Bayesian Fuzzy Clustering with Robust Weighted Distance for Multiple ARIMA and Multivariate Time-Series

Pacifico, Antonio

LUISS Guido Carli University, CEFOP-LUISS, LUISS Guido Carlo University

2020

Bayesian Fuzzy Clustering with Robust Weighted Distance for Multiple ARIMA and Multivariate Time-Series

Antonio Pacifico*

Abstract

The paper suggests and develops a computational approach to improve hierarchical fuzzy clustering time-series analysis when accounting for high dimensional and noise problems in dynamic data. A Robust Weighted Distance measure between pairs of sets of Auto-Regressive Integrated Moving Average models is used. It is *robust* because Bayesian Model Selection methodology is performed with a set of conjugate informative priors in order to discover the most probable set of clusters capturing different dynamics and interconnections among time-varying data, and *weighted* because each time-series is 'adjusted' by own Posterior Model Size distribution in order to group dynamic data objects into 'ad hoc' homogenous clusters. Monte Carlo methods are used to compute exact posterior probabilities for each cluster chosen and thus avoid the problem of increasing the overall probability of errors that plagues classical statistical methods based on significance tests. Empirical and simulated examples describe the functioning and the performance of the procedure. Discussions with related works and possible extensions of the methodology to jointly deal with endogeneity issues and misspecified dynamics in high dimensional multicountry setups are also displayed.

JEL classification: A1; C01; E02; H3; N01; O4

Keywords: Distance Measures; Fuzzy Clustering; ARIMA Time-Series; Bayesian Model Selection; MCMC Integrations.

*Corresponding author: Antonio Pacifico, Postdoc in `Applied Statistics & Econometrics`, LUISS Guido Carli University, CEFOP-LUISS (Rome), CSS Scientific Institute (Tuscany). Email: antonio.pacifico86@gmail.com or apacifico@luiss.it. ORCID:https://orcid.org/0000-0003-0163-4956

1 Introduction

With increasing power of data storages and processors, several applications have found the chance to store and keep data for a long time. Thus, data in many applications have been stored in the form of time-series data. Generally, a time-series is classified as dynamic data because its feature values change as a function of time, which means that the values of each point of a time-series are one or more observations that are made chronologically (see, e.g., Keogh and Kasetty (2003) and Rani and Sikka (2012)). The intrinsic nature of a time-series is usually that the observations are dependent or correlated. The Auto-Regressive Integrated Moving Average (ARIMA) processes are a very general class of parametric models useful for describing dynamic data and their correlations. This amount of time-series data has provided the opportunity of analysing time-series for many researchers in data mining communities in the last decades. Consequently, many researches and projects relevant to analyze time-series have been performed in various areas for different purposes such as: subsequence matching; clustering; identifying patterns; trend analysis; and forecasting. It has carried the need to develop many on-going research projects aimed to improve the existing techniques (see, for instance, Zakaria et al. (2012) and Rakthanmanon et al. (2012)).

Time-series data are of interest because of its pervasiveness in various areas such as: business; finance; economic; health care; and government. Given a set of unlabeled time series, it is often desirable to determine groups (or clusters) of similar time-series. These time-series could be monitoring data collected during different periods from a particular process or from more than one process. Generally, clusters are formed by grouping objects that have maximum similarity with other objects within the group, and minimum similarity with objects in other groups. It is a useful approach for exploratory data analysis as it identifies structures in an unlabelled dataset by objectively organizing data into similar groups. The goal of clustering is to identify structure in an unlabeled data set by grouping data objects into a tree of homogeneous clusters, where the within-group-object similarity is minimize and the between-group-object dissimilarity is maximized. Nevertheless, works devoting to the cluster analysis of time-series are relatively scant compared with those focusing on static data. In addition, a pure hierarchical clustering method suffers from its inability to perform adjustment once a merge or split decision has been executed. Thus, for improving the clustering quality of hierarchical methods, there is a trend of increased activity to integrate hierarchical clustering with other clustering techniques.

My approach and empirical application aim to give a valid contribution to such topics. More precisely, when dealing with time-series, a suitable measure to evaluate the similarities and dissimilarities within the data becomes necessary and subsequently it exhibits a significant impact on the results of clustering. This selection should be based upon the nature of time-series and the application itself. In this context, hierarchical fuzzy clustering tends to hold a relevant competitive position. It is a data mining technique where similar data are placed into related or homogeneous groups without advanced knowledge of the groups' definitions. Fuzzy clustering is one of the widely used clustering techniques where one data object is allowed to be in more than one cluster to a different degree. Fuzzy C-Means (FCM) and Fuzzy C-Medoids (FCMdd) are the two well-known and representative fuzzy clustering methods (see, e.g., Kannan et al. (2012), Izakian et al. (2015), Kaufman and Rousseeuw (2009), and Liao (2005)). In both techniques, the objective is to form a number of cluster centers and a partition matrix so that a given performance index becomes minimized. FCM generates a set of cluster centers using a weighted average of data, whereas FCMdd selects the cluster centers as some of the existing data points (medoids). They aim is to minimize a weighted sum of distances between data points and cluster centers. The empirical analysis conducted in this paper focuses on the only FCM algorithm for effective clustering of ARIMA time-series.

My methodological contribution is based on two lines of research. First, the study on comparative aspects of time-series clustering experiments (see, e.g., Keogh and Kasetty (2003), Aghabozorgi et al. (2015), Kavitha and Punithavalli (2010), Liao (2005), D'Urso and Maharaj (2009), Ramoni et al. (2002), and Rani and Sikka (2012)). Second, the use of measures of similarity/dissimilarity between univariate linear models (see, e.g., Piccolo (1990), Corduas and Piccolo (2008), Maharaj (1996), Martin (2000), and Triacca (2016)). Thus, the main thrust of this study is to provide an updated investigation on the trend of improvements in efficiency, quality and complexity of fuzzy clustering time-series approaches, and highlight new paths for future works. I adapt the framework of Pacifico (2020b), who develops a Robust Open Bayesian procedure for implementing Bayesian Model Averaging (BMA) strategy and Markov Chain Monte Carlo (MCMC) methods in order to deal with endogeneity issues and functional form misspecification in multiple linear regression models. In particular, the paper has three specific objectives. First, I develop a computational approach to improve hierarchical fuzzy clustering time-series analysis when accounting for high dimensional and noise problems in dynamic data and jointly dealing with endogeneity issues and misspecified dynamics in multicountry setups. Second, in order to make operative the notion of distance between finite sets of ARIMA models, a Robust Weighted Distance (RWD) measure between stationary and invertible ARIMA processes is performed. It is $robust$ because the Bayesian Model Selection (BMS) is performed with a set of conjugate informative priors in order to discover the most probable set of clusters capturing different dynamics and interconnections among time-varying data and $weighted$ because each unlabeled time-series is 'adjusted', on average, by own Posterior Model Size (PMS) distribution[1] in order to group dynamic data objects into 'ad hoc' homogenous clusters where the within-group-object similarity is minimize and the between-group-object dissimilarity is maximized. Third, a MCMC approach is used to move through the model space and the parameter space at the same time in order to obtain a reduced set containing $best$ (possible) model solutions and thus exact posterior probabilities for each cluster chosen, dealing with the problem of increasing the overall probability of errors that plagues classical statistical methods based on significance tests. Here, $best$ stands for the clustering model providing the most accurate group of homogeneous time-series over all (possible) candidate processes. In this context, Bayesian methods are used to reduce the dimensionality of the model, structure the time variations, evaluate issues of endogeneity and structured model uncertainty, with one or more parameters posited as the source of model misspecification problems. In this way, policies designed to protect the economy against the worst-case consequences of misspecified dynamics are less aggressive and good approximations of the estimated rule. The dimensionality reduction is greatly important in clustering of time-series because: (i) it reduces memory requirements as all time-series cannot fit in the main memory; (ii) distance calculation among dynamic data is computationally expensive and thus dimensionality reduction significantly speeds up clustering; and (iii) one may cluster time-series which are similar in noise instead of clustering them based on similarity in shape.

Empirical and simulated examples describe the functioning and the performance of the procedure. More precisely, I perform an empirical application for moderate time-varying data ($k \leq 15$) and a simulated experiment for larger sets ($k > 15$) on a database of multiple ARIMA time-series in order to display the performance and usefulness of BHFC procedure and RWD measure, with k denoting the number of time-series. In addition, I extend and implement the BHFC procedure in order to deal with endogeneity issues and functional form misspecifications when accounting for dynamics of the economy in high dimensional time-varying multicountry data. In this context, the RWD measure is used to group multiple data objects that are generated from different

[1] They correspond to the sum of Posterior Inclusion Probability between all the grouped time-series according to their membership values.

series among a pool of advanced European economies. I build on Pacifico (2019a) and estimate a simplified version of the Structural Panel Bayesian VAR (SPBVAR) by defining hierarchical prior specification strategy and MCMC implementations in order to extend variable selection procedure for clustering time-series to a wide array of candidate models. A discussion with different distance measures is also accounted for.

The outline of this paper is as follows. Section 2 reviews the proposed methods for clustering time-series. Section 3 displays a brief description of various concepts and definitions used throughout the paper. Section 4 describes Bayesian framework and conjugate hierarchical setups in ARIMA time-series. Section 5 illustrates MCMC algorithm and the proposed RWD measure for clustering ARIMA time-series. Section 6 illustrates empirical and simulated examples by comparing the performance of BHFC procedure with related works. Section 7 extends the methodology to deal with endogeneity issues and misspecified dynamics in high dimensional time-varying multicountry setups. Finally, Section 8 contains some concluding remarks.

2 Literature Review on Distance Measures

When dealing with time-series data which are usually serially correlated, one needs to extract the significant features from them. Thus, measuring the similarity and the dissimilarity between models becomes crucial in many field of time-series analysis. An appropriate distance function to evaluate similarities and dissimilarities of time-series has a significant impact on the clustering algorithms and their final results produced by them. This selection may depend upon the nature of the data and the specificity of the application.

In most partition-based time-series data fuzzy clustering techniques, the Euclidean Distance (ED) is commonly used to quantify the similarities and dissimilarities of time-series. However, it compares the points of time-series in a fixed order and cannot take into account existing time shifts. In addition, the ED is applicable only when comparing equal-length time-series and, in most feature-based clustering techniques, the representatives of clusters cannot be reconstructed in the original time-series domain and in such a way they are not useful for data summarization. Izakian et al. (2013) and Izakian and Pedrycz (2014) propose an augmented version of ED function for fuzzy clustering of time-series data. The original time-series as well as different representation techniques have been examined for clustering purpose. D'Urso and Maharaj (2009) transform the time-series data through their autocorrelation representation, and use the Euclidean distance to compare data in the new feature space. Thus, a FCM technique has been employed to cluster the transformed data. Keogh et al. (2001) propose a hierarchical clustering technique of time-series data to quantify the dissimilarity of time-series.

Nevertheless, when clustering by dynamics and measuring the distance between multiple time-series, highly unintuitive results may be obtained since some distance measures may be highly sensitive to some distortions in the data and thus, by using raw time-series, one may cluster time-series which are similar in noise instead of clustering them based on similarity in shape (see, e.g., Keogh and Ratanamahatana (2005) and Ratanamahatana et al. (2005)). High dimensional and noise are characteristics of most time-series data and thus dimensionality reduction methods are usually used in whole time-series clustering in order to address these issues and promote the performance (see, e.g., Keogh and Kasetty (2003), Keogh and Ratanamahatana (2005), and Ghysels et al. (2006)). Thus, the potential to group time-series into clusters so that the elements of each cluster have similar dynamics is the reason why choosing the appropriate approach for dimension reduction (feature extraction), and an appropriate data representation method is a challenging task. In fact, it is a trade-off between speed and quality and all efforts must be made to obtain a proper balance point between quality and execution time. Dimensionality reduction represents the raw time-series in another space by transforming time-series to a lower dimensional space or by feature extraction. Time-series dimensionality reduction techniques have progressed a long way and are widely used with large scale time-series dataset and each has its own features and drawbacks (see, e.g., Lin et al. (2007) and Keogh et al. (2001)).

In this context, a Bayesian approach is particularly well suited to cluster by dynamics and deal with raw time-series since it provides a principled way to integrate prior and current evidence. In addition, because the posterior probability of a partition is the scoring metric, it avoids the problem of increasing the overall probability of errors that plagues classical statistical methods based on significance tests. Bayesian clustering methods has been pioneered by Cheeseman et al. (1996) for static databases, under the assumption that the data are independent and identically distributed. Poulsen (1990), Cooper and Herskovits (1992), and Visser et al. (2000) extend the original method to temporal data using an approximate mixture-model approach to cluster discrete MCs within a pre-specified number of clusters. More recently, Ramoni et al. (2002) propose a Bayesian method to cluster time-series through modelling the time-series as Markov chains and using a symmetric Kullback-Libler[2] distance between transition matrices. Here, the clustering

[2] It is a well-known statistical indicator useful in evaluating the similarity of time-series represented by their Markov chains.

has been considered as a Bayesian Model Selection (BMS) problem to find the most suitable set of clusters. However, these studies only focus on standard clustering techniques of time-series and thus they are unable to merge different dynamic data objects in more than one similar clusters or split them in different groups correctly when dealing with endogeneity issues and unmodelled dynamic interconnections.

This paper gives a valid contribution to that literature by performing a BHFC methodology with RWD measure between stationary and invertible ARIMA processess. The proposed approach displays three important features which makes it ideal for clustering by dynamics and measuring the distance between multiple time-series. First, hierarchical conjugate informative priors are able to discover the most probable set of clusters capturing different dynamics and interconnections among time-varying data. Second, full posterior distributions in fuzzy clustering of time-series data are able to avoid the problem of increasing the overall probability of errors that plagues classical statistical methods based on significance tests. Third, by construction, the proposed procedure is able to jointly deal with high dimensional and noise problems, especially in extending the analysis to multicountry setups.

