创业与管理学院

School of Entrepreneurship and Management

SHANGHAITECH SEM WORKING PAPER SERIES
No. 2020-003

An Infinite Hidden Markov Model for Short-term Interest Rates

John M. Maheu

DeGroote School of Business, McMaster University

Qiao Yang

ShanghaiTech University

May, 2016

https://ssrn.com/abstract=3521099

School of Entrepreneurship and Management
ShanghaiTech University
http://sem.shanghaitech.edu.cn

An Infinite Hidden Markov Model for Short-term Interest Rates*

John M. Maheu† Qiao Yang‡

May 2016

Abstract

The time-series dynamics of short-term interest rates are important as they are a key input into pricing models of the term structure of interest rates. In this paper we extend popular discrete time short-rate models to include Markov switching of infinite dimension. This is a Bayesian nonparametric model that allows for changes in the unknown conditional distribution over time. Applied to weekly U.S. data we find significant parameter change over time and strong evidence of non-Gaussian conditional distributions. Our new model with an hierarchical prior provides significant improvements in density forecasts as well as point forecasts. We find evidence of recurring regimes as well as structural breaks in the empirical application.

key words: hierarchical Dirichlet process prior, beam sampling, Markov switching, MCMC

JEL: C58, C14, C22, C11

*We are grateful for helpful comments from Yong Song and seminar participants at the University of Toronto. Maheu thanks the SSHRC for financial support.

†DeGroote School of Business, McMaster University, 1280 Main Street W., Hamilton, ON, Canada, L8S4M4 and University of Toronto, Canada and RCEA, Italy, maheujm@mcmaster.ca

‡Department of Economics, University of Toronto, bill.yang@mail.utoronto.ca

1 Introduction

Models of the term structure of interest rates are important in finance. They are used to price contingent claims, manage financial risk and assess the cost of capital. In most models the short-rate plays a very important role (Lhabitant et al. 2001, Musiela & Rutkowski 2005, Canto 2008). The time-series dynamics of the short-rate are important and difficult to model over long periods due to changes in monetary regimes and economic shocks. In this paper we extend the popular short-rate models to include Markov switching of infinite dimension. This is a Bayesian nonparametric model that allows for changes in the unknown conditional distribution over time. Applied to weekly data we find significant parameter change over time and strong evidence of non-Gaussian conditional distributions. Our new model with an hierarchical prior provides significant improvements in density forecasts as well as point forecasts.

Markov switching model have been used extensively to model interest rates. Early applications include Hamilton (1988), Albert & Chib (1993) and Garcia & Perron (1996). Markov switching and GARCH or stochastic volatility are combined by Cai (1994), Gray (1996) and Kalimipalli & Susmel (2004) to better capture volatility dynamics. However, Smith (2002) finds that stochastic volatility and Markov switching are substitutes with the latter being preferred. In related work Lanne & Saikkonen (2003) combine a mixture autoregressive process with time-varying transition probabilities and GARCH.

Ang & Bekaert (2002) show that a state dependent Markov switching model can capture the non-linearities in the drift and volatility function of the US short-rate. Evidence for nonlinear behaviour in the drift term is also found in Pesaran et al. (2006) using a model of structural change. In contrast, Durham (2003) finds no significant evidence of nonlinearity in the drift and concludes that volatility is the critical component. Guidolin & Timmermann (2009) use a four-state Markov switching model to capture the dynamics in US spot and forward rates. They improve point forecasts by combining forecasts of future spot rates with forecasts from time-series models or macroeconomic variables.

What is clear from this literature is that some form of regime switching is necessary to capture changes in the short-rate dynamics over time. Volatility clustering is important and simple two-state models are insufficient to deal with this. In addition, the papers that consider forecasting have focused on point forecasts and ignored density forecasts.

This paper contributes to this literature by designing an infinite hidden Markov model (IHMM) to capture the dynamics of U.S. short-term interest rate. IHMM can be thought of as a first-order Markov switching model with a countably infinite number of states. Given a finite dataset, the number of states is estimated along with all the other parameters. This is essentially a nonparametric model and part of our focus is to flexibly model the conditional distributions of the short-rate and investigate density forecasts. The unbounded nature of the transition matrix allows for both recurring states from the past as well as new states to capture structure change. An advantage

of this approach is that as new data arrives, if a new state is needed to capture new features of the conditional density, it is automatically introduced and incorporated into forecasts. These type of dynamic features cannot be captured by fixed state MS models.

The prior for the infinite transition matrix is a special case of the hierarchical Dirichlet process of Teh et al. (2006). Each row of the transition matrix is centered around a common draw from a top level Dirichlet process. This also aids in posterior simulation and centers the model around a standard Dirichlet process mixture model (Escobar & West 1995) which does not allow for time dependence. A *sticky* version of the infinite hidden Markov model which favours self transitions between states was introduced by Fox et al. (2011) and applied to inflation (Jochmann 2014, Song 2013), ARMA models (Carpantier & Dufays 2014) and conditional correlations (Dufays 2012). The sticky version is less attractive for financial data in which rapid switching between states is necessary to capture the unknown distribution as well as changes in this distribution over time.

An IHMM extension is applied to the Vasicek (1977) (VSK) model and the Cox et al. (1985) (CIR) model. Applied to weekly data from 1954 to 2014, on average, the model uses about 8 states to capture the unknown conditional distribution. Overall, the CIR specification performs the best while the VSK version requires a few more states on average to fit the data. There is evidence of states reoccurring from the past as well as new unique states being introduced over time. This is especially true from 2008 on as this interest rate regime is historically unique and represents a structural break.

We find evidence of parameter change in both the conditional mean as well as the conditional variance with the latter showing the largest moves. Predictive density plots display significant asymmetry and fat tails that are frequently multimodal. For instance, in 2009 the predictive density has local modes in the right tail. These account for the small probability of returning to a higher interest rates regime.

Consistent with Song (2013), adding a hierarchical prior for the data density parameters leads to gains in out-of-sample forecast accuracy. The model provides large gains in density forecasts compared to several finite state Markov switching models and a GARCH specification. All of the benchmark models we consider are strongly rejected by predictive Bayes factors in favour of the new nonparametric models. Point forecasts are competitive with existing models.

This paper is organized as follows. The next section discusses benchmark models used for model comparison and extension. Section 3 introduces the Dirichlet process, hierarchical Dirichlet process and the infinite hidden Markov model. Posterior sampling is discussed in Section 4 while empirical results are found in Section 5. The Appendix collects the details of posterior simulation.

2 Benchmark Models

Chan et al. (1992) show that the following specification for the short-term riskless rate

y_t nests many popular models

$$dy_t = (\lambda + \beta y_{t-1})dt + \sigma y_{t-1}^x dW_t. \tag{1}$$

dW_t is a Brownian motion and both the drift term $\lambda+\beta y_{t-1}$ and spot variance σy_{t-1}^x can be a function of the level of y_{t-1}. The discrete time version of the model, conditional on $y_{1:t-1} = \{y_1, \ldots, y_{t-1}\}$, is

$$\Delta y_t = \lambda + \beta y_{t-1} + \sigma y_{t-1}^x \epsilon_t, \quad \epsilon_t \sim N(0,1). \tag{2}$$

In the empirical application we will consider a rolling window version of this model and focus on $x = 0$ and $1/2$ which correspond to the Vasicek (1977) (VSK) model and the Cox et al. (1985) (CIR) model, respectively.

The next specification is a finite state Markov switching model. Let $s_t \in \{1, \ldots, K\}$ the model is

$$\begin{aligned} \Delta y_t &= \lambda_{s_t} + \beta_{s_t} y_{t-1} + \sigma_{s_t} y_{t-1}^x \epsilon_t, \quad \epsilon_t \sim N(0,1), & (3)\\ s_t | s_{t-1} &\sim P_{s_{t-1}}, & (4) \end{aligned}$$

where $P_{s_{t-1}}$ denotes row s_{t-1} of the $K \times K$ transition matrix P and is the discrete distribution governing the move from state s_{t-1} to s_t. This model can be estimated following Chib (1996). The VSK model ($x = 0$) is labelled as MS-K-VSK while the CIR version ($x = 1/2$) is MS-K-CIR. We also consider specifications with a hierarchical prior (MS-K-VSK-H,MS-K-CIR-H). This is discussed in more detail in the next section.

