hypothesis Testing; Sample Splitting; Selective Conditional Inference.

JEL classification: C12, C23.

Emails: oguzhan.akgun@u-bourgogne.fr (O. Akgün), alain.pirotte@assas-universite.fr (A. Pirotte), g.urga@city.ac.uk (G. Urga), zlyang@smu.edu.sg (Z. Yang).

1. Introduction

Despite the large and ever-growing literature[1] on testing equal predictive ability (EPA) using time series data, testing EPA with panels has attracted attention among econometricians only recently. To the best of our knowledge, the only contributions are those of Akgun, Pirotte, Urga and Yang (2024, APUY hereafter) and Qu, Timmermann and Zhu (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* refers to equivalence of $K \geq 2$ clusters of units.

In many applied forecasting contexts, the predictive performances of different forecasting models or agents vary across units, such as countries or firms. For example, Dreher et al. (2008) show that IMF forecasts significantly differ in quality depending on whether countries received IMF assistance or were aligned with major donors in international forums. Similarly, forecasting accuracy may differ systematically across units grouped by income level, geography, political alignments, or development status. This suggests that forecasting accuracy is heterogeneous across clusters, often in ways that are not directly observable to the researcher. In such cases, testing for EPA must accommodate clustered heterogeneity, often without knowing the clusters in advance.

The primary contribution of this paper is the development of novel conditional C-EPA tests for panel data where the cluster structure is unknown. Our framework advances beyond the recent papers by APUY and QTZ in several important directions. First, inspired by Giacomini and White (2006), we consider a general setting with conditioning variables, which provides a very flexible environment for practitioners. Although it may be considered trivial, this is an important extension that has not yet been considered in the panel data literature on EPA testing. Second and most importantly, we allow the clusters to be learned from the data using the Panel Kmeans algorithm which generalizes the classical Kmeans algorithm to panel settings. Unlike clustering on time averages of observations, our method fully exploits the time variation in the data. Third, to ensure valid inference after cluster estimation, we develop a selective conditional inference framework based on the polyhedral method (see, e.g., Lee et al., 2016). We develop a Wald-type test of a homogeneity on the centers of a pair of clusters and show that its asymptotic distribution is equivalent to that of a truncated χ-variate after conditioning on the estimated clusters. We derive the analytical characterization of the truncation region under the

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

Panel Kmeans algorithm. This is nontrivial because changes on the clustering algorithm changes the selection region, necessitating new derivations. Fourth, we prove that using an information criterion to select the number of clusters pre-testing does not invalidate our proposed procedure. Furthermore, we demonstrate through Monte Carlo experiments that using multiple random initializations also results in valid inference without additional conditioning. This fills in a gap in the literature as previous studies did not address two well known shortcomings of the Kmeans-type algorithms: the difficulty in selecting the number of clusters and the possibility of converging to local minima. Last, rather than relying on a Wald test for the joint C-EPA hypothesis, we propose a p-value combination approach that aggregates $n_p = K(K-1)/2$ pairwise equality tests and an O-EPA test. Although Wald tests for linear hypotheses post-clustering are easy to generalize to our setting, they tend to be anti-conservative when the number of constraints is large, which is the case in our specific problem of C-EPA testing (see Yun and He, 2024; Akgun and Okui, 2025, for an overall homogeneity test for Kmeans and for tests of general linear hypotheses in panels with latent clusters, respectively). Our p-value combination strategy avoids this complication and maintains proper control of the Type I error.

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 clustering, sequential binary segmentation, and Kmeans.[2] In this paper, we focus on the Panel Kmeans estimator 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; Patton and Weller, 2023, 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, Panel Kmeans can detect these clusters under general conditions. That is, the Panel 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 all units effectively belong to a single cluster. Although it remains potentially consistent, the asymptotic distribution is generally unknown. This leads to the problem of *double dipping* (see Kriegeskorte et al., 2009), where the same data are used for both model selection (via clustering, in our context) and inference.

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 hypotheses on cluster centers. However, the time dimension of a panel provides a solution to this, as in

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

Patton and Weller (2023) where a Split Sample test is proposed to test the homogeneity of the mean of a panel process among clusters chosen by Panel Kmeans. Despite its theoretical and computational simplicity, sample splitting has its drawbacks. First, Split Sample tests rely on the selection of two sub-samples. One sample, called *training* sample, is used for the estimation of the clusters and another, called *test* sample, for inference on the centers of these estimated clusters. However, this selection can be arbitrary in practice and there is no guidance on how to split the sample.[3] Second, structural breaks in the time period under consideration may completely invalidate the sample splitting method. In standard validation and cross-validation methods, samples are often divided randomly. In panel setting however, sample is divided at a given date which may exactly or approximately correspond to the date of breaks which can grossly affect the results of the tests. 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 $l \geq 1$ lags are allowed but impose independence beyond l lags. They show that, in this case, sample splitting continues to be valid if l periods between the two sub-samples are discarded. This validates the use of the Split Sample tests for this particular type of dependence but may result in loss of power.

In this paper, we develop an alternative selective conditional inference framework which uses the full sample of observations for both estimating the unknown clusters and making inference on their centers. In particular, we follow the recently developing literature on polyhedral method for inference after selection (Lee et al., 2016; Gao et al., 2024; Chen and Witten, 2023). Our main motivation source is the papers by Gao et al. (2024) and Chen and Witten (2023) who propose calculating selective p-values to test the equality of two cluster means post-clustering in a cross-sectional framework. The generalization of the methods followed in these papers to our context requires solving several non-trivial problems. These papers focus on the equality of a pair of cluster centers whereas we are interested in the joint null which states that all clusters have a zero mean. A recent attempt to generalize the framework to the joint equality of all cluster means has been made by by Yun and He (2024). Although it would not be too complicated to depart from this paper for testing our null, this would potentially lead to very poor performance in small samples due to the large number of constraints to be tested.

The methodology that we follow to overcome these difficulties can be summarized as follows.

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

First, we use a panel data version of the Lloyd's Kmeans algorithm (Lloyd, 1982) inspired by the influential work of Bonhomme and Manresa (2015). This algorithm is called Panel Kmeans as it differs substantially from the classical Kmeans which is developed for single indexed variables. Using this algorithm, we estimate the cluster membership variables as well as the cluster centers which measure the average predictive ability differences between competing forecasts for a given cluster. Second, based on the square root of a Wald statistic, we construct a test statistic measuring the difference in forecast loss differentials between estimated clusters. The usual critical values for a χ-variate are not valid for the problem in hand, as described above. Instead, after conditioning on the estimated clusters, we show that the test statistic follows a truncated χ distribution where the truncation sets are polygons in the data space. We derive the analytical formulae for the calculation of the truncation set. Third, we observe that the null hypothesis of C-EPA can be decomposed into n_p unique pairwise equality hypothesis and an O-EPA hypothesis which states that the overall mean of the panel is zero. Based on this simple observation, we apply a p-value combination method following the recent advances in the literature (Vovk and Wang, 2020; Vovk et al., 2022; Gasparin et al., 2025) using the p-values of the pairwise equality tests together with that of the O-EPA test.

Most of the literature on selective inference, and in particular Gao et al. (2024) and Chen and Witten (2023), are based on strong assumptions on the data generating process such as normality, homoskedasticity and independent observations. This is, of course, a very important constraint for our purpose. We derive the limiting theory of the proposed test statistics for potentially heteroskedastic, dependent and non-Gaussian panel data. 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 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 Panel 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 considerable power even under weak deviations from the C-EPA null.

We illustrate the empirical relevance of our methodology through an application to exchange rate forecasting, comparing the performance of traditional time series models with that of modern machine learning approaches. Drawing on a large dataset of bilateral exchange rates against the U.S. dollar, we evaluate the predictive ability of each method relative to a benchmark AR(1) model. Our results reveal substantial heterogeneity across exchange rate clusters and show that nonlinear models that incorporate macroeconomic fundamentals significantly outperforming conventional benchmarks. These findings complement recent evidence by Spreng and Urga (2023), who highlight the importance of powerful multivariate forecast comparison methods, and by Hillebrand et al. (2023), who emphasize the forecasting gains from utilizing macroeconomic fundamentals to forecast exchange rates.

All testing procedures developed in this paper are implemented in two dedicated `R` packages. The `clusteredEPA` package provides tools for testing EPA in the presence of latent clusters, including Selective Inference and Split Sample procedures. The companion package `PanelKmeansInference` focuses on post-clustering inference for coefficient homogeneity across clusters estimated via Panel Kmeans. Both packages and the replication material of the paper can be downloaded on `https://github.com/akoguzhan/`.

Organization of the paper. Section 2 presents the null and the alternative hypotheses of interest, three motivating examples, a generalized test of C-EPA with predetermined clusters, and the Panel Kmeans estimator of the unknown clusters. Section 3 discusses basic regularity conditions for our new tests and presents two useful lemmas. Section 4 introduces the tests of C-EPA with unknown clusters and presents their asymptotic properties. Section 5 presents essential Monte Carlo results. An empirical illustration is reported in Section 6. Section 7 concludes. Appendices A-D contain the proofs of the theoretical result and the description of the Split Sample tests.

Notation. Random variables are denoted by upper-case letters and their realizations by the corresponding lower-case letters, e.g., $w(\cdot)$ denotes a realization of the test statistic $W(\cdot)$. $\|\cdot\|$ denotes Euclidean norm, $\mathbf{1}\{\cdot\}$ is indicator function, $\text{diag}(\cdot)$ forms a diagonal matrix by given elements, $\text{tr}(\cdot)$ is the trace of a square matrix, $[\,\cdot\,]$ returns an integer by rounding, $|\cdot|$ denotes its cardinality when applied to a set, $\lfloor\,\cdot\,\rfloor$ returns the closest integer smaller than its argument, $\otimes$ denotes Kronecker product.

2. Setup and Preliminaries

In this section, we introduce the testing framework, the C-EPA null and alternative hypotheses. Then we present three motivating examples, and introduce a conditional C-EPA test with predetermined clusters that generalize APUY. Finally, we present the Panel Kmeans estimator and the associated algorithm which will serve as a tool to select the clusters based on which we will conduct the C-EPA tests with unknown clusters.

2.1. Testing framework and hypotheses

Let $\widehat{Y}_{a,it}$ be the τ-steps ahead, $\tau \geq 1$, forecast of agents $a = 1, 2$ 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 as in APUY and QTZ, or a forecasting model. To the best of our knowledge, there are no papers focusing on the theoretical comparison of out-of-sample forecasts of panel data models but the corresponding time series literature is rich (see Clark and McCracken, 2001, 2013, 2014, 2015; Giacomini and White, 2006, among others). A generic loss function is denoted by $L(\cdot, \cdot)$. 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 defined on a complete probability space $(\Omega, \mathcal{E}, \mathbb{P})$.

The null hypothesis of interest is the generalized C-EPA hypothesis. By "generalized" we mean that it allows conditioning variables, contrary to the unconditional null hypotheses considered in recent papers by APUY and QTZ. This null hypothesis is stated as

$$\mathcal{H}_0 : \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbb{E}(\Delta L_{it} \mid \mathcal{F}_{t-\tau}) = 0, \text{ a.s., for all } k = 1, 2, \ldots, K \tag{1}$$

where $\mathcal{F}_t \subseteq \mathcal{E}$ is a conditioning set (see the description below), and $\mathcal{C}_k$, $k = 1, \ldots, K$, are the sets of panel unit indexes. More concretely, $\mathcal{C}_k = \{i : k_i = k\}$ where $k_i \in \{1, 2, \ldots, K\}$ is the cluster membership indicator of unit i. These sets are mutually exclusive and exhaustive, that is, $\mathcal{C}_k \bigcap \mathcal{C}_g = \emptyset$ for all $k \neq g$, and $\bigcup_{k=1}^{K} \mathcal{C}_k = \{1, \ldots, N\}$. The alternative hypothesis is

$$\mathcal{H}_1 : \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \mathbb{E}(\Delta L_{it} \mid \mathcal{F}_{t-\tau}) \neq 0, \text{ for at least one } k = 1, 2, \ldots, K. \tag{2}$$

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, we could also focus on

the averages of these expectations over time. However, this requires different ways of estimating the variance or clustering (see Harvey et al., 2024, and Remark 3 below).

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 $\mathcal{F}_t = \{\emptyset, \Omega\}$. 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, consider the σ-field $\sigma(\{W_{is}\}_{i=1}^{N}, s \leq t)$ generated by the present and the past of the measurable-$\mathcal{E}$ 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 $\mathcal{F}_t = \sigma(\{W_{is}\}_{i=1}^{N}, s \leq t)$. Second, a researcher may be interested in the conditional EPA with respect to the realization of a vector of measurable-$\mathcal{E}$ common factors F_t. In this case, we let $\mathcal{F}_t = \sigma(F_s, s \leq t)$. 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.

Remark 1. The two conditional schemes described above, namely conditioning on observed covariates vs. common factors need not be mutually exclusive. In practice, one may estimate forecast errors from alternative panel data models that include both external predictors and observed or estimated common factors, and whose errors also exhibit spatial or network dependence. These general forecasting models nest both strong cross-sectional dependence via common factors and weak cross-sectional dependence via spatial interactions (see Chudik et al., 2011, for different types of cross-sectional dependence). In turn, one would expect the resulting loss differentials to contain different types of cross-sectional dependence via differences in alternative models. While our current theoretical framework accommodates general forms of cross-sectional dependence, formal treatment of parametric structures of cross-dependence in loss differentials could lead to more powerful inference (see APUY for further insights on the distinction of weak vs. strong cross-sectional dependence in EPA testing).

Remark 2. In some cases, one may want to compare a large number of forecasts made for a given unit with a base forecast. For example, in comparing the inflation forecasts of the survey of professional forecasters with that of the IMF for the Euro area, the framework remains similar but the meaning of the indexes change. This case can be described as follows. Y_t denotes the Euro area inflation. $\widehat{Y}_{1,it}$

is the forecast of the i-th forecaster for period t and $\widehat{Y}_{2,t}$ is the IMF forecast for period t. The loss differentials are then $\Delta L_{it} = L(\widehat{Y}_{1,it}, Y_t) - L(\widehat{Y}_{2,t}, Y_t)$ which still depend on the two indexes i and t. As far as its assumptions on the loss differentials are satisfied, our framework is applicable in these situations. In the following section, we give further examples in detail which justify the practical importance of testing the C-EPA null with unknown clusters.

The null hypothesis $\mathcal{H}_0$ implies that $|\mathcal{C}_k|^{-1} \sum_{i \in \mathcal{C}_k} \mathbb{E}(\widetilde{H}_{i,t-\tau} \Delta L_{it}) = 0$, for any measurable-$\mathcal{E}$ vector of random variables $\widetilde{H}_{it}$ (Giacomini and White, 2006). Here, by taking expectations with respect to the vector of measurable-$\mathcal{E}$ vector $\widetilde{H}_{i,t-\tau}$, we obtain an unconditional moment condition. Let H_{it} be such a $P \times 1$ vector, called a "testing function" by Giacomini and White (2006) and $Z_{it} = H_{i,t-\tau} \Delta L_{it}$ with $\mu_i^0 = \mathbb{E}(Z_{it})$. Define also $\theta_k^0(\mathcal{C}) = |\mathcal{C}_k|^{-1} \sum_{i \in \mathcal{C}_k} \mu_i^0$ where $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$. Then, $\mathcal{H}_0$ implies

$$\mathcal{H}_0' : \theta_k^0(\mathcal{C}) = 0, \text{ for all } k = 1, 2, \ldots, K. \tag{3}$$

This transformation from a conditional to an unconditional moment is standard in forecast evaluation and GMM-type testing frameworks. It enables tractable estimation and inference without explicitly modeling the conditioning σ-field $\mathcal{F}_{t-\tau}$. While this approach does not retain all the information contained in the full conditional distribution of ΔL_{it}, it preserves enough structure for hypothesis testing provided that the specified test function is sufficiently informative. In practice, the choice of $H_{i,t-\tau}$, for instance lagged loss differentials, regressors, or cluster-specific moments, alters the power properties and interpretation of the resulting test.

2.2. Examples

The usefulness of testing the conditional EPA hypothesis has been widely documented in the literature starting with Giacomini and White (2006) (see also the excellent review by Clark and McCracken, 2013). We now present examples highlighting the importance of accounting for unknown clusters when testing the C-EPA hypothesis.

Example 1: Time series forecasting. In time series forecasting, it is common to compare the predictive accuracy of a benchmark model, such as an AR(1), against alternative specifications with additional flexibility. For example, Marcellino et al. (2006) compare direct and iterated autoregressive forecasts across a wide set of macroeconomic variables.

Suppose we observe N bivariate time series $\{Y_{it}, X_{it}\}_{t=0}^{T}$. The true data-generating process (DGP)

of the series belongs to one of two latent clusters:

$$Y_{it} = \begin{cases} \alpha_i + \beta_i X_{i,t-1} + U_{it}, & i \in \mathcal{C}_1, \\ \beta_i X_{i,t-1} + U_{it}, & i \in \mathcal{C}_2, \end{cases}$$

where $U_{it} \sim iid(0, \sigma^2)$ and assume that the predictor is fixed and $X_{i,T}$ is observed.

Two forecasters have imperfect knowledge of the DGP and make the following forecasts:

$$\text{Forecaster 1:} \quad \widehat{Y}^{(1)}_{i,T+1} = \hat{\alpha}_i + \hat{\beta}_i X_{i,T},$$
$$\text{Forecaster 2:} \quad \widehat{Y}^{(2)}_{i,T+1} = \tilde{\beta}_i X_{i,T}.$$

The least squares estimators $\hat{\alpha}_i$, $\hat{\beta}_i$, and $\tilde{\beta}_i$ are computed from a fixed estimation window and are therefore subject to sampling variability.

Each forecaster performs well on one cluster and poorly on the other. Specifically, for units in $\mathcal{C}_1$, Forecaster 1 correctly specifies the model by including an intercept, while Forecaster 2 omits this term and incurs bias. In contrast, for units in $\mathcal{C}_2$, the true DGP has no intercept, and Forecaster 2 is correctly specified, while Forecaster 1 overfits by including a superfluous constant term.

This setup yields systematic differences in forecast accuracy across clusters. In Appendix A we derive the expected loss differential between the two forecasters, conditional on the true cluster membership, which is given by

$$\frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \{\mathbb{E}[(\widehat{Y}^{(1)}_{i,T+1} - Y_{i,T+1})^2] - \mathbb{E}[(\widehat{Y}^{(2)}_{i,T+1} - Y_{i,T+1})^2]\} = \begin{cases} \dfrac{1}{|\mathcal{C}_1|} \displaystyle\sum_{i \in \mathcal{C}_1} [\mathbb{V}(\hat{\alpha}_i) + \mathbb{B}(\hat{\alpha}_i)^2 - \alpha_i^2 + \Delta_i], \\ \dfrac{1}{|\mathcal{C}_2|} \displaystyle\sum_{i \in \mathcal{C}_2} [\mathbb{V}(\hat{\alpha}_i) + \mathbb{B}(\hat{\alpha}_i)^2 + \Delta_i], \end{cases} \tag{4}$$

where $\Delta_i := [\mathbb{V}(\hat{\beta}_i) - \mathbb{V}(\tilde{\beta}_i) + \mathbb{B}(\hat{\beta}_i)^2 - \mathbb{B}(\tilde{\beta}_i)^2] X_{i,T}^2 + 2X_{i,T}\text{Cov}(\hat{\alpha}_i, \hat{\beta}_i)$ with $\mathbb{B}(\cdot)$ denoting the bias of an estimator.

This decomposition illustrates how heterogeneity in model specification and estimation precision across clusters induce systematic differences in forecast performance. These loss differentials motivate the C-EPA hypothesis as a natural testable implication of latent cluster structure in predictive performance.

Example 2: Panel data forecasting. Latent group structures became popular in panel data analysis in the last decade (see Bonhomme and Manresa, 2015; Su et al., 2016; Ando and Bai, 2017; Lumsdaine et al., 2023). Suppose that two forecasters are interested in a variable Y_{it} whose DGP is

given by

$$Y_{it} = \beta_{k_i}' X_{i,t-1} + U_{it}, \quad U_{it} \sim iid(0, \sigma^2), \quad k_i \in \{1, \ldots, K\}.$$

We assume that the vector of predictors $X_{i,t-1}$ is known and fixed, and that the forecast errors U_{it} are independent of all regressors and estimators.

Two forecasters make the following two forecasts:

$$\begin{aligned} \text{Forecaster 1:} \quad & \widehat{Y}_{i,T+1}^{\text{pooled}} = \hat{\beta}' X_{i,T}, \\ \text{Forecaster 2:} \quad & \widehat{Y}_{i,T+1}^{\text{het}} = \hat{\beta}_i' X_{i,T}. \end{aligned}$$

While the pooled estimator $\hat{\beta}$ suffers from misspecification bias if $\beta_{k_i} \neq \beta$, the individual estimator $\hat{\beta}_i$ is unbiased but suffers from increased variance due to limited time series observations.

Under standard regularity conditions, the cluster-level expected loss differential is

$$\begin{aligned} &\frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \{\mathbb{E}[(\widehat{Y}_{i,T+1}^{\text{pooled}} - Y_{i,T+1})^2] - \mathbb{E}[(\widehat{Y}_{i,T+1}^{\text{het}} - Y_{i,T+1})^2]\} \\ &\qquad = [\mathbb{E}(\hat{\beta}) - \beta_k]' \Sigma_X [\mathbb{E}(\hat{\beta}) - \beta_k] + \text{tr}\{[\mathbb{V}(\hat{\beta}) - \overline{\mathbb{V}(\hat{\beta}_i)}] \Sigma_X\}, \end{aligned} \tag{5}$$

where $\Sigma_X = |\mathcal{C}_k|^{-1} \sum_{i \in \mathcal{C}_k} X_{i,T} X_{i,T}'$ is the empirical second moment matrix of the regressors in cluster $\mathcal{C}_k$, and $\overline{\mathbb{V}(\hat{\beta}_i)} = |\mathcal{C}_k|^{-1} \sum_{i \in \mathcal{C}_k} \mathbb{V}(\hat{\beta}_i)$ is the average variance of the unit-specific estimators. The proof of this expression is given in Appendix A. This highlights how clusters with large group-level heterogeneity induce systematic differences in forecast performance across units.

Example 3: Forecasting with machine learning methods. Machine learning methods are becoming increasingly popular in economic applications (see, for instance, Athey (2018) for a discussion and Haghighi et al. (2025) for the recent special issue of the Journal of Econometrics). In high dimensional forecasting tasks, researchers often compare linear methods such as Lasso with nonlinear alternatives like random forests. For instance, Goulet Coulombe et al. (2022) compared a large set of data-rich and data-poor models. The authors found that the main advantage of machine learning methods for macroeconomic forecasting is their ability to capture nonlinearities associated with macroeconomic uncertainty, financial stress and housing bubbles. Two methods are trained and evaluated using validation MSE:

- Method 1: linear forecast (e.g., Lasso),
- Method 2: nonlinear forecast (e.g., random forests).

If some units display nonlinear patterns while others do not, the average MSE over the panel units may

be misleading. Then, to check the relative performance of two different machine learning methods, one might need to apply a second machine learning method, namely clustering, because clustering forecast loss differentials and testing the C-EPA null allows for identification of cluster-level model dominance. If it was not reserved for another econometric method, we would call the application our testing framework to this case "double machine learning."

2.3. 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 $\mathcal{F}_t = \{\emptyset, \Omega\}$, the C-EPA null reduces to $|\mathcal{C}_k|^{-1} \sum_{i \in \mathcal{C}_k} \mathbb{E}(\Delta L_{it}) = 0$ for all $k = 1, 2, \ldots, K$ 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 $\mathcal{F}_t \neq \{\emptyset, \Omega\}$ together with a small sample adjustment.