3 ARIMA Time-Series and Bayesian Framework

A stationary time-series is one whose probability distribution is time-invariant. On the contrary, a non-stationary time-series may have its mean μ_t or variance σ_t varying with time. Normally, a time-series has four components: (i) a trend, (ii) a cycle, (iii) a stochastic persistence component, and (iv) a random element. ARIMA models constitute a broad class of parsimonious time-series processes which are useful in describing a wide variety of time-series. For example, the process x_t is said to be an Auto-Regressive Integrated Moving Average process of order p, d, q, with mean μ, if it is generated by:

$$\phi(B)(x_t - \mu) = \theta(B)\epsilon_t \tag{1}$$

Letting $y_t = x_t - \mu$, the ARIMA(p, d, q) model in (1) can be written as:

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{2}$$

The model in equation (2) is called ARIMA(p,d,q) model, where B is the backward shift operator, $\epsilon_t \sim WN(0,\sigma^2)$ is a Gaussian white noise process, $y_t \sim N(\mu 1_N, \sigma_\epsilon^2 C_N)$, with 1_N is a $k \cdot 1$ vector of ones and $(C_N)_{ij} = cov(y_i, y_j) = \rho(i-j) = \rho(|i-j|)$, $i = 1,2,\ldots,p$ and $j = 1,2,\ldots,q$ denote generic Auto-Regressive (AR) and Moving Average (MA) lag orders, respectively, $d = 1,\ldots,D$ refers to higher differentiation order to obtain a stationary time-series, $t = 1,2,\ldots,n$ denote time periods, $\phi_i(B) = (1-\phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$ represents the correlation of x_t on its preceding values, and $\theta_j(B) = (1-\theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)$ represents the MA component.

Collecting the components ϕ_i and θ_j in a coefficient vector δ, where $\delta = (\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q)'$, two conditions need to be assessed. First, if the roots of $\phi(B) = 0$ and $\theta(B) = 0$ lie outside the unit circle, the process in (1) is said to be stationary and invertible, respectively, and thus there is a unique model corresponding to the likelihood function (see, for instance, Li and McLeod (1986)). Second, if the stationarity and invertibility conditions hold, the parameter vector δ is constrained to lie in regions C_p and C_q, respectively, corresponding to the polynomial operator root conditions. Here, the region $C_p \cdot C_q$ contains allowable values of (ϕ,θ) which are simple to identify for $p \leq 2$ and $q \leq 2$. These identifiability conditions enforce a unique parameterization of the model in terms of μ, σ^2, and the ARMA components in δ.

In Bayesian framework, given a stationary and invertible ARIMA(p,d,q) time-series model of the form (1), the region $C_p \cdot C_q$ determines the ranges of integration for obtaining joint and marginal distributions of the parameters and for evaluating posterior expected values. Generally, Bayesian analysis of these models ignores this region in order to obtain convenient distributional results for the posterior densities (see, e.g., Zellner (1983), Carlin et al. (1992b), and Carlin et al. (1992a)). However, when $p+q \geq 4$, with unknown μ and σ^2, such techniques become unfeasible. Hereafter, unless otherwise specified, I refer to ARIMA model simply as time-series.

In hierarchical models, many problems involve multiple parameters which can be regarded as related in some way by the structure of the problem. A joint probability model for those parameters should reflect their mutual dependence. Typically, the dependence can be summarized by viewing these parameters as a sample from a common population distribution. Thus, the problem can be modelled hierarchically, with observable outcomes (y_t) created conditionally on the unknown parameters (ψ), which themselves are assigned a joint distribution in terms of further (possibly common) parameters, hyperparameters, with $\psi = (\phi,\theta,\mu,\sigma^2)$. In addition, 'common' parameters would change meaning from one model to another, so that prior distributions must change in a corresponding fashion. This hierarchical thinking may play an important role in developing computational strategies.

Given a set of k time-series, a partitioning method constructs τ partitions of the dynamic object data, where each partition represents a cluster containing at least one object and $\tau \leq k$. Let $\{M_k, k \in \mathcal{K}, M_k \in \mathcal{M}\}$ be a countable collection of k time-series, where $k = 1, 2, \ldots, m$ and M_k contains the vector of the unknown parameters δ, $\{\Delta_k, \delta_k \in \Delta_k, \Delta_k \in \Delta\}$ be the set of all possible values for the parameters of model M_k, and $f(M_k)$ be the prior probability of model M_k, the Posterior Model Probability (PMP)[3] is given by:

$$f(M_k|y) = \frac{f(M_k) \cdot f(y|M_k)}{\sum_{M_k \in \mathcal{M}} f(M_k) \cdot f(y|M_k)} \qquad with \qquad M_k \in \mathcal{M} \tag{3}$$

where $f(y|M_k)$ is the marginal likelihood corresponding to $f(y|M_k) = \int f(y|M_k, \delta_k) \cdot f(\delta_k|M_k, y) d\delta_k$ and $f(\delta_k|M_k, y)$ is the conditional prior distribution of δ_k. The conditional likelihood is obtained from the factorization:

$$\begin{aligned} f(y|\delta) &= f(y_1|\delta) f(y_2|y_1, \delta) \cdots f(y_n|y_1, y_2, \ldots, y_{n-1}, \delta) = \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \cdot exp\left\{-\frac{1}{2\sigma^2} \cdot \sum_{t=1} n(y_t - \mu_t)^2\right\} \end{aligned} \tag{4}$$

where

$$\mu_t = \begin{cases} \sum_{i=1}^p \phi_i y_{t-i} - \sum_{i=1}^p \delta_i (y_{t-i} - \mu_{t-i}) - \sum_{j=1}^q \theta_j \epsilon_{t-j} & \text{for } t = 2, \ldots, q \\ \sum_{i=1}^p \phi_i y_{t-i} - \sum_{i=1}^p \delta_i (y_{t-i} - \mu_{t-i}) & \text{for } t = q+1, \ldots, n \end{cases} \tag{5}$$

Finally, the natural parameter space and model space for (M_k, δ_k) are, respectively:

$$\Delta = \bigcup_{M_k \in \mathcal{M}} \{M_k\} \cdot \Delta_k \tag{6}$$

$$\mathcal{M} = \bigcup_{k \in \mathcal{K}} \{k\} \cdot M_k \tag{7}$$

When the size of the set of possible model solutions $\mathcal{M}$ is high dimensional, the calculation of

[3] See, for instance, Pacifico (2020b).

the integral $f(y|M_k)$ becomes unfeasible. Thus, a MCMC method is required in order to generate observations from the joint posterior distribution $f(M_k,\delta_k|y)$ of (M_k,δ_k) for estimating $f(M_k|y)$ and $f(\delta_k|M_k, y)$.

4 Conjugate Hierarchical Priors and Posterior Distributions

The main thrust of the fuzzy clustering algorithm is to find the set of clusters that gives the best partition according to some measure and assign each time-series to one or more homogeneous clusters. A fuzzy partition is an assignment of MCs to cluster such that each time-series is grouped on the basis of their dynamics. In this study, I regard the task of clustering Markov Chains (MCs) as a BMS problem. More precisely, the selected model is the most probable way of partitioning MCs according to their similarity, given the dynamic data. I use the PMP in (3) of the fuzzy partition as a scoring metric and I select the model with maximum PMP. Formally, it is done by regarding a fuzzy partition as a hidden discrete variable W. Each state W_τ of W represents a cluster of time-series and thus determines a transition matrix. Each fuzzy partition identifies a clustering model M_τ, with $p(M_\tau)$ being its prior probability. The directed link from the node W and the node containing the MCs represents the dependence of the transition matrix $y_t|y_{t-l}$, with l denoting the number of states of W. The latter is unknown, but the number ρ of available MCs imposes an upper bound, as $l \leq \rho$.

Given the model in equation (2), the full model class set is:

$$\mathscr{F} = \Big\{M_k : M_k \subset \mathscr{F}, M_k \in \mathscr{M}, k \in \mathscr{K},\ \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t\Big\} \tag{8}$$

where $\mathscr{M} = [\{k\} \cdot M_k]$ represents the natural model space for each t.

By Bayes' Theorem, the posterior probability of M_τ, given the sample $\mathscr{F}$, is:

$$\pi(M_\tau|\mathscr{F}) = \frac{\pi(M_\tau) \cdot \pi(\mathscr{F}|M_\tau)}{\pi(\mathscr{F})} \tag{9}$$

where, by construction, $M_\tau < M_k$, $\tau \leq k$, $\{1 \leq \tau \leq k\}$.

The quantity $\pi(\mathscr{F})$ is the marginal probability of the dynamic data and constant over time since all models are compared over the same data objects. In addition, since I consider informative proper priors, all models are *a priori* equally likely and thus the comparison can be based on the marginal likelihood $\pi(\mathscr{F}|M_\tau)$, which is a measure of how likely the dynamic data are if a clustering model M_τ is true. This quantity can be computed from the marginal distribution of W and the conditional distribution of $y_t|y_{t-l}$. In this context, W_τ would correspond to the cluster membership (see, for instance, Cooper and Herskovits (1992)).

The main thrust of the BHFC procedure is to identify time-series with similar dynamics. However, the variable selection problem arises when there is some unknown subset of k time-series so small that it would be preferable to ignore them. Thus, I introduce an auxiliary indicator variable $\beta = (\beta_k)$, with $\beta = (\beta_1, \beta_2, \dots, \beta_m)^{'}$, corresponding to the ARMA parameters δ_k, where $\beta_k = 1$ if δ_k is sufficiently large (presence of the time-series y_t in the clustering procedure). When $\beta_k = 0$, the variable δ_k would be sufficiently small so that the time-series y_t should be ruled out from a clustering model M_τ.

The BMS procedure entails estimating the parameters β and thus finding the $best$ subset containing the fuzzy partitions of dynamic data objects. Here, $best$ stands for the clustering model providing the most accurate group of homogeneous time-series over all (possible) candidate k series. The posterior probability that a time-series y_t is in the BHFC procedure can be simply calculated as the mean value of the indicator β. Since the appropriate value of β is unknown, one could model the uncertainty underlying variable selection by a mixture prior:

$$\pi(\beta, \mu, \sigma|y_t) = \underbrace{\pi(\beta|y_t)}_{conjugate} \cdot \pi(\mu|y_t) \cdot \pi(\sigma|y_t) \cdot \pi(\beta) \tag{10}$$

Bayesian inference proceeds by obtaining marginal posterior distributions of the components of δ as well as features of these distributions. In this context, the τ-th subset model is described by modelling β as a realization from a multivariate normal prior:

$$\pi(\beta) = N_\tau\Big(0, \Sigma_\tau\Big) \tag{11}$$

where $\Sigma_\tau = diag(\gamma_0 I_{p\cdot p}, \sigma^2 I_{q\cdot q})$ denotes the $[(p+q)\cdot(p+q)]$ covariance matrix, with γ_0 be the variance of the stationary ARIMA time-series and σ^2 be the assumed error variance. It is a restriction I assume to deal with computational problems, without loss of generality making inference on ψ. In addition, let $n_{\tau,ls}$ be the observed frequencies of transitions from $l \to s$ in cluster W_τ, with $\{l, s\}$ denoting generic states of W, the transition probability matrix of a cluster W_τ can be estimated as:

$$\hat{\pi}(\beta) = \frac{\psi_{\tau,ls} + n_{\tau,ls}}{\alpha_{\tau,l} + n_{\tau,l}} \tag{12}$$

where $\psi_{\tau,ls}$ are hyperparameters associated with the prior estimates of $\pi(\beta)$, according to the non-0 components of β, restricted to a benchmark prior $max(N, |\beta|)$, with $|\beta|$ denoting the model size[4] and $n_{\tau,l} = \sum_s n_{\tau,ls}$ referring to the number of transitions observed from state l in cluster W_τ.

The joint posterior density for the ARIMA parameters given the process y_t in (2) is:

$$\pi(\beta|y_t) = \frac{1}{(\sigma^2)^{\frac{n+2}{2}}} \cdot exp\Big\{-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \mu_t)^2\Big\} \cdot \hat{\pi}(\beta) \tag{13}$$

where μ_t has been defined in (5). About the unknown parameters μ and σ^2 contained in ψ, the complete conditional distribution is:

$$\pi(\mu|y_t) = N_\tau\Big(\frac{1}{n}\sum_{t=1}^{n}(y_t - \mu_t), \frac{\sigma^2}{n}\Big) \tag{14}$$

$$\pi(\sigma|y_t) = IG\Big(\frac{n}{2}, \frac{1}{2}\sum_{t=1}^{n}(y_t - \mu_t)^2\Big) \tag{15}$$

where the Inverse Gamma (IG) distribution is a two-parameter family of continuous probability distributions denoting the distribution of the reciprocal of a variable distributed according to the Gamma distribution, which provides the probabilities of occurrence of different possible outcomes in an experiment.

The complete conditional densities for the ϕ_i's and θ_j's are proportional to (13) and have to be sampled subject to the restriction to $C_p \cdot C_q$. Finally, given the hierarchical setup, the marginal posterior distribution $\pi(\beta)$ contains the relevant information for variable selection. Based on the data y_t, the posterior density $\pi(\beta|y_t)$ updates the prior probabilities on each of the W_τ possible clusters. Identifying each β with a submodel via $\beta_k = 1$ if and only if β_k is included, the β's with higher posterior probability will identify the most accurate group of homogeneous time-series and thus supported mainly by the data and the prior distributions. A reasonable choice might have the β_k's independent with marginal Posterior Model Size distribution:

[4] See, for instance, Pacifico (2020b).

$$\pi(\beta_k) = w_{|\beta|} \cdot \binom{k}{|\beta|}^{-1} \tag{16}$$

where $w_{|\beta|}$ denotes the model prior choice related to the Prior Inclusion Probability (PIP) with respect to the model size $|\beta|$, through which the β_k's will require a non-0 estimate or to be included in the cluster. In this way, one would weight more according to model size and, by setting $w_{|\beta|}$ large for smaller $|\beta|$, assign more weight to parsimonious models. Such priors would work well when k is either moderate (e.g., equal to or less than 15) or large (e.g., more than 15), yielding sensible results.