The final comparison model is a GARCH specification

$$\begin{aligned} \Delta y_t &= \lambda + \beta y_{t-1} + y_{t-1}^x \epsilon_t, & (5)\\ \epsilon_t &= \sigma_t z_t \quad z_t \sim N(0,1) & (6)\\ \sigma_t^2 &= \omega_0 + \omega_1 \epsilon_{t-1}^2 + \omega_2 \sigma_{t-1}^2, & (7) \end{aligned}$$

with $\omega_0 > 0$ and $\omega_1 \geq 0, \omega_2 \geq 0$. This gives the GARCH-VSK ($x = 0$) and the GARCH-CIR ($x = 1/1$) models.

3 Infinite Hidden Markov Model (IHMM)

An infinite Hidden Markov Model (IHMM) is a Bayesian nonparametric extension of the Markov-switching (MS) model. This builds on the Dirichlet process (DP), the hierarchical Dirichlet process (HDP) and their associated mixture models. We first discuss the Dirichlet process and the Dirichlet process mixture model before moving onto the HDP and the IHMM.

3.1 The Dirichlet Process Mixture Model

The Dirichlet process (DP) was introduced by Ferguson (1973) and is a distribution of probability measures over a measurable space Θ. $G \sim DP(\alpha, H)$ denotes a distribution G drawn from the DP with base measure H and $\alpha > 0$ a concentration parameter. A key property of the DP is for any finite partition $\{A_1, \ldots, A_K\}$ of Θ,

$$G(A_1), \ldots, G(A_K)|\alpha, H \sim \text{Dir}\Big(\alpha H(A_1), \ldots, \alpha H(A_K)\Big), \tag{8}$$

where $\text{Dir}\Big(\alpha H(A_1), \ldots, \alpha H(A_K)\Big)$ denotes a Dirichlet distribution with parameter vector $(\alpha H(A_1), \ldots, \alpha H(A_K))$. Therefore, by the properties of the Dirichlet distribution $E[G(A_i)] = H(A_i)$ and $\text{Var}(G(A_i)) = G(A_i)(1 - G(A_i))/(1 + \alpha)$.

A constructive definition of the DP is due to Sethuraman (1994) who defined a stick-breaking representation. The stick-breaking representation considers a unit-length stick that has been divided into multiple sub-sticks, where each sub-stick (π_k) is a random proportion (v_k) of the remaining stick. Let δ_{θ_k} denote a probability measure concentrated at θ_k then the stick-breaking construction of $G \sim DP(\alpha, H)$ is,

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i} \quad \text{where} \quad \theta_i \overset{iid}{\sim} H \quad i = 1, 2, \ldots, \infty \tag{9}$$

$$\pi_i = v_i \prod_{l=1}^{i-1} (1 - v_l), \quad v_i \overset{iid}{\sim} \text{Beta}(1, \alpha). \tag{10}$$

We denote the construction of the weights in (10) as $\{\pi_i\}_{i=1}^{\infty} \sim Stick(\alpha)$ and they form a distribution over the natural numbers.

The parameter α governs the distribution of the unit mass over the weights π_i. Large values of α spread the mass over many clusters θ_i while small values concentrate most of the mass on a few clusters. The DP is a distribution over a discrete probability measure, which guarantees that the probability measure G from equation (2) is a subset of base distribution H.

With this we can now define the Dirichlet process mixture (DPM) model for $y_t, t = 1, 2, \ldots,$ as

$$s_t|\alpha \sim Stick(\alpha) \tag{11}$$

$$y_t|s_t, \Theta \sim F(y_t|\theta_{s_t}) \tag{12}$$

where $\Theta = \{\theta_i\}_{i=1}^{\infty}$, $F(\cdot|\cdot)$ is the distribution of y_t given parameter θ_{s_t} and $Stick(\alpha)$ and θ_{s_t} are defined above. This is the basic model for Bayesian density estimation (Escobar & West 1995) and is appropriate for modeling an unconditional distribution. However, if there is time variation in the distribution of y_t this model is not suitable. We will introduce time-varying weights in (9) that result in s_t following a first-order Markov chain.

3.2 Hierarchical Dirichlet process

The hierarchical Dirichlet process (HDP) prior is introduced by Teh et al. (2006) as an extension to the DP prior. The HDP is a family of Dirichlet processes that share a common base measure which is also distributed according to a DP prior. Thus, the HDP has a hierarchical structure which is constructed by two DPs,

$$G_0|\eta, H \sim DP(\eta, H) \tag{13a}$$

$$G_j|\alpha, G_0 \overset{iid}{\sim} DP(\alpha, G_0), \quad j = 1, \ldots, \infty, \tag{13b}$$

where the process defines group-specific probability measures G_j conditional on a global probability measure G_0. α and η are concentration parameters and H is the base measure. Using the stick-breaking construction, we have the following representations of the HDP.

$$G_0 = \sum_{i=1}^{\infty} \gamma_i \delta_{\theta_i}, \quad \Gamma = \{\gamma_i\}_{i=1}^{\infty} \sim Stick(\eta), \quad \theta_i \overset{iid}{\sim} H, \tag{14}$$

$$G_j = \sum_{i=1}^{\infty} \pi_{ji} \delta_{\theta_i}, \quad \Pi_j = \{\pi_{ji}\}_{i=1}^{\infty} \overset{iid}{\sim} Stick2(\alpha, \Gamma), \tag{15}$$

where the weights of the latter measure G_j, are denoted as $Stick2(\alpha, \Gamma)$, and are constructed as

$$\pi_{ji} = \hat{\pi}_{ji} \prod_{l=1}^{i-1} (1 - \hat{\pi}_{jl}), \qquad \hat{\pi}_{ji} \overset{iid}{\sim} \text{Beta}\left(\alpha\gamma_i, \alpha\left(1 - \sum_{l=1}^{i} \gamma_l\right)\right), \tag{16}$$

$i = 1, 2, \ldots$. Note that all G_j shar the same atoms but have different weights, π_{ji}. Each G_j can serve as a prior for row j of the transition matrix an an infinite Markov chain. That is, if $s_t \in \{1, 2, \ldots\}$, then row j of the transition matrix is Π_j and directs the possible moves of $s_t = j$ to s_{t+1}. The model is completed with a conditional data distribution $F(y_t|\theta_{s_t})$.

3.3 Polya Urn Process for HDP

Before moving to the full specification of the model it is insightful to understand the conditional sampling distribution of the states in the IHMM induced by the HDP. First, we begin with the Polya urn process for the basic DP. In the DPM model of (11)-(12), the sequential sampling of the states s_t and their associated parameter θ_{s_t} obeys a Polya urn sampling scheme (Blackwell & MacQueen 1973). The Polya urn sampling scheme is obtained by integrating out $G \sim DP(\alpha, H)$. Let n_i be the current number of sampled states i (initially 0) and let the current number of different states be K. If $\delta_{i,j}$ denotes the Kronecker delta then draws from G are obtained as follows.

1. Set $s_1 = 1$, $K = 1$, $n_1 = 1$ and $\theta_1 \sim H$.

2. Given $s_{1:t-1}$ sample s_t as

$$s_t|s_{1:t-1} \sim \sum_{i=1}^{K} \frac{n_i}{\left(\alpha + \sum_j n_j\right)} \delta_{i,s_t} + \frac{\alpha}{\left(\alpha + \sum_j n_j\right)} \delta_{K+1,s_t}. \tag{17}$$

3. $n_{s_t} \leftarrow n_{s_t} + 1$ and if $s_t = K + 1$ draw $\theta_{s_t} \sim H$ and set $K \leftarrow K + 1$. Increment t and go to 2.

This process is equivalent to starting with an empty urn and placing a ball of colour θ_1 in and incrementing n_1. Thereafter, a ball of colour θ_i, $i = 1, \ldots, K$ is randomly drawn with probability $n_i/(\alpha + \sum_j n_j)$ and otherwise a new ball is drawn from $\theta_i \sim H$ with probability $\alpha/(\alpha + \sum_j n_j)$. Then if an existing ball was selected it is replaced in the urn along with another copy, while if a new colour ball is selected it is put in the urn and its count set to 1 (that is, n_i is incremented). This process is repeated. Note that states that are sampled frequently have a larger count n_i and reinforce their likelihood of being sampled in the future.