Consider the following test statistic for (3):

$$W(\mathcal{C}) = \frac{B - KP + 1}{KPB} T \hat{\theta}'(\mathcal{C}) \widehat{\Omega}^{-1}(\mathcal{C}) \hat{\theta}(\mathcal{C}), \tag{6}$$

where $\hat{\theta}(\mathcal{C}) = [\hat{\theta}_1'(\mathcal{C}), \ldots, \hat{\theta}_K'(\mathcal{C})]'$ with $\hat{\theta}_k(\mathcal{C}) = (|\mathcal{C}_k|T)^{-1} \sum_{i \in \mathcal{C}_k} \sum_{t=1}^{T} Z_{it}$ and $\widehat{\Omega}(\mathcal{C})$ is an orthonormal series (OS) variance-covariance estimator defined as follows

$$\begin{aligned} \widehat{\Omega}(\mathcal{C}) &= \frac{1}{B} \sum_{j=1}^{B} \widehat{\Lambda}_j(\mathcal{C}) \widehat{\Lambda}_j'(\mathcal{C}), \\ \widehat{\Lambda}_j(\mathcal{C}) &= \sqrt{\frac{2}{T}} \sum_{t=1}^{T} [\bar{Z}_t(\mathcal{C}) - \hat{\theta}(\mathcal{C})] \cos\left[\pi j \left(\frac{t - 1/2}{T}\right)\right], \end{aligned} \tag{7}$$

with $\bar{Z}_t(\mathcal{C}) = [\bar{Z}_{1,t}'(\mathcal{C}), \ldots, \bar{Z}_{K,t}'(\mathcal{C})]'$, $\bar{Z}_{k,t}(\mathcal{C}) = |\mathcal{C}_k|^{-1} \sum_{i \in \mathcal{C}_k} Z_{it}$, and B is the number of orthonormal basis functions used in its estimation. The first factor in (6), $(B - KP + 1)/KPB$, is a small sample correction obtained through the connection between Hotelling's T^2 distribution and the F-distribution, and using the asymptotic property of the proposed variance estimator.

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 $W(\mathcal{C}) \xrightarrow{d} \mathbb{F}_{KP, B-KP+1}$ 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 $B \longrightarrow \infty$, a generalization of the usual results of APUY hold such that $W(\mathcal{C}) \xrightarrow{d} \chi^2_{KP}$. The results of Sun (2013) show that when B is

not too large, using the $\mathbb{F}_{KP,B-KP+1}$ critical values instead of (scaled) χ^2_{KP} critical values results in better size properties. We leave the formal discussion of the theoretical and numerical results to next sections.

With some abuse of notation, let $p[w(\mathcal{C})] = \mathbb{P}_{\mathcal{H}_0}[\mathbb{F}_{KP,B-KP+1} \geq w(\mathcal{C})]$ be the p-value associated with $w(\mathcal{C})$. Here, as in the rest of the paper, we do not show the dependency of $p[\cdot]$ 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 $p[w(\mathcal{C})] \leq \alpha$ where $\alpha \in (0,1)$ is the predetermined Type I error rate.

2.4. Panel Kmeans estimator

If there is no a priori information on the clusters $\mathcal{C}_k$, $k = \{1, \ldots, K\}$, one may use the Panel Kmeans estimator applied to the stacked panel $Z = (Z'_{11}, Z'_{12} \ldots, Z'_{NT})'$, denoted $\mathcal{C}(Z)$, to learn these clusters from the data. For a given K, the Panel Kmeans estimators of the cluster membership sets and cluster centers are defined respectively as:

$$\begin{aligned}(\widehat{\mathcal{C}}_1, \ldots, \widehat{\mathcal{C}}_K) &= \underset{(\mathcal{C}_1,\ldots,\mathcal{C}_K)}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} \left\| Z_{it} - \frac{1}{|\mathcal{C}_k|T} \sum_{j\in\mathcal{C}_k}\sum_{s=1}^{T} Z_{js} \right\|^2, \\ \hat{\theta}_k(\widehat{\mathcal{C}}) &= \frac{1}{|\widehat{\mathcal{C}}_k|T} \sum_{i\in\widehat{\mathcal{C}}_k}\sum_{t=1}^{T} Z_{it}.\end{aligned} \tag{8}$$

The optimization problem in (8) is typically solved by an iterative algorithm, similar to those proposed by Lloyd (1982) or Hartigan (1975). The Panel Kmeans estimates of the cluster membership variables and the cluster centers can be calculated using Algorithm 1 which is a generalization of that of Lloyd's.

Algorithm 1 is a generalization of Lloyd's classical Kmeans clustering algorithm in several respects. Moreover, this generalization has important consequences in our framework. First, Lloyd's algorithm clusters the observations by minimizing within-cluster Euclidean distances to static centroids whereas the Panel Kmeans algorithm clusters units based on their entire time series profile. Specifically, it minimizes the total within-cluster sum of squared deviations over time, thereby extending the clustering criterion to sequences rather than points. This introduces a temporal structure absent in classical Kmeans, while preserving the iterative structure of centroid updating and cluster reassignment. Second, like classical Kmeans, the Panel Kmeans algorithm solves a non-convex minimization problem.

Algorithm 1: Panel Kmeans

Input: Data matrix $Z = (Z'_{11}, Z'_{12} \ldots, Z')'$, number of clusters K

Output: Cluster assignments k_i, cluster centers θ_k

Initialize $\theta_k^{(0)}$ for $k = 1, \ldots, K$; set $m \leftarrow 0$;

repeat

  for $i \leftarrow 1$ **to** N **do**

    $k_i^{(m+1)} \leftarrow \arg \min_{k \in \{1,\ldots,K\}} \sum_{t=1}^{T} \|Z_{it} - \theta_k^{(m)}\|^2$;

  for $k \leftarrow 1$ **to** K **do**

    Update cluster $\mathcal{C}_k^{(m+1)} \leftarrow \{i : k_i^{(m+1)} = k\}$;

    $\theta_k^{(m+1)} \leftarrow \frac{1}{|\mathcal{C}_k^{(m+1)}|T} \sum_{i \in \mathcal{C}_k^{(m+1)}} \sum_{t=1}^{T} Z_{it}$;

  $m \leftarrow m + 1$;

until $k_i^{(m)} = k_i^{(m-1)}$ *for all* $i = 1, \ldots, N$;

The objective function is piecewise quadratic and discontinuous in the assignment variables, leading to the possibility of converging to local minima. This motivates the use of multiple random initializations. Third, the key challenge in the generalization of the selective inference framework for classical Kmeans developed by Chen and Witten (2023) to the Panel Kmeans lies in the dependency structure of the data. Observations for a given unit are temporally dependent and potentially cross-sectionally correlated, making standard theoretical arguments more delicate.

Different initialization methods for Algorithm 1 exist. Chen and Witten (2023) initializes the Kmeans algorithm by choosing K random cluster centers from data. Then they condition on the initial assignments based on these centers as well as each cluster assignment in the Kmeans iterations. In our setting, we first assign each unit randomly to a cluster and then calculate the corresponding cluster centers. Hence, first cluster centers are not chosen to minimize a distance metric. This is a subtle but important difference. The method of Chen and Witten (2023) results in two sets of analytical formulae: one for the initialization and one for the canonical assignments. In our method, we rely only on the second because the initialization does not use a distance metric. Hence, truncation set calculations we provide in Appendix D are simpler than those of Chen and Witten (2023).

Remark 3. Our selective inference framework can, in principle, be extended to models with group fixed effects (GFE) varying over time. Specifically, suppose each unit's outcome is a $P \times 1$ vector Z_{it}, and follows the model $Z_{it} = \mu_{k_i,t} + V_{it}$, where $\mu_{k,t}$ is a time-varying grouped fixed effect and V_{it} is the innovation. The testing problem then concerns the C-EPA null hypothesis defined by

$\mathcal{H}_0 : T^{-1}\sum_{t=1}^{T}\mu_{k,t} = 0$ for all k. This is an important extension because it allows instabilities in relative forecast superiority over time while focusing still on the average equivalence. In a time series setting, Harvey et al. (2024) handled this problem by nonparametric local demeaning to estimate the LRV whereas this GFE modeling strategy may simplify obtaining a consistent variance estimator by assuming that the heterogeneity of the instability is fixed and low dimensional with respect to N.

In this case, a multidimensional GFE generalization of the Panel Kmeans clustering algorithm continues to yield a polyhedral selection region in $\mathbb{R}^{NTP}$. This generalization preserves the logic of the polyhedral approach but potentially introduces new computational burdens as the truncation region grows in complexity and evaluating exact p-values requires new techniques. Developing efficient and scalable inference methods in this multivariate GFE setting is an important direction for future work.

3. Assumptions and Two Useful Lemmas

In this section, we present the assumptions and two preliminary results that will be instrumental in developing the asymptotic theory for the proposed testing procedures. The formulation of these assumptions requires some additional notation. Throughout this paper, we use C to denote a generic positive constant, and we write $(T, N) \to \infty$ to indicate the joint divergence of both dimensions. Asymptotic results under $(T, N) \to \infty$ are understood to hold for any sequence $N = N(T)$ such that $N(T)$ is increasing in T and diverges as $T \to \infty$. We also define $V_{it} = Z_{it} - \mu_i^0$, where $V_{p,it}$, for $p = 1, \ldots, P$, denotes the pth element of the vector V_{it}.

The following assumptions will be referred to throughout the paper. They are grouped into two: first three are the generic assumptions, labeled as G#, which are required for both size and power properties of the tests, and the other three are specific assumptions, labeled as S#, required only for the power properties of the test under the alternative hypothesis $\mathcal{H}_1$.

Assumption G1. (a) $\|\mathbb{E}(\mu_i^0)\| < \infty$, (b) $\mathbb{E}\|V_{it}\|^4 \leq C$, (c) $T^{-1}\sum_{t,s=1}^{T}\mathbb{E}\|V_{it}V_{is}'\| \leq C$.

Assumption G2. $|\mathcal{C}_k|/N \longrightarrow \pi_k \in (0, 1)$ for each $k = 1, \ldots, K$ as $N \longrightarrow \infty$.

Assumption G3. V_{it} is weakly stationary for all $i = 1, \ldots, N$ with $\Omega_i = \sum_{j=-\infty}^{\infty}\mathbb{E}[V_{it}V_{i,t-j}']$ being positive definite and either of the following two holds:

(a) $\mathbb{E}(|V_{p,i1}|^\zeta) < \infty$ for all $p = 1, \ldots, P$ and for $\zeta \geq 2$, V_{it} is φ-mixing with $\sum_{l=1}^{\infty}\varphi_l^{1-1/\zeta} < \infty$,

(b) $\mathbb{E}(|V_{p,i1}|^\zeta) < \infty$ for all $p = 1, \ldots, P$ and for $\zeta > 2$, V_{it} is α-mixing with $\sum_{l=1}^{\infty}\alpha_l^{1-2/\zeta} < \infty$.

Assumption S1. $\mu_i^0 = \theta_k^0$ for all $i \in \mathcal{C}_k^0$ and $k = 1, \ldots, K^0$, where θ_k^0 is the true cluster center of the kth cluster and $\mathcal{C}_k^0$ is the set of units belonging to the true kth cluster.

Assumption S2. Let $K^0 \geq 2$. Then for all $k, g \in \{1, \ldots, K^0\}$, $k \neq g$, there exists $C_{k,g} > 0$ such that $\|\theta_k^0 - \theta_g^0\|^2 \geq C_{k,g}$.

Assumption S3. There exist constants $a_1 > 0$ and $b_1 > 0$ such that, for each $i = 1, \ldots, N$, 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 $\mathbb{P}(|V_{p,it}| > C) \leq e^{1-(C/a_2)^{b_2}}$ for all p, i, t and $C > 0$.

Assumptions G1(a) and G1(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 so that the cluster centers as well as their variances are finite and can be estimated consistently with further regularity conditions. Assumption G1(c) limits the time-series dependence in the sense that $\sum_{t \neq s} \mathrm{E}\|V_{it} V_{is}'\| = O(T)$. We do not place any restriction on the CD characteristics of the panel and allow for both strong and weak CD (see the discussion following Lemma 1 below).

Assumption G2 controls the asymptotic number of units per cluster. 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 to the population. This assumption can be relaxed at the expense of more complicated notation.

Assumption G3 states standard mixing conditions. Here, Ω_i 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, 2022, 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.

Assumption S1 states that the centers of panel units are homogeneous within clusters but heterogeneous between them. Assumption S2 complements the previous assumption by putting a lower bound to the differences between cluster centers. It formalizes implicitly 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 $K^0 = 1$, 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 S3 places additional constraints on the dependence properties and tail probabilities of the process V_{it} over Assumptions G1 and G3. These conditions are imposed for the consistent estimation of cluster membership and the asymptotic equivalence of the cluster center estimators based on Panel Kmeans to the one based on true clusters.

Now we state two useful lemmas which will serve for our theoretical analysis of the test statistics which we will develop. This requires newly introduced notation. Let $\theta_k^0(\mathcal{C}) = |\mathcal{C}_k|^{-1} \sum_{i \in \mathcal{C}_k} \mu_i^0$ be the true center of the k-th cluster implied by the partition $\mathcal{C}$. Define the $KP \times 1$ vector of true cluster centers: $\theta^0(\mathcal{C}) = [\theta_1^{0\prime}(\mathcal{C}), \ldots, \theta_K^{0\prime}(\mathcal{C})]'$, $\Omega(\mathcal{C}) \in \mathbb{R}^{KP \times KP}$ denote the variance-covariance matrix of the vector $\hat{\theta}(\mathcal{C})$ after scaling, that is, $\Omega(\mathcal{C}) = \mathbb{V}\{\sqrt{T}[\hat{\theta}(\mathcal{C}) - \theta^0(\mathcal{C})]\}$, and let $\mathcal{N}(\mathcal{C}) = \text{diag}(|\mathcal{C}_1|, \ldots, |\mathcal{C}_K|) \otimes I_P$. The following result summarizes the usual properties of the sample mean for a fixed $\mathcal{C}$. It will prove useful for our theory even though we focus on estimated clusters because of the fact that we will condition on the estimated cluster for valid C-EPA testing.

Lemma 1. Let $\mathcal{C}$ be a fixed set of cluster memberships and ϵ a fixed real such that $\epsilon \in [1/2, 1]$. Under Assumptions G1–G3, the following results hold as $(T, N) \to \infty$:

(a) $\hat{\theta}(\mathcal{C}) - \theta^0(\mathcal{C}) = o_p(1)$,

(b) $\widetilde{\Omega}(\mathcal{C})^{-1/2} \mathcal{N}(\mathcal{C})^{1-\epsilon} T^{1/2} [\hat{\theta}(\mathcal{C}) - \theta^0(\mathcal{C})] \xrightarrow{d} \mathcal{N}(0, I_{KP})$, where $\widetilde{\Omega}(\mathcal{C}) = \mathcal{N}(\mathcal{C})^{2(1-\epsilon)} \Omega(\mathcal{C})$.

Part (a) is a law of large numbers which shows that Assumptions G1-G3 are sufficient for the consistency of the sample means for the cluster centers defined by a given $\mathcal{C}$. Part (b) is the corresponding central limit theorem. The real scalar $\epsilon \in [1/2, 1]$ measures the degree of CD in V_{it}. The case of $\epsilon = 1$ corresponds to the case of strong CD of the loss differentials as in the factor models; the case of $\epsilon \in [1/2, 1)$ corresponding to the case of weak CD as in the spatial models, or independence across the cross-sections. 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 $\epsilon = (1/2, 1]$.

Remark 4. This general formulation using the parameter ϵ encompasses both strong and weak CD and aligns with models of loss differentials incorporating common factors and spatial interactions as

discussed above. The result concerns with the case of a fixed $\mathcal{C}$ 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.

Lemma 2. Under Assumptions G1-S2 and if $K = K^0$, as $(T, N) \to \infty$,

(a) $\hat{\theta}(\widehat{\mathcal{C}}) - \theta^0 = o_p(1)$.

(b) If Assumption S3 also holds, for all $\xi > 0$, $\mathbb{P}(\sup_{i\in\{1,\ldots,N\}} |\widehat{k}_i - k_i^0| > 0) = o(1) + o(NT^{-\xi})$,

(c) $\hat{\theta}(\widehat{\mathcal{C}}) - \hat{\theta}(\mathcal{C}^0) = o_p(T^{-\xi})$.

(d) If also $N/T^{\xi} \to 0$, $\widetilde{\Omega}(\mathcal{C}^0)^{-1/2}\mathcal{N}(\mathcal{C}^0)^{1-\epsilon}T^{1/2}[\hat{\theta}(\widehat{\mathcal{C}}) - \theta^0] \xrightarrow{d} \mathbb{N}(0, I_{KP})$.

Based on this result, a naive attempt to test the null hypothesis of C-EPA would be to estimate the unknown clusters using the Panel Kmeans estimator and then to use these estimates to construct a Wald test statistic. Let $W(\widehat{\mathcal{C}})$ be the usual Wald test statistic calculated using the Panel Kmeans estimates obtained using the above algorithm. Consider the test which rejects the associated null if $p[w(\widehat{\mathcal{C}})] \leq \alpha$ for some $\alpha \in (0, 1)$. 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 shown by Gao et al. (2024), Patton and Weller (2023), and Chen and Witten (2023). This occurs because the selection of clusters is a data-dependent procedure that implicitly favors detecting heterogeneity, even under the null. When clustering is applied under the null, the algorithm will typically partition the data to minimize within-cluster loss, resulting in estimated cluster means that are artificially separated. This induces a form of selection bias in the test statistic, leading to severe inflation in Type I error rates if this selection is not accounted for. As explained in Section 4 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 with simulations in Section 5.

4. Tests of Generalized C-EPA with Unknown Clusters

In this section, we develop a valid test for the C-EPA null hypothesis when clusters are estimated via the Panel Kmeans algorithm. As it is mentioned in the previous section, using the estimated

clusters for testing in a naive manner results in over rejection of the null hypothesis. Here, a selective conditional inference approach will be employed to control for the Type I error rate by conditioning on the estimated clusters. We consider an approach based on sample splitting in Appendix B.

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

$$\mathcal{H}_0^{homo} : \{\theta_k^0(\mathcal{C}) = \theta_g^0(\mathcal{C})\} \text{ for all } k, g \in \{1, \ldots, K\},\ k \neq g, \tag{9}$$

being the homogeneity hypothesis and

$$\mathcal{H}_0^{oepa} : \frac{1}{N} \sum_{k=1}^{K} |\mathcal{C}_k| \theta_k^0(\mathcal{C}) = 0, \tag{10}$$

the O-EPA hypothesis, as named by APUY where the overall predictive performance is represented as a weighted average of cluster means. We note that the parameter of interest in the O-EPA hypothesis is invariant to the clusters chosen.

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 under known clusters. 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 4.1, we first develop a selective conditional inference framework to test the homogeneity of a pair of clusters selected by Panel Kmeans. Then, we propose a p-value combination test of $\mathcal{H}_0^{homo}$. In Section 4.2, an O-EPA test and the main test statistic of $\mathcal{H}_0$ are presented. Finally, in Section 4.3, we cover the case of an unknown number of clusters and develop a method for its estimation.

4.1. Testing the null of homogeneity

We develop a test for (9). First, tests for each pairwise equality sub-hypothesis is developed and their theoretical properties are presented. Then a homogeneity test is developed via a p-value combination method.

Testing pairwise equality. The homogeneity null $\mathcal{H}_0^{homo}$ is the intersection of $n_p = K(K-1)/2$ unique pairwise equality hypotheses. For each of these pairwise equality nulls, we define the test

statistic $D_{k,g}(\widehat{\mathcal{C}})$ as the square root of the associated Wald test statistic. That is,

$$D^2_{k,g}(\widehat{\mathcal{C}}) = T[\hat{\theta}_k(\widehat{\mathcal{C}}) - \hat{\theta}_g(\widehat{\mathcal{C}})]'\widehat{\Sigma}^{-1}_{k,g}(\widehat{\mathcal{C}})[\hat{\theta}_k(\widehat{\mathcal{C}}) - \hat{\theta}_g(\widehat{\mathcal{C}})], \tag{11}$$

where

$$\widehat{\Sigma}_{k,g}(\widehat{\mathcal{C}}) = \widehat{\omega}_{k,k}(\widehat{\mathcal{C}}) + \widehat{\omega}_{g,g}(\widehat{\mathcal{C}}) - 2\widehat{\omega}_{k,g}(\widehat{\mathcal{C}}),$$

with $\widehat{\omega}_{k,g}(\widehat{\mathcal{C}})$ being the $\{k,g\}$th $P\times P$ block of $\widehat{\Omega}(\widehat{\mathcal{C}})$. It is easily seen that, under appropriate conditions, $D_{k,g}(\mathcal{C}) \xrightarrow{d} \chi_P$ as $T \longrightarrow \infty$, where χ_P is a random variable distributed as a χ variate with P 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. For a pair of clusters $k, g \in \{1,\dots,K\},\ k \neq g$ a test of $\mathcal{H}_0^{k,g} : \{\theta_k^0(\mathcal{C}) = \theta_g^0(\mathcal{C})\}$ controls the selective Type I error rate asymptotically as $(T,N) \to \infty$ at level $\alpha \in (0,1)$ if

$$\lim_{(T,N)\to\infty} \mathbb{P}_{\mathcal{H}_0}\left[\text{Reject } \mathcal{H}_0 \text{ at level } \alpha \;\middle|\; \bigcap_{i=1}^{N}\{\hat{k}_i(Z) = \hat{k}_i(z)\}\right] \leq \alpha, \tag{12}$$

where $\hat{k}_i(Z)$, $i = 1,\dots,N$ is the output of Algorithm 1 and $\hat{k}_i(z)$ is its sample counterpart associated with the realization z of Z.

The definition states that a valid test of the pairwise equality hypothesis $\mathcal{H}_0^{k,g}$ is the one that controls the selective Type I error rate α given the clusters estimated by the Panel Kmeans algorithm. More specifically, the conditioning event in (12) implies that $\mathcal{H}_0^{k,g}$ 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), characterizing this condition is not trivial but we can instead condition on the clusters estimated at all $m = 1,\dots,M$ steps of the algorithm. Two more terms to be conditioned on will be easily seen through a decomposition of the random matrix Z into a term associated with the test statistic $D_{k,g}(\mathcal{C})$ and a term which is orthogonal to this one. We have the following expression:

$$Z = \Pi_{k,g}Z + D_{k,g}(\mathcal{C})\frac{\nu_{k,g}}{\sqrt{T}\|\nu_{k,g}\|^2}\{\text{dir}[\widehat{\Sigma}^{-1/2}_{k,g}(\mathcal{C})Z'\nu_{k,g}]\}'\widehat{\Sigma}^{1/2}_{k,g}(\mathcal{C}) \tag{13}$$

where

$$\Pi_{k,g} = I - \frac{\hat{\nu}_{k,g}\hat{\nu}'_{k,g}}{\|\hat{\nu}_{k,g}\|^2}$$

is the orthogonal projection matrix onto the subspace orthogonal to $\hat{\nu}_{k,g} = (\hat{\nu}'_{k,g,1}, \ldots, \hat{\nu}'_{k,g,N})'$, $\hat{\nu}_{k,g,i} = \iota_T \hat{\delta}_{k,g,i}$, ι_T being a $T \times 1$ vector of ones and $\hat{\delta}_{k,g,i} = \mathbf{1}\{\hat{k}_i = k\}/|\widehat{\mathcal{C}}_k| - \mathbf{1}\{\hat{k}_i = g\}/|\widehat{\mathcal{C}}_g|$. This equality is derived in Equation (32) of Appendix C. The derivation of the conditional distribution of $D_{k,g}(\widehat{\mathcal{C}})$ given $\widehat{\mathcal{C}}$ will be based on this expression. Now we define the following asymptotic p-value

$$p_\infty[d_{k,g}(\widehat{\mathcal{C}})] = \lim_{(T,N)\to\infty} P_{\mathcal{H}_0}\left[D_{k,g}(\widehat{\mathcal{C}}) \geq d_{k,g}(\widehat{\mathcal{C}}) \;\middle|\; \mathcal{A}\right], \tag{14}$$

for $k, g \in \{1, \ldots, K\}$, where

$$\mathcal{A} = \left\{ \bigcap_{m=1}^{M} \bigcap_{i=1}^{N} \{k_i^{(m)}(Z) = k_i^{(m)}(z)\}, \Pi_{k,g} Z = \Pi_{k,g} z, \operatorname{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\widehat{\mathcal{C}}) Z' \hat{\nu}_{k,g}] = \operatorname{dir}[\widehat{S}_{k,g}^{-1/2}(\widehat{\mathcal{C}}) z' \hat{\nu}_{k,g}] \right\} \tag{15}$$

with $\widehat{S}_{k,g}(\mathcal{C})$ being a realization of $\widehat{\Sigma}_{k,g}(\widehat{\mathcal{C}})$ associated with the realization z of Z.