Finally, the exact and final solution will correspond to one of the submodels M_τ with higher log natural Bayes Factor (lBF)[5]

$$lBF_{\tau,k} = log\left\{\frac{\pi(M_\tau|Y_t = y_t)}{\pi(M_k|Y_t = y_t)}\right\} \tag{17}$$

where $\tau \leq k$. In this procedure, the lBF would also be called the log Weighted Likelihood Ratio (lWLR) factor of M_τ to M_k with the priors being the weighting functions. The corresponding scale of evidence[6] is:

$$\begin{cases} 0 < lB_{\xi,l} < 2 & \text{no evidence for submodel } M_\xi \\ 2 < lB_{\xi,l} < 6 & \text{moderate evidence for submodel } M_\xi \\ 6 < lB_{\xi,l} < 10 & \text{strong evidence for submodel } M_\xi \\ lB_{\xi,l} > 10 & \text{very strong evidence for submodel } M_\xi \end{cases} \tag{18}$$

5 MCMC Algorithm and RWD Measure

According to Pacifico (2020b), the factorization in μ and σ, given $|\beta|$, allows for the easy construction of MCMC algorithms for simulating a Markov chain:

$$\psi^{(0)}, \psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(m)} | \psi_k, y_t \quad \xrightarrow{d} \quad \pi(\beta|y_t) \tag{19}$$

[5] See, for instance, Pacifico (2020b).

[6] See, for instance, Kass and Raftery (1995).

where $\psi^{(0)}$ is automatically assigned to the model selection procedure in absence of any relationship between the ARMA parameters. The sequence in equation (19) is converging in distribution to $\pi(\beta|y_t)$ in (13) and exactly contains the information relevant to variable selection. Thus, an ergodic Markov chain in which it is embedded is:

$$\beta^{(0)},\mu^{(0)},\sigma^{(0)},\beta^{(1)},\mu^{(1)},\sigma^{(1)},\beta^{(2)},\mu^{(2)},\sigma^{(2)},\ldots, \quad \xrightarrow{d} \quad \pi(\beta,\mu,\sigma|y_t) \tag{20}$$

where $\beta^{(0)},\mu^{(0)},\sigma^{(0)}$ are automatically assigned to the model selection procedure in absence of any relationship between the ARMA parameters, where $\mu^{(0)}$ denotes the mean of ARIMA processes y_t given $\alpha \neq 0$ and $\sigma^{(0)} = \Sigma_{\hat{\phi},\hat{\theta}}$ corresponds to the asymptotic covariance matrix of Maximum Likelihood Estimates (MLEs) for ϕ and θ. The sequence in equation (20) converges in distribution to the full posterior $\pi(\beta,\mu,\sigma|y_t)$ and would correspond to an auxiliary Gibbs sequence.

In general variable selection problems, where the number of potential predictors k is small (e.g., equal to or less than 15), the sequence in equation (19) can be used to evaluate the full posterior $\pi(\beta|y_t)$ in (13). In large problems (e.g., when k is more than 15), it will still provide useful and faster information, performing more with respect to model size and being more effective than a $per-iteration$ basis for learning about $\pi(\beta|y_t)$.

The main advantage of using the conjugate hierarchical prior is that it enables analytical marging out of β and σ from $\pi(\beta,\mu,\sigma|y_t)$. Combining the likelihood (equation (13)) with the marginal (equation (11)) and conditional (equations (14) and (15)) distributions, it yields to the joint posterior:

$$\pi(\beta,\mu,\sigma|y_t) \propto |\Sigma_{\hat{\phi},\hat{\theta}}|^{-1/2} \cdot exp\Big\{-\frac{1}{2\sigma^2}\cdot|\bar{y}-\bar{\mathcal{Z}}_\tau\beta|^2\Big\}\cdot exp\Big\{\frac{1}{2}\sum_{t=1}^{n}(y_t-\mu_t)^2\Big\}\cdot\hat{\pi}(\beta) \tag{21}$$

where $\bar{y} = [y \quad 0]'$ and $\bar{\mathcal{Z}}_\tau = \Big[\mathcal{Z}_\tau \quad (\Sigma_{\hat{\phi},\hat{\theta}})^{-1/2}\Big]'$ are $2\cdot 1$ vectors, with $\mathcal{Z}_\tau$ being a $n\cdot k$ matrix whose columns correspond to the time-series components δ associated to the non-0 components of β.

Integrating out μ and σ yields:

$$\pi(\beta|y_t) \propto \pi(\beta) \equiv |\bar{\mathcal{Z}}_\tau'\bar{\mathcal{Z}}_\tau|^{-1/2}\cdot|\Sigma_{\hat{\phi},\hat{\theta}}|^{-1/2}\cdot\Big(\nu+\Sigma^2_\beta\Big)^{-(n+k+1)/2}\cdot\pi(\beta) \tag{22}$$

where $\nu = n-k-1$ are the degrees of freedom, $\pi(\beta)$ denotes the marginal distribution of the β_k's, and S_{β}^2 is the decomposition of the variance in the procedure selection and defined as:

$$S_{\beta}^2 = \bar{y}^{'}\bar{y} - \bar{y}^{'}\bar{\mathcal{Z}}_{\tau}\left(\bar{\mathcal{Z}}_{\tau}^{'}\bar{\mathcal{Z}}_{\tau}\right)^{-1}\bar{\mathcal{Z}}_{\tau}^{'}\bar{y} = y^{'}y - y^{'}\mathcal{Z}_{\tau}\left(\mathcal{Z}_{\tau}^{'}\mathcal{Z}_{\tau} + \left(\Sigma_{\hat{\phi},\hat{\theta}}\right)^{-1}\right)^{-1}\mathcal{Z}_{\tau}^{'}y \tag{23}$$

The generation of the components in equation (19) in conjuction with $\pi(\beta)$ in equation (22) can be obtained trivially as simulations of Bernoulli draws (e.g., with $\delta = 0$ for small β and $\delta \neq 0$ otherwise). The required sequence of Bernoulli probabilities can be computed fast and efficiently by exploiting the appropriate updating scheme for $\pi(\beta)$ in function of the time-series components δ:

$$\frac{\pi\left(\delta_k = 1, \beta_{(k)}|y\right)}{\pi\left(\delta_k = 0, \beta_{(k)}|Y\right)} = \frac{\pi\left(\delta_k = 1, \beta_{(k)}\right)}{\pi\left(\delta_k = 0, \beta_{(k)}\right)} \tag{24}$$

At each step of the iterative simulation from (19), one of the values of $\pi(\beta)$ in equation (24) will be available from the previous component simulation.

The attractive feature of the conjugate prior is in the availability of the exact $\pi(\beta)$ values, providing useful informations about $\pi(\beta|y_t)$. For example, the exact relative probability of two time-series, with one ($\delta_{k=1}$) and two ($\delta_{k=2}$) AR and MA lag orders, is obtained as $\left[\pi(\beta_1)/\pi(\beta_2)\right]$. This allows for more accurate identification of the high probability models among those selected. Thus, only minimal additional effort is required to obtain these relative probabilitites since $\pi(\beta)$ must be calculated for each of the visited M_{τ} models containing the allowable values of (ϕ,θ) in the execution of the MCMC algorithm.

Finally, to complete the BHFC method, I need to evaluate all possible partitions and return the one with the highest posterior probability. Since the number of possible partitions grows exponentially with the number of MCs, a heuristic method is required to make the search feasible. I use a measure of similarity between estimated transition probability matrices ($\hat{\pi}(\beta)$) to guide the search process. The resulting algorithm is called Robust Weighted Distance measure. The algorithm performs a bottom-up search by recursively merging the closest MCs, denoting either a cluster or a single time-series, and evaluating whether the resulting model is more probable than the model where these MCs are kept distinct. The similarity measure that guides the process can be any distance between probability distributions.

Let Q_1 and Q_2 be matrices of transition probabilities of two MCs, and $q_{1,ls}$ and $q_{2,ls}$ be the probabilities of the transition $l \to s$ in Q_1 and Q_2, the RWD of these two transition probabilities from

Q_1 to Q_2 is:

$$D_{rwd}(Q_1||Q_2) = \sum_{s=1}^{J} \bar{\omega}_s \frac{D(q_{1,l}, q_{2,l})}{J} \quad with \quad D(q_{1,l}, q_{2,l}) = \frac{[d(q_{1,l}, q_{2,l}) + d(q_{2,l}, q_{1,l})]}{2} \tag{25}$$

where $\bar{\omega}$ is the PMS distribution, on average, between the two probabilities q_1 and q_2 obtained by the transition probability matrix in equation (12). Here, the distance of the probability distribution $d(q_{1,l}, q_{2,l})$ is not symmetric, $d(q_{1,l}, q_{2,l}) \neq d(q_{2,l}, q_{1,l})$, and is:

$$d(q_{1,l}, q_{2,l}) = \sum_{s=1}^{J} \Big(q_{1,ls} \cdot log_2 \Big(\frac{q_{1,ls}}{q_{2,ls}} \Big) \Big) \tag{26}$$

The distance in equation (25) is an implemented version of the symmetric Kullback-Leibler Distance (KLD)[7]. More precisely, since each of the two matrices (Q_1 and Q_2) is a collection of J probability distributions and rows with the same index are probability distributions conditional on the same event, the measure of similarity that RWD uses is an average of their own PMS distribution between corresponding rows. In addition, the distance in (25) is zero when $Q_1 = Q_2$ and greater than zero otherwise. The main thrust behind the RWD measure is that merging more similar MCs, more probable homogeneous models (M_τ) shoul be found sooner and the conditional likelihood in (4) used as a scoring metric by the algorithm should increase.

6 Applications

This Section discusses an empirical application for moderate time-varying data ($k \leq 15$) and a simulated experiment for larger sets ($k > 15$) in order to display the performance and usefulness of BHFC procedure and RWD measure defined in equation (25) for effective clustering of Multiple ARIMA (MARIMA) time-series.

6.1 Real GDP Growth Rate Data

The empirical application consists of measuring the distance between MARIMA time-series by means of the proposed RWD measure between the productivity – in terms of real GDP per capita in logarithmic form – for G7 economies [Canada (CA), France (FR), Germany (DE), Italy (IT), Japan

[7] See, for instance, Do and Vetterli (2002).

(JP), the United Kingdom (UK), and the United States (US)] and for two non-G7 European countries [Ireland (IE) and Spain (ES)], spanning the period 1995$q1$ to 2016$q4$. All the k series are expressed in quarters and seasonally adjusted. All data points are obtained from OECD data source.

I consider a process $y = \{y_{it}\ ;\ i \in \mathbb{N}\ ,\ t \in \mathbb{T}\ \}$ that admits an ARIMA representation:

$$y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} \quad with \quad \epsilon_t \sim WN(0, \sigma^2) \tag{27}$$

All the series show an increasing trend and thus are not stationary over time (Figure 1). The results find confirmations in the corresponding highly significant and decreasing Auto-Correlation Functions (ACFs) in Figure 2.

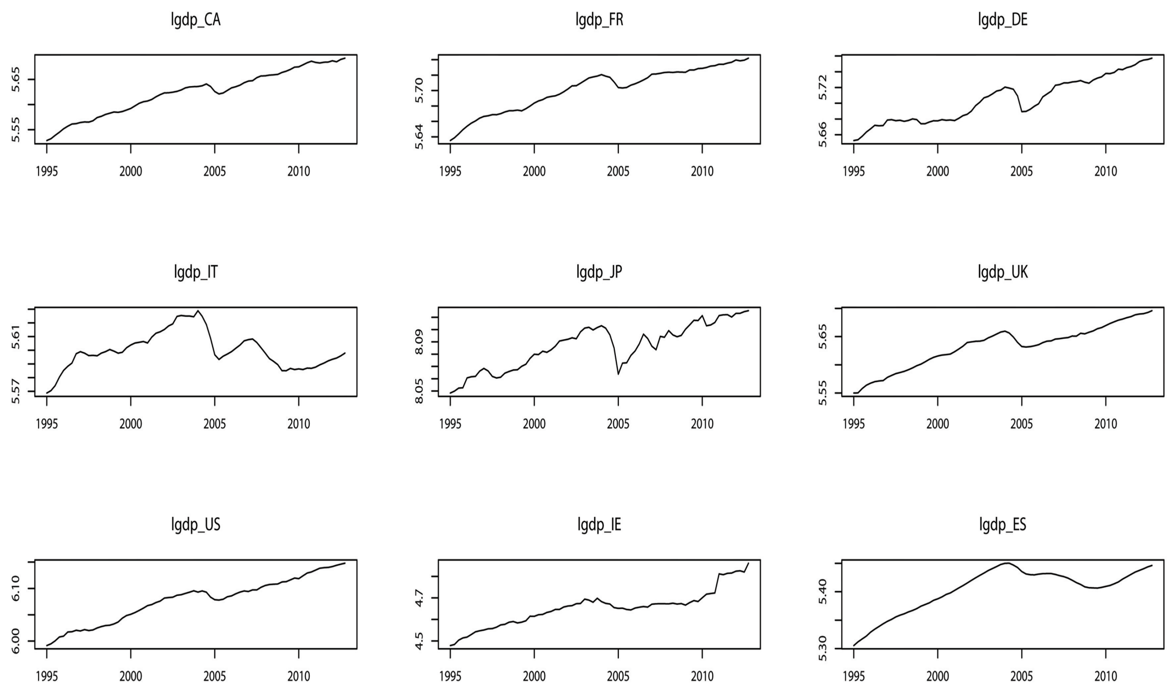


Figure 1: Time-series for the real GDP per capita for a pool of advanced European economies are shown, spanning the period 1995$q1$ to 2016$q4$. They account for G7 economies (CA, FR, DE, IT, JP, UK, and US) and for two non-G7 European countries (IE and ES). The Y and X axis represent the series and sampling time, respectively. All the series are expressed in quarters and seasonally adjusted. All data points are obtained from OECD data source.

