Autoregressive Moving Average Infinite Hidden Markov-Switching Models

Luc Bauwens*
CORE, Université catholique de Louvain
SKEMA Business School - Université de Lille
and
Jean-François Carpantier
CREA, University of Luxembourg
and
Arnaud Dufays†
Department of Economics, Université Laval
Ecole Nationale de la Statistique et de l'Administration Economique, CREST

October 20, 2015

Abstract

Markov-switching models are usually specified under the assumption that all the parameters change when a regime switch occurs. Relaxing this hypothesis and being able to detect which parameters evolve over time is relevant for interpreting the changes in the dynamics of the series, for specifying models parsimoniously, and may be helpful in forecasting. We propose the class of sticky infinite hidden Markov-switching autoregressive moving average models, in which we disentangle the break dynamics of the mean and the variance parameters. In this class, the number of regimes is possibly infinite and is determined when estimating the model, thus avoiding the need to set this number by a model choice criterion. We develop a new Markov chain Monte Carlo estimation method that solves the path dependence issue due to the moving average component. Empirical results on macroeconomic series illustrate that the proposed class of models dominates the model with fixed parameters in terms of point and density forecasts.

Keywords: ARMA, Bayesian inference, Dirichlet process, Forecasting.

*Luc Bauwens acknowledges support of the "Communauté française de Belgique" through contract "Projet d'Actions de Recherche Concertées" 12/17-045", granted by the "Académie universitaire Louvain".

†Arnaud Dufays has been partly supported by the contract "Investissement d'Avenir" ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047 granted by the Centre de Recherche en Economie et Statistique (CREST) and by the FSR grant rewarded by the Université catholique de Louvain

1 Introduction

Econometricians have developed models with changing parameters at least since Goldfled & Quandt (1973) introduced the idea of Markov-switching (MS) to model the changes in the parameters of a regression equation. This idea consists in enriching the regression with a discrete latent variable process indexing the parameters so that they can switch from one value to another. Hamilton (1989) updated the idea and introduced in particular a filtering algorithm that enables a direct evaluation of the likelihood function. A few years later, Chib (1998) proposed change-point (CP) models, where the transitions from one value to another are not reversible, as a convenient way to model structural breaks at unknown break dates. The estimation of all these models relies on algorithms that are not applicable to models exhibiting path dependence, such as the autoregressive moving average (ARMA) and the generalized autoregressive conditional heteroskedastic (GARCH) models. The difficulty occurs because an unobservable variable at date t (the lagged error term in ARMA models, the lagged conditional variance in GARCH) depends on the entire path of states that have been followed until that date. The computational time thus exponentially grows with the number of time-series observations (for the MS version) and is practically infeasible even for relatively short series.[1]

Selecting the number of states (or regimes) in these models is an important issue. This is typically done by using a model choice criterion after estimating the model with different numbers of regimes. Bayesian inference by Markov chain Monte Carlo (MCMC) is practical even though the evaluation of the likelihood function is infeasible due to path dependence. Although several numerical tools are available for computing the marginal likelihood (Ardia et al. (2009)), it still remains a tedious calculation for complex models (see e.g. Bauwens et al. (2013)). The sticky infinite hidden Markov-switching (sticky IHMS) modelling framework (Fox et al. (2011)) allows us to bypass this demanding computation by assuming a Markov chain with a potentially infinite number of regimes, thus encompassing any finite number of them. This setting has been successfully applied in genetics (Beal

[1]Several papers propose estimation methods for the MS and CP-GARCH models, either circumventing the path dependence issue (e.g. Gray (1996), Klaassen (2002) Haas et al. (2004)) or tackling the issue upfront (e.g. Francq & Zakoian (2008), Henneke et al. (2011), Bauwens et al. (2013)).

& Krishnamurthy (2006)), visual recognition (Kivinen et al. (2007)), and economics, with in the last area in particular autoregressive (AR) models (Song (2014), Jochmann (2015)) and volatility models (e.g. Jensen & Maheu (2010)).

With respect to this background, our contribution is threefold. To begin with, we propose a simple solution to relax the classical assumption of MS models, which states that all the parameters must change whenever a break occurs. To do so, we separate the break dynamics of the mean and variance parameters and use a hierarchical Dirichlet process to drive each of them. Although not based on Dirichlet processes, similar approaches relying on finite-state Markov chains are proposed by Doornik (2013), Goutte (2014), and Eo (2012). While the first two assume a fixed number of regimes, the latter uses the marginal log-likelihood to select the optimal specification. This method is impractical for a large number of regimes or of parameters. In comparison, we only need one estimation to determine the optimal number of regimes. Moreover, our forecasts take the uncertainty of the number of regimes into account without resorting to Bayesian model averaging. Our empirical forecasting results indicate that this feature contributes positively to the predictive performance of the proposed models. Our modeling approach therefore generalizes the conventional MS approach in two related directions: an unbounded number of states (as in existing IHMS-AR models) and a flexibility on the dynamics of the parameters, by allowing different break dates in the parameters of the mean and of the variance.

Secondly, as our baseline model is an ARMA one, we need an estimation method that operates for models subjected to path dependence. We develop a new MCMC algorithm that solves this issue. In addition, the sampling of the ARMA parameters is performed with the manifold Metropolis adjusted Langevin algorithm (MALA) introduced by Girolami & Calderhead (2011).

As a third contribution, we introduce in the econometric literature the steppingstone algorithm (see Xie et al. (2011)), which provides a new way to estimate the marginal log-likelihood (MLL) from the MCMC output.

The rest of the paper is organised in six sections. The model is presented in Section 2 and the estimation procedure in Section 3. The steppingstone algorithm is exposed in Section 4. The prior elicitation and the label switching problem are addressed in Section 5.

Applications, including forecasting evaluations, are presented in Section 6. The last section contains our conclusions. A supplementary appendix (SA) provides additional empirical results.

2 MS-ARMA Models

We start by defining the model with a finite number of states in order to discuss its limitations. In subsection 2.2, we present the Dirichlet process mixture model and the related Dirichlet process. These processes are the building blocks of the infinite hidden Markov-switching framework (IHMS) on which the IHMS-ARMA model class is built. This model class is defined in subsection 2.3. For simplicity, the exposition is limited to the ARMA(1,1) model, the term ARMA being used to designate shortly ARMA(1,1) in the rest of the paper. Similarly, $y_{1:T} = \{y_1, ..., y_T\}$ denotes a generic time series.

2.1 The Model With a Finite Number of Regimes

The MS-ARMA model is defined by

$$y_t = \mu_{s_t} + \beta_{s_t} y_{t-1} + \phi_{s_t} \epsilon_{t-1} + \epsilon_t, \tag{1}$$

$$\epsilon_t \sim N(0, \sigma^2_{s_t}), \tag{2}$$

where $\mu_{s_t} \in \Re$, $\sigma^2_{s_t} > 0$, $|\beta_{s_t}| < 1$ (for stationarity), $|\phi_{s_t}| < 1$ (for invertibility) and $\beta_{s_t} + \phi_{s_t} \neq 0$ (no root cancellation). The elements of the vector $s_{1:T} = \{s_1, ..., s_T\}$ take integer values from 1 to K and denote which regime (also called state hereafter) is active at each period of time. They are assumed to follow a first-order Markov chain with a homogeneous transition probability matrix given by

$$P_{MS} = \begin{pmatrix} p_{11} & p_{12} & ... & p_{1K} \\ p_{21} & p_{22} & ... & p_{2K} \\ & & ... & \\ p_{K1} & p_{K2} & ... & p_{KK} \end{pmatrix},$$

in which p_{ij} denotes the probability of moving from state i to j with the constraint that $\sum_{j=1}^{K} p_{ij} = 1$, $\forall i \in [1, K]$. This setting is similar to that of Hamilton (1989).

Although flexible, the MS-ARMA model has two limitations that motivate our contributions in this paper. Firstly, the number of regimes K must be fixed before the estimation. Indeed, the standard inference method consists in estimating the MS-ARMA model for several a priori plausible values of K, and then choosing the optimal number of regimes by using a model choice criterion. This approach requires several estimations that are tedious when K is large. Moreover, it does not take the uncertainty on the number of regimes into account. Secondly, at each regime switch, all the parameters change simultaneously. However a break may affect only a subset of the parameters. As a consequence, the standard model may be over-parameterized, resulting in imprecise estimates, in particular of the parameters of short regimes, and deteriorated forecast performance.

2.2 The Dirichlet Process Mixture Model

To address the two issues, we adopt a setting that potentially allows for an infinite number of regimes and we disentangle the dynamics of the model parameters by allowing for different break dates for the mean function parameters and for the variance. To do so, we rely on the sticky IHMS framework of Fox et al. (2011) (see also Teh et al. (2006)) which is based on the Dirichlet process (DP) introduced by Ferguson (1973). We introduce first the DP and highlight its interest, before defining the complete model in the next subsection.

The Dirichlet process can be used as a non-parametric prior on model parameters. Its most popular use is the DP mixture model. An example of this for an ARMA model is the following:

$$
\begin{aligned}
y_t &= \mu + \beta y_{t-1} + \phi \epsilon_{t-1} + \epsilon_t, && (3)\\
\epsilon_t &\sim N(0, \sigma_t^2), && (4)\\
\sigma_t^2 | G_0 &\sim G_0, && (5)\\
G_0 | \eta, H &\sim DP(\eta, H). && (6)
\end{aligned}
$$

The Dirichlet process is denoted by $\mathrm{DP}(\eta, H)$, where η is a positive 'concentration' parameter and H a continuous 'base' distribution (for example, an inverse-gamma in this context). The DP expectation is given by H which indicates that any draw G_0 of the process can be seen as a distribution over the same support as H. The concentration parameter η controls

its dispersion with respect to the base distribution. In particular, the larger is η, the more similar are the distributions G_0 and H.

To simplify the presentation, in the model (3)-(6), we only introduce the possibility of breaks in the variance. A more general mixture model can be straightforwardly designed by adding a DP layer to the mean function parameters (as we do in the next subsection). Two useful properties, as well as two related representations, of the Dirichlet process are worth mentioning.

Firstly, the DP non-parametric prior is parsimonious. Indeed, from its Pólya urn representation that allows to integrate the DP (see Blackwell & MacQueen (1973)), the time-varying variance σ_t^2 of (5) is distributed, conditionally on the previous realizations, as

$$\sigma_t^2|\sigma_1^2,...,\sigma_{t-1}^2 \quad \sim \quad \sum_{i=1}^{t-1} \frac{1}{\eta+t-1}\delta_{\sigma_i^2} + \frac{\eta}{\eta+t-1}H, \tag{7}$$

where $\delta_{\sigma_i^2}$ is the probability measure concentrated at σ_i^2. This result shows that the probability of drawing a new value from H decreases as the time index grows. If we denote all the n_i identical values σ_i^2 by $\tilde{\sigma}_i^2$ and assume that at time t, only K different variances have been drawn, then the DP property saying that 'rich regime gets richer' becomes transparent, since Equation (7) is equivalent to

$$\sigma_t^2|\sigma_1^2,...,\sigma_{t-1}^2 \quad \sim \quad \sum_{i=1}^{K} \frac{n_i}{\eta+t-1}\delta_{\tilde{\sigma}_i^2} + \frac{\eta}{\eta+t-1}H. \tag{8}$$

The probability that $\sigma_t^2 = \tilde{\sigma}_i^2$ is given by $n_i/(\eta+t-1)$ and increases with the number of realizations that have already been assigned to regime i. This feature highlights the time-varying nature of the model variance.

Secondly, the Dirichlet process is discrete, which is essential to build a Markov chain with an infinite number of regimes. Sethuraman (1994) shows an alternative to the Pólya urn to construct a Dirichlet process. From two independent sequences of i.i.d. random variables $\{\pi_i\}_{i=1}^{\infty}$ and $\{\sigma_i^2\}_{i=1}^{\infty}$ built as follows

$$\beta_i \sim Beta(1,\eta), \qquad \pi_i = \beta_i \prod_{l=1}^{i-1}(1-\beta_l), \tag{9}$$

$$\sigma_i^2 \sim H, \qquad G_0 = \sum_{i=1}^{\infty} \pi_i \delta_{\sigma_i^2}, \tag{10}$$

it turns out that G_0 is distributed as a Dirichlet process with concentration parameter η and base distribution H. The sequence $\{\pi_i\}_{i=1}^{\infty}$, conveniently written $\pi \sim Stick(\eta)$, satisfies $\sum_{i=1}^{\infty} \pi_i = 1$ with probability one and therefore defines a distribution over the positive integers. The explicit form of G_0 highlights that the DP support is discrete. From the stick-breaking representation (9)-(10), the conditional predictive density of the DP mixture model (3)-(6) is given by

$$f(y_t|y_{1:t-1}, \mu, \beta, \phi, \{\sigma_i^2\}_{i=1}^{\infty}, \{\pi_i\}_{i=1}^{\infty}) = \sum_{i=1}^{\infty} \pi_i f_N(y_t|\mu + \beta y_{t-1} + \phi\epsilon_{t-1}, \sigma_i^2), \quad (11)$$

where $f_N(x|a, b)$ stands for the Normal density function with expectation a and variance b evaluated at x. The predictive density (11) shows that the Dirichlet process helps to move to an infinite number of regimes but also highlights that the transition probabilities to switch from one state to another are independent of time. Teh et al. (2006) were the first to restore the Markovian property in the state transitions by introducing the infinite hidden Markov-switching framework. Afterwards, Fox et al. (2011) developed the sticky-IHMS setting that copes with the high regime persistence typical in a time series context.

2.3 The Model With an Infinite Number of Regimes

The sticky IHMS-ARMA model is defined as

$$\begin{aligned}
y_t &= \mu_t + \beta_t y_{t-1} + \phi_t \epsilon_{t-1} + \epsilon_t, && (12)\\
\epsilon_t &\sim N(0, \sigma_t^2), && (13)\\
\psi_t \equiv \{\mu_t, \beta_t, \phi_t\}|\psi_{t-1}, G_{\psi_{t-1}} &\sim G_{\psi_{t-1}}, && (14)\\
G_{\psi_{t-1}}|G_0 &\sim \mathrm{DP}(\alpha_\psi + \kappa_\psi, \frac{\alpha_\psi G_0 + \kappa_\psi \delta_{\psi_{t-1}}}{\alpha_\psi + \kappa_\psi}), && (15)\\
G_0|\eta_\psi, H_\psi &\sim DP(\eta_\psi, H_\psi), && (16)\\
\sigma_t^2|\sigma_{t-1}^2, G_{\sigma_{t-1}^2} &\sim G_{\sigma_{t-1}^2}, && (17)\\
G_{\sigma_{t-1}^2}|G_1 &\sim \mathrm{DP}(\alpha_\sigma + \kappa_\sigma, \frac{\alpha_\sigma G_1 + \kappa_\sigma \delta_{\sigma_{t-1}^2}}{\alpha_\sigma + \kappa_\sigma}), && (18)\\
G_1|\eta_\sigma, H_\sigma &\sim DP(\eta_\sigma, H_\sigma), && (19)
\end{aligned}$$

where $\delta_{\psi_{t-1}}$ and $\delta_{\sigma_{t-1}^2}$ are probability measures concentrated at ψ_{t-1} and σ_{t-1}^2 respectively and, as in the model with finite number of states, we impose $|\beta_t| < 1$, $|\phi_t| < 1$, and $\beta_t + \phi_t \neq 0$ for all t. The parameters α_ψ, κ_ψ, η_ψ, α_σ, κ_σ, η_σ must be positive.