The Polya urn process for the infinite Markov chain model in Teh et al. (2006) is closely related and results from integrating out G_0 and G_j in (14) and (15). However, it involves a separate urn for each state (ball) sampled, as well as a top level *oracle* urn. Sampling states involves a two step process. First, we decide whether to sample from the existing urns (states) or to sample from the *oracle* urn. Sampling from an existing state is influenced by previous state counts and results in the recurrence of a past state. This is the next state in the Markov chain and we are done. On the other hand, if the *oracle* urn is selected it allows for previous states to be sampled, with probabilities related to previous counts from oracle draws, as well as new states. New states (balls) are only obtained from the oracle. Therefore, the next state in the Markov chain is either selected from the existing urns which contain pre-existing states or it is selected from the *oracle* urn in which case a previous state could be sampled or a new state sampled.

Let c_i denote the counts of the sampled states (balls) $i = 1, 2, \ldots,$ in the oracle urn and let n_{ij} be the current number of times state j was sampled conditional on being in urn i. Both c_i and n_{ij} are initialized to 0. New states s_t and the associated parameters θ_{s_t} are generated as follows.

1. Set $s_1 = 1$, $K = 1$, $n_1 = 1$, $c_1 = 1$ and $\theta_1 \sim H$.

2. Sample $\tilde{s}$ according to

$$\tilde{s}|s_{t-1} = i, s_{1:t-2} \sim \sum_{j=1}^{K} \frac{n_{ij}}{(\alpha + \sum_i n_{ij})} \delta_{j,\tilde{s}} + \frac{\alpha}{(\alpha + \sum_i n_{ij})} \delta_{K+1,\tilde{s}}. \tag{18}$$

3. If $\tilde{s} \leq K$ then $s_t = \tilde{s}$, $n_{is_t} \leftarrow n_{is_t} + 1$ and go to 4. If $\tilde{s} = K + 1$ then s_t is obtained by sampling from the oracle.

$$s_t|c_1, \ldots, c_K \sim \sum_{j=1}^{K} \frac{c_j}{(\eta + \sum_l c_l)} \delta_{j,s_t} + \frac{\eta}{(\eta + \sum_l c_l)} \delta_{K+1,s_t}, \tag{19}$$

$n_{is_t} \leftarrow n_{is_t} + 1$ and $c_{s_t} \leftarrow c_{s_t} + 1$. If $s_t = K + 1$, $\theta_{s_t} \sim H$ and $K \leftarrow K + 1$.

4. Increment t and go to 2.

State transitions are largely governed by the past count of transitions n_{ij} and α, except for when a transition is queried from the oracle. In this case, the past counts of states sampled from the oracle affects the choices. As in the DP case, states that have a larger number of transitions between themselves will reinforce this in future moves. Clearly, the parameters α and η play an important role in governing the likelihood of querying the oracle and the likelihood of new states being introduced over time. The Polya urn sampling scheme of the IHMM is important for setting priors on α and η but it also plays a critical role in posterior sampling for these parameters and other components in the model.

Larger values of α and η increase the likelihood of new states occurring. A small α but large η will favor the infrequent introduction of new states over time. Small values of α and η promote parsimony of states.

3.4 Infinite Hidden Markov Model

If $s_t \in \{1, 2, 3, \dots\}$ is the unobserved state variable and $\Pi_j = (\pi_{j1}, \pi_{j2}, \dots)$ is the jth row of the transition matrix, then π_{ji} becomes the prior probability of s_t moving from state j to i at time $t+1$. The infinite hidden Markov (IHMM) model, a time dependent version of (11)-(12), can be written as

$$\Gamma|\eta \sim Stick(\eta) \qquad \theta_i \overset{iid}{\sim} H \quad i = 1, 2, \dots, \tag{20a}$$

$$\Pi_j|\alpha, \Gamma \overset{iid}{\sim} Stick2(\alpha, \Gamma), \qquad j = 1, 2, \dots, \tag{20b}$$

$$s_t|s_{t-1}, \Pi_{s_{t-1}} \sim \Pi_{s_{t-1}}, \quad t = 1, \dots, T \tag{20c}$$

$$y_t|s_t, \Theta \sim F(y_t|\theta_{s_t}). \tag{20d}$$

Each row Π_j, of the transition matrix Π is assumed to be drawn from a DP prior with a common base measure $\Gamma = (\gamma_1, \gamma_2, \dots)$ and precision α. As a result, it can be shown that $E[\pi_{ji}] = E[\gamma_i] = \eta^{i-1}/(1+\eta)^i$. In other words, the prior on the transition matrix is centered around equal values of row probabilities. This corresponds to the DPM model in (11)-(12) and is a natural starting point for inference.

The parameters η and α play an important role in the distribution of the weights. Different combinations of η and α can be used to enforce various prior beliefs about the Markov chain. As discussed in the last section, larger values of η favour more active states while larger values of α allow for the consideration of new states more often. Rather that set these to specific values we place a prior on them and estimate them along with all other model parameters.

3.5 IHMM with Hierarchical Prior

When new states are introduced in the model there may be benefits to learning about the distributional features of the data density parameters, $\Theta = \{\theta_1, \theta_2, \dots\}$ if they share a common distribution. To allow the prior to be centered and concentrated in empirically important regions of the parameter space we introduce a hierarchical prior for H.

In the empirical work we will focus our attention on the following extension of the CIR ($x = 1/2$) and VSK ($x = 0$) to the infinite hidden Markov model with hierarchical prior (IHMM-CIR-H and IHMM-VSK-H).

$$\Gamma|\eta \sim Stick(\eta), \qquad \theta_i \overset{iid}{\sim} H(\xi), \qquad \xi \sim Q, \quad i = 1, 2, \dots, \tag{21a}$$

$$\Pi_j|\alpha, \Gamma \overset{iid}{\sim} Stick2(\alpha, \Gamma), \qquad j = 1, 2, \dots, \tag{21b}$$

$$s_t|s_{t-1}, \Pi_{s_{t-1}} \sim \Pi_{s_{t-1}}, \quad t = 1, \dots, T \tag{21c}$$

$$\Delta y_t|Y_{t-1}, s_t, \Theta \sim N(\lambda_{s_t} + \beta_{s_t} y_{t-1}, \sigma^2_{s_t} y^x_{t-1}), \tag{21d}$$

where $\theta_i = (\lambda_i, \beta_i, \sigma_i)$. The hierarchical priors for η and α are,

$$\eta \sim \text{Gamma}(a_1, b_1) \quad \alpha \sim \text{Gamma}(a_2, b_2) \tag{22}$$

where $\text{Gamma}(a, b)$ denotes a gamma distribution with mean a/b. Given ξ the prior for θ_i is $H(\xi)$ and if $\vartheta_i = (\lambda_i, \beta_i)$ it follows

$$\vartheta_i \sim N(\phi, B) \qquad \sigma_i^{-2} \sim \text{Gamma}(\chi, \nu). \tag{23}$$

Finally, the prior for $\xi = (\phi, B, \chi, \nu)$ follows

$$\boldsymbol{\phi} \sim N(h_0, H_0) \quad B^{-1} \sim W(a_0, A_0) \quad \chi \sim \text{Exp}(\rho_0) \quad \nu \sim Gamma(c_0, d_0), \tag{24}$$

where $W(a_0, A_0)$ is a Wishart distribution with scale matrix A_0, degree of freedom parameter a_0 and mean of $a_0 A_0$. $\text{Exp}(\rho_0)$ is the exponential distribution with parameter ρ_0. According to this parametrization $E(\sigma^{-2}) = \frac{\chi}{\nu}$ and $E(\chi) = \rho_0$. ϕ and h_0 are 2×1 vector. H_0 and A_0 are 2×2 matrices.

4 Posterior Sampling

If the IHMM model was replaced with a finite state hidden Markov model much of the posterior sampling complexity would be eliminated and existing methods of sampling such as Chib (1996) could be used. The beam sampler of Van Gael et al. (2008) is designed to exactly achieve this. It is a stochastic truncation that reduces the infinite state space to a finite one and allows for the forward-filter backward sampler (FFBS) of Chib (1996) to be applied.