A common model building strategy is to select the exact differentiation order and thus plausible values of AR (p) and MA (q) lag orders on statistics calculated from the data to assess the stationarity and invertibility of the process in (27) for each selected country. More precisely, I use the Bayesian Information Criterion (BIC) in equation (28) to select the optimal lag lengths in AR

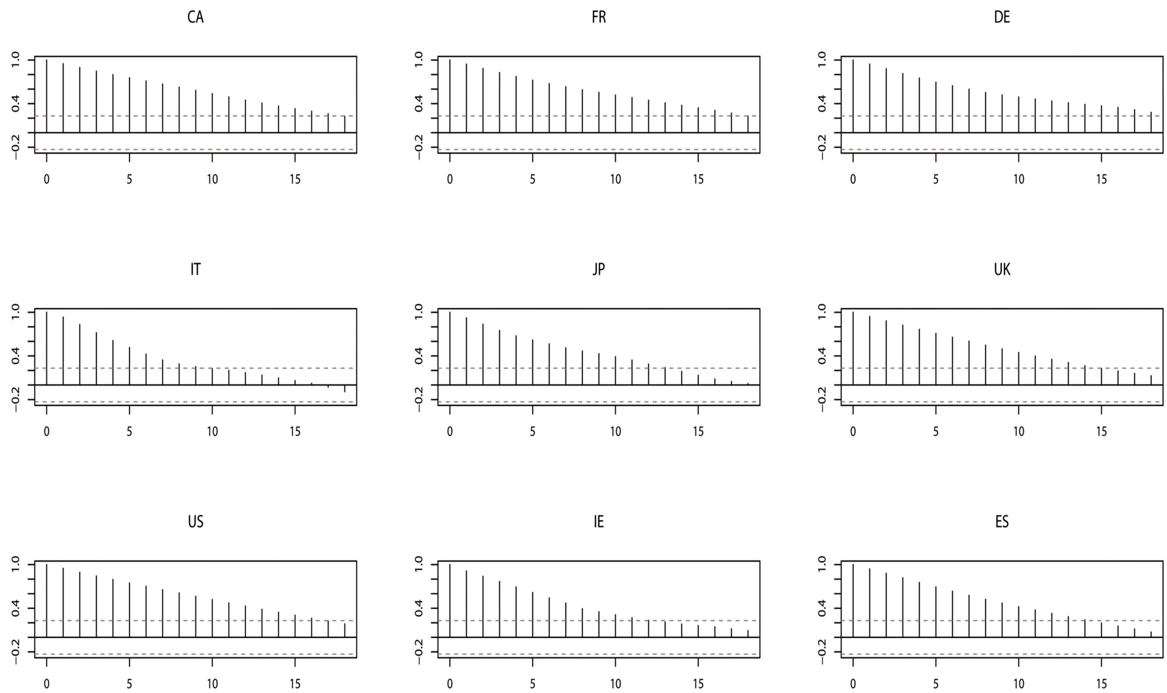


Figure 2: Sample Autocorrelation Functions for the real GDP per capita for a pool of advanced European economies are shown, spanning the period 1995q1 to 2016q4. They account for G7 economies (CA, FR, DE, IT, JP, UK, and US) and for two non-G7 European countries (IE and ES). The dashed lines display the Bartlett's bands. The Y and X axis represent the ACFs values and lags, respectively. All lags are expressed in quarters.

(p) component (see, for instance, Schwarz (1978)) and the Augmented Dickey-Fuller (ADF) test in equation (29) to choose the order of integration to ensure stationarity (see, for instance, Dickey and Fuller (1979)).

$$BIC(p,q) = log(\hat{\sigma}^2) + \frac{(p+q)\cdot log(T)}{T} \tag{28}$$

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \ldots + \delta_{p-1} \Delta y_{t-p+1} + \epsilon_t \tag{29}$$

where $\hat{\sigma}^2$ denotes the MLE of σ^2, α is a constant, β is the coefficient on a time trend, and p and q denote the lag orders of the ARIMA process in (27).

By equations (28) and (29), I estimate 9 different ARIMA processes. In Table 1, I display the ARIMA time-series, the ADF tests in terms of p-values, and Ljung-Box test statistics of the series to jointly assess the robustness of the estimates and investigate linear dependencies among series. The maximum differencing order to test stationarity sets 3 in order to highlight (possible) endogeneity issues.

Table 1: Estimates, Stationarity, and Diagnostic Test

Country	CA	FR	DE	IT	JP	UK	US	IE	ES
$ARIMA(p,d,q)$	(3,3,1)	(2,3,2)	(2,3,3)	(2,3,2)	(2,3,3)	(3,3,2)	(3,3,2)	(3,3,1)	(3,3,2)
ADF_d	.018**	.047 **	.030**	.022**	.013**	.013**	.018**	.026**	.096*
LGB_{π_1}	.00***	.00***	.00***	.00***	.00***	.00***	.00***	.00***	.00***
LGB_{π_2}	0.22	0.38	0.70	0.66	0.36	0.37	0.52	0.67	0.10

The Table is so split: the first row displays the country indices; the second row refers to ARIMA(p,d,q) models; the third row stands for the ADF tests in terms of p-values and significant codes, with $d = 3$; and the last two rows stand for Ljung-Box test statistics of the series ($\pi_1 = p$) and residuals ($\pi_2 = q$) in terms of p-values and significant codes. The significant codes are: *** significance at 1%, ** significance at 5%, and * significance at 10%.

According to Bayesian inference, higher PMP distribution (equation (3)) and log Bayes Factor (equation (17)) among series are obtained by performing a BHFC procedure with a maximum of three clusters ($c = 3$). They are so split: (i) CA, FR, DE, IT, UK, US, and ES; (ii) IE; and (iii) JP. Table 2 shows the membership values and the corresponding cluster for each series[8]. The findings address three important issues. First, when accounting for dynamics of the economy, an accurate BMA strategy, as implied in BHFC procedure, is required in order to group dynamic data objects into more probable homogenous clusters (M_τ). Second, the use of conjugate hierarchical informative priors in fuzzy clustering algorithm is able to highlight similarity – in terms of cross-country homogeneity – among series and thus group them in 'ad hoc' clusters. Third, membership values show the presence of relevant endogeneity issues (e.g., DE, FR, IT, CA, UK, SE, and US) and heterogeneity (e.g., JP and IE) among countries when performing fuzzy clustering of time-series (Table 2). Finally, given the ROB strategy implied in the procedure[9], the Conditional Posterior Sign (CPS)[10] can be obtained by the Posterior Inclusion Probabilities[11] in order to observe how the GDP time-series for each country evolve over time. Most countries tend to show negative effects (except for FR, JP, and ES) that, in a context of economic interactions and international spillover effects, would be interpreted as net receivers given an unexpected shock on GDP. Conversely, FR, JP, and ES seem to be net senders[12]. Thus, it would be interesting to extend the fuzzy clustering analysis in a multivariate context (Section 7).

[8] I use Fuzzy C-Means with 2 parameter of fuzziness, 100 random starts of the algorithm, and 100 iterations per each random start.

[9] See, for instance, Pacifico (2020b) for further specifications on the econometric methodology.

[10] The CPS refers to the posterior probability of a positive coefficient expected value conditional on inclusion. It indicates a positive effect on GDP whether it is close to 1 and a negative effect whether it is close to 0.

[11] They correspond to the sum of PMPs between all the series according to their membership values.

[12] See, for instance, Pacifico (2019a).

Table 2: Membership Values and Clusters

Country	MEMB1	MEMB2	MEMB3	Cluster	CPS
CA	0.001	0.992	0.007	2	0.007
FR	0.000	1.000	0.000	2	1.000
DE	0.002	0.989	0.009	2	0.028
IE	0.000	0.003	0.997	3	0.000
IT	0.002	0.985	0.013	2	0.000
JP	0.998	0.001	0.001	1	1.000
ES	0.010	0.838	0.152	2	0.979
UK	0.001	0.994	0.005	2	0.043
US	0.236	0.626	0.138	2	0.057

The Table is so split: the first column refers to the countries; the following three columns display the membership values; the fifth column displays the corresponding cluster membership; and the sixth column shows the CPSs.

The previous results are better highlighted graphically (Figure 3). Indeed, by focusing on the first cluster[13], it is clear the presence of consistent heterogeneity among series, but with some common components (e.g., DE, FR, IT, CA, and UK). However, a persistent homogeneity matters over time among European countries, including US (misspecified dynamics). These findings confirm the efficacy of the BMA strategy implied in the BHFC with endogeneity issues when grouping multiple dynamic processes. The remaining two clusters highlight that JP and IE series tend to evolve in a heterogeneous way over time with respect to the others.

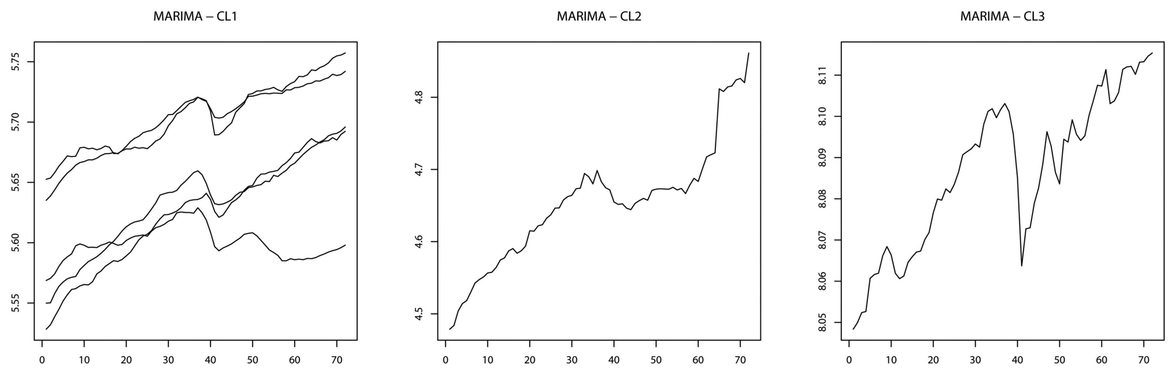


Figure 3: The real GDP per capita series are drawn and grouped according to the RWD measure, spanning the period 1995q1 to 2016q4. The three clusters are so split: (i) CA, FR, DE, IT, UK, US, and ES; (ii) IE; and (iii) JP. The Y and X axis represent the series and sampling distribution in quarters, respectively.

In Table 3, I compare the performance of the BHFC procedure with some related distance measures[14] for effective clustering of ARIMA time-series: (i) the most natural metric for computing

[13] Here, the series ES and US have been dropped to better scale the plot.

[14] Own computations.

distance in most applications such as Euclidean Distance (ED)[15]; (ii) Dynamic Time Warping (DTW)[16]; (iii) Longest Common Sub-Sequence (LCSS)[17]; and (iv) Minimal Variance Matching (MVM)[18]. Here, some considerations are in order. These clustering approaches would be classified as 'shaped-based similarity' measures, suitable for short-length time-series and used to find similar time-series in time and shape. More precisely, the ED measure is one of the most used time-series dissimilarity measures, favored by its computational simplicity and indexing capabilities. The DTW and LCSS approaches tend to be very appropriate in the case which the similarity between time-series is based on 'similarity in shape' or there are time-series with different length. They focus on the averaging method for time-series clustering and thus define a cluster by their combinations hierarchically or sequentially. However, their drawback is the strong dependency on the ordering of choosing pairs which result in different final clusters. Finally, MVM measure computes the distance value between different time-series directly based on the distances of corresponding elements, just as DTW, and allows the query sequence to match to only subsequence of the target sequence, just as LCSS. The main difference between LCSS and MVM is that LCSS requires the distance threshold to optimize over the length of the longest common subsequence, while MVM directly optimizes the sum of distances of corresponding elements without any distance threshold. Instead, the main difference between DTW and MVM is that MVM is able to skip some elements of the target series when computing the correspondence.

The main thrust of this example is to prove that RWD measure gets the higher cluster similarity metric than the other related methods by dealing with either model uncertainty and overfitting (implied in Bayesian framework) or endogeneity issues and misspecified dynamics (implied in the BMA strategy) when clustering dynamic data. All approaches would perform better by choosing two clusters. The JP series are grouped in a unique cluster with respect to the others (including the IE series). By running the lBF (equation (17)) between the submodels M_τ and the submodels related to the alternative approaches (M_*)[19], I find moderate support with DTW and MVM measures and strong evidence with LCSS measure by supporting elastic distances and unequal size time-series in fuzzy clustering approach. No evidence (or weak support) is found for ED measure due to its inability in clustering data sets containing many time series.

[15]See, e.g., Chan et al. (2003).

[16]See, e.g., Chu et al. (2002).

[17]See, e.g., Vlachos et al. (2002) and Banerjee and Ghosh (2001).

[18]See, e.g., Latecki et al. (2005).

[19]In this context, the submodels would correspond to the subsequences – which result in different clusters – between time-series having maximum similarity with other objects within that group and minimum similarity with objects in other groups.

Table 3: Performance Comparison

Distance Measure	Cluster	CSM	lBF	Evidence
RWD	3	0.923	-	-
ED	2	0.424	1.84	weak
DTW	2	0.594	5.07	moderate
LCSS	2	0.763	7.13	strong
MVM	2	0.605	5.49	moderate

The Table is so split: the first column describes the distance measure used; the second column displays the optimal number of clusters; the third column accounts for Cluster Similarity Metric; and the last two columns refer to the log Bayes Factor and the corresponding scale of evidence.

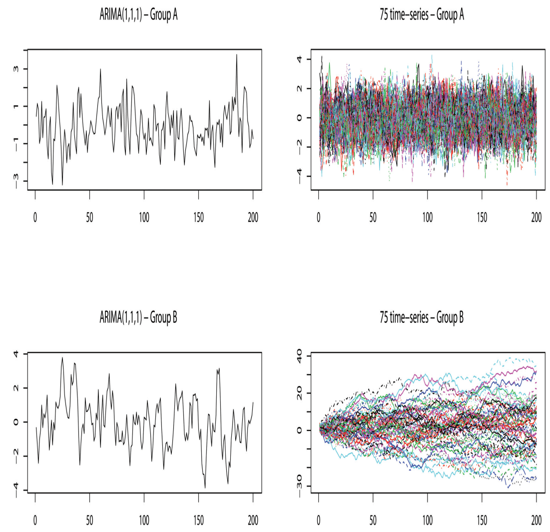
6.2 Simulated Example

I perform fuzzy clustering on a database of ARIMA(1,1,1) time-series and analyze the results. More precisely, I generate four groups (A, B, C, and D), each with $k = 75$ ARIMA(1,1,1) time-series, where $t = 1,2,\ldots,200$ and the parameter vectors (ϕ,θ) are uniformly distributed in the ranges $[(1.30,0.30)\pm0.01]$, $[(1.34,0.34)\pm0.01]$, $[(1.60,0.60)\pm0.01]$, and $[(1.64,0.64)\pm0.01]$, respectively. The white noise ϵ_t used has mean zero and variance 0.01. All the simulated time-series are stationary, invertible, and integrated of order one, $ARIMA(1,1,1) \sim I(1)$. I construct 10 collections from these series and run fuzzy clustering on each of the groups. Collections $1-5$ have been built by selecting 15 time-series, each from groups A and B. Similarly, collections $6-10$ have been built by selecting 15 time-series, each from groups C and D (Figure 4).

