Hierarchical Bayesian Fuzzy Clustering Approach for High Dimensional Linear Time-Series

Pacifico, Antonio

University of Macerata, LUISS Guido Carli University, University of Bruxellex

December 2022

Hierarchical Bayesian Fuzzy Clustering Approach for High Dimensional Linear Time-Series

Antonio Pacifico*

Abstract

This paper develops a computational approach to improve fuzzy clustering and forecasting performance when dealing with endogeneity issues and misspecified dynamics in high dimensional dynamic data. Hierarchical Bayesian methods are used to structure linear time variations, reduce dimensionality, and compute a distance function capturing the most probable set of clusters among univariate and multivariate time-series. Nonlinearities involved in the procedure look like permanent shifts and are replaced by coefficient changes. Monte Carlo implementations are also addressed to compute exact posterior probabilities for each cluster chosen and then minimize the increasing probability of outliers plaguing traditional clustering time-series techniques. An empirical example highlights the strengths and limitations of the estimating procedure. Discussions with related works are also displayed.

Keywords: ARIMA Models; Forecasting; Distance Measures; Bayesian Model Averaging; Monte Carlo Algorithms; Dynamic Data.

*Corresponding author: Antonio Pacifico, Assistant Professor at the University of Macerata, Department of Economics and Law, Macerata (IT). Email: antonio.pacifico@unimc.it. ORCID: https://orcid.org/0000-0003-0163-4956

1 Introduction

Let a set of linear time-series be serially correlated, it is often desirable to extract their significant features and then determine groups (or clusters) of similar time-varying data during different periods from a particular process or from more than one (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)). Thus, appropriate distance functions to evaluate similarities and dissimilarities between models become crucial – mainly dealing with high dimensional dynamic data – and have a significant impact on the clustering algorithms. Their selection may depend upon the nature of the data and the specificity of the application.

In that context, hierarchical fuzzy clustering holds a relevant competitive position referring to a data mining technique where similar data are placed into related or homogeneous groups without advanced knowledge of the groups' definitions. Moreover, one data object is allowed to be in more than one cluster to a different degree, and then it is very useful to quantify similarities and dissimilarities of time-series (see, for instance, Izakian et al. (2013), Izakian and Pedrycz (2014), and D'Urso and Maharaj (2009)). Fuzzy C-Means (FCM) and Fuzzy C-Medoids (FCMdd) are the two well-known and representative fuzzy clustering methods, whose objective is to minimize a weighted sum of distances between data points and cluster centers (see, e.g., Kannan et al. (2012), Keogh et al. (2001), Izakian et al. (2015), Kaufman and Rousseeuw (2009), and Liao (2005)).

However, high dimensional data and noise problems – e.g., the difficulty in obtaining appropriate measures or accurately identifying the correct model to represent the data – are characteristics of most time-series, entailing inconsistent estimates and inaccurate forecasts. Thus, dimensionality reduction methods are usually used in whole time-series clustering in order to address these issues and promote forecasting performance by transforming time-series to a lower dimensional space or by feature extraction (see, e.g., Keogh and Kasetty (2003), Keogh and Ratanamahatana (2005), and Ghysels et al. (2006)). Nevertheless, there 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. For instance, when clustering by dynamics and measuring the distance between multiple time-series, unintuitive results may be obtained since some distance measures may be highly sensitive to some distortions in the data. 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., Izakian et al. (2015), Keogh and Ratanamahatana (2005), and Ratanamahatana et al. (2005) (dynamic time warping); and Izakian et al. (2013), Izakian and Pedrycz (2014), and D'Urso and Maharaj (2009) (Euclidean distance and weighted Euclidean distance-based approach)).

The method proposed in this paper consists of overtaking these drawbacks in order to improve fuzzy clustering linear time-series and forecasting performance in either high dimensional univariate or multivariate model settings. The former are addressed through Auto-Regressive Integrated Moving Average (ARIMA) processes better fitting with the Bayesian (parametric) hierarchical framework for describing a wide variety of dynamic data and their correlations (see, e.g., Singh and Mahmoud (2019), Triacca (2016), Zakaria et al. (2012), and Rakthanmanon et al. (2012)). Multivariate observations are involved accounting for Structural Panel Vector Auto-Regressive (SPVAR) models in order to investigate cross-country heterogeneity, interdependence, and commonality among different high dimensional endogenous factors. The computational approach takes the name of Hierarchical

Bayesian Fuzzy Clustering (HBFC). Its main thrust is to deal with endogeneity issues and misspecified dynamics among different time-series and perform accurate forecasting. It would be useful when studying – for example – multiple macroeconomic–financial data serially correlated among them.

Methodologically, a hierarchical Bayesian approach is applied to estimate the dependency structure of stationary linear time-series. It usually consists of choosing a class of parametric functions, based on some `a priori` information about the field under study, and estimating the unknown parameters from a set of observations. In this way, the proposed methodology is able to avoid the problem of making critical dependency assumptions by the possibility of choosing a `best` model solution (or suitable set of clusters). Here, `best` stands for the clustering model providing the most accurate group of homogeneous time-series – over all (possible) candidate set of clusters – which involves accurate forecasting performance. Conjugate Informative Proper Mixture (CIPM) priors are used to discover the most probable set of clusters capturing different unmodelled (or misspecified) dynamics and interactions among time-varying data. The CIPM priors are an implementation of the conjugate informative proper priors in Pacifico (2020b) and act as a `strong` model selection when dealing with high dimensional model classes, where `strong` highlights the ability to minimize the probability of outliers plaguing traditional clustering time-series techniques. Markov Chain Monte Carlo (MCMC) algorithms and implementations are performed to compute exact posterior probabilities for each cluster chosen and then address accurate forecasts. However, the investigation of multiple structural breaks (or change-points) and related nonlinearities are unfeasible due to the state-space structure supposed for the time-varying data, where volatility changes look like permanent shifts and are replaced by coefficient changes.

The contributions of this paper are threefold. First, I build on and implement the Pacifico (2020b)'s analysis, who develops a robust open Bayesian procedure in two stages for implementing Bayesian Model Averaging (BMA) and Bayesian Model Selection (BMS) when accounting for dynamics of the economy in either time-invariant moderate data or time-varying high dimensional multivariate data. More precisely, I implement the prior specification strategy in a dynamic context in order to make inference on univariate and multivariate high dimensional settings, and then obtain the `best` subsets of clusters where the within-group-object similarity is minimized (e.g., because of misspecified dynamics) and the between-group-object dissimilarity is maximized (e.g., because of endogeneity issues). Then, the subset with the highest Bayes Factor would correspond to the final solution containing the most suitable subset of clusters quantifying similarities and dissimilarities of time-series. Finally, because the posterior probability of a partition is the scoring metric, it avoids the problem of increasing the overall probability of errors that plagues frequentist statistical methods based on significance tests.

Second, in most partition-based fuzzy clustering time-series techniques, when grouping 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. Thus, by using raw time-series, one may cluster them because similar in noise rather than in shape. This study overtakes these problems by performing a Robust Weighted Distance (RWD) measure to group finite sets of ARIMA processes (univariate case) and multicountry dynamics for SPVAR models (multivariate case). It is *robust* because the BMS is performed with a set of CIPM priors in order to discover the most probable set of clusters capturing different dynamics and interconnections among time-varying data, and *weighted* because each unlabelled time-series is 'adjusted', on average, by own Posterior Model Size (PMS) distribution in order to group dynamic data objects into 'ad hoc' homogenous clusters. They correspond to the sum of Posterior Inclusion Probability between all grouped time-series according to their membership values.

Third, a MCMC approach is used to move through the model space and the parameter space at the same time in order to transform time-series to a lower dimensional space and compute exact posterior probabilities for each cluster chosen. Better evidence-based forecasting is involved in HBFC because of two main features: the use of a hierarchical Bayes approach with informative mixture priors and dimensionality reduction. The latter is greatly important in fuzzy clustering time-series analysis because: (*i*) it reduces memory requirements as all time-series cannot fit in the main memory; and (*ii*) distance calculation among dynamic data is computationally expensive and thus dimensionality reduction significantly speeds up clustering.

Empirical examples describe the functioning and the performance of the estimating procedure. More precisely, I build on Pacifico (2020a) and perform an empirical experiment for large time-varying data ($k > 15$) on a database of Multiple ARIMA (MARIMA) models, with k denoting the number of time-series. Then, I address an empirical case-study extending and implementing the HBFC procedure in high dimensional time-varying multicountry data. A simplified version of the SPBVAR is accounted for grouping multiple data objects to a wide array of candidate models that are generated from different series among a pool of advanced European economies.

The outline of this paper is as follows. Section 2 discusses the proposed methodology with related works. Section 3 introduces the computational approach, the Bayesian inference, and their features. Section 4 illustrates the dynamic analysis describing prior specification strategy, posterior distributions, and MCMC algorithms. Section 5 discusses an empirical application for effective clustering of MARIMA time-series and better forecasts. Section 6 extends the methodology in a high dimensional multivariate context. The final section contains some concluding remarks.

2 Discussion with Related Works

This study is linked to several strands of the literature in Dynamic Linear Models (DLMs). As regards Bayesian approaches and tools including MCMC algorithms, closely related works addressing forecasting techniques and dynamic modelling are West and Harrison (1999), West et al. (1985), Cargnoni et al. (1996), Frühwirth-Schnatter (1994), Harrison and West (1991), Durbin and Koopman (2012), and Prado and West (2010) (normal and generalised linear models); Carlin et al. (1992c), Carter and Kohn (1994), McCulloch and Tsay (1994), and Pole and West (1990) (linear and non-linear state-space modelling); and Quintana and West (1987) (multivariate DLMs). Overall, DLMs provide a probabilistic information on the parameters and observables at any given time in order to incorporate additional information relevant to the development of the time-series, and include a parametric (or state-space) estimates and point forecasts by modelling joint distributions and future values of a time-series. However, they suffer from some basic and – from a practical viewpoint – highly important drawbacks. For instance, the DLMs focus on a sequential model definition for time-series which describes how the parameters change in time conditional on existing information. Thus, if state-space models do not provide good forecasts, they might not be taken too seriously although offering good descriptions of the data generation process. Moreover, the complexity of the likelihoods is such as exact inferences about observed relationships (estimations) and further observations (predictions) are precluded, and then the extent to which they are adequate in any particular application is usually unclear. Finally, possible extensions involving issues of endogeneity and unmodelled dynamics would increase the complexity of the analysis by requiring more special treatments and extensions (either theoretical or methodological).

Three key features matter concerning the HBFC methodology. First, CIPM prior distributions used to obtain the most probable set of clusters deal with dynamic interactions and feedback effects among time-varying data for better forecasts and predictive distributions. Second, the RWD measure implicit in the procedure acts as a `strong` model selection and then able to group unmodelled dynamic data objects into 'ad hoc' homogenous clusters. Third, exact and consistent posterior distributions for each subset of clusters are computed by jointly modelling the model space (because of high dimensional and noise problems) and the parameter space (because of endogeneity issues and functional forms of misspecification).

This study is also correlated to the literature concerning Bayesian model probabilities and averaging with dynamic data. Some main studies developing Bayesian approaches to forecasting are Harrison and Stevens (1976), Hoeting et al. (1999), Doan et al. (1992), Giannone et al. (2015), Huerta and West (1999), and Kleibergen and Paap (2002) (Bayesian forecasting and prior specification in the selection of the model); Litterman (1986a,b), Gamerman and Migon (1993), and Prado and West (1997) (forecasting with Bayesian Vector autoregressive models); Albert and Chib (1993), Kitegawa (1987), McCulloch and Tsay (1993), and Zellner et al. (1991) (Bayesian inference and prediction in autoregressive time-series); and West (1992a,b) (Bayesian mixture models). In this context, the forecasting problem is dealt with two separate stages: (*i*) estimating the current parameters and observables through a prior specification strategy of the model probabilities according to the availability of the data; and (*ii*) extrapolating this information forward in time to make inferences and then construct posterior model probabilities. However, under ideal conditions, only a small neighborhood of model configurations will be consistent with both the observed data and the domain expertise one encodes in the prior model, resulting in precise inferences and accurate posterior distributions strongly concentrated along each parameter (`identifiability`). Under more realistic conditions, measurements and domain expertise might be much less informative allowing posterior distributions to stretch across more expansive and complex neighborhoods of the model configuration space. These intricate uncertainties would complicate not only the utility of these inferences, but also the ability to quantify them computationally (`degeneracy`). More precisely, the `degeneracy` denotes a qualitative description of how strongly and uniformly a realized likelihood function or posterior density function concentrates around a single point in the model configuration space. The more uniformly a realized likelihood function concentrates around a single point, the more informed inferences about the current parameter values will be.

