THE RIMINI CENTRE FOR ECONOMIC ANALYSIS

WP 28_12

Yong Song
University of Technology Sydney, Australia
The Rimini Centre for Economic Analysis (RCEA), Italy

MODELLING REGIME SWITCHING AND STRUCTURAL BREAKS WITH AN INFINITE HIDDEN MARKOV MODEL

The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Rimini Centre for Economic Analysis.

The Rimini Centre for Economic Analysis
Legal address: Via Angherà, 22 – Head office: Via Patara, 3 - 47900 Rimini (RN) – Italy
www.rcfea.org - secretary@rcfea.org

Modelling Regime Switching and Structural Breaks with an Infinite Hidden Markov Model [*]

Yong Song[†]
University of Technology, Sydney
RCEA

Abstract

This paper proposes an infinite hidden Markov model to integrate the regime switching and the structural break dynamics in a single, coherent Bayesian framework. Two parallel hierarchical structures, one governing the transition probabilities and another governing the parameters of the conditional data density, keep the model parsimonious and improve forecasts. This flexible approach allows for regime persistence and estimates the number of states automatically. A global identification methodology for structural changes versus regime switching is presented. An application to U.S. real interest rates compares the new model to existing parametric alternatives.

Keywords: Markov switching, structural break, Dirichlet process, infinite hidden Markov model
JEL classification: C11, C53

[*] I am sincerely thankful to my supervisor, Professor John M. Maheu, for guidance and many helpful comments. I am also grateful to Herman K. van Dijk(coeditor), two anonymous referees, an anonymous commenter, Martin Burda, John Geweke, Christian Gourieroux, Robert Kohn, Thomas McCurdy, James Morley, Rodney Strachan and seminar participants at Australian National University, Bank of Canada, CEA conference, CenSoC at University of Technology Sydney, University of Melbourne, University of New South Wales, University of Toronto, and Wilfrid Laurier University. A previous version is titled as *Modelling Regime Switching and Structural Breaks with an Infinite Dimension Markov Switching Model*

[†] Yong Song; Center for the Study of Choice, University of Technology, Sydney and RCEA, P.O. Box 123 Broadway, NSW 2007, Australia; Phone: (61)02 9514 9782; Email: Yong.Song@uts.edu.au

1 Introduction

This paper contributes to the current literature by accommodating regime switching and structural break dynamics in a unified framework. Current regime switching models are not suitable for capturing instability of dynamics because they assume a finite number of states and that the future is like the past. Structural break models allow the dynamics to change over time, however, they may incur loss in the estimation precision because the past states cannot recur and the parameters in each state are estimated separately. An infinite hidden Markov model (iHMM) is proposed to accommodate both types of the models and provide much richer dynamics. This paper shows how to identify structural breaks versus regime switching. The estimation and forecasting are based on a Bayesian framework. The model is applied to U.S. real interest rates.

Regime switching models were first applied by Hamilton (1989). It is an important methodology to model nonlinear dynamics and widely applied to economic data including business cycles (Hamilton 1989), bull and bear markets (Maheu et al. forthcoming), interest rates (Ang and Bekaert 2002) and inflation (Evans and Wachtel 1993). There are two common features of these models. First, past states can recur over time. Second, the number of states is finite (it is usually 2 and at most 4). In the rest of the paper, a regime switching model is assumed to have both features. In practice, the second feature may cause biased out-of-sample forecasts if sudden changes of the dynamics exist.

In contrast to the regime switching models, structural break models can capture dynamic instability by assuming an infinite or a much larger number of states at the cost of extra restrictions. For example, Koop and Potter (2007) proposed a structural break model with an infinite number of states. If there is a change in the data dynamics, it will be captured by a new state. The restriction in their model is that the parameters in a new state are different from those in the previous ones. This condition is imposed for estimation tractability. However, it prevents the data divided by break points from sharing the same model parameters, and could incur some loss in estimation precision. In the current literature, structural break models such as Chib (1998), Wang and Zivot (2000), Pesaran et al. (2006) and Maheu and Gordon (2008) have the same feature as Koop and Potter (2007); namely that the states cannot recur. In the rest of the paper, a structural break model is assumed to have non-recurring states and an infinite or a large number of states.

As we can see, regime switching and structural break dynamics have different implications for data fitting and forecasting. What is missing in the current literature is a method to reconcile them. For instance, a common practice is to use one approach or the other in applications to specific problems. Garcia and Perron (1996) used a three-regime Markov switching model for U.S. real interest rates while Wang and Zivot (2000) applied a model with structural breaks in the mean and the volatility. Did the real interest rates in 1981 have distinct dynamics or return to a historical state with the same dynamics? Existing econometric models have difficulty answering such question.

This paper provides a solution by proposing an infinite hidden Markov model. It incorporates regime switching and structural break dynamics in a unified framework. Recurring states are allowed to improve estimation and forecasting precision. An unknown number of states is embedded in the infinite dimension structure and estimated endogenously to capture the dynamic instability. Different from the Bayesian model averaging methodology, this model combines different dynamics in the estimation.

The proposed model builds on and extends Fox et al. (2011). They used Dirichlet processes as the prior on the transition probabilities of an infinite hidden Markov model. The key innovation in their work is introducing a *sticky* parameter that favors state persistence and avoids the saturation of states. Jochmann (2010) applies their model to detect the number of regimes in U.S. inflation. Their model is denoted by FSJW in the rest of the paper.[1]

The contributions of this paper are as follows. First, a second hierarchical structure in addition to FSJW is introduced to allow learning and sharing of information for the parameters of the conditional data density in each state. Second, I present a methodology to identify structural breaks versus regime switching dynamics.

Lastly, the model is applied to U.S. real interest rates, which is a canonical problem in the existing literature. Specifically, the iHMM is compared to the regime switching model by Garcia and Perron (1996) in a Bayesian framework, the structural break model by Wang and Zivot (2000) with minor modifications and other parametric models. The model comparison shows that the iHMM provides the best out-of-sample forecasts and the regime switching dynamics dominate the structural break dynamics for U.S. real interest rates.

The rest of the paper is organized as follows. Section 2 introduces the Dirichlet process and its stick breaking representation to make this paper self-contained. Section 3 outlines the infinite hidden Markov model and discusses its structure and implications. Section 4 sketches the posterior sampling algorithm, explains how to identify the regime switching and the structural break dynamics, and describes the forecasting method. Section 5 studies the dynamics of U.S. real interest rates and checks the prior sensitivity. Section 6 concludes.

2 Dirichlet Process

The Dirichlet process was introduced by Ferguson (1973) as the extension of the Dirichlet distribution from a finite dimension to an infinite dimension. It is a distribution of

[1] For an infinite hidden Markov model without regime persistence, see Teh et al. (2006). For other nonparametric modeling methods by using the Dirichlet processes but without the hidden Markov representation, see Bassetti et al. (2011); Griffin (2011); Griffin and Steel (2011).

distributions and has two parameters: a shape parameter G_0, which is a probability measure over a sample space Ω, and a scalar concentration parameter $\alpha_0 > 0$. Ferguson (1973) has shown that the random distribution F drawn from a Dirichlet process is almost surely discrete, although the shape parameter G_0 can be continuous. So F can be written as $F = (\Theta, p)$, where $\Theta = (\theta_1, \theta_2, \cdots)'$, $p = (p_1, p_2, \cdots)'$ with $p_i > 0$ and $\sum\limits_{i=1}^{\infty} p_i = 1$. Each distinct value is represented by θ_i and its corresponding probability is p_i, for $i = 1, 2, \cdots$. Sethuraman (1994) found the stick breaking representation of the Dirichlet process as follows:

$$V_i \overset{iid}{\sim} \mathbf{B}(1, \alpha_0); \quad p_i = V_i \prod_{j=1}^{i-1} (1 - V_j) \tag{1}$$

$$\theta_i \overset{iid}{\sim} G_0 \tag{2}$$

The notation $\mathbf{B}$ represents the beta distribution. This representation shows that p and Θ are independent. The process (1), which generates p, is called the *stick breaking process* and denoted by $\mathbf{SBP}(\alpha_0)$ in the rest of the paper.