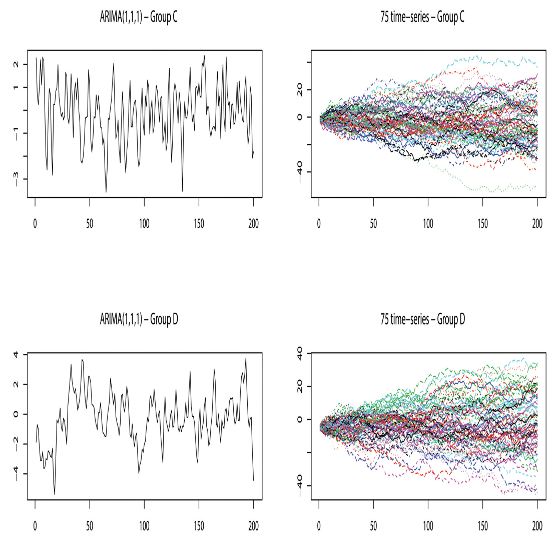
According to Bayesian inference, higher PMP distribution (equation (3)) and log Bayes Factor (equation (17)) among series are obtained by performing a BHFC procedure with a maximum of three clusters ($c = 3$).

In Table 4, I compare the cluster similarity metric obtained using RWD measure[20] with four different traditional similarity measures[21]: (i) the most popular multidimensional scaling method such as Weighted Euclidean Distance (WED)[22]; (ii) Discrete Wavelet Transform (DWT)[23]; (iii) Discrete Fourier Transform (DFT)[24]; and (iv) Principal Component Analysis (PCA)[25]. Here, some considerations are in order. These clustering approaches would be classified as 'structural level'

[20] I use Fuzzy C-Means with 2 as parameter of fuzziness, 100 random starts of the algorithm, and 100 as maximum iterations per each random start.
[21] Own computations.
[22] See, e.g., Horan (1969) and Carroll and Chang (1970).
[23] See, e.g., Struzik and Sibes (1999).
[24] See, e.g., Agrawal et al. (1993).
[25] See, e.g., Gavrilov et al. (2000).



(a) Groups A and B



(b) Groups C and D

Figure 4: Simulated ARIMA(1,1,1) time-series, with $k = 75$ and $t = 200$, from the groups A, B, C, and D. The Y and X axis represent the series and sampling time, respectively.

similarity measures, based on global and high level structure and used for long-length time-series data. More precisely, the WED depends on the combination of the weights used and the model parameters. Thus, it tends to be more sensitive to the position of the AR coefficients and very close to the cepstral distance for effective clustering of ARIMA time-series. The DWT technique is used in wide range of applications from computer graphic to speech, image, and signal processing. The advantage of using DWT is its ability of representation of time-series as multi-resolution. Additionally, the location of time and frequency can be gained by means of the time-frequency localization property of DWT. The DFT uses the Euclidean distance between time-series of equal length as the measure of their similarity. Then, it reduces their sequences into points in low-dimensional space. The approach tends to improve upon the measurement of similarity between time-series since the effects of high frequency components, which usually correspond to noise, are discarded. Finally, the PCA is a well-known data analysis technique widely used to measure time-series similarity in multivariate analysis. The particular feature of PCA is the main tendency of observed data compactly and thus the ability to be used as a method to follow up a clue when any significant structure in the data is not obvious. However, when the data does not have a structure that PCA can capture, satisfactory results cannot be obtained due to the uniformity of the data structure and thus any significant and accumulated proportion for the principal components cannot be found.

The higher cluster similarity metric is found for the RWD measure (close to 1) and thus the BHFC procedure is able to provide an accurate clustering for all of the ARIMA(1,1,1) collections, by jointly dealing with endogeneity issues and structural model uncertainty, where one or more parameters are posited as the source of model misspecification problems. The results also indicate that the dynamic data objects are clustered with a high confidence level and thus the RWD measure would be better that the other distance measures.

Table 4: Cluster Similarity Metric

Collection	RWD	WED	DWT	DFT	PCA
1	0.980	0.813	0.596	0.561	0.533
2	0.989	0.817	0.599	0.553	0.578
3	0.990	0.835	0.598	0.595	0.583
4	0.960	0.807	0.565	0.547	0.521
5	0.969	0.821	0.583	0.556	0.569
6	0.969	0.823	0.553	0.537	0.534
7	0.985	0.817	0.579	0.603	0.606
8	0.964	0.825	0.589	0.621	0.625
9	0.978	0.871	0.632	0.650	0.633
10	0.998	0.853	0.661	0.669	0.665
lBF	-	8.45	6.71	6.57	1.33
Evidence	-	strong	moderate	moderate	weak

The Table shows the cluster similarity metric for five different distance measures. The columns display all of the ARIMA(1,1,1) time-series collections and the related distance measures used for the fuzzy clustering analysis. The last two columns refer to the log Bayes Factor and the corresponding scale of evidence, respectively.

In Appendix A, I draw the clustering plots accounting for all of the 10 collections from the ARIMA(1,1,1) time-series (Figure 9). The first five collections $(1-5)$ have been built by selecting 15 time-series, each from groups A and B. The others $(6-10)$ have been built by selecting 15 time-series, each from groups C and D. All the grouped series show a highly low average dissimilarity within a cluster and a highly strong dissimilarity between clusters. Thus, the three clusters are well defined and separated. In addition, given the BMA implied in the BHFC procedure, dynamics between series are well highlighted and grouped. By plotting all of the ARIMA(1,1,1) collections within each cluster, these dynamics are better observed and, in a context of multicountry setups, they would correspond to time-varying cross-country linkages such as interdependence, heterogeneity, and commonality (Figure 10 in Appendix B).

7 RWD Measure in High Dimensional Time-Varying Multicountry Data

In this section, I extend the analysis to Vector Autoregressive (VAR) models in order to test the performance of the BHFC procedure when investigating endogeneity issues, misspecified dynamics, and linear interdependencies among multiple time-series.

The empirical application builds on Pacifico (2019a) and focuses on a simplified version of the SPBVAR accounting for 9 advanced countries, the G7 economies [Canada (CA), France (FR), Germany (DE), Italy (IT), Japan (JP), the United Kingdom (UK), and the United States (US)] and two non-G7 European countries [Ireland (IE) and Spain (ES)], and some of the core variables of the real [real GDP growth rate ($rgdpg$) and general government spending (gov)] and financial [general government debt ($debt$) and the current account balance ($curr$)] business cycles. The overall series are expressed in quarters and all data points are originated from the OECD database (Table 6).

The simplified version of the time-varying SPBVAR developed in Pacifico (2019a) takes the form:

$$Y_{it}^{\ddot{m}} = A_{it,j}^{\ddot{m}}(L)\, Y_{i,t-1}^{\ddot{m}} + \ddot{\varepsilon}_{it} \tag{30}$$

where $i, j = 1, 2, \ldots, 9$ are country indices, $t = 1, 2, \ldots, T$ denotes time, $\ddot{m} = 1, \ldots, 4$ denotes the set of endogenous variables, $A_{it,j}$ is a $36 \cdot 36$ matrix of real and financial variables for each pair of countries (i, j) for a given $\ddot{m}$, $Y_{i,t-1}$ is a $36 \cdot 1$ vector of lagged variables of interest that accounts for the real and financial dimensions for each i for a given $\ddot{m}$, and $\ddot{\varepsilon}_{it} \sim i.i.d.N(0, \ddot{\Sigma})$ is an $36 \cdot 1$ vector of disturbance terms. For convenience, I suppose one lag and no intercept.

The estimation sample covers the period from March 1995 to December 2016. It amounts, without restrictions, to $3,168$ regression parameters. More precisely, each equation of the time-varying SPBVAR in (30) has 36 coefficients and there are 88 equations in the system. According to the BMA strategy implied in the BHFC procedure, there are $2^{36} = 68,719,476,736$ possible model solutions. Thus, it is very costly to select subsets of clusters (M_τ) via MCMC algorithms since the number of coefficients is increased by $N\ddot{M}$ factors, with $\ddot{M}$ denoting the set of the lagged endogenous variables accounted for. Moreover, all series vary over time and thus dynamic relationships, cross-unit lagged interdependencies, and dynamic feedback matter. Then, in order to be able in applying the specifications underlying the BHFC procedure (Section 4), I need to express the time-varying SPBVAR in (30) in terms of a multivariate normal distribution.

By adapting the framework of Pacifico (2019a) and following the Bayesian implementations in Pacifico (2020b), I can able to express the time-varying SPBVAR in terms of multivariate normal distribution:

$$Y_t = (I_{N\ddot{M}} \otimes \ddot{X}_t)\ddot{\gamma}_t + \ddot{E}_t \tag{31}$$

where $Y_t = (Y_{1t}^{\ddot{m}'}, \dots, Y_{Nt}^{\ddot{m}'})'$ is an $36 \cdot 1$ vector containing the set of real and financial variables for each i for a given $\ddot{m}$, $\ddot{X}_t = (Y_{i,t-1}^{'\ddot{m}}, Y_{i,t-2}^{'\ddot{m}}, \dots, Y_{i,t-l}^{'\ddot{m}})'$ is an $1 \cdot \ddot{k}$ vector containing all lagged variables for each i, with $\ddot{k} = N\ddot{M}$ be the number of all matrix coefficients in each equation of the model (30) for each pair of countries (i, j), $\ddot{\gamma}_{it,j}^{\ddot{k}} = vec(\ddot{g}_{it,j}^{\ddot{k}})$ is an $N\ddot{M}\ddot{k} \cdot 1$ vector containing all columns, stacked into a vector[26], of the matrix $A_t(L)$ for each pair of countries (i, j) for a given $\ddot{k}$, and $\ddot{E}_t = (\ddot{\varepsilon}_{1t}^{'}, \dots, \ddot{\varepsilon}_{Nt}^{'})'$ is an $36 \cdot 1$ vector containing the random disturbances of the model. In model (31) there is no subscript i since all lagged variables in the system are stacked in $\ddot{X}_t$.

In this study, the implementation is adapted to the thrust of the RWD measure in merging more similar MCs and thus identifying – as fast as possible – more probable homogeneous models M_τ among dynamic data. More precisely, the coefficient vectors in $\ddot{\gamma}_t$ represent all the model solutions counted in the natural model space $\mathcal{M}$ (equation (7)), and each factor would correspond to a clustering model M_τ identifying a distinct cluster of (potential) combination of the series $\ddot{m}$. The underlying logic is to exploit the Bayesian hierarchical framework implied in the BHFC procedure in order to extend the clustering analysis by dealing with misspecified dynamics (structural uncertainty) and endogeneity issues when studying and conducting fuzzy clustering analysis in

[26] Here, the vec operator is used to transform a matrix into a vector by stacking the columns of the matrix one underneath the other, with $\ddot{g}_{it,j}^{\ddot{k}} = (A_{it,j}^{1'}, A_{it,j}^{2'}, \dots, A_{it,j}^{\ddot{M}'})'$ and $\ddot{\gamma}_t = (\ddot{\gamma}_{1t}^{'}, \ddot{\gamma}_{2t}^{'}, \dots, \ddot{\gamma}_{Nt}^{'})'$ denoting the time-varying coefficient vectors, stacked for i, for each country–variable pair.

time-varying multicountry setups. However, because the coefficient vectors in $\ddot{\gamma}_t$ vary in different time periods for each country–variable pair and there are more coefficients than data points, to avoid the dimensionality problem, I adapt the framework in Pacifico (2019a) and assume $\ddot{\gamma}_t$ has the following factor structure:

$$\ddot{\gamma}_t = \sum_{f=1}^{\ddot{F}} \ddot{G}_f \cdot \beta_{\ddot{f}t} + \ddot{u}_t \quad \text{with} \quad \ddot{u}_t \sim N(0, \Sigma_{\ddot{u}}) \tag{32}$$

where $\ddot{F} \ll N\ddot{M}\ddot{k}$ and $dim(\beta_{\ddot{f}t}) \ll dim(\ddot{\gamma}_t)$ by construction, $\ddot{G}_{\ddot{f}} = [\ddot{G}_1, \ddot{G}_2, \ldots, \ddot{G}_{\ddot{F}}]$ are $N\ddot{M}\ddot{k} \cdot 1$ matrices obtained by multiplying the matrix coefficients ($\ddot{g}^{\ddot{k}}_{it,j}$) stacked in the vector $\ddot{\gamma}_t$ by conformable matrices $\mathfrak{B}_{\ddot{f}}$ with elements equal to zero and one[27], u_t is an $N\ddot{M}\ddot{k} \cdot 1$ vector of unmodelled variations present in $\ddot{\gamma}_t$, and $\Sigma_{\ddot{u}} = \Sigma_{\ddot{e}} \otimes \ddot{V}$, where $\Sigma_{\ddot{e}}$ is the covariance matrix of the vector $\ddot{E}_t$ and $\ddot{V} = (\sigma^2 I_{\ddot{k}})$ as in Kadiyala and Karlsson (1997). The idea is to shrink $\ddot{\gamma}_t$ to a much smaller dimensional vector β_t, with $\beta_t = (\beta_{1t}', \beta_{2t}', \ldots, \beta_{\ddot{F}t}')'$ denoting the adjusted auxiliary indicator defined in Section 4 ($\beta = \{\beta_k\}$), and containing all regression coefficients stacked into a vector.