The idea behind beam sampling is closely related to slice sampling for DPM models (Walker 2007) and involves the introduction of latent variables u_t, $t = 1, \ldots, T$ such that the conditional density of u_t is

$$p(u_t|s_{t-1}, s_t, \Pi) = \frac{\mathbb{1}(0 < u_t < \pi_{s_{t-1},s_t})}{\pi_{s_{t-1},s_t}}. \tag{25}$$

The u_t are sampled along with the other parameters but the sampling of the states given $u_{1:t}$ in the filter step of the FFBS becomes,

$$\begin{aligned} p(s_t|y_{1:t}, u_{1:t}, \Pi) &\propto p(y_t|y_{1:t-1}, s_t) \sum_{s_{t-1}=1}^{\infty} \mathbb{1}(u_t < \pi_{s_{t-1},s_t}) p(s_{t-1}|y_{1:t-1}, u_{1:t-1}, \Pi) \quad (26) \\ &\propto p(y_t|y_{1:t-1}, s_t) \sum_{s_{t-1}:u_t<\pi_{s_{t-1},s_t}} p(s_{t-1}|y_{1:t-1}, u_{1:t-1}, \Pi). \quad (27) \end{aligned}$$

The u_t slices out states with small π_{s_{t-1},s_t} and results in a finite summation in (27) of dimension K, since the number of states s_{t-1} that satisfy $u_t < \pi_{s_{t-1},s_t}$ is finite. This turns the infinite summation into a finite one. Once the forward pass is computed for $t = 1, \ldots, T$ the backward pass follows from,

$$p(s_t|s_{t+1}, y_{1:T}, u_{1:T}) \propto p(s_t|y_{1:t}, u_{1:t}) \mathbb{1}(u_{t+1} < \pi_{s_t,s_{t+1}}), \qquad t = T-1, \ldots, 1. \tag{28}$$

This is initiated with a draw of s_T from the last value of the filter $p(s_T|y_{1:T}, u_{1:T}, \Pi)$.

In each MCMC iteration the slice sampler effectively truncates the system to a dimension of size K. We order each of the states that receive a non-zero weight as the first K states and keep track of the $K+1$ state as the residual probability. As such, posterior sampling is over the quantities $\Gamma = (\gamma_1, \ldots, \gamma_K, \gamma^r_{K+1})$ with $\gamma^r_{K+1} = \sum_{l=K+1}^{\infty} \gamma_l$ and $\Pi_j = (\pi_{j1}, \ldots, \pi_{jK}, \pi^r_{jK+1})$, $j = 1, \ldots, K$, where $\pi^r_{jK+1} = \sum_{l=K+1}^{\infty} \pi_{jl}$.

After initializing the parameters the full MCMC routine involves the following steps:

1. sample $s_{1:T}|y_{1:T}, u_{1:T}, \Pi$
2. sample $\Pi_j|s_{1:T}, \Gamma$, $j = 1, \ldots, K$.
3. sample $u_{1:T}|s_{1:T}, \Pi$ and update K
4. sample $\theta_j|s_{1:T}, y_{1:T}, \xi$, $j = 1, \ldots, K$
5. sample $\Gamma|s_{1:T}, \eta$
6. sample $\xi|\theta_1, \ldots, \theta_K$, $\eta|s_{1:T}, \Gamma$ and $\alpha|s_{1:T}, \Gamma$.

Iterating over these sampling steps gives one draw from the posterior denoted as $\Omega = \{\Gamma, \Pi, K, \Theta, s_{1:T}, \xi, \eta, \alpha\}$. Full details of each of the posterior sampling steps can be found in the Appendix. After a suitable burn-in period a large number of draws are collected from which features of the posterior and predictive densities can be estimated.

For example, given N posterior draws of $\{\theta_{s_t}^{(i)}\}_{i=1}^N$ then a simulation consistent estimate of $E[\theta_{s_t}|y_{1:T}]$ is $\frac{1}{N}\sum_{i=1}^N \theta_{s_t}^{(i)}$. For full sample estimates we drop the first 80,000 draws and collect the next 100,000 for posterior inference. For computing predictive likelihoods and predictive means sequentially for the out-of-sample period we use a burn-in of 5000 and use the next 20,000 for posterior/predictive inference.

4.1 Predictive Density

Given a sequence of posterior draws, $\{\Omega^{(i)}\}_{i=1}^N$, from the IHMM models using T observations, the predictive density can be computed at Δy_{T+1} (or y_{T+1}) as follows.

1. For each i, randomly draw a state s_{T+1} according to the multinomial distribution Π_{s_T}.
2. If $s_{T+1} \leq K^{(i)}$ set $(\lambda^{(i)}, \beta^{(i)}, \sigma^{2(i)}) \equiv \theta_{s_{T+1}}$ and otherwise set $(\lambda^{(i)}, \beta^{(i)}, \sigma^{2(i)}) \equiv \theta$, where $\theta \sim H(\xi^{(i)})$.

The predictive density for y_{T+1} is estimated as

$$p(y_{T+1}|y_{1:T}) \approx \frac{1}{N}\sum_{i=1}^N N(y_{T+1}|\lambda^{(i)} + (1+\beta^{(i)})y_T, \sigma^{(i)2} y_T^x), \tag{29}$$

where $N(\cdot|\cdot,\cdot)$ is the normal probability density function. Similarly, the predictive mean of y_{T+1} can be estimated as

$$E[y_{T+1}|y_{1:T}] \approx \frac{1}{N}\sum_{i=1}^N (\lambda^{(i)} + (1+\beta^{(i)})y_T). \tag{30}$$

The predictive density and prediction means can be used to compare and evaluate the performance of several models in the out-of-sample period.

5 Application to Short Term T-Bill Rate

5.1 Data

The data is the 3-month T-bill of secondary market obtained from the Board of Governors of the Federal Reserve System. The data is weekly, Friday to Friday from Jan-15-1954 to Mar-28-2014 (3142 observations). Figure 1 displays the data while Table 1 reports summary statistics.

5.2 Model Priors

For the IHMM-CIR-H and IHMM-VSK-H models h_0 is set to a vector of zeros, H_0 and A_0 to an identity matrix, $a_0 = 2$, $\rho_0 = 1$, $c_0 = 5$, and $d_0 = 1$. For the precision parameters η and α the hyperparameters are $a_1 = a_2 = 5$ and $b_1 = b_2 = 1$. The first 3 columns of Table 2 summarize the prior. In some cases the 0.9 density intervals are obtained by simulation.

For the CIR, VSK, the MS-K-CIR, MS-K-VSK and GARCH-CIR and GARCH-VSK the priors are $\lambda \sim N(0, 25)$, $\beta \sim N(0, 1)$, $\sigma^2 \sim IG(2, 0.5)$. For the MS models, each row of the transition matrix has the prior $Dir(1, \ldots, 1)$ where K is the fixed number of states. We also consider MS models with the identical hierarchical prior of the IHMM-CIR-H and IHMM-VSK-H specifications over θ and ξ. Finally, for the GARCH-CIR and GARCH-VSK models all priors are assumed to be independent $N(0, 1)$ with the restrictions $\omega_0 > 0, \omega_1 \geq 0$, and $\omega_2 \geq 0$ imposed.

5.3 Posterior Analysis

Table 2 reports posterior summary statistics for the IHMM-CIR-H and IHMM-VSK-H models. Compared to the prior (columns 2 and 3) there is significant learning in all the parameters as the density intervals change in location and their length decreases. On average, the two models are using 8 components in the mixture to capture the shape and changes in the distribution. Figure 2 is a histogram of the sampled active states K. Although the distribution is concentrated around 8 there is uncertainty in both models. Forecasts from this model automatically incorporate this regime uncertainty.

Figures 3 and 4 display the posterior mean of the model parameters subject to regime change ($E[\lambda_{s_t}|y_{1:T}]$, $E[\beta_{s_t}|y_{1:T}]$ and $E[\sigma_{s_t}|y_{1:T}]$). There is considerable time variation in all parameters. A fixed parameter model would have difficulty fitting such a long time period with these parameter shifts. Some of the state changes appear to return to previous values in the time history while some appear to be new and unique. This is most apparent for the VSK model in Figure 4 where somewhere in 2008 the model displays a combination of parameter estimates that are unique. This is like a structure break in that the model identifies dynamics for the short term interest rates that differ from all past states. The evidence for a structural changes is much less for the CIR specification.