The Dirichlet process was not widely used for continuous random variables until West et al. (1994) and Escobar and West (1995) proposed the Dirichlet process mixture model (DPM) as follows:

$$p \sim \mathbf{SBP}(\alpha_0); \quad \theta_i \overset{iid}{\sim} G_0 \text{ for } i = 1, 2, \cdots; \quad g(y) = \sum_{i=1}^{\infty} p_i f(y \mid \theta_i). \tag{3}$$

The probability density function $g(y)$, is an infinite mixture of the probability density functions $f(y \mid \theta_i)$'s. If $f(y \mid \theta_i)$ is the normal distribution density function and θ_i represents the mean and the variance, y is distributed as an infinite mixture of normal distributions. Hence, continuous random variables can be modelled nonparametrically by the DPM model. Otranto and Gallo (2002) applied this approach to detect the number of regimes.

The DPM models are usually applied to cross sectional data because of the exchangeability of the observations. However, it is not appropriate for time series modelling since it lacks of state persistence.

3 Infinite Hidden Markov Model

The infinite hidden Markov model is introduced as follows:

$$\pi_0 \sim \mathbf{SBP}(\gamma); \quad \gamma > 0 \tag{4}$$

$$\pi_i \mid \pi_0 \sim \mathbf{DP}\left(c,\ (1-\rho)\pi_0 + \rho\delta_i\right); \quad c > 0, \rho \in (0,1) \tag{5}$$

$$\lambda \sim \mathcal{G} \tag{6}$$

$$\theta_i \overset{iid}{\sim} G_0(\lambda) \tag{7}$$

$$s_t \mid s_{t-1} = i \sim \pi_i \tag{8}$$

$$y_t \mid s_t, Y_{1,t-1} \sim f(y_t \mid \theta_{s_t}, Y_{1,t-1}) \tag{9}$$

where $i = 1, 2, \cdots$. The notation $Y_{1,t} = (y_1, \cdots, y_t)'$ represents the information up to time t. δ_i is the degenerate distribution at integer i.

The *first* hierarchical structure, which governs the transition probabilities, comprises (4) and (5). π_0 is the hierarchical distribution drawn from a stick breaking process and represents a discrete distribution with support on the natural numbers. Each infinite dimensional vector π_i is drawn from a Dirichlet process with the concentration parameter c and the shape parameter $(1-\rho)\pi_0 + \rho\delta_i$. There are three points worth noticing for clarity. First, because the shape parameter $(1-\rho)\pi_0 + \rho\delta_i$ has support only on the natural numbers and each number is associated with a non-zero probability, the random distribution π_i can only take values on the natural numbers and each value has positive probability. After combining the same integer values and sorting them in the ascending order, we can use the vector $\pi_i = (\pi_{i1}, \pi_{i2}, \cdots)'$ to represent a distribution drawn from $\mathbf{DP}(c, (1-\rho)\pi_0 + \rho\delta_i)$. Each element π_{ij} is the probability of s_t taking the integer value j given that $s_{t-1} = i$. Second, by stacking π_i's, we can construct the infinite dimensional transition matrix $P = (\pi_1, \pi_2, \cdots)'$ to obtain the hidden Markov model representation.

Lastly, if ρ is larger, π_i is expected to have a larger probability at integer i. This implies s_t, the state at time t, is more likely to be the same as s_{t-1}. Hence, ρ captures state persistence. In the rest of the paper, ρ is referred as the sticky coefficient. Conditional on π_0 and ρ, the mean of the transition matrix $E(P \mid \pi_0, \rho) = (1-\rho)(\pi_0, \pi_0, \cdots)' + \rho\mathbf{I}_\infty$ is a convex combination of two infinite dimensional matrices. The sticky coefficient ρ increases the self-transition probabilities by adding weights to the infinite dimensional identity matrix $\mathbf{I}_\infty$. The concentration parameter c controls how close P is to $E(P \mid \pi_0, \rho)$.

The *second* hierarchical structure, which governs the parameters of the conditional data density, includes (6) and (7). $G_0(\lambda)$ is the hierarchical distribution from which the state dependent parameter θ_i is drawn independently; $\mathcal{G}$ is the prior of λ. This structure provides a way of learning λ from past values of θ_i to improve estimation and forecasting. If a new state is born, the conditional data density parameter θ_{new} is drawn from $G_0(\lambda)$. Without this hierarchical structure, the new parameter only depends on the assumption. Pesaran et al. (2006) argued the importance of modelling the hierarchical distribution in the presence of structural breaks. This paper adopts their method to estimate the hierarchical distribution $G_0(\lambda)$.

In comparison to the iHMM, FSJW is comprised of (4)-(5) and (7)-(9). The stick breaking representation of the Dirichlet process is not fully exploited by FSJW, since it has only one hierarchical structure on the transition probabilities. In fact, the stick breaking representation (1)-(2) decomposes the random probability measure F drawn from a Dirichlet process into two independent parts: the probabilities are generated from a stick breaking process and the parameter values are drawn from the shape parameter. The iHMM takes fuller advantage of this structure than FSJW by modelling two parallel hierarchical structures.

A common practice of setting the prior for the transition matrix of a finite Markov switching model assumes each row of the transition matrix is drawn from a Dirichlet distribution independently. If extending to the infinite dimension, each row π_i should be drawn from a stick breaking process. However, Teh et al. (2006) argued that this prior may have an overparametrization problem without a hierarchical structure similar to (4) and (5), because it precludes each π_i from sharing information between each other. In terms of parsimony, the iHMM only needs one stick breaking process for the hierarchical distribution π_0, instead of assuming an infinite number of the stick breaking processes for the whole transition matrix P.

The iHMM is also related to the DPM model (3), because (8) and (9) imply $y_t \mid s_{t-1} = i, Y_{1,t-1} \sim \sum_{j=1}^{\infty} \pi_{ij} f(y_t \mid \theta_j, Y_{1,t-1})$. In contrast to the DPM model, the mixture probability π_{ij} is state dependent. This feature allows the iHMM to capture time varying dynamics.

4 Estimation, Inference and Forecast

I assume the conditional density of y_t follows a Gaussian AR(q) process, $y_t \mid s_t, Y_{1,t-1} \sim \mathbf{N}(\phi_{s_t,0} + \phi_{s_t,1} y_{t-1} + \cdots + \phi_{s_t,q} y_{t-q}, \sigma^2_{s_t})$. By definition, the conditional data density parameter is $\theta_i = (\phi_i', \sigma_i)'$ with $\phi_i = (\phi_{i0}, \phi_{i1}, \cdots, \phi_{iq})'$ for $i = 1, 2, \cdots$.

The hierarchical distribution $G_0(\lambda)$ in (7) is assumed as the standard conjugate normal-gamma distribution in the Bayesian literature (Geweke, 2009).

$$\sigma_i^{-2} \sim \mathbf{G}\left(\frac{\chi}{2}, \frac{\nu}{2}\right), \quad \phi_i \mid \sigma_i \sim \mathbf{N}\left(\phi, \sigma_i^2 H^{-1}\right). \tag{10}$$

By definition, the collection of the hierarchical parameters is $\lambda = (\phi, H, \chi, \nu)$. ϕ is a $(q+1) \times 1$ vector, and H is a $(q+1) \times (q+1)$ positive definite matrix. $\frac{\chi}{2} > 0$ is the scalar and $\frac{\nu}{2} > 0$ is the degree of freedom of the gamma distribution.