Equations (12) and (13) correspond to an ARMA model with time-varying parameters. To ensure tractability and to mimic the abrupt switches of the parameters in a Markov-switching model, we use different Dirichlet processes as priors on the mean and the variance parameters. Referring to the Pólya urn Equation (8), the model parameters at time t, conditionally on the previous ones, can stay in or move to an existing regime, or switch to a new one whose values are generated from the base distribution. The two structures, (14)-(16) and (17)-(19), constitute two hierarchical Dirichlet processes. To differentiate them from the Pólya urn Equation (8), as the model parameters are now driven by several distributions (instead of one in the DPM model) coming from multiple Dirichlet processes, the probabilities of moving from one state to another become time-varying and directly depend on the previous parameter. Additionally, as the discrete distribution G_0 is shared among the different Dirichlet processes of Equations (15), the distributions of the mean function parameters share the same possible states. The same comment holds for the variance parameters as the base distribution G_1 is also common to the Dirichlet processes, see Equation (18). The sticky hierarchical Dirichlet framework suggested by Fox et al. (2011) can account for the persistence in the regimes compared to the hierarchical Dirichlet structure proposed by Teh et al. (2006). This is done by introducing the parameters κ_ψ and κ_σ that generate the persistence in the regimes by increasing the probability of picking the parameter of the previous state (hence the qualifier 'sticky').

Using the stick-breaking formulation of the sticky IHMS framework (see Fox et al. (2011)) leads to the following way to formulate the sticky IHMS-ARMA process:

$$
\begin{aligned}
y_t &= \mu_{s_t^\psi} + \beta_{s_t^\psi} y_{t-1} + \phi_{s_t^\psi} \epsilon_{t-1} + \epsilon_t, && (20)\\
\epsilon_t &\sim N(0, \sigma^2_{s_t^\sigma}), && (21)\\
s_t^\psi | s_{t-1}^\psi, \{p_i^\psi\}_{i=1}^\infty &\sim p^\psi_{s_{t-1}^\psi}, && (22)\\
p_i^\psi | \pi^\psi &\sim DP(\alpha_\psi + \kappa_\psi, \frac{\alpha_\psi \pi^\psi + \kappa_\psi \delta_i}{\alpha_\psi + \kappa_\psi}), && (23)\\
\pi^\psi &\sim Stick(\eta_\psi), && (24)\\
\psi_{s_t^\psi} \equiv \{\mu_{s_t^\psi}, \beta_{s_t^\psi}, \phi_{s_t^\psi}\} &\sim H_\psi, && (25)\\
s_t^\sigma | s_{t-1}^\sigma, \{p_i^\sigma\}_{i=1}^\infty &\sim p^\sigma_{s^\sigma_{t-1}}, && (26)
\end{aligned}
$$

$$p_i^{\sigma}|\pi^{\sigma} \sim DP(\alpha_\sigma + \kappa_\sigma, \frac{\alpha_\sigma \pi^{\sigma} + \kappa_\sigma \delta_i}{\alpha_\sigma + \kappa_\sigma}), \tag{27}$$

$$\pi^{\sigma} \sim Stick(\eta_\sigma), \tag{28}$$

$$\sigma^2_{s_t^{\sigma}} \sim H_\sigma, \tag{29}$$

where s_t^{ψ}, s_t^{σ} are discrete random variables that can take any positive integer value. Let us define $\Theta = \{\{\mu_i\}_{i=1}^{\infty}, \{\beta_i\}_{i=1}^{\infty}, \{\phi_i\}_{i=1}^{\infty}, \{\sigma_i\}_{i=1}^{\infty}\}$ and $F_{t-1} = \{y_{1:t-1}, \Theta, \{p_i^{\psi}\}_{i=1}^{\infty}, \{p_i^{\sigma}\}_{i=1}^{\infty}\}$. From the above representation of the IHMS-ARMA model, we can obtain the following predictive densities:

$$f(y_t|F_{t-1}, s^{\psi}_{1:t-1}, s^{\psi}_t, s^{\sigma}_{1:t-1}) = \sum_{j=1}^{\infty} p_{s^{\sigma}_{t-1}j} f_N(y_t|\mu_{s_t} + \beta_{s_t} y_{t-1} + \phi_{s_t}\epsilon_{t-1}, \sigma_j^2), \tag{30}$$

$$f(y_t|F_{t-1}, s^{\psi}_{1:t-1}, s^{\sigma}_{1:t-1}) = \sum_{i=1}^{\infty} p_{s^{\psi}_{t-1}i} \Big[\sum_{j=1}^{\infty} p_{s^{\sigma}_{t-1}j} f_N(y_t|\mu_i + \beta_i y_{t-1} + \phi_i \epsilon_{t-1}, \sigma_j^2) \Big]. \tag{31}$$

Equation (30) highlights that the conditional distribution given the current state of the mean function parameters is an infinite mixture of Normal distributions with time-varying probabilities. When we integrate over the current mean state value (s_t^{ψ}), Equation (31) emphasizes that the model is equivalent to a MS-ARMA model with an infinite number of regimes for the mean function parameters and for the variance. This avoids the two drawbacks of the MS-ARMA model with finite number of states mentioned at the end of subsection 2.1.

Due to the DP assumptions (23) and (27), the expected values of the i-th rows (p_i^{ψ} and p_i^{σ}) of the infinite dimensional transition probability matrices are given by

$$E(p_i^{\psi}|\alpha_\psi, \pi^{\psi}, \kappa_\psi) = \frac{\alpha_\psi \pi^{\psi} + \kappa_\psi \delta_i}{\alpha_\psi + \kappa_\psi}, \quad E(p_i^{\sigma}|\alpha_\sigma, \pi^{\sigma}, \kappa_\sigma) = \frac{\alpha_\sigma \pi^{\sigma} + \kappa_\sigma \delta_i}{\alpha_\sigma + \kappa_\sigma}. \tag{32}$$

These formulas show that the expected self-transition probability is inflated compared to the probability of moving to another state thanks to the positive sticky parameters κ_ψ and κ_σ. We therefore have an infinite dimensional Markovian structure encouraging regime persistence.

Regarding the base distributions H_ψ and H_σ – see (25) and (29)– a third hierarchical layer is introduced in Section 5 in order to update them with information stemming from the active regimes. As advocated by Song (2014), this layer improves the birth of new

regimes by drawing realistic parameters from the common distributions. The types and hyper-parameters of these and all other prior distributions are defined in Section 5.

The IHMS framework has been used in several empirical applications. Jochmann (2015) and Song (2014) use it to model macroeconomic series with an autoregressive model (thus without path dependence). In volatility modeling, Jensen & Maheu (2010), Jensen & Maheu (2013), Jensen & Maheu (2014), Dufays (2015) and Jin & Maheu (2014) also apply this kind of structure to GARCH, stochastic volatility and realized volatility processes. All these papers provide empirical evidence in favour of IHMS models compared to the existing alternatives. The proposed sticky IHMS-ARMA model makes no exception as shown by the forecasting results reported in Section 6. Furthermore, we innovate in two directions with respect to those papers. Firstly, the model relies on two sticky IHMS structures, improving its flexibility. Secondly, we extend the model to include a MA component, thus we face the complications due to the path dependence issue. Large sample properties of time-varying ARMA parameters have been the focus of another strand of the literature (Basawa & Lund (2001), Francq & Gautier (2004) among others). However, contrary to Basawa & Lund (2001), our approach does not assume periodic changes in the ARMA time-varying parameters. Our approach also departs from Francq & Gautier (2004) by relaxing the assumption that changes in the mean and variance parameters are simultaneous. As a final remark, the exposition has used an ARMA(1,1) specification with Normal errors, but a higher order ARMA model or a different innovation distribution can be handled without complications.

3 Estimation by MCMC

Two issues need to be addressed to estimate the IHMS-ARMA model: the path dependence issue, and the infinite number of regimes revealed by the predictive distribution shown in Equation (31). We deal with the former in the same way as Dufays (2015) proposed in the GARCH context; for details, see Section 3.1. To tackle the second issue, one can rely on the beam sampler (Van Gael et al. (2008)), which augments the posterior distribution with a set of auxiliary variables that truncate the infinite number of states to a finite one. As the posterior distribution marginalized with respect to the auxiliary variables corresponds to

the targeted posterior one, the MCMC algorithm is correct. A simpler alternative consists in truncating the infinite sum to a large number of states L without embedding auxiliary random variables, a technique known as the *degree L weak limit approximation* (Ishwaran & Zarepour (2002)). Despite the truncation, if the chosen number L is large enough, the error is negligible, see Kurihara et al. (2007) and Fox et al. (2011). In this paper, we rely on this approximation as it eases the algorithm implementation and its exposition. However, with slight modifications, the MCMC scheme would also operate with the auxiliary variable approach.

The estimation is based on the stick-breaking representation in Equations (20)-(29). Under the degree L weak limit approximation, every row of each transition matrix is truncated to a finite Dirichlet distribution (denoted by Dir), instead of being driven by a Dirichlet process as in Equations (23) and (27). Thus, (23) and (27) are replaced by

$$\begin{aligned} p_i^{\psi} = \{p_{i1}^{\psi}, p_{i2}^{\psi}, ..., p_{iL}^{\psi}\} | \pi^{\psi}, \alpha_{\psi}, \kappa_{\psi} &\sim Dir(\alpha_{\psi}\pi_1^{\psi}, \alpha_{\psi}\pi_2^{\psi}, ..., \alpha_{\psi}\pi_i^{\psi} + \kappa_{\psi}, ..., \alpha_{\psi}\pi_L^{\psi}), \\ p_i^{\sigma} = \{p_{i1}^{\sigma}, p_{i2}^{\sigma}, ..., p_{iL}^{\sigma}\} | \pi^{\sigma}, \alpha_{\sigma}, \kappa_{\sigma} &\sim Dir(\alpha_{\sigma}\pi_1^{\sigma}, \alpha_{\sigma}\pi_2^{\sigma}, ..., \alpha_{\sigma}\pi_i^{\sigma} + \kappa_{\sigma}, ..., \alpha_{\sigma}\pi_L^{\sigma}), \end{aligned}$$

respectively. In the same spirit, the probability distributions π^{ψ} and π^{σ} of the stick-breaking representation are truncated to L elements following symmetric Dirichlet prior distributions given the parameters (η_{ψ} and η_{σ}):

$$\begin{aligned} \pi^{\psi} = \{\pi_1^{\psi}, \pi_2^{\psi}, ..., \pi_L^{\psi}\} | \eta_{\psi} &\sim Dir(\frac{\eta_{\psi}}{L}, \frac{\eta_{\psi}}{L}, ..., \frac{\eta_{\psi}}{L}), \\ \pi^{\sigma} = \{\pi_1^{\sigma}, \pi_2^{\sigma}, ..., \pi_L^{\sigma}\} | \eta_{\sigma} &\sim Dir(\frac{\eta_{\sigma}}{L}, \frac{\eta_{\sigma}}{L}, ..., \frac{\eta_{\sigma}}{L}). \end{aligned}$$

For notational ease, the sticky IHMS parameters $\alpha = \{\alpha_{\psi}, \alpha_{\sigma}\}, \kappa = \{\kappa_{\psi}, \kappa_{\sigma}\}, \eta = \{\eta_{\psi}, \eta_{\sigma}\}$ are brought together in the set $H_{Dir} = \{\alpha, \eta, \kappa\}$ and the truncated transition matrices are denoted by

$$P^{\psi} = \begin{pmatrix} p_{11}^{\psi} & p_{12}^{\psi} & ... & p_{1L}^{\psi} \\ p_{21}^{\psi} & p_{22}^{\psi} & ... & p_{2L}^{\psi} \\ & & ... & \\ p_{L1}^{\psi} & p_{K2}^{\psi} & ... & p_{LL}^{\psi} \end{pmatrix}, \quad P^{\sigma} = \begin{pmatrix} p_{11}^{\sigma} & p_{12}^{\sigma} & ... & p_{1L}^{\sigma} \\ p_{21}^{\sigma} & p_{22}^{\sigma} & ... & p_{2L}^{\sigma} \\ & & ... & \\ p_{L1}^{\sigma} & p_{K2}^{\sigma} & ... & p_{LL}^{\sigma} \end{pmatrix}.$$

Bayesian estimation is feasible by explicitly treating $s_{1:T}^{\psi}, s_{1:T}^{\sigma}$ as parameters. To simulate the posterior distribution, we use a Gibbs sampler that cycles between eight full

conditional distributions, as summarised in Table 1, where $\bar{\mu}, \bar{\Sigma}, \bar{e}, \bar{f}$ are the parameters of the base distributions H_ψ and H_σ. Details about the prior distributions are provided in Section 5.

Table 1: Sticky IHMS-ARMA Gibbs sampler

1. $f(s^{\psi}_{1:T}\|\Theta, P^{\psi}, s^{\sigma}_{1:T}, y_{1:T})$	5. $f(\Theta\|\bar{\mu}, \bar{\Sigma}, \bar{e}, \bar{f}, H_{Dir}, s^{\psi}_{1:T}, s^{\sigma}_{1:T}, y_{1:T})$
2. $f(P^{\psi}\|\Theta, H_{Dir}, \pi^{\psi}, s^{\psi}_{1:T}, y_{1:T})$	6. $f(\bar{\mu}, \bar{\Sigma}, \bar{e}, \bar{f}\|\Theta, H_{Dir}, s^{\psi}_{1:T}, s^{\sigma}_{1:T}, y_{1:T})$
3. $f(s^{\sigma}_{1:T}\|\Theta, P^{\sigma}, s^{\psi}_{1:T}, y_{1:T})$	7. $f(\pi^{\psi}, \pi^{\sigma}\|\Theta, P^{\psi}, P^{\sigma}, H_{Dir}, s^{\psi}_{1:T}, s^{\sigma}_{1:T}, y_{1:T})$
4. $f(P^{\sigma}\|\Theta, H_{Dir}, \pi^{\sigma}, s^{\sigma}_{1:T}, y_{1:T})$	8. $f(H_{Dir}\|P^{\psi}, P^{\sigma}, \pi^{\psi}, \pi^{\sigma}, s^{\psi}_{1:T}, s^{\sigma}_{1:T}, y_{1:T})$

Except in steps 1 and 5 in Table 1, the full conditional distributions can be directly simulated, as detailed in Appendix 1. In the rest of this section, we concentrate on the most challenging item of the sampler (step 1), and we expose how we sample the ARMA parameters in subsection 3.2. For model comparisons based on forecasts, we detail how we compute and evaluate the predictive density at any horizon in subsection 3.3.

3.1 Sampling the State Vector of the Mean Function Parameters

Sampling the state vector is usually done by a forward-backward algorithm (Rabiner (1989), Chib (1998)). The algorithm is applicable if there is no path dependence, as in the case of an AR model, but not of an ARMA model. In an AR model, the likelihood of observation t only depends on the current state, while in the MA one, it depends on the whole path of past states. The forward-backward method is then infeasible since the computations exponentially grow with the time index t. For the ARMA model, we adapt the method of Dufays (2015) and follow a two-step procedure. We first sample an entire state vector from an approximate model, which is a modified MS-ARMA model adapted from Klaassen (2002). Such a model is free from the path dependence problem, so that the forward-backward algorithm can be used to sample a state vector. We then implement the Metropolis-Hastings (MH) step to accept or reject the draw. Although an approximate model is used to sample the state vector, the MH step ensures that the targeted distribution remains the posterior one of the IHMS-ARMA model.

GARCH and ARMA models are closely related when tackling the path dependence problem. Reliable approximations of the MS-GARCH process have been proposed by Gray (1996), Klaassen (2002), and Haas et al. (2004). We adapt the approach of Klaassen (2002) to the ARMA case. This replaces the unobserved error ϵ_{t-1} in the ARMA equation by its conditional expectation $E_{s_{t-1}^{\psi}}[\epsilon_{t-1}|y_{1:t-1}, s_t^{\psi}, \Theta, P^{\psi}, s_{1:t-1}^{\sigma}]$ denoted by $\tilde{\epsilon}_{t-1,s_t^{\psi}}$ below. The approximate model is

$$y_t = \mu_{s_t^{\psi}} + \beta_{s_t^{\psi}} y_{t-1} + \phi_{s_t^{\psi}} \tilde{\epsilon}_{t-1,s_t^{\psi}} + \epsilon_t(s_t^{\psi}),$$

where $\tilde{\epsilon}_{t-1,s_t^{\psi}} = \sum_{i=1}^{L} \epsilon_{t-1}(i) f(s_{t-1} = i|y_{1:t-1}, s_t^{\psi}, \Theta, P^{\psi}, s_{1:t-1}^{\sigma})$ and $\epsilon_{t-1}(s_{t-1}^{\psi}) = y_{t-1} - \mu_{s_{t-1}^{\psi}} - \beta_{s_{t-1}^{\psi}} y_{t-2} - \phi_{s_{t-1}^{\psi}} \tilde{\epsilon}_{t-2,s_{t-1}^{\psi}}$ (see Appendix 2 for the computation of $\tilde{\epsilon}_{t-1,s_t^{\psi}}$). The approximation eliminates the path dependence problem since the error term $\epsilon_t(s_t^{\psi})$ only depends on the current state and not also on the past sequence of states.