The second last panel in these figures displays $E[\beta_{s_t} > 0|y_{1:T}]$ over time. This corresponds to the probability of an explosive regime. This probability is non-negligible over the sample period for both models. In the case of the VSK model, Figure 4 shows that it is almost certainly explosive from 2008 and onward.

The final panel of the plot displays the posterior evidence for state changes at each point in the sample. There are a few periods where states persist but generally there is regular changes in states. This is expected if the Gaussian assumption for the data density does not fit the data as the model will resort to mixture of the Gaussian densities to approximate the unknown distribution.

Figures 5 and 6 display a heat map for the latent states s_t, $t = 1, \ldots, T$ for IHMM-CIR-H and IHMM-VSK-H. A heat map is an estimate of $P(s_i = s_j | y_{1:T})$ and reported in a table in which colour differences denote different probabilities over the range of $i = 1, \ldots, T$ and $j = 1, \ldots, T$. High probability values (light colour) on the main diagonal indicate new regimes that are unique to that time period. Many of these occur for both models and indicate a relatively high number of states. On the other hand, high probability values (light colour) off the main diagonal indicate regimes that reoccur historically. Both models display significant periods of unique regimes which is indicative of structure change, however, it is more pronounced in the VSK version. What these figures show is that there are new states being introduced into the model over time and there are frequent regimes changes back to previous states. These type of dynamic features cannot be captured by fixed state MS models.

5.4 Out-of-Sample Analysis

The main out-of-sample results are based on the sample 1955-Dec-09 to 2014-Mar-28 (3043 observations). At each point t in the sample, the models are estimated using data up to t and model forecasts are computed for y_{t+1}. To compare models we focus on log-predictive likelihood values and root-mean squared forecast errors based on predictive means.

The predictive likelihood of y_{t+1} for the infinite hidden Markov models are estimated by plugging in the data y_{t+1} into (29). Similarly, predictive likelihoods are estimated for other models. We report $LPL_M = \sum_{i=\tau_1}^{\tau_2} \log p(y_i | y_{1:i-1}, M)$ for model M. Several models can be compared by log-predictive Bayes factors. The log-predictive Bayes factors for model A against B is $LPL_A - LPL_B$ where positive values favour A and values in excess of 5 are considered strong support for A.

Forecast performance is reported in Table 3. Two sample periods are included, along with various benchmark comparison models. Log-predictive Bayes factors can be computed by subtracting any two log-predictive values in the table. The best model in terms of LPL_M, for the larger out-of-sample period is the IHMM-CIR-H. The log-predictive Bayes factor against the next best model (IHMM-VSK-H) is 44 representing a substantial improvement. The evidence for the IHMM-CIR-H against the best parametric model (MS-3-CIR-H) is also strong with a log-Bayes factor of 52. The RMSE of the IHMM-CIR-H is also the best but the gains over the benchmark models are modest. It is 0.7% lower than the MS-3-CIR-H model.

The second portion of Table 3 record the forecast performance for a shorter more recent sample. Consistent with the previous results the quality of density forecasts are superior for the IHMM-CIR-H specification while the best point forecasts come from IHMM-VSK model, but again, gains are small in the RMSE.

Figures 7 and 8 display the cumulative log-predictive likelihoods for various models at each point in the out-of-sample period. All the specifications have difficulty with high volatility episodes. The IHHM hierarchical versions recover the fastest. The IHMM-CIR-H and IHMM-VSK-H models provide significant gains in density forecasts

throughout the sample. The dominance of these models is not confined to any particular period or outliers but involves steady ongoing gains over time.

Closely related to this is the estimate of the number of effective states (posterior mean of K) at each point in the out-of-sample period. Figures 9 and 10 display these estimates. It shows a regular increase in the dimension of the model over time. However, there are periods in which the number of states level off or even drop (just before 97-July, Figure 9). The VSK version of the model requires more discrete jumps in the number of states to deal with heteroskedasticity. This can be seen after 72-Aug and after 2007 in Figure 10.

Finally, to see why the IHMM-CIR-H performs so well in density forecasts Figure 11 produces the predictive density for this model along with a rolling window version of the basic CIR model. For each of the selected dates the density and log-density are displayed. It is clear that the IHMM-CIR-H is often very different than the parametric specification. The model displays fat tails and asymmetry. The log-density plots show the IHMM-CIR-H model to have several modes in the right tail. This is expected as the model allows for the possibility of moving from low interest rate regimes back to higher interest rate regimes.

5.5 Robustness

Table 4 reports the out-of-sample results for the shorter period for both IHMM models and various prior configurations. In general, the top panel presents results for priors that have an increased variance but centered in the same location. The second panel reports the impact from more concentrated priors on η and α. These favor fewer clusters in the model.

Different priors do lead to differences in the log-predictive likelihoods. However, the IHMM-CIR-H specification continues to be the best model for density forecasts and is always significantly better than the best performing benchmark model MS-3-CIR-H. For instance, the MS-3-CIR-H model has a log-predictive likelihood of 1306.158 while the least favorable prior gives the IHMM-CIR-H model a log-predictive likelihood value of 1324.724.

The density forecasts for the IHMM-VSK-H show a bit less variation. The RMSE for both models is very robust to changes in the prior. In summary, the new nonparametric models continue to dominate the benchmarks for different prior configurations.

5.6 Sticky Prior

According to the forecast performance, the IHMM-CIR-H is the best performed model. The sticky version will be based on this model. The sticky IHMM for CIR with hierarchical priors is denoted as Sticky-IHMM-CIR-H. The equation (21b) is replaced by the following,

$$\Pi_j|\alpha,\Gamma \overset{iid}{\sim} Stick2\big(\alpha+\kappa, \frac{\alpha\Gamma+\kappa\delta_j}{\alpha+\kappa}\big), \qquad j=1,2,\ldots. \tag{31}$$

The $\delta_j = 1$ when $j = i$ is the self state. Otherwise, $\delta_j = 0$. The α is be sampled together with κ and a new prior p will be sampled as well. For example:

$$\alpha + \kappa \sim Gamma(a_2, b_2) \qquad p \sim Beta(a_3, b_3) \tag{32}$$

We evaluate two set of priors of the IHMM-CIR-H-S for forecasting performance. The loose set of prior is let h_0 is set to a vector of zeros, H_0 and A_0 to an identity matrix, $a_0 = 2$, $\rho_0 = 1$, $c_0 = 5$, and $d_0 = 1$. For the precision parameters of η, $\alpha+\kappa$ and p, the hyperparameters are $a_1 = a_2 = a_3 = 5$ and $b_1 = b_2 = b_3 = 1$. The less loose set of prior is set h_0 to a vector of zeros, H_0 and A_0 to an identity matrix, $a_0 = 3$, $\rho_0 = 1$, $c_0 = 2$, and $d_0 = 1$. For the precision parameters of η, $\alpha+\kappa$ and p, the hyperparameters are $a_1 = a_2 = a_3 = 2$ and $b_1 = b_2 = b_3 = 1$. The log-predictive likelihood for the less loose set of priors is 2863.519 and the RMSE is 0.1928. The log-predictive likelihood for the loose set of priors is 2858.66 and RMSE is 0.1929. The out-of-sample size is from 1955-Dec-09 to 2014-Mar-28 (3043 observations).

6 Conclusion

This paper extends popular discrete time short interest rate models to include Markov switching of infinite dimension. This is a Bayesian nonparametric model that allows for changes in the unknown conditional distribution over time. Applied to weekly U.S. data we find significant parameter change over time and strong evidence of non-Gaussian conditional distributions. Our new model with an hierarchical prior provides significant improvements in density forecasts as well as point forecasts. The empirical study finds recurring regimes as well as unique regimes (structural breaks) in the US short-term rate.

7 Appendix

Several of the sampling steps are based on Teh et al. (2006), Van Gael et al. (2008) and Fox et al. (2011).