Running equations (31) and (32) for the model described in (30), I assume that the coefficient vectors in $\ddot{\gamma}_t$ depends on three factors:

$$\ddot{G}_{\ddot{f}} \beta_{\ddot{f}t} = \ddot{G}_1 \beta_{1t} + \ddot{G}_2 \beta_{2t} + \ddot{G}_3 \beta_{3t} + \ddot{u}_t \tag{33}$$

where, stacking for t, $\beta_{\ddot{f}} = (\beta_1, \beta_2, \beta_3)'$ contains all the series to be estimated and grouped in distinct clusters for each clustering model M_τ. Given the factorization in equation (33), the reduced-form SPBVAR model in equation (31) can be transformed into a Structural Normal Linear Regression (SNLR) model[28] written as

$$Y_t = \ddot{\Theta}_t \left(\sum_{\ddot{f}=1}^{3} \ddot{G}_{\ddot{f}} \beta_{\ddot{f}t} + \ddot{u}_t \right) + \ddot{E}_t \equiv \ddot{\chi}_{\ddot{f}t} \beta_{\ddot{f}t} + \ddot{\eta}_t \quad \text{with} \quad \ddot{\Theta}_t = \left(I_{N\ddot{M}} \otimes \ddot{X}_t \right) \tag{34}$$

where $\ddot{\Theta}_t$ contains all the lagged series in the system by construction, $\ddot{\chi}_{\ddot{f}t} \equiv \ddot{\Theta}_{\ddot{f}t} \ddot{G}_{\ddot{f}t}$ is an $N\ddot{M} \cdot 1$ matrix that stacks all coefficients of the system, and $\ddot{\eta}_t \equiv \ddot{\Theta}_{\ddot{f}t} \ddot{u}_t + \ddot{E}_t \sim N(0, \sigma_t \cdot \Sigma_{\ddot{u}})$, with $\sigma_t = (I_N + \sigma^2 \ddot{\Theta}_t' \ddot{\Theta}_t)$.

[27] Here, $\mathfrak{B}_{\ddot{f}}$ would correspond to the auxiliary indicator $\beta = (\beta_k)$ in matrix form.

[28] See, for instance, Pacifico (2019b) for an illustration of the conformation of the time-varying multicountry SPBVAR model and the exact form of the $\beta_{\ddot{f}t}$'s, $\ddot{\chi}_{\ddot{f}t}$'s, and the $\ddot{G}_{\ddot{f}}$'s.

Let $\ddot{M}_{\ddot{v}} = \ddot{M}_{\ddot{v}1}, \ddot{M}_{\ddot{v}2}, \ddot{M}_{\ddot{v}3}$ be the number of variable-specific factors. The specified factors in equation (33) I assume for the empirical analysis are: (i) $\ddot{\chi}_{1t}\beta_{1t}$ is a $N\ddot{M}\cdot\ddot{M}_{\ddot{v}1}$ observable cross-country variable-specific indicator for Y_t that accounts for all the series with close similarity metric in order to highlight cross-country commonality among real and financial variables, with $\ddot{\chi}_{1t} = \sum_i \sum_{\ddot{M}_{\ddot{v}1}} y_{i\ddot{M}_{\ddot{v}1}t-1}$; (ii) $\ddot{\chi}_{2t}\beta_{2t}$ is a $N\ddot{M}\cdot\ddot{M}_{\ddot{v}2}$ observable cross-country variable-specific indicator for Y_t that accounts for all the series with higher and similar membership values in order to investigate interdependence and endogeneity issues among countries and sectors, with $\ddot{\chi}_{2t} = \sum_i \sum_{\ddot{M}_{\ddot{v}2}} y_{i\ddot{M}_{\ddot{v}2}t-1}$; and (iii) $\ddot{\chi}_{3t}\beta_{3t}$ is a $N\ddot{M}\cdot\ddot{M}_{\ddot{v}3}$ observable cross-country variable-specific indicator for Y_t that accounts for all the series with similar MCs and higher log BF in order to deal with endogeneity issues and misspecified dynamics, with $\ddot{\chi}_{3t} = \sum_i \sum_{\ddot{M}_{\ddot{e}}} y_{i\ddot{M}_{\ddot{e}}t-1}$.

By construction, conjugate hierarchical informative priors (Section 4) and MCMC simulations (Section 5) can be used without loss of estimation efficiency. Nevertheless, further specifications and MCs implementations need to be done. For example, since the coefficient vectors β_t vary over time, I suppose the following state-space structure:

$$\beta_t = \beta_{t-1} + \ddot{v}_t \quad \text{with} \quad \ddot{v}_t \sim N(0, \ddot{H}_t) \tag{35}$$

where $\beta_t = (\beta_{1t}, \beta_{2t}, \ldots)'$, $\ddot{H}_t = diag(\ddot{\ddot{H}}_{1t}, \ddot{\ddot{H}}_{2t}, \ddot{\ddot{H}}_{3t})$ is a block diagonal matrix, and $\ddot{\ddot{H}}_{\ddot{f}t} = (\ddot{h}_{\ddot{f}t} \cdot I)$, where $\ddot{h}_{\ddot{f}t}$ controls the tightness (stringent conditions) of the factorization ($\ddot{f}$) of the time-varying parameters (β_t) in order to make it estimable[29]. The errors $\ddot{E}_t$, $\ddot{u}_t$, and $\ddot{v}_t$ are mutually independent.

Supposing exact factorization[30], mixture prior distributions in (12) need to be improved. Thus, let $\ddot{\phi}_0 = (\Sigma_{\ddot{e}}^{-1}, \ddot{h}_{\ddot{f}0}, \beta_0)$ be a vector containing mixture density of all the conjugate priors, the likelihood function can be derived from the sampling density $p(Y|\ddot{\phi}_0)$ by using a mixture hierarchical distribution[31]. The posterior distributions for $\ddot{\phi} = (\Sigma_{\ddot{e}}^{-1}, \ddot{h}_{\ddot{f}t}, \{\beta_t\}_{t=1}^{T})$ are calculated by combining the prior with the (conditional) likelihood for the initial conditions of the data. The resulting function for the full model M_k, containing all coefficient vectors $\ddot{\gamma}_t$ in (32), is then proportional to

[29] The random-walk assumption in (35) is very common in the time-varying VAR literature and has the advantage of focusing on permanent shifts and reducing the number of parameters in the estimation procedure.

[30] The factorization of γ_t becomes exact as long as σ^2 converges to zero.

[31] See, for instance, Pacifico (2020b) fur further specifications on prior and posterior computations.

$$L(Y^T|\ddot{\phi}, M_k) \propto (\Sigma_{\ddot{e}})^{-\frac{T}{2}} exp\Big\{-\frac{1}{2}\Big[\Sigma_t(Y_t - (\ddot{\Theta}_t\ddot{G})\beta_t)^{'}\Big]\Sigma_{\ddot{e}}^{-1}\Big[\Sigma_t(Y_t - (\ddot{\Theta}_t\ddot{G})\beta_t)\Big]\Big\} \tag{36}$$

where $Y^T = (Y_1,\ldots,Y_T)$ denotes the data, and $\ddot{\phi}$ refers to the unknowns whose joint probability distribution needs to be found. According to equations (34) and (35), dynamic analysis in BHFC procedure can be conducted by MCs implementations[32] in order to construct exact mixed posterior distributions of the time-varying β_t. More precisely, a variant of the Gibbs sampler approach is used in this analysis by making use of the Kalman filter[33], so it only requires knowledge of the conditional posterior of $(\beta_1,\ldots,\beta_T|Y^T,\ddot{\phi}_{-\beta_t})$. The result would correspond to a Bayesian Model Averaging implied in the BHFC procedure.

In this framework, dynamic feedbacks are obtained by recursively calculating the first two moments of posterior distributions on a set of $1,000$ until $5,000$ draws. The total number of draws has been $1,000+4,000 = 5,000$, which corresponds to the sum of the final number of draws to discard and save, respectively. A total of $4,000$ retained replications have been used to conduct posterior inference at each t and the convergence is obtained at about $1,000$ draws. In addition, let $\ddot{\delta} = (\ddot{z}_1, \ddot{\omega}_0, \ddot{S}_0, \ddot{\beta}_1)$ be a vector collecting all the known hyperparameters used in the empirical analysis[34], these latter are treated as fixed and are either obtained from the data to tune the prior to the specific applications (this is the case for $\ddot{\omega}_0$ and $\ddot{\beta}_1$) or selected a priori to produce relatively loose priors (this is the case for $\ddot{z}_1$ and $\ddot{S}_0$). The values are: $\ddot{z}_1 = NM = 36$, $\ddot{\omega}_0 = 9^0$, $\ddot{S}_0 = 0.91$, and $\ddot{\beta}_1 = diag(\ddot{Q}_{11},\ldots,\ddot{Q}_{1N})$, where $\ddot{Q}_{1i}$ is the estimated covariance matrix for each i.

Given the above specifications, I am able to discriminate among all the series in M_k by directly choosing a pool of *best* submodels (M_τ) that contain the only regression parameters with higher posterior means ($\hat{\beta}_{\ddot{f}t}$'s) and different from zero. Thus, I can jointly deal with overestimation of effect sizes (or individual contributions) and structured model uncertainty (implied in the procedure) without loss of estimation efficiency. The log Bayes Factor in (17) is computed as:

$$lBF_{k,\tau} = log\Big(\frac{L(Y_T|M_k)}{L(Y_T|M_\tau)}\Big) \tag{37}$$

where Y_T denotes the data and $L(Y_T|M_k)$ refers to the (conditional) likelihood function conducted on submodels M_τ by MCs implementations. Support for discovering the most probable set of

[32] See, for instance, Kaufman and Gupta (1991), Heilpern (1997), and Chen and Hsieh (2000) for interesting MC methods for fuzzy clustering analysis.
[33] See, e.g., Chib and Greenberg (1995).
[34] Own computations.

clusters capturing different dynamics and interconnections among time-varying data is obtained by comparing the marginal likelihoods of the unrestricted model (M_k) and a vector of the sub-models (M_τ).

Let Q_{im} and Q_{jm} be matrices of transition probabilities between distinct MCs among countries and sectors, and $q_{im,ls}$ and $q_{jm,ls}$ be the probabilities of the transition $l \to s$ in Q_{im} and Q_{jm}, the RWD from Q_{im} to Q_{jm} is:

$$D_{rwd}(Q^f_{im}||Q^f_{jm}) = \sum_{s=1}^{J} \ddot{\omega}^f_s \frac{D(q^f_{im,l}, q^f_{jm,l})}{J} \tag{38}$$

where $D(q^f_{im,l}, q^f_{jm,l}) = \frac{[d(q^f_{im,l}, q^f_{jm,l}) + d(q^f_{jm,l}, q^f_{im,l})]}{2}$ and $\ddot{\omega}^f$ is the PMS distribution, on average, between the probabilities $q^f_{im,l}$ and $q^f_{jm,l}$ obtained by the transition probability matrices in equation (33), with f denoting the factorization according to the needs of the investigation.

7.1 Empirical Results

By focusing on the main time-series components, a highly persistent cyclical or seasonal changes are observed among real and financial variables (Figure 5). Same results are found accounting for the ACFs for both real and financial series (Figure 6).

In Table 5, I show the membership values and clusters, and the CPS for each country and variable[35]. The optimal number of clusters in order to achieve higher PMP distribution and log Bayes Factor among series is three ($c = 3$). Three main considerations are in order. First, the results find confirmation with the preliminary findings in Pacifico (2019a), focusing on the only real and financial variables. More precisely, most countries tend to show negative and positive posterior probability signs in real economy and financial dimension, respectively. In a context of international business cycles, it would correspond to net receivers (negative CPS) and net senders (positive CPS). Second, according to the cluster membership, there is a consistent degree of heterogeneity among countries in the financial dimension and even more in the real economy. Third, the membership values are close to 1 and thus the RWD measure provides an accurate clustering among multiple series by jointly dealing with endogeneity issues and structural model uncertainty.

[35] I use Fuzzy C-Means with 2 parameter of fuzziness, 500 random starts of the algorithm, and 500 iterations per each random start.

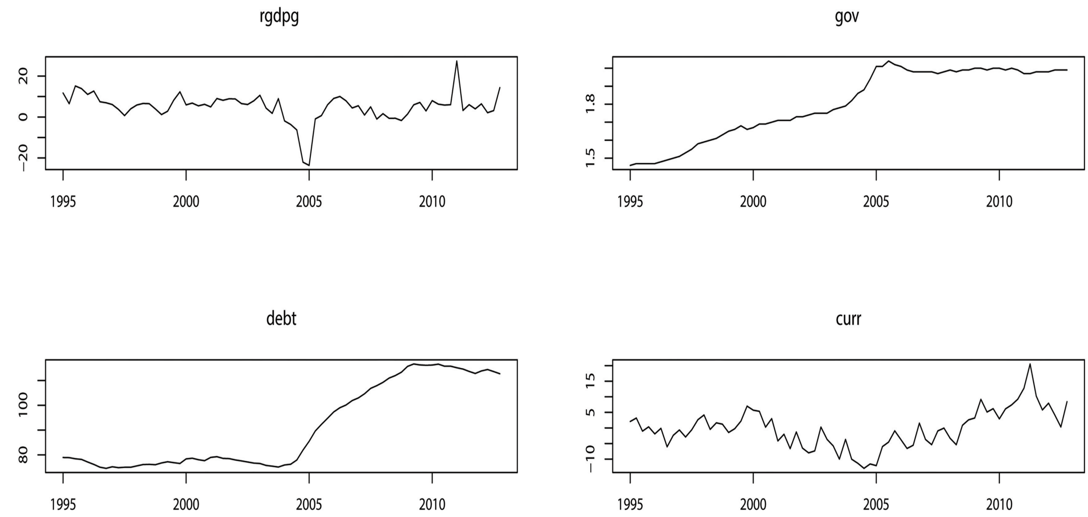


Figure 5: Time-series for the real and financial variables among a pool of advanced European economies are shown, spanning the period 1995q1 to 2016q4. They account for real GDP growth rate ($rgdpg$), general government spending (gov), general government debt ($debt$), and the current account balance ($curr$). The Y and X axis represent the series and sampling time, respectively. All the series are expressed in quarters and seasonally adjusted. All data points are obtained from OECD data source.