The prior on the collection of the hierarchical parameters λ in (6) is set as:

$$H \sim \mathbf{W}(A_0, a_0); \quad \phi \mid H \sim \mathbf{N}(m_0, \tau_0 H^{-1}); \quad \chi \sim \mathbf{G}(\frac{d_0}{2}, \frac{c_0}{2}); \quad \nu \sim \mathbf{Exp}(\rho_\nu). \tag{11}$$

H is drawn from a Wishart distribution with parameter A_0, which is a $(q+1) \times (q+1)$ positive definite matrix and the degree of freedom $a_0 > 0$. m_0 is a $(q+1) \times 1$ vector representing the mean of ϕ, and $\tau_0 > 0$ is a scalar. χ is distributed as a gamma distribution with the scalar $\frac{d_0}{2}$ and the degree of freedom $\frac{c_0}{2}$. ν has an exponential distribution with mean ρ_ν.

The posterior sampling is based on Markov chain Monte Carlo (MCMC) methods. Fox et al. (2011) showed the block sampler based on the approximation by a finite number of states is more efficient than the individual sampler. [2] To apply the block sampler, the iHMM is approximated by a finite but large number of states as follows:

$$\pi_0 \sim \mathbf{Dir}\left(\frac{\gamma}{L}, \cdots, \frac{\gamma}{L}\right) \tag{12}$$

$$\pi_i \mid \pi_0 \sim \mathbf{Dir}\left((1-\rho)c\pi_{01}, ..., (1-\rho)c\pi_{0i} + \rho c, \cdots, (1-\rho)c\pi_{0L}\right) \tag{13}$$

$$\lambda \sim \mathcal{G} \tag{14}$$

[2] Consistency of the approximation was proved by Ishwaran and Zarepour (2000, 2002). Ishwaran and James (2001) compared the individual sampler with the block sampler and found the latter is more efficient in terms of mixing.

$$\theta_i \overset{iid}{\sim} G_0(\lambda) \tag{15}$$

$$s_t \mid s_{t-1} = i \sim \pi_i \tag{16}$$

$$y_t \mid s_t, Y_{1,t-1} \sim \mathbf{N}(\phi_{s_t,0} + \phi_{s_t,1} y_{t-1} + \cdots + \phi_{s_t,q} y_{t-q},\ \sigma^2_{s_t}) \tag{17}$$

where L is the maximal number of states in the approximation and $i = 1, 2, \cdots, L$. The approximation in (12) follows Ishwaran and Zarepour (2000), who have shown that $\mathbf{Dir}\left(\frac{\gamma}{L}, \cdots, \frac{\gamma}{L}\right)$ converges to $\mathbf{SBP}(\gamma)$ as $L \to \infty$. Notice that (13) is not an approximation because a Dirichlet process is equivalent to a Dirichlet distribution if its shape parameter only has support on a finite number of elements. The hierarchical distribution $G_0(\lambda)$ and its prior are set as (10) and (11), respectively.

From the empirical point of view, the essence of the iHMM is not only its infinite dimension, but also its sensible hierarchical structure of the prior. If L is large enough, the finite approximation (12)-(17) is equivalent to the original model (4)-(9) in practice. In the application, I check whether L is large enough to preserve the implications from the infinite dimensionality.

4.1 Estimation

Appendix A shows the posterior sampling algorithm. The parameter space is partitioned into four parts: (S, I), (Θ, P, π_0), (ϕ, H, χ) and ν. S, I and Θ are the collections of s_t, a binary auxiliary variable I_t and θ_i, respectively.[3] Each part is sampled conditional on the other parts and the whole data sample Y as follows:

1. Sample $(S, I) \mid \Theta, P, Y$

   (a) Sample $S \mid \Theta, P, Y$ by using the forward filter and backward sampler in Chib (1996).

[3] I_t is an auxiliary variable for sampling π_0. The details are in the appendix A.

   (b) Sample $I \mid S$ by using a Polya Urn scheme.

2. Sample $(\Theta, P, \pi_0) \mid S, I, Y$

   (a) Sample $\Theta \mid S, Y$ by using the regular linear model result.

   (b) Sample $\pi_0 \mid I$ from a Dirichlet distribution.

   (c) Sample $P \mid \pi_0, S$ from Dirichlet distributions.

3. Sample $(\phi, H, \chi) \mid S, \Theta, \nu$

   (a) Sample $(\phi, H) \mid S, \Theta$ by using the conjugacy of the normal-Wishart distribution.

   (b) Sample $\chi \mid \nu, S, \Theta$ from a gamma distribution.

4. Sample $\nu \mid \chi, S, \Theta$ with the random walk Metropolis-Hastings algorithm.

After initializing the parameter values, the algorithm is applied iteratively many times to obtain a large sample of the model parameters. The first block of samples is discarded as the burn-in samples. The rest of the samples, $\{S^{(i)}, \Theta^{(i)}, P^{(i)}, \pi_0^{(i)}, \phi^{(i)}, H^{(i)}, \chi^{(i)}, \nu^{(i)}\}_{i=1}^{N}$, are used for inferences as if they were drawn from the posterior distribution. Simulation consistent posterior statistics are computed as sample averages. For example, the posterior mean of ϕ, $E(\phi \mid Y)$, is calculated by using $\frac{1}{N}\sum_{i=1}^{N}\phi^{(i)}$.

The label switching problem associated with the mixture models is not considered in this paper.[4] Instead, I follow Geweke's (2007) methodology by focusing on the label-invariant statistics and ignoring label permutations during the MCMC simulations.

4.2 Identification of Regime Switching and Structural Breaks

The current literature does not study the identification of regime switching and structural breaks for infinite hidden Markov models. This paper proposes an identification

[4] For the label switching problem, see Celeux et al. (2000) and Frühwirth-Schnatter (2001).

methodology to identify regime switching and structural breaks based on whether a state is recurrent or not. Heuristically, if a state only appears for one consecutive period, it is classified as a non-recurrent state. Otherwise, it is defined as recurrent. The starting time of a recurrent (non-recurrent) state is identified as a regime switching (structural break) point.

More rigorously, if there exist times t_0 and t_1 with $t_0 \leq t_1$ such that $s_t = i$ if and only if $t_0 \leq t \leq t_1$, then state i is non-recurrent and t_0 is identified as a break point. On the other hand, if $s_{t_0} \neq s_{t_0 - 1}$ and t_0 is not a break point, then t_0 is identified as a regime switching point.

Two issues are worth noticing for this identification criterion. First, a true path of states from a regime switching model can have non-recurrent states because of randomness or a small sample size. Hence, this identification approach may label a switching point of a regime switching model as a structural break even if the true states were observed. However, this is simply accidental. As more data are observed, a regime switching model will have all its states identified as recurrent. Second, the purpose of the identification is not to decompose the infinite hidden Markov model into several regime switching and structural break sub-models (there is no unique way even if we wanted to), but to study the richer dynamics which allow recurrent states while accommodating structural breaks.

4.3 Forecast and Model Comparison

Defining $\Psi^{(i)} = \{S^{(i)}, \Theta^{(i)}, P^{(i)}, \pi_0^{(i)}, \phi^{(i)}, H^{(i)}, \chi^{(i)}, \nu^{(i)}\}$ as one sample of the parameters from the posterior distribution conditional on data $Y_{1,t}$, the out-of-sample conditional predictive density at time $t+1$ is $p(\tilde{y}_{t+1} \mid \Psi^{(i)}, Y_{1,t}) = \sum_{j=1}^{L} \pi^{(i)}_{s_t^{(i)}, j} f(\tilde{y}_{t+1} \mid \theta_j^{(i)}, Y_{1,t})$, where $\tilde{y}_{t+1}$ is the random variable at time $t+1$. After integrating out the model parameters by taking the average over the posterior samples, we can obtain the predictive density as $\hat{p}(\tilde{y}_{t+1} \mid Y_{1,t}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}\pi^{(i)}_{s_t^{(i)},j} f(\tilde{y}_{t+1} \mid \theta_j^{(i)}, Y_{1,t})$. If replacing $\tilde{y}_{t+1}$ by the observed value y_{t+1}, the value $\hat{p}(y_{t+1} \mid Y_{1,t}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}\pi^{(i)}_{s_t^{(i)},j} f(y_{t+1} \mid \theta_j^{(i)}, Y_{1,t})$ is called the predictive likelihood of y_{t+1}. Similarly, the predictive mean is obtained as $\hat{E}(\tilde{y}_{t+1} \mid Y_{1,t}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}\pi^{(i)}_{s_t^{(i)},j} E(\tilde{y}_{t+1} \mid \theta_j^{(i)}, Y_{1,t})$.