We therefore sample a new state vector $s_{1:T}^{\psi'}$ from the MS-ARMA approximation employing the forward-backward algorithm. The proposed parameter is accepted according to the MH ratio:

$$\begin{aligned}
\alpha(s_{1:T}^{\psi}, s_{1:T}^{\psi'}|y_{1:T}, \Theta, s_{1:T}^{\sigma}, P^{\psi}) &= \min\{1, \frac{f(s_{1:T}^{\psi'}|y_{1:T}, \Theta, s_{1:T}^{\sigma}, P^{\psi}) q(s_{1:T}^{\psi}|y_{1:T}, \Theta, s_{1:T}^{\sigma}, P^{\psi})}{f(s_{1:T}^{\psi}|y_{1:T}, \Theta, s_{1:T}^{\sigma}, P^{\psi}) q(s_{1:T}^{\psi'}|y_{1:T}, \Theta, s_{1:T}^{\sigma}, P^{\psi})}\} \\
&= \min\{1, \frac{f(y_{1:T}|s_{1:T}^{\psi'}, \Theta, s_{1:T}^{\sigma}) f(s_{1:T}^{\psi'}|P^{\psi}) q(s_{1:T}^{\psi}|y_{1:T}, \Theta, s_{1:T}^{\sigma}, P^{\psi})}{f(y_{1:T}|s_{1:T}^{\psi}, \Theta, s_{1:T}^{\sigma}) f(s_{1:T}^{\psi}|P^{\psi}) q(s_{1:T}^{\psi'}|y_{1:T}, \Theta, s_{1:T}^{\sigma}, P^{\psi})}\}
\end{aligned}$$

where $q(s_{1:T}^{\psi}|y_{1:T}, \Theta, s_{1:T}^{\sigma}, P^{\psi})$ is the proposal distribution of $s_{1:T}^{\psi}$ derived from the forward-backward algorithm.

A proposed $s_{1:T}^{\psi}$ is very likely to be rejected if it is drawn as one block from the MS-ARMA approximation. To ensure good MCMC mixing properties, we sample the state vector in blocks of random sizes sampled from a uniform distribution with lower bound equal to 40 and upper bound to 150 (see e.g. Jensen & Maheu (2010) in a stochastic volatility context). This avoids situations where the MCMC algorithm always rejects the proposed state vector and also enhances the acceptance rate. In our empirical applications, the average of the acceptance rate over the random block sizes is always above 40%.

3.2 Sampling the ARMA Parameters

Due to the unobserved lagged error term, the full conditional distribution of the ARMA parameters cannot be simulated directly. Nevertheless, given the mean function parameters and the state vectors, the full conditional distribution of the variances is a product of (conditionally) independent inverse-gamma distributions if the base distribution H_σ is itself an inverse-gamma (see Section 5 for prior densities and Appendix 1 for details). We therefore split the block into two pieces and sample first the variances conditionally on all the other parameters. Then the mean function parameters are drawn from their full conditional distributions using an MH algorithm.

Focusing on the mean function parameters, we adapt the Riemannian Manifold Metropolis Adjusted Langevin Algorithm (RMMALA) to define a proposal distribution (see Girolami & Calderhead (2011) and its corrected version, Xifara et al. (2014)). The RMMALA algorithm is a discrete version of an Ito stochastic differential equation of the Langevin diffusion, which exhibits the full conditional distribution as unique stationary one. Focusing on the mean function parameters $\psi_i = \{\mu_i, \beta_i, \phi_i\}$ of the i-th regime and denoting by $\log f(\psi_i|D)$ the logarithm of the full conditional density, where $D = \{\{\psi_j\}_{j\neq i, j=1}^{L}, \{\sigma_i\}_{i=1}^{L}, \bar{\mu}, \bar{\Sigma}, \bar{e}, \bar{f}, H_{Dir}, s_{1:T}^{\psi}, s_{1:T}^{\sigma}, y_{1:T}\}$, the RMMALA proposal distribution, given the current MCMC realization ψ_i, is defined by

$$\tilde{\psi}|\psi_i \quad \sim \quad N(\xi(\psi_i, \gamma), \gamma^2 G^{-1}(\psi_i)),$$

where $G(\psi_i)$ denotes the Hessian of minus the logarithm of $f(\psi_i|D)$, γ is a discretization tuning constant, and $\xi(.,.)$ stands for a function of the gradient, the Hessian and the third derivative of minus the logarithm of $f(\psi_i|D)$. If we assume that the curvature is locally constant, the proposal distribution is simplified as follows:

$$\tilde{\psi}|\psi_i \quad \sim \quad N(\psi_i + \frac{\gamma^2}{2} G^{-1}(\psi_i) \nabla \log f(\psi_i|D), \gamma^2 G^{-1}(\psi_i)), \tag{33}$$

where ∇ denotes the gradient operator. The proposal distribution (33) is called the simplified manifold MALA (smMALA). Intuitively, the proposal expectation lies in a high density area of the posterior distribution thanks to the gradient. Moreover, the translation takes the local curvature into account through the Hessian matrix. Girolami & Calderhead

(2011) provide several examples in which the proposal distribution (33) allows to update strongly correlated parameters in one block.

Although very appealing, the proposal distribution requires the computation of the Hessian, which can be negative definite. To circumvent the issue as well as to speed up the Hessian computation, we use its Gauss-Newton approximation, as suggested by Vakilzadeh et al. (2014). It is given by $G(\psi_i) \approx J'(\psi_i)J(\psi_i) + C_{\text{prior}}^{-1}$, where $J(\psi_i)$ is the Jacobian of the standardized error terms and C_{prior}^{-1} is the inverse of the covariance matrix of the prior distribution.

The tuning constant γ has also an impact on the proposal density and therefore on the MCMC mixing properties. If its value is too large, the MCMC sampler can be stuck for long periods, while if it is too small, the proposed update is likely to be accepted but the posterior support exploration is very slow. We solve this issue by adapting the rule of Atchadé & Rosenthal (2005): at the r-th MCMC iteration, the constant γ_r is updated as $\gamma_r = \max\left(\zeta, \gamma_{r-1} + (\alpha - \alpha_{opt})/(0.6^r)\right)$, where ζ is a very small positive constant to avoid a negative value of γ_r, α is the current acceptance rate of the MH algorithm, and α_{opt} stands for the user-defined one. In the empirical applications, α_{opt} is set to 40% and ζ to 10^{-8}.

3.3 Predictive Densities and the Continuously Ranked Probability Score

The usefulness of a model can be assessed by its forecasting ability. We explain how to obtain draws from the predictive density $f(y_{T+h}|y_{1:T})$ where h is the forecast horizon. The predictive density can be estimated by

$$\begin{aligned} f(y_{T+h}|y_{1:T}) &\approx \frac{1}{R}\sum_{r=1}^{R} f(y_{T+h}|\{\Theta, P^{\psi}, P^{\sigma}, s_{1:T+h}^{\psi}, s_{1:T+h}^{\sigma}, y_{T+1:T+h-1}\}^r, y_{1:T}), \\ &\approx \frac{1}{R}\sum_{r=1}^{R} f_N(y_{T+h}|\mu_{s_{T+h}^{\psi,r}}^r + \beta_{s_{T+h}^{\psi,r}}^r y_{T+h-1}^r + \phi_{s_{T+h}^{\psi,r}}^r \epsilon_{T+h-1}^r, (\sigma_{s_{T+h}^{\sigma,r}}^2)^r), \end{aligned}$$

where the R draws (indexed by the superscript r) come from the posterior distribution $f(\Theta, P^{\psi}, P^{\sigma}, s_{1:T+h}^{\psi}, s_{1:T+h}^{\sigma}, y_{T+1:T+h-1}|y_{1:T})$. Consequently, the computation of the predictive density requires to sample future observations and states in the Gibbs sampler sketched

in Table 1. To do that, we add a step in the sampler to draw the future states and observations. Table 2 documents how this is done.

Table 2: Sampling from $f(y_{T+1:T+h-1}, s^{\sigma}_{T+1:T+h}, s^{\psi}_{T+1:T+h}|\Theta, P^{\psi}, P^{\sigma}, s^{\psi}_{1:T}, s^{\sigma}_{1:T}, y_{1:T})$

For $j = 1$ to $h-1$ Do
Sample $s^{\sigma}_{T+j} \sim \text{Mult}(p_{s^{\sigma}_{T+j-1}1}, p_{s^{\sigma}_{T+j-1}2}, ..., p_{s^{\sigma}_{T+j-1}L})$
Sample $s^{\psi}_{T+j} \sim \text{Mult}(p_{s^{\psi}_{T+j-1}1}, p_{s^{\psi}_{T+j-1}2}, ..., p_{s^{\psi}_{T+j-1}L})$
Sample $\epsilon_{T+j} \sim N(0, \sigma^2_{s^{\sigma}_{T+j}})$
Set $y^{\psi}_{T+j} = \mu_{s^{\psi}_{T+j}} + \beta_{s^{\psi}_{T+j}} y_{T+j-1} + \phi_{s^{\psi}_{T+j}} \epsilon_{T+j-1} + \epsilon_{T+j}$
EndFor
Sample $s^{\sigma}_{T+h} \sim \text{Mult}(p_{s^{\sigma}_{T+h-1}1}, p_{s^{\sigma}_{T+h-1}2}, ..., p_{s^{\sigma}_{T+h-1}L})$
Sample $s^{\psi}_{T+h} \sim \text{Mult}(p_{s^{\psi}_{T+h-1}1}, p_{s^{\psi}_{T+h-1}2}, ..., p_{s^{\psi}_{T+h-1}L})$

'Mult' stands for Multinomial distribution.

The predictive density evaluated at the realized data is a prominent metric to assess the predictive performance of a model but other loss functions exist and can provide additional information about the performance of a model. For this purpose, we also use the mean squared forecast errors (MSFE) and the continuously ranked probability score (CRPS) popularized by Gneiting & Raftery (2007) in our empirical applications.

The CRPS is based on the Brier probability score and relies on the idea that any density forecast f of a random variable Y induces a probability forecast for the binary event $\{Y \leq z\}$ through its cumulative density function, i.e. $F(z) = \int_{-\infty}^{z} f(u)du$. The CRPS is defined as

$$S(f, y) = \int_{-\infty}^{\infty} (F(z) - 1_{\{y \leq z\}})^2 dz. \tag{34}$$

The score function is strictly proper if the random variable Y has finite first moment and goes to an infinite value otherwise. The value of the function (34) can be computed by simulation since $S(f, y) = E_F(|Y - y|) - \frac{1}{2}E_F(|Y - Y'|)$, where Y and Y' are independent random variables from the same distribution F (see, e.g., Gneiting & Raftery (2007)). The

loss function assesses the forecast performance of a model through the distance between the predictive cumulative distribution function of the chosen model and the empirical one.

In forecast evaluations, we are interested in the mean of the score function for the k-ahead predictive density over a time window, i.e. $\bar{S}_k^\tau(\hat{f}_{k+1:k+\tau}, y_{k+1:k+\tau}) = \frac{1}{\tau}\sum_{t=1}^{\tau} S(\hat{f}_{t+k}, y_{t+k})$, where $\hat{f}_{t+k}$ denotes the value of the predictive probability density function and y_{t+k} the observed value of the time series. Gneiting & Ranjan (2011) emphasize two appealing features of the mean of the CRPS. Firstly, the equation (34) becomes

$$\bar{S}_k^\tau(\hat{f}_{k+1:k+\tau}, y_{k+1:k+\tau}) = \int_{-\infty}^{\infty} \frac{1}{\tau}\sum_{t=1}^{\tau}(\hat{F}_{t+k}(z) - 1_{\{y_{t+k}\leq z\}})^2 dz. \tag{35}$$

The term inside the integral can be plotted with respect to z in order to see where the cumulative density function deviates from the empirical distribution.

Secondly, as in Amisano & Giacomini (2007), one can derive a statistical test to assess if a model M_1 produces a better score function than an alternative M_2. The test consists in comparing the mean of the score functions as follows:

$$t_\tau^k = \frac{\bar{S}_k^\tau(\hat{f}_{k+1:k+\tau}^{M_1}, y_{k+1:k+\tau}) - \bar{S}_k^\tau(\hat{f}_{k+1:k+\tau}^{M_2}, y_{k+1:k+\tau})}{\sqrt{\sigma_\tau^2/\tau}} \quad \rightarrow \quad N(0,1), \tag{36}$$

where $\sigma_\tau^2 = \frac{1}{\tau}[\sum_{t=1}^{\tau}\Delta_{t,k}^2 + 2\sum_{j=1}^{k-1}\sum_{t=1}^{\tau-k-j}\Delta_{t,k}\Delta_{t+j,k}]$ and $\Delta_{t,k} = S(\hat{f}_{t+k}^{M_1}, y_{t+k}) - S(\hat{f}_{t+k}^{M_2}, y_{t+k})$. In the predictive exercises, we apply this test to our three score functions.

4 Model Selection

In Bayesian inference, model comparison is often carried out through Bayes factors that require the computation of the marginal likelihood (ML). In this section, we adapt the steppingstone sampling (see Xie et al. (2011)) used in phylogenetics to estimate the marginal likelihood from a MCMC output. The approach relies on multiple importance sampling steps and is actually a generalization of the bridge sampling algorithm (e.g. Fruhwirth-Schnatter (2004)).

The ML is the normalizing constant of the posterior distribution and, gathering all the random parameters of the model in Ψ, is defined as $f(y_{1:T}) = \int f(y_{1:T}|\Psi)f(\Psi)d\Psi$.

As this integration is intractable for most models, the importance sampling (IS) approach makes use of a proposal distribution defined on the support of Ψ to obtain an

estimator of the marginal likelihood by simulation, as follows:

$$f(y_{1:T}) = \int f(y_{1:T}|\Psi)f(\Psi)\frac{q(\Psi|y_{1:T})}{q(\Psi|y_{1:T})}d\Psi, \quad (37)$$

$$\approx \frac{1}{N}\sum_{r=1}^{N}\frac{f(y_{1:T}|\Psi^r)f(\Psi^r)}{q(\Psi^r|y_{1:T})}, \quad (38)$$

where the realizations $\{\Psi^r\}_{r=1}^N$ constitute an ergodic sample from the proposal density $q(\Psi|y_{1:T})$. The IS estimator defined above is almost surely consistent under conditions stated by Geweke (1989). The precision of the IS estimator depends on the quality of the proposal distribution, which should be a good enough approximation of the posterior, so that the variance of the ratio of the posterior to the proposal is finite and as small as possible. This is typically very hard to achieve if Ψ is of high dimension.

Instead of applying importance sampling using a single proposal candidate, the steppingstone algorithm considers a sequence of IS steps using tempered posterior distributions as proposal ones. Let $\Phi(x) : [0, 1, ..., p] \to [0, 1]$ with $\Phi(0) = 0$ and $\Phi(p) = 1$ be an increasing function. The tempered posterior distributions are specified as

$$f_x(\Psi|y_{1:T}) = \frac{f(y_{1:T}|\Psi)^{\Phi(x)}f(\Psi)}{Z_x} \propto f(y_{1:T}|\Psi)^{\Phi(x)}f(\Psi),$$

where $Z_x = \int f(y_{1:T}|\Psi)^{\Phi(x)}f(\Psi)d\Psi$ stands for the normalizing constant (or the marginal likelihood) of the tempered posterior distribution f_x. When $x = 0$, the distribution coincides with the prior one and when $x = p$, it coincides with the targeted posterior distribution. The steppingstone method aims to build a sequence of bridging distributions from the prior to the posterior. Note that the ML is given by $\prod_{k=1}^{p}(Z_k/Z_{k-1})$ when the prior distribution is proper (i.e. integrates to one, so that $Z_0 = 1$). Then using the IS approach, we have that

$$\frac{Z_k}{Z_{k-1}} \approx \frac{1}{N}\sum_{r=1}^{N} f(y_{1:T}|\Psi^r)^{\Phi(k)-\Phi(k-1)}, \quad (39)$$

where the realizations $\{\Psi^r\}_{i=1}^N$ form an ergodic sample drawn from $f_{k-1}(\Psi|y_{1:T})$. For a given function $\Phi(x)$, Equation (39) provides a simple tool to sequentially compute the ML of any model by MCMC. Indeed, one just needs to adapt the MCMC scheme in order to obtain draws from the distribution $f_x(\Psi|y_{1:T})$.