Some remarks follow. The first condition in (15) is the most crucial to the selective conditional inference framework. It states that the cluster to which each panel unit i is assigned in every iteration m of the Panel Kmeans algorithm using the realization z, namely $k_i^{(m)}(z)$, corresponds to the cluster obtained using Z, that is $k_i^{(m)}(Z)$. 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 Panel Kmeans algorithm applied to the particular realization z in hand. The next two conditions allow us to remove the nuisance parameters $\Pi_{k,g} Z$ and $\operatorname{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\widehat{\mathcal{C}}) Z' \hat{\nu}_{k,g}]$ which appear in (13). Otherwise the conditional distribution of $D_{k,g}(\widehat{\mathcal{C}})$ given $\widehat{\mathcal{C}}$ is not tractable. These are standard conditions in selective conditional inference literature (see Fithian et al., 2017; Gao et al., 2024; Chen and Witten, 2023).

The asymptotic p-value $p_\infty[d_{k,g}(\widehat{\mathcal{C}})]$ is based on the selective conditional 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 $i \neq j$ or $t \neq s$, but only across different variables of the same observation, i.e. between $Z_{p,it}$ and $Z_{c,it}$, the p-th and the c-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 $k \in \{2, \ldots, K\}$ with $K \geq 2$ given, and $B \to \infty$ as $(T, N) \to \infty$ such that $B/T \to 0$. Under Assumptions G1-G3 and $\mathcal{H}_0^{k,g}$, a p-value following the asymptotic principle (14) can be

calculated using

$$p[d_{k,g}(\widehat{\mathcal{C}})] = 1 - F_{\chi_P}[\, d_{k,g}(\widehat{\mathcal{C}}); \mathcal{T}\,], \tag{16}$$

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

$$\mathcal{T} = \left\{ \phi \in \mathbb{R}_{\geq 0} : \bigcap_{m=1}^{M} \bigcap_{i=1}^{N} \{k_i^{(m)}[z(\phi)] = k_i^{(m)}(z)\} \right\}, \tag{17}$$

and

$$z(\phi) = \Pi_{k,g} z + \phi \frac{\hat{\nu}_{k,g}}{\sqrt{T}\|\hat{\nu}_{k,g}\|^2} \{\mathrm{dir}[\widehat{S}_{k,g}^{-1/2}(\widehat{\mathcal{C}}) z' \hat{\nu}_{k,g}]\}' \widehat{S}_{k,g}^{1/2}(\widehat{\mathcal{C}}). \tag{18}$$

The equation in (18) 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 k and g are either pushed towards each other or pulled further apart in the direction of $\widehat{S}_{k,g}^{-1/2}(\widehat{\mathcal{C}}) z' \hat{\nu}_{k,g}$. If $\phi = d_{k,g}(\widehat{\mathcal{C}})$ then $z(\phi) = z$. If $\phi > d_{k,g}(\widehat{\mathcal{C}})$ then the two clusters are pulled apart. If instead $\phi < d_{k,g}(\widehat{\mathcal{C}})$, 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). The change in the clustering algorithm from Kmeans to Panel Kmeans leads to substantial differences in the geometry of the selection region, requiring us to derive new formulae for the truncation sets. We document the steps of the calculation of this p-value in Appendix D through a characterization of the truncation set $\mathcal{T}$ in our case of Panel Kmeans. The following result establishes the asymptotic validity of $p[d_{k,g}(\widehat{\mathcal{C}})]$ for the pairwise null hypothesis $\mathcal{H}_0^{k,g}$ defined in Definition 1.

Proposition 1. Let $k \in \{2, \ldots, K\}$, $K = K^0 \geq 2$ given, and $B \to \infty$ as $(T, N) \to \infty$ such that $B/T \to 0$.

(a) Under Assumptions G1-G3, and $\mathcal{H}_0^{k,g}$,

$$\lim_{(T,N)\to\infty} \mathbb{P}\{\, p[D_{k,g}(\widehat{\mathcal{C}})] \leq \alpha \,\} = \alpha, \ \forall.$$

(b) Suppose now that $K = K^0 \geq 2$, and $N/T^{\xi} \to 0$ for some $\xi > 0$. Under Assumptions G1-S3, and if $\mathcal{H}_0^{k,g}$ fails,

$$\lim_{(T,N)\to\infty} \mathbb{P}\{\, p[D_{k,g}(\widehat{\mathcal{C}})] \leq \alpha \,\} = 1, \ \forall \alpha \in (0, 1).$$

Part (a) of the proposition states that the random variable $p[D_{k,g}(\widehat{\mathcal{C}})]$ satisfies the definition of a p-variable of Vovk and Wang (2020) asymptotically, under the null of pairwise cluster equality. The

p-value $p[d_{k,g}(\widehat{\mathcal{C}})]$ 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 $D_{k,g}(\widehat{\mathcal{C}})$ is consistent whenever $\mathcal{H}_0^{k,g}$ fails. Here, it is required that the number of clusters K is correctly chosen to be equal to K^0. We relax this assumption in Section 4.3 by proposing an information criterion to estimate K^0.

Remark 5. The framework described here can be modified to test the null of significance of each cluster center. Namely, to test $\mathcal{H}_0^k : \theta_k^0(\mathcal{C}) = 0$ for $k \in \{1, \ldots, K\}$, one can consider $D_k^2(\widehat{\mathcal{C}}) = T\hat{\theta}_k(\widehat{\mathcal{C}})'\widehat{\omega}_{k,k}(\widehat{\mathcal{C}})^{-1}\hat{\theta}_k(\widehat{\mathcal{C}})$ and set $\Pi_k = I - \hat{\nu}_k\hat{\nu}_k'/\|\hat{\nu}_k\|^2$ where $\hat{\nu}_k = (\hat{\nu}_{k,1}', \ldots, \hat{\nu}_{k,N}')'$, $\hat{\nu}_{k,i} = \iota_T\hat{\delta}_{k,i}$ and $\hat{\delta}_{k,i} = \mathbf{1}\{\hat{k}_i = k\}/|\widehat{\mathcal{C}}_k|$. The results concerning the statistical properties of the test statistic, in particular the asymptotic truncated distribution, remain seemingly unchanged.

Testing homogeneity. We construct a p-value combination test for the homogeneity null (9) by aggregating the n_p selective p-values $p[D_{k,g}(\widehat{\mathcal{C}})]$ from all unique pairwise equality tests. Following the recent studies of Vovk and Wang (2020) and Vovk et al. (2022) on the M-family of merging functions, our proposed test is based on the generalized mean of order $r \in \mathbb{R} \setminus 0$ defined as:

$$F_r = b_{r,n_p}\left\{\frac{1}{n_p}\sum_{\substack{k,g\in\{1,\ldots,K\}\\ k\neq g}}\{p[D_{k,g}(\widehat{\mathcal{C}})]\}^r\right\}^{1/r} \wedge 1$$

where b_{r,n_p} is a calibration constant chosen to ensure that F_{r,n_p} is a valid p-value under arbitrary dependence among the p-values.

M-family nests classical combination rules as special cases. In particular, the cases of $r = 1$, $r \to 0$ and $r = -1$ correspond to arithmetic mean, geometric mean and harmonic mean, respectively. Furthermore, the Bonferroni p-merging function is obtained as $r \to -\infty$ (Vovk and Wang, 2020). However, not all of these preserve the merging or precision properties under arbitrary dependence, especially for small numbers of p-values. In our selective inference framework where the p-values are dependent due to overlapping clustering and shared data, we choose a value of r within the admissible range $r \in [-\infty, -1)$ to ensure that the resulting M-mean is a valid p-merging function under dependence. With simulation exercises, we found out that this choice provides the best finite-sample accuracy among a large number of other choices considered by Vovk and Wang (2020) and Vovk et al. (2022). Following Proposition 5 of Vovk and Wang (2020), we set $b_{r,n_p} = [r/(r+1)]n_p^{1+1/r}$

for this choice of the interval of r. The resulting homogeneity test statistic is given by:

$$F_{homo,r} = \frac{r}{r+1} n_p^{1+1/r} \left\{ \frac{1}{n_p} \sum_{\substack{k,g \in \{1,\dots,K\} \\ k \neq g}} \{p[D_{k,g}(\widehat{\mathcal{C}})]\}^r \right\}^{1/r} \wedge 1 \tag{19}$$

with $r \in [-\infty, -1)$.

This test statistic belongs to the class of precise merging functions, satisfying both monotonicity and sharpness properties under arbitrary dependence of the input p-values. The normalization factor $[r/(r+1)]n_p^{1+1/r}$ guarantees that the statistic in (19) defines a valid p-value under the global null hypothesis. This is shown in Theorem 2 of Vovk and Wang (2020) and generalized in Theorem 3 of Vovk et al. (2022), where they establish the admissibility and optimality of such M-family-based merging functions. In particular, the proposed $F_{homo,r}$ controls the family-wise Type I error under any form of dependence between the constituent p-values.

Remark 6. Unlike Fisher's method (Fisher, 1925), which assumes independence, or Bonferroni's p-merging function, which is conservative, this choice of merging function maintains optimal Type I control under general dependence structures.

Remark 7. A similar p-merging function was recently used by Spreng and Urga (2023) in a multiple forecast comparison setting. The difference between our proposal and that of the authors lies on the choice of the calibration constant b_{r,n_p}. While Spreng and Urga (2023) sets $b_{r,n_p} = r/(r+1)$, we follow exactly the constant suggested by Proposition 5 of of Vovk and Wang (2020), we set $b_{r,n_p} = [r/(r+1)]n_p^{1+1/r}$ which we found to be resulting in smaller size distortions in our particular framework with a small number of p-values combined.

The asymptotic properties of the test statistic $F_{homo,r}$ are formally stated in the following result.

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

(a) Under Assumptions G1-G3, and $\mathcal{H}_0^{homo}$,

$$\limsup_{(T,N)\to\infty} p(F_{homo,r}) \leq \alpha, \ \forall \alpha \in (0,1).$$

(b) Suppose now that $K = K^0 \geq 2$ and $N/T^\xi \to 0$ for some $\xi > 0$. Under Assumptions G1-S3, and if $\mathcal{H}_0^{homo}$ fails,

$$\lim_{(T,N)\to\infty} \mathbb{P}[\, p(F_{homo,r}) \leq \alpha \,] = 1, \ \forall \alpha \in (0,1).$$

Although non-crucial for the development of our C-EPA test statistic with unknown clusters, the test statistic $F_{homo,r}$ 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 $\mathcal{H}_0^{k,g}$ fails.

4.2. 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

$$W_{oepa} = \frac{B-P+1}{PB} T \bar{Z}_o' \widehat{\Omega}_o^{-1} \bar{Z}_o, \tag{20}$$

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

$$\begin{aligned} \widehat{\Omega}_o &= \frac{1}{B}\sum_{j=1}^{B} \widehat{\Lambda}_{o,j}\widehat{\Lambda}_{o,j}', \\ \widehat{\Lambda}_{o,j} &= \sqrt{\frac{2}{T}}\sum_{t=1}^{T}[\bar{Z}_t - \bar{Z}_o]\cos\left[\pi j\left(\frac{t-1/2}{T}\right)\right]. \end{aligned} \tag{21}$$

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

Proposition 2. Suppose that Assumptions G1 and G3 hold with $\mathcal{C} = (1, \ldots, 1)$, that is $K = 1$. Then, for B fixed as $(T, N) \to \infty$, the following results hold.

(a) Under $\mathcal{H}_0^{oepa}$, $W_{oepa} \xrightarrow{d} \mathbb{F}_{P,B-P+1}$.

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

The test rejects the null of O-EPA if $p(w_{oepa}) = \mathbb{P}_{\mathcal{H}_0}[\, \mathbb{F}_{P,B-P+1} \geq w_{oepa} \,] \leq \alpha$ where $\alpha \in (0,1)$ is the predetermined Type I error rate. When $B = T$ and $P = 1$, 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 $S^{(3)}$ 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 n_p

pairwise equality tests and the O-EPA test:

$$F_{SI,r} = \frac{r}{r+1}(n_p+1)^{1+1/r}\left\{\frac{1}{n_p+1}\sum_{\substack{k,g\in\{1,\ldots,K\}\\ k\neq g}}\{p[D_{k,g}(\widehat{\mathcal{C}})]\}^r + \frac{1}{n_p+1}p(W_{oepa})^r\right\}^{1/r} \wedge 1 \qquad (22)$$

where $r \in [-\infty, -1)$.

The following main result of the paper summarizes the desired asymptotic properties of (22).

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

(a) Under Assumptions G1-G3, and $\mathcal{H}_0$,

$$\limsup_{(T,N)\to\infty} p(F_{SI,r}) \leq \alpha, \ \forall \alpha \in (0,1).$$

(b) Suppose now that $K = K^0 \geq 2$ and $N/T^{\xi} \to 0$ for some $\xi > 0$. Under Assumptions G1-S3, and if either $\mathcal{H}_0^{homo}$ or $\mathcal{H}_0^{oepa}$ fails, then,

$$\lim_{(T,N)\to\infty} \mathbb{P}[\, p(F_{SI,r}) \leq \alpha\,] = 1, \ \forall \alpha \in (0,1).$$

The asymptotic result shows that the proposed selective conditional 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 5 where the simulation results confirm these theoretical expectations.

4.3. 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 (22) for $K = 2, \ldots, K_{max}$ 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:

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

where $\widehat{V}_{it}(K) = Z_{it} - \hat{\theta}_{K,\hat{k}_i}$ with $\hat{\theta}_{K,\hat{k}_i}$ being the solution to (8) with K clusters, and ς is a tuning constant. The IC estimate of the number of clusters is given by

$$\widehat{K}_{IC} = \underset{K \in \{2,\ldots,K_{max}\}}{\arg\min} \; IC(K). \tag{23}$$

For the Split Sample test, this IC can be adapted by using only the training portion of the data. Penalty functions other than the one used here can also be employed (see, for instance Bai and Ng, 2002, for different penalties for estimating 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 $\widehat{K}_{IC}$ is consistent for $K^0 \geq 2$ under Assumptions G1-S2 if N and T diverge at the same rate.

In our simulations we found that the values satisfying $\varsigma = [1.5, 3]$ works well with smaller values tending to over-estimate the number of clusters when the signal in the data is weak. The upper bound $\varsigma = 3$ is also suggested by the results of Lumsdaine et al. (2023). As in our particular setting homogeneity testing is embedded in the framework, we set $\varsigma = 1.5$ which sacrifices some precision by over-estimating the true number.

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 conditional inference tests. This is because the computation of the conditioning set $\mathcal{T}$ is time consuming, and contrary to the Bonferroni p-value, an information criterion requires only the Panel Kmeans estimates for different values of K and not $\mathcal{T}$.

An alternative to IC approach is cross-validation (CV) (Li et al., 2025). CV works as follows. The data is repeatedly split into training and validation sets, and for each possible number of clusters K, the within-cluster prediction error on the validation set is evaluated using parameters estimated from the training set. The number of clusters that minimizes the average out-of-sample prediction error across folds is then the estimated number of clusters which we denote $\widehat{K}_{CV}$. An attractive feature of CV is that it is data-driven and does not require tuning parameters, unlike the IC (23). However, it is computationally more intensive, especially when used in conjunction with procedures like Selective Inference. For this reason, we employ CV only in our empirical application, while relying on the IC estimates for our simulation exercises.

An important concern about using a data-dependent choice of the number clusters is that it might invalidate the selective conditional inference procedure because one might need further conditioning on the particular choice of the information criterion. For instance, while developing valid inference pro-

cedures on Lasso, choosing the tuning parameter of the objective function requires extra conditioning (Markovic et al., 2017). The following result shows that this is not the case in our framework.

Proposition 3. Let $\widehat{\mathcal{C}}$ be a clustering with K clusters and assume that $\widehat{\mathcal{C}}$ is the unique output of Algorithm 1. The, inference procedures that condition on the clustering assignment $\widehat{\mathcal{C}}$ implicitly condition on $\widehat{K}_{IC}$ as well. That is, for any test statistic T,

$$\mathbb{P}\left[D_{k,g}(\widehat{\mathcal{C}}) \in \mathcal{T} \,\middle|\, \bigcap_{i=1}^{N}\{\hat{k}_i = k_i\}\right] = \mathbb{P}\left[D_{k,g}(\widehat{\mathcal{C}}) \in \mathcal{T} \,\middle|\, \widehat{K}_{IC} = K, \bigcap_{i=1}^{N}\{\hat{k}_i = k_i\}\right],$$

where $\widehat{K}_{IC}$ is given by (23).

Remark 8. The important assumption of the proposition is that $\widehat{\mathcal{C}}$ is the unique output of Algorithm 1 for given K. As mentioned in the discussion following Algorithm 1, the iterative optimization method does not guarantee the uniqueness of $\widehat{\mathcal{C}}$. In practice, to converge to the global minimum one has to use a large number of initializations to implement the optimization via Algorithm 1. This justifies the use of the IC estimate $\widehat{K}_{IC}$ but raises another question: do we need further conditioning while using multiple initializations to find the best fitting partition of the data? Intuitively this is not the case because our selective conditional framework provides Type I error control in the sense of Definition 1 uniformly over the space of initial partitions. We leave the formal treatment of this question to future work noting that our simulation study supports our conjecture.

5. Monte Carlo Study

We study the finite sample size and power properties of the test statistics. In Section 5.1 we describe the Monte Carlo design and in Section 5.2 we report and comment on the results.

5.1. Design

To investigate the finite sample properties of the testing procedures, we generate observations from a panel AR(1) process given by:

$$Y_{it} = \alpha(1-\rho_{k_i}) + \rho_{k_i} Y_{i,t-1} + U_{it}, \quad U_{it} \sim iid\,\mathbb{N}(0,1). \tag{24}$$

This DGP, as well as our setup that we describe below, is similar to that of Hoga and Dimitriadis (2023) except that their focus is on measurement errors in the target variable whereas ours is on clustered heterogeneity.

There are forecasters, indexed by $a = 1, 2$, who do not observe the true data-generating process but aim to construct one-step-ahead forecasts of Y_{it}. Forecaster 1 includes an intercept but also makes noisy forecasts. Whereas Forecaster 2 omits the intercept altogether. Their forecasting models are given by:

$$\begin{aligned} &\text{Forecaster 1: } \widehat{Y}_{1,it} = \alpha(1-\rho_{k_i}) + \rho_{k_i} Y_{i,t-1} + \varepsilon_{it}, \\ &\text{Forecaster 2: } \widehat{Y}_{2,it} = \rho_{k_i} Y_{i,t-1}, \end{aligned} \tag{25}$$

for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, where $k_i \in \{1, 2, 3\}$ denotes the latent cluster membership of unit i. For computational efficiency and following Hoga and Dimitriadis (2023), we assume that they use the true slope parameter, and the true intercept, if included. This is justified with the noise term in the first set of forecasts which might be due to over parametrization of heterogeneity in real applications, and the omission of the intercept in the second set of forecasts.

The noise term ε_{it} is constructed to preserve zero mean and a cluster-specific forecast variance. It evolves according to the following stationary process:

$$\varepsilon_{it} = \phi \varepsilon_{i,t-1} + \lambda F_t + \sqrt{\sigma^2_{\varepsilon,k_i}(1-\phi^2) - \lambda^2} \cdot \xi_{it}, \qquad \xi_{it} \sim iid\, \mathbb{N}(0,1),$$

where $F_t \sim iid\, \mathbb{N}(0,1)$ is a common factor independent of ξ_{it}. The parameter $\phi \in (-1, 1)$ governs the AR(1) persistence of ε_{it} whereas λ controls the strength of CD through the common factor F_t. The variance of the noise of Forecaster 1 for cluster k_i is given by $\sigma^2_{\varepsilon,k_i} = \alpha^2(1-\rho_{k_i})^2 + \psi_{k_i}$. This construction ensures that the forecast noise incorporates both time dependence via ϕ, and CD via λ.

We implement both unconditional and conditional EPA tests. These are associated with the choices $H_{i,t-1} = 1$ and $H_{i,t-1} = (1, Y_{i,t-1})'$, respectively. Set $\Delta L_{it} = (Y_{it} - \widehat{Y}^{(1)}_{it})^2 - (Y_{it} - \widehat{Y}^{(2)}_{it})^2$. Straightforward calculations (see Appendix C of Hoga and Dimitriadis, 2023) show that the expected quadratic loss differentials in these two cases are given by

$$\mathbb{E}(H_{i,t-1}\Delta L_{it}) = \begin{cases} \psi_{k_i}, & \text{if } H_{i,t-1} = 1, \\ (\psi_{k_i}, \mu \cdot \psi_{k_i})', & \text{if } H_{i,t-1} = (1, Y_{i,t-1})'. \end{cases}$$

This expression illustrates that the magnitude of the expected loss differential depends solely on the noise variance in the case of unconditional EPA testing, and the noise variance and the unconditional mean μ in the case of conditional EPA testing.

Parameter choices. In all experiments, we set $\mu = 1$, $\phi = 0.2$ and $\lambda = 0.2$. For the AR(1) dynamics of the observed outcome Y_{it} we assume that the panel units form three latent clusters of heterogeneous

sizes which correspond to the clusters of the loss differentials. We set,

$$k_i^0 = \begin{cases} 1, & \text{if } i \in \{1, \ldots, N/4\}, \\ 2, & \text{if } i \in \{N/4+1, \ldots, N/2\}, \\ 3, & \text{if } i \in \{N/2+1, \ldots, N\}, \end{cases} \tag{26}$$

and $(\rho_1, \rho_2, \rho_3) = (0.1, 0.2, 0.3)$. That is the size of the third cluster is twice the size of the first and the second clusters. To investigate the empirical size of the tests, we set $(\psi_1, \psi_2, \psi_3) = (0, 0, 0)$. The power is analyzed under two cases both with $K^0 = 3$:

- **Case 1– O-EPA hypothesis fails:** $(\psi_1, \psi_2, \psi_3) = \psi/2 + \psi \cdot (-1.2, -0.8, 1)$,
- **Case 2– O-EPA hypothesis holds:** $(\psi_1, \psi_2, \psi_3) = \psi \cdot (-1.2, -0.8, 1)$.