This paper compares the iHMM to the existing alternative models by using the predictive likelihood of the last 80% of the data. The predictive likelihood of a model M_i can be decomposed as the product of successive predictive likelihoods $p(Y_{t+1,T} \mid M_i, Y_{1,t}) = \prod_{\tau=t+1}^{T} p(y_\tau \mid Y_{1,\tau}, M_i)$. A large value means a better out-of-sample forecasting ability for model M_i. The overparametrization is penalized by the Bayesian method, so the model comparison obeys Ockham's razor.

Kass and Raftery (1995) compared model M_i and M_j by the log Bayes factor: $\log(BF_{ij}) = \log\frac{p(Y_{1,T}|M_i)}{p(Y_{1,T}|M_j)}$. They suggested interpreting the evidence for M_i versus M_j as: not worth more than a bare mention if $0 \leq \log(BF_{ij}) < 1$; positive if $1 \leq \log(BF_{ij}) < 3$; strong if $3 \leq \log(BF_{ij}) < 5$; very strong if $\log(BF_{ij}) \geq 5$. Geweke and Amisano (2010) argued that the predictive likelihood comparison is more robust to the prior elicitation than the marginal likelihood. Meanwhile, the interpretation of the log predictive Bayes factor $\log(BF_{ij} \mid Y_{1,t}) = \log\left(\frac{p(Y_{t+1,T}|Y_{1,t},M_i)}{p(Y_{t+1,T}|Y_{1,t},M_j)}\right)$ is the same as the log Bayes factor if the initial sample $Y_{1,t}$ is regarded as a training sample. This paper uses the predictive likelihood for model comparison with Kass and Raftery's (1995) criterion.

5 Application to U.S. Real Interest Rates

The dynamic stability was tested by Fama (1975), Rose (1988) and Walsh (1987). While Fama (1975) found the *ex ante* real interest rate as a constant, Rose (1988) and Walsh (1987) cannot reject the existence of an integrated component. Garcia and Perron (1996) reconciled these results using a three-regime Markov switching model and found switching points at the beginning of 1973 (the oil crisis) and the middle of 1981 (the federal budget deficit) using quarterly U.S. real interest rates of Huizinga and Mishkin (1986) from 1961Q1-1986Q3. The real interest rate dynamics in each state are characterized by a Gaussian AR(2) process. Wang and Zivot (2000) used the same data to investigate structural breaks and found support of four states (3 breaks) by Bayes factors.

This paper constructs U.S. quarterly real interest rates in the same way as Huizinga and Mishkin (1986) and extends their data set to a total of 252 observations from 1947Q1 to 2009Q4. The summary statistics are in Table 1.

5.1 Models and Priors

The iHMM uses the following prior: $\gamma = 1, c = 10, \rho = 0.9, A_0 = 0.2\mathbf{I}, a_0 = 5, m_0 = \underline{0}, \tau_0 = 1, d_0 = 1, c_0 = 5$ and $\rho_\nu = 5$. This prior is informative but covers a wide range of the parameter space. I assume each regime has an AR(2) representation. The block sampler uses the truncation of $L = 10$.[5]

A K-regime Markov switching model of Garcia and Perron (1996), which is denoted by MS(K), is put in a Bayesian framework. Each regime has an AR(2) representation.

$$P(s_t = j \mid s_{t-1} = i) = p_{ij} \tag{18}$$

$$y_t \mid s_t, Y_{1,t-1} \sim \mathbf{N}(\phi_{s_t,0} + \phi_{s_t,1} y_{t-1} + \phi_{s_t,2} y_{t-2},\ \sigma^2_{s_t}) \tag{19}$$

The prior for the transition probabilities are set as K independent uniform Dirichlet distributions as $(p_{i1}, \cdots, p_{iK}) \overset{iid}{\sim} \mathbf{Dir}(1, \cdots, 1)$ for $i = 1, \cdots, K$. The prior for the data density parameters are set as a normal gamma distribution for each regime as (10). I set $\chi = 5, \nu = 5, H = \mathbf{I}$ and $\phi = \underline{0}$, which are the means of the aforementioned hierarchical

[5] $L = 10$ is chosen to represent a potentially large number of states and keep a reasonable amount of computation. Some larger L's are also tried and produce similar results.

prior for the iHMM.

The structural break model follows Wang and Zivot (2000) with minor modifications. The difference is that Wang and Zivot (2000) only allow the intercept and the volatility to change, while I assume all parameters change simultaneously. Each regime has an AR(2) representation.

$$P(s_t = i \mid s_{t-1} = i) = \begin{cases} p & \text{if } i < K \\ 1 & \text{if } i = K \end{cases} \tag{20}$$

$$P(s_t = i + 1 \mid s_{t-1} = i) = 1 - p \quad \text{if } i < K \tag{21}$$

$$y_t \mid s_t, Y_{1,t-1} \sim \mathbf{N}(\phi_{s_t,0} + \phi_{s_t,1} y_{t-1} + \phi_{s_t,2} y_{t-2},\ \sigma^2_{s_t}) \tag{22}$$

where $i = 1, \cdots, K$. The prior of p is a beta distribution $\mathbf{B}(9, 1)$. The prior for the data density parameters, (ϕ_i, σ_i), is the same as the MS(K) model.

A linear AR(q) model is applied as a benchmark for model comparison:

$$y_t \mid Y_{t-1}, \phi, \sigma \sim \mathbf{N}(\phi_0 + \phi_1 y_{t-1} + \cdots + \phi_q y_{t-q}, \sigma^2). \tag{23}$$

The prior of (ϕ, σ) is the same as the MS(K) and the SB(K) model, where the dimensionality of ϕ depends on the number of lags q.

The AR(q) models with rolling windows are also estimated to control for structural instability. The model and the prior are the same as the aforementioned AR(q) model. The windows used are 3, 5, 10 and 20 years.

The last candidate is the Bayesian model averaging (BMA) approach. The first BMA, denoted by BMA:MS, includes 10 Markov switching models from MS(1) to MS(10), among which MS(1) is simply the linear AR(2) model and MS(10) is the iHMM without the hierarchical structures. The second BMA, denoted by BMA:SB, includes 20 structural break models from SB(1) to SB(20). The last BMA, denoted by BMA:MS+SB, includes the above 10 Markov switching models and 20 structural break models.

5.2 Results

Table 2 shows the log predictive likelihoods of different models. First, the table shows that the linear models are dominated by the nonlinear models. Second, the AR models with a 5-year rolling window performs competitively, which support the importance of the nonlinearity. Third, the log predictive likelihoods strongly support the Markov switching models against the structural break models. The log predictive likelihood of the four-regime or five-regime Markov switching model is larger than that of any K-regime structural break model by more than 5, which is very strong evidence based on Kass and Raftery (1995). At last, the iHMM performs the best among all models, including the BMA approaches, even though the evidence is not strong. Since the iHMM nests the MS(K) models, the results imply that the real interest rates are better described by the regime switching dynamics.

To study the in-sample dynamics, the full sample is estimated by the iHMM. Figure 1 plots the posterior means of different parameters over time, including the regime switching and structural break probabilities. There is no sign of structural breaks from the bottom panel. Hence, the regime switching dynamics prevail, which is consistent with Table 2's results. Three important regimes can be visually detected: one has high volatility and high persistence, one has low volatility and intermediate persistence and the last one has intermediate volatility and low persistence.

Figure 2 plots the posterior mean of the cumulative number of active states over time. A state is defined as active if it is occupied by data. The posterior mean of the total number of active states is 3.4. Recalling that the finite approximation uses $L = 10$

in the estimation, this value implies that the truncation restriction is not binding.