The steppingstone method is unsatisfactory in the sense that the accuracy of the estimator depends on a function $\Phi(x)$ that is model-dependent. In the literature, different functions have been proposed (Xie et al. (2011) and Herbst & Schorfheide (2012)) and a consensus has emerged to suggest that more IS steps should be devoted to very small values of Φ. However, this does not help much to select the function. Instead of fixing it, we use the Sequential Monte Carlo (SMC) theory to make the steppingstone algorithm function-free. Indeed, although not recognized by their authors, the steppingstone algorithm is actually an adaptation of the ML computation proposed in the SMC sampler (see Del Moral et al. (2006)) to MCMC methods. We suggest therefore to build the tempered function from one IS step to the next by using the effective sample size (ESS) criterion (see Doucet et al. (2001)). The ESS is defined as $N/\sum_{r=1}^{N} W_r^2$, where W_r is the normalised weight given by $W_r \propto f(y_{1:T}|\Psi^r)^{\Phi(k)-\Phi(k-1)}$ in the present context. Hence, the ESS is a function of $\Phi(k)$ and we set the value of $\Phi(k)$ by solving the following optimisation program:

$$\Phi(k) \quad = \quad \text{argmax}_{\Phi(k)} ESS(\Phi(k)) \;\; \text{s.t.} \;\; ESS(\Phi(k)) < 0.75.$$

This optimization is standard in the SMC literature (see Jasra et al. (2011) or Dufays (2014)) and avoids the difficult choice of the tempering function.

As a final note, marginal likelihoods of Markov-switching models can be biased if computed by the MCMC technique of Chib (1995). The issue coming from the label switching problem is addressed in Fruhwirth-Schnatter (2004) where the bridge sampling method is introduced. Nevertheless, the bridge sampling accuracy highly depends on an user-defined proposal distribution and therefore can be difficult to use in practice, especially in high dimension. The steppingstone sampling solves the problem as no extra distribution is required and as it does not fix the posterior distribution at a specific value of the parameters as in Chib (1995).

5 Prior Elicitation and Label Switching Problem

Table 3 reports the prior distributions and their hyper-parameters. Regarding the sticky IHMS parameter set H_{Dir}, the priors are conjugate, as suggested by Fox et al. (2011). Table 3 suggests two different values for the hyper-parameters of the persistence variables ρ_ψ and

ρ_σ. The first one ($\omega_{\text{MS/CP}} = 10$) implies a weak state persistence. The break dynamics is likely to rapidly switch from one state to another, hence the name Markov-switching type. The second value ($\omega_{\text{MS/CP}} = 1000$) induces high state persistence (hence the name change-point type). The posterior results are likely to be easier to interpret as only few changes will occur in the break dynamics. This value therefore induces a kind of change-point behaviour. The two cases are considered in the empirical applications, and we can discriminate between the two prior types by the marginal likelihood criterion.

The prior on the parameters of the ARMA mean function is a hierarchical Normal-Wishart distribution that provides an additional layer on the base distributions of the Dirichlet processes. The marginal prior density of the AR and MA parameters, taking into account their truncation to the interval (-1,+1), is almost uniform on that interval (with mean 0 and standard deviation 0.57). Gathering information of existing regimes, this structure facilitates the birth of new regimes without complicating the MCMC simulation (since the prior is conjugate). A similar idea is applied to the variances.

The posterior distribution is invariant to the label of the state vector. If a label switch occurs in the state vectors during the MCMC simulation, the usual summary statistics such as the posterior means and standard deviations are misleading. Indeed these statistics depend on the label of the state. Different solutions exist to solve this issue. The prior distributions can be chosen to rule out the label switching problem by constraining the support of the parameters given the regimes. However, finding appropriate constraints to preclude all the possible switches without truncating heavily the posterior distribution can be difficult. Otherwise, as advocated by Geweke (2007), the label switching issue can be completely ignored in the MCMC simulations. In this case, either a loss function is used to sort the posterior draws in one specific label ordering (see e.g. Marin et al. (2005), Bauwens et al. (2013)) or the reported summary statistics must be label invariant (see Song (2014), Dufays (2015)). We apply the latter approach. For instance, the posterior draws deliver, among other things, the probabilities of having a number of regimes for the mean function parameters and the variance. From the MCMC sample, we can also compute the posterior means and the confidence intervals of the model parameters as they evolve through time. These summary statistics do not depend on a specific label and can therefore be reported

Table 3: Prior distributions

Prior distributions of the Dirichlet parameters

For the Mean:	$\eta_\psi \sim G(1,10)$	$\alpha_\psi + \kappa_\psi \sim G(1,10)$
		$\rho_\psi = \frac{\kappa_\psi}{\alpha_\psi+\kappa_\psi} \sim Beta(\omega_{\text{MS/CP}},1)$
For the Variance:	$\eta_\sigma \sim G(1,10)$	$\alpha_\sigma + \kappa_\sigma \sim G(1,10)$
		$\rho_\sigma = \frac{\kappa_\sigma}{\alpha_\sigma+\kappa_\sigma} \sim Beta(\omega_{\text{MS/CP}},1)$
Markov-switching type: $\omega_{\text{MS/CP}} = 10$		Change-point type: $\omega_{\text{MS/CP}} = 1000$

Prior distributions of the ARMA parameters

For each regime i: $\{\mu_i, \beta_i, \phi_i\} \sim H_\psi \equiv \text{N}(\bar{\mu},\bar{\Sigma})\ \delta_{\{|\beta_i|<1,|\phi_i|<1\}}$

Hierarchical parameter: $\bar{\mu}$	Hierarchical parameter: $\bar{\Sigma}$
$\bar{\mu} \sim \text{N}(\underline{\mu},\underline{\Sigma})$	$\bar{\Sigma}^{-1} \sim \text{W}(\underline{V},\underline{v})$
$\underline{\mu} = \{0,0,0\}$, $\underline{\Sigma} = 0.1 I_3$	$\underline{V} = \frac{1}{5\underline{v}} I_3$, $\underline{v} = 5$

Prior distributions of the variances

For each regime i: $\{\sigma_i^{-2}\} \sim H_\sigma \equiv \text{G}(\bar{e},\bar{f})$

Hierarchical parameter: $\bar{e}$	Hierarchical parameter: $\bar{f}$
$\bar{e} \sim \text{Exp}(\underline{e_a})$	$\bar{f}^{-1} \sim \text{G}(\underline{f_a},\underline{f_b})$
$\underline{e_a} = 2$	$\underline{f_a} = 10$, $\underline{f_b} = 1/5$

G(a,b) stands the Gamma density with shape parameter a and scale b (see Appendix 1). W(S,ν) stands for the Wishart density with scale matrix parameter S and shape parameter ν. I_d stands for the identity matrix of dimension d and $\delta_{\{a>0\}}$ is the Dirac function taking the value one if the constraint $a > 0$ holds and zero otherwise.

without bothering about the label invariance issue.

6 Applications

In this section, several estimation and forecasting results for the sticky IHMS-ARMA model are provided for the quarterly U.S. GDP growth rate and a monthly U.S inflation series. Afterwards, a forecasting exercise on eighteen macroeconomic series is reported to compare the performances of the sticky IHMS-ARMA and of the ARMA model with fixed parameters. Regarding the MCMC implementation, the starting values are the maximum likelihood estimates of the model without any break. The burn-in period uses 7,500 iterations and the next 22,500 draws are stored to compute the posterior results. The number L of the degree L weak limit approximation is fixed to ten. Additional results are provided in the SA.

6.1 U.S. GDP Growth Rate

Hamilton (1989) applied the MS model using an AR mean function to the U.S. quarterly GDP growth rate series. We revisit this example by using the sticky IHMS-ARMA model and its AR version on the series from 1947Q2 to 2014Q1 (268 observations). The graph of the series (visible in Figure 1) suggests that breaks should be taken into account, in particular due to the well-known great moderation phenomenon.

Table 4 (see also Table 2 of the SA) provides overwhelming evidence (with increases by at least 23 points of the MLL) in favour of the IHMS models compared to the ARMA model with fixed parameters. Moreover, the IHMS models with prior hyper-parameters implying weak regime persistence (MS-prior) slightly improve the fit over the models implying high regime persistence (CP-prior), whatever the dynamic specification (AR or ARMA). The differences of the MLL are equal to 1.53 and 1.23 respectively. Assuming prior odds equal to unity, these values imply posterior probabilities of the models with short regime persistence amounting to 0.82 and 0.77 compared to their CP alternatives. There is also similar evidence that, in the IHMS class, the ARMA specification dominates the corresponding AR one (with average improvements of 1.94 and 1.64 leading to posterior probabilities of

Table 4: U.S. GDP: marginal log-likelihood values

	ARMA(1,1)	IHMS-AR(1)		IHMS-ARMA(1,1)	
		CP	MS	CP	MS
Average	-378.57	-355.07	-353.54	-353.13	-351.90
Min.	-378.73	-355.14	-353.76	-353.36	-352.50
Max.	-378.53	-355.01	-353.37	-352.97	-351.45

CP and MS refer to the IHMS hyper-parameters: CP (MS) imply high (weak) regime persistence. Average, minimum and maximum values are computed from ten different estimations.

0.87 and 0.84 in favour of the ARMA function). Whatever the model, the MLL values of the ten replications are very similar.

Focusing on the IHMS-ARMA model for the two kinds of prior, Table 5 reports the posterior probabilities of the number of regimes for the mean function parameters and the variance (see also Table 3 in the SA). As expected, the uncertainty on the number of regimes is much more important for the MS-prior than for the CP-prior. With both types of prior, there is no evidence of breaks in the mean function parameters. The variance mainly evolves over time through two different states for the CP-prior while more regimes are found for the MS one.

Figure 1 displays the series together with the probabilities of a break in the previous or in the next year computed from the posterior samples. The probability of having a break in the mean function parameters is null for the CP-prior. For the MS-prior, there are small positive break probabilities in the beginning of the series, in 1971, at the beginning of the great moderation era and during the financial crisis. Regarding the breaks in the variance, the MS-prior leads to much more instabilities than the CP one. However, both priors agree on a quiet period starting with the great moderation and ending at the financial crisis.

The corresponding estimated time-varying parameters are displayed on Figure 2, where instead of showing the graph for the constant term (μ_t) of the ARMA equation, we show the implied 'unconditional' expectation $\mu_t/(1-\beta_t)$. Overall, both types of prior deliver similar results. The mean function parameters are relatively constant over time, especially

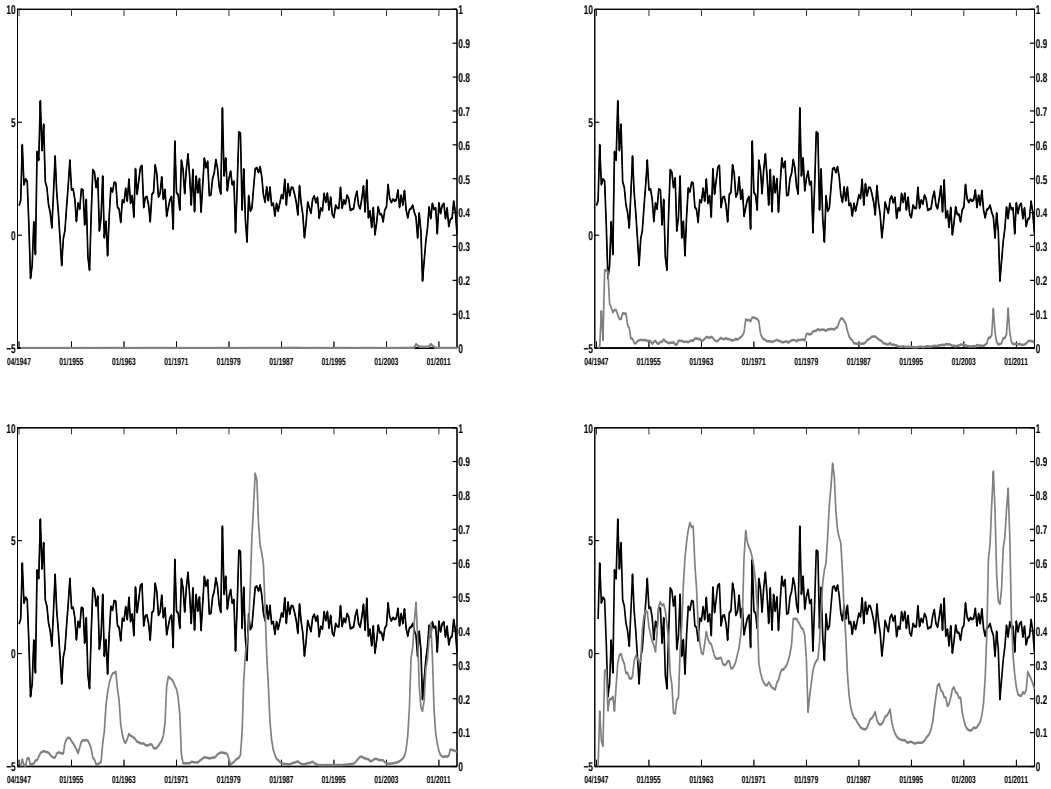
Table 5: U.S. GDP: posterior probabilities of having a specific number of regimes

IHMS-ARMA with MS prior									
# Regimes	1	2	3	4	5	6	7	8	9
μ, β, ϕ	**0.62**	0.20	0.06	0.07	0.03	0.01	0	0	0
σ^2	0	0.06	0.19	**0.26**	0.23	0.15	0.07	0.03	0.01
IHMS-ARMA with CP prior									
# Regimes	1	2	3	4	5	6	7	8	9
μ, β, ϕ	**0.99**	0.01	0.00	0	0	0	0	0	0
σ^2	0	**0.80**	0.16	0.03	0.01	0	0	0	0

with the CP-prior. The variance evolves with changes that are smooth or sharp depending on when they occur. As expected due to the differences between the priors, the variance dynamics obtained with the CP-prior is close to a change-point model, as the level is more or less constant before and after the great moderation, with a blip during the financial crisis. On each figure the corresponding posterior median of the ARMA model with fixed parameters is also shown. Obviously, the variance of this model cannot accommodate both the high and low volatility levels of the error terms and therefore its posterior estimate lies in between. For the mean function parameters, there are small differences between the posterior estimates of the simple ARMA and IHMS-ARMA models, even if there is no break in the mean function parameters of the latter. The occurrence of breaks in the variance and the fact that the variance and the parameters of the mean functions are not independent a posteriori is a reason for such differences.