The parameter ψ measures the deviation from the null hypotheses. We consider the values $\psi \in \{0.125, 0.25, 0.375, 0.5\}$.

We investigate the size of the tests for all possible pairs (T, N) such that $N \in \{80, 120, 160\}$ and $T \in \{20, 50, 100, 200\}$. As the testing procedures we propose are computationally costly, we analyze the power for $N = 80$ and $T \in \{50, 200\}$. We note that the loss differentials carry strong CD, hence the number of cross-sectional units do not have an effect on the power of the tests. For the same efficiency reason, we set the number of replications as 1000.

Implementation of the tests. We implement four different types of tests. These are labeled called Predetermined, Naive, Split Sample, and Selective Inference. For each of these we consider the unconditional and conditional tests. The details of the implementation are as follows.

- **Predetermined:** As described in Section 2.3 by setting $k_i = k_i^0$ for all $i = 1, \ldots, N$.
- **Naive:** As described in Section 2.3 by setting $k_i = \hat{k}_i$ for all $i = 1, \ldots, N$ where $\hat{k}_i$ is the output of Algorithm 1.
- **Split Sample:** As described in Appendix B by setting $\mathcal{S}_1 = \{1, \ldots, 0.2 \cdot T\}$ and $\mathcal{S}_2 = \{0.2 \cdot T + 1 + l, \ldots, T\}$ with $l = \lfloor \sqrt{0.2 \cdot T} \rfloor$ to minimize the statistical dependence between the two portions of the sample, and $k_i = \hat{k}_i$ with $\hat{k}_i$ being estimated from $\mathcal{S}_1$ using Algorithm 1.
- **Selective Inference:** As described in Section 4 by setting $k_i = \hat{k}_i$ for all $i = 1, \ldots, N$ where $\hat{k}_i$ is the output of Algorithm 1.

All tests are robust to arbitrary autocorrelation and CD. The number of cosines in the OS estimator of the LRV is chosen as $B = \min(\lfloor PT^{2/3} \rfloor, T)$ in the case of full sample tests and as

$B = \min(\lfloor P|\mathcal{S}_2|^{2/3} \rfloor, |\mathcal{S}_2|)$ for the Split Sample tests. As the latent nature of the clusters is central to our framework, all tests except Predetermined are implemented using $\widehat{K}_{IC}$ given in (23). When relevant, Algorithm 1 is run with 10 random initialization and maximum 100 iterations.

5.2. Results

We report the results in three parts. We comment on the size properties, the power properties and the results of a robustness check to structural breaks in the process.

Size properties. Table 1 reports the rejection rates of four C-EPA testing procedures under the null hypothesis, evaluated at the 5% nominal level. The results are presented separately for unconditional and conditional tests.

The Naive test, which conducts inference as if cluster assignments were known and fixed, exhibits dramatic size distortions across all configurations. Its rejection rate is exactly 1.00 in every design, both in unconditional and conditional setups. This complete failure to control size reflects the well-known danger of ignoring model selection when clusters are estimated from the data.

By contrast, the Predetermined test which uses exogenous, fixed clusters for testing delivers rejection rates close to the nominal level, ranging from 0.04 to 0.07. For instance, when $N = 120$ and $T = 100$, the rejection rate is 0.05 and 0.06 for the unconditional and conditional cases, respectively. This method offers a reasonable benchmark, but its feasibility is limited by the requirement of pre-specified cluster structure.

The Split Sample test shows relatively accurate size control as well, with rejection rates between 0.03 and 0.11. For example, when $N = 160$ and $T = 20$, the rejection rate is 0.07 unconditionally and 0.09 conditionally, slightly above the nominal level but still within acceptable range given the small sample. Its reliance on data splitting reduces bias from re-using the same sample, though at the cost of potential power loss due to reduced sample size for both estimation and testing.

Finally, the Selective Inference test, which corrects for the randomness introduced by cluster estimation via truncation-based conditioning, consistently delivers accurate size control. Rejection rates are always very close to the nominal level 0.05. For example, at $N = 80$ and $T = 50$, the selective test rejects the null at a rate of 0.05 unconditionally and 0.06 conditionally. These results confirm that the proposed selective procedure successfully accounts for the data-dependent nature of cluster formation, without relying on sample splitting or external cluster information.

In summary, the simulations demonstrate that naive inference leads to massive over-rejection, while

Table 1: Rejection rates of C-EPA tests under the null

N	T	Predetermined	Naive	Split Sample	Selective Inference
Unconditional tests ($H_{i,t-1} = 1$)					
80	20	0.07	1.00	0.07	0.05
80	50	0.05	1.00	0.05	0.05
80	100	0.06	1.00	0.06	0.07
80	200	0.05	1.00	0.03	0.05
120	20	0.07	1.00	0.07	0.04
120	50	0.05	1.00	0.07	0.03
120	100	0.05	1.00	0.06	0.04
120	200	0.05	1.00	0.06	0.04
160	20	0.06	1.00	0.07	0.04
160	50	0.06	1.00	0.06	0.04
160	100	0.06	1.00	0.06	0.04
160	200	0.05	1.00	0.04	0.03
Conditional tests ($H_{i,t-1} = (1, Y_{i,t-1})'$)					
80	20	0.05	1.00	0.11	0.05
80	50	0.04	1.00	0.06	0.06
80	100	0.06	1.00	0.06	0.05
80	200	0.05	1.00	0.05	0.05
120	20	0.06	1.00	0.10	0.03
120	50	0.06	1.00	0.07	0.04
120	100	0.06	1.00	0.07	0.04
120	200	0.05	1.00	0.06	0.05
160	20	0.06	1.00	0.09	0.04
160	50	0.06	1.00	0.07	0.04
160	100	0.05	1.00	0.06	0.02
160	200	0.05	1.00	0.04	0.03

Note: Rejection rates are calculated from 1000 Monte Carlo replications under the null hypothesis with nominal size: $\alpha = 0.05$. Predetermined tests are described in Section 2.3 and calculated with $k_i = k_i^0$ given in Equation (26). Naive tests are similar except they use the estimated clusters. Split Sample tests are described in Appendix B and Selective Inference tests in Section 4. All tests are robust to arbitrary autocorrelation and CD. The number of clusters for Naive, Split Sample and Selective Inference tests is determined using Equation (23).

both sample splitting and selective approaches control size effectively. Among the feasible procedures, the selective test offers the most robust and accurate size behavior across a wide range of panel dimensions and test designs.

Power properties. Table 2 reports the rejection frequencies of the four C-EPA procedures under the alternative hypothesis, in a setting where the O-EPA hypothesis fails. As expected, all tests exhibit increasing power as ψ and T grow, but there are important differences in how quickly this increase occurs.

The Naive test continues to reject 100% of the time, regardless of the strength of the alternative or the sample size. The Predetermined test uses the true cluster assignments. It performs well overall but is infeasible. Its performance is good especially for moderate to large deviations from the null. For example, with $T = 50$ and $\psi = 0.125$, it achieves 87% rejection in the unconditional case and 74% in the conditional one. With $T = 200$, power reaches 100% across all configurations. This performance illustrates that when cluster assignments are correctly specified in advance, the test can reliably detect violations of EPA even when they are small.

The Split Sample test shows lower power in small samples, particularly for weak signals. For instance, when $T = 50$ and $\psi = 0.125$, rejection rates are just 20% (unconditional) and 15% (conditional). However, its power improves substantially with longer time series and stronger alternatives: for $T = 200$ and $\psi = 0.25$, rejection rates reach 100% in both cases. This highlights the trade-off inherent in data splitting—robust size control comes at the cost of efficiency in small samples.

The Selective Inference test behaves similarly to the Split Sample test. With $T = 50$ and $\psi = 0.125$, it rejects the null in only 19% of simulations (unconditional) and 16% (conditional), but power increases rapidly with larger T and stronger violations. For example, with $T = 200$ and $\psi = 0.25$, the rejection rate reaches 100%. In addition, we find that its power is consistently higher than that of the Split Sample test in conditional testing. These confirm that while Selective Inference test is conservative in small samples, it is capable of detecting meaningful deviations when sufficient information is available.

In summary, when the O-EPA assumption fails, both the Split Sample and Selective Inference procedures deliver high power while preserving valid size. The predetermined test provides a useful upper bound on power when cluster structure is known. The selective procedure offers a robust alternative that adapts to increasing signal strength without inflating false positives.

Table 3 displays the rejection rates of the four C-EPA tests under the alternative hypothesis when the overall EPA hypothesis holds. In this setting, the deviation from the null occurs within the

Table 2: Rejection rates of C-EPA tests under the alternative: Case 1– O-EPA fails

T	ψ	Predetermined	Naive	Split Sample	Selective Inference
Unconditional tests ($H_{i,t-1} = 1$)					
50	0.125	0.87	1.00	0.20	0.19
200	0.125	1.00	1.00	0.79	0.72
50	0.250	1.00	1.00	0.68	0.62
200	0.250	1.00	1.00	1.00	1.00
50	0.375	1.00	1.00	0.98	0.91
200	0.375	1.00	1.00	1.00	1.00
50	0.500	1.00	1.00	1.00	0.99
200	0.500	1.00	1.00	1.00	1.00
Conditional tests ($H_{i,t-1} = (1, Y_{i,t-1})'$)					
50	0.125	0.76	1.00	0.15	0.16
200	0.125	1.00	1.00	0.68	0.71
50	0.250	1.00	1.00	0.50	0.58
200	0.250	1.00	1.00	1.00	1.00
50	0.375	1.00	1.00	0.88	0.91
200	0.375	1.00	1.00	1.00	1.00
50	0.500	1.00	1.00	0.99	0.99
200	0.500	1.00	1.00	1.00	1.00

Note: Rejection rates are calculated from 1000 Monte Carlo replications under the alternative hypothesis for different values of ψ which measures the strength of the deviation from the null. Nominal size: $\alpha = 0.05$ and $N = 80$. See Table 1 notes.

cluster centers, but the global equality of predictive ability holds across clusters. This configuration is particularly relevant for assessing whether inference procedures can detect heterogeneous predictive content even when overall forecast performance is similar across clusters.

The Predetermined test, which assumes known cluster structure, provides an upper bound on feasible power. Its rejection rates are high across all scenarios, reaching 1.00 in nearly every case, including small samples and weak deviations (e.g., $T = 50$, $\psi = 0.125$, the power equals 0.78 unconditionally, 0.65 conditionally). These values confirm that informative deviations are present and detectable with idealized cluster knowledge.

As before, the Naive test rejects in all cases, with power equal to 1.00 even when the deviation is weak. The Split Sample test performs notably well in this scenario. Although power is low for weak deviations and small samples (e.g., $T = 50$, $\psi = 0.125$, power equals only 0.07 both unconditionally

and conditionally), it rises rapidly with increasing signal strength. For instance, with $T = 50$ and $\psi = 0.375$, the Split Sample test reaches 80% (unconditional) and 57% (conditional) power. When $T = 200$, rejection rates exceed 95% for $\psi \geq 0.25$ in both test types. This strong performance reflects the fact that the test can leverage more power when overall EPA holds and cluster selection happens to align well with the underlying structure.

By contrast, the Selective Inference test exhibits lower power in this setting. When deviations are small, rejection rates remain near the nominal level (e.g., $T = 50$, $\psi = 0.125$, power equals 0.06). Even with stronger deviations and longer panels, the increase in power is more gradual. For example, with $T = 200$ and $\psi = 0.375$, the test rejects 31% (unconditionally) and 53% (conditionally); for $\psi = 0.5$, power improves to 64% and 67% respectively. This pattern reflects some key features of the selective procedure. First, by accounting for cluster estimation uncertainty, it sacrifices power in settings where selection aligns with true structure but the null hypothesis is close to being true. Second, it relies on additional conditions due to the nuisance parameters in the conditional distribution which result in lower power than the Split Sample test in certain scenarios. Third, it uses O-EPA test as a component in the p-value combination step. Since O-EPA holds in this case, the power of the resulting C-EPA test is below ideal. However, as we discuss later in the section, it still is the only viable procedure of C-EPA testing in most empirical settings. To conclude, while the test is robust to false positives, it may under-reject in cases where the alternative is subtle and the overall structure is well behaved.

To sum up the findings of this experiment, the Split Sample approach dominates in terms of power when the O-EPA assumption holds, especially for moderate to large deviations. The Selective Inference test remains valid but conservative, offering protection against size distortions at the expense of some power loss. This trade-off highlights a core message of our framework: inference that accounts for model selection can be more reliable, but necessarily faces a trade-off between robustness and sensitivity to weak signals.

Alternative DGPs and additional results. First, we conduct a robustness analysis to draw attention to a situation which is quite realistic in practice where Selective Inference test stands out as the only available method to test the C-EPA hypotheses with unknown clusters. This is when there are breaks in the cluster centers such that even if the C-EPA hypothesis holds, the Split Sample tests grossly over-reject the true null hypothesis. Second, we discuss the small sample properties of the O-EPA and homogeneity tests which shed light on the finding on the power of the Selective Inference tests when O-EPA holds.

Table 3: Rejection rates of C-EPA tests under the alternative: Case 2– O-EPA holds

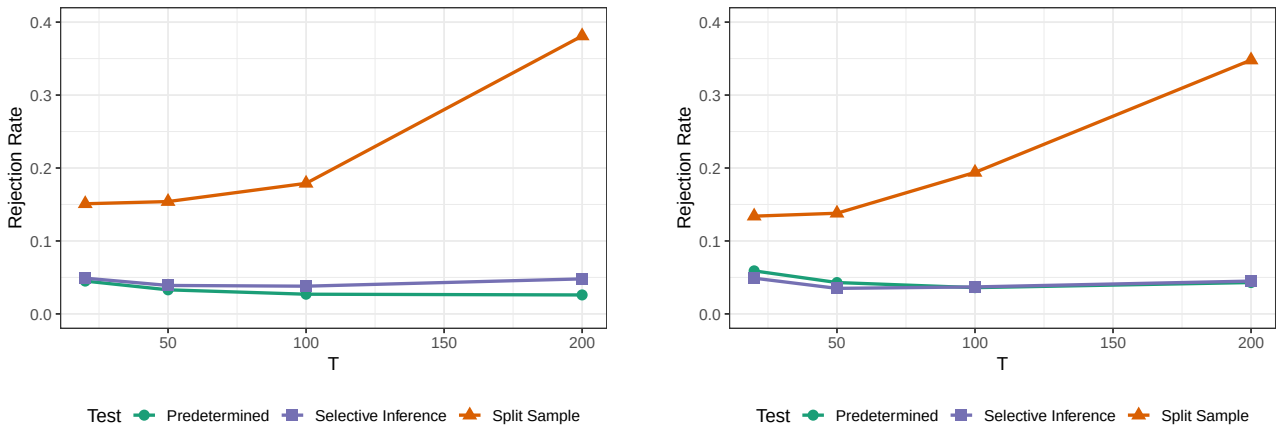
T	ψ	Predetermined	Naive	Split Sample	Selective Inference
Unconditional tests ($H_{i,t-1} = 1$)					
50	0.125	0.78	1.00	0.07	0.06
200	0.125	1.00	1.00	0.32	0.07
50	0.250	1.00	1.00	0.30	0.07
200	0.250	1.00	1.00	1.00	0.18
50	0.375	1.00	1.00	0.80	0.10
200	0.375	1.00	1.00	1.00	0.31
50	0.500	1.00	1.00	0.99	0.14
200	0.500	1.00	1.00	1.00	0.64
Conditional tests ($H_{i,t-1} = (1, Y_{i,t-1})'$)					
50	0.125	0.65	1.00	0.07	0.06
200	0.125	1.00	1.00	0.20	0.08
50	0.250	1.00	1.00	0.20	0.07
200	0.250	1.00	1.00	0.96	0.27
50	0.375	1.00	1.00	0.57	0.11
200	0.375	1.00	1.00	1.00	0.53
50	0.500	1.00	1.00	0.95	0.20
200	0.500	1.00	1.00	1.00	0.67

Note: Nominal size: $\alpha = 0.05$ and $N = 80$. See Table 2 notes.

Figure 1 reports the empirical size of various C-EPA testing procedures under the null hypothesis when the data-generating process includes structural breaks in the relative forecast performance across clusters. The true O-EPA null as well as the C-EPA null hold on average over time period under consideration. In particular, for $t \in \{1, \ldots, T/2\}$ we set $(\psi_1, \psi_2, \psi_3) = \psi/2 + \psi \cdot (-1.2, -0.8, 1)$ as in the main Monte Carlo design Case 1, and for $t \in \{T + 2 + 1, \ldots, T\}$ we set $(\psi_1, \psi_2, \psi_3) = -\psi/2 - \psi \cdot (-1.2, -0.8, 1)$. That is, there is no global improvement in predictive ability.

The figure reveals a stark contrast in the behavior of the testing procedures. The Selective Inference test maintains excellent size control across all sample sizes, with rejection rates consistently close to the nominal 5% level in both unconditional and conditional settings. This confirms that the method appropriately accounts for the randomness introduced by data-driven cluster estimation, even in the presence of structural instability.

In contrast, the Split Sample test shows substantial over-rejection, with empirical size rising sharply



(a) Unconditional tests ($H_{i,t-1} = 1$)

(b) Conditional tests ($H_{i,t-1} = (1, Y_{i,t-1})'$)

Figure 1: Empirical size of C-EPA tests under the null hypothesis ($\alpha = 0.05$) for different time dimensions T. The tests are applied to simulated data with $N = 80$ and $\psi = 0.25$. Each line corresponds to a different version of the test procedure.

with the time dimension T. In the unconditional test, its rejection rate increases from roughly 15% at $T = 20$ to over 35% at $T = 200$. The conditional version follows a similar trajectory. This pronounced size distortion reflects the inability of the split sample test to account for changes in the structure of predictive accuracy. By separating the sample for estimation and testing, the procedure fails to recognize time-varying cluster centers and exaggerates evidence against the null.

The Predetermined test, which assumes known clusters, also exhibits good size control, as expected, but is not feasible in practice when clusters are unknown. The results thus highlight the danger of using Split Sample approaches in the presence of temporal instability, and the value of selective inference procedures that condition properly on the estimated cluster structure using the full sample.

In summary, when the null hypothesis holds but structural breaks induce heterogeneous forecast patterns, the selective inference test is the only feasible method among those considered that maintains reliable control over false positives.

Table 4 presents additional simulation results on the performance of the O-EPA test and the homogeneity test. The results on the size of the tests confirm that all procedures maintain appropriate size control, with rejection rates close to the nominal level of 5%. In the first power scenario, where the O-EPA hypothesis fails, both tests exhibit increasing power with larger signal strength and time dimensions, as expected. The final block reports results under a setting where the O-EPA null holds but clusters are heterogeneous. Importantly, the conditional homogeneity test consistently rejects in

Table 4: Rejection rates of O-EPA and Homogeneity tests

			Unconditional		Conditional	
N	Tobs	ψ	O-EPA	Homogeneity	O-EPA	Homogeneity
			Size			
80	20	0	0.06	0.05	0.06	0.04
80	50	0	0.07	0.05	0.06	0.05
80	100	0	0.06	0.08	0.05	0.05
80	200	0	0.04	0.05	0.04	0.06
120	20	0	0.06	0.03	0.05	0.03
120	50	0	0.05	0.04	0.04	0.05
120	100	0	0.05	0.03	0.05	0.05
120	200	0	0.06	0.03	0.06	0.06
160	20	0	0.07	0.02	0.06	0.03
160	50	0	0.05	0.03	0.05	0.03
160	100	0	0.06	0.04	0.04	0.03
160	200	0	0.04	0.04	0.04	0.03
			Power: Case 1– O-EPA hypothesis fails			
80	50	0.125	0.33	0.27	0.06	0.05
80	200	0.125	0.87	0.81	0.07	0.11
80	50	0.250	0.83	0.73	0.07	0.10
80	200	0.250	1.00	1.00	0.18	0.28
80	50	0.375	0.98	0.97	0.08	0.11
80	200	0.375	1.00	1.00	0.27	0.52
80	50	0.500	1.00	1.00	0.12	0.19
80	200	0.500	1.00	1.00	0.44	0.67
			Power: Case 2– O-EPA hypothesis holds			
80	50	0.125	0.07	0.06	0.06	0.06
80	200	0.125	0.04	0.04	0.07	0.11
80	50	0.250	0.06	0.06	0.07	0.10
80	200	0.250	0.04	0.04	0.19	0.30
80	50	0.375	0.06	0.06	0.11	0.14
80	200	0.375	0.04	0.04	0.33	0.55
80	50	0.500	0.06	0.07	0.15	0.23
80	200	0.500	0.05	0.04	0.65	0.69

Note: O-EPA test is described in Section 4.2 and Homogeneity test is described in Section 4.1. See Table 1-3 for the details on simulation design.

this case, especially for large T, indicating that it remains powerful for detecting latent heterogeneity even when O-EPA is valid. On the other hand, O-EPA test still provides correct Type I error control, as expected. This sheds light on the relatively poor performance of the Selective Inference test of C-EPA in this case: since it combines p-values of pairwise homogeneity tests as well as the O-EPA test, it results in lower power because of this second component.

6. Empirical Application: Forecasting Exchange Rate Returns

This section applies a variety of forecasting methods to monthly exchange rate returns using conventional time series models as well as more modern machine learning methods with and without macroeconomic predictors from the FRED-MD database. The forecasting objective is to predict one-month-ahead log returns of a panel of 131 bilateral exchange rates over dates from January 1999 to December 2023.

6.1. Data preparation

We use monthly bilateral exchange rates from the IMF. The data set spans from January 1999 to December 2023. Although the IMF Exchange Rates data set provides a longer history on some series, we focus on this particular period to obtain a balanced panel. The starting date particularly reflects the availability of Euro/Dollar exchange rate. Log returns are computed as first differences of the natural logarithm of the exchange rate levels, multiplied by 100 to express them in percentage terms. Each series is then standardized. As our objective is model comparison instead of making real forecasts, we do not de-standardize the series before presenting the results. Series with missing observations or near-zero standard deviation are excluded from the analysis. This results in 131 monthly bilateral exchange rates against the US Dollar.

We obtain monthly macroeconomic indicators from the FRED-MD dataset. Variables are transformed using the `tw_apc` procedure with `kmax`$= 8$ of the `fbi` package (Chan et al., 2023), which uses the Tall-Wide method to impute the missing values in a given panel data. To avoid look-ahead bias, each predictor matrix is lagged appropriately within the recursive forecasting window. We remove exchange rate variables that overlap with the dependent variables (`EXSZUSx`, `EXJPUSx`, `EXUSUKx`, `EXCAUSx`).

Forecasts are produced using a recursive window of length $r = 60$ months. For each forecast date $t = r + 1, \ldots, T - 1$, we re-estimate model parameters and predict the return in $t + 1$. The final

sample of forecast errors covers the period from February 2004 to December 2023, that is $T = 238$ and $N = 131$ in our final sample of forecast comparison.

6.2. Forecasting methods

We compare the performance of five forecasting models that span linear and nonlinear approaches, with and without macroeconomic predictors. All models are estimated separately for each exchange rate series using a recursive forecasting design with a fixed window of $r = 60$ months and a one-month-ahead forecast horizon. Forecast accuracy is evaluated via quadratic loss function. For EPA tests, we use quadratic loss differentials relative to the AR(1) benchmark. All the other details on the implementation of the tests correspond exactly to those of the Monte Carlo simulations.

We classify the five methods under consideration into two categories: data poor and data rich methods. We now describe these methods.

Data poor methods. These methods are considered "data poor" in the sense that they rely solely on the history of the dependent variable. The two models we consider are described in what follows.