The clustering of the regimes is shown in Figure 3, which is the temperature plot as in Geweke and Jiang (2011). This 2D plot is equivalent to a $T \times T$ matrix, in which the value of the tth row and τth column is the probability of two periods being in the same regime, $p(s_t = s_\tau \mid Y)$. A large(small) value is represented by a dark(light) color. There are three visible regimes. The first is associated with the very beginning of the data period, early and mid 1980s and the very end of the data period. The second regime occupies the longest period, which are the early 1950s, from the late 1950s to the early 1970s and from the mid 1980s to the early 2000s. The last regime is short and has some uncertainties, which is associated with a short period around the mid 1950s, from the early 1970s to the late 1970s and a short period around early 2000s. For the last regime, the uncertainty comes from whether the last two episodes are associated with the first one. It is quite certain that the last two are in the same regime from the figure, since the color is dark for the off-diagonal block which is located at the intersection of these two episodes. However, the color for the off-diagonal blocks, which are located at the intersections of the first episode and the last two episodes, are gray. Because the gray color means a probability around 0.5, loosely speaking, these three episodes have half probability to be in one regime and half probability to be in two regimes. This explains that the posterior mean of the total number of regimes is 3.4

Garcia and Perron (1996) found switching points at the beginning of 1973 and the middle of 1981. From the iHMM, the probability of regime switching in 1973Q1 is 0.39, which is consistent with their finding. From 1980Q2 to 1981Q1, the probabilities of regime switching are 0.18, 0.13, 0.32 and 0.19, respectively. There are many uncertainties in the switching point identification at these times. However, it is quite likely that the state changed in one of these episodes, which is only slightly earlier than in Garcia and Perron (1996). On the other hand, Huizinga and Mishkin (1986) identified

October 1979 and October 1982 as the turning points. Probabilities of regime switching or structural breaks in 1979Q3 and Q4 are less than 0.02 and 0.04 respectively, while in 1982Q3 and 1982Q4 they are both less than 0.01. Thus, the iHMM supports Garcia and Perron (1996) against Huizinga and Mishkin (1986).

As an attempt to locate potential state changing points, I define a time as a candidate turning point if the sum of regime switching and structural break probability is greater than 0.3. There are 9 points in total: 1952Q1, 1952Q3, 1956Q2, 1958Q2, 1973Q1, 1980Q4, 1986Q2, 2002Q1, and 2005Q3. Among those points, 1973Q1 and 1980Q4 are consistent with Garcia and Perron (1996). Wang and Zivot (2000) found 1970Q3, 1980Q2 and 1985Q4 as structural break points. 1980Q4 and 1986Q2 are close to their finding. However, the iHMM does not identify late 1970 as neither a break nor a switching point, which contradicts their result.

Another interesting result is shown in Figure 4 by using the same data length as Garcia and Perron (1996), which is from 1961Q1 to 1986Q3 with 103 observations. If we ignore the last few data points, it is exact like the temperature plot of a structural break model as in Geweke and Jiang (2011). A structural break model by Wang and Zivot (2000) is indeed a good fit for this sub-sample.

5.3 Marginal Likelihood and Prior Sensitivity

For completeness, I also calculate the marginal likelihoods for different models.[6] The model with the largest marginal likelihood is the BMA:MS model (-549.0) comparing to the iHMM(-558.3). This is because of the high marginal likelihoods for the MS(K) models when K is large. For example, the marginal likelihood of MS(10) is -548.8. Its dominance is clearly from the first 50 observations instead of the last 200, since the iHMM has a larger predictive likelihood in Table 2. This is due to the fact that the initial

[6] It is not shown in the paper and can be requested from the author.

samples are more sensitive to the prior elicitation. The hierarchical prior is more diffuse than the non-hierarchical one, hence it tends to produce smaller predictive likelihoods for the first several observations. On the other hand, the hierarchical structure is able to learn more information from the data, so it could have better forecasting ability after a certain training sample.

To further investigate the prior sensitivity, I focus on the inverse of the covariance matrix of the regression coefficients H (up to a scalar σ^2). If H is large, the prior shrinks the estimates of ϕ_i's towards the hyper-parameter ϕ. So I call a prior tight if H is large and flat if it is small. The original prior of H is a Wishart $\mathbf{W}(0.2\mathbf{I}, 5)$ in the iHMM. The MS(K) model set $H = \mathbf{I}$, which is the mean of the iHMM's prior.

For the tight prior, the inverse of the covariance matrix H is set as $10\mathbf{I}$ for the MS(K) models, which means the covariance matrix of ϕ_i is divided by 10. The prior of H in the iHMM is given by scaling A_0 by 10, which means $H \sim \mathbf{W}(2\mathbf{I}, 5)$. So the mean of the new hierarchical prior in the iHMM equals to the new value of H in the MS(K) model. The scalar τ_0 is scaled by 10 in order to compensate for the change of the covariance matrix of ϕ. For a even tighter prior, $H = 100\mathbf{I}$ for the MS(K), and $A_0 = 20\mathbf{I}$ and $\tau_0 = 100$ for the iHMM.

For the flat prior, I set $H = 0.1\mathbf{I}$ for the MS(K) models. Correspondingly, A_0 and τ_0 are set as $0.02\mathbf{I}$ and 0.1 for the iHMM. A even flatter prior is $H = 0.01\mathbf{I}$ for the MS(K) models and $A_0 = 0.002\mathbf{I}, \tau_0 = 0.01$ for the iHMM.

Table 3 shows the marginal and predictive likelihoods for the MS(K) models and the iHMM. In terms of marginal likelihood, the flat and the flatter prior favor the parsimonious MS(3) model. The MS(10) model is beaten because of a *bad* prior elicitation, which is in line with Maheu and McCurdy (2009). For the tight and tighter priors, the model with a larger number of regimes such as MS(5) or the iHMM is more acceptable.

The most important inference from Table 3 is that the iHMM always performs the best in the predictive likelihood. The power of the hierarchical structure resides in its ability to learn extra information from the data. In this application, the iHMM learns quickly and shows a superior predictive ability for the last 80% of the sample.

Another important finding is that the original prior is the most disadvantageous prior in the sensitivity check, because, with the original prior, the predictive Bayes factor of the iHMM against the best MS(K) model is the smallest. This finding is positive evidence to support the robustness of the results of the model comparison.

Figure 5 monitors the evolution of the cumulative log predictive Bayes factor, $\log(\frac{p(Y_{1,t}|\text{iHMM})}{p(Y_{1,t}|\text{AR}(2))})$, which is the solid line on the top panel. The value in the end of the line is the log Bayes factor by definition. The dashed line in the top panel represents the regime change probability. The bottom panel plots the smoothed volatility implied by the iHMM.

The learning of the hierarchical structure can be illustrated by focusing on the high volatility phases, which includes the beginning of the data period, early and mid 1980s and the end of the data period. The first segment has a jump of the Bayes factor because of the prior setting. After the jump, the Bayes factor is quite stable since there is no structural change for the first several observations. Before the second segment, the Bayes factor begins to fall in 1973 since a new regime is entered and the iHMM is learning. The Bayes factor decreases mildly in the second high volatility phase, because the iHMM is still learning the dynamic structure. Finally, the Bayes factor does not decrease in the last phase of high volatility, since the iHMM has accumulated enough information about the high volatility phase. For the rest of the time, the Bayes factor is increasing, simply because the nonlinear dynamics outperform the linear dynamics and the iHMM is able to capture the regime changes while AR(2) model can not.

5.4 Predictive Mean

Table 4 shows the root mean square errors from different approaches. The best model is AR(3). This could be attributed to the fact that the iHMM focuses on the overall density forecasting instead of the point forecast.[7] This is analogous to forecasting at different horizons. A model being good at short-run forecast is not necessarily good at long-run forecasting.

Meanwhile, we can still learn from Table 4 that the iHMM with the tighter prior performs the best among all nonlinear models including the AR(q) models with rolling windows. The tighter prior means that additional shrinkage hierarchical prior methods may provide improvement in the iHMM for both density and mean prediction.