We compare the out-of-sample forecast performances of the five models appearing in Table 4, to which we add three models: an AR(16) model with fixed parameters, this lag order resulting from the AIC criterion, and the restricted IHMS-ARMA models in which the mean parameters jointly move with the variance one. The forecasts start in the first quarter of 1987 (at 60% of the sample) and are computed until the end of the sample, adding one new observation at a time. At each step, all models are re-estimated and predictive densities as well as the CRPS are computed for different horizons. The mean

Figure 1: U.S. GDP series and posterior probabilities of having a break in the previous year or in the next year



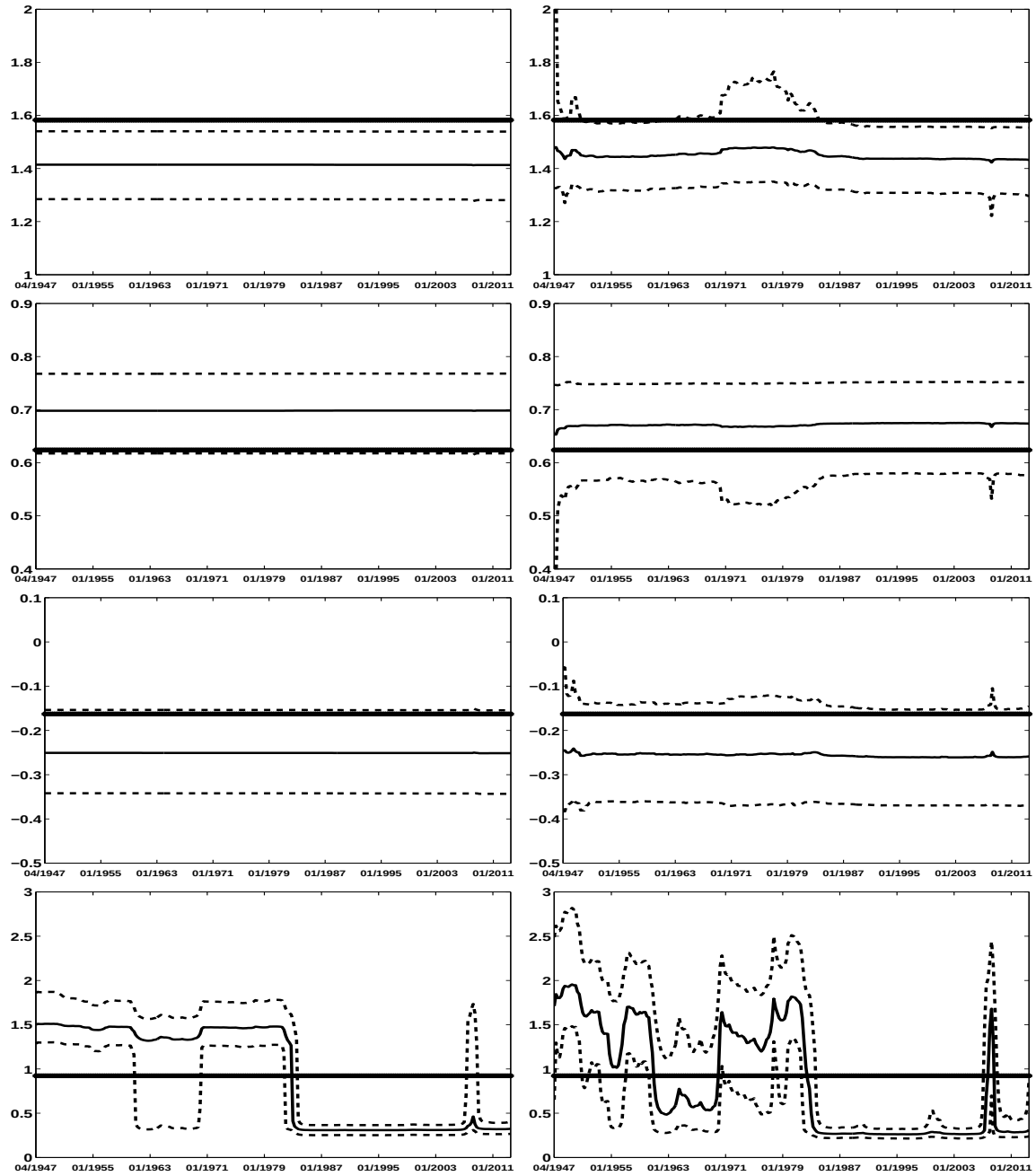
The left column corresponds to the CP-prior, the right one to the MS-prior. Break probabilities (right vertical axis) over the past or the next year are in grey. First row: break probabilities of the mean function parameters. Last row :break probabilities of the variance parameter.

squared forecast errors (MSFE) using the predictive means, the average values (over the forecast period) of the predictive densities (APD), and the CRPS values evaluated at the observed data are reported in Table 6. The main conclusions from these results follow.

- For each criterion and forecast horizon, all the IHMS models values dominate substantially the ARMA model with fixed parameters. The AR(16) model performs better than this ARMA model but is still performing less well than the IHMS models (with very few minor exceptions).

- Regarding the statistical test (36), for the APD and for the CRPS, the flexible IHMS

Figure 2: U.S. GDP: posterior medians and 70% credible intervals



Left column: CP-prior. Right: MS-prior. Thick horizontal line: posterior median of the ARMA model with fixed parameters. Thin continuous and dotted lines: IHMS-ARMA posterior median and the limits of the 70% posterior credible interval. Top row: long term mean $\mu_t/(1-\beta_t)$. Second row: AR coefficient. Third row: MA coefficient. Last row: variance.

models statistically outperform the ARMA model at all horizons (except for APD in the case of the four year horizon of the IHMS-ARMA model with MS prior), while for MSFE, most tests are significant only for forecast horizons up to two years. Table 4 in the SA contains the test results for comparing the flexible IHMS models with respect to the AR(16) model and shows not surprisingly in view of the previous comment, that there are of course less cases of significant differences than in the comparisons with the ARMA model.

- Concentrating on the unrestricted IHMS models, the ARMA versions slightly dominate their AR counterparts, and similarly the models with the CP prior dominate those with the MS prior.

- Incorporating an MA term is more relevant than including a flexible dynamic structure for the breaks since the restricted IHMS-ARMA model with CP prior setting is often the best competitor.

Going one step further, the integrand of Equation (35) is helpful to assess when a model produces better forecasts than another. For example, a model could very well forecast the growth rate when the U.S. economy is in expansion and could be a bad predictor in recessions. To evaluate this, Figure 3 displays the integrand of Equation (35) with respect to z for one-step ahead predictions. We observe that the integrand of the ARMA model envelops the integrand of all the IHMS models meaning that whatever the state of the U.S. economy, the distance between the cumulative density functions of the IHMS models and the empirical one are always smaller than for the ARMA model.

Finally, Figure 4 shows the differences between the one-quarter ahead predictive densities (without taking logarithms) of the IHMS-ARMA (CP-prior) model and the fixed parameter ARMA one, both evaluated at the realized outcome. The gains are spread over the entire period and cannot therefore be associated to a specific sub-period.

6.2 U.S. Inflation

The inflation measure is computed from the personal consumption expenditure deflator and spans the period from February 1959 to November 2012 (646 observations). Table 7

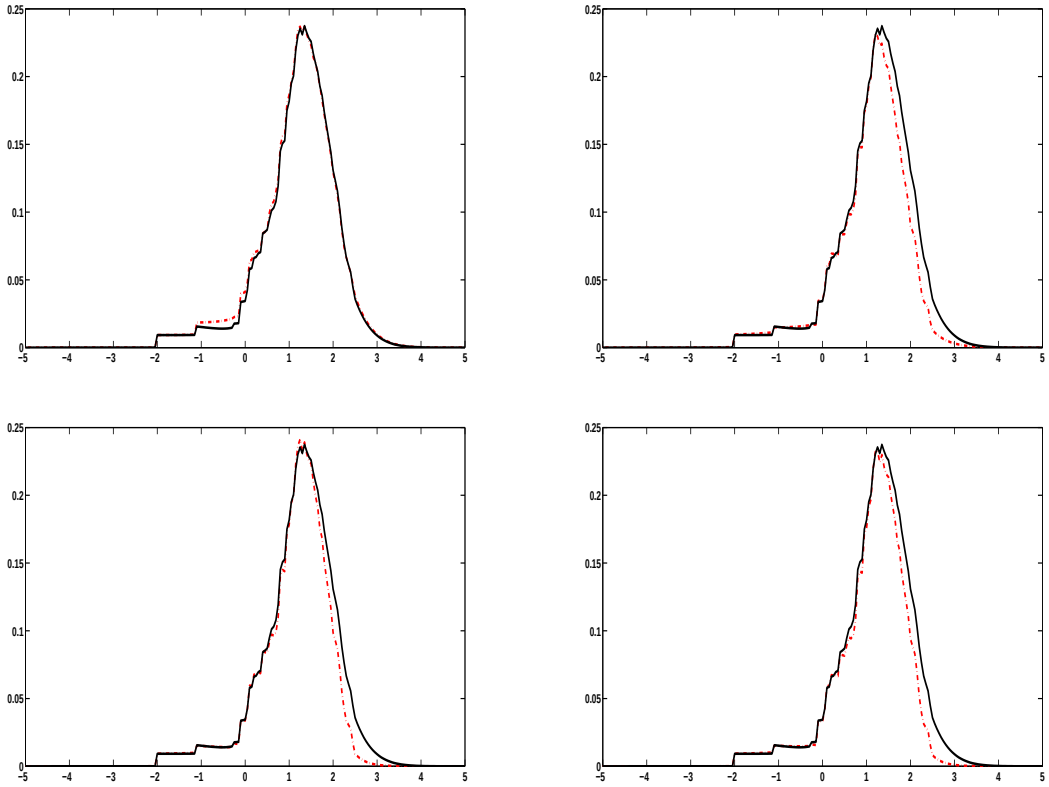
Table 6: U.S. GDP Growth Rate: APD, MSFE and CRPS

Forecast Horizons	One quarter	Two quarters	One year	Two years	Three years	Four years
			APD			
AR(16)	0.36	0.33	0.32	0.31	0.31	0.31
ARMA	0.33	0.30	0.28	0.27	0.27	0.27
Rest. IHMS-ARMA (CP)	0.46	0.44	0.40	0.39	0.38	**0.39**
Rest. IHMS-ARMA (MS)	0.45	0.42	0.38	0.35	0.34	0.34
IHMS-AR (CP)	0.46**	0.43**	0.39**	0.38**	0.37*	0.38*
IHMS-AR (MS)	0.43**	0.39**	0.36**	0.34**	0.33*	0.33*
IHMS-ARMA (CP)	**0.47****	**0.44****	**0.40****	**0.39****	**0.38***	0.38*
IHMS-ARMA (MS)	0.44**	0.41**	0.37**	0.34**	0.33**	0.33*
			MSFE			
AR(16)	0.43	0.52	0.58	0.60	0.61	0.58
ARMA	0.40	0.53	0.68	0.78	0.82	0.82
Rest. IHMS-ARMA (CP)	**0.36****	**0.43****	**0.52****	**0.57***	**0.61***	**0.57**
Rest. IHMS-ARMA (MS)	0.37**	0.45**	0.55**	0.64*	0.69	0.69
IHMS-AR (CP)	0.39	0.49**	0.60**	0.67*	0.70	0.68
IHMS-AR (MS)	0.40	0.52	0.64*	0.70*	0.73	0.73
IHMS-ARMA (CP)	0.37**	0.45**	0.56**	0.65*	0.69	0.68
IHMS-ARMA (MS)	0.38**	0.47**	0.60**	0.69*	0.73	0.73
			CRPS			
AR(16)	0.37	0.40	0.42	0.43	0.44	0.43
ARMA	0.37	0.42	0.47	0.50	0.51	0.51
Rest. IHMS-ARMA (CP)	0.33**	0.36**	**0.39****	**0.42****	**0.42***	**0.42**
Rest. IHMS-ARMA (MS)	0.33**	0.37**	0.41**	0.44**	0.45*	0.45
IHMS-AR (CP)	0.34**	0.38**	0.42**	0.44**	0.45**	0.44*
IHMS-AR (MS)	0.35**	0.39**	0.43**	0.46**	0.47**	0.47*
IHMS-ARMA (CP)	**0.33****	**0.36****	0.41**	0.44**	0.45**	0.45*
IHMS-ARMA (MS)	0.34**	0.37**	0.42**	0.46**	0.47**	0.47

APD: average of predictive densities. MSFE: mean squared forecast error. CRPS: continuous ranked probability score. CP and MS refer to the prior IHMS hyper-parameters. Forecasts are from 1987Q1 to 2014Q1. Bold numbers identify the best performing model. A star indicates that there exists a significant statistical difference with respect to the ARMA model at the 10% level (two-sided test). A double star denotes significance at 5%.

reports the MLL values of several models including the ARMA model with fixed parameters. Additionally to these results, Table 5 of the SA includes the results of the random walk process and the AR(12) chosen by the AIC but these models are dominated by the ARMA process with fixed parameters. Like for the U.S. GDP growth rate, there is a strong evidence in favour of the IHMS models with respect to the simple ARMA one (differences of MLL larger than 62). Moreover, the models with the IHMS CP-type prior implying weak regime persistence have the highest MLL (differences equal to 2.96 and 2.24 leading to posterior probabilities of 0.95 and 0.9). The inclusion of MA terms in the IHMS models increases slightly the average MLL (1.73 and 1.01) but this conclusion should be tempered (especially

Figure 3: U.S. GDP growth rate: Integrands of Equation (35) for several models

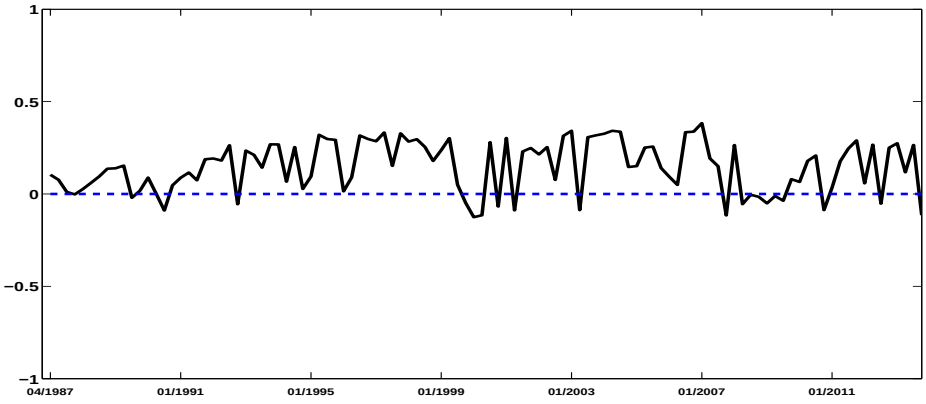


Horizontal axis: value of z in Equation (35). Vertical axis: value of the integrand in Equation (35) for one-step ahead forecasts. In clockwise order starting from the upper-left plot, the dashed line corresponds to the AR model, the restricted IHMS-ARMA (CP) one, the IHMS-ARMA (CP) model and the IHMS-AR (CP) one, and the continuous line to the ARMA model.

for the model with MS-type prior) given the overlap of the min-max ranges.

Table 8 reports the posterior probabilities of the number of regimes of the restrictedIHMS-ARMA models (see also Table 6 of the SA which includes the IHMS-ARMA models based on a unique HDP). The mean function parameters and the variance clearly experience structural breaks. Furthermore, the uncertainty on the number of regimes is visible, especially if the MS-prior is used. This highlights the limitation of the standard method consisting in picking up the model with a fixed number of regimes exhibiting the highest marginal likelihood. With the CP-prior, two regimes seem sufficient for each set of parameters, while

Figure 4: U.S. GDP: difference between one-quarter ahead predictive densities of the IHMS-ARMA (CP-prior) model and the fixed parameter ARMA model



three to five seem useful with the MS-prior.

The estimated break probabilities over the past and the next year are displayed on Figure 5. First of all, the IHMS model with a priori long-lasting regimes (CP-prior) only detects one break for the mean function parameters, which happens in the early 2000's. On the contrary, many instabilities on the mean function parameters are visible for the IHMS model with a priori weakly persistent regimes. Some of these changes correspond to known historical episodes. For example, the breaks detected around 1973 and 1979 capture the oil crisis era and the change of the monetary policy of the Fed, both marked by a rise of U.S. inflation (see the top right graph). The break dynamics of the variance is more volatile in both configurations. Even if less switches are detected by the IHMS model with a priori high persistence, the two graphs do not show very different results. Two quiet periods spanning from 1967 to 1973 and from 1991 to 1998 do not exhibit breaks.