7.1 Sampler Steps

Recall the notation: full sample of $y_{1:T}$, state variables $s_{1:T}$, $\Gamma = (\gamma_1, \ldots, \gamma_K, \gamma^r_{K+1})$ with $\gamma^r_{K+1} = \sum_{l=K+1}^{\infty} \gamma_l$, $\Pi_j = (\pi_{j1}, \ldots, \pi_{jK}, \pi^r_{jK+1})$, $j = 1, \ldots, K$, where $\pi^r_{jK+1} = \sum_{l=K+1}^{\infty} \pi_{jl}$ and $\theta = \{\lambda, \beta, \sigma\}$.

1. Sample $u_{1:T}$: $u_1 \sim U(0, \gamma_{s_1})$ and $u_t \sim U(0, \pi_{s_{t-1}s_t})$, $t = 2, \ldots, T$

2. Expand Π and K: If

$$\max\left\{\pi^r_{1K+1}, \ldots, \pi^r_{KK+1}\right\} > \min\left\{u_1, \ldots, u_T\right\}, \tag{33}$$

   then repeat the following steps:

   (a) Increment $K \leftarrow K + 1$ and $\theta_K \sim H(\xi)$,

   (b) Update γ. Note that with the increment in K, γ_K is the old residual probability which is broken as

$$\tau \sim \text{Beta}(1, \eta), \quad \gamma^r_{K+1} \leftarrow (1 - \tau)\gamma_K, \quad \gamma_K \leftarrow \tau\gamma_K.$$

   (c) $\Pi_K = (\pi_{K,1}, \ldots, \pi^r_{KK+1}) \sim \text{Dir}(\alpha\gamma_1, \ldots, \alpha\gamma^r_{K+1})$.

   (d) Update Π_j, $j = 1, \ldots, K$ as

$$\tau_j \sim \text{Beta}\left(\alpha\gamma_K, \alpha\gamma^r_{K+1}\right), \quad \pi^r_{jK+1} \leftarrow (1 - \tau_j)\pi_{jK}, \quad \pi_{jK} \leftarrow \tau_j\pi_{jK}$$

   Steps a) – d) are repeated until (33) does not hold. This process expands Π until the possibility of new state is too small to consider in the FFBS steps next.

3. Forward filter of $s_{1:T}$. Let $f(.|.)$ be the density function of y_t.

   (a) Initial forecast step for s_1:

$$p(s_1 = k|u_{1:T}, \Gamma, \Theta) \propto \mathbb{1}(u_1 < \gamma_k), \quad \text{for } k = 1, \ldots, K$$

   (b) The updating step for $s_{1:T}$: and $k = 1, \ldots, K$.

$$p(s_t = k|y_{1:t}, u_{1:T}, \Pi, \Theta) \propto f(y_t|y_{1:t-1}, \theta_k)p(s_t = k|y_{1:t-1}, u_{1:T}, \Pi, \Theta)$$

   (c) The forecasting step for $k = 1, \ldots, K$

$$p(s_{t+1} = k|y_{1:t}, u_{1:T}, \Pi, \Theta) \propto \sum_{l=1}^{K} \mathbb{1}(u_{t+1} < \pi_{lk})p(s_t = l|y_{1:t}, u_{1:t}, \Pi, \Theta)$$

(d) Iterate step b and c until $t = T$.

4. Backward Sampler for $s_{1:T}$:

   (a) Sample the initial state of s_T from $p(s_T|y_{1:T}, u_{1:T}, \Pi, \Theta)$.

   (b) Sample s_t, $t = T-1, T-2, \ldots, 1$ recursively from

$$P(s_t = i|s_{t+1} = j, y_{1:t}, u_{1:T}, \Pi, \Theta) \propto \mathbb{1}(\pi_{ij} < u_{t+1}) p(s_t = i|y_{1:t}, u_{1:T}, \Pi, \Theta).$$

   After this step we perform a cleanup step to remove any empty states (states with no observations allocated to them). This results in a redefined K and transition matrix Π, and other model parameters are adjusted accordingly so that states $1, \ldots, K$ are active states.

5. Sample $c_{1:K}$: Recall c_i is number of counts of state i sampled from the oracle. It is an auxiliary variable that facilities several sampling steps below. Let o_{ji} be the number of oracle draws out of the total transitions n_{ji}, of state j to state i. Following Fox et al. (2011) we simulate the sequence of o_{ji} to obtain a draw of c_i. For each $i = 1 \ldots, K$, $j = 1, \ldots, K$, do the following steps.

   (a) set $o_{ji} = 0$

   (b) draw $x_l \sim \text{Bernoulli}\big(\frac{\alpha\gamma_i}{l-1+\alpha\gamma_i}\big)$, for $l = 1, \ldots, n_{ji}$: if $x_l = 1$ increment o_{ji}.

   Then $c_i = \sum_{j=1}^{K} o_{ji}$.

6. Sample η: Two auxiliary variables $\underline{\nu}$ and $\underline{\lambda}$ are introduced to sample η. We apply a Gibbs sampler to sample $\underline{\nu}$ and $\underline{\lambda}$ conditional on $c = \sum_{i=1}^{K} c_i$ and previous η first, then sample η through a gamma distribution.

   (a) $\underline{\nu} \sim \text{Bernoulli}\big(\frac{c}{c+\eta}\big)$

   (b) $\underline{\lambda} \sim \text{Beta}\big(\eta + 1, c\big)$

   (c) $\eta \sim \text{Gamma}(a_1 + \overline{K} - \underline{\nu}, b_1 - log\underline{\lambda})$

7. Sample α: The derivations follows directly from Fox et al. (2011).

   (a) $\overline{\nu}_j \sim \text{Bernoulli}(\frac{n_{j.}}{n_{j.}+\alpha})$ for $j = 1, \ldots, K$. $\quad n_{j.} = \sum_{i=1}^{K} n_{ji}$

   (b) $\overline{\lambda}_j \sim \text{Beta}(\alpha + 1, n_{j.})$ for $j = 1, \ldots, K$.

   (c) $\alpha \sim \text{Gamma}\big(a_2 + c - \sum_{j=1}^{K} \overline{\nu}_j, b_2 - \sum_{j=1}^{K} log(\overline{\lambda}_j)\big)$

8. Sample Γ: Given the counts $c_{1:K}$, which are from a sample from Γ, conjugacy gives the update for Γ as

$$\Gamma|c_{1:K}, \eta \sim Dir(c_1, c_2, \ldots, c_K, \eta).$$

9. Sample Π_j: for $j = 1, \ldots, K$, given $(n_{j1}, \ldots, n_{jK})$ and the conjugate property of the Dirichlet distribution,

$$\Pi_j | \alpha, n_{j1:K} \sim Dir(\alpha\gamma_1 + n_{j1}, \ldots, \alpha\gamma_K + n_{jK}, \alpha\gamma_{K+1}^r).$$

10. Sample $\theta_{1:K}$: Let $\theta_k = (\vartheta_k, \sigma_k)$ and $\vartheta_k = (\lambda_k, \beta_k)^T$, for k=1.. . . ,K:

    Let $\hat{Y}_k \equiv \{\Delta y_t | s_t = k\}$, $\hat{X}_k \equiv \{(1, y_{t-1}) | s_t = k\}$ and $T_k = \{\#t | s_t = k\}$. Therefore, $\hat{Y}_k$ is a vector with dimension $T_k \times 1$ and $\hat{X}_k$ is a matrix with dimension $T_k \times 2$. This gives the linear model

$$\hat{Y}_k = \hat{X}_k \vartheta_k + u \quad u \sim N(0, \sigma_k^2 I)$$

    with the following Gibbs sampling steps

$$\vartheta_k \sim N\Big(V(\frac{1}{\sigma^2}\hat{X}_k^T\hat{Y}_k + \phi B^{-1}), V\Big), \quad V = \Big(\frac{1}{\sigma^2}\hat{X}_k^T\hat{X}_k + B^{-1}\Big)^{-1}$$

$$\sigma_k^{-2} \sim \text{Gamma}\left(\frac{T + 2\chi}{2}, \frac{(\hat{Y}_k - \hat{X}_k\vartheta_k)^T(\hat{Y}_k - \hat{X}_k\vartheta_k) + 2\nu}{2}\right).$$