5.5 Dynamics in Each Regime

In order to check the sensitivity of the dynamics in each regime. I estimated another two versions of the iHMM's by assuming AR(1) and AR(3) in each regime, respectively. The priors are the same as the original prior of the iHMM.

The marginal likelihoods for the iHMM with the AR(1) and AR(3) dynamics are -556.8 and -552.5. Their respective predictive likelihoods are -423.0 and -420.8. These values are close to the results of the iHMM with the AR(2) dynamics and do not change the qualitative implications of Table 2 and 3. Again, the original setting is a disadvantageous specification, hence the AR(1) and AR(3) results support the model robustness.

[7] Pesaran et al. (2011) discuss the mean forecasting in the presence of structural instability

6 Conclusion

This paper proposes to apply an infinite hidden Markov model (iHMM) to integrate current Markov switching and structural break models in a single, coherent framework. Two parallel hierarchical structures, one governing the transition probabilities and the other governing the parameters of the conditional data density, are imposed for parsimony and to improve forecasts. A methodology for the identification of regime switching and structural breaks is proposed.

The application to U.S. real interest rates shows the iHMM is robust to model uncertainty and provides superior out-of-sample forecasts than existing Markov switching and structural break models. The second hierarchical structure is robust to prior elicitation and able to learn extra information from the data quickly. From both the density forecasts and the posterior probabilities of regime switching and structural breaks, U.S. real interest rates are better described by a regime switching model.

A Block Sampling

A.1 Sample $(S, I) \mid \Theta, P, Y$

$S \mid \Theta, P, Y$ is sampled by the forward and backward smoother in Chib (1996).

I is introduced to facilitate the π_0 sampling. From (12) and (13), the filtered distribution of π_i conditional on $S_t = (s_1, \cdots, s_t)$ and π_0 is a Dirichlet distribution:

$$\pi_i \mid S_t, \pi_0 \sim \textbf{Dir}\left(c(1-\rho)\pi_{01} + n_{i1}^{(t)}, \cdots, c(1-\rho)\pi_{0i} + c\rho + n_{ii}^{(t)}, \cdots, c(1-\rho)\pi_{0L} + n_{iL}^{(t)}\right)$$

where $n_{ij}^{(t)}$ is the number of $\{\tau \mid s_\tau = j, s_{\tau-1} = i, \tau \leq t\}$. After integrating out π_i, the conditional distribution of s_{t+1} given S_t and π_0 is $p(s_{t+1} = j \mid s_t = i, S_t, \pi_0) \propto c(1-\rho)\pi_{0j} + c\rho\delta_i(j) + n_{ij}^{(t)}$

Construct a variable I_t with a Bernoulli distribution

$$p(I_{t+1} \mid s_t = i, S_t, \pi_0) \propto \begin{cases} c\rho + \sum\limits_{j=1}^{L} n_{ij}^{(t)} & \text{if } I_{t+1} = 0, \\ c(1-\rho) & \text{if } I_{t+1} = 1. \end{cases}$$

and the conditional distribution

$$p(s_{t+1} = j \mid I_{t+1} = 0, s_t = i, S_t, \beta) \propto n_{ij}^{(t)} + c\rho\delta_i(j),$$

$$p(s_{t+1} = j \mid I_{t+1} = 1, s_t = i, S_t, \beta) \propto \pi_{0j}.$$

This construction preserves the same conditional distribution of s_{t+1} given S_t and π_0. To sample $I \mid S$, use the Bernoulli distribution $I_{t+1} \mid s_t = i, s_{t+1} = j, \pi_0 \sim \mathbf{Ber}(\frac{c(1-\rho)\pi_{0j}}{n_{ij}^{(t)} + c\rho\delta_i(j) + c(1-\rho)\pi_{0j}})$.

A.2 Sample $(\Theta, P, \pi_0) \mid S, I, Y$

After sampling I and S, write $m_i = \sum\limits_{s_t=i} I_t$. By construction, the conditional posterior of π_0 given S and I only depends on I and is given by $\pi_0 \mid S, I \sim \mathbf{Dir}(\frac{\gamma}{L} + m_1, \cdots, \frac{\gamma}{L} + m_L)$. This approach of sampling π_0 is easier than Fox et al. (2011).

Conditional on π_0 and S, the sampling of π_i is given by $\pi_i \mid \pi_0, S \sim \mathbf{Dir}(c(1-\rho)\pi_{01} + n_{i1}, \cdots, c(1-\rho)\pi_{0i} + c\rho + n_{ii}, \cdots, c(1-\rho)\pi_{0L} + n_{iL})$, where n_{ij} is the number of $\{\tau \mid s_\tau = j, s_{\tau-1} = i\}$.

The sampling of $\Theta \mid S, Y$ uses the result of regular linear models with the conjugate priors.

A.3 Sample $(\phi, H, \chi) \mid S, \Theta, \nu$

The conditional posterior is $\phi, H \mid \{\phi_i, \sigma_i\}_{i=1}^K \sim \mathbf{NW}(m_1, \tau_1, A_1, a_1)$, where K is the number of active states. ϕ_i and σ_i are the parameters associated with these states. We can derive $m_1 = \frac{1}{\tau_0^{-1} + \sum_{i=1}^K \sigma_i^{-2}} \left(\tau_0^{-1} m_0 + \sum_{i=1}^K \sigma_i^{-2} \phi_i \right)$, $\tau_1 = \frac{1}{\tau_0^{-1} + \sum_{i=1}^K \sigma_i^{-2}}$, $A_1 = \left(A_0^{-1} + \sum_{i=1}^K \sigma_i^{-2} \phi_i \phi_i' + \tau_0^{-1} m_0 m_0' - \tau_1^{-1} m_1 m_1' \right)^{-1}$ and $a_1 = a_0 + K$.

The conditional posterior of χ is $\chi \mid \nu, \{\sigma_i\}_{i=1}^K \sim \mathbf{G}(d_1/2, c_1/2)$, where $d_1 = d_0 + \sum_{i=1}^K \sigma_i^{-2}$ and $c_1 = c_0 + K\nu$.

A.4 Sample $\nu \mid \chi, S, \Theta$

The conditional posterior of ν is $p(\nu \mid \chi, \{\sigma_i\}_{i=1}^K) \propto \left(\frac{(\chi/2)^{\nu/2}}{\Gamma(\nu/2)} \right)^K \left(\prod_{i=1}^K \sigma_i^{-2} \right)^{\nu/2} \exp\{-\frac{\nu}{\rho_\nu}\}$. A Metropolis-Hastings method is applied to sample ν, with the proposal distribution of $\nu \mid \nu' \sim \mathbf{G}(\frac{\zeta_\nu}{\nu'}, \zeta_\nu)$. ζ_ν is fine tuned to produce a reasonable acceptance rate around 0.5, as suggested by Müller (1991); Roberts et al. (1997).

References

Ang, A. and Bekaert, G. Regime switches in interest rates. *Journal of Business & Economic Statistics*, 20(2):163–182, 2002.

Bassetti, F., Casarin, R., and Leisen, F. Beta-product poisson-dirichlet processes. *Arxiv preprint arXiv:1109.4777*, 2011.

Celeux, G., Hurn, M., and Robert, C.P. Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association*, 95 (451), 2000.

Chib, S. Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, 75(1):79–97, 1996.

Chib, S. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241, 1998.

Escobar, MD and West, M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 1995.

Evans, M. and Wachtel, P. Inflation regimes and the sources of inflation uncertainty. *Journal of Money, Credit and Banking*, pages 475–511, 1993.

Fama, E.F. Short-term interest rates as predictors of inflation. *The American Economic Review*, 65(3):269–282, 1975.

Ferguson. A bayesian analysis of some nonparametric problem. *The Annals of Statistics*, 1(2):209–230, 1973.

Fox, E.B., Sudderth, E.B., Jordan, M.I., and Willsky, A.S. A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

Frühwirth-Schnatter, S. Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of the American Statistical Association*, 96(453), 2001.

Garcia, R. and Perron, P. An analysis of the real interest rate under regime shifts. *The Review of Economics and Statistics*, 78(1):111–125, 1996.