- **AR(p) selected by BIC:** An autoregressive model with lag length p selected via the Bayesian Information Criterion (BIC).
- **Elastic Net:** A linear penalized regression combining ℓ_1 and ℓ_2 penalties (Zou and Hastie, 2005), applied to the lags of the dependent variable. The method balances variable selection and shrinkage, mitigating overfitting in high dimensional settings. The penalty parameters are selected via 5-fold cross-validation, which is used to jointly determine both the overall regularization strength and the mixing parameter governing the weight between LASSO and Ridge penalties. The model is implemented using the `glmnet` package (Friedman et al., 2010).
- **XGBoost:** An ensemble of gradient-boosted decision trees applied to the the lags of the target variable (Chen and Guestrin, 2016). XGBoost captures nonlinearities and interaction effects by sequentially fitting trees to the residuals of prior iterations. Forecasts are generated using the past 6 lags of the target variable as features. The model is trained for 50 boosting rounds using default hyperparameters and the squared error loss. The model is implemented using the `xgboost` package (Chen and Guestrin, 2016).

For all three models, we allow a maximum lag length of 6. The AR(p) selects the optimal lag within this range using BIC while Elastic Net allows for a more general model structure such that all

consecutive lags do not necessarily appear in the model. XGBoost further allows for nonlinearities in the relationship of the target and its lags. These data poor approaches provide useful baselines to assess the marginal value of more flexible, data-rich machine learning methods.

Data rich methods. These methods are considered "data rich" as they exploit high dimensional information from a large set of macroeconomic predictors. Unlike the data poor models, which rely primarily on univariate dynamics, these methods are designed explicitly to extract predictive signals from complex interactions and nonlinearities in the covariate space. Their flexibility makes them particularly well-suited in environments characterized by structural change, unknown functional forms, or unstable predictor relevance.

- **Support Vector Machine (SVM):** A kernel-based machine learning method applied to macroeconomic features. The SVM solves a regularized minimization problem that fits the data within a margin of tolerance (Smola and Schölkopf, 2004). The implementation uses an ε-insensitive regression formulation with a radial basis function (RBF) kernel. The design matrix includes the first lag of the target variable and the contemporaneous values of the scaled macro predictors. Hyperparameters are selected via cross-validation. We use the `e1071` package to implement the support vector regression with a radial basis function kernel (Meyer et al., 2024).
- **Random Forest:** A nonparametric ensemble method based on bagged decision trees (Breiman, 2001). The model is trained using the lagged target variable and standardized macro predictors as features. Each tree is fit on a bootstrap sample of the training data with random feature selection at each split. The implementation uses the `randomForest` package with default hyperparameters and no tuning. Forecasts are based on the most recent observation of macroeconomic predictors. The model is estimated using the `randomForest` package (Liaw and Wiener, 2002).

The use of default hyperparameters reflects a deliberate emphasis on simplicity and replicability. While further tuning could improve the performance of certain methods, our approach is conservative and avoids complication by applying standard practices such built-in bagging in Random Forests. The resulting forecasts serve as a benchmark for the potential gains from machine learning with a large sample of macroeconomic features. We note that we implemented several other methods such as the factor augmented regressions following the targeted predictors methodology of Bai and Ng (2008) as well as the macro-feature-augmented versions of Elastic Net and XGBoost which resulted in objectively worse performance than the methods we report here. Hence, to save space, we ignore these methods.

6.3. Results

6.3.1. Descriptive analysis

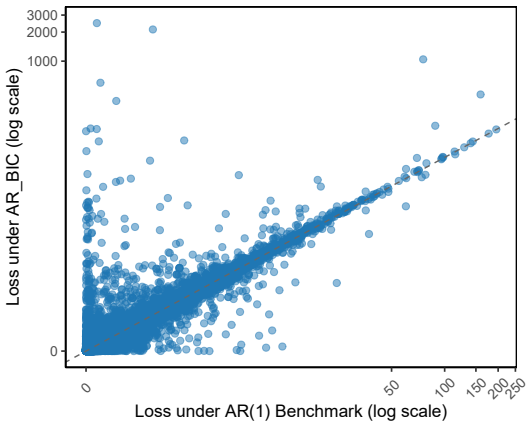
Figure 2 presents log-log scatterplots of forecast losses across a large panel of prediction tasks. The horizontal axis in each panel shows the loss under the AR(1) benchmark, while the vertical axis displays the loss under a competing method. Each point corresponds to a unique predictive task (e.g., variable-horizon-variable combination), allowing for a granular comparison of relative performance.

Panel (a) compares the AR(1) model with an AR(p) model selected via the Bayesian Information Criterion (BIC). While the AR(p) model occasionally outperforms the benchmark, evidenced by points below the 45-degree line, a substantial share of forecasts perform worse. This illustrates the tradeoff between increased model flexibility and estimation uncertainty (Inoue and Kilian, 2006), especially under limited sample sizes or structural instability.

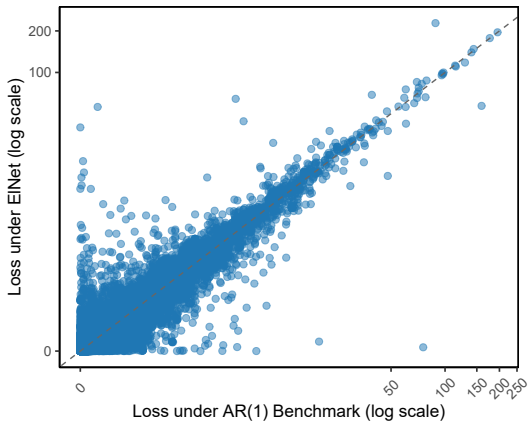
Panel (b) reports results for the Elastic Net estimator (Zou and Hastie, 2005), applied to a broad set of macroeconomic predictors. The majority of points lie below the 45-degree line, suggesting that regularized linear models consistently outperform the benchmark. The relatively tight distribution around the diagonal further indicates that the Elastic Net achieves a favorable bias-variance balance, likely due to its dual shrinkage mechanism.

Panels (c) through (e) show results from nonlinear machine learning methods, namely XGBoost (Chen and Guestrin, 2016), Support Vector Machines, and Random Forests. These models also achieve superior performance in the majority of forecasting tasks, particularly in cases where the AR(1) model yields high losses. However, the scatter of outcomes is more dispersed than under Elastic Net, reflecting the higher variance typically associated with flexible, nonparametric learners (Athey and Imbens, 2019). Despite this, the lower-left clustering of many points suggests that machine learning models excel particularly in regimes where linear benchmarks fail.

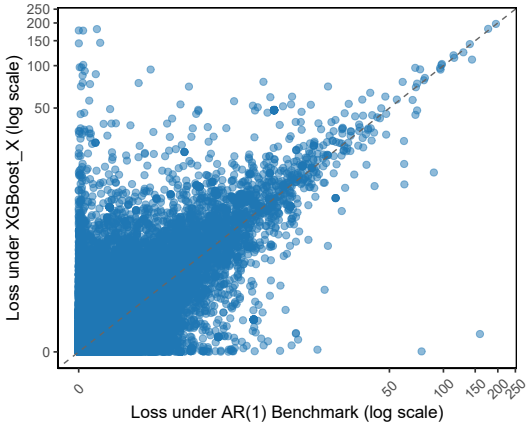
Collectively, the evidence underscores three key findings. First, the AR(1) model is difficult to outperform uniformly but can be outperformed substantially in specific environments. Second, the inclusion of macro predictors, when guided by regularization or adaptive learning, can materially improve forecast accuracy. Third, while more flexible methods may incur higher variance, they exhibit considerable upside, especially when benchmark models are misspecified or under-fit. These results contribute to a growing literature that highlights the potential of machine learning in macroeconomic and financial forecasting (Medeiros et al., 2021; Welch and Goyal, 2008).



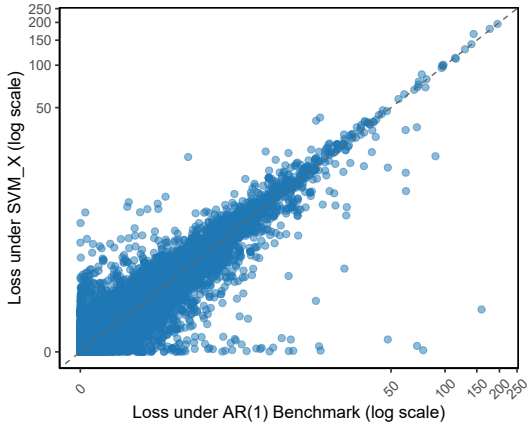
(a) AR(p) chosen by BIC



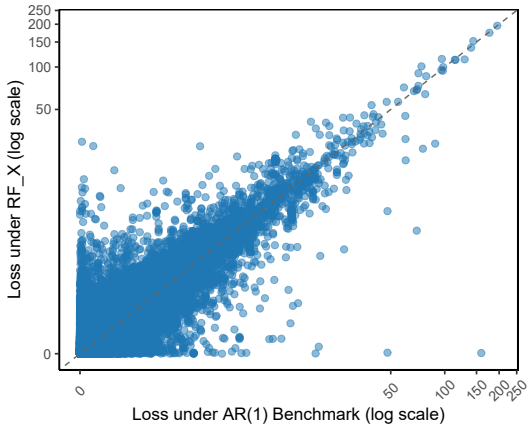
(b) Elastic Net



(c) XGBoost



(d) SVM with Macro Predictors



(e) Random Forest with Macro Predictors

Figure 2: Scatter plots of quadratic forecast losses of alternative forecasting models vs. the benchmark AR(1) model. The 45-degree dashed line indicates equality in forecast performance.

Table 5 presents summary statistics of forecast loss differentials relative to the AR(1) benchmark. Negative values indicate improved forecast performance relative to AR(1). Among the methods considered, XGBoost shows the largest average improvement, with a mean loss differential of -0.54, and a substantial left-skew in its distribution (first quartile = -0.52). This suggests that it often delivers strong gains in cases where AR(1) performs poorly. AR(p) also yields a negative mean (-0.34), but with very high variance (standard deviation = 19.98), indicating occasional large outliers likely due to overfitting in small samples.

Table 5: Summary statistics of loss differentials of different methods with respect to AR(1)

Variable	Mean	Std. Dev.	1st Quartile	Median	3rd Quartile
AR(p)	-0.34	19.98	-0.08	0.00	0.05
Elastic Net	0.03	1.40	-0.06	0.00	0.09
XGBoost	-0.54	3.93	-0.52	-0.03	0.08
SVM	0.00	1.47	-0.12	0.00	0.08
Random Forest	-0.07	1.62	-0.18	0.00	0.09

Note: The results are based on 31178 observations ($T = 238$, $N = 131$) on loss differentials. A negative mean signifies an overall improvement over AR(1) forecasts.

In contrast, the remaining methods, namely Elastic Net, SVM, and Random Forest, have mean loss differentials close to zero, but all display modest left tails. For instance, Elastic Net has a first quartile of -0.06 and third quartile of 0.09, indicating small but frequent gains over AR(1) with little risk of large deterioration. Random Forest shows similar patterns. Taken together, these statistics suggest that flexible methods like XGBoost can offer substantial upside at the cost of some variability, while regularized linear models such as Elastic Net deliver more stable but smaller improvements.

6.3.2. Test results

Table 6 reports the p-values from a series of C-EPA tests applied to loss differentials between five forecasting models and the AR(1) benchmark. The aim is to detect whether the models improve predictive accuracy overall or within specific clusters of currency pairs.

We first look at the O-EPA test results. We see that for all models but SVM, the O-EPA hypothesis is rejected at least at the 10% level in all settings. Recall that AR(p), XGBoost and Random Forest perform better overall with respect to AR(1), whereas Elastic Net is worse, according to Table 5. Hence, in an unconditional setting, the superiority of the first three methods and the inferiority of the last, against AR(1), are confirmed by the O-EPA test results.

Across all settings, SVM stands out as the only method consistently associated with very high p-values in the O-EPA test (e.g., 0.90, 0.95, 0.97), indicating no statistically significant improvement over AR(1) on average over all units and time periods. However, these high p-values do not imply poor performance; rather, they reflect that gains are not homogeneous across all cross-sectional units. This interpretation is supported by the rejection of the homogeneity test at the 10% level in the conditional test with the lagged target and when the number of clusters is chosen by CV (p-value = 0.07). This suggests that SVM's performance is heterogeneous conditional on the past realization of the target variable. Moreover, the selective inference C-EPA test is significant at the 10% level (p-value = 0.09).

More generally, the rejection of the homogeneity null in several cases justifies the use of our Selective Inference C-EPA testing procedure. For example, when conditioning on the lagged target variable, the homogeneity test rejects for SVM and XGBoost depending on the clustering method, and in many cases selective C-EPA p-values very low (e.g., Random Forest yields a p-value of 0.00 in all settings.). These results confirm that forecast gains may vary across clusters, making clustered tests essential to discover such patterns.

We finally note an important implication of the empirical results. The choice of the method for estimating the number of clusters plays a crucial role in test results. CV tends to yield more frequent rejection of the homogeneity null and higher power in the selective C-EPA test compared to the IC approach. This pattern is particularly evident for models like SVM, where the IC-based procedure fails to detect group heterogeneity, but CV-based clustering leads to a borderline or significant result. Hence, the findings underscore the importance of flexible and data-driven clustering in enhancing the sensitivity of selective forecast evaluation procedures.

Overall, these results highlight that clustered inference can detect model improvements that are missed by aggregate tests, and that conditioning and clustering are both essential tools in evaluating forecast performance in panel settings with heterogeneous effects.

7. 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 Panel 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.

Table 6: p-values from C-EPA tests across models and conditioning variables

	Test	AR(p)	Elastic Net	XGBoost	SVM	Random Forest
			Unconditional Tests			
	O-EPA	0.01	0.03	0.00	0.90	0.00
$K = \widehat{K}_{CV}$	Homogeneity	1.00	0.93	0.45	1.00	1.00
	Naive	0.00	0.00	0.00	0.00	0.00
	Split Sample	0.00	0.09	0.00	0.15	0.13
	Selective Inference	0.03	0.11	0.00	1.00	0.00
$K = \widehat{K}_{IC}$	Homogeneity	0.45	0.01	0.91	0.99	0.67
	Naive	0.00	0.01	0.00	0.14	0.00
	Split Sample	0.00	0.11	0.00	0.17	0.00
	Selective Inference	0.01	0.02	0.00	1.00	0.00
			Conditional Tests - Lagged Target			
	O-EPA	0.01	0.09	0.00	0.95	0.00
$K = \widehat{K}_{CV}$	Homogeneity	0.00	1.00	1.00	0.07	0.21
	Naive	0.00	0.02	0.00	0.51	0.02
	Split Sample	0.00	0.02	0.00	0.13	0.01
	Selective Inference	0.00	0.40	0.00	0.09	0.00
$K = \widehat{K}_{IC}$	Homogeneity	0.15	0.67	0.31	0.80	0.21
	Naive	0.00	0.04	0.00	0.72	0.02
	Split Sample	0.00	0.20	0.00	0.16	0.08
	Selective Inference	0.03	0.20	0.00	1.00	0.00
			Conditional Tests - Post Global Financial Crisis Dummy			
	O-EPA	0.00	0.09	0.00	0.97	0.00
$K = \widehat{K}_{CV}$	Homogeneity	0.23	1.00	0.35	0.19	0.93
	Naive	0.00	0.03	0.00	0.00	0.01
	Split Sample	0.00	0.08	0.00	0.11	0.10
	Selective Inference	0.01	0.37	0.00	0.25	0.00
$K = \widehat{K}_{IC}$	Homogeneity	0.23	0.36	0.02	0.07	0.97
	Naive	0.00	0.03	0.00	0.34	0.01
	Split Sample	0.00	0.08	0.00	0.11	0.10
	Selective Inference	0.01	0.18	0.00	0.14	0.00

Note: The results are based on 31178 observations ($T = 238$, $N = 131$) on loss differentials. All tests are robust to arbitrary autocorrelation and CD. Panel Kmeans tests use 10000 initializations. $\widehat{K}_{CV}$ denotes the 10-fold cross-validated estimate of K. $\widehat{K}_{CV} = 2$ in all cases. $\widehat{K}_{IC}$ uses $K_{max} = 5$. Training portion for Split Sample tests is $\gamma = 0.1$. O-EPA test is described in Section 4.2. See Table 1 for the detail on all other testing procedure.

Our proposed method is a conditional testing procedure based on recent developments in the area of selective conditional 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 Panel 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 conditional 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 a battery of time series models as well as more modern machine learning methods with the AR(1) benchmark in terms of their predictive ability, using a large data set of exchange rates. The results showed that taking the latent clusters in the loss differentials between alternative methods and the AR(1) can help the practitioner to improve their forecasts.

Appendices

A. Derivations of the Loss Differentials in Section 2.2

A.1. Proof of Equation (4)

We begin by showing (4). Let $X_{i,T}$ be known and fixed at the time of forecasting. The true data-generating process is given by

$$Y_{i,T+1} = \begin{cases} \alpha_i + \beta_i X_{i,T} + U_{i,T+1}, & i \in \mathcal{C}_1, \\ \beta_i X_{i,T} + U_{i,T+1}, & i \in \mathcal{C}_2, \end{cases}$$

where $U_{i,T+1} \sim iid(0, \sigma^2)$ and is independent of all other variables. The predictors $\hat{\alpha}_i$, $\hat{\beta}_i$, and $\tilde{\beta}_i$ are estimated from a fixed window of past observations and are thus random, while $X_{i,T}$ is treated as fixed. We analyze the two clusters separately.

Case 1: $i \in \mathcal{C}_1$ (True DGP with intercept). In this case, Forecaster 1 correctly includes both an intercept and a slope, whereas Forecaster 2 omits the intercept and thus suffers from misspecification

bias. The one-step-ahead forecast errors can be written as

$$\widehat{Y}^{(1)}_{i,T+1} - Y_{i,T+1} = (\hat{\alpha}_i - \alpha_i) + (\hat{\beta}_i - \beta_i)X_{i,T} + U_{i,T+1},$$
$$\widehat{Y}^{(2)}_{i,T+1} - Y_{i,T+1} = -\alpha_i + (\tilde{\beta}_i - \beta_i)X_{i,T} + U_{i,T+1}.$$

The expectation of squared forecast error of Forecaster 1 is, by a bias–variance decomposition:

$$\begin{aligned}\mathbb{E}[(\widehat{Y}^{(1)}_{i,T+1} - Y_{i,T+1})^2] &= \mathbb{E}[(\hat{\alpha}_i - \alpha_i)^2] + X^2_{i,T}\mathbb{E}[(\hat{\beta}_i - \beta_i)^2] + 2X_{i,T}\mathbb{E}[(\hat{\alpha}_i - \alpha_i)(\hat{\beta}_i - \beta_i)] + \mathbb{E}[U^2_{i,T+1}] \\ &= \mathbb{V}(\hat{\alpha}_i) + \mathbb{B}(\hat{\alpha}_i)^2 + X^2_{i,T}[\mathbb{V}(\hat{\beta}_i) + \mathbb{B}(\hat{\beta}_i)^2] + 2X_{i,T}\operatorname{Cov}(\hat{\alpha}_i, \hat{\beta}_i) + \sigma^2.\end{aligned}$$

Now, let us turn to Forecaster 2, which omits the intercept. This model is misspecified for units in the cluster $\mathcal{C}_1$. Since $\tilde{\beta}_i$ is the OLS estimator from a regression without intercept, it absorbs some of the variation of the omitted constant. The resulting forecast error has a fixed bias term $-\alpha_i$, in addition to the slope estimation error and innovation. Taking the expectation of its square, we have

$$\mathbb{E}[(\widehat{Y}^{(2)}_{i,T+1} - Y_{i,T+1})^2] = \alpha_i^2 + X^2_{i,T}[\mathbb{V}(\tilde{\beta}_i) + \mathbb{B}(\tilde{\beta}_i)^2] + \sigma^2.$$

Subtracting these two expressions yields the expected forecast loss differential:

$$\begin{aligned}&\mathbb{E}[(\widehat{Y}^{(1)}_{i,T+1} - Y_{i,T+1})^2] - \mathbb{E}[(\widehat{Y}^{(2)}_{i,T+1} - Y_{i,T+1})^2] \\ &= \mathbb{V}(\hat{\alpha}_i) + \mathbb{B}(\hat{\alpha}_i)^2 - \alpha_i^2 + [\mathbb{V}(\hat{\beta}_i) - \mathbb{V}(\tilde{\beta}_i)]X^2_{i,T} + [\mathbb{B}(\hat{\beta}_i)^2 - \mathbb{B}(\tilde{\beta}_i)^2]X^2_{i,T} + 2X_{i,T}\operatorname{Cov}(\hat{\alpha}_i, \hat{\beta}_i) \\ &= \mathbb{V}(\hat{\alpha}_i) + \mathbb{B}(\hat{\alpha}_i)^2 - \alpha_i^2 + \Delta_i,\end{aligned}$$

where $\Delta_i := [\mathbb{V}(\hat{\beta}_i) - \mathbb{V}(\tilde{\beta}_i) + \mathbb{B}(\hat{\beta}_i)^2 - \mathbb{B}(\tilde{\beta}_i)^2]X^2_{i,T} + 2X_{i,T}\operatorname{Cov}(\hat{\alpha}_i, \hat{\beta}_i)$. Averaging over $i \in \mathcal{C}_1$ establishes the first line of (4).

Case 2: $i \in \mathcal{C}_2$ (True DGP without intercept). Here, Forecaster 2 correctly specifies the model by excluding the intercept. Forecaster 1, on the contrary, includes an unnecessary intercept term, which leads to overparameterization. The forecast errors are:

$$\widehat{Y}^{(1)}_{i,T+1} - Y_{i,T+1} = \hat{\alpha}_i + (\hat{\beta}_i - \beta_i)X_{i,T} + U_{i,T+1},$$
$$\widehat{Y}^{(2)}_{i,T+1} - Y_{i,T+1} = (\tilde{\beta}_i - \beta_i)X_{i,T} + U_{i,T+1}.$$

Again, we compute the expected squared forecast errors under each model. For Forecaster 1, who estimates both an intercept and slope, we have

$$\mathbb{E}[(\widehat{Y}^{(1)}_{i,T+1} - Y_{i,T+1})^2] = \mathbb{V}(\hat{\alpha}_i) + \mathbb{B}(\hat{\alpha}_i)^2 + X^2_{i,T}[\mathbb{V}(\hat{\beta}_i) + \mathbb{B}(\hat{\beta}_i)^2] + 2X_{i,T}\operatorname{Cov}(\hat{\alpha}_i, \hat{\beta}_i) + \sigma^2.$$

Now, we turn to Forecaster 2, which correctly omits the intercept. The expected forecast loss is:

$$\mathbb{E}[(\widehat{Y}_{i,T+1}^{(2)} - Y_{i,T+1})^2] = \mathbb{V}(\tilde{\beta}_i)X_{i,T}^2 + \mathbb{B}(\tilde{\beta}_i)^2 X_{i,T}^2 + \sigma^2.$$

Subtracting the two, we obtain the loss differential:

$$\mathbb{E}[(\widehat{Y}_{i,T+1}^{(1)} - Y_{i,T+1})^2] - \mathbb{E}[(\widehat{Y}_{i,T+1}^{(2)} - Y_{i,T+1})^2] = \mathbb{V}(\hat{\alpha}_i) + \mathbb{B}(\hat{\alpha}_i)^2 + \Delta_i,$$

where Δ_i is the same as previously defined. Averaging over $i \in \mathcal{C}_2$ yields the second line of (4), completing the derivation.