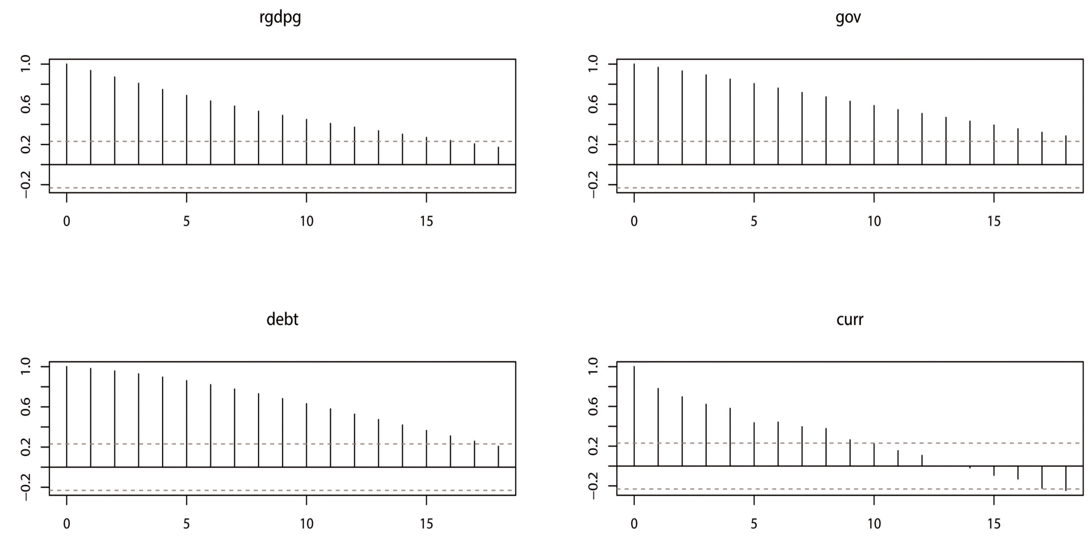


Figure 6: Sample Autocorrelation Functions for the real and financial variables among a pool of advanced European economies are shown, spanning the period 1995q1 to 2016q4. They account for real GDP growth rate ($rgdpg$), general government spending (gov), general government debt ($debt$), and the current account balance ($curr$). The dashed lines display the Bartlett's bands. The Y and X axis represent the ACFs values and lags, respectively. All lags are expressed in quarters.

The previous results are highlighted and enhanced by drawing the clustering plots (Figure 7). A relevant common component matters more in the real economy, but with stronger heterogeneity

Table 5: Membership Values and Clusters

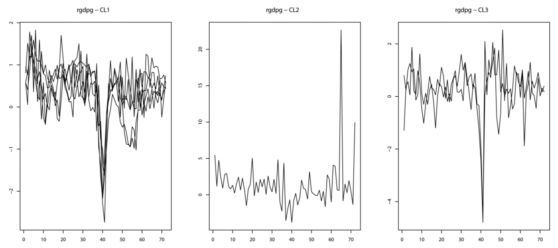
Series	Country	CA	FR	DE	IE	IT	JP	ES	UK	US
	$Memb$	0.99	0.89	0.96	0.99	0.90	0.91	0.97	0.93	0.99
$rgdpg$	$Cluster$	1	1	3	2	1	3	1	1	1
	CPS	0.00	1.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00
	$Memb$	0.92	0.98	0.90	0.92	0.96	0.93	0.99	0.97	0.94
gov	$Cluster$	3	1	3	2	3	1	3	2	2
	CPS	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
	$Memb$	0.98	0.96	0.93	0.95	0.94	0.99	0.99	0.95	0.98
$debt$	$Cluster$	1	1	2	2	1	3	2	2	1
	CPS	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00
	$Memb$	0.99	0.93	0.96	0.94	0.98	0.95	0.98	0.97	0.92
$curr$	$Cluster$	2	2	1	3	2	1	3	3	3
	CPS	1.00	1.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00

Here, $Memb$ refers to the membership values for each country and variable, $Cluster$ denotes the membership cluster for each country, and CPS stands for the Conditional Posterior Sign according to the posterior probability for each pair of country and variable.

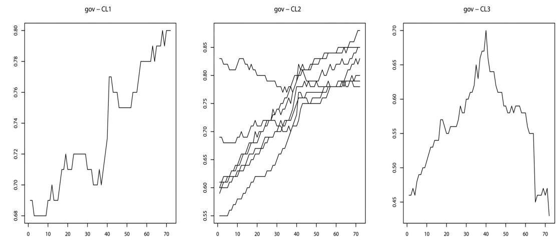
among series between clusters (Figures 7a and 7b). Larger cross-country homogeneity and interdependence are found among financial variables within and between clusters, respectively (Figures 7c and 7d). Moreover, according to Bayesian inference, some drawbacks of robust policies designed to protect the economy against the worst-case consequences of misspecified dynamics can be assessed. For example, stronger homogeneity is found among countries, except for Japan and Ireland. Then, the fuzzy clustering analysis shows two area-specific common groups and two specific clusters: (i) North American area (US, Canada); (ii) continental European area (France, Germany, Italy); (iii) Ireland; and (iv) Japan. These last two countries tend to be separate from the two areas in most cases, possibly due to different economic dynamics.

These findings are robust and consistent with the more recent business cycle studies, which recognize the importance of accounting for additional time-varying factors when jointly studying and quantifying a large set of series in dynamic multicountry setups (see, for instance, Pacifico (2019a,b) and Pacifico (2020a)). Thus, possible implementations of the BHFC procedure would be to group real and financial variables by dealing with not directly observed endogenous time-variant factors, such as transmission channels and economic–institutional linkages.

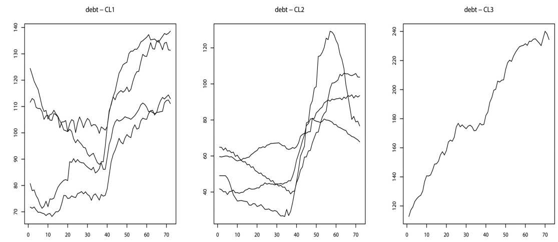
Finally, in Figure 8, I draw the clustering plots according to both real and financial variables for each of the cross-country indicators ($\ddot{\chi}_{\ddot{f}t}\hat{\beta}_{\ddot{f}t}$). They are so split: (i) the first cluster ($c = 1$) contains all the real variables and the current account balance for all countries; (ii) the second ($c = 2$)



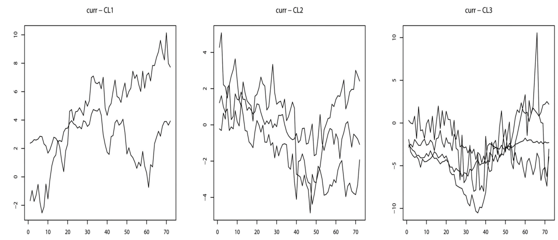
(a) Real GDP Growth Rate



(b) Government Spending



(c) Government Debt



(d) Current Account Balance

Figure 7: Clustering plots accounting for both real (plots a and b) and financial (plots c and d) variables among a pool of advanced European economies are drawn and grouped according to the RWD measure, spanning the period 1995q1 to 2016q4. The cluster membership for each series is displayed in Table 5. The Y and X axis represent the series and sampling distribution in quarters, respectively.

cluster contains the government debt for Japan, North American area, and continental European area; and (iii) the third cluster ($c = 3$) contains the government debt for the other countries. The findings can be summarized in three main results. First, larger commonality and heterogeneity matter in real economy among countries (cluster 1). Second, there are stronger cross-country interdependencies among real and financial dimension (cluster 1). It finds confirmation with the more recent literature on business cycles highlighting that an unexpected country-specific shock directly affects a country in the financial dimension and then in the real economy. Third, although relevant cross-country interdependencies among sectors, mainly in the financial dimension, the empirical analysis suggests that area-specific common factors play an important role, separating the sample into North American and continental European area, including Japan (cluster 2), with other developed economies (cluster 3).

Overall, the persistent heterogeneity observed in the real and financial dimensions highlights the relevant impact of significant endogeneity issues and misspecified dynamics among countries and sectors and thus the need for forecasters and policymakers to conduct a deeper analysis by accounting for both group-specific and global factors when formulating policy implications and feedback effects associated with temporary or persistent long-run effects[36].

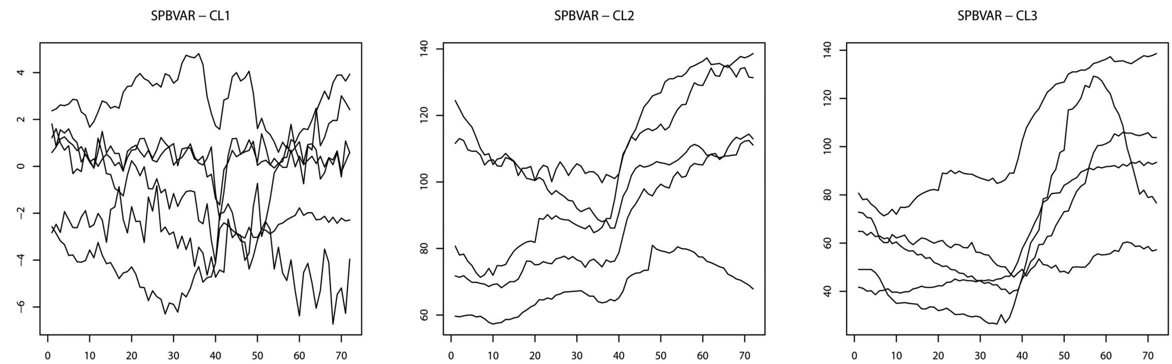


Figure 8: Clustering plots according to both real and financial variables for each of the cross-country indicators ($\ddot{\chi}_{\ddot{f}t}\hat{\beta}_{\ddot{f}t}$) are drawn. They account for clustering models M_τ identifying three distinct cluster of (potential) combination of the series $\ddot{m}$. The Y and X axis represent the series and sampling distribution in quarters, respectively.

[36] See, for instance, Pacifico (2020a), Koop and Korobilis (2012), and Raftery et al. (2010) for interesting implementations when studying macroeconomic-financial linkages in multicountry setups.

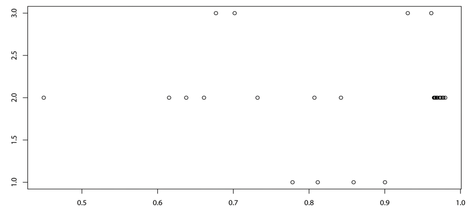
8 Concluding Remarks

The paper develops a computational approach to improve hierarchical fuzzy clustering time-series analysis when accounting for high dimensional and noise problems in dynamic data and jointly dealing with endogeneity issues and misspecified dynamics. A Robust Weighted Distance measure between stationary and invertible ARIMA processes is performed in order to make operative the notion of distance between finite sets of ARIMA models. It is *robust* because the Bayesian Model Selection is performed with a set of conjugate informative priors in order to discover the most probable set of clusters capturing different dynamics and interconnections among time-varying data and *weighted* because each unlabeled time-series is 'adjusted', on average, by own Posterior Model Size distribution in order to group dynamic data objects into 'ad hoc' homogenous clusters where the within-group-object similarity is minimize and the between-group-object dissimilarity is maximized.

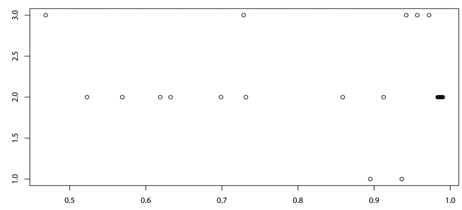
The main thrust of the BHFC procedure is the use of conjugate hierarchical informative priors in order to discover the most probable set of clusters capturing different dynamics and interconnections among time-varying data. Full posterior distributions for effective clustering of ARIMA time-series are obtained by MCMC algorithms in order to avoid the problem of increasing the overall probability of errors that plagues classical statistical methods based on significance tests. In this context, Bayesian methods are used to reduce the dimensionality of the model, structure the time variations, evaluate issues of endogeneity and structured model uncertainty, with one or more parameters posited as the source of model misspecification problems.

In this study, empirical and simulated examples describe the functioning and the performance of the procedure. More precisely, I perform an empirical application for moderate time-varying data ($k \leq 15$) and a simulated experiment for larger setups ($k > 15$) on a database of multiple ARIMA time-series in order to display the performance and usefulness of BHFC procedure and RWD measure, with k denoting the number of time-series. In addition, I extend and implement the BHFC procedure in order to deal with endogeneity issues and functional form misspecifications when accounting for dynamics of the economy in high dimensional time-varying multi-country data. In this context, the RWD measure is used in order to group multiple data objects that are generated from different series among a pool of advanced European economies. I build on Pacifico (2019a) and estimate a simplified version of the Structural Panel Bayesian VAR (SPBVAR) by defining hierarchical prior specification strategy and MCMC implementations in order to extend variable selection procedure for clustering time-series to a wide array of candidate models.