Geweke, J. Interpretation and inference in mixture models: Simple MCMC works. *Computational Statistics & Data Analysis*, 51(7):3529–3550, 2007.

Geweke, J. *Complete and Incomplete Econometric Models*. Princeton Univ Pr, 2009.

Geweke, J. and Amisano, G. Comparing and evaluating bayesian predictive distributions of asset returns. *International Journal of Forecasting*, 26(2):216–230, 2010.

Geweke, J. and Jiang, Y. Inference and prediction in a multiple structural break model. *Journal of Econometrics*, 163(2):172–185, 2011.

Griffin, J.E. Inference in infinite superpositions of non-gaussian ornstein–uhlenbeck processes using bayesian nonparametic methods. *Journal of Financial Econometrics*, 9(3):519, 2011.

Griffin, J.E. and Steel, M.F.J. Stick-breaking autoregressive processes. *Journal of Econometrics*, 162(2):383–396, 2011.

Hamilton, J.D. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, 57(2): 357–384, 1989.

Huizinga, J. and Mishkin, F.S. Monetary policy regime shifts and the unusual behavior of real interest rates. *Carnegie-Rochester Conference Series on Public Policy*, 24(1): 231–274, 1986.

Ishwaran, H. and James, L.F. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453), 2001.

Ishwaran, H. and Zarepour, M. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371, 2000.

Ishwaran, H. and Zarepour, M. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12(3):941–963, 2002.

Jochmann, M. Modeling U S Inflation Dynamics: A Bayesian Nonparametric Approach. Working Paper, 2010.

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

Koop, G. and Potter, S.M. Estimation and forecasting in models with multiple breaks. *Review of Economic Studies*, 74(3):763, 2007.

Maheu, J.M. and Gordon, S. Learning, forecasting and structural breaks. *Journal of Applied Econometrics*, 23(5):553–584, 2008.

Maheu, J.M. and McCurdy, T.H. How useful are historical data for forecasting the long-run equity return distribution? *Journal of Business and Economic Statistics*, 27(1):95–112, 2009.

Maheu, J.M., McCurdy, T.H., and Song, Y. Components of bull and bear markets: bull corrections and bear rallies. *Journal of Business and Economic Statistics*, forthcoming.

Müller, P. A generic approach to posterior integration and Gibbs sampling. *Rapport technique*, pages 91–09, 1991.

Otranto, E. and Gallo, G.M. A nonparametric bayesian approach to detect the number of regimes in markov switching models. *Econometric Reviews*, 21(4):477–496, 2002.

Pesaran, M.H., Pettenuzzo, D., and Timmermann, A. Forecasting time series subject to multiple structural breaks. *Review of Economic Studies*, 73(4):1057–1084, 2006.

Pesaran, M.H., Pick, A., and Pranovich, M. Optimal forecasts in the presence of structural breaks. Cambridge Working Papers in Economics, 2011.

Roberts, GO, Gelman, A., and Gilks, WR. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

Rose, A.K. Is the real interest rate stable? *Journal of Finance*, 43(5):1095–1112, 1988.

Sethuraman, J. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Teh, Y.W., Jordan, M.I., Beal, M.J., and Blei, D.M. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Walsh, C.E. Three questions concerning nominal and real interest rates. *Economic Review*, (Fall):5–19, 1987.

Wang, J. and Zivot, E. A Bayesian time series model of multiple structural changes in level, trend, and variance. *Journal of Business & Economic Statistics*, 18(3):374–386, 2000.

West, M., Müller, P., and Escobar, M.D. Hierarchical priors and mixture models, with application in regression and density estimation. *Aspects of uncertainty: A Tribute to DV Lindley*, pages 363–386, 1994.

Tables

Table 1: Summary statistics

mean	0.96
variance	9.91
skewness	-0.76
excess kurtosis	3.60

There are 252 observations from 1947Q1 to 2009Q4 for U.S. quarterly real interest rate.

Table 2: Log predictive likelihoods

AR(q)	q=2	q=3	q= 4			
	-456.5	-450.2	-455.7			
rolling-AR(2)	3yr	5yr	10yr	20 yr		
	-440.9	-432.7	-448.1	-475.6		
rolling-AR(3)	3yr	5yr	10yr	20 yr		
	-451.7	-436.0	-441.8	-467.9		
rolling-AR(4)	3yr	5yr	10yr	20 yr		
	-462.4	-439.0	-444.7	-467.6		
MS(K)	K=3	K=4	K=5			
	-436.7	-427.4	-426.8			
SB(K)	K=3	K=4	K=5	K=10	K=15	K=20
	-452.6	-454.6	-443.3	-436.8	-432.9	-430.0
BMA:MS		-423.6				
BMA:SB		-433.1				
BMA:MS+SB		-423.6				
iHMM		**-423.3**				

There are 252 observations from 1947Q1 to 2009Q4 for U.S. quarterly real interest rate. The last 200 observations are used to calculate the predictive likelihoods.

Table 3: Robustness check for prior elicitation

	MS(3)	MS(4)	MS(5)	MS(10)	iHMM
original:log ML	-563.7	-553.2	-551.9	**-548.8**	-558.3
original:log PL	-436.7	-427.4	-426.8	-423.6	**-423.3**
tight:log ML	-556.5	-552.5	**-551.1**	-553.2	-553.9
tight:log PL	-430.1	-426.4	-425.1	-426.5	**-422.6**
tighter:log ML	-572.5	-568.7	-563.9	-568.7	**-550.7**
tighter:log PL	-437.6	-434.1	-428.9	-426.9	**-420.5**
flat:log ML	**-564.3**	-567.6	-570.6	-586.5	-568.9
flat:log PL	-430.0	-432.1	-434.9	-447.8	**-425.6**
flater:log ML	**-579.7**	-580.5	-582.8	-600.7	-580.2
flater:log PL	-439.3	-433.2	-433.0	-436.9	**-431.0**

There are 252 observations from 1947Q1 to 2009Q4 for U.S. quarterly real interest rate. The last 200 observations are used to calculate the predictive likelihoods.

Table 4: Root mean square error(RMSE)

AR(q)	q=2	q=3	q= 4			
	2.31	**2.26**	2.39			
rolling-AR(2)	3yr	5yr	10yr	20 yr		
	2.45	2.42	2.41	2.36		
rolling-AR(3)	3yr	5yr	10yr	20 yr		
	2.64	2.47	2.40	2.33		
rolling-AR(4)	3yr	5yr	10yr	20 yr		
	2.83	2.59	2.49	2.42		
MS(K)	K=3	K=4	K=5			
	2.34	2.37	2.37			
SB(K)	K=3	K=4	K=5	K=10	K=15	K=20
	2.42	2.49	2.44	2.65	2.87	2.75
iHMM		2.45				
iHMM tighter prior		2.33				

There are 252 observations from 1947Q1 to 2009Q4 for U.S. quarterly real interest rate. The last 200 observations are used to calculate the forecasting errors.