A.2. Proof of Equation (5)

The forecast error under pooled estimation is

$$\widehat{Y}_{i,T+1}^{\text{pooled}} - Y_{i,T+1} = (\hat{\beta} - \beta_{k_i})' X_{i,T} - U_{i,T+1}.$$

Squaring and taking expectation:

$$\begin{aligned}\mathbb{E}[(\widehat{Y}_{i,T+1}^{\text{pooled}} - Y_{i,T+1})^2] &= \mathbb{E}\{[(\hat{\beta} - \beta_{k_i}]' X_{i,T})^2\} + \mathbb{E}(U_{i,T+1}^2) \\ &= \mathbb{E}[(\hat{\beta} - \beta_{k_i})' X_{i,T} X_{i,T}' (\hat{\beta} - \beta_{k_i})] + \sigma^2.\end{aligned}$$

Using the bias–variance decomposition:

$$\mathbb{E}[(\hat{\beta} - \beta_{k_i})(\hat{\beta} - \beta_{k_i})'] = \mathbb{V}(\hat{\beta}) + [\mathbb{E}(\hat{\beta}) - \beta_{k_i}][\mathbb{E}(\hat{\beta}) - \beta_{k_i}]',$$

we obtain

$$\mathbb{E}[(\widehat{Y}_{i,T+1}^{\text{pooled}} - Y_{i,T+1})^2] = [\mathbb{E}(\hat{\beta}) - \beta_{k_i}]' X_{i,T} X_{i,T}' [\mathbb{E}(\hat{\beta}) - \beta_{k_i}] + \text{tr}[\mathbb{V}(\hat{\beta}) X_{i,T} X_{i,T}'] + \sigma^2.$$

For Forecaster 2, the forecast error is

$$\widehat{Y}_{i,T+1}^{\text{het}} - Y_{i,T+1} = (\hat{\beta}_i - \beta_{k_i})' X_{i,T} - U_{i,T+1}.$$

Assuming $\mathbb{E}(\hat{\beta}_i) = \beta_{k_i}$, the expected squared forecast error is

$$\mathbb{E}[(\widehat{Y}_{i,T+1}^{\text{het}} - Y_{i,T+1})^2] = \text{tr}[\mathbb{V}(\hat{\beta}_i) X_{i,T} X_{i,T}'] + \sigma^2.$$

Taking the difference yields

$$\begin{aligned}\mathbb{E}[(\widehat{Y}_{i,T+1}^{\text{pooled}} - Y_{i,T+1})^2] - \mathbb{E}[(\widehat{Y}_{i,T+1}^{\text{het}} - Y_{i,T+1})^2] &= [\mathbb{E}(\hat{\beta}) - \beta_{k_i}]' X_{i,T} X_{i,T}' [\mathbb{E}(\hat{\beta}) - \beta_{k_i}] \\ &\quad + \text{tr}\{[\mathbb{V}(\hat{\beta}) - \mathbb{V}(\hat{\beta}_i)] X_{i,T} X_{i,T}'\}.\end{aligned}$$

Letting $\Sigma_X = \frac{1}{|\mathcal{C}_k|}\sum_{i\in\mathcal{C}_k} X_{i,T}X_{i,T}'$ and $\overline{\mathbb{V}(\hat{\beta}_i)} = \frac{1}{|\mathcal{C}_k|}\sum_{i\in\mathcal{C}_k}\mathbb{V}(\hat{\beta}_i)$, we have

$$\frac{1}{|\mathcal{C}_k|}\sum_{i\in\mathcal{C}_k}\{\mathbb{E}[(\widehat{Y}_{i,T+1}^{\text{pooled}} - Y_{i,T+1})^2] - \mathbb{E}[(\widehat{Y}_{i,T+1}^{\text{het}} - Y_{i,T+1})^2]\}$$
$$= [\mathbb{E}(\hat{\beta}) - \beta_k]'\Sigma_X[\mathbb{E}(\hat{\beta}) - \beta_k] + \text{tr}\{[\mathbb{V}(\hat{\beta}) - \overline{\mathbb{V}(\hat{\beta}_i)}]\Sigma_X\},$$

noting that $\beta_{k_i} = \beta_k$ for all $i \in \mathcal{C}_k$, which establishes equation (5).

B. Split-sample Test Statistic

In the main text, the selective conditional 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 $\mathcal{S}_1$ and $\mathcal{S}_2$ be two mutually exclusive but not necessarily exhaustive subsets of $\mathcal{S} = \{1,\dots,T\}$ given by $\mathcal{S}_1 = \{1,2,\dots,\lfloor\gamma\cdot T\rfloor\}$ and $\mathcal{S}_2 = \{\lfloor\gamma\cdot T\rfloor+1+l, \lfloor\gamma\cdot T\rfloor+2+l,\dots,T\}$ where $l \geq 1$ is an integer which ensures independence between the two subsets and $\gamma \in (0,1)$ is the proportion of the time series observation in the training set. γ is typically chosen to satisfy $\gamma < 0.5$ because the Panel Kmeans estimator of the cluster membership is super-consistent (Bonhomme and Manresa, 2015) whereas the power of the test statistics crucially depend on a large number of time series observations in the test set.

Let $\widehat{\mathcal{C}}_{\mathcal{S}_1}$ be the partition of the panel units obtained from the Panel Kmeans estimator given in (8) using the sample of N cross-sectional units and the training set $\mathcal{S}_1$. We define $\hat{\theta}_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) = [\hat{\theta}_{1,\mathcal{S}_2}'(\widehat{\mathcal{C}}_{\mathcal{S}_1}),\dots,\hat{\theta}_{K,\mathcal{S}_2}'(\widehat{\mathcal{C}}_{\mathcal{S}_1})]$, and $\hat{\theta}_{k,\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) = |\mathcal{S}_2|^{-1}\sum_{t\in\mathcal{S}_2}\bar{Z}_{k,t}(\widehat{\mathcal{C}}_{\mathcal{S}_1})$, $\bar{Z}_{k,t}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) = |\widehat{\mathcal{C}}_{k,\mathcal{S}_1}|^{-1}\sum_{i\in\widehat{\mathcal{C}}_{k,\mathcal{S}_1}}^{N} Z_{it}$. A Split Sample test statistic for $\mathcal{H}_0$ is

$$W_{SS}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) = \frac{B-KP+1}{KPB}|\mathcal{S}_2|\hat{\theta}_{\mathcal{S}_2}'(\widehat{\mathcal{C}}_{\mathcal{S}_1})\widehat{\Omega}_{\mathcal{S}_2}^{-1}(\widehat{\mathcal{C}}_{\mathcal{S}_1})\hat{\theta}_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}), \tag{27}$$

with $\widehat{\Omega}_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) = \frac{1}{B}\sum_{j=1}^{B}\widehat{\Lambda}_j(\widehat{\mathcal{C}}_{\mathcal{S}_1})\widehat{\Lambda}_j'(\widehat{\mathcal{C}}_{\mathcal{S}_1})$, $\widehat{\Lambda}_j(\widehat{\mathcal{C}}_{\mathcal{S}_1}) = \sqrt{\frac{2}{|\mathcal{S}_2|}}\sum_{t\in\mathcal{S}_2}[\bar{Z}_t(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \hat{\theta}_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1})]\cos\left[\pi j\left(\frac{t-1/2}{P}\right)\right]$.

Let $\mathcal{E}_t = \sigma(\{V_{it}\}_{i=1}^N, s \leq t)$ be the σ-algebra generated by the past and present of V_{it}. The asymptotic properties of the Split Sample test crucially depend on the following assumption.

Assumption G4. V_{it} is independent of all measurable-$\mathcal{E}_{t-l}$ random variables for some $l \geq 1$ and for all $t = 1,\dots,T$, $i = 1,\dots,N$.

According to Assumption G4, time series dependence in the process V_{it} is limited such that V_{it} is

independent of V_{js} whenever $|t-s| \geq l$ for all i and j. This assumption is somewhat restrictive as it rules out many mixing processes for V_{it}. 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 G1-G3 and G4 hold. Then, for B fixed, $|\mathcal{S}_1|, |\mathcal{S}_2| \to \infty$ as $(T, N) \to \infty$, the following results hold.

(a) Under $\mathcal{H}_0$, $W_{SS}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) \xrightarrow{d} \mathbb{F}_{KP,B-KP+1}$.

(b) Suppose now that $K = K^0 \geq 2$. Under Assumptions G1-S2 and G4, and if $\mathcal{H}_0$ fails, then, for any $C > 0$, $\mathbb{P}[W_{SS}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) > C] \to 1$.

The result above leads us to the following remarks. First, the Split Sample test statistics rely on the selection of the two sub-samples $\mathcal{S}_1$ and $\mathcal{S}_2$ which can be arbitrary in practice. Furthermore, since inference is based on a reduced sample size, the associated test statistics may have low power. However, we note that the selective conditional inference approach has extra conditioning due to the nuisance parameters in the conditional distribution of interest. 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.

C. Proofs

C.1. Proof of Lemma 1

To prove Part (a), we show that each $K \times 1$ component of $\hat{\theta}(\mathcal{C})$ satisfies $\hat{\theta}_k(\mathcal{C}) - \theta_k^0(\mathcal{C}) = o_p(1)$. By definition, $Z_{it} = \mu_i^0 + V_{it}$ and $\mathbb{E}(V_{it}) = 0$. Since $\hat{\theta}_k(\mathcal{C}) = (|\mathcal{C}_k|T)^{-1} \sum_{i \in \mathcal{C}_k} \sum_{t=1}^{T} Z_{it}$ and noting that $\theta_k^0(\mathcal{C}) = |\mathcal{C}_k|^{-1} \sum_{i \in \mathcal{C}_k} \mu_i^0$, we have

$$\hat{\theta}_k(\mathcal{C}) - \theta_k^0(\mathcal{C}) = \frac{1}{|\mathcal{C}_k|T} \sum_{i \in \mathcal{C}_k} \sum_{t=1}^{T} V_{it}, \tag{28}$$

which gives $\mathbb{E}[\hat{\theta}_k(\mathcal{C}) - \theta_k^0(\mathcal{C})] = 0$. Turning to the variance, we have

$$\begin{aligned}\left\|\mathbb{E}\{[\hat{\theta}_k(\mathcal{C}) - \theta_k^0(\mathcal{C})][\hat{\theta}_k(\mathcal{C}) - \theta_k^0(\mathcal{C})]'\}\right\| &= \left\|\frac{1}{(|\mathcal{C}_k|T)^2}\sum_{i,j\in\mathcal{C}_k}\sum_{t,s=1}^{T}\mathbb{E}(V_{it}V_{js}')\right\| \\ &\leq \frac{1}{|\mathcal{C}_k|^2 T}\sum_{i,j\in\mathcal{C}_k}\left(\frac{1}{T}\sum_{t,s=1}^{T}\mathbb{E}\|V_{it}V_{js}'\|\right) \\ &\leq \frac{1}{|\mathcal{C}_k|^2 T}\sum_{i,j\in\mathcal{C}_k}\left(\frac{1}{T}\sum_{t,s=1}^{T}\mathbb{E}\|V_{it}V_{js}'\|\right) = O\left(\frac{1}{\pi_k^2 T}\right),\end{aligned} \tag{29}$$

as $(T,N) \longrightarrow \infty$ where the summability of the double sum over t,s follows from the moment conditions of Assumption G1 which ensure that $T^{-1}\sum_{t,s=1}^{T}\mathbb{E}\|V_{it}V_{js}'\|$ is uniformly bounded, and the result follows since G2 ensures that $\pi_k \geq \underline{\pi} > 0$ uniformly in N. This concludes Part (a).

For Part (b), we write

$$\begin{aligned}\widetilde{\Omega}(\mathcal{C})^{-1/2}\mathcal{N}^{1-\epsilon}(\mathcal{C})T^{1/2}[\hat{\theta}(\mathcal{C}) - \theta^0(\mathcal{C})] &= \widetilde{\Omega}(\mathcal{C})^{-1/2}\mathcal{N}^{1-\epsilon}(\mathcal{C})T^{-1/2}\sum_{t=1}^{T}\bar{V}_t(\mathcal{C}) \\ &= \widetilde{\Omega}(\mathcal{C})^{-1/2}T^{-1/2}\sum_{t=1}^{T}\mathcal{N}^{1-\epsilon}(\mathcal{C})\bar{V}_t(\mathcal{C})\end{aligned}$$

where $\bar{V}_t(\mathcal{C}) = [\bar{V}_{1t}(\mathcal{C}),\ldots,\bar{V}_{Kt}(\mathcal{C})]'$ with $\bar{V}_{kt}(\mathcal{C}) = (|\mathcal{C}_k|T)^{-1}\sum_{i\in\mathcal{C}_k}V_{it}$. Since the mixing properties are hereditary, $\mathcal{N}^{1-\epsilon}(\mathcal{C})\bar{V}_t(\mathcal{C})$ satisfies the same mixing conditions satisfied by V_{it} by Assumption G3 (see, for instance, Result 1 of Driscoll and Kraay, 1998). Hence, $\mathcal{N}^{1-\epsilon}(\mathcal{C})\bar{V}_t(\mathcal{C})$ satisfies the conditions of Corollary 2.2 of Phillips and Durlauf (1986) and a multivariate invariance principle holds. The CLT of Part (b) follows directly from this result.

C.2. Proof of Lemma 2

Let

$$\widehat{\mathcal{Q}}(\theta,\mathcal{C}) = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|Z_{it} - \theta_{k_i}\|^2,$$

be the objective function of the Panel Kmeans estimator divided by NT where $\theta_k = |\mathcal{C}_k|^{-1}\sum_{i\in\mathcal{C}_k}\mu_i$, and

$$\widetilde{\mathcal{Q}}(\theta,\mathcal{C}) = N^{-1}\sum_{i=1}^{N}\|\theta_{k_i^0}^0 - \theta_{k_i}\|^2 + (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|V_{it}\|^2,$$

the auxiliary objective function where $\theta_k^0 = |\mathcal{C}_k^0|^{-1} \sum_{i \in \mathcal{C}_k} \mu_i^0$. We also define the Hausdorff distance between $\hat{\theta}(\mathcal{C})$ and $\theta^0(\mathcal{C})$ as

$$d_H[\hat{\theta}(\mathcal{C}), \theta^0(\mathcal{C})] = \max \left\{ \max_{k \in \{1,\dots,K\}} \min_{g \in \{1,\dots,K\}} \left\| \hat{\theta}_k(\mathcal{C}) - \theta_g^0(\mathcal{C}) \right\|^2, \right.$$
$$\left. \max_{g \in \{1,\dots,K\}} \min_{k \in \{1,\dots,K\}} \left\| \hat{\theta}_k(\mathcal{C}) - \theta_g^0(\mathcal{C}) \right\|^2 \right\}.$$

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

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

(b) $\widetilde{\mathcal{Q}}[\hat{\theta}(\widehat{\mathcal{C}}), \widehat{\mathcal{C}}] - \widetilde{\mathcal{Q}}(\theta^0, \mathcal{C}^0) = o_p(1)$.

Proof. To prove (a), we write

$$\left| \widehat{\mathcal{Q}}(\theta, \mathcal{C}) - \widetilde{\mathcal{Q}}(\theta, \mathcal{C}) \right| = \left| \frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} V_{it}' (\theta_{k_i^0}^0 - \theta_{k_i}) \right| \leq 2 \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \theta_{k_i^0}^0 - \theta_{k_i} \right\| \left\| \frac{1}{T} \sum_{t=1}^{T} V_{it} \right\| \right) = o_p(1),$$

which follows directly from Assumption G1(a)-(c). To show (b), we first note that $\widetilde{\mathcal{Q}}(\theta, \mathcal{C})$ is uniquely minimized at true values. To see this, it suffices to write

$$\begin{aligned} \widetilde{\mathcal{Q}}(\theta, \mathcal{C}) - \widetilde{\mathcal{Q}}(\theta^0, \mathcal{C}^0) &= \frac{1}{N} \sum_{i=1}^{N} \left\| \theta_{k_i^0}^0 - \theta_{k_i} \right\|^2 \\ &= \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{g=1}^{K} \mathbf{1}\{k_i^0 = k\} \mathbf{1}\{k_i = g\} \left\| \theta_k^0 - \theta_g(\mathcal{C}) \right\|^2 \\ &\geq \sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{k_i^0 = k\} \min_{g \in \{1,\dots,K\}} \left\| \theta_k^0 - \theta_g(\mathcal{C}) \right\|^2 \\ &= \sum_{k=1}^{K} \frac{|\mathcal{C}_k^0|}{N} \min_{g \in \{1,\dots,K\}} \left\| \theta_k^0 - \theta_g(\mathcal{C}) \right\|^2, \end{aligned} \tag{30}$$

where $|\mathcal{C}_k^0|/N \longrightarrow \pi_k^0 \in (0,1)$ by Assumption G2. Note that, by definition, Panel Kmeans estimator satisfies $\widehat{\mathcal{Q}}[\hat{\theta}(\widehat{\mathcal{C}}), \widehat{\mathcal{C}}] \leq \widehat{\mathcal{Q}}(\theta, \mathcal{C})$. Combining this with (a), we find $\widetilde{\mathcal{Q}}[\hat{\theta}(\widehat{\mathcal{C}}), \widehat{\mathcal{C}}] + o_p(1) \leq \widetilde{\mathcal{Q}}(\theta, \mathcal{C}) + o_p(1)$. Hence, by (30), we have $\widetilde{\mathcal{Q}}[\hat{\theta}(\widehat{\mathcal{C}}), \widehat{\mathcal{C}}] - \widetilde{\mathcal{Q}}(\theta^0, \mathcal{C}^0) = o_p(1)$ which ends the proof. □

For Part (a), we will show the consistency of the Panel 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 $d_H[\hat{\theta}(\widehat{\mathcal{C}}), \theta^0] = o_p(1)$. Define the permutation $\upsilon : \{1, \ldots, K\} \longrightarrow \{1, \ldots, K\}$ as $\upsilon(k) = \arg\min_{g \in \{1,\ldots,K\}} \|\theta_k^0 - \hat{\theta}_g(\widehat{\mathcal{C}})\|^2$. Following steps similar to those in (30), it is easy to show that $\|\theta_k^0 - \hat{\theta}_g(\widehat{\mathcal{C}})\|^2$ is bounded away from zero. It follows that $\upsilon(k) \neq \upsilon(g)$ for all $k \neq g$, with probability approaching to one. Thus, for all $g \in \{1, \ldots, K\}$, $\min_{g \in \{1,\ldots,K\}} \|\theta_k^0 - \hat{\theta}_g(\widehat{\mathcal{C}})\|^2 \leq \|\theta_{\upsilon^{-1}(g)}^0 - \hat{\theta}_g(\widehat{\mathcal{C}})\|^2 = \min_{\tilde{g} \in \{1,\ldots,K\}} \|\theta_{\upsilon^{-1}(g)}^0 - \hat{\theta}_{\tilde{g}}(\widehat{\mathcal{C}})\|^2 = o_p(1)$ where the last equality follows from (30) and Lemma C.1(b). This in turn implies that

$$\max_{k \in \{1,\ldots,K\}} \min_{g \in \{1,\ldots,K\}} \|\theta_k^0 - \theta_g\|^2 = o_p(1).$$

Combining this with the definition of the Hausdorff distance, we find $d_H[\hat{\theta}(\widehat{\mathcal{C}}), \theta^0] = o_p(1)$ which shows that there exists a permutation $\upsilon(k)$ such that $\|\theta_{\upsilon(k)}^0 - \hat{\theta}_k(\widehat{\mathcal{C}})\|^2 = o_p(1)$ which ends the proof of Part (a).

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

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

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

Proof. As in the proof of Lemma B.4 of Bonhomme and Manresa (2015), we first note that, by the definition of $\hat{k}_i(Z)$ in (23), $\mathbf{1}\{\hat{k}_i(Z) = k\} \leq \mathbf{1}\left\{\sum_{t=1}^T \|Z_{it} - \theta_k\|^2 \leq \sum_{t=1}^T \|Z_{it} - \theta_{k_i^0}\|^2\right\}$. Notice also that we can write $N^{-1}\sum_{i=1}^N \mathbf{1}\{\hat{k}_i(Z) \neq k_i^0\} = \sum_{k=1}^K N^{-1}\sum_{i=1}^N \mathbf{1}\{k_i^0 \neq k\}\mathbf{1}\{\hat{k}_i(Z) = k\}$. Combining these two gives $N^{-1}\sum_{i=1}^N \mathbf{1}\{\hat{k}_i(Z) \neq k_i^0\} \leq \sum_{k=1}^K N^{-1}\sum_{i=1}^N Q_{ik}(\theta)$ where $Q_{ik}(\theta) = \mathbf{1}\{k_i^0 \neq k\}\mathbf{1}\left\{\sum_{t=1}^T \|Z_{it} - \theta_k\|^2 \leq \sum_{t=1}^T \|Z_{it} - \theta_{k_i^0}\|^2\right\}$. We will bound $Q_{ik}(\theta)$. By the fact that $Z_{it} = \theta_{k_i^0}^0 + V_{it}$, we have

$$\begin{aligned} Q_{ik}(\theta) &= \mathbf{1}\{k_i^0 \neq k\}\mathbf{1}\left\{\sum_{t=1}^T \sum_{p=1}^P \left[2V_{p,it}(\theta_{p,k_i^0} - \theta_{p,k}) + (\theta_{p,k_i^0}^0 - \theta_{p,k})^2 - (\theta_{p,k_i^0}^0 - \theta_{p,k_i^0})^2\right] \leq 0\right\} \\ &\leq \max_{k \neq g} \mathbf{1}\left\{\sum_{t=1}^T \sum_{p=1}^P \left[2V_{p,it}(\theta_{p,g} - \theta_{p,k}) + (\theta_{p,g}^0(\mathcal{C}^0) - \theta_{p,k})^2 - (\theta_{p,g}^0(\mathcal{C}^0) - \theta_{p,g})^2\right] \leq 0\right\}. \end{aligned}$$

Define

$$\begin{aligned} A = \Bigg| &\sum_{t=1}^T \sum_{p=1}^P \left[2V_{p,it}(\theta_{p,g} - \theta_{p,k}) + (\theta_{p,g}^0(\mathcal{C}^0) - \theta_{p,k})^2 - (\theta_{p,g}^0(\mathcal{C}^0) - \theta_{p,g})^2\right] \\ &- \sum_{t=1}^T \sum_{p=1}^P \left[2V_{p,it}(\theta_{p,g}^0(\mathcal{C}^0) - \theta_{p,k}^0(\mathcal{C}^0)) + (\theta_{p,g}^0(\mathcal{C}^0) - \theta_{p,k}^0(\mathcal{C}^0))^2\right]\Bigg|. \end{aligned}$$

Rearranging and using the triangular inequality, we find,

$$A \leq |A_1| + |A_2| + |A_3| + |A_4|,$$

where

$$A_1 = 2\sum_{t=1}^{T}\sum_{p=1}^{P} V_{p,it}(\theta_{p,g} - \theta^0_{p,g}(\mathcal{C}^0))$$
$$A_2 = 2\sum_{t=1}^{T}\sum_{p=1}^{P} V_{p,it}(\theta^0_{p,k}(\mathcal{C}^0) - \theta_{p,k})$$
$$A_3 = T\sum_{p=1}^{P}(\theta^0_{p,g}(\mathcal{C}^0) - \theta_{p,g})^2$$

and

$$\begin{aligned} A_4 &= T\sum_{p=1}^{P}\left[(\theta^0_{p,g}(\mathcal{C}^0) - \theta_{p,k})^2 - (\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0))^2\right] \\ &= T\sum_{p=1}^{P}\left[\theta^2_{p,k} - \theta^0_{p,k}(\mathcal{C}^0)^2 - 2\theta^0_{p,g}(\mathcal{C}^0)(\theta_{p,k} - \theta^0_{p,k}(\mathcal{C}^0))\right] \\ &= T\sum_{p=1}^{P}\left[\theta^2_{p,k} - \theta^0_{p,k}(\mathcal{C}^0)^2\right] - 2T\sum_{p=1}^{P}\left[\theta^0_{p,g}(\mathcal{C}^0)(\theta_{p,k} - \theta^0_{p,k}(\mathcal{C}^0))\right]. \end{aligned}$$