11. Sample ξ, hierarchical priors, which are ϕ, B, χ, ν. $H(\xi)$ is defined as follows,

$$\vartheta \sim N(\phi, B) \qquad \sigma^{-2} \sim \text{Gamma}(\chi, \nu).$$

    Q is defined as follows,

$$\phi \sim N(h_0, H_0) \quad B^{-1} \sim W(a_0, A_0) \quad \chi \sim \text{Exp}(\rho_0) \quad \nu \sim \text{Gamma}(c_0, d_0).$$

    (a) Sample $\phi \,| B, h_0, H_0, \vartheta_{1:K} \sim N(\mu_\phi, \Sigma_\phi)$ where

$$\overline{\vartheta} = \frac{1}{K}\sum_{j=1}^{K}\vartheta_j, \ \mu_\phi = \Sigma_\phi\Big(H_0^{-1}h_0 + KB^{-1}\overline{\vartheta}\Big), \ \Sigma_\phi = \Big(H_0^{-1} + KB^{-1}\Big)^{-1}.$$

    (b) Sample $B^{-1} | \phi, a_0, A_0, \vartheta_{1:K} \sim \text{Wishart}(\omega_B, \Omega_B)$ where,

$$\omega_B = K + \alpha_0 \quad and \quad \Omega_B = \Big(A_0^{-1} + \sum_{j=1}^{K}(\vartheta_j - \phi)(\vartheta_j - \phi)^T\Big)^{-1}$$

    (c) Sample $\nu | \chi, c_0, d_0, \sigma_{1:K}^{-2} \sim \text{Gamma}\Big(c_0 + K\chi, d_0 + \sum_{j=1}^{K}\sigma_j^{-2}\Big)$

    (d) Sample $\chi | \nu, \rho_0, \sigma_{1:K}^{-2}$. There is no conjugate prior so we apply a Metropolis-Hastings step. The conditional posterior is

$$\pi(\chi | \nu, \rho_0, \sigma_{1:K}^{-2}) = \text{Exp}(\chi | \rho_0)\prod_{j=1}^{K} G(\sigma_j^{-2} | \chi, \nu).$$

The proposal is

$$q(\chi^{new}|\chi^{old}) \sim \text{Gamma}(\zeta, \zeta/\chi^{old})$$

and we choose ζ so the rejection rate is between 0.3 and 0.6. The proposal χ^{new}, is accepted with probability

$$\min\left[\frac{\pi(\chi^{new}|\nu,\rho_0,\sigma_{1:K}^{-2})/q(\chi^{new}|\chi^{old})}{\pi(\chi^{old}|\nu,\rho_0,\sigma_{1:K}^{-2})/q(\chi^{old}|\chi^{new})},1\right].$$

12. Repeat 1-11.

7.2 Sticky IHMM-CIR-H

All the sampler steps are the same as IHMM-CIR-H and IHMM-VSK-H except the sampling step of c, η, α, Γ and $\Pi_{1:K}$ are replaced by the following steps:

1. Sample $c_{1:K}$, $\overline{c}_{1:K}$ and p:

   Recall c_i is number of counts of state i sampled from the oracle. It is an auxiliary variable that facilities several sampling steps below. Let o_{ji} be the number of oracle draws out of the total transitions n_{ji}, of state j to state i. Following Fox et al. (2011) we simulate the sequence of o_{ji} to obtain a draw of c_i. For each $i = 1\ldots, K$, $j = 1, \ldots, K$, do the following steps.

   (a) set $o_{ji} = 0$

   (b) draw $x_l \sim \text{Bernoulli}\big(\frac{\alpha\gamma_i+\kappa I(j=i)}{l-1+\alpha\gamma_i+\kappa I(j=i)}\big)$, for $l = 1, \ldots, n_{ji}$: if $x_l = 1$ increment o_{ji}.

   (c) $w_j = \text{Binomial}\big(o_{ji}, \frac{p}{p+\gamma_j(1-p)}\big)$, $p = \frac{\kappa}{\alpha+\kappa}$ and $j = i$.

   (d) $\overline{o}_{ji} = o_{ji}$ if $j \neq i$, and $\overline{o}_{ji} = o_{ji} - w_j$ if $j = i$.

   (e) $c_i = \sum_{j=1}^K o_{ji}$ and $\overline{c}_i = \sum_{j=1}^K \overline{o}_{ji}$

   (f) $p \sim Beta(\sum_j w_j + a_3, c - \sum_j w_j + b_3)$

2. Sample $\alpha + \kappa$: The derivations follows directly from Fox et al. (2011).

   (a) $\overline{\nu}_j \sim \text{Bernoulli}(\frac{n_{j.}}{n_{j.}+\alpha+\kappa})$ for $j = 1, \ldots, K$. $\quad n_{j.} = \sum_{i=1}^K n_{ji}$

   (b) $\overline{\lambda}_j \sim \text{Beta}(\alpha + \kappa + 1, n_{j.})$ for $j = 1, \ldots, K$.

   (c) $\alpha + \kappa \sim \text{Gamma}\big(a_2 + c - \sum_{j=1}^K \overline{\nu}_j, b_2 - \sum_{j=1}^K log(\overline{\lambda}_j)\big)$ and $c = \sum_{i=1}^K c_i$

3. Sample η: Two auxiliary variables $\underline{\nu}$ and $\underline{\lambda}$ are introduced to sample η. Let $\overline{c} = \sum_{i=1}^K \overline{c}_i$:

   (a) $\overline{K} = \sum_{i=1}^K I(\overline{c}_i > 0)$

(b) $\underline{\nu} \sim \text{Bernoulli}\big(\frac{\overline{c}}{\overline{c}+\eta}\big)$

(c) $\underline{\lambda} \sim \text{Beta}\big(\eta+1, \overline{c}\big)$

(d) $\eta \sim \text{Gamma}(a_1 + \overline{K} - \underline{\nu}, b_1 - log\underline{\lambda})$

4. Sample Γ: Given the counts $c_{1:K}$, which are from a sample from Γ, conjugacy gives the update for Γ as

$$\Gamma|\overline{c}_{1:K}, \eta \sim Dir(\overline{c}_1, \overline{c}_2, \ldots, \overline{c}_K, \eta).$$

5. Sample Π_j: for $j = 1, \ldots, K$, given $(n_{j1}, \ldots, n_{jK})$ and the conjugate property of the Dirichlet distribution,

$$\Pi_j|\alpha, n_{j1:K} \sim Dir(\alpha\gamma_1 + n_{j1}, \ldots, \alpha\gamma_j + \kappa + n_{jj}, \ldots, \alpha\gamma_K + n_{jK}, \alpha\gamma^r_{K+1}).$$

References

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

Ang, A. & Bekaert, G. (2002), 'Short rate nonlinearities and regime switches', *Journal of Economic Dynamics and Control* **26**(78), 1243 – 1274.

Blackwell, D. & MacQueen, B. (1973), 'Ferguson distributions via polya urn schemes', *The Annals of Statistics* pp. 353–355.

Cai, J. (1994), 'A markov model of switching-regime arch', *Journal of Business & Economic Statistics* **12**(3), 309–316.

Canto, R. (2008), Modelling the term structure of interest rates: A literature review. Available at SSRN: http://ssrn.com/abstract=1640424.

Carpantier, J.-F. & Dufays, A. (2014), Specific markov-switching behaviour for arma parameters. CORE discussion paper 2014/14.

Chan, K., Karolyi, G., Longstaff, F. & Sanders, A. (1992), 'An empirical comparison of alternative models of the short-term interest rate', *Journal of Finance* **47**, 1209–1227.

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

Cox, J., Ingersoll, J. & Ross, R. (1985), 'A theory of the term structure of interest rates', *Econometrica* **53**, 385–407.

Dufays, A. (2012), 'Infinite state markov switching for dynamic volatility and correlation models', *CORE discussion paper 2012/4* .

Durham, G. B. (2003), 'Likelihood-based specification analysis of continuous-time models of the short-term interest rate', *Journal of Financial Economics* **70**(3), 463 – 487.

Escobar, M. & West, M. (1995), 'Bayesian density estimation and inference using mixtures', *Journal of the American Statistical Association* **90**, 577–588.