Figures

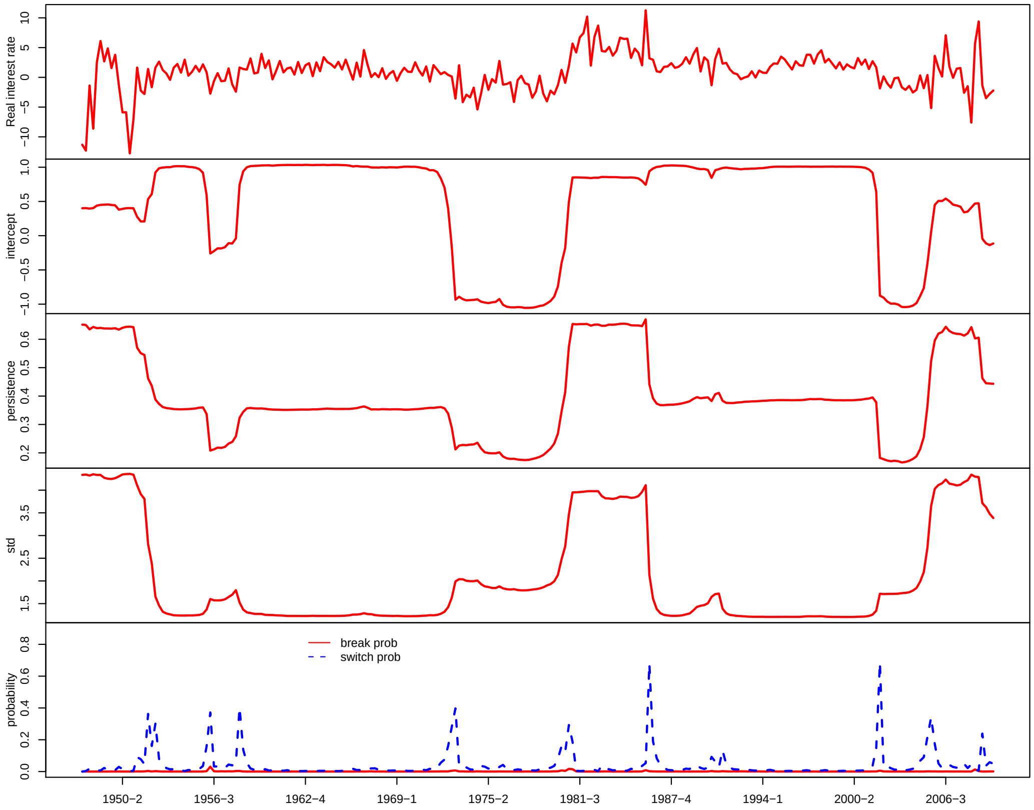

Figure 1: There are 252 observations from 1947Q1 to 2009Q4 for U.S. quarterly real interest rate. The data are estimated by the iHMM and each state has Gaussian AR(2) dynamics: $y_t = \phi_{s_t,0} + \phi_{s_t,1}y_{t-1} + \phi_{s_t,2}y_{t-2} + \sigma_{s_t}\varepsilon_t$. The first panel plots the data and the rest plots the posterior mean of different parameters: the second panel plots the intercepts $\phi_{s_t,0}$, the third panel plots the persistence parameters $\phi_{s_t,1} + \phi_{s_t,2}$, the fourth panel plots the conditional standard deviations σ_{s_t} and the last panel plots the probabilities of regime switching and structural breaks.

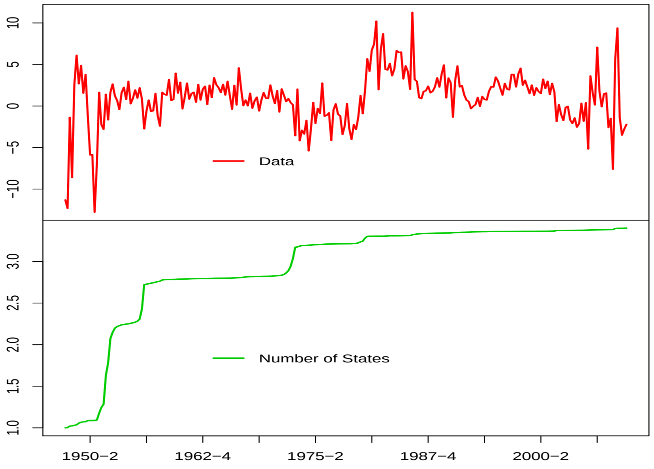

Figure 2: There are 252 observations from 1947Q1 to 2009Q4 for U.S. quarterly real interest rate. The data are estimated by the iHMM and each state has Gaussian AR(2) dynamics: $y_t = \phi_{s_t0} + \phi_{s_t1} y_{t-1} + \phi_{s_t2} y_{t-2} + \sigma_{s_t}\varepsilon_t$. The top panel plots the data and the bottom panel plots the posterior mean of the cumulative number of active states (active state means it has been visited at least once).

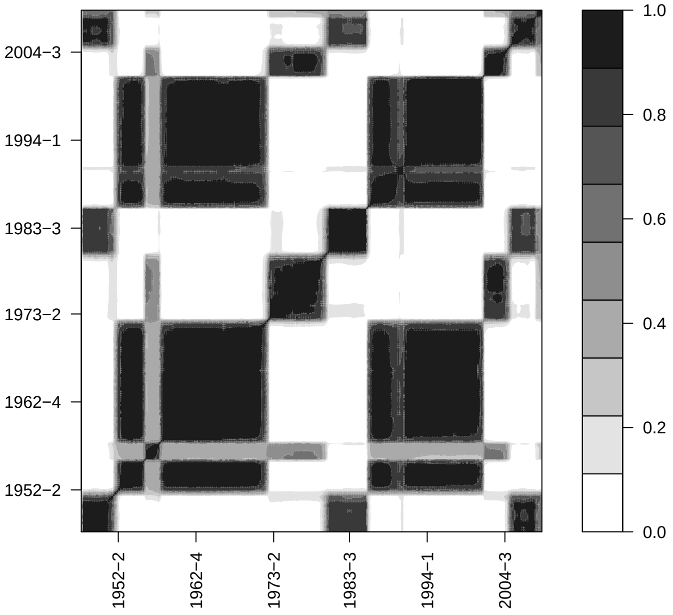

Figure 3: The temperature plot of the clustering of regimes by the iHMM. The data is U.S. quarterly real interest rates from 1947Q1 to 2009Q4.

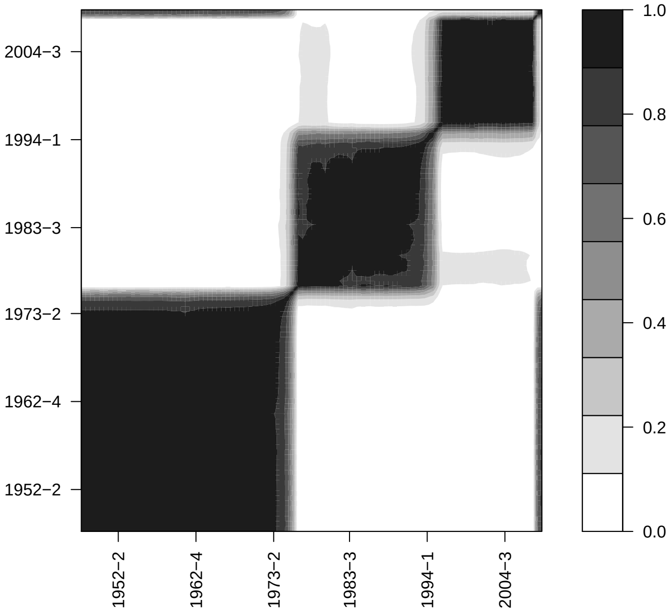

Figure 4: The temperature plot of the clustering of regimes by the iHMM. The data is U.S. quarterly real interest rates from 1961Q1 to 1986Q3, which is the same as in Garcia and Perron (1996).

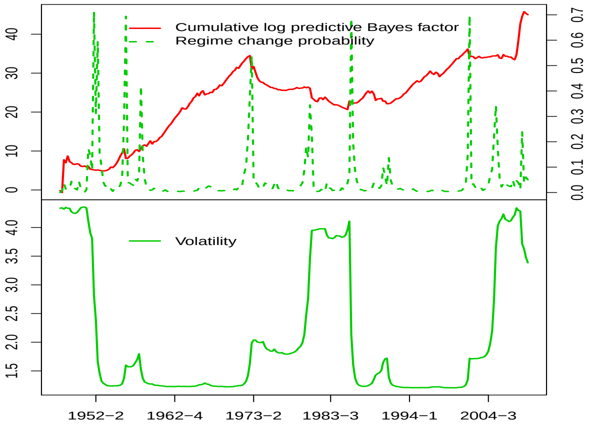

Figure 5: The top panel plots the cumulative log predictive Bayes factors(solid line). The dashed line represents the smoothed regime change probabilities implied by the iHMM. The bottom panel plots the posterior mean of the volatilities implied by the iHMM.