Figure 6 documents the time-varying posterior medians of the model parameters along with the 70% credible intervals. Interestingly, the mean function parameters and the variance do not exhibit the same dynamics, emphasizing the relevance of disentangling the two sets of parameters. Obviously, these variations cannot be accommodated by the parameters of the standard ARMA model, as highlighted by its posterior medians (thick lines). The model with a priori high persistence (CP-prior) exhibits a change-point behaviour for

Table 7: U.S. Inflation: marginal log-likelihood values

	ARMA(1,1)	IHMS-AR(1)		IHMS-ARMA(1,1)	
		CP	MS	CP	MS
Average	-1398.11	-1335.81	-1332.85	-1334.08	-1331.84
Min.	-1398.25	-1336.31	-1333.36	-1334.77	-1332.51
Max.	-1397.92	-1335.38	-1331.66	-1332.83	-1331.53

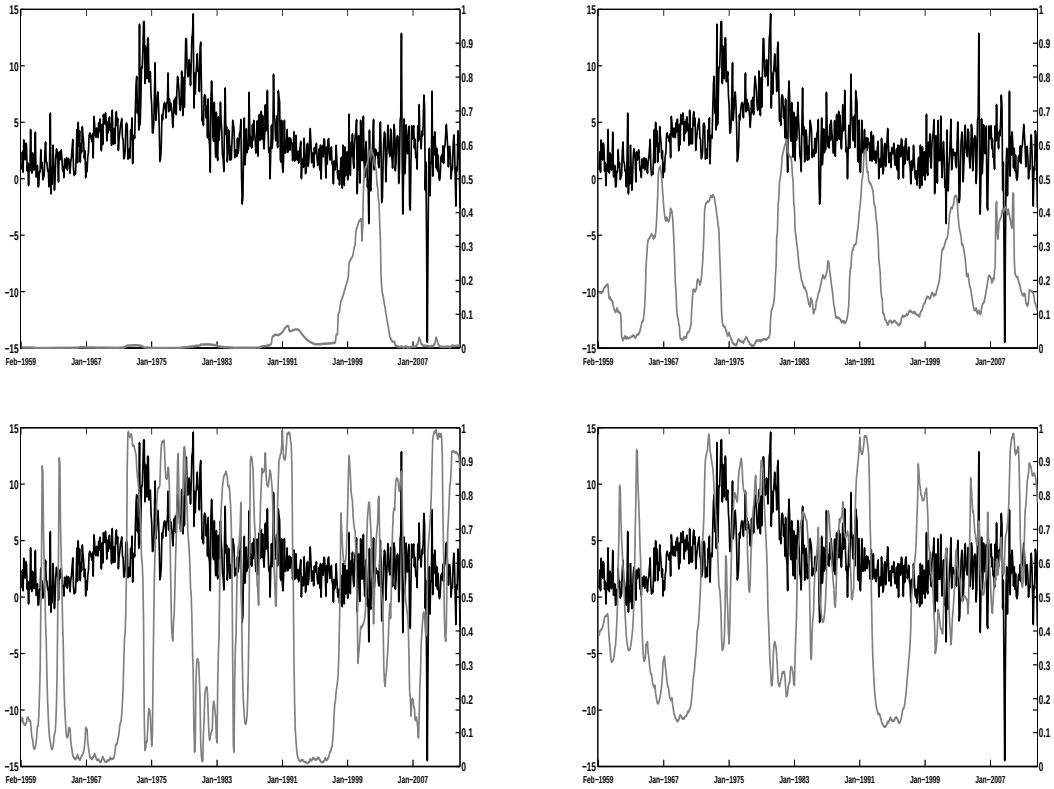
CP and MS refer to the IHMS hyper-parameters: CP (MS) imply high (weak) regime persistence). Average, minimum and maximum are computed from ten different estimations.

Table 8: U.S. Inflation: posterior probabilities of having a specific number of regimes

IHMS-ARMA with MS prior									
# Regimes	1	2	3	4	5	6	7	8	9
μ, β, ϕ	0	0.08	**0.28**	0.27	0.20	0.13	0.03	0.01	0.00
σ^2	0	0.11	0.22	**0.26**	0.20	0.13	0.06	0.02	0.00
IHMS-ARMA with CP prior									
# Regimes	1	2	3	4	5	6	7	8	9
μ, β, ϕ	0	**0.70**	0.29	0.01	0	0	0	0	0
σ^2	0	**0.87**	0.12	0.01	0.00	0.00	0	0	0

the mean function parameters, while the variance recurrently switches between two states. The estimated mean function is close to a unit root process until the early 2000's (with the AR parameter slightly below 1); then until the end of the sample, it corresponds to a weakly persistent ARMA process (with the AR parameter close to 0.2). The ARMA model with fixed parameters obviously cannot capture the break. Its estimated AR coefficient is also close to 1 over the entire sample (but slightly less than in the IHMS model during the first regime of the latter). The MA coefficient of the IHMS model is clearly different from zero in both regimes, with a strong negative value (around -0.75) during the first regime, and a positive value (about 0.3) in the second regime. The estimated MA parameter of the

Figure 5: U.S. Inflation series and posterior probabilities of having a break in the last year or in the next year.



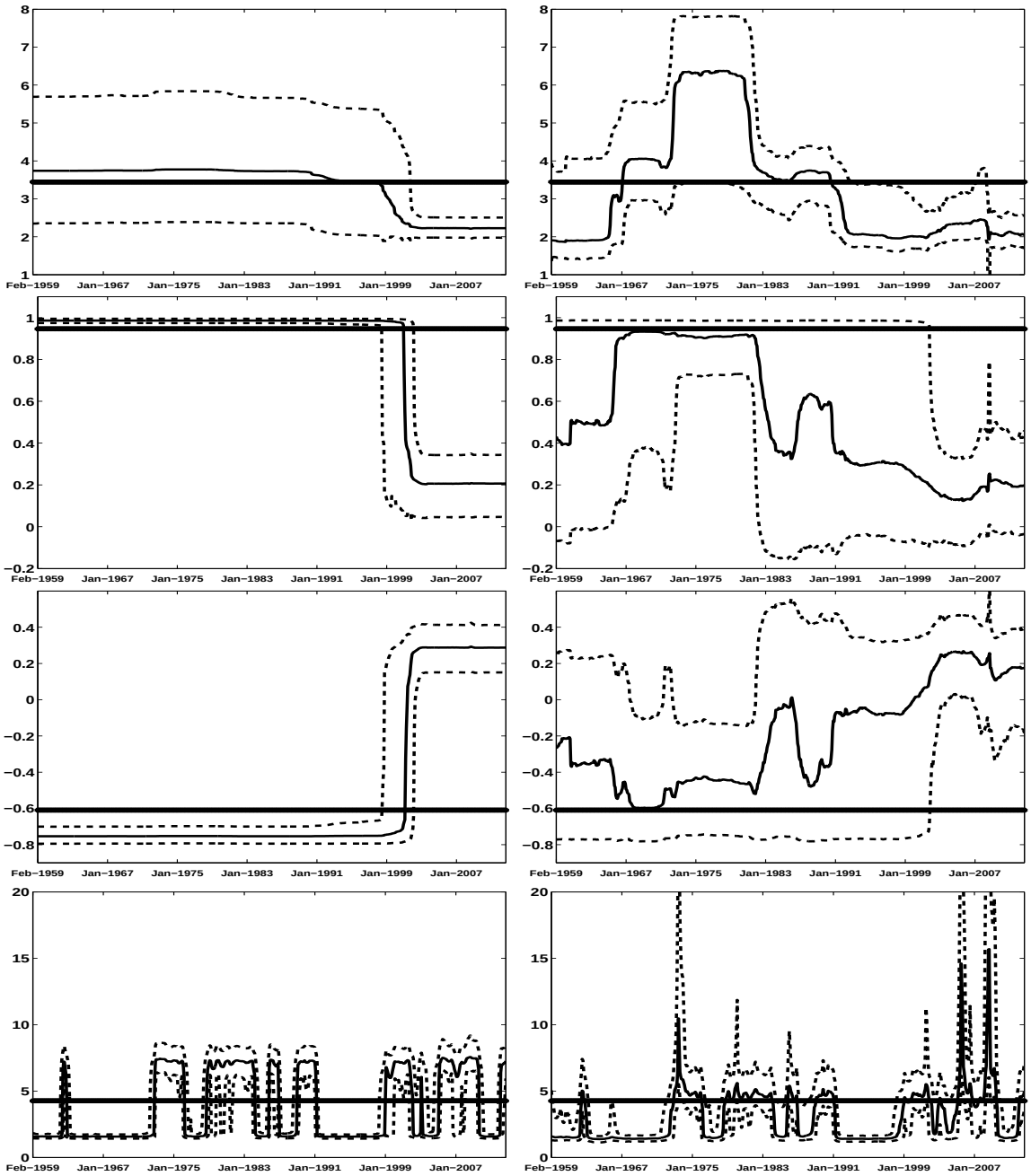
The left column corresponds to the CP-prior, the right one to the MS-prior. Break probabilities (right vertical axis) over the past or the next year are in grey. First row: break probabilities of the mean function parameters. Last row: break probabilities of the variance parameter.

simple ARMA model (at 0.6) is dominated by the first regime.[2]

In the IHMS-ARMA model with a priori weak persistence (MS-prior), the AR and the MA coefficients have broadly a similar time series evolution as for the CP-prior, but they are much more instable. For instance, the AR coefficient is about equal to 0.4 at the start of the sample, then it increases to about 0.95 between 1967 and 1983 (covering in particular the stagflation period of the 70's), then drops to its initial level, and after a short rebound,

[2]The results for the first regime of the IHMS-ARMA model with CP-prior are close to those of Stock & Watson (2007). These authors find that the *first difference* of (quarterly) inflation is well captured by a MA(1) model (with heteroskedasticity) on the period 1960-1983, with estimated MA coefficient of -0.25.

Figure 6: U.S. Inflation: posterior medians and 70% credible intervals



Left column: CP-prior. Right: MS-prior. Thick horizontal line: posterior median of the ARMA model with fixed parameters. Thin continuous and dotted lines: IHMS-ARMA posterior median and the limits of the 70% posterior credible interval. Top row: long term mean $\mu_t/(1-\beta_t)$. Second row: AR coefficient. Third row: MA coefficient. Last row: variance.

gradually drops to 0.2 (as in the CP-prior graph). Not surprisingly, the 70% credible intervals are wider when the MS-prior is used (since the MS-prior is less informative). The MA parameter is time-varying and negative (around -0.4) until 1982, then increases until 2003 when it is around 0.2. Its 70% credible interval does not cover the value zero only after 2003.

Finally, as for the U.S. GDP growth rate, we report the results of forecast comparisons over the last 60% of the sample, starting the forecasts in April 1991. In addition to the models considered in Table 7, the set of models includes a random walk (RW) process that is sometimes considered to be relevant for this series, the AR(12) model, since this lag order results from the AIC, and the restricted IHMS-ARMA models. Table 9 provides the APD, MSFE and CRPS values for six forecast horizons and inspires the following comments.

- For each criterion and forecast horizon, the APD of the all the IHMS models values are higher than those of the ARMA and AR(12) models with fixed parameters. For the MSFE, the IHMS models are also performing better, with very few exceptions for some of them at the horizons of one year and sixteen months. Concerning the CRPS, their values are smallest for the IHMS models. On all criteria, the RW model is clearly performing much less well than the other models, except for the APD at forecast horizon of one month where its performance is less inferior than at the other horizons. The RW model is particularly badly performing for the forecasts at large horizons.

- The different IHMS models have about the same performance according to the APD criterion. Judging by the MSFE, the best forecasting models have an AR mean function at four horizons and an ARMA one at the other two horizons. For the CRPS, the models with the AR mean function dominate at five horizons out of six.

- Regarding the statistical test (36), for the APD, the four flexible IHMS models statistically outperform the ARMA process at all horizons. For the MSFE, there are less significant tests, but the IHMS-ARMA model with the MS prior outperforms the ARMA model at all horizons except at one month. For the CRPS, the unrestricted IHMS models statistically dominate the ARMA models except in three cases out of

twenty-four. Table 7 in the SA contains the test results for comparing the flexible IHMS models with respect to the AR(12) model and shows that there are less cases of significant differences than in the comparisons with the ARMA model. The AR(12) model has ten parameters more than the standard ARMA model. This contributes to increase the denominator of the test statistic (36) and renders them less significant.

Table 9: U.S. Inflation: APD, MSFE and CRPS

Forecast Horizons	One month	Two months	Four months	Eight months	One year	Sixteen months
			APD			
RW	0.13	0.10	0.08	0.06	0.05	0.04
AR(12)	0.14	0.14	0.14	0.13	0.12	0.12
ARMA	0.15	0.14	0.14	0.13	0.12	0.12
Rest. IHMS-ARMA (CP)	0.16**	0.15**	0.15*	0.14*	0.13	0.13
Rest. IHMS-ARMA (MS)	0.16**	0.15**	0.15**	0.14**	0.14**	0.13*
IHMS-AR (CP)	0.16**	0.15**	0.15**	**0.14****	0.14**	**0.14****
IHMS-AR (MS)	0.16**	0.15**	0.15**	0.14**	**0.14****	0.14**
IHMS-ARMA (CP)	0.16**	0.15**	0.15**	0.14**	0.13**	0.13**
IHMS-ARMA (MS)	**0.16****	**0.16****	**0.15****	0.14**	0.13**	0.13**
			MSFE			
RW	6.65	10.26	10.33	12.10	12.99	9.94
AR(12)	6.58	6.73	6.95	6.87	7.01	6.62
ARMA	5.52	6.60	6.52	6.93	7.24	6.53
Rest. IHMS-ARMA (CP)	5.25	6.19	**5.79**	6.29	6.94	6.56
Rest. IHMS-ARMA (MS)	5.28	6.32	5.98	6.53	7.20	6.82
IHMS-AR (CP)	5.00*	6.09*	5.84*	**6.15**	6.55*	6.26
IHMS-AR (MS)	**4.94****	**6.04***	5.85*	6.17	**6.51***	6.33
IHMS-ARMA (CP)	5.19	6.33	6.08*	6.56	7.13	6.62
IHMS-ARMA (MS)	5.26	6.32**	5.98*	6.35*	6.78**	**6.12***
			CRPS			
RW	1.37	1.66	1.74	2.11	2.35	2.46
AR(12)	1.29	1.25	1.28	1.31	1.29	1.30
ARMA	1.19	1.28	1.24	1.34	1.37	1.33
Rest. IHMS-ARMA (CP)	1.16	1.22	**1.18**	1.25	1.32	1.30
Rest. IHMS-ARMA (MS)	1.16	1.24	1.20	1.28	1.34	1.33
IHMS-AR (CP)	1.13**	1.22*	1.19	1.25*	1.29**	**1.27**
IHMS-AR (MS)	**1.12****	**1.21***	1.18*	**1.24***	**1.28****	1.28
IHMS-ARMA (CP)	1.15*	1.24**	1.20**	1.28**	1.33	1.30
IHMS-ARMA (MS)	1.14**	1.22**	1.18**	1.27**	1.31**	1.27*

APD: average of predictive densities. MSFE: mean squared forecast error. CRPS: continuous ranked probability score. CP and MS refer to the prior IHMS hyper-parameters. Forecasts are from April, 1991 to November 2012. Bold numbers identify the best performing model. A star indicates that there exists a significant statistical difference with respect to the ARMA model at the 10% level (two-sided test). A double star denotes significance at 5%.

6.3 Comparison of predictive performance for other series

We extend our study of the IHMS-ARMA model by applying to other U.S. series the same type of comparison of forecast performance as in the previous subsections. Table 10 lists eighteen quarterly macroeconomic series (from 1959Q1 to 2011Q3) also used in Bauwens et al. (2015). All series are transformed in logarithm and differenced once, except series 9. The first series is the same as in subsection 6.1 but on a shorter period. The fourth one is the quarterly version of the inflation series analysed in subsection 6.2. For each of these series, the forecast implementation is the same as for the quarterly GDP growth series of subsection 6.1, except that the forecast period starts in 1990Q3.

For each series, Table 11 reports the model that delivers the best APD, MSFE and CRPS values. It also gives the percentage of improvement (or deterioration) with respect to the fixed parameter ARMA model. The following conclusions emerge.