More recently, similar to this paper, Yao et al. (2018) (linear regressions and mixture models), McAlinn and West (2019) (univariate time-series), and McAlinn et al. (2020) (multivariate settings) have discussed and addressed these problems by developing a formal Bayesian framework for synthesizing densities in a dynamic context. They implement Bayesian predictive synthesis in a sequential and dynamic setting for density forecast combination based on agent opinion analysis theory. Their extension involves dynamic latent factor models in which sequences of forecast densities define time-varying priors for inherent latent factor processes linked to the time-series of interest. However, the standard practice of 're-normalizing' latent variables for each period is over-identifying and restrictive when addressing a set of lagged dynamic and endogenous variables.

The proposed methodology differ to the previous studies because of: (*i*) the use of a multidimensional (or panel data) structural framework for multivariate time-series in order to group similar time-series distinguishing from homogeneity, interdependency, and commonality; and (*ii*) the construction of MCMC algorithms in order to obtain a reduced set of homogeneous clusters for better conditional forecasts dealing with overfitting (or overestimation of effect sizes) and dimensionality reduction.

Finally, this paper is related to some alternative techniques dealing with high dimensional time-series such as Harrison and West (1987), Aguilar and West (2000), West (2003), West and Harrison (1999), and Lopes and West (2004) (dynamic factor models); and De Mol et al. (2008) (sparse models). Concerning dynamic factor models, they build on informative prior distributions for model parameters based on a rather informal look at some initial data, and develop MCMC methods of model fitting and computation in the chosen class of dynamic factor models. However, in high dimensional time-varying data, the choice of appropriate f-factor models – where f refers to a specific number of factors – is not immediate and easily achievable due to the related issues of model uncertainty and specification. More precisely, whether f is too large, MCMC algorithms would not work well by experiencing convergence difficulties in models. Conversely, choices of f sufficiently small would overlook important information on time-series distribution falling into issues of endogeneity. In this study, the optimal number of clusters is chosen by performing a model selection on all possible model solutions according to similarity in dynamics and components.

On the other hand, sparse models focus on Bayesian regression methods where the setting of predictors is performed using double-exponential priors (variable selection) instead of Gaussian priors (variable aggregation). More precisely, under Gaussian case, the posterior distribution generating coefficients is maximized, implying that all variables in the panel are given non-zero coefficients. Conversely, under double-exponential case, one would put more mass near zero and in the tails, allowing those coefficients to be either large or zero. The regressors, as in principal component analysis, are linear combinations of all variables in the panel. However, further implementations need to be addressed because the variable selection problem involved in the panel is not clearly interpretable and developable, mainly in dynamic data where the selected variables tend to change over time implying 'parameter instability'. By construction, the Bayesian approach proposed in this paper avoid it acting as a strong model selection and then capturing the most probable set of clusters among time-varying data dealing with issues of model uncertainty and endogeneity.

3 Econometric Model Specification

3.1 Hierarchical Framework and RWD Measure

A stochastic process $(x_t)_{t \geq -p-d}$ is said to be an ARIMA(p, d, q) if it satisfies the following equation:

$$\Phi(B)(1-B)^d x_t = \alpha + \Theta(B)\epsilon_t \quad \forall t \geq 0 \tag{1}$$

The stochastic process x_t also writes:

$$\left(1-\sum_{i=1}^{p}\phi_i B^i\right)y_t=\alpha+\left(1+\sum_{j=1}^{q}\theta_j B^j\right)\epsilon_t \tag{2}$$

where $y_t = \Delta^d x_t = (1-B)^d x_t$ is asymptotically equivalent to an ARMA(p,q) process, B is the backward shift operator, $d = 1,\ldots,\tilde{d}$ refers to higher differentiation order to obtain a stationary time-series, $t = 1,2,\ldots,n$ denotes time periods, $i = 1,2,\ldots,p$ and $j = 1,2,\ldots,q$ denote generic Auto-Regressive (AR) and Moving Average (MA) lag orders, respectively, $\Phi_i(B) = (1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p)$ and $\Theta_j(B) = (1+\theta_1 B+\theta_2 B^2+\ldots+\theta_q B^q)$ represent the AR and MA components, respectively, α refers to a constant term, and $\epsilon_t \sim WN(0,\sigma^2)$ is a *i.i.d.* Gaussian white noise process. Hereafter, unless otherwise specified, I refer to ARIMA model simply as time-series.

In equation (2), 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 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 time-series components are 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 mean (μ), variance (σ^2), and the ARMA components.

In Bayesian framework, given a stationary and invertible time-series of the form (2), the region $C_p \cdot C_q$ determines the ranges of integration for obtaining joint and marginal distributions of the parameters and 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.

Let $k = 1,2,\ldots,m$ denote linear time-series, I define an $m\nu \cdot 1$ vector $\delta_{kt} = vec(\gamma_{k,t})$ containing (stacked) all AR and MA components for each time-series for a given ν, with $\nu = i+j$, $\gamma_{k,t} = (\phi'_{i1,t},\phi'_{i2,t},\ldots,\phi'_{im,t},\theta'_{j1,t},\theta'_{j2,t},\ldots,\theta'_{jm,t})'$, and $\delta_t = (\delta'_{1t},\delta'_{2t},\ldots,\delta'_{mt})'$ denoting the time-varying parameters, stacked for k, for each time-series. With these specifications, I can extend equation (2) to MARIMA models using a simultaneous-equation form:

$$y_t = \dot{x}_t\delta_t + \tilde{\epsilon}_t \tag{3}$$

where $\dot{x}_t = (I_k \otimes \tilde{x}_t)$ contains all lagged y_t's and random disturbances within the MARIMA processes, with $\tilde{x}_t = (y'_{1,t-i}, y'_{2,t-i},\ldots,y'_{m,t-i},\epsilon'_{1,t-j},\epsilon'_{2,t-j},\ldots,\epsilon'_{m,t-j})'$ being a $1 \cdot m\nu$ vector, and $y_t = (y'_{1,t},y'_{2,t},\ldots,y'_{m,t})'$ and $\tilde{\epsilon}_t = (\epsilon'_{1,t},\epsilon'_{2,t},\ldots,\epsilon'_{m,t})'$ are $m \cdot 1$ vectors containing variables of interest and random white noises for each k.

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 $(\phi,\theta,\mu,\sigma^2)$, which themselves are assigned a joint distribution in terms of further (possibly common) parameters, called hyperparameters. 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 M_k contains the vector of the unknown parameters δ_t, $\{\Delta_k, \delta_{k,t} \in \Delta_k, \Delta_k \in \Delta\}$ 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) is given by:

$$f(M_k|y) = \frac{f(M_k) \cdot f(y|M_k)}{\sum_{M_k \in \mathcal{M}} f(M_k) \cdot f(y|M_k)} \qquad \text{with} \qquad M_k \in \mathcal{M} \tag{4}$$

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

$$\begin{aligned} f(y_t|\delta_t) &= f(y_1|\delta_t) f(y_2|y_1, \delta_t) \cdots f(y_n|y_1, y_2, \ldots, y_{n-1}, \delta_t) = \\ &= \Big(2\pi\sigma^2\Big)^{-\frac{n}{2}} \cdot exp\Big\{ -\frac{1}{2\sigma^2} \cdot \sum_{t=1} n (y_t - \mu_t)^2 \Big\} \end{aligned} \tag{5}$$

where

$$\mu_t = \begin{cases} \sum_{i=1}^{p} \delta_{-\theta_{j,t}} y_{t-i} - \sum_{i=1}^{p} \delta_{-\theta_{j,t}} (y_{t-i} - \mu_{t-i}) - \sum_{j=1}^{q} \delta_{-\phi_{i,t}} \epsilon_{t-j} & \text{for } t = 2, \ldots, q \\ \sum_{i=1}^{p} \delta_{-\theta_{j,t}} y_{t-i} - \sum_{i=1}^{p} \delta_{-\theta_{j,t}} (y_{t-i} - \mu_{t-i}) & \text{for } t = q+1, \ldots, n \end{cases} \tag{6}$$

where $\delta_{-\theta_{j,t}}$ and $\delta_{-\phi_{i,t}}$ are the time-series components excluding MA and AR lag orders, respectively.

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

$$\Delta = \bigcup_{M_k \in \mathcal{M}} \{M_k\} \cdot \Delta_k \tag{7}$$

$$\mathcal{M} = \bigcup_{k \in \mathcal{K}} \{k\} \cdot M_k \tag{8}$$

When the size of the set of possible model solutions $\mathcal{M}$ is high dimensional, the calculation of the integral $f(y_t|M_k)$ becomes unfeasible. Thus, a MCMC method is required in order to generate observations from the joint posterior distribution $f(M_k, \delta_t|y_t)$ of (M_k, δ_t) for estimating $f(M_k|y_t)$ and $f(\delta_t|M_k, y_t)$.

The main thrust of the HBFC procedure is to find the set of clusters that gives the best fuzzy partition and then assign each linear time-series to one or more homogeneous clusters dealing with endogeneity issues and model misspecification problems. Here, a fuzzy partition denotes an assignment of Markov Chains (MCs) to cluster such that each time-series is grouped on the basis of their components and unmodelled dynamics. In addition, the task of clustering MCs are treated 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 (4) 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 (3), the full model class set is:

$$\mathcal{F} = \left\{ M_k : M_k \subset \mathcal{F}, M_k \in \mathcal{M}, k \in \mathcal{K},\ \alpha + \sum_{i=1}^{p} \phi_i y_{k,t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{k,t-j} + \epsilon_{k,t} \right\} \tag{9}$$

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

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

$$\pi(M_\tau|\mathcal{F}) = \frac{\pi(M_\tau) \cdot \pi(\mathcal{F}|M_\tau)}{\pi(\mathcal{F})} \tag{10}$$

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

The quantity $\pi(\mathcal{F})$ 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 $\pi(\mathcal{F}|M_\tau)$, 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 exact and final solution will correspond to one of the submodels M_τ with higher natural log Bayes Factor (lBF):

$$lBF_{\tau,k} = log\left\{\frac{\pi(M_\tau|Y_t = y_t)}{\pi(M_k|Y_t = y_t)}\right\} \tag{11}$$

where $\tau \leq k$. In this procedure, the lBF would also be called the log weighted likelihood ratio factor of M_τ to M_k with the priors being the weighting functions. The corresponding scale of evidence[1] is:

$$\begin{cases} 0.00 < lBF_{\tau,k} < 4.99 & \text{no evidence for submodel } M_\tau \\ 5.00 < lBF_{\tau,k} < 9.99 & \text{moderate evidence for submodel } M_\tau \\ 10.00 < lBF_{\tau,k} < 14.99 & \text{strong evidence for submodel } M_\tau \\ lBF_{\tau,k} \geq 15.00 & \text{very strong evidence for submodel } M_\tau \end{cases} \tag{12}$$

Finally, to complete the HBFC 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

[1] It is a generalization of Kass and Raftery (1995).

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 $(\hat{y}_t|\hat{y}_{t-l})$ to guide the search process. The resulting algorithm is called robust weighted distance (RWD) 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_{k_1} and Q_{k_2} be matrices of transition probabilities between distinct MCs among time-series, and $q_{k_1,ls}$ and $q_{k_2,ls}$ be the probabilities of the transition $l \to s$ in Q_{k_1} and Q_{k_2}, the RWD from Q_{k_1} to Q_{k_2} is:

$$D_{rwd}\left(Q^c_{k_1}||Q^c_{k_2}\right) = \sum_{s=1}^{J} \bar{\omega}^c_s \frac{D(q^c_{k_1,l}, q^c_{k_2,l})}{J} \tag{13}$$

where $\bar{\omega}^c$ is the PMS distribution, on average, between the probabilities $q^c_{k_1}$ and $q^c_{k_2}$ designed in the clustering procedure and obtained by the estimated transition probability matrices, with c denoting the optimal number of clusters according to BMS procedure.

The distance in equation (13) is an implemented version of the symmetric Kullback-Leibler distance[2]. More precisely, since each of the two matrices ($Q^c_{k_1}$ and $Q^c_{k_2}$) 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 (13) is zero when $Q^c_{k_1} = Q^c_{k_2}$, 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 (5) used as a scoring metric by the algorithm should increase.

3.2 Bayesian Inference and Variable Selection Problem

The BMA procedure entails estimating the time-varying parameters δ_t to find the `best` subset that contains the fuzzy partitions of dynamic data objects. However, the variable selection problem arises because, in equation (3), dynamic feedback and interactions among time-series are possible. Thus, even if these features would ensure better and more accurate forecasts, it is very costly making the number of coefficients of (3) very large. Indeed, they are increased by $[m\nu]$ factors. Moreover, because the coefficient vectors in δ_t vary in different time periods for each time-series, in high dimensional case (more coefficients than data), it is impossible to eliminate δ_t. To avoid the curse of dimensionality, I implement the framework in Pacifico (2020b) by adapting it in a time-varying context. More precisely, the variable selection problem is addressed by assuming δ_t to have the following factor structure:

$$\delta_t = \sum_{c=1}^{\bar{c}} D_c \cdot \beta_{ct,k} + u_t \qquad \text{with} \qquad u_t \sim N(0, \Sigma_u) \tag{14}$$

[2] It is a well-known statistical indicator useful in evaluating the similarity of time-series represented by their Markov chains. See, for instance, Do and Vetterli (2002).

where $\bar{c}$ denotes the maximum number of clusters, with $\bar{c} \ll m$ and $dim(\beta_{ct,k}) \ll dim(\delta_t)$ by construction, $D_c = [d_1, d_2, \ldots, d_{\bar{c}}]$ is an $m\nu \cdot c$ conformable matrix with elements equal to zero (absence of $k-$th time-series in the $c-$th cluster) and one (presence of $k-$th time-series in the $c-$th cluster), u_t is a $m\nu \cdot 1$ vector of unmodelled variations among time-series present in δ_t, and $\Sigma_u = \sigma^2_\epsilon \otimes V$, with $\sigma^2_\epsilon = diag(\epsilon^{2'}_{1t}, \epsilon^{2'}_{2t}, \ldots, \epsilon^{2'}_{mt})$ denoting the variance of the vector $\tilde{\epsilon}$ that includes stochastic volatility terms and $V = (\sigma^2 \cdot I_{m\nu})$.

In this framework, endogeneity issues and functional forms of misspecification are absorbed in the $(mc \cdot 1)$ time-varying coefficient vectors β_{ct} (stacked for k). They are observable smooth linear functions of the lagged variables and then easily estimable without loss of efficiency and accuracy. The correct choice of clusters across and within series is obtained through the RWD measure by assigning a membership value in the range $[0, 1]$ to all clusters. The generalization from the set $\{0, 1\}$ to the interval $[0, 1]$ is called `fuzzification`, where the $c = m$ case would mean that each time-series has been assigned to its own cluster (no homogeneity among time-series). Let the framework be dynamic, either the clustering membership or MARMA components change dynamically over time and are modelled via Bayesian inference.

The idea is to shrink δ_t to a much smaller dimensional vector β_{ct}, with $\beta_{ct} = (\beta'_{1t,k}, \beta'_{2t,k}, \ldots, \beta'_{\bar{c}t,k})'$, containing all the univariate linear regression coefficients stacked into a vector. In this way, further investigations can be performed (e.g., linear dependencies, dynamic feedback effects, business cycles, interactions). Finally, the `fuzzification` and factorization of δ_t become exact as long as σ^2 converges to zero. In equation (14), all factors are permitted to be time-varying, and the computational costs involved in using that specification are moderate since the high dimensionality is avoided via Bayesian inference and MCMC implementations. For instance, Kalman-Filter technique is used to get appropriate posterior distributions for time-varying coefficients β_{ct}. According to (14), the equation (3) can be written as:

$$y_t = \dot{x}_t \left(\sum_{c=1}^{\bar{c}} D_c \beta_{ct} + u_t \right) + \tilde{\epsilon}_t \equiv \chi_{ct} \beta_{ct} + \eta_t \tag{15}$$

where $\chi_{ct} \equiv \dot{x}_t D_c$ is an $m \cdot c$ matrix that stacks all time-varying coefficients (β_{ct}) and their possible interactions among MARIMA processes, with $\chi_t = diag(\chi'_{1t,k}, \chi'_{2t,k}, \ldots, \chi'_{\bar{c}t,k})$, and $\eta_t \equiv \dot{x}_t u_t + \tilde{\epsilon}_t \sim N(0, \sigma^2_\epsilon \otimes \Sigma_u)$. If the `fuzzification` and factorization are exact, the covariance matrix of η_t is homoskedastic, the variance in error terms is allowed to be time-variant, and volatility changes are replaced by coefficient changes. The equation (15) takes the name of Seemingly Unrelated Regression (SUR) model.

To complete the specification, I suppose the following state-space structure for the time-varying regression coefficients:

$$\beta_{ct} = \beta_{ct-1} + \upsilon \qquad \text{with} \qquad \upsilon \sim N(0, \Sigma_\upsilon) \tag{16}$$

where $\beta_{ct} = (\beta_{1t,k}, \beta_{2t,k}, \ldots)'$, $\Sigma_\upsilon = diag(\tilde{\Sigma}_{1t}, \tilde{\Sigma}_{2t}, \ldots, \tilde{\Sigma}_{\bar{c}t})$ is a block diagonal matrix, and $\tilde{\Sigma}_{ct} = (\tilde{s}_{ct} \cdot I_{m\nu})$, where $\tilde{s}_{ct}$ is an unknown indicator managing the stringent conditions of the time-varying parameters under the factorization (14) in order to make this latter estimable. The random-walk assumption in (16) is very common in the time-varying literature and has the advantage of focusing on permanent shifts (e.g., unmodelled dynamics in a given time period) and reducing the number of parameters in the estimation procedure. The errors $\tilde{\epsilon}_t$ and u_t are mutually independent.

3.3 Model Features

To illustrate the conformation of the time-varying vectors and conformable matrices in (15), I suppose there are $m = 3$ time-series following ARIMA$(2, 1, 1)$ processes, and the optimal number of clusters is $\bar{c} = 2$, with the first two series belonging to cluster 1 and the third one to cluster 2. For convenience, I suppose no intercept. Thus, the MARIMA models grouped in (3) assume the form:

$$\begin{aligned} y_{1t} &= \phi_{1,1} y_{1,t-1} + \phi_{2,1} y_{1,t-2} + \theta_{1,1} \epsilon_{1,t-1} + \epsilon_{1,t} \\ y_{2t} &= \phi_{1,2} y_{2,t-1} + \phi_{2,2} y_{2,t-2} + \theta_{1,2} \epsilon_{2,t-1} + \epsilon_{2,t} \\ y_{3t} &= \phi_{1,3} y_{3,t-1} + \phi_{2,3} y_{3,t-2} + \theta_{1,3} \epsilon_{3,t-1} + \epsilon_{3,t} \end{aligned} \tag{17}$$

Let $\delta_t = \Big(vec(\Phi'(B)), vec(\Theta'(B)) \Big)$ be the $9 \cdot 1$ vector containing – stacked into columns – the AR and MA components for each time-series, and let $\tilde{x}_t = (y'_{1,t-1}, y'_{2,t-1}, y'_{3,t-1}, y'_{1,t-2}, y'_{2,t-2}, y'_{3,t-2}, \epsilon'_{1,t-1}, \epsilon'_{2,t-1}, \epsilon'_{3,t-1})'$ be the $1 \cdot 9$ vector containing all lagged y_t's and white noises within the MARIMA models in (17), the factorization is:

$$\delta_t = \sum_{c=1}^{2} D_c \beta_{ct,k} + u_t \tag{18}$$

where u_t is a 9×1 vector capturing unaccounted features, $\gamma_{k,t}$ is a 9×1 vector containing (stacked) all AR and MA components for each time-series given ν, and, stacking for k, $\beta_{ct} = (\beta'_{1t}, \beta'_{2t})$ is a $6 \cdot 1$ vector containing all time-varying coefficient vectors (ϕ_i, θ_j) to be estimated. More precisely, the factors $\beta_{1t} = (\beta_{1t,1}, \beta_{1t,2}, \beta_{1t,3})$ and $\beta_{2t} = (\beta_{2t,1}, \beta_{2t,2}, \beta_{2t,3})$ are $m \cdot M_c$ mutually orthogonal vectors capturing unmodelled dynamics in δ_t among time-series belonging to the first and the secon cluster, respectively, where $M_c = (M_{c1}, M_{c2})$ denotes the `best` set of assigned clusters.

Letting $i_1 = (1, 1, 1)'$ and $i_2 = (0, 0, 0)'$, the factorization in (18) can be rewritten as:

$$\underset{(9\cdot1)}{\delta_t} = \underset{(9\cdot2)}{\begin{pmatrix} i_1 & i_2 \\ i_1 & i_2 \\ i_2 & i_1 \end{pmatrix}} \cdot \underset{(6\cdot1)}{\begin{pmatrix} \beta_{1t} \\ \beta_{2t} \end{pmatrix}} + \underset{(9\cdot1)}{u_t} \tag{19}$$

Thus, following some arrangements, the SUR model is:

$$\underset{(3\cdot1)}{\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{pmatrix}} = \underset{(3\cdot2)}{\begin{pmatrix} \chi_{1t,1} & 0 \\ \chi_{1t,2} & 0 \\ 0 & \chi_{2t,3} \end{pmatrix}} \cdot \underset{(6\cdot1)}{\begin{pmatrix} \beta_{1t,1} \\ \beta_{1t,2} \\ \beta_{1t,3} \\ \beta_{2t,1} \\ \beta_{2t,2} \\ \beta_{2t,3} \end{pmatrix}} + \underset{(3\cdot1)}{\eta_t} \tag{20}$$

where $\chi_{1t} = (\chi_{1t,1}, \chi_{1t,2}, 0)$ and $\chi_{2t} = (0, 0, \chi_{2t,3})$ are observable specific indicators for y_t capturing dynamics and potential interactions among ARIMA processes assigned to clusters 1 and 2, with $\chi_{1t,1} = \sum_k y_{1kt-i}$ and $\chi_{1t,2} = \sum_k y_{2kt-i}$.

4 Dynamic Analysis

4.1 Prior Specification Strategy

Before specifying prior setups and assumptions, I recall the state-space structure in which the Kalman Filter is employed:

$$y_t = \Big(\dot{x}_t \cdot G_c\Big)\beta_{ct} + \eta_t \qquad \Big(\text{'Measurement} \quad \text{Equation'}\Big) \tag{21}$$

$$\beta_{ct} = \beta_{ct-1} + \upsilon \qquad \Big(\text{'State-Transition} \quad \text{Equation'}\Big) \tag{22}$$

Supposing exact `fuzzification` and factorization in (14), I need to define prior moments on $\psi_0 = (\sigma^2_\epsilon, \tilde{s}_{c0}, \beta_{ct})$, where ψ_0 is a vector collecting the prior densities. Since their appropriate values are unknown, one can hierarchically model the uncertainty underlying variable selection through CIPM priors:

$$\pi(\sigma^2_\epsilon, \tilde{s}_{c0}, \beta_{ct}|y_t) = \pi(\sigma^2_\epsilon|y_t) \cdot \prod_c \pi(\tilde{s}_{c0}|y_t) \cdot \pi(\beta_{ct}|\sigma^2_\epsilon, y_t) \tag{23}$$

Nevertheless, β_{ct} and σ^2_ϵ are not independent of one another. General priors that do not involve the restrictions inherent in (23) are the independent normal-Wishart and the independent inverted Gamma distributions (see, for instance, Chib and Greenberg (1995, 1996)). They are common choices to estimate parametric SUR models – as with (15) – by allowing cross-equation independence of the coefficient distributions and remove the dependence of β_{ct} and σ^2_ϵ. Thus, the CIPM priors involved in (23) can be written as:

$$\pi(\beta_{ct}|y_t) = N\Big(\bar{\beta}_{t|t}, R_{t|t}\Big) \tag{24}$$

$$\pi(\sigma^2_\epsilon|y_t) = iW\Big(\varpi, \zeta\Big) \tag{25}$$

$$\pi(\tilde{s}_{c0}|y_t) = IG\Big\{\frac{\varphi_0}{2}, \frac{\omega_0}{2}\Big\} \tag{26}$$

All hyperparameters are known. More precisely, they are treated as fixed and are either obtained from the data to tune the prior to the specific applications (such as ϖ and φ_0) or selected `a priori` to produce relatively loose priors (such as ζ and ω_0). The hyperparameters $\bar{\beta}_{t|t}$ and $R_{t|t}$ are obtained by running the conditional posterior distribution of $(\beta_{ct}|y_t)$ given the state-space structure in equations (21) and (22).

4.2 Posterior Distributions and MCMC Algorithms

The posterior distributions for $\psi = (\sigma_\epsilon^2, \tilde{s}_{ct}, \{\beta_{ct}\}_{t=1}^T)$ are obtained by combining the prior with the conditional likelihood in (5). The resulting function is then proportional to:

$$L\Big(y^T|\psi\Big) \propto \Big(\sigma_\epsilon^2\Big)^{\frac{n}{2}} \cdot exp\Bigg\{-\frac{1}{2\sigma_\epsilon^2}\Big[\Sigma_t\Big(y_t - (\dot{x}_t G)\beta_{ct}\Big)^{'}\Big] \cdot \Big[\Sigma_t\Big(y_t - (\dot{x}_t G)\beta_{ct}\Big)\Big]\Bigg\} \tag{27}$$

where $y^T = (y_1, \ldots, y_T)$ denotes the data and ψ refers to the unknowns whose joint distribution needs to be found. However, the analytical computation of posterior distributions $(\psi|y^T)$ is unfeasible. In this study, the Kalman-Filter technique is used through MCMC algorithms to generate a random trajectory for $\{\beta_{ct}\}$. More precisely, for the conditional posterior distribution of $(\beta_{11}, \ldots, \beta_{\bar{c}T}|y^T)$, the forward recursions for posterior means and the covariance matrix are, respectively:

$$\bar{\beta}_{t|t} = \bar{\beta}_{t-1|t-1} + \Big[\Big(y_t - (\dot{x}_t D)^{'}\bar{\beta}_{t-1|t-1}\Big) \cdot P_{t|t-1}^{-1}\Big] \tag{28}$$

$$R_{t|t} = \Big[I_{m\nu} - \Big(P_{t|t-1}^{-1} \cdot (\dot{x}_t D)^{'}\Big)\Big] \cdot (R_{t-1|t-1}) \tag{29}$$

with

$$P_{t|t-1} = \Big[(\dot{x}_t D)^{'} \cdot R_{t-1|t-1} \cdot (\dot{x}_t D)\Big] + \sigma_\epsilon^2 \tag{30}$$

Thus, the marginal distributions of β_{ct} can be computed by averaging over draws in the nuisance dimensions, and the Kalman filter backward can be run to characterise posterior distributions for ψ:

$$\pi(\beta_{ct}|\beta_{ct-1}, y^T, \mathfrak{F}_{-p-d}) = N(\bar{\beta}_{t|t+h}, R_{t|t+h}) \tag{31}$$

with

$$\bar{\beta}_{t|t+h} = \Big(R_{t|t+h}^{-1} \cdot \bar{\beta}_{t|t}\Big) + \Big[\sum_{t=1}^{T}(\dot{x}_t D)^{'} \cdot (\sigma_\epsilon^2)^{-1} \cdot (\dot{x}_t D)\Big] \tag{32}$$

$$R_{t|t+h} = \Big[I_{m\nu} - \Big(R_{t|t} \cdot R_{t+h|t}^{-1}\Big)\Big] \cdot (R_{t|t}) \tag{33}$$

where $\bar{\beta}_{t|t+h}$ and $R_{t|t+h}$ are smoothed h-period-ahead forecasts of β_{ct} and variance-covariance matrix of the forecast error, respectively, and $\mathfrak{F}_{-p-d}$ refers to the information on the parameters and observables at time $(-p-d)$.

Finally, the other posterior distributions can be defined as:

$$\pi(\sigma_\epsilon^2|y_t) = iW\Big(\hat{\varpi}, \hat{\zeta}\Big) \tag{34}$$

$$\pi(\tilde{s}_{ct}|y_t) = IG\left\{\frac{\varphi_{\tilde{s}}}{2}, \frac{\omega_{\tilde{s}}}{2}\right\} \quad (35)$$

where $\hat{\varpi} = \varpi + n$ and $\varphi_{\tilde{s}} = \varphi_0 + m\nu$ are arbitrary degrees of freedom, and $\hat{\zeta} = \zeta + \sum_t u_t' u_t$ and $\omega_{\tilde{s}} = \omega_0 + \sum_t (\beta_{ct} - \beta_{ct-1})^{-1} \cdot (\beta_{ct} - \beta_{ct-1})$ are arbitrary scale parameters. In this study, $\varpi \cong n \cdot M_c$, $\varphi_0 \cong 0.1 \cdot exp(M_c)$, $\zeta \cong 1.0$, and $\omega_0 \cong 0.01$.

5 Empirical Example

The empirical application consists of measuring the distance between MARIMA time-series in a high dimensional context ($k > 15$). Its main thrust is to highlight performance and limitations of the proposed method. More precisely, I build on Pacifico (2020a) by clustering the productivity – in terms of real GDP per capita in logarithmic form – for 20 country-specific models, including the United States. They are so split: 12 European advanced economies such as Austria (*AT*), Belgium (*BE*), Finland (*FI*), France (*FR*), Germany (*DE*), Greece (*GR*), Ireland (*IE*), Italy (*IT*), Netherlands (*NL*), Portugal (*PT*), Slovenia (*SL*), and Spain (*ES*); and 7 European emerging economies such as Czech Republic (*CZ*), Hungary (*HU*), Poland (*PO*), Slovak Republic (*SK*), Estonia (*EE*), Latvia (*LV*), and Lithuania (*LT*). The estimation sample covers the period from March, 1995 to December, 2018 ($T = 96$). All the k series are expressed in quarters[3] and seasonally adjusted. All data points are obtained from the Eurostat and OECD database. The methodology supports FCM clustering algorithm using weighted averages of the data, but close results are obtained accounting for FCMdd clustering as well. The only difference resides in higher membership values assigned to every clusters because of selecting their centers by some of the existing data points (medoids). Thus, it implies that the PMS distribution – corresponding to the parameter $\bar{\omega}^c$ in (13) – would be stretched across more expansive and complex neighborhoods of the model configuration space.

Let a process $y = \{y_{it} \; ; \; i \in \mathbb{N} \; , \; t \in \mathbb{T} \; \}$ admit the following time-series representation:

$$x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} \quad with \quad \epsilon_t \sim WN(0, \sigma^2) \quad (36)$$

Stated the series to be not stationary over time, 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. In this context, I estimate 20 differenced time-series with appropriate lag lengths in AR and then MA components according to the Bayesian inference. More precisely, l run multiple random trajectories for any time-varying coefficient vector (β_{ct}) up to find the exact `fuzzification` and factorization in (14). The RWD measure is computed by assigning a membership value in the range $[0, 1]$ in order to obtain the optimal number of clusters grouping the MARIMA time-series.

The maximum differencing order to test stationarity sets 3, and higher PMS (2.99) and lBF (16.21) among series are obtained by performing a HBFC procedure with a maximum of three clusters ($c = 3$) and $100,000$ iterations per each random start[4]. The associated computational costs are minimized, ensuring consistent posterior estimates and dimension reduction. The clusters are so split: (i) CZ,

[3] The proposed methodology would also perform well using different time periods (e.g., dailies and monthlies data).
[4] According to 100 and $1,000$ iterations, the corresponding PMS/lBF are 1.05/5.57 and 2.50/6.71, respectively.

FR, DE, HU, IT, PO, and US; (*ii*) EE, LV, LT, SK, and SL; and (*iii*) AT, BE, FI, GR, IE, NL, PT, and ES. Table 1 shows membership values and clusters for each time-series.

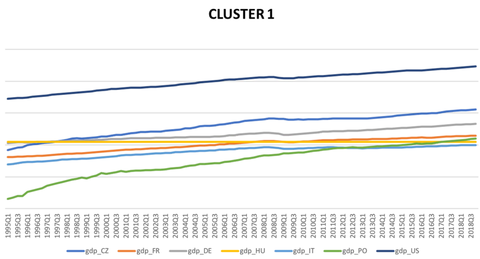
These findings address three important issues. First, when accounting for dynamics of the economy, an accurate BMA strategy is required in order to group time-varying data objects into more probable homogenous clusters (M_τ). Second, the use of conjugate hierarchical informative priors in fuzzy clustering algorithms is able to highlight similarity/dissimilarity among series dealing with unobserved cross-country homogeneity/heterogeneity, and thus group them in 'ad hoc' clusters. Third, given the BMA strategy implied in the HBFC procedure, the Conditional Posterior Sign (CPS) can be computed in order to observe how the GDP time-series for each country evolve over time. It refers to the posterior probability of a coefficient expected value conditional on inclusion. Positive and negative effects on GDP matter whether it is close to 1 and 0, respectively. The CPS is obtained by the Posterior Inclusion Probabilities, corresponding to the sum of PMPs between all series according to their membership values. In a context of economic dynamic interactions, let the CPSs be exactly equal to 1 or 0, the performance of the estimating procedure is highlighted by carefully addressing endogeneity issues and misspecified dynamics. Thus, it would be interesting to extend the fuzzy clustering analysis in a multivariate context (Section 6).

Table 1: Membership Values and Clusters

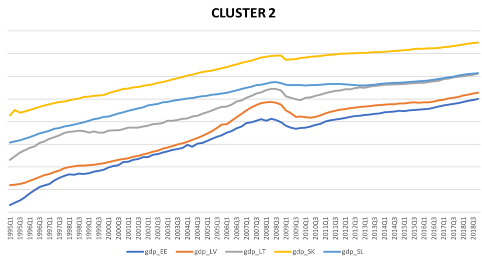
Country	MEMB1	MEMB2	MEMB3	Cluster	CPS	MEMB-FCMdd
AT	0.02	0.00	0.98	3	1.00	1.00
BE	0.01	0.01	0.98	3	1.00	0.99
CZ	1.00	0.00	0.00	1	0.00	1.00
EE	0.01	0.95	0.04	2	0.00	0.96
FI	0.02	0.04	0.94	3	0.00	0.95
FR	0.87	0.02	0.11	1	1.00	0.89
DE	0.96	0.01	0.03	1	1.00	0.97
GR	0.02	0.03	0.95	3	0.00	0.96
HU	0.82	0.05	0.13	1	0.00	0.84
IE	0.03	0.06	0.91	3	1.00	0.93
IT	0.76	0.03	0.21	1	0.01	0.78
LV	0.01	0.97	0.02	2	1.00	0.97
LT	0.00	0.99	0.01	2	0.00	1.00
NL	0.15	0.04	0.81	3	0.99	0.82
PO	0.57	0.05	0.38	1	0.93	0.58
PT	0.02	0.06	0.92	3	1.01	0.94
SK	0.03	0.74	0.23	2	1.00	0.79
SL	0.00	0.98	0.02	2	0.00	0.98
ES	0.41	0.06	0.53	3	0.99	0.55
US	0.87	0.04	0.09	1	1.00	0.87

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; the sixth column shows the CPSs; and the seventh column displays the membership value according to the FCMdd.

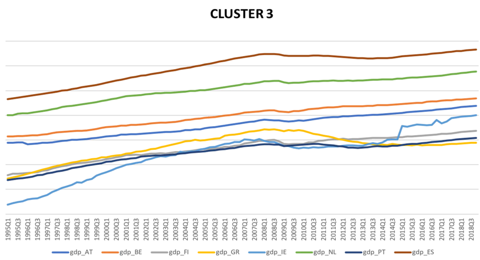
The previous results are better highlighted graphically (Figure 1). Indeed, the first cluster shows consistent similarity over time among US and some European advanced (FR, DE, and IT) and emerging (CZ, HU, and PO) economies (plot 1a), although with different levels of productivity (omitted factors). The second cluster corresponds to Baltic and other two Central-Eastern European countries (plot 1b), displaying more persistent divergence with some common behaviors (unobserved heterogeneity). The last cluster accounts for West European countries (plot 1c), showing stringent interdependency with higher divergence in terms of economic productivity and structure (misspecified dynamics). These findings confirm the efficacy of the HBFC procedure when grouping multiple dynamic processes, but limited to linear univariate settings.



(a) Cluster 1



(b) Cluster 2



(c) Cluster 3

Figure 1: The real GDP per capita series are drawn and grouped among 20 country-specific models, spanning the period 1995q1 to 2018q4. The Y and X axis represent the series and sampling distribution in quarters, respectively.

In Table 2, I compare the stimating performance of the HBFC procedure with three related distance measures for effective clustering of MARIMA time-series: (i) the most popular multidimensional scaling method such as Weighted Euclidean Distance (WED)[5]; (ii) Discrete Wavelet Transform (DWT)[6]; and (iii) Discrete Fourier Transform (DFT)[7]. Here, some considerations are in order. These clustering approaches would be classified as 'structural level' 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. The DWT technique has the advantage of representing time-series as multi-resolution. Additionally, the location of time and frequency can be gained by means of the time-frequency localization[8]. 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 problems – are discarded.

The main thrust of this example is to highlight that RWD measure gets the highest 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

[5]See, e.g., Horan (1969) and Carroll and Chang (1970).
[6]See, e.g., Struzik and Sibes (1999).
[7]See, e.g., Agrawal et al. (1993).
[8]It refers to the property of a function which minimizes the spreads or the variance in time and frequency domains.

BMA strategy) when clustering linear dynamic data. Every related approaches would perform well by choosing two clusters: emerging economies and advanced countries including US. By running the lBF in (11) between the submodels M_τ and the ones related to the alternative approaches (M_*)[9], I find moderate support with DWT and DFT measures, and strong evidence with WED measure by supporting unequal size time-series in fuzzy clustering.

Table 2: Performance Comparison

Distance Measure	Cluster	CSM	lBF	Evidence
RWD	3	9.61	-	-
WED	2	8.45	11.07	strong
DWT	2	6.71	7.47	moderate
DFT	2	6.01	8.60	moderate

The Table is so split: the first column refers to the distance measures; 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, respectively.

Finally, in Figure 2, I draw density forecast combinations between MARIMA time-series, spanning the period 1995q1 to 2021q1. They correspond to the projections of every subsequences drawn in the sample, according to the three clusters identified in Figure 1. The yellow and red curves denote the 95% confidence bands, and the blue and purple curves denote the conditional[10] and unconditional[11] projection of outcomes y_t for each time period t, respectively. Here, the outcomes absorb the conditional forecasts computed for a time frame of 9 quarters (2 years and a quarter) in order to also address potential findings concerning the impact of triggering events (e.g., the Great Recession and the ongoing pandemic crisis on the global economy). The natural conjugate prior refers to two subsamples: 2006q1–2009q1 and 2009q2–2018q4 in order to investigate (potential) macroeconomic–financial linkages.

The results emphasize the findings obtained in Table 1 and Figure 1. (*i*) Conditional projections lie in the confidence interval; conversely, unconditional projections tend to diverge over time. Thus, when studying time-varying (economic) factors, structural time variations and dynamic feedback have to be accounted for. (*ii*) Density forecasts in cluster 1 (top-plot) tend to show a similar evolution to the ones according to cluster 2 (middle-plot) because of not directly observed cross-country interdependencies (e.g., similar dynamics, co-movements, economic structure). (*iii*) Density forecasts in cluster 3 (bottom-plot) tend to display larger divergences (in terms of projections) because of potential functional forms of misspecification (e.g., economic dynamic interactions among advanced economies in driving the transmission of unexpected shocks on the productivity). (*iv*) Most positive projections in clusters 1 and 2 higlight the accuracy of the methodology by dealing with endogeneity issues. For instance, developed countries tend to be net senders because of consistent interdependencies and – in turn – emerging economies because of stringent economic–institutional

[9]Here, the submodels would correspond to the subsequences – grouped in two clusters – between time-series having maximum similarity with other objects within that group and minimum similarity with objects in other groups.

[10]Generally, the conditional projection in density forecasts is the one that the model would have obtained over the same period conditionally on the actual path of unexpected dynamics for that period.

[11]Generally, the unconditional projection in density forecasts is the one that the model would obtain for output growth for that period only on the basis of historical information, and it is consistent with a model-based forecast path for the other variables.

implications (see, e.g., Pacifico (2020a)). Thus, the need for forecasters and policymakers to extend the methodology in a multivariate context and improve the analysis constructing appropriate multicountry VAR and panel VAR models.

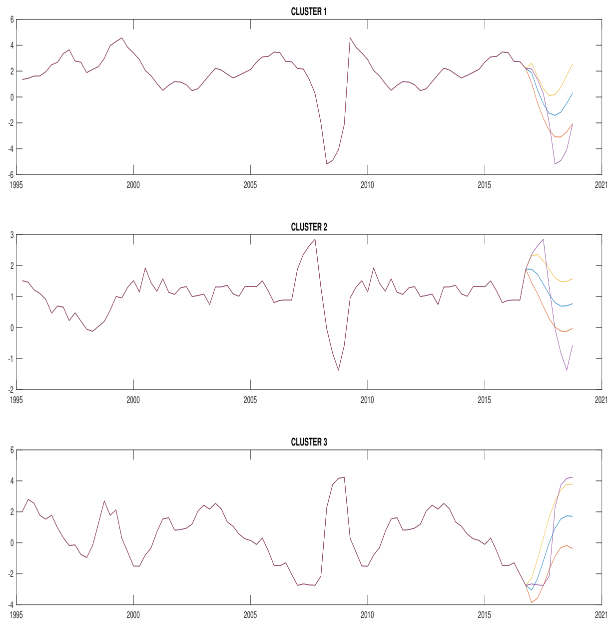


Figure 2: The plot draws density forecast combinations for real GDP per capita series among 20 country-specific models, spanning the period 1995$q1$ to 2021$q1$. It accounts for clustering models M_τ according to the three clusters identified through the HBFC procedure. The Y and X axis represent the projections (blue line) and sampling distribution in quarters (purple line), respectively. They correspond to conditional (blue line) and unconditional (purple line) projections of every subsequences drawn in the sample.

6 HBFC Procedure in a Multivariate Case

In this section, the proposed methodology is extended to panel VAR models in order to test the estimating performance of the HBFC procedure in high dimensional time-varying multicountry settings. The empirical application builds on Pacifico (2020a) and focuses on a simplified version of the SPBVAR accounting for the 20 country-specific models. It involves the productivity ($prod$) – in terms of real GDP per capita in logarithmic form – and four additional endogenous variables: general government spending (gov) and gross fixed capital formation ($gfcf$) describing real economy; and general government debt ($debt$) and current account balance ($curr$) denoting the financial dimension.

The aim of this analysis is to jointly cluster different country- and variable-specific models in order to identify possible homogeneity, commonality, and interdependence among countries and sectors by dealing with endogeneity issues and misspecified dynamics.

The simplified version of the time-varying SPBVAR takes the form:

$$Y_{i,t}^{\ddot{m}} = A_{it,j}^{\ddot{m}}(L)Y_{i,t-1}^{\ddot{m}} + \ddot{\varepsilon}_{it}^{\ddot{m}} \tag{37}$$

where $i, j = 1, 2, \ldots, 20$ are country indices, $t = 1, 2, \ldots, T$ denotes time, $\ddot{m} = 1, \ldots, 6$ denotes the set of endogenous variables, $A_{it,j}$ is a $[(20 \cdot 6) \cdot (20 \cdot 6)]$ matrix of real and financial variables for each pair of countries (i,j) for a given $\ddot{m}$, $Y_{i,t-1}$ is a $[(20 \cdot 6) \cdot 1]$ vector of lagged variables of interest accounting for 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 $[(20 \cdot 6) \cdot 1]$ vector of disturbance terms. For convenience, I suppose one lag and no intercept.

The estimation sample covers the period from March 1995 to December 2019 and amounts – without restrictions – to $26,000$ regression parameters (260 coefficients and 100 equations). According to the BMA strategy implied in the HBFC procedure, there are 2^{120} possible model solutions. Thus, MCMC algorithms need to be improved in order to select $M_{\ddot{\tau}}$ submodels, with $M_{\ddot{\tau}}$ denoting the `best` subset of clusters involved in the multivariate analysis. Indeed, the number of coefficients is increased by $N\ddot{M}$ factors, with $\ddot{M}$ denoting the set of the lagged endogenous variables accounted for. Thus, in order to apply the specifications underlying the HBFC procedure (Section 3), I need to express the time-varying SPBVAR in (37) in terms of a multivariate normal distribution:

$$Y_t = (I_{N\ddot{M}} \otimes \ddot{X}_t)\ddot{\delta}_t + \ddot{E}_t \tag{38}$$

where $Y_t = (Y_{1t}^{\ddot{m}'}, \ldots, Y_{Nt}^{\ddot{m}'})'$ is an $[(20 \cdot 6) \cdot 1]$ vector containing the set of real and financial variables for each i for a given $\ddot{m}$, $\ddot{X}_t = (Y_{i,t-1}^{'\ddot{m}}, Y_{i,t-2}^{'\ddot{m}}, \ldots, Y_{i,t-l}^{'\ddot{m}})'$ 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 (37) for each pair of countries (i,j), $\ddot{\delta}_{it,j}^{\ddot{k}} = vec(\ddot{\gamma}_{it,j}^{\ddot{k}})$ is an $N\ddot{M}\ddot{k} \cdot 1$ vector containing all columns stacked into a vector, with $\ddot{\gamma}_{it,j}^{\ddot{k}} = (A_{it,j}^{1'}, A_{it,j}^{2'}, \ldots, A_{it,j}^{\ddot{M}'})'$ and $\ddot{\delta}_t = (\ddot{\delta}_{1t}', \ddot{\delta}_{2t}', \ldots, \ddot{\delta}_{Nt}')'$ denoting the time-varying coefficient vectors, stacked for i, for each country–variable pair, and $\ddot{E}_t = (\ddot{\varepsilon}_{1t}', \ldots, \ddot{\varepsilon}_{Nt}')'$ is an $[(20 \cdot 6) \cdot 1]$ vector containing the random disturbances of the model. In model (38) there is no subscript i since all lagged variables in the system are stacked in $\ddot{X}_t$.

In this study, the implementation is addressed to the thrust of the HBFC procedure in merging more similar MCs and thus identifying – as fast as possible – more probable homogeneous submodels $M_{\ddot{\tau}}$ among $M_{\ddot{k}}$, with $M_{\ddot{k}}$ denoting all possible linear model solutions involved in the multivariate analysis. More precisely, the coefficient vectors in $\ddot{\delta}_t$ represent all the model solutions counted in the natural model space $\mathcal{M}$ (equation (8)), and each factor would correspond to a clustering model $M_{\ddot{\tau}}$ identifying a distinct cluster of (potential) combination of the series $\ddot{m}$. Thus, I adapt the framework in Pacifico (2020a) and assume $\ddot{\delta}_t$ has the following factor structure:

$$\ddot{\delta}_t = \sum_{\ddot{c}=1}^{\ddot{c}} \ddot{G}_{\ddot{c}} \cdot \beta_{\ddot{c}t} + \ddot{u}_t \quad \text{with} \quad \ddot{u}_t \sim N(0, \Sigma_{\ddot{u}}) \tag{39}$$

where $\ddot{c}$ denotes the maximum number of clusters according to the multivariate analysis, with $\ddot{c} \ll N\ddot{M}\ddot{k}$ and $dim(\beta_{\ddot{c}t}) \ll dim(\ddot{\delta}_t)$ by construction, $\ddot{G}_{\ddot{c}} = [\ddot{G}_1, \ddot{G}_2, \ldots, \ddot{G}_{\ddot{c}}]$ are $N\ddot{M}\ddot{k} \cdot 1$ matrices obtained by multiplying the matrix coefficients $(\ddot{\gamma}_{it,j}^{\ddot{k}})$ stacked in the vector $\ddot{\delta}_t$ by conformable matrices $\ddot{D}_{\ddot{c}}$ with elements equal to zero and one, $\ddot{u}_t$ is an $N\ddot{M}\ddot{k} \cdot 1$ vector of unmodelled variations present in $\ddot{\delta}_t$, 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 vector $\beta_{\ddot{c}t} = (\beta_{1t}', \beta_{2t}', \ldots, \beta_{\ddot{c}t}')'$ denotes the adjusted auxiliary indicator defined in Section 3.2 and contains all time-varying regression coefficients stacked into a vector.

Let the factorization be exact ($\sigma^2 \to 0$), the reduced-form SPBVAR model in equation (38) can be transformed into a Structural Normal Linear Regression (SNLR) model[12] written as

[12]It is similar to the SUR model described in (15).

$$Y_t = \ddot{\ddot{X}}_t\left(\sum_{\ddot{c}=1}^{\ddot{\ddot{c}}} \ddot{G}_{\ddot{c}}\beta_{\ddot{c}t} + \ddot{u}_t\right) + \ddot{E}_t \equiv \ddot{\chi}_{\ddot{c}t}\beta_{\ddot{c}t} + \ddot{\eta}_t \quad \text{with} \quad \ddot{\ddot{X}}_t = \left(I_{N\ddot{M}} \otimes \ddot{X}_t\right) \tag{40}$$

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

Finally, to apply conjugate hierarchical informative priors and MCMC integrations, I adapt the same state-space structure for $\beta_{\ddot{c}t}$ as with (16):

$$\beta_{\ddot{c}t} = \beta_{\ddot{c}t-1} + \ddot{v}_t \quad \text{with} \quad \ddot{v}_t \sim N(0, \Sigma_{\ddot{v}}) \tag{41}$$

where $\beta_{\ddot{c}t} = (\beta_{1t}, \beta_{2t}, \ldots)'$, $\Sigma_{\ddot{v}} = diag(\ddot{\ddot{\Sigma}}_{1t}, \ddot{\ddot{\Sigma}}_{2t}, \ldots, \ddot{\ddot{\Sigma}}_{\ddot{c}t})$ is a block diagonal matrix, and $\ddot{\ddot{\Sigma}}_{\ddot{c}t} = (\ddot{\ddot{s}}_{\ddot{c}t} \cdot I_{\ddot{k}})$, where $\ddot{\ddot{s}}_{\ddot{c}t}$ controls the tightness (stringent conditions) of the factorization of the time-varying parameters ($\beta_{\ddot{c}t}$) in order to make them estimable. The errors $\ddot{E}_t$, $\ddot{u}_t$, and $\ddot{v}_t$ are mutually independent.

Supposing exact factorization, the new prior moments are $\ddot{\psi}_0 = (\Sigma_{\ddot{e}}, \ddot{\ddot{s}}_{\ddot{c}0}, \beta_{\ddot{c}t})$ and the CIPM priors become:

$$\pi(\beta_{\ddot{c}t}|Y^T) = N\left(\ddot{\ddot{\beta}}_{t|t}, \ddot{R}_{t|t}\right) \tag{42}$$

$$\pi(\Sigma_{\ddot{e}}^{-1}|Y^T) = iW\left(z_1, \beta_1\right) \tag{43}$$

$$\pi(\ddot{\ddot{s}}_{\ddot{c}0}|Y^T) = IG\left\{\frac{\ddot{\varphi}_0}{2}, \frac{\ddot{\omega}_0}{2}\right\} \tag{44}$$

The posterior distributions for $\ddot{\psi} = (\Sigma_{\ddot{e}}, \ddot{\ddot{s}}_{\ddot{c}t}, \{\beta_{\ddot{c}t}\}_{t=1}^T)$ are obtained by combining the above adjusted priors and conditional likelihood.

$$L\left(Y^T|\ddot{\psi}\right) \propto \left(\Sigma_{\ddot{e}}^{\frac{T}{2}}\right) \cdot exp\left\{-\frac{1}{2}\left[\sum_{t=1}^{T}\left(Y_t - (\ddot{\ddot{X}}_t\ddot{G})\beta_{\ddot{c}t}\right)'\right] \cdot \Sigma_{\ddot{e}}^{-1} \cdot \left[\sum_{t=1}^{T}\left(Y_t - (\ddot{\ddot{X}}_t\ddot{G})\beta_{\ddot{c}t}\right)\right]\right\} \tag{45}$$

The forward recursions for posterior means and the covariance matrix for $\{\beta_{\ddot{c}t}\}$ are, respectively:

$$\ddot{\ddot{\beta}}_{t|t} = \ddot{\ddot{\beta}}_{t-1|t-1} + \left[\ddot{R}_{t-1|t-1}(\ddot{\ddot{X}}_t\ddot{G})\ddot{P}_{t|t-1}^{-1}\right]\left[Y_t - (\ddot{\ddot{X}}_t\ddot{G})'\ddot{\ddot{\beta}}_{t-1|t-1}\right] \tag{46}$$

$$\ddot{R}_{t|t} = \left[I_{\ddot{k}} - \left(\ddot{R}_{t-1|t-1}(\ddot{\ddot{X}}_t\ddot{G})\ddot{P}_{t|t-1}^{-1}(\ddot{\ddot{X}}_t\ddot{G})'\right)\right] \cdot \ddot{R}_{t-1|t-1} \tag{47}$$

with

$$\ddot{P}_{t|t-1} = \left[(\ddot{\ddot{X}}_t\ddot{G})'\ddot{R}_{t-1|t-1}(\ddot{\ddot{X}}_t\ddot{G})\right] + \Sigma_{\ddot{e}} \tag{48}$$

Thus, the posterior distributions for $\ddot{\psi}$ are so computed:

$$\pi(\beta_{\ddot{c}t}|\beta_{\ddot{c}t-1}, Y^T, \ddot{\psi}_{-\beta_{\ddot{c}t}}) = N(\ddot{\tilde{\beta}}_{t|t+h}, \ddot{R}_{t|t+h}) \tag{49}$$

where

$$\ddot{\tilde{\beta}}_{t|t+h} = \tilde{R}_{t|t+h} \cdot \left[\left(\ddot{R}^{-1}_{t|t+h} \cdot \ddot{\tilde{\beta}}_{t|t} \right) + \left(\sum_{t=1}^{T} (\ddot{\tilde{X}}_t \ddot{G})' \Sigma^{-1}_{\ddot{e}} (\ddot{\tilde{X}}_t \ddot{G}) \right) \right] \tag{50}$$

$$\ddot{R}_{t|t+h} = \left[I_{\ddot{k}} - \left(\ddot{R}_{t|t} \cdot \ddot{R}^{-1}_{t+h|t} \right) \right] \cdot (\ddot{R}_{t|t}) \tag{51}$$

with

$$\tilde{R}_{t|t+h} = \left[\ddot{R}^{-1}_{t|t+h} + \left(\sum_{t=1}^{T} (\ddot{\tilde{X}}_t \ddot{G})' \Sigma^{-1}_{\ddot{e}} (\ddot{\tilde{X}}_t \ddot{G}) \right) \right]^{-1} \tag{52}$$

Finally, the multicountry HBFC procedure is completed by defining the other posterior distributions as:

$$\pi(\Sigma_{\ddot{e}}|Y^T, \ddot{\psi}_{-\Sigma_{\ddot{e}}}) = iW\left(\hat{z}_1, \hat{\beta}_1 \right) \tag{53}$$

$$\pi(\ddot{\tilde{s}}_{\ddot{c}t}|Y^T, \ddot{\psi}_{-\ddot{\tilde{s}}_{\ddot{c}t}}) = IG\left\{ \frac{\ddot{\varphi}_t}{2}, \frac{\ddot{\omega}_t}{2} \right\} \tag{54}$$

where $\hat{z}_1 = z_1 + T$ and $\ddot{\varphi}_t = \ddot{\varphi}_0 + \ddot{k}$ are arbitrary degrees of freedom, and $\hat{\beta}_1 = \beta_1 + \sum_t u_t' u_t$ and $\ddot{\omega}_t = \ddot{\omega}_0 + \sum_t (\beta_t^{\ddot{c}} - \beta_{t-1}^{\ddot{c}})^{-1} \cdot (\beta_t^{\ddot{c}} - \beta_{t-1}^{\ddot{c}})$ are arbitrary scale parameters, with $\beta_t^{\ddot{c}}$ denoting the $\ddot{c}^{th}$ subvector of $\beta_{\ddot{c}t}$. In this study, $z_1 \cong N \cdot (\ddot{M} + \ddot{M}_{\ddot{c}})$, $\ddot{\varphi}_0 \cong 0.1 \cdot exp(\ddot{M} + \ddot{M}_{\ddot{c}})$, $\beta_1 \cong 1.0$, and $\ddot{\omega}_0 \cong 0.1$, where $\ddot{M}_{\ddot{c}}$ denotes the `best` subset of clusters according to the multivariate analysis.

Given the above specifications, I am able to discriminate among all series in $M_{\ddot{k}}$ by directly choosing a pool of *best* submodels ($M_{\ddot{\tau}}$) that contain the only regression parameters with higher posterior means ($\hat{\beta}_{\ddot{c}t}$'s) and different from zero. Thus, I can jointly deal with overestimation of effect sizes (or individual contributions) and model uncertainty (implied in the procedure) without loss of estimator efficiency. The log Bayes Factor in (11) is computed as:

$$lBF_{\ddot{k},\ddot{\tau}} = log\left(\frac{L(Y_T|M_{\ddot{k}})}{L(Y_T|M_{\ddot{\tau}})} \right) \tag{55}$$

where Y_T denotes the data and $L(Y_T|M_{\ddot{k}})$ refers to the (conditional) likelihood function conducted on submodels $M_{\ddot{\tau}}$ by MCs implementations. Support for discovering the most probable set of clusters capturing different dynamics and interconnections among time-varying data is obtained by comparing the marginal likelihoods of the unrestricted models ($M_{\ddot{k}}$) and a vector of the submodels ($M_{\ddot{\tau}}$).

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

$$D_{rwd}\Big(Q^{\ddot{c}}_{i\ddot{m}}||Q^{\ddot{c}}_{j\ddot{m}}\Big) = \sum_{s=1}^{J} \ddot{\omega}^{\ddot{c}}_{s} \frac{D(q^{\ddot{c}}_{i\ddot{m},l}, q^{\ddot{c}}_{j\ddot{m},l})}{J} \tag{56}$$

where $D(q^{\ddot{c}}_{i\ddot{m},l}, q^{\ddot{c}}_{j\ddot{m},l}) = \frac{[d(q^{\ddot{c}}_{i\ddot{m},l}, q^{\ddot{c}}_{j\ddot{m},l}) + d(q^{\ddot{c}}_{j\ddot{m},l}, q^{\ddot{c}}_{i\ddot{m},l})]}{2}$ and $\ddot{\omega}^{\ddot{c}}$ is the PMS distribution, on average, between the probabilities $q^{\ddot{c}}_{i\ddot{m},l}$ and $q^{\ddot{c}}_{j\ddot{m},l}$.

6.1 Empirical Results

The optimal convergence is obtained with a maximum of three clusters ($c = 3$) and $1,000$ iterations per each random start, corresponding to the highest PMS/lBF: 2.89/16.70 about *prod*; 1.89/16.35 about *gov*; 1.82/15.97 about $gfcf$; 1.99/15.95 about *debt*; and 1.98/15.93 about $curr$. Increasing the number of iterations, the procedure would perform better, but with larger computational costs. To avoid that trade-off, the estimating procedure may be improved by supposing – for example – a less stringent state-space structure in order to not only allow for time-varying shifts, but also deal with (potential) nonlinearities (e.g., volatility issues). In this way, alternative MCMC integrations would also matter improving the number of iterations with low computational costs. For instance, Markov-Switching mixed-frequency Bayesian processes are able to model covariance matrices of country-specific Markov chains, mainly when transition probabilities differ among time-varying data (see, e.g., Casarin et al. (2018)).

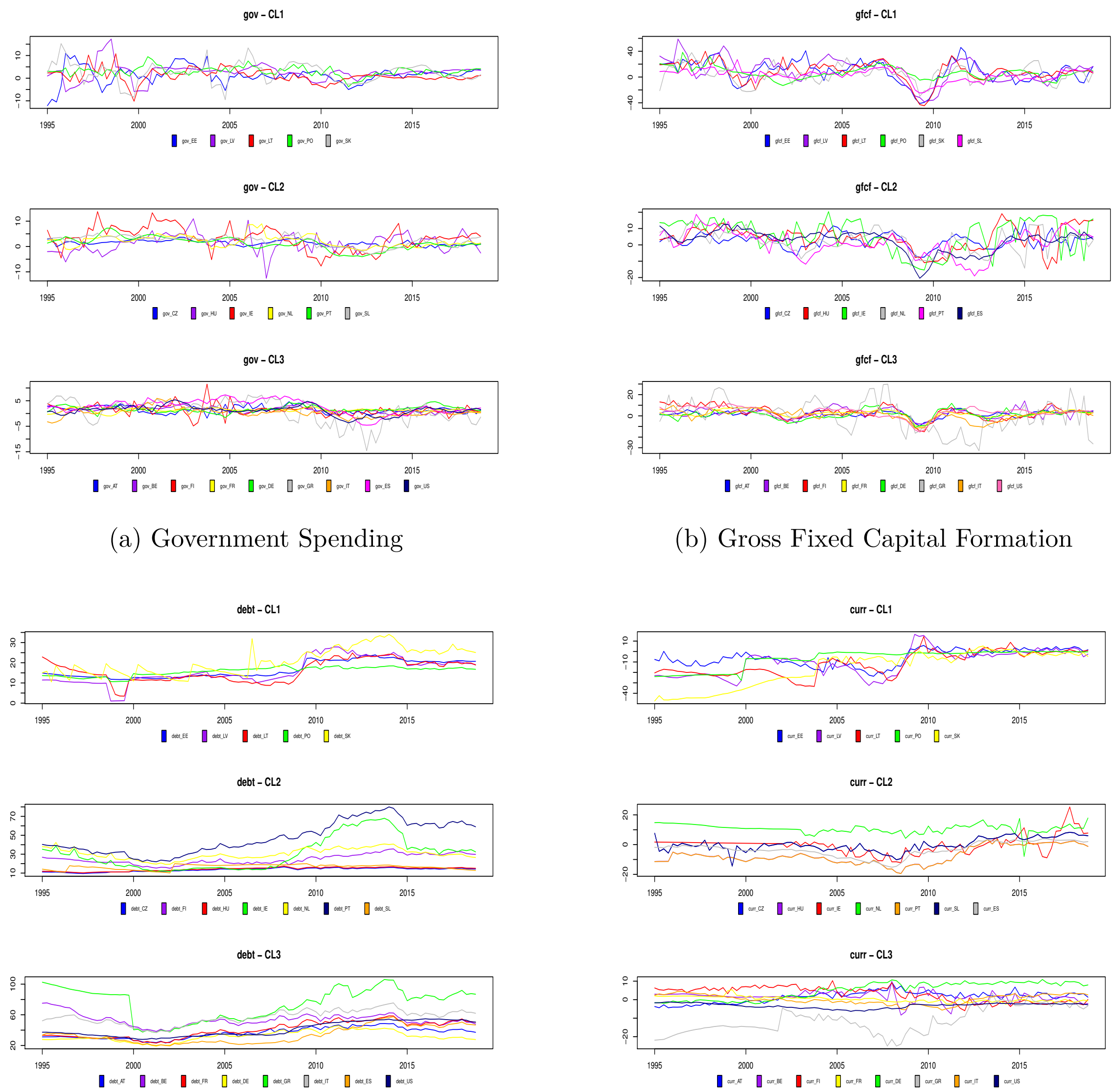
Furthermore, I compare the estimating performance of the HBFC procedure with a data analysis technique widely used to measure time-series similarity in multivariate analysis: the Principal Component Analysis (PCA)[13]. Its particular feature is the main tendency of observing data compactly and thus the 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 do 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. Indeed, it finds two maximum clusters: (*i*) real economy and (*ii*) financial sector.

In Figure 3, the clustering plots are drawn. I dropped the *prod* showing similar results compared to the univariate case-study. Three main considerations are in order. First, a relevant common component matters more in the real economy, but with stronger heterogeneity among series between clusters (Figures 3a and 3b). Second, larger cross-country homogeneity and interdependency matter among financial variables within and between clusters, respectively (Figures 3c and 3d). The former would be larger among Baltic and other emerging countries (clusters 1 and 2); while stringent interdependency – even if with stronger divergences – concerns mainly advanced economies (clusters 2 and 3). Third, the factorization and `fuzzification` implied in the HBFC procedure are exact highlighting three main area-specific common groups: (*i*) Baltic European countries and other emerging economies (such as PO and SK) because of misspecified dynamics (e.g., commonality due to stringent financial exposures with advanced countries); (*ii*) Central-Eastern European countries and some advanced economies (such as IE, NL, and PT) because of unobserved heterogeneity (e.g., divergent responses to a common unexpected shock); and (*iii*) Western European countries and US because of hidden factors (e.g., strong economic–institutional linkages). Concerning ES, it would tend to belong to either the first or third group due to large trade exposures (see, for instance,

[13] See, e.g., Gavrilov et al. (2000).

Pacifico (2020a)).

These findings are robust and consistent with the more recent business cycle literature, which recognizes the importance of dealing with noise and high dimensional problems when quantifying similarities/dissimilarities among dynamic data (see, for instance, Izakian et al. (2013), Ratanamahatana et al. (2005), Ramoni et al. (2002), McAlinn et al. (2020), and Kaufman and Rousseeuw (2009)).



(a) Government Spending

(b) Gross Fixed Capital Formation

(c) Government Debt

(d) Current Account Balance

Figure 3: Clusters for both real (plots a and b) and financial (plots c and d) variables are drawn among 20 country-specific models, spanning the period 1995q1 to 2018q4. The Y and X axis represent the series and sampling distribution in quarters, respectively.

Finally, in Figure 4, I draw density forecast combinations for real GDP per capita series dealing with both real and financial variables for the three country-specific common groups, spanning the period 1995q1 to 2021q1. The findings can be summarized in four main results. (i) Larger heterogeneity matters in the first two groups mainly concerning emerging economies (top- and middle-plot). (ii) There are stronger cross-country interdependencies among advanced economies showing similar dynamics over time (bottom-plot). (iii) Stringent interlinkages between Central-Eastern and Western countries seemingly absorbe an unexpected shock because of net senders in their financial dimension (middle-plot). Then, the relevant common component in real economy would affect economic dynamics through unobserved transmission channels, negatively during triggering events (e.g., the global financial crisis). (*iv*) An increasing degree of divergence matters among countries during the ongoing pandemic crisis, mainly among emerging economies showing slower economic recovery. Thus, the HBFC procedure would be able to extract significant features of multivariate time-series and thus obtain related or homogeneous clusters without avdanced knowledge of the data.

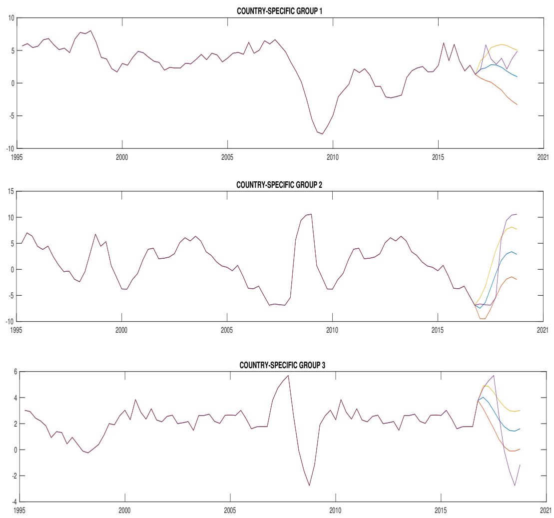
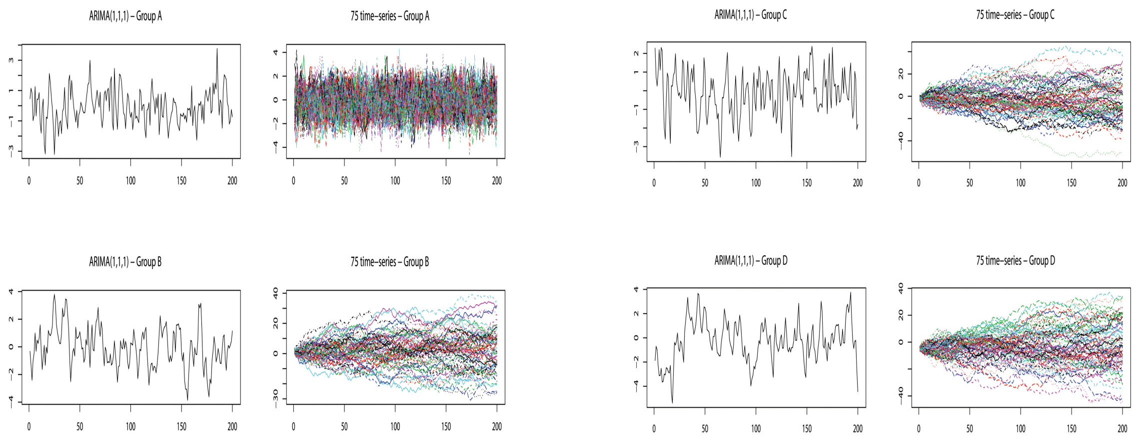


Figure 4: Density forecast combinations for real GDP per capita series are drawn accounting for both real and financial variables. It deals with the three country-specific common groups, spanning the period $1995q1$ to $2021q1$. The Y and X axis represent the projections and sampling distribution in quarters, respectively.

7 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 $[(1.30, 0.30) \pm 0.01]$, $[(1.34, 0.34) \pm 0.01]$, $[(1.60, 0.60) \pm 0.01]$, and $[(1.64, 0.64) \pm 0.01]$, respectively. The white noise ϵ_t used has mean zero and variance 0.01. All 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 every 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 5).



(a) Groups A and B (b) Groups C and D

Figure 5: Simulated MARIMA(1,1,1) models are drawn from the groups A, B, C, and D, with $k = 75$ and $t = 200$. The Y and X axis represent the series and sampling time, respectively.

According to Bayesian inference, higher PMP distribution and log Bayes Factor among MARIMA series are obtained by performing a HBFC procedure with three clusters ($c = 3$) and 100,000 iterations for each random start.

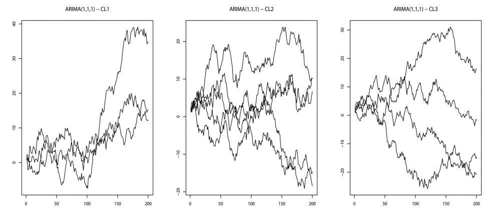
In Table 3, I compare the cluster similarity metric – obtained using RWD measure – with the three related similarity measures: (*i*) Weighted Euclidean Distance (WED); (*ii*) Discrete Wavelet Transform (DWT); and (*iii*) Discrete Fourier Transform (DFT). The highest cluster similarity metric is found for the RWD measure (close to 1), providing an accurate clustering for all of the MARIMA(1,1,1) collections.

Table 3: Cluster Similarity Metric

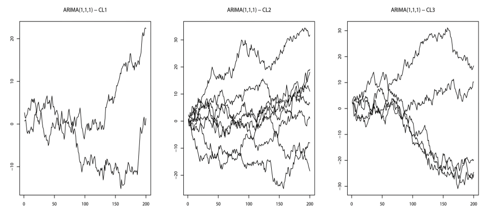
Collection	RWD	WED	DWT	DFT
1	9.82	8.127	4.76	5.20
2	9.78	8.21	4.49	5.63
3	9.81	8.47	5.01	5.83
4	9.54	8.17	4.85	5.51
5	9.65	8.35	4.83	5.48
6	9.56	8.41	4.60	5.38
7	9.83	8.33	4.83	5.91
8	9.72	8.22	5.02	5.83
9	9.75	8.80	5.25	5.45
10	9.93	8.46	5.15	5.74
lBF	-	8.71	5.57	5.83
Evidence	-	strong	moderate	moderate

The Table shows the cluster similarity metric for five different distance measures. The last two columns refer to the log Bayes Factor and the corresponding scale of evidence, respectively.

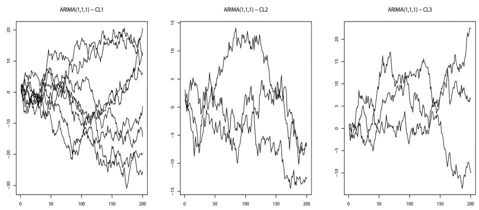
In Figure 6, I draw the clustering plots accounting for all of the 10 collections from MARIMA(1,1,1) models. All the grouped series show lowly average dissimilarity within a cluster and highly strong dissimilarity between clusters. Thus, the three clusters are well defined and separated. In addition, given the BMA implied in the HBFC procedure, dynamics between series are well highlighted and clustered.



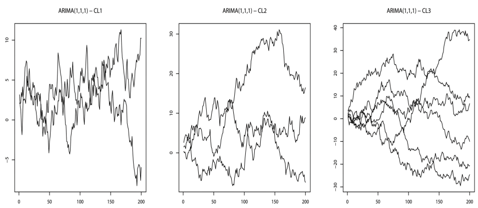
(a) Collection 1



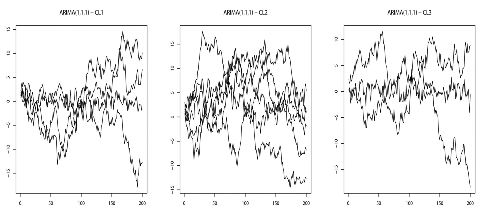
(b) Collection 2



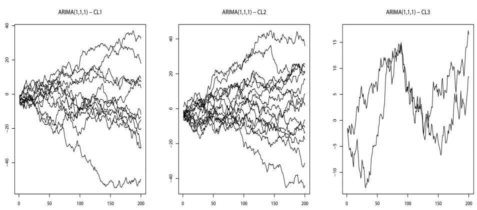
(c) Collection 3



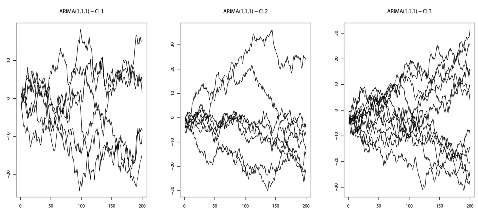
(d) Collection 4



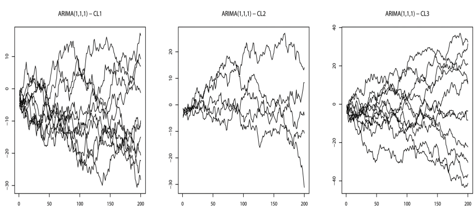
(e) Collection 5



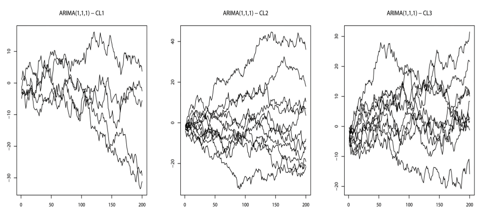
(f) Collection 6



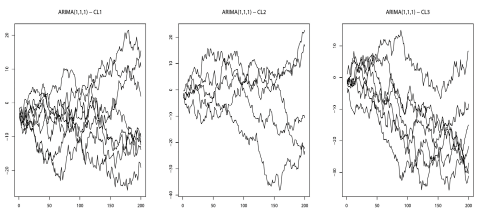
(g) Collection 7



(h) Collection 8



(i) Collection 9



(j) Collection 10

Figure 6: Clustering plots accounting for all of the 10 collections from MARIMA(1,1,1) models are drawn. 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.

8 Concluding Remarks

The paper develops a computational approach to improve hierarchical fuzzy clustering time-series analysis and forecasting performance in high dimensional dynamic data when dealing with endogeneity issues and misspecified dynamics. The main thrust of the proposed procedure is the use of conjugate hierarchical informative proper priors in order to discover the most probable set of clusters capturing different dynamics and interconnections among linear time-varying data. Full posterior distributions for effective clustering univariate and multivariate time-series are obtained by MCMC implementations in order to avoid the problem of increasing the overall probability of errors that plagues classical statistical methods based on significance tests. Bayesian methods are used to reduce the dimensionality of the model acting as a strong variable selection.

In this study, empirical examples describe the estimating procedure and forecasting performance. More precisely, an empirical application for high dimensional time-varying data on a database of multiple ARIMA time-series is accounted for. Then, the methodology is extended in a multivariate context in order to improve the variable selection procedure for clustering time-series to a wide array of candidate models.

Compliance with Ethical Standards

Funding: No fund was used to conduct this study.
Conflict of Interest: The author declares no conflict of interest.
Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

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.

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business and Economic Statistics*, 18(3):338–357.

Albert, J. H. and Chib, S. (1993). Bayesian inference via gibbs sampling of autoregressive time series subject to markov mean and variance shifts. *Journal of Business and Economic Statistics*, 11:1–15.

Cargnoni, C., Müller, P., and West, M. (1996). Bayesian forecasting of multinomial time series through conditionally gaussian dynamic models. *Journal of the American Statistical Association*, 92(438):640–647.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992a). Hierarchical bayesian analysis of changepoint 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.

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

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.

Carter, C. K. and Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81:541–553.

Casarin, R., Foroni, C., Marcellino, M., and Ravazzolo, F. (2018). Uncertainty through the lenses of a mixed-frequency bayesian panel markov-switching model. *The Annals of Applied Statistics*, 12(4):2559–2586.

Chib, S. and Greenberg, E. (1995). Hierarchical analysis of sur models with extensions to correlated serial errors and time-varying parameter models. *Journal of Econometrics*, 68(2):339–360.

Chib, S. and Greenberg, E. (1996). Markov chain monte carlo simulation methods in econometrics. *Econometric Theory*, 12(3):409–431.

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

De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328.

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

Doan, T., Litterman, R. B., and Sims, C. (1992). Forecasting and conditional projections using realistic prior distributions. *Econometric Reviews*, 3:1–100.

Durbin, T. and Koopman, S. (2012). Time series analysis by state space methods, 2nd edn. *Oxford Statistical Science Series, Oxford University Press.*

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

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15(2):183–202.

Gamerman, D. and Migon, H. S. (1993). Dynamic hierarchical models. *Journal of the Royal Statistical Society (Series B)*, 55:629–642.

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.

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *The Review of Economics and Statistics*, 97(2):436–451.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society: Series B*, 38(3):205–228.

Harrison, P. J. and West, M. (1987). Practical bayesian forecasting. *The Statistician*, 36:115–125.

Harrison, P. J. and West, M. (1991). Dynamic linear model diagnostics. *Biometrika*, 78:797–808.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, pages 382–401.

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

Huerta, G. and West, M. (1999). Priors and component structures in autoregressive time series models. *Journal of the Royal Statistical Society: Series B*, 61:881–899.

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 the American Statistical Association*, 90(430):773–795.

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.

Kitegawa, G. (1987). A smoothness priors time varying ar coefficient modeling of nonstationary time series. *Journal of the American Statistical Association*, 82:1032–1063.

Kleibergen, F. R. and Paap, R. (2002). Priors, posteriors and bayes factors for a bayesian analysis of cointegration. *Journal of Econometrics*, 111(2):223–249.

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.

Litterman, R. B. (1986a). Forecasting with bayesian vector autoregressions: Five years of experience. *Journal of Business and Economic Statistics*, 4:25–38.

Litterman, R. B. (1986b). Specifying vector autoregressions for macroeconomic forecasting. *In Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, P.K. Goel and A. Zellner (Eds.). North-Holland, Amsterdam.*

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, pages 41–67.

McAlinn, K., Aastveit, K. A., Nakajima, J., and West, M. (2020). Multivariate bayesian predictive synthesis in macroeconomic forecasting. *Journal of the American Statistical Association*, 115(531):1092–1110.

McAlinn, K. and West, M. (2019). Dynamic bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, 210(1):155–169.

McCulloch, R. E. and Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association*, 88(423):968–978.

McCulloch, R. E. and Tsay, R. S. (1994). Bayesian analysis of autoregressive time series via the gibbs sampler. *Journal of Time Series Analysis*, 15:235–250.

Pacifico, A. (2020a). Fiscal implications, misspecified dynamics, and international spillover effects across europe: A time-varying multicountry analysis. *International Journal of Statistics and Economics*, 21(2):18–40.

Pacifico, A. (2020b). Robust open bayesian analysis: Overfitting, model uncertainty, and endogeneity issues in multiple regression models. *Econometric Reviews*, 40(2):148–176. DOI: `https://doi.org/10.1080/07474938.2020.1770996`.

Pole, A. and West, M. (1990). Efficient bayesian learning in non-linear dynamic models. *Journal of Forecasting*, 9:119–136.

Prado, R. and West, M. (1997). Bayesian analysis and decomposition of multiple non-stationary time series. *In Modelling Longitudinal and Spatially Correlated Data, T. Gregoire (Ed.). Springer-Verlag, New York.*

Prado, R. and West, M. (2010). Time series: Modelling, computation & inference. *Chapman & Hall, CRC Press.*

Quintana, J. M. and West, M. (1987). Multivariate time series analysis: new techniques applied to international exchange rate data. *The Statistician*, 36:275–281.

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.

Singh, H. V. P. and Mahmoud, Q. H. (2019). Evaluation of arima models for human-machine interface state sequence prediction. *Machine Learning and Knowledge Extraction*, 1:287–311.

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.

West, M. (1992a). Approximating posterior distributions with mixtures. *Journal of the Royal Statistical Society (Series B)*, 55:409–422.

West, M. (1992b). Mixture models, monte carlo, bayesian updating and dynamic models. *Computing Science and Statistics*, 24:325–333.

West, M. (2003). Bayesian factor regression models in the large p, small n paradigm. *Bayesian Statistics*, 7:733–742.

West, M. and Harrison, J. (1999). Bayesian forecasting and dynamic models, 2nd edition. *Springer Verlag.*

West, M., Harrison, J., and Migon, H. S. (1985). Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1003.

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.

Zellner, A., Hong, C., and Min, C. (1991). Forecasting turning points in international output growth rates using bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques. *Journal of Econometrics*, 49:275–304.