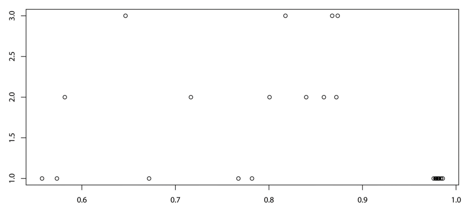
A Clustering Plot for ARIMA(1,1,1) - Cluster Membership



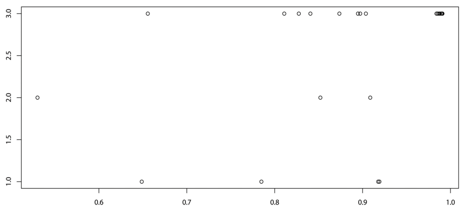
(a) Collection 1



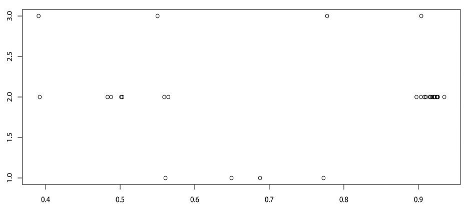
(b) Collection 2



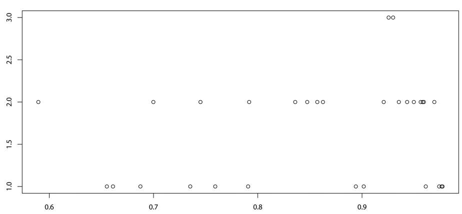
(c) Collection 3



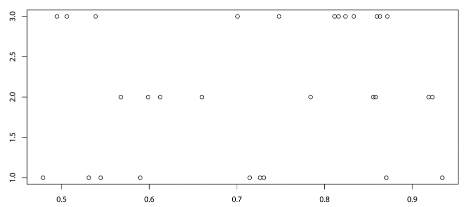
(d) Collection 4



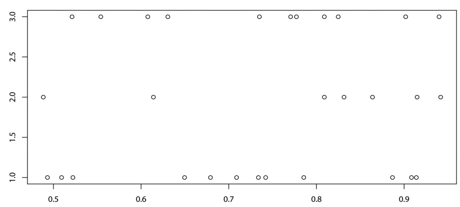
(e) Collection 5



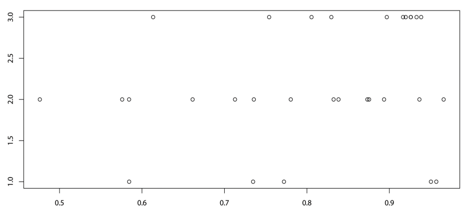
(f) Collection 6



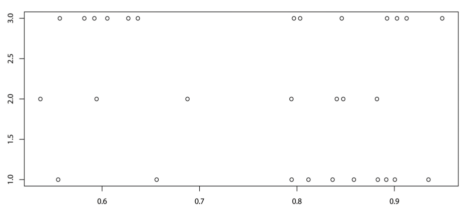
(g) Collection 7



(h) Collection 8



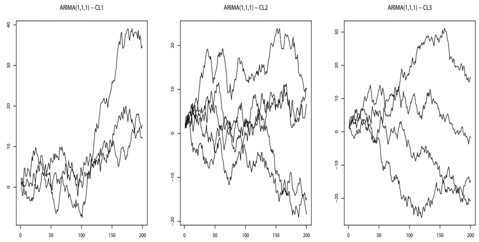
(i) Collection 9



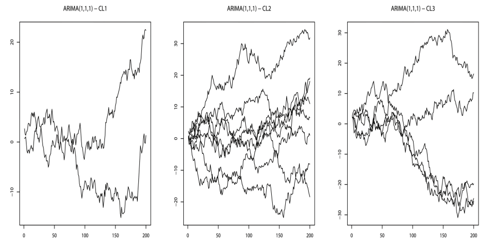
(j) Collection 10

Figure 9: Clustering plots accounting for all of the 10 collections from ARIMA(1,1,1) time-series. Collections 1−5 have been built by selecting 15 time-series, each from groups A and B. Collections 6 – 10 have been built by selecting 15 time-series, each from groups C and D. The Y and X axis represent the membership clusters and values for each series, respectively.

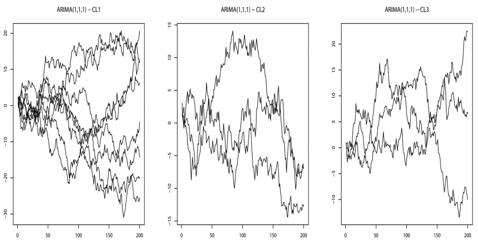
B Clustering Plot for ARIMA(1,1,1) - Series Collection



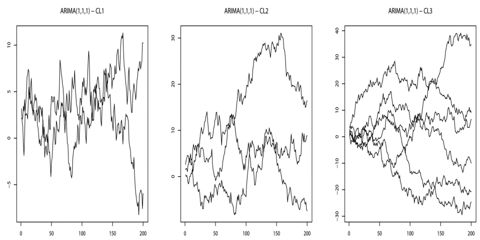
(a) Collection 1



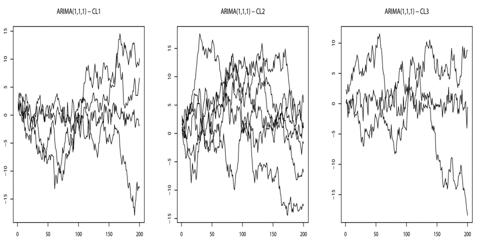
(b) Collection 2



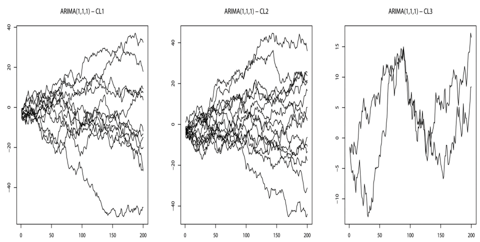
(c) Collection 3



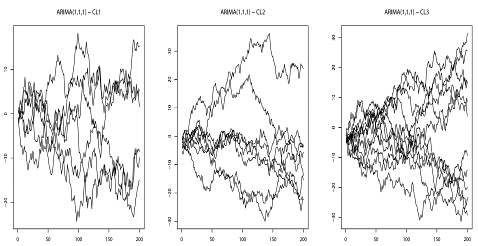
(d) Collection 4



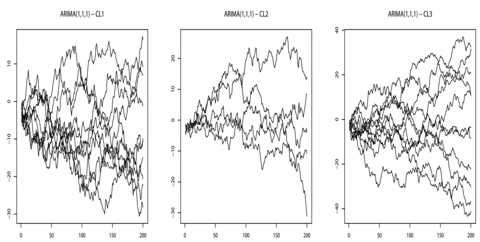
(e) Collection 5



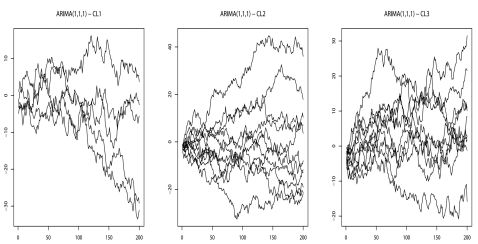
(f) Collection 6



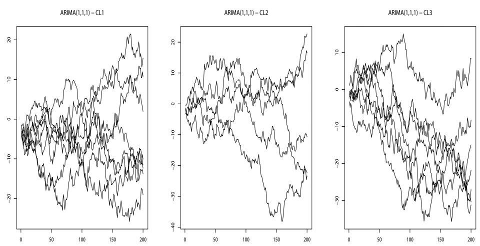
(g) Collection 7



(h) Collection 8



(i) Collection 9



(j) Collection 10

Figure 10: Clustering plots accounting for all of the 10 collections from ARIMA(1,1,1) time-series. Collections 1−5 have been built by selecting 15 time-series, each from groups A and B. Collections 6 – 10 have been built by selecting 15 time-series, each from groups C and D. The Y and X axis represent the series and sampling time, respectively.

C Data Description

Table 6: Data

VARIABLES	DESCRIPTION
GDP Growth Rate	It is calculated as: $Log\left(\frac{GDP_{it,j}}{GDP_{it-1,j}}\right)$.
General Government Spending	Financial accounts for general government spending.
General Government Debt	Non-financial accounts for general government debt.
Current Account Balance	Non-financial accounts for general government net.

Here, government spending and debt, and current account balance are weighted for the GDP.

Compliance with Ethical Standards

Funding: No funding was used for this study.

Conflict of Interest: The author declares no conflict of interest.

References

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering – a decade review. *Information Systems*, 53:16–38.

Agrawal, R., Faloutsos, C., and Swami, A. (1993). Efficient similarity search in sequence databases. *Proceedings of FODO*, pages 69–84.

Banerjee, A. and Ghosh, J. (2001). Clickstream clustering using weighted longest common subsequences. *Proceedings of SIAM Conference on Data Mining*, pages 33–40.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992a). Hierarchical bayesian analysis of change-point problems. *Journal of the Royal Statistical Society (Series C)*, 41(2):389–405.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992b). A monte carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, 87(418):493–500.

Carroll, J. D. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35:283–320.

Chan, F. K.-P., Fu, A. W.-C., and Yu, C. (2003). Haar wavelets for efficient similarity search of time-series: with and without time warping. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):686–705.

Cheeseman, J. R., Trucks, G. W., Keith, T. A., and Frisch, M. J. (1996). A comparison of models for calculating nuclear magnetic resonance shielding tensors. *The Journal of Chemical Physics*, 104(14):5497–5509.

Chen, S. H. and Hsieh, C. H. (2000). Representation, ranking, distance, and similarity of l-r type fuzzy number and application. *Australian Journal of Intelligent Processing Systems*, 6(4):217–229.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Chu, S., Keogh, E., Hart, D., and Pazzani, M. (2002). Iterative deepening dynamic time warping for time series. *Proceedings of the 2nd SIAM International Conference on Data Mining*, pages 195–212.

Cooper, G. F. and Herskovits, E. (1992). A bayesian method for the introduction of probabilistic networks from data. *Machine Learning*, 9:309–347.

Corduas, M. and Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Computational Statistics & Data Analysis*, 52:1860–1872.

Dickey, D. A. and Fuller, W. A. (1979). Distributions of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 75:427–431.

Do, M. N. and Vetterli, M. (2002). Wavelet-based texture retrieval using generalized gaussian density and kullbackâĂŞleibler distance. *IEEE Transactions on Image Processing*, 11(2):146–158.

D'Urso, P. and Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*, 160(24):3565–3589.

Gavrilov, M., Anguelov, D., Indyk, P., and Motwani, R. (2000). Mining the stock market: Which measure is best? *Proceedings of the KDD*, pages 487–496.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, 131(1-2):59–95.

Heilpern, S. (1997). Representation and application of fuzzy numbers. *Fuzzy Sets and Systems*, 91(2):259–268.

Horan, C. B. (1969). Multidimensional scaling: Combining observations when individuals have different perceptual structures. *Psychometrika*, 34:139–165.

Izakian, H. and Pedrycz, W. (2014). Agreement-based fuzzy c-means for clustering data with blocks of features. *Neurocomputing*, 127:266–280.

Izakian, H., Pedrycz, W., and Jamal, I. (2013). Clustering spatiotemporal data: An augmented fuzzy c-means. *IEEE Transactions of Fuzzy Systems*, 21(5):855–868.

Izakian, H., Pedrycz, W., and Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, 39:235–244.

Kadiyala, R. K. and Karlsson, S. (1997). Numerical methods for estimation and inference in bayesian var models. *Journal of Applied Econometrics*, 12(2):99–132.

Kannan, S. R., Ramathilagam, S., and Chung, P. C. (2012). Effective fuzzy c-means clustering algorithms for data clustering problems. *Expert Systems with Applications*, 39(7):6292–6300.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of American Statistical Association*, 90(430):773–795.

Kaufman, A. and Gupta, M. M. (1991). Introduction to fuzzy arithmetic: Theory and applications. *Van Nostrand Reinhold Co.*

Kaufman, L. and Rousseeuw, P. J. (2009). Finding groups in data: An introduction to cluster analysis. pages 349–371.

Kavitha, V. and Punithavalli, M. (2010). Clustering time series data stream – a literature survey. *International Journal of Computer Science and Information Security*, 8(1):289–294.

Keogh, E., Chakrabarti, K., Pazzani, M., and Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3:263–286.

Keogh, E. and Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371.

Keogh, E. and Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7:358–386.

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Latecki, L. J., Megalooikonomou, V., Wang, Q., Lakaemper, R., Ratanamahatana, C., and Keogh, E. (2005). Elastic partial matching of time series. *Knowledge Discovery in Databases PKDD 2005*, pages 577–584.

Li, W. K. and McLeod, A. I. (1986). Fractional time series modelling. *Biometrika*, 73(1):217–221.

Liao, T. W. (2005). Clustering of time series data - a survey. *Pattern Recognition*, 38(11):1857–1874.

Lin, J., Keogh, E., Wei, L., and Lonardi, S. (2007). Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144.

Maharaj, E. A. (1996). A significance test for classifying arma models. *Journal of Statistical Computation and Simulation*, 54:305–331.

Martin, R. J. (2000). A metric for arma processes. *IEEE transactions on Signal Processing*, 48(4):1164–1170.

Pacifico, A. (2019a). International co-movements and business cycles synchronization across advanced economies: A spbvar evidence. *International Journal of Statistics and Probability*, 8(4):68–85.

Pacifico, A. (2019b). Structural panel bayesian var model to deal with model misspecification and unobserved heterogeneity problems. *Econometrics*, 7(1):1–24.

Pacifico, A. (2020a). Fiscal implications, misspecified dynamics, and international spillover effects across europe: A time-varying multicountry analysis. *Forthcoming in International Journal of Statistics and Economics.*

Pacifico, A. (2020b). Robust open bayesian analysis: Overfitting, model uncertainty, and endogeneity issues in multiple regression models. *Forthcoming in Econometric Reviews.*

Piccolo, D. (1990). A distance measure for classifying arima models. *Journal of Time Series Analysis*, 11(2):153–164.

Poulsen, C. (1990). Mixed markov and latent markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing*, 7(1):5–19.

Raftery, A., Karny, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52(1):52–66.

Rakthanmanon, T., Campana, A. B., Batista, G., Zakaria, J., and Keogh, E. (2012). Searching and mining trillions of time series subsequences under dynamic time warping. *Proceedings of the Conference on Knowledge Discovery and Data Mining*, pages 262–270.

Ramoni, M., Sebastiani, P., and Cohen, P. (2002). Bayesian clustering by dynamics. *Machine Learning*, 47(1):91–121.

Rani, S. and Sikka, G. (2012). Recent techniques of clustering of time series data: A survey. *International Journal of Computer Applications*, 52(15):1–9.

Ratanamahatana, C., Keogh, E., Bagnall, A. J., and Lonardi, S. (2005). A novel bit level time series representation with implication of similarity search and clustering. *Advances in Knowledge Discovery and Data Mining*, pages 771–777.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Struzik, Z. and Sibes, A. (1999). Measuring time series similarity through large singular features revealed with wavelet transformation. *Proceedings of the 10th Intl. Workshop on Database and Expert Systems Appl.*, pages 162–166.

Triacca, U. (2016). Measuring the distance between sets of arma models. *Econometrics*, 32(4):1–11.

Visser, I., Raijmakers, M., and Molenaar, P. (2000). Confidence intervals for hidden markov model parameters. *British Journal of Mathematical and Statistical Psychology*, 53:317–327.

Vlachos, M., Kollios, G., and Gunopulos, D. (2002). Discovering similar multidimensional trajectories. *Proceedings of 18th International Conference on Data Engineering*, pages 673–684.

Zakaria, J., Rotschafer, S., Mueen, A., Razak, K., and Keogh, E. (2012). Mining massive archives of mice sounds with symbolized representations. *SIAM International Conference on Data Mining*, pages 1–10.

Zellner, A. (1983). Applications of bayesian analysis in econometrics. *Journal of the Royal Statistical Society (Series D)*, 32(1-2):23–34.