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

$$A \leq TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{p=1}^{P} V^2_{p,it}\right)^{1/2} + TC_2\eta + TC_3\sqrt{\eta},$$

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

$$\begin{aligned} Q_{ik}(\theta) &\leq \max_{g\neq k} \mathbf{1}\left\{\sum_{t=1}^{T}\sum_{p=1}^{P}\left[2V_{p,it}(\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0)) + (\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0))^2\right]\right. \\ &\qquad \left. \leq TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{p=1}^{P} V^2_{p,it}\right)^{1/2} + TC_2\eta + TC_3\sqrt{\eta}\right\}. \end{aligned}$$

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

$$\begin{aligned}\widetilde{Q}_{ik} \leq \max_{g\neq k} \mathbf{1}\Bigg\{ & \sum_{t=1}^{T}\sum_{p=1}^{P} 2V_{p,it}(\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0)) \\ & \leq -T\sum_{p=1}^{P}(\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0))^2 + TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{p=1}^{P} V^2_{p,it}\right)^{1/2} + TC_2\eta + TC_3\sqrt{\eta}\Bigg\}.\end{aligned}$$

This gives $\sup_{\theta\in\Theta_\eta} N^{-1}\sum_{i=1}^{N}\mathbf{1}\{\hat{k}_i(Z)\neq k_i^0\} \leq N^{-1}\sum_{i=1}^{N}\sum_{k=1}^{K}\widetilde{Q}_{ik}$. Now we have

$$\begin{aligned}\mathbb{P}(\widetilde{Q}_{ik}=1) \leq \sum_{g\neq k}\mathbb{P}\Bigg(& \sum_{t=1}^{T}\sum_{p=1}^{P} 2V_{p,it}(\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0)) \\ & \leq -T\sum_{p=1}^{P}(\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0))^2 + TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{p=1}^{P} V^2_{p,it}\right)^{1/2} + TC_2\eta + TC_3\sqrt{\eta}\Bigg) \\ \leq \sum_{g\neq k}\Bigg[& \mathbb{P}\left(\sum_{t=1}^{T}\sum_{p=1}^{P} 2V_{p,it}(\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0)) \leq -TC_{k,g} + TC_1\sqrt{\eta}\sqrt{C} + TC_2\eta + TC_3\sqrt{\eta}\right) \\ & + \mathbb{P}\left(\sum_{p=1}^{P}(\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0))^2 < C_{k,g}\right) + \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{p=1}^{P} V^2_{p,it} > C\right)\Bigg].\end{aligned}$$

By Assumption S2, the second term above is null, and by Lemma B.5 of Bonhomme and Manresa (2015) and under Assumption S3, $\mathbb{P}\left(T^{-1}\sum_{t=1}^{T}\sum_{p=1}^{P} V^2_{p,it} > C\right) = o(T^{-\xi})$, for all $\xi > 0$. Furthermore, by choosing η suitably, we find

$$\begin{aligned}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{p=1}^{P} 2V_{p,it}(\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0)) \leq -C_{k,g} + C_1\sqrt{\eta}\sqrt{C} + C_2\eta + C_3\sqrt{\eta}\right) \\ \leq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{p=1}^{P} V_{p,it}(\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0)) \leq -\frac{C_{k,g}}{2}\right) = o(T^{-\xi})\end{aligned}$$

where we obtain the last equality by applying Lemma B.5 of Bonhomme and Manresa (2015) with $z_t = V_{p,it}(\theta^0_{p,g}(\mathcal{C}^0) - \theta^0_{p,k}(\mathcal{C}^0))$ and $z = C_{k,g}/2$. This in turn implies that $N^{-1}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{P}(\widetilde{Q}_{ik}=1) = o(T^{-\xi})$. Finally we note that, for all $\xi > 0$ and $\tilde{\xi} > 0$,

$$\begin{aligned}\mathbb{P}\left(\sup_{\theta\in\Theta_\eta}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\hat{k}_i(Z)\neq k_i^0\} > \tilde{\xi}T^{-\xi}\right) & \leq \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\widetilde{Q}_{ik} > \tilde{\xi}T^{-\xi}\right) \\ & \leq \frac{\mathbb{E}\left(N^{-1}\sum_{i=1}^{N}\sum_{k=1}^{K}\widetilde{Q}_{ik}\right)}{\tilde{\xi}T^{-\xi}} = o(1).\end{aligned}$$

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 in our case. Part (c) also follows similar lines to the proof of Theorem 2 and Corollary 1 of Bonhomme and Manresa (2015), with the difference that we have (31) in place of their analogous condition.

First, as in the proof of Part (a), define $\mathcal{Q}^*(\theta) = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\|Z_{it} - \theta_{\hat{k}_i}\|^2$, the concentrated version of $\widehat{\mathcal{Q}}(\theta, \mathcal{C})$, and $\mathcal{Q}^\dagger(\theta) = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\|Z_{it} - \theta_{k_i^0}\|^2$. By choosing η small enough, Lemma C.2 implies that

$$\sup_{\theta\in\Theta_\eta}\left|\mathcal{Q}^*(\theta) - \mathcal{Q}^\dagger(\theta)\right| = o_p(T^{-\xi}),$$

for all $\xi > 0$. Furthermore, by consistency of $\hat{\theta}(\widehat{\mathcal{C}})$ and $\hat{\theta}(\mathcal{C}^0)$, as $(T, N) \to \infty$, $\mathcal{Q}^*[\hat{\theta}(\widehat{\mathcal{C}})] - \mathcal{Q}^\dagger[\hat{\theta}(\widehat{\mathcal{C}})] = o_p(T^{-\xi})$ and $\mathcal{Q}^*[\hat{\theta}(\mathcal{C}^0)] - \mathcal{Q}^\dagger[\hat{\theta}(\mathcal{C}^0)] = o_p(T^{-\xi})$ which in turn gives $\mathcal{Q}^\dagger[\hat{\theta}(\widehat{\mathcal{C}})] - \mathcal{Q}^\dagger[\hat{\theta}(\mathcal{C}^0)] = o_p(T^{-\xi})$. Now, as in (30),

$$\begin{aligned}
\mathcal{Q}^\dagger[\hat{\theta}(\widehat{\mathcal{C}})] - \mathcal{Q}^\dagger[\hat{\theta}(\mathcal{C}^0)] &= \frac{1}{N}\sum_{i=1}^N\left\|\hat{\theta}_{\hat{k}_i} - \hat{\theta}_{k_i}\right\|^2 \\
&= \frac{1}{N}\sum_{i=1}^N\sum_{k=1}^K\sum_{g=1}^K \mathbf{1}\{\hat{k}_i = k\}\mathbf{1}\{k_i = g\}\left\|\hat{\theta}_{\hat{k}_i} - \hat{\theta}_{k_i}\right\|^2 \\
&\geq \sum_{k=1}^K\frac{1}{N}\sum_{i=1}^N \mathbf{1}\{\hat{k}_i = k\}\min_{g\in\{1,\dots,K\}}\left\|\hat{\theta}_{\hat{k}_i} - \hat{\theta}_g\right\|^2 \\
&= \sum_{k=1}^K\frac{|\widehat{\mathcal{C}}_k|}{N}\min_{g\in\{1,\dots,K\}}\left\|\hat{\theta}_{\hat{k}_i} - \hat{\theta}_g\right\|^2,
\end{aligned} \tag{31}$$

where $|\widehat{\mathcal{C}}_k|/N \longrightarrow \pi_k^0 \in (0,1)$ by Assumption G2. We thus obtain $\hat{\theta}_{\hat{k}_i} - \hat{\theta}_{k_i} = o_p(T^{-\xi})$ which ends the proof of Part (c). Part (d) then follows from the consistency of the estimator and Assumption G3.

C.3. Proof of Lemma 3

Let $\mathcal{C}$ be a partition of the panel units with $K \geq 2$, and $\nu_{k,g}$ the associated $NT \times 1$ vector. For convenience, we remind that, $\nu_{k,g} = (\nu_{k,g,1}', \dots, \nu_{k,g,N}')'$, $\nu_{k,g,i} = \iota_T\delta_{k,g,i}$, ι_T being a T-vector of ones, $\delta_{k,g,i} = \mathbf{1}\{k_i = k\}/|\mathcal{C}_k| - \mathbf{1}\{k_i = g\}/|\mathcal{C}_g|$. As in the main text, we also have $\Pi_{k,g} = I - \nu_{k,g}\nu_{k,g}/\|\nu_{k,g}\|^2$. The following lemmas will be referred to in the proof of our result.

Lemma C.3. Suppose that Assumptions G1-G3 hold. Then, as $(T, N) \to \infty$ and $B \to \infty$ with $B/T \to 0$, $\widehat{\Omega}(\mathcal{C}) - \Omega(\mathcal{C}) = o_p(1)$.

Proof. See Sun (2013). □

Lemma C.4. Suppose that Assumptions G1-G3, and $\mathcal{H}_0^{k,g} : \theta_k^0(\mathcal{C}) = \theta_g^0(\mathcal{C})$ hold. Then, $D_{k,g}(\mathcal{C}) \xrightarrow{d} \chi_K$ for all $k, g \in \{2, \ldots, K\}$, $k \neq g$, as $B \to \infty$, $(T, N) \to \infty$ such that $B/T \to 0$.

Proof. For the result to hold, it suffices to show that $D_k^2 \xrightarrow{d} \chi_P^2$ under the assumptions. Let $R_{k,g}$ be the $P \times KP$ selection matrix such that $R_{k,g}\hat{\theta}(\mathcal{C}) = \hat{\theta}_k(\mathcal{C}) - \hat{\theta}_g(\mathcal{C})$. Namely, the matrix $R_{k,g}$ contains an identity matrix I_P in the kth block, $-I_P$ in the gth block, and zeros elsewhere. Using Lemma 1, we find $\Sigma^{-1/2}(\mathcal{C})T^{1/2}R_{k,g}[\hat{\theta}(\mathcal{C}) - \theta^0(\mathcal{C})] \xrightarrow{d} \mathbb{N}(0, I_P)$ where $\Sigma(\mathcal{C}) = R_{k,g}\Omega(\mathcal{C})R_{k,g}'$. Under $\mathcal{H}_0^{k,g}$, this in turn gives

$$T[\hat{\theta}_k(\mathcal{C}) - \hat{\theta}_g(\mathcal{C})]'\Sigma_{k,g}^{-1}(\mathcal{C})[\hat{\theta}_k(\mathcal{C}) - \hat{\theta}_g(\mathcal{C})] \xrightarrow{d} \chi_P^2,$$

where $\Sigma_{k,g}(\mathcal{C}) = \omega_{k,k}(\mathcal{C}) + \omega_{g,g}(\mathcal{C}) - 2\omega_{k,g}(\mathcal{C})$ with $\omega_{k,g}(\mathcal{C})$ begin the $\{k, g\}$th $P \times P$ block of $\Omega(\mathcal{C})$. But by Lemma C.3, we have $\widehat{\omega}_{k,g}(\mathcal{C}) - \omega_{k,g}(\mathcal{C}) = o_p(1)$ from which the result follows. □

Lemma C.5. Suppose that Assumptions G1-G3, and $\mathcal{H}_0^{k,g} : \theta_k^0(\mathcal{C}) = \theta_g^0(\mathcal{C})$ hold. Then, as $(T, N) \to \infty$, $\Pi_{k,g}Z$, $D_{k,g}(\mathcal{C})$ and $\text{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})Z'\nu_{k,g}]$ are asymptotically pairwise independent.

Proof. Notice first that we can write $D_{k,g}(\mathcal{C}) = \|\sqrt{T}\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})Z'\nu_{k,g}\|$ and under Assumptions G1-G3, $\sqrt{T}\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})Z'\nu_{k,g} \xrightarrow{d} \mathbb{N}(0, I_P)$ if $\mathcal{H}_0^{k,g}$ holds, as in Lemma C.4. It follows that $D_{k,g}(\mathcal{C})$ is asymptotically independent of $\text{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})Z'\nu_{k,g}]$ since the length and direction of a standard normal random vector are independent of each other.

To show that $D_{k,g}(\mathcal{C})$ is asymptotically independent of $\Pi_{k,g}Z$, we first note that $\Pi_{k,g}\nu_{k,g} = 0$. This implies by the properties of the matrix normal distribution that $Z'\nu_{k,g}$ is independent of $\Pi_{k,g}Z$ from which the desired result follows 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

$$\begin{aligned} Z &= \Pi_{k,g}Z + (I - \Pi_{k,g})Z = \Pi_{k,g}Z + \frac{\nu_{k,g}\nu_{k,g}'Z\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})\widehat{\Sigma}_{k,g}^{1/2}(\mathcal{C})}{\|\nu_{k,g}\|^2} \\ &= \Pi_{k,g}Z + \frac{\|\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})Z'\nu_{k,g}\|}{\|\nu_{k,g}\|^2}\nu_{k,g}[\text{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}]Z'\nu_{k,g})]'\widehat{\Sigma}_{k,g}^{1/2}(\mathcal{C}) \\ &= \Pi_{k,g}Z + D_{k,g}(\mathcal{C})\frac{\nu_{k,g}}{\sqrt{T}\|\nu_{k,g}\|^2}[\text{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}]Z'\nu_{k,g})]'\widehat{\Sigma}_{k,g}^{1/2}(\mathcal{C}). \end{aligned} \tag{32}$$

By placing this equation in (14), we find that

$$
\begin{aligned}
p_\infty[d_{k,g}(\mathcal{C})] = \lim_{(T,N)\to\infty} \mathbb{P}_{\mathcal{H}_0} \Bigg[& D_{k,g}(\mathcal{C}) \geq d_{k,g}(\mathcal{C}) \;\Bigg| \\
& \bigcap_{m=1}^{M} \bigcap_{i=1}^{N} \left\{ k_i^{(m)} \left(\Pi_{k,g} Z + D_{k,g}(\mathcal{C}) \frac{\nu_{k,g}}{\sqrt{T}\|\nu_{k,g}\|^2} [\mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}] Z' \nu_{k,g})]' \widehat{\Sigma}_{k,g}^{1/2}(\mathcal{C}) \right) = k_i^{(m)}(z) \right\}, \\
& \Pi_{k,g} Z = \Pi_{k,g} z, \mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) Z' \nu_{k,g}] = \mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) z' \nu_{k,g}] \Bigg] \\
= \lim_{(T,N)\to\infty} \mathbb{P}_{\mathcal{H}_0} \Bigg[& D_{k,g}(\mathcal{C}) \geq d_{k,g}(\mathcal{C}) \;\Bigg| \\
& \bigcap_{m=1}^{M} \bigcap_{i=1}^{N} \left\{ k_i^{(m)} \left(\Pi_{k,g} z + D_{k,g}(\mathcal{C}) \frac{\nu_{k,g}}{\sqrt{T}\|\nu_{k,g}\|^2} [\mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}] z' \nu_{k,g})]' \widehat{\Sigma}_{k,g}^{1/2}(\mathcal{C}) \right) = k_i^{(m)}(z) \right\}, \\
& \Pi_{k,g} Z = \Pi_{k,g} z, \mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) Z' \nu_{k,g}] = \mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) z' \nu_{k,g}] \Bigg],
\end{aligned}
$$

where we used the two conditions $\Pi_{k,g} Z = \Pi_{k,g} z$ and $\mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) Z' \nu_{k,g}] = \mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) z' \nu_{k,g}]$ to obtain the second equality. By Lemma C.5, this implies

$$
\begin{aligned}
p_\infty[d_{k,g}(\mathcal{C})] = \lim_{(T,N)\to\infty} \mathbb{P}_{\mathcal{H}_0} \Bigg[& D_{k,g}(\mathcal{C}) \geq d_{k,g}(\mathcal{C}) \;\Bigg| \\
& \bigcap_{m=1}^{M} \bigcap_{i=1}^{N} \left\{ k_i^{(m)} \left(\Pi_{k,g} z + D_{k,g}(\mathcal{C}) \frac{\nu_{k,g}}{\sqrt{T}\|\nu_{k,g}\|^2} [\mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}] z' \nu_{k,g})]' \widehat{\Sigma}_{k,g}^{1/2}(\mathcal{C}) \right) = k_i^{(m)}(z) \right\} \Bigg].
\end{aligned}
$$

Next, by plugging the definition of $\Pi_{k,g}$ into the first term of (32), we have

$$
\begin{aligned}
z(\phi) &\equiv z - \frac{\|z'\nu_{k,g}\|}{\|\nu_{k,g}\|^2} \nu_{k,g} [\mathrm{dir}(z'\nu_{k,g})]' + D_{k,g}(\mathcal{C}) \frac{\nu_{k,g}}{\sqrt{T}\|\nu_{k,g}\|^2} [\mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}] z' \nu_{k,g})]' \widehat{\Sigma}_{k,g}^{1/2}(\mathcal{C}) \\
&= z - \frac{\|z'\nu_{k,g}\|}{\|\nu_{k,g}\|^2} \nu_{k,g} [\mathrm{dir}(z'\nu_{k,g})]' + D_{k,g}(\mathcal{C}) \frac{\nu_{k,g}}{\sqrt{T}\|\nu_{k,g}\|^2} \frac{\|z'\nu_{k,g}\|}{\|\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) z'\nu_{k,g}\|} [\mathrm{dir}(z'\nu_{k,g})]' \\
&= z - \frac{\|z'\nu_{k,g}\|}{\|\nu_{k,g}\|^2} \nu_{k,g} [\mathrm{dir}(z'\nu_{k,g})]' + \phi \frac{\nu_{k,g}}{\sqrt{T}\|\nu_{k,g}\|^2} \frac{\|z'\nu_{k,g}\|}{\|\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) z'\nu_{k,g}\|} [\mathrm{dir}(z'\nu_{k,g})]'.
\end{aligned} \tag{33}
$$

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

$$
p_\infty[d_{k,g}(\mathcal{C})] = \mathbb{P}_{\mathcal{H}_0} \left[\phi \geq d_{k,g}(\mathcal{C}) \;\middle|\; \bigcap_{m=1}^{M} \bigcap_{i=1}^{N} \left\{ k_i^{(m)}[z(\phi)] = k_i^{(m)}(z) \right\} \right],
$$

which shows that $p_\infty(d_{k,g}(\mathcal{C}))$ can be calculated as the survival function of a χ_q variable truncated to the set $\mathcal{T} = \left\{ \phi \in \mathbb{R}_{\geq 0} : \bigcap_{m=1}^{M} \bigcap_{i=1}^{N} k_i^{(m)}[z(\phi)] = k_i^{(m)}(z) \right\}$, that is, $p(d_{k,g}) = 1 - F_{\chi_q}[d_{k,g}; \mathcal{T}]$. This completes the proof.

C.4. Proof of Proposition 1

Lemma C.6. Suppose that Assumptions G1-S3, and $\mathcal{H}_1^{k,g}: \theta_k^0(\mathcal{C}) \neq \theta_g^0(\mathcal{C})$ hold. Then, $D_{k,g}$ diverges as $B \to \infty$, $(T, N) \to \infty$ such that $B/T \to 0$.

Proof. We first note that by Assumption G3, $\Sigma_{k,g}(\mathcal{C}^0)$ is positive definite, so its inverse square root $\Sigma_{k,g}^{-1/2}(\mathcal{C}^0)$ exists and is also positive definite. Moreover, by Assumption S2 $\|\theta_k^0 - \theta_g^0\| > 0$, which means that the difference $\theta_k^0 - \theta_g^0$ is a nonzero vector. Then, since $\Sigma_{k,g}^{-1/2}(\mathcal{C}^0)$ is positive definite and the argument is nonzero, we have

$$\|\Sigma_{k,g}^{-1/2}(\mathcal{C}^0)[\theta_k^0 - \theta_g^0]\|^2 = [\theta_k^0 - \theta_g^0]'\Sigma_{k,g}^{-1}(\mathcal{C}^0)[\theta_k^0 - \theta_g^0] > 0.$$

Taking square roots on both sides gives $\|\Sigma_{k,g}^{-1/2}(\mathcal{C}^0)\left[\theta_k^0 - \theta_g^0\right]\| > 0$.

Now note that $T^{-1/2}D_{k,g} = \|\widehat{\Sigma}_{k,g}^{-1/2}(\widehat{\mathcal{C}})[\hat{\theta}_k(\widehat{\mathcal{C}}) - \hat{\theta}_g(\widehat{\mathcal{C}})]\|$. Moreover, by Lemma 1(a), Lemma 2

$$\|\widehat{\Sigma}_{k,g}^{-1/2}(\widehat{\mathcal{C}})[\hat{\theta}_k(\widehat{\mathcal{C}}) - \hat{\theta}_g(\widehat{\mathcal{C}})]\| - \|\Sigma_{k,g}^{-1/2}(\mathcal{C}^0)[\theta_k^0 - \theta_g^0]\| = o_p(1).$$

Then, it follows that

$$T^{-1/2}D_{k,g} = \|\widehat{\Sigma}_{k,g}^{-1/2}(\widehat{\mathcal{C}})[\hat{\theta}_k(\widehat{\mathcal{C}}) - \hat{\theta}_g(\widehat{\mathcal{C}})]\| \xrightarrow{p} \|\Sigma_{k,g}^{-1/2}(\mathcal{C}^0)[\theta_k^0 - \theta_g^0]\| > 0,$$

which implies that $D_{k,g} \to \infty$ as $T \to \infty$. □

To prove the first part of the proposition, we write

$$\begin{aligned}
&\limsup_{(T,N)\to\infty} \mathbb{P}\Bigg[p[D_{k,g}(\mathcal{C})] \leq \alpha \,\Bigg|\, \bigcap_{m=1}^{M}\bigcap_{i=1}^{N}\left\{k_i^{(m)}\left(\Pi_{k,g}Z + D_{k,g}(\mathcal{C})\frac{\nu_{k,g}}{\sqrt{T}\|\nu_{k,g}\|^2}\{\mathrm{dir}[\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}]Z'\nu_{k,g})\}'\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})\right) = k_i^{(m)}(z)\right\}, \\
&\qquad \Pi_{k,g}Z = \Pi_{k,g}z, \mathrm{dir}(\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})Z'\nu_{k,g}) = \mathrm{dir}(\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})z'\nu_{k,g})\Bigg] \\
&= \limsup_{(T,N)\to\infty} \mathbb{P}\left[p[D_{k,g}(\mathcal{C})] \leq \alpha \,\middle|\, \bigcap_{m=1}^{M}\bigcap_{i=1}^{N}\left\{k_i^{(m)}\left(z[D_{k,g}(\mathcal{C})]\right) = k_i^{(m)}(z)\right\}\right] \\
&= \limsup_{(T,N)\to\infty} \mathbb{P}\left[1 - F_{\chi_q}\left[D_{k,g}(\mathcal{C}); \mathcal{T}\right] \leq \alpha \,\middle|\, \bigcap_{m=1}^{M}\bigcap_{i=1}^{N}\left\{k_i^{(m)}\left(z[D_{k,g}(\mathcal{C})]\right) = k_i^{(m)}(z)\right\}\right],
\end{aligned}$$

which follows lines similar to those above and the definition of $F_{\chi_q}\left[D_{k,g}(\mathcal{C}); \mathcal{T}\right]$ as the cumulative

distribution function of a χ_q variate truncated to the set $\mathcal{T}$. It remains to show that

$$\limsup_{(T,N)\to\infty} \mathbb{P}\left[1 - F_{\chi_q}\left[D_{k,g}(\mathcal{C}); \mathcal{T}\right] \leq \alpha \;\middle|\; \bigcap_{m=1}^{M}\bigcap_{i=1}^{N}\left\{k_i^{(m)}\left(z[D_{k,g}(\mathcal{C})]\right) = k_i^{(m)}(z)\right\}\right] = \alpha.$$

Note that, under $\mathcal{H}_0$, the conditional distribution of $D_{k,g}(\mathcal{C})$ given $\bigcap_{m=1}^{M}\bigcap_{i=1}^{N}\left\{k_i^{(m)}\left(z[D_{k,g}(\mathcal{C})]\right) = k_i^{(m)}(z)\right\}$ is $F_{\chi_q}(\cdot, \mathcal{T})$.

$$\begin{aligned}
&\limsup_{(T,N)\to\infty} \mathbb{P}\left[p[D_{k,g}(\mathcal{C})] \leq \alpha \;\middle|\; \bigcap_{i=1}^{N}\left\{k_i^{(M)}(Z) = k_i^{(M)}(z)\right\}\right] \\
&\quad= \lim_{(T,N)\to\infty} \mathbb{E}\left[\mathbf{1}\left\{p[D_{k,g}(\mathcal{C})] \leq \alpha\right\} \;\middle|\; \bigcap_{i=1}^{N}\left\{k_i^{(M)}(Z) = k_i^{(M)}(z)\right\}\right] \\
&\quad= \lim_{(T,N)\to\infty} \mathbb{E}\left[\mathbb{E}\left(\mathbf{1}\left\{p[D_{k,g}(\mathcal{C})] \leq \alpha\right\} \;\middle|\; \bigcap_{m=1}^{M}\bigcap_{i=1}^{N}\left\{k_i^{(m)}(Z) = k_i^{(m)}(z)\right\}, \Pi_{k,g}Z = \Pi_{k,g}z,\right.\right. \\
&\qquad \left.\left.\operatorname{dir}(\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})Z'\nu_{k,g}) = \operatorname{dir}(\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C})z'\nu_{k,g})\right) \;\middle|\; \bigcap_{i=1}^{N}\left\{k_i^{(M)}(Z) = k_i^{(M)}(z)\right\}\right] \\
&\quad= \lim_{(T,N)\to\infty} \mathbb{E}\left[\alpha \;\middle|\; \bigcap_{i=1}^{N}\left\{k_i^{(M)}(Z) = k_i^{(M)}(z)\right\}\right] = \alpha,
\end{aligned}$$

which concludes the proof of Part (a).