Ferguson, T. S. (1973), 'A bayesian analysis of some nonparametric problems', *The Annals of Statistics* **1**, 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**, 1020–1056.

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

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

Guidolin, M. & Timmermann, A. (2009), 'Forecasts of {US} short-term interest rates: A flexible forecast combination approach', *Journal of Econometrics* **150**(2), 297 – 311.

Hamilton, J. D. (1988), 'Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates', *Journal of Economic Dynamics and Control* **12**, 385–423.

Jochmann, M. (2014), 'Modeling U.S. inflation dynamics: a bayesian nonparametric approach', *forthcoming Econometric Reviews* .

Kalimipalli, M. & Susmel, R. (2004), 'Regime-switching stochastic volatility and short-term interest rates', *Journal of Empirical Finance* **11**(3), 309 – 329.

Lanne, M. & Saikkonen, P. (2003), 'Modeling the u.s. short-term interest rate by mixture autoregressive processes', *Journal of Financial Econometrics* **1**(1), 96–125.

Lhabitant, F., Gibson, R. & Talay, D. (2001), Modeling the term structure of interest rates: A review of the literature. Available at SSRN: http://ssrn.com/abstract=275076.

Musiela, M. & Rutkowski, M. (2005), Short-term rate models, *in* 'Martingale Methods in Financial Modelling', Vol. 36 of *Stochastic Modelling and Applied Probability*, Springer Berlin Heidelberg, pp. 383–416.

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

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

Smith, D. R. (2002), 'Markov-switching and stochastic volatility diffusion models of short-term interest rates', *Journal of Business & Economic Statistics* **20**(2), 183–197.

Song, Y. (2013), 'Modelling regime switching and structural breaks with an infinite hidden markov model', *forthcoming Journal of Applied Econometrics* .

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

Van Gael, J., Saatci, Y., Teh, Y. & Ghahramani, Z. (2008), Beam sampling for the infinite hidden markov model, *in* 'Proceedings of the 25th International Conference on Machine Learning:', pp. 1088–1095.

Vasicek, O. (1977), 'An equilibrium characterization of the term structure', *Journal of Financial Economics* **5**, 177–188.

Walker, S. (2007), 'Sampling the dirichlet mixture model with slices', *Communications in Statistics Simulation and Computation* **36**, 123–145.

Table 1: Statistical Summaries of 3-Month T-Bill

Name	Mean	SD	Median	25%Q	75%Q	Skewness	Kurtosis
Level	4.658	3.030	4.615	2.720	6.110	0.840	1.260
Change	0.000	0.190	0.000	-0.050	0.050	-0.730	24.150

This table reports summary statistics for weekly 3 month T-bill rates from Jan-15-1954 to Mar-28-2014 (3142 observations).

Table 2: Posteriors and Prior Summary of Hierarchical-Priors

	Prior		Posterior of IHMM-CIR-H		Posterior of IHMM-VSK-H	
	Mean	90% DI	Mean	90% DI	Mean	90% DI
η	5.000	(2.459, 8.110)	3.810	(2.258, 5.591)	4.104	(2.490, 5.944)
α	5.000	(2.541, 7.797)	0.768	(0.530, 1.023)	0.717	(0.503, 0.953)
χ	1.000	(0.996, 2.081)	0.678	(0.368, 1.046)	0.502	(0.283, 0.755)
ν	1.000	(0.093, 2.180)	8.07e-4	(3.11e-4, 1.41e-3)	7.02e-4	(2.6e-4,1.23e-3)
$\boldsymbol{\phi}_1$	0.000	(-0.857, 0.892)	0.002	(-0.220, 0.229)	2.62e-4	(-0.196, 0.196)
$\boldsymbol{\phi}_2$	0.000	(-0.881, 0.980)	-0.003	(-0.211, 0.204)	6.23e-4	(-0.167, 0.170)
$\boldsymbol{B}_{11}$		(0.227, 4.333)	0.396	(0.093, 0.770)	0.354	(0.079, 0.669)
$\boldsymbol{B}_{22}$		(0.206, 4.572)	0.341	(0.081, 0.673)	0.203	(0.069, 0.332)
$\boldsymbol{B}_{12}$		(-1.515, 1.603)	-0.157	(-0.428, 0.054)	-0.097	(-0.234, 0.048)
Regimes #			7.985	(7.000, 9.000)	8.771	(8.000, 10.000)

This table reports the mean and 0.90 density intervals (DI) from the benchmark prior and the posteriors of the IHMM-CIR-H and IHMM-VSK-H models.

Table 3: Forecast Performance

Model	Log-Predictive Likelihood	RMSE
1955-Dec-09 to 2014-Mar-28 (3043 observations)		
IHMM-CIR-H	**2860.079**	**0.1937**
IHMM-CIR	2432.647	0.1951
MS-3-CIR-H	2805.243	0.1947
MS-3-CIR	2435.441	0.1949
MS-2-CIR-H	2631.190	0.1949
MS-2-CIR	2422.794	0.1950
GARCH-CIR	2118.011	0.1946
CIR-Roll	1761.328	0.1954
iHMM-VSK-H	2815.764	0.1954
iHMM-VSK	2381.192	0.1956
MS-3-VSK-H	2647.157	0.1951
MS-3-VSK	2371.533	0.1956
MS-2-VSK-H	2177.383	0.1953
MS-2-VSK	2143.111	0.1952
GARCH-VSK	2055.751	0.1954
VSK-roll	1332.411	0.1955
2000-Jan-07 to 2014-Mar-28 (743 observations)		
IHMM-CIR-H	**1333.426**	0.09638
IHMM-CIR	1244.576	0.09717
MS-3-CIR-H	1306.158	0.09679
MS-3-CIR	1247.913	0.09744
MS-2-CIR-H	1244.722	0.09671
MS-2-CIR	1230.406	0.09681
GARCH-CIR	1054.613	0.09714
CIR-Roll	965.030	0.09677
IHMM-VSK-H	1309.384	0.09628
IHMM-VSK	1015.873	**0.09611**
MS-3-VSK-H	1205.062	0.09660
MS-3-VSK	1002.516	0.09650
MS-2-VSK-H	908.486	0.09694
MS-2-VSK	877.123	0.09686
GARCH-VSK	1112.603	0.09715
VSK-Roll	677.907	0.09671

This table displays log-predictive likelihoods and root-mean squared forecast errors over two sample periods for various models. MS-K denotes a Markov switching model of dimension K, IHMM an infinite hidden Markov model, H a hierarchical prior, CIR-Roll and VSK-Roll refer models estimation with a rolling window of size 500. VSK models set $x = 0$ while CIR set $x = 0.5$. For additional details see the text.

Table 4: Sensitivity Analysis of 2000-Jan-07 to 2014-Mar-28 (743 observations)

Changes to Prior	IHMM-CIR-H		IHMM-VSK-H	
	LPL	RMSE	LPL	RMSE
benchmark prior	1333.426	0.09638	1309.384	0.09628
$a_1 = 2.5$, $b_1 = 0.5$	1324.724	0.09723	1306.500	0.09632
$a_2 = 2.5$, $b_2 = 0.5$	1331.880	0.09630	1305.404	0.09672
$h_0 = (0,0)^T$ Diag$(H_0) = 5$	1327.240	0.09721	1303.035	0.09575
$a_0 = 2$, Diag$(A_0) = 2$	1330.226	0.09626	1307.847	0.09561
$\rho_0 = 3$,$c_0 = 1$, $d_0 = 0.5$	1333.953	0.09534	1304.624	0.09550
All Above Combined	1324.741	0.09677	1308.507	0.09610
$a_1 = 2$, $b_1 = 8$	1326.314	0.09597	1308.049	0.09545
$a_2 = 2$, $b_2 = 8$	1333.778	0.09632	1300.221	0.09626
$a_1 = a_2 = 2$, $b_1 = b_2 = 8$	1328.790	0.09600	1308.854	0.09515

This table displays log-predictive likelihoods (LPL) and root-mean squared forecast errors (RMSE) for the two models, IHMM-CIR-H and IHMM-VSK-H, for various changes in the prior parameters from the assumed benchmark prior listed in Section 5.2.

Figure 1: Weekly 3-Month T-Bill Rate Level (Top) and Change (Bottom)