- Overall, the IHMS-ARMA model appears much more often than the three other ones.
- For the APD criterion, the IHMS-ARMA process is the best one in more than 56% of the cases for each horizon.[3]
- For the MSFE criterion, the performance of the simple ARMA model and the restricted IHMS-ARMA one gets better but these competitors are still dominated by the flexible IHMS processes in more than 60% of the cases at horizon one. For farther horizons, the restricted IHMS-ARMA model becomes as good as the most flexible ones. This result can be explained by the new regimes created during the forecast of the series. In particular, the parameters of the new regimes are sampled from the hierarchical distributions and can therefore generate bad forecasts.
- Focusing on the CRPS, the IHMS-ARMA models dominate the other models in more than 55% of the cases at all horizons.
- Considering the degree of improvement, the IHMS model drastically enhances the APD and the MSFE for several series. For instance, the relative APD of GDP (series

[3]The results for the forecasts at horizons of three and four years are not reported, as they are very close to those shown in Table 11.

Table 10: Macroeconomic series for predictive performance comparison

Number	Name
1	Real Gross Domestic Product
2	Personal Income
3	Real Personal Consumption Expenditures
4	Personal Consumption Expenditures: Chain-type Price Index
5	Real Gross Private Domestic Investment
6	Business Sector: Output Per Hour of All Persons
7	Real Imports of Goods and Services
8	Real Exports of Goods and Services
9	Real Change in Private Inventories
10	Real Government Consumption Expenditures and Gross Inv.
11	Compensation of Employees: Wages and Salary Accruals
12	Net Corporate Dividends
13	Personal Saving
14	Real Disposable Personal Income
15	Gross Domestic Product: Implicit Price Deflator
16	Nonfarm Business Sector: Unit Labor Cost
17	Private Residential Fixed Investment
18	Gross Saving

1), Real Imports and Exports of Goods and Services (series 7 and 8) and Private Residential Fixed Investment (series 17) are between 32% and 62% for all horizons. Similar conclusions hold for series 17 when we look at the MSFE. Finally, considering the CRPS, the improvements are also high for the same series as in the APD. On the contrary, when the ARMA model dominates the time-varying ones, the improvements are rather modest. They stay below 6% for all the series, whatever the criterion.

Table 11: Predictive performance of IHMS-ARMA with respect to ARMA and Restricted IHMS-ARMA.

Forecast Horizons	One quarter	Two quarters	One year	Two years
Series		**APD**		
1	36.31 (CP)	33.78 (CP)	32.16 (CP)	33.05 (CP)
2	5.36 (R. CP)	1.77 (R. CP)	2.05 (MS)	2.57 (R. CP)
3	11.64 (CP)	10.53 (CP)	7.79 (CP)	7.63 (CP)
4	-0.44 (ARMA)	3.18 (MS)	3.91 (MS)	3.56 (MS)
5	20.54 (CP)	17.50 (CP)	17.35 (CP)	16.24 (CP)
6	9.29 (CP)	9.32 (CP)	10.16 (CP)	8.41 (CP)
7	58.84 (CP)	49.45 (CP)	42.09 (CP)	43.64 (CP)
8	62.45 (CP)	55.08 (CP)	54.89 (CP)	56.87 (CP)
9	-2.43 (ARMA)	-2.35 (ARMA)	-2.08 (ARMA)	-1.01 (ARMA)
10	7.89 (MS)	6.18 (MS)	5.61 (CP)	5.61 (CP)
11	4.09 (R. CP)	2.87 (R. CP)	2.74 (MS)	-0.54 (ARMA)
12	19.96 (R. CP)	2.03 (R. CP)	-5.82 (ARMA)	-4.15 (ARMA)
13	1.69 (R. CP)	-0.11 (ARMA)	-1.94 (ARMA)	1.61 (R. CP)
14	12.54 (R. CP)	6.56 (MS)	5.39 (MS)	5.77 (MS)
15	1.23 (MS)	1.15 (MS)	1.49 (MS)	1.25 (CP)
16	13.02 (R. MS)	14.58 (R. MS)	13.09 (R. MS)	10.49 (R. MS)
17	46.16 (CP)	40.94 (CP)	37.99 (CP)	35.09 (CP)
18	2.54 (MS)	2.94 (MS)	1.33 (R. CP)	0.52 (R. MS)
Perc. IHMS-ARMA	56 %	67 %	72 %	61 %
Series		**MSFE**		
1	-6.56 (CP)	-7.23 (R. CP)	-2.32 (R. CP)	-2.62 (R. CP)
2	0.20 (ARMA)	-2.81 (R. CP)	2.17 (ARMA)	2.95 (ARMA)
3	-1.89 (CP)	-3.90 (CP)	-0.06 (CP)	-0.42 (MS)
4	-3.46 (MS)	-5.86 (MS)	-5.91 (MS)	-4.35 (CP)
5	-7.80 (MS)	-0.98 (R. CP)	4.84 (ARMA)	-0.51 (MS)
6	-1.47 (R. CP)	-0.39 (R. CP)	1.50 (ARMA)	-1.27 (R. CP)
7	-30.64 (R. MS)	-5.79 (R. CP)	3.67 (ARMA)	0.40 (ARMA)
8	-37.30 (R. MS)	5.98 (ARMA)	2.10 (ARMA)	-0.06 (R. CP)
9	-2.16 (R. CP)	-0.66 (R. CP)	4.75 (ARMA)	1.55 (ARMA)
10	-3.07 (MS)	-5.22 (MS)	-1.29 (CP)	-1.69 (R. MS)
11	-1.00 (MS)	-4.14 (MS)	-6.65 (MS)	-3.25 (MS)
12	-9.97 (R. MS)	-3.74 (MS)	4.65 (ARMA)	-0.64 (MS)
13	-8.54 (MS)	-13.64 (CP)	-7.47 (CP)	-4.13 (R. CP)
14	-2.49 (MS)	-2.04 (R. CP)	-1.31 (R. CP)	-2.48 (CP)
15	-1.42 (MS)	-1.82 (R. CP)	-3.41 (R. CP)	-2.43 (R. CP)
16	-2.95 (MS)	-7.09 (R. CP)	-5.66 (CP)	-5.68 (R. MS)
17	-14.71 (MS)	-16.37 (MS)	-15.25 (CP)	-4.36 (MS)
18	0.59 (ARMA)	-2.01 (MS)	2.18 (ARMA)	0.17 (ARMA)
Perc. IHMS-ARMA	61 %	39 %	39 %	39 %
Series		**CRPS**		
1	-7.51 (CP)	-5.42 (MS)	-2.51 (MS)	-3.37 (CP)
2	-0.17 (MS)	0.93 (ARMA)	-0.71 (MS)	0.55 (ARMA)
3	-1.49 (CP)	-1.22 (CP)	1.07 (ARMA)	-0.20 (R. CP)
4	-1.88 (MS)	-3.48 (MS)	-4.31 (MS)	-1.87 (MS)
5	-8.29 (MS)	-6.91 (R. CP)	-3.43 (R. CP)	-4.19 (MS)
6	-1.65 (R. CP)	-2.66 (R. MS)	-0.94 (R. MS)	-1.03 (R. MS)
7	-20.62 (R. MS)	-11.22 (CP)	-7.72 (CP)	-8.93 (CP)
8	-23.28 (R. MS)	-8.93 (R. CP)	-8.96 (R. MS)	-9.37 (MS)
9	-2.77 (R. CP)	-2.64 (R. CP)	-2.27 (MS)	-1.41 (MS)
10	-3.08 (MS)	-3.70 (MS)	-2.58 (MS)	-3.59 (MS)
11	-1.06 (CP)	-1.99 (MS)	-4.46 (MS)	-1.65 (MS)
12	-11.91 (MS)	-7.01 (MS)	-0.10 (MS)	-0.75 (MS)
13	-8.00 (MS)	-5.67 (MS)	-3.12 (R. CP)	-2.46 (MS)
14	-3.57 (R. CP)	-0.17 (R. CP)	-0.70 (MS)	-2.64 (CP)
15	-1.07 (R. CP)	-3.04 (MS)	-2.32 (MS)	-1.56 (R. MS)
16	-3.79 (MS)	-2.45 (R. MS)	-1.43 (R. MS)	-1.73 (R. MS)
17	-15.43 (MS)	-13.56 (MS)	-10.78 (CP)	-4.35 (R. CP)
18	-0.95 (R. CP)	-1.89 (R. CP)	-0.22 (R. CP)	-0.10 (CP)
Perc. IHMS-ARMA	61 %	56 %	61 %	67 %

The Table reports the best model according to each criterion and the percentage difference with respect to the ARMA model. 'Perc. IHMS-ARMA' is the number of times in percentage that the best model is the IHMS-ARMA one with the flexible dynamic in the break.MS (CP) means IHMS-ARMA with MS (CP) prior and R. MS (CP) stands for the restricted IHMS-ARMA with MS (CP) prior.

7 Conclusions

The Markov-switching modelling framework is a powerful tool to capture occasional changes in the parameter values of dynamic econometric models at a priori unknown dates. Such models may suffer from a potential over-parameterization issue due to the assumption that all the parameters must change when a break occurs. We propose a solution to this problem by relying on the hierarchical Dirichlet process. The break dynamics of the mean function parameters is separated from the one of the variance and thanks to the a priori infinite number of states, only one estimation is sufficient to determine the number of regimes in the two parameter structures. Consequently, the proposed IHMS framework extends the MS class in two related directions as it both allows for an unbounded number of regimes and a flexible dynamics for the model parameters. In addition to that, this modeling approach is operational for complex models as we solve the path dependence problem due to the moving average component of ARMA models. This is achieved by using a Metropolis-Hastings step with a proposal density based on an approximate model inspired by the solution that Klaassen (2002) proposed for the MS-GARCH model.

Empirical applications on the quarterly U.S. GDP growth rate and on the monthly U.S. inflation rate highlight the relevance of allowing for the possibility of different structural breaks in the mean and in the variance. In particular, for the U.S. GDP growth rate, we find a structural break in the variance at the beginning of the great moderation era, but no simultaneous break in the mean function parameters. The latter parameters are therefore estimated from the entire sample. The inference on the U.S. inflation delivers similar results as the breaks of the mean function parameters are different from those of the variance. This example additionally highlights that assuming only two regimes for a time series is not always satisfactory as the number of breaks in the mean and the variance are larger than two when the prior information favours weakly persistent regimes. A forecasting comparison on eighteen quarterly series illustrates that the IHMS-ARMA models perform better than the ARMA one with fixed parameters for a majority of series, in some case by a wide margin.

Future research could be devoted to relaxing the geometric duration of the regimes implied by the Markov chains. We could also investigate if allowing a different break

structure for each model parameter further improves the predictions. Another avenue of research could be to extend the approach to VARMA and factor models.

Appendix 1: Sticky IHMS-ARMA Gibbs Sampler

Before detailing the implementation of the sampler, we introduce some notations. Sums over one indices are denoted by dots: for instance $\sum_a x_{a,b} = x_{.,b}$ and $\sum_a \sum_b x_{a,b} = x_{.,.}$. The vector $\{x_1, x_2, ..., x_r\}$ is denoted by $x_{1:r}$ and is a row vector, while its transpose is denoted by $x'_{1:r}$. The symbol L stands for the number of regimes. If a positive random variable X follows a Gamma distribution with positive parameters k (shape) and θ (scale), we write that $X \sim G(k, \theta)$, and the corresponding density function is written

$$f(x|k, \theta) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{\frac{-x}{\theta}}.$$

The vector $\Theta = \{\mu_1, \ldots, \mu_L, \beta_1, \ldots, \beta_L, \phi_1, \ldots, \phi_L, \sigma_1, \ldots, \sigma_L\}'$ contains all the ARMA parameters of the IHMS-ARMA model of subsection 2.3 of the paper, and $\theta_j = \{\mu_j, \beta_j, \phi_j, \sigma_j\}'$ the parameters of regime j.

One iteration of the Gibbs sampler algorithm sketched in Table 1 passes through between the following steps:

1. Sampling $s^{\psi}_{1:T}$ from $f(s^{\psi}_{1:T}|\Theta, P^{\psi}, s^{\sigma}_{1:T}, y_{1:T})$: see sub-section 3.1.

2. Sampling P^{ψ} from $f(P^{\psi}|H_{Dir}, \pi^{\psi}, s^{\psi}_{1:T}, y_{1:T})$: for $j = 1, ..., L$, sample $p^{\psi}_{j,1:L} \sim \text{Dir}(\alpha_{\psi}\pi^{\psi}_1 + n^{\psi}_{j,1}, ..., \alpha_{\psi}\pi^{\psi}_j + \kappa_{\psi} + n^{\psi}_{j,j}, ..., \alpha_{\psi}\pi^{\psi}_L + n^{\psi}_{j,L})$, where $n^{\psi}_{j,k}$ denotes the number of transitions from state j to k observed in the state vector $s^{\psi}_{1:T}$.

3. Sampling $s^{\sigma}_{1:T}$ from $f(s^{\sigma}_{1:T}|\Theta, P^{\sigma}, s^{\psi}_{1:T}, y_{1:T})$ by the forward-backward algorithm (see Chib (1996)).

4. Sampling P^{σ} from $f(P^{\sigma}|H_{Dir}, \pi^{\sigma}, s^{\sigma}_{1:T}, y_{1:T})$: for $j = 1, ..., L$, sample $p^{\sigma}_{j,1:L} \sim \text{Dir}(\alpha_{\sigma}\pi^{\sigma}_1 + n^{\sigma}_{j,1}, ..., \alpha_{\sigma}\pi^{\sigma}_j + \kappa_{\sigma} + n^{\sigma}_{j,j}, ..., \alpha_{\sigma}\pi^{\sigma}_L + n^{\sigma}_{j,L})$, where $n^{\sigma}_{j,k}$ denotes the number of transitions from state j to k observed in the state vector $s^{\sigma}_{1:T}$.

5. Sampling $\{\alpha_{\psi}, \kappa_{\psi}, \eta_{\psi}\}$ from $f(\alpha_{\psi}, \kappa_{\psi}, \eta_{\psi}|P^{\psi}, \pi^{\psi}, s^{\psi}_{1:T}, y_{1:T})$:

(a) Introduce auxiliary variables:

- Sampling m: For $j = 1, ..., L$, and $k = 1, ..., L$. Set $m_{j,k} = 0$. For $i = 1, ..., n_{j,k}$ sample $x_i \sim \text{Bernoulli}(\frac{\alpha_\psi \pi_k^\psi + \kappa_\psi 1_{\{j=k\}}}{i-1+\alpha_\psi \pi_k^\psi + \kappa_\psi 1_{\{j=k\}}})$ and increment $m_{j,k} = 0$ if $x_i = 1$.
- Sampling r: For j=1,...,L. $r_j \sim \text{Binomial}(m_{j,j}, \frac{\rho}{(1-\rho)\pi_j^\psi + \rho})$ where $\rho = \frac{\alpha_\psi}{\alpha_\psi + \kappa_\psi}$
- Set $\bar{m}_{j,k} = m_{j,k}$ if $j \neq k$ and $\bar{m}_{j,k} = m_{j,k} - r_j$ if $j = k$.
- Set $\bar{K} = 0$, for $k = 1, ..., L$, if $\bar{m}_{.,k} > 0$ then increment $\bar{K}$.

(b) Sampling α_ψ and κ_ψ:

- Sample auxiliary variables: for $i = 1, ..., L$, $q_i \sim \text{Beta}(\alpha_\psi + \kappa_\psi + 1, n_{i,.})$ and $s_i \sim \text{Bernoulli}(\frac{n_{i,.}}{n_{i,.} + \alpha_\psi + \kappa_\psi})$.
- Sample $\rho = \frac{\kappa_\psi}{\alpha_\psi + \kappa_\psi} \sim \text{Beta}(\rho_{\text{hyp1}} + r_., \rho_{\text{hyp2}} + m_{.,.} - r_.)$ where ρ_{hyp1} and ρ_{hyp2} denotes the hyper-parameters of ρ (see Table 3).
- Sample $\alpha_\psi + \kappa_\psi \sim \text{G}(a_{\text{hyp}} + m_{.,.} - s_., (\frac{1}{b_{\text{hyp}}} - \log q_.)^{-1})$ where a_{hyp} and b_{hyp} denotes the hyper-parameters of $\alpha_\psi + \kappa_\psi$ (see Table 3).
- Set $\alpha_\psi = (1 - \rho)(\alpha_\psi + \kappa_\psi)$ and $\kappa_\psi = \rho(\alpha_\psi + \kappa_\psi)$.