Part (b) follows directly from Lemma C.6 which implies that $D_{k,g} \to \infty$ under the alternative hypothesis, hence, for any $\alpha \in (0,1)$

$$\lim_{(T,N)\to\infty} \mathbb{P}\{p[D_{k,g}(\mathcal{C})] \leq \alpha\} = 1,$$

and noting that under Lemma 2(b), the conditioning event holds with probability 1.

C.5. Proof of Theorem 1

Lemma C.7. Let $G_{NT} = (G_{1,NT}, \ldots, G_{n,NT})'$ be a random n-vector such that $G_{NT} \xrightarrow{d} G$ as $(T,N) \to \infty$. Define

$$f(x_1, \ldots, x_n) = \frac{r}{r+1} n^{1+1/r}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^r\right)^{1/r}$$

where $x_i > 0$ for all $i = 1, \ldots, n$ and $r \in [-\infty, -1)$. Then $f(G_{NT}) \xrightarrow{d} f(G)$.

Lemma C.8. Let $G_{NT} = (G_{1,NT}, \ldots, G_{n,NT})'$ be a random n-vector such that $G_{NT} \xrightarrow{d} G$ as $(T,N) \to \infty$. Define

$$\mathcal{R}_\alpha = \{(x_1, \ldots, x_n) \in [0,1]^n : F(x_1, \ldots, x_n) \leq \alpha\}$$

for all $\alpha \in (0,1)$, where $F(x_1,\ldots,x_n) = f(x_1,\ldots,x_n) \wedge 1$ for some continuous function $f : [0,1]^n \to \mathbb{R}$ and $r \in (1,\infty)$. Then $\lim_{(T,N)\to\infty} \mathbb{P}\left(G_{NT} \in \mathcal{R}_\alpha\right) \leq \mathbb{P}\left(G \in \mathcal{R}_\alpha\right)$.

Proof. Since f is continuous and bounded above by construction, the function $F = f \wedge 1$ is also continuous. Then the set $\mathcal{R}_\alpha = \{x \in [0,1]^n : F(x) \leq \alpha\}$ is closed. The result follows from the Portmanteau Theorem (see, Section 3.4 of Gasparin et al., 2025). □

Define $p^*[D_{k,g}(\widehat{\mathcal{C}})]$ as the limit of the random variable $p[D_{k,g}(\widehat{\mathcal{C}})]$ which satisfies $p[D_{k,g}(\widehat{\mathcal{C}})] \xrightarrow{d} p^*[D_{k,g}(\widehat{\mathcal{C}})] \sim \mathbb{U}[0,1]$ as $(T,N) \to \infty$ for all $k,g \in \{1,\ldots,K\}$, $k \neq$, which holds by Proposition 1(a). By Theorem 1 of Spreng and Urga (2023), we have

$$\mathbb{P}\left[\frac{r}{r+1} n_p^{1+1/r} \left\{\frac{1}{n_p} \sum_{\substack{k,g\in\{1,\ldots,K\} \\ k\neq g}} \{p^*[D_{k,g}(\widehat{\mathcal{C}})]\}^r\right\}^{1/r} \leq \alpha\right] \leq \alpha,$$

Then, part (a) is proved directly by Lemma C.8.

Part (b) now follows from Proposition 1(a) under which at least for one pair $k,g \in \{1,\ldots,K\}$, $k \neq$, the p-value satisfies $p(D_{k,g}) \xrightarrow{p} 1$.

C.6. Proof of Proposition 2

Part (a) follows directly from Theorem 3.1 of Sun (2013) under our Assumptions G1 and G3 by setting $\mathcal{C} = (1,\ldots,1)'$. Part (b) follows from Section 4.1 of Sun (2011) under the same assumptions.

C.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 $k,g \in \{1,\ldots,K\}$, $k \neq g$, the p-value satisfies $p(D_{k,g}) \xrightarrow{p} 1$ and the conditioning event holds w.p.a. 1 by Lemma 2(b), or the O-EPA test statistic diverges.

C.8. Proof of Proposition 3

Consider the mapping $Z \mapsto \widehat{\mathcal{C}}$ where Z is the input of Algorithm 1 and $\widehat{\mathcal{C}}$ is the partition of the panel units which is the output of it. Notice that $Z \mapsto \widehat{\mathcal{C}}$ is the composition of two deterministic procedures: 1. selection of the number of clusters $\widehat{K}_{IC}$ via the minimization of $IC(K)$ in (23), and 2. estimation

of the clustering assignment $\widehat{\mathcal{C}}$ by solving the Panel Kmeans problem (8) with $K = \widehat{K}_{IC}$. Since both steps are deterministic functions of the data, the composite map $Z \mapsto \widehat{\mathcal{C}}$ is itself deterministic.

Now fix a particular realization $\mathcal{C}^*$ of the clustering. The number of clusters in $\mathcal{C}^*$ is fixed. Denote this number by K^*. Then,

$$\{\widehat{\mathcal{C}} = \mathcal{C}^*\} \subseteq \{\widehat{K}_{IC} = K^*\},$$

by the uniqueness of the output $\mathcal{C}^*$ for a given K^*. Hence, conditioning on the event $\{\widehat{\mathcal{C}} = \mathcal{C}^*\}$ implicitly restricts us to the subset of the sample space where $\widehat{K}_{IC} = K^*$. This yields

$$\mathbb{P}\left[D_{k,g}(\widehat{\mathcal{C}}) \in \mathcal{T} \;\middle|\; \widehat{\mathcal{C}} = \mathcal{C}^*\right] = \mathbb{P}\left[D_{k,g}(\widehat{\mathcal{C}}) \in \mathcal{T} \;\middle|\; \widehat{K}_{IC} = K^*,\ \widehat{\mathcal{C}} = \mathcal{C}^*\right],$$

as claimed.

C.9. Proof of Theorem 3

The proof begins algebraically similar to the proof of Lemma 1 except that we will establish a CLT conditional on $\mathcal{C}_{\mathcal{S}_1} = \sigma(\{Z_{it}\}_{i=1}^N, t \in \mathcal{S}_1)$. First, we will show that each $P \times 1$ sub-vector of $\hat{\theta}_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1})$ satisfies $\hat{\theta}_{k,\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) = \theta_k^0(\widehat{\mathcal{C}}_{\mathcal{S}_1}) + o_p(1)$. By Assumption G4, we have

$$\begin{aligned}\mathbb{E}(\hat{\theta}_{k,\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \theta_k^0(\widehat{\mathcal{C}}_{\mathcal{S}_1}) \mid \mathcal{C}_{\mathcal{S}_1}) &= \mathbb{E}\left(\frac{1}{|\widehat{\mathcal{C}}_k||\mathcal{S}_2|}\sum_{i=1}^N\sum_{t\in\mathcal{S}_2} V_{it}\{\hat{k}_{i,\mathcal{S}_1} = k\} \;\middle|\; \mathcal{C}_{\mathcal{S}_1}\right) \\ &= \frac{1}{|\widehat{\mathcal{C}}_k||\mathcal{S}_2|}\sum_{i=1}^N\sum_{t\in\mathcal{S}_2} \mathbb{E}(V_{it} \mid \mathcal{C}_{\mathcal{S}_1})\{\hat{k}_{i,\mathcal{S}_1} = k\} = 0,\end{aligned} \tag{34}$$

For the conditional variance, we find

$$\begin{aligned}&\left\|\mathbb{E}\left[(\hat{\theta}_{k,\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \theta_k^0(\widehat{\mathcal{C}}_{\mathcal{S}_1})(\hat{\theta}_{k,\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \theta_k^0(\widehat{\mathcal{C}}_{\mathcal{S}_1})' \;\middle|\; \mathcal{C}_{\mathcal{S}_1}\right]\right\| \\ &\quad= \left\|\mathbb{E}\left[\frac{1}{(|\widehat{\mathcal{C}}_k||\mathcal{S}_2|)^2}\sum_{i,j=1}^N\sum_{t,s\in\mathcal{S}_2} V_{it}V_{js}'\mathbf{1}\{\hat{k}_{i,\mathcal{S}_1} = k\}\mathbf{1}\{\hat{k}_{j,\mathcal{S}_1} = k\} \;\middle|\; \mathcal{C}_{\mathcal{S}_1}\right]\right\| \\ &\quad\le \frac{1}{|\widehat{\mathcal{C}}_k|^2|\mathcal{S}_2|}\sum_{i,j=1}^N\left\|\frac{1}{|\mathcal{S}_2|}\sum_{t,s\in\mathcal{S}_2}\mathbb{E}\left(V_{it}V_{js}' \mid \mathcal{C}_{\mathcal{S}_1}\right)\right\|\mathbf{1}\{\hat{k}_{i,\mathcal{S}_1} = k\}\mathbf{1}\{\hat{k}_{j,\mathcal{S}_1} = k\} \\ &\quad\le \frac{1}{|\widehat{\mathcal{C}}_k|^2|\mathcal{S}_2|}\sum_{i,j=1}^N\left\|\frac{1}{|\mathcal{S}_2|}\sum_{t,s\in\mathcal{S}_2}\mathbb{E}\left(V_{it}V_{js}' \mid \mathcal{C}_{\mathcal{S}_1}\right)\right\| = O_p\left(\frac{1}{\pi_k^2|\mathcal{S}_2|}\right),\end{aligned} \tag{35}$$

by Assumptions G1 and G2 from which it follows that $\hat{\theta}_{k,\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) = \theta_k^0(\widehat{\mathcal{C}}_{\mathcal{S}_1}) + o_p(1)$. Now, by Assump-

tion G3, conditional on $\mathcal{C}_{\mathcal{S}_1}$ and under $\mathcal{H}_0$, as $|\mathcal{S}_1|, |\mathcal{S}_2| \to \infty$, $(T, N) \to \infty$ we have

$$\begin{aligned}\Omega_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1})^{-1/2}[\hat{\theta}_{k,\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \theta_k^0(\widehat{\mathcal{C}}_{\mathcal{S}_1})] \\ = \Omega_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1})^{-1/2}|\mathcal{S}_2|^{-1/2}\sum_{t\in\mathcal{S}_2}\bar{V}_t(\widehat{\mathcal{C}}_{\mathcal{S}_1}) \xrightarrow{d} \mathbb{N}(0, I_K),\end{aligned}$$

with $\Omega_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) = P^{-1}\sum_{t,s\in\mathcal{S}_2}\mathbb{E}[\bar{V}_t(\widehat{\mathcal{C}}_{\mathcal{S}_1})\bar{V}_s'(\widehat{\mathcal{C}}_{\mathcal{S}_1})]$. Part (a) then follows from Theorem 1 of Sun (2013) noting that $\widehat{\Omega}_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \Omega(\widehat{\mathcal{C}}_{\mathcal{S}_1}) = o_p(1)$, conditional on $\mathcal{C}_{\mathcal{S}_1}$.

For Part (b), we first write

$$\begin{aligned}\hat{\theta}_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \theta^0 &= [\hat{\theta}_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \hat{\theta}_{\mathcal{S}_1}(\widehat{\mathcal{C}}_{\mathcal{S}_1})] + [\hat{\theta}_{\mathcal{S}_1}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \theta^0] \\ &= [\hat{\theta}_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \hat{\theta}_{\mathcal{S}_1}(\widehat{\mathcal{C}}_{\mathcal{S}_1})] + o_p(1),\end{aligned}$$

as $(R, N) \longrightarrow \infty$, 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 $K \times 1$ subvectors of the term:

$$\begin{aligned}\hat{\theta}_{k,\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) - \hat{\theta}_{k,\mathcal{S}_1}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) &= \frac{1}{|\widehat{\mathcal{C}}_k||\mathcal{S}_2|}\sum_{i=1}^{N}\sum_{t\in\mathcal{S}_2} Z_{it}\mathbf{1}\{\hat{k}_{i,\mathcal{S}_1} = k\} - \frac{1}{|\widehat{\mathcal{C}}_k||\mathcal{S}_1|}\sum_{i=1}^{N}\sum_{t\in\mathcal{S}_1} Z_{it}\mathbf{1}\{\hat{k}_{i,\mathcal{S}_1} = k\} \\ &= \frac{1}{|\widehat{\mathcal{C}}_k||\mathcal{S}_2|}\sum_{i=1}^{N}\sum_{t\in\mathcal{S}_2} V_{it}\mathbf{1}\{\hat{k}_{i,\mathcal{S}_1} = k\} - \frac{1}{|\widehat{\mathcal{C}}_k||\mathcal{S}_1|}\sum_{i=1}^{N}\sum_{t\in\mathcal{S}_1} V_{it}\mathbf{1}\{\hat{k}_{i,\mathcal{S}_1} = k\} \\ &= \frac{1}{|\mathcal{S}_2|}\sum_{t\in\mathcal{S}_2}\bar{V}_{k,t} - \frac{1}{\mathcal{S}_1}\sum_{t\in\mathcal{S}_1}\bar{V}_{k,t} \\ &= O_p\left(\frac{\pi_k^{\epsilon-1}}{N^{1-\epsilon}\sqrt{|\mathcal{S}_2|}}\right) + O_p\left(\frac{\pi_k^{\epsilon-1}}{N^{1-\epsilon}\sqrt{|\mathcal{S}_1|}}\right) = o_p(1).\end{aligned}$$

This in turn gives

$$\hat{\theta}_{\mathcal{S}_2}'(\widehat{\mathcal{C}}_{\mathcal{S}_1})\widehat{\Omega}_{\mathcal{S}_2}^{-1}(\widehat{\mathcal{C}}_{\mathcal{S}_1})\hat{\theta}_{\mathcal{S}_2}(\widehat{\mathcal{C}}_{\mathcal{S}_1}) \xrightarrow{p} \theta^{0\prime}\Omega^{-1}\theta^0 > 0,$$

by Assumptions G3 and S1 from which it follows that $W_{SS}(\widehat{\mathcal{C}}_{\mathcal{S}_1})$ diverges w.p.a. 1 which completes the proof.

D. Calculation of the Truncation Set $\mathcal{T}$

For convenience, we restate the expression for the truncation set $\mathcal{T}$:

$$\mathcal{T} = \left\{\phi \in \mathbb{R}_{\geq 0} : \bigcap_{m=1}^{M}\bigcap_{i=1}^{N} k_i^{(m)}[z(\phi)] = k_i^{(m)}(z)\right\}.$$

According to the second step (assignment) of Algorithm 1, the equality inside the braces holds if and only if the cluster center which is closest to z_{it} in total over t, coincides with the cluster center of the

previous iteration that is closest to $[z(\phi)]_{it}$ in total over t, for all $i = 1, \ldots, N$. Using Proposition 2 of Chen and Witten (2023) we can then write:

$$\mathcal{T} = \bigcap_{m=1}^{M} \bigcap_{i=1}^{N} \bigcap_{g=1}^{G} \left\{ \phi \in \mathbb{R}_{\geq 0} : \frac{1}{T} \sum_{t=1}^{T} \left\| [z(\phi)]_{it} - \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} w_j^{(m-1)} [k_i^{(m)}(z)][z(\phi)]_{jt} \right\|^2 \right.$$
$$\left. \leq \sum_{t=1}^{T} \left\| [z(\phi)]_{it} - \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} w_j^{(m-1)}(k)[z(\phi)]_{jt} \right\|^2 \right\} \tag{36}$$

where $w_i^{(m)}(k) = \mathbf{1}\left\{k_i^{(m)}(z) = k\right\} / \sum_{j=1}^{N} \mathbf{1}\left\{k_j^{(m)}(z) = k\right\}$. By (33), we see that

$$[z(\phi)]_{it} = z_{it} - \hat{\delta}_{k,g,i} \frac{\|z'\hat{\nu}_{k,g}\|}{\|\hat{\nu}_{k,g}\|^2} \mathrm{dir}(z'\hat{\nu}_{k,g}) + \left(\frac{\|z'\hat{\nu}_{k,g}\|}{\|\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) z'\hat{\nu}_{k,g}\|} \frac{\hat{\delta}_{k,g,i}}{\sqrt{T}\|\hat{\nu}_{k,g}\|^2} \phi \right) \mathrm{dir}(z'\hat{\nu}_{k,g}). \tag{37}$$

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

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

where

$$\tilde{a}_{ij} = \left(\frac{\|z'\hat{\nu}_{k,g}\|}{\|\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) z'\hat{\nu}_{k,g}\|} \right)^2 \left(\frac{\hat{\delta}_{k,g,i} - \sum_{j=1}^{N} w_j^{(m-1)}(k)\hat{\delta}_{k,g,j}}{\sqrt{T}\|\hat{\nu}_{k,g}\|^2} \right)^2,$$

$$\tilde{b}_{ijt} = 2 \left(\frac{\|z'\hat{\nu}_{k,g}\|}{\|\widehat{\Sigma}_{k,g}^{-1/2}(\mathcal{C}) z'\hat{\nu}_{k,g}\|} \right)$$
$$\times \left\{ \frac{\hat{\delta}_{k,g,i} - \sum_{j=1}^{N} w_j^{(m-1)}(k)\hat{\delta}_{k,g,j}}{\sqrt{T}\|\hat{\nu}_{k,g}\|^2} \left\langle z_{it} - \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} w_j^{(m-1)}(k) z_{jt}, \mathrm{dir}(z'\hat{\nu}_{k,g}) \right\rangle \right.$$
$$\left. - \frac{(\hat{\delta}_{k,g,i} - \sum_{j=1}^{N} w_j^{(m-1)}(k)\hat{\delta}_{k,g,j})^2}{\sqrt{T}\|\hat{\nu}_{k,g}\|^4} \|z'\hat{\nu}_{k,g}\| \right\},$$

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

These in turn show that the truncation set $\mathcal{T}$ can be analytically calculated as the inequalities defined in the two components of (36) are all quadratic in ϕ.

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, the 29$^{\text{th}}$ International Panel Data Conference (Orléans, 3-5 July 2024), and the participants of the GIAM Seminar at Galatararay University (Istanbul, 29 April, 2025). The usual disclaimer applies.

References

Akgun, O. and Okui, R. (2025). Testing parameter homogeneity in panel models with latent group structure: A selective conditional inference approach. Unpublished manuscript.

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.

Ando, T. and Bai, J. (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association*, 112(519):1182–1198.

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*, pages 507–547. University of Chicago Press.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

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

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

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.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chan, B. C. Y., Ng, S., and Bai, J. (2023). fbi: Factor-based imputation and fred-md/qd data set. https://github.com/cykbennie/fbi. R package version 0.7.0.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

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. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of econometrics*, 105(1):85–110.

Clark, T. E. and McCracken, M. W. (2014). Tests of equal forecast accuracy for overlapping models. *Journal of Applied Econometrics*, 29(3):415–430.

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.

Dreher, A., Marchesi, S., and Vreeland, J. R. (2008). The political economy of IMF forecasts. *Public Choice*, 137:145–171.

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. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

Fithian, W., Sun, D., and Taylor, J. (2017). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597v4*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

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. (2025). Combining exchangeable P-values. *Proceedings of the National Academy of Sciences*, 122(11):e2410849122.

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.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5):920–964.

Haghighi, M., Joseph, A., Kapetanios, G., Kurz, C., Lenza, M., and Marcucci, J. (2025). Machine learning for economic policy. *Journal of Econometrics*, 249(Part C):105970.

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.

Harvey, D. I., Leybourne, S. J., and Zu, Y. (2024). Testing for equal average forecast accuracy in possibly unstable environments. *Journal of Business & Economic Statistics*, 43(3):643–656.

Hillebrand, E., Mikkelsen, J. G., Spreng, L., and Urga, G. (2023). Exchange rates and macroeconomic fundamentals: Evidence of instabilities from time-varying factor loadings. *Journal of Applied Econometrics*, 38(6):857–877.

Hoga, Y. and Dimitriadis, T. (2023). On testing equal conditional predictive ability under measurement error. *Journal of Business & Economic Statistics*, 41(2):364–376.

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.

Inoue, A. and Kilian, L. (2006). On the selection of forecasting models. *Journal of Econometrics*, 130(2):273–306.

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.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the Lasso. *The Annals of Statistics*, 44(3):907–927.

Li, Z., Zhu, X., and Zou, C. (2025). Consistent selection of the number of groups in panel models via cross-validation. *arXiv preprint arXiv:2209.05474v3*.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3):18–22.

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*.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1-2):499–526.

Markovic, J., Xia, L., and Taylor, J. (2017). Unifying approach to selective inference with applications to cross-validation. *arXiv preprint arXiv:1703.06559*.

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

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, A. d. P., and Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2024). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-16.

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

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. (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.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14:199–222.

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., Wang, B., and Wang, R. (2022). Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1):351–375.

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

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508.

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

Yun, Y. and He, Y. (2024). Selective inference for multiple pairs of clusters after k-means clustering. *arXiv preprint arXiv:2405.16379*.

Zhu, Y. and Timmermann, A. (2022). Can two forecasts have the same conditional expected accuracy? *arXiv preprint arXiv:2006.03238v2*.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.