(c) Sampling η_ψ:

- Sample auxiliary variables: $\tilde{q} \sim \text{Beta}(\eta_\psi + 1, \bar{m}_{.,.})$ and $\tilde{s} \sim \text{Bernoulli}(\frac{\bar{m}_{.,.}}{\bar{m}_{.,.} + \eta_\psi})$.
- Sample $\eta_\psi \sim \text{G}(\eta_{\psi,\text{hyp1}} + \bar{K} - \tilde{s}, \{\frac{1}{\eta_{\psi,\text{hyp2}}} - \log \tilde{q}\}^{-1})$ where $\eta_{\psi,\text{hyp1}}$ and $\eta_{\psi,\text{hyp2}}$ denotes the hyper-parameters of η_ψ (see Table 3).

6. Sampling $\{\alpha_\sigma, \kappa_\sigma, \eta_\sigma\}$ from $f(\alpha_\sigma, \kappa_\sigma, \eta_\sigma | P^\sigma, \pi^\sigma, s^\sigma_{1:T}, y_{1:T})$: similar to previous item.

7. Sampling π^ψ from $f(\pi^\psi | P^\psi, H_{Dir}, s^\psi_{1:T}, y_{1:T}) \sim \text{Dir}(\frac{\eta_\psi}{L} + \bar{m}_{.,1}, ..., \frac{\eta_\psi}{L} + \bar{m}_{.,L})$.

8. Sampling π^σ from $f(\pi^\sigma | P^\sigma, H_{Dir}, s^\sigma_{1:T}, y_{1:T})$: similar to previous item.

9. For $j = 1, ..., L$, sampling σ_j^{-2} from
$f(\sigma_j^{-2} | \bar{e}, \bar{f}, s^\psi_{1:T}, s^\sigma_{1:T}, y_{1:T}) \sim G(0.5 n^\sigma_{.,j} + \bar{e}, (0.5 \sum_{t=1}^T \epsilon_t^2 \delta_{\{s^\sigma_t = j\}} + \frac{1}{\bar{f}})^{-1})$
where $\delta_{\{s^\sigma_t = j\}}$ is the Dirac function equal to one if $s^\sigma_t = j$ and zero otherwise.

10. For $j = 1, ..., L$, sampling μ_j, β_j, ϕ_j from $f(\mu_j, \beta_j, \phi_j | \Theta \backslash \{\mu_j, \beta_j, \phi_j\}, \bar{\mu}, \bar{\Sigma}, s^\psi_{1:T}, s^\sigma_{1:T}, y_{1:T})$ using the MH algorithm described in subsection 3.2.

11. Sampling $\bar{\mu}, \bar{\Sigma}$ from $f(\bar{\mu}, \bar{\Sigma}|\Theta)$:

    - Drawing $\bar{\mu}$ from $f(\bar{\mu}|\bar{\Sigma}, \Theta) \sim N(\mu_{\text{post}}, \Sigma_{\text{post}})$ where $\Sigma_{\text{post}} = (\underline{\Sigma}^{-1} + L\bar{\Sigma}^{-1})^{-1}$ and $\mu_{\text{post}} = \Sigma_{\text{post}}(\underline{\Sigma}^{-1}\underline{\mu} + \sum_{j=1}^{L} \bar{\Sigma}^{-1}\theta_j)$.
    - Drawing $\bar{\Sigma}^{-1}$ from $f(\bar{\Sigma}^{-1}|\bar{\mu}, \Theta) \sim \text{W}((\underline{V}^{-1} + \sum_{j=1}^{L}(\theta_j - \bar{\mu})(\theta_j - \bar{\mu})')^{-1}, \underline{v} + L)$.

12. Sampling $\bar{e}, \bar{f}^{-1}$ from $f(\bar{e}, \bar{f}^{-1}|\{\sigma_1^2, ...\sigma_L^2\}, y_{1:T})$:

    - Drawing $\bar{e}$ from $f(\bar{e}|\bar{f}, \{\sigma_1^2, ...\sigma_L^2\}, y_{1:T})$ by a MH step. The proposal distribution is a random walk Normal one with variance equal to 0.5.
    - Drawing $\bar{f}^{-1}$ from $f(\bar{f}^{-1}|\bar{e}, \{\sigma_1^2, ...\sigma_L^2\}, y_{1:T}) \sim G(L\bar{e} + \underline{f_a}, (\frac{1}{\underline{f_b}} + \sum_{j=1}^{L} \sigma_j^{-2})^{-1})$.

Appendix 2: Approximate MS-ARMA Model

In this appendix, we provide the computation of $\tilde{\epsilon}_{t-1,s_t^\psi}$ that is used in the approximate model (see subection 3.1). By definition of the expectation,

$$E_{s_{t-1}^\psi}[\epsilon_{t-1}|y_{1:t-1}, s_t^\psi, \Theta, P^\psi, s_{1:t-1}^\sigma] = \sum_{i=1}^{L} \epsilon_{t-1}(i) f(s_{t-1}^\psi = i|y_{1:t-1}, s_t^\psi, \Theta, P^\psi, s_{1:t-1}^\sigma),$$

where $\epsilon_{t-1}(i) = y_{t-1} - \mu_i - \beta_i y_{t-2} - \phi_i \tilde{\epsilon}_{t-2,i}$.

The conditional probabilities $f(s_{t-1}^\psi|y_{1:t-1}, s_t^\psi, \Theta, P^\psi, s_{1:t-1}^\sigma)$ are given by

$$\begin{aligned} f(s_{t-1}^\psi|y_{1:t-1}, s_t^\psi, \Theta, P^\psi, s_{1:t-1}^\sigma) &= \frac{f(s_t^\psi, s_{t-1}^\psi|y_{1:t-1}, \Theta, P^\psi, s_{1:t-1}^\sigma)}{f(s_t^\psi|y_{1:t-1}, \Theta, P^\psi, s_{1:t-1}^\sigma)} \\ &= \frac{f(s_{t-1}^\psi|y_{1:t-1}, \Theta, P^\psi, s_{1:t-1}^\sigma) f(s_t^\psi|s_{t-1}^\psi, P^\psi)}{f(s_t^\psi|y_{1:t-1}, \Theta, P^\psi, s_{1:t-1}^\sigma)} \\ &= \frac{f(s_{t-1}^\psi|y_{1:t-1}, \Theta, P^\psi, s_{1:t-1}^\sigma) f(s_t^\psi|s_{t-1}^\psi, P^\psi)}{\sum_{i=1}^{L} f(s_{t-1}^\psi = i|y_{1:t-1}, \Theta, P^\psi, s_{1:t-1}^\sigma) f(s_t^\psi|s_{t-1}^\psi = i, P^\psi)}. \end{aligned}$$

From the last line of the above formula, the forward step of the forward-backward algorithm provides all the quantities to compute at each time index the conditional distribution $f(s_{t-1}^\psi|y_{1:t-1}, s_t^\psi, \Theta, P^\psi, s_{1:t-1}^\sigma)$.

References

Amisano, G. & Giacomini, R. (2007), 'Comparing density forecasts via weighted likelihood ratio tests', *Journal of Business and Economic Statistics* **25**, 177–190.

Ardia, D., Hoogerheide, L. & van Dijk., H. (2009), 'To bridge, to warp or to wrap? A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihoods', *Tinbergen Institute Discussion paper TI 2009-017/4* .

Atchadé, Y. & Rosenthal, J. (2005), 'On adaptive Markov chain Monte Carlo algorithms', *Bernoulli* **11(5)**, 815–828.

Basawa, I. V. & Lund, R. B. (2001), 'Large sample properties of parameter estimates for periodic ARMA models', *Journal of Time Series Analysis* **22**, 651–663.

Bauwens, L., Dufays, A. & Rombouts, J. (2013), 'Marginal likelihood for Markov switching and change-point GARCH models', *Journal of Econometrics* **178**(3), 508–522.

Bauwens, L., Koop, G., Koribilis, D. & Rombouts, J. (2015), 'The contribution of structural break models to forecasting macroecomic series', *Journal of Applied Econometrics* **30**, 596–620.

Beal, M. J. & Krishnamurthy, P. (2006), Gene expression time course clustering with countably infinite hidden Markov models, *in* 'Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)', AUAI Press, Arlington, Virginia, pp. 23–30.

Blackwell, D. & MacQueen, J. (1973), 'Ferguson distributions via Pólya urn schemes', *Annals of Statistics* **1**, 353–355.

Chib, S. (1995), 'Marginal likelihood from the Gibbs output', *Journal of the American Statistical Association* **90**, 1313–1321.

Chib, S. (1996), 'Calculating posterior distributions and modal estimates in Markov mixture models', *Journal of Econometrics* **75**, 79–97.

Chib, S. (1998), 'Estimation and comparison of multiple change-point models', *Journal of Econometrics* **86**, 221–241.

Del Moral, P., Doucet, A. & Jasra, A. (2006), 'Sequential Monte Carlo samplers', *Journal of the Royal Statistical Society (Series B)* **68**, 411–436.

Doornik, J. (2013), 'A Markov-switching model with component structure for US GNP', *Economics Letters* **118**(2), 265–268.

Doucet, A., de Freitas, N. & Gordon, N. (2001), *Sequential Monte Carlo Methods in Practice*, Springer Verlag, Berlin.

Dufays, A. (2014), 'On the conjugacy of off-line and on-line sequential Monte Carlo samplers', *National Bank of Belgium Working Paper No. 263* .

Dufays, A. (2015), 'Infinite-state Markov-switching for dynamic volatility', *Journal of Financial Econometrics* **forthcoming**.

Eo, Y. (2012), Bayesian inference about the types of structural breaks when there are different breaks in many parameters, Technical report, University of Sidney, School of Economics.

Ferguson, T. S. (1973), 'A Bayesian analysis of some nonparametric problem', *The Annals of Statistics* **1**(2), 209–230.

Fox, E., Sudderth, E., Jordan, M. & Willsky, A. (2011), 'A sticky HDP-HMM with application to speaker diarization', *Annals of Applied Statistics* **5**(2A), 1020–1056.

Francq, C. & Gautier, A. (2004), 'Estimation of time-varying ARMA models with markovian changes in regime', *Statistics & Probability Letters* **70**(4), 243–251.

Francq, C. & Zakoian, J. (2008), 'Deriving the autocovariances of powers of Markov-switching GARCH models, with applications to statistical inference', *Computational Statistics and Data Analysis* **52**, 3027–3046.

Fruhwirth-Schnatter, S. (2004), 'Estimating marginal likelihoods for mixture and Markov-switching models using bridge sampling techniques', *The Econometrics Journal* **7**, 143–167.

Geweke, J. (1989), 'Bayesian inference in econometric models using Monte Carlo integration', *Econometrica* **57**, 1317–1339.

Geweke, J. (2007), 'Interpretation and inference in mixture models: Simple MCMC works', *Computational Statistics and Data Analysis* **51**, 3259–3550.

Girolami, M. & Calderhead, B. (2011), 'Riemann manifold Langevin and Hamiltonian Monte Carlo methods', *Journal of the Royal Statistical Society (Series B)* **73**(2), 123–214.

Gneiting, T. & Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**, 359–378.

Gneiting, T. & Ranjan, R. (2011), 'Comparing density forecasts using threshold- and quantile-weighted scoring rules', *Journal of Business and Economic Statistics* **29**(3), 411–422.

Goldfled, S. M. & Quandt, R. E. (1973), 'A Markov model for switching regressions', *Journal of Econometrics* **1**, 3–16.

Goutte, S. (2014), 'Conditional Markov regime switching model applied to economic modelling', *Economic Modelling* **38**(C), 258–269.

Gray, S. (1996), 'Modeling the conditional distribution of interest rates as a regime-switching process', *Journal of Financial Economics* **42**, 27–62.

Haas, M., Mittnik, S. & Paolella, M. (2004), 'A new approach to Markov-switching GARCH models', *Journal of Financial Econometrics* **2**, 493–530.

Hamilton, J. (1989), 'A new approach to the economic analysis of nonstationary time series and the business cycle', *Econometrica* **57**, 357–384.

Henneke, J., Rachev, S., Fabozzi, F. J. & Nikolov, M. (2011), 'MCMC-based estimation of Markov switching ARMA-GARCH models', *Applied Economics* **43**(3), 259–271.

Herbst, E. & Schorfheide, F. (2012), 'Sequential Monte Carlo sampling for DSGE models', *Working Paper No12-27, Federal reserve Bank of Philadelphia* .

Ishwaran, H. & Zarepour, M. (2002), 'Exact and approximate sum representations for the Dirichlet process', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **30**(2), 269–283.

Jasra, A., Stephens, D. A., Doucet, A. & Tsagaris, T. (2011), 'Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo', *Scandinavian Journal of Statistics* **38**, 1–22.

Jensen, M. J. & Maheu, J. M. (2010), 'Bayesian semiparametric stochastic volatility modeling', *Journal of Econometrics* **157**(2), 306–316.

Jensen, M. J. & Maheu, J. M. (2013), 'Bayesian semiparametric multivariate GARCH modeling', *Journal of Econometrics* **176**(1), 3–17.

Jensen, M. J. & Maheu, J. M. (2014), 'Estimating a semiparametric asymmetric stochastic volatility model with a Dirichlet process mixture', *Journal of Econometrics* **178**(P3), 523–538.

Jin, X. & Maheu, J. M. (2014), Bayesian semiparametric modeling of realized covariance matrices, Mpra paper, University Library of Munich, Germany.

Jochmann, M. (2015), 'Modeling U.S. inflation dynamics: A Bayesian nonparametric approach', *Econometric Reviews* **34**(5), 537–558.

Kivinen, J., Sudderth, E. & Jordan, M. (2007), 'Learning multiscale representations of natural scenes using Dirichlet processes', *Proceedings of the IEEE International Conference on Computer Vision* .

Klaassen, F. (2002), 'Improving GARCH volatility forecasts with regime-switching GARCH', *Empirical Economics* **27**(2), 363–394.

Kurihara, K., Welling, M. & Teh, Y. (2007), Collapsed variational Dirichlet process mixture models, *in* 'Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)'.

Marin, J.-M., Mengersen, K. & Robert, C. P. (2005), Bayesian modelling and inference on mixtures of distributions, Handbook of Statistics 25, North Holland, pp. 459–507.

Rabiner, L. R. (1989), 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE* pp. 257–286.

Sethuraman, J. (1994), 'A constructive definition of Dirichlet priors', *Statistica Sinica* **4**, 639–650.

Song, Y. (2014), 'Modelling regime switching and structural breaks with an infinite hidden Markov model', *Journal of Applied Econometrics* **29**(5), 825–842.

Stock, J. H. & Watson, M. W. (2007), 'Why has U.S. inflation become harder to forecast?', *Journal of Money, Credit and Banking* **39**(s1), 3–33.

Teh, Y., Jordan, M., Beal, M. & Blei, D. M. (2006), 'Hierarchical Dirichlet processes', *Journal of the American Statistical Association* **101**, 1566–1581.

Vakilzadeh, M., Yaghoubi, V., Johansson, A. & Abrahamsson, T. (2014), 'Manifold Metropolis adjusted Langevin algorithm for high-dimensional Bayesian FE', *Proceedings of the 9th International Conference on Structural Dynamics* pp. 3029–3036.

Van Gael, J., Saatci, Y., Teh, Y. & Ghahramani, Z. (2008), 'Beam sampling for the infinite hidden Markov model', *Proceedings of the 25th International Conference on Machine learning* .

Xie, W., Lewis, P., Fan, Y., Kuo, L. & Chen, M.-H. (2011), 'Improving marginal likelihood estimation for Bayesian phylogenetic model selection', *Systematic Biology* **60**(2), 150–160.

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S. & Girolami, M. (2014), 'Langevin diffusions and the Metropolis-adjusted Langevin algorithm', *Statistics & Probability Letters* **91**, 14 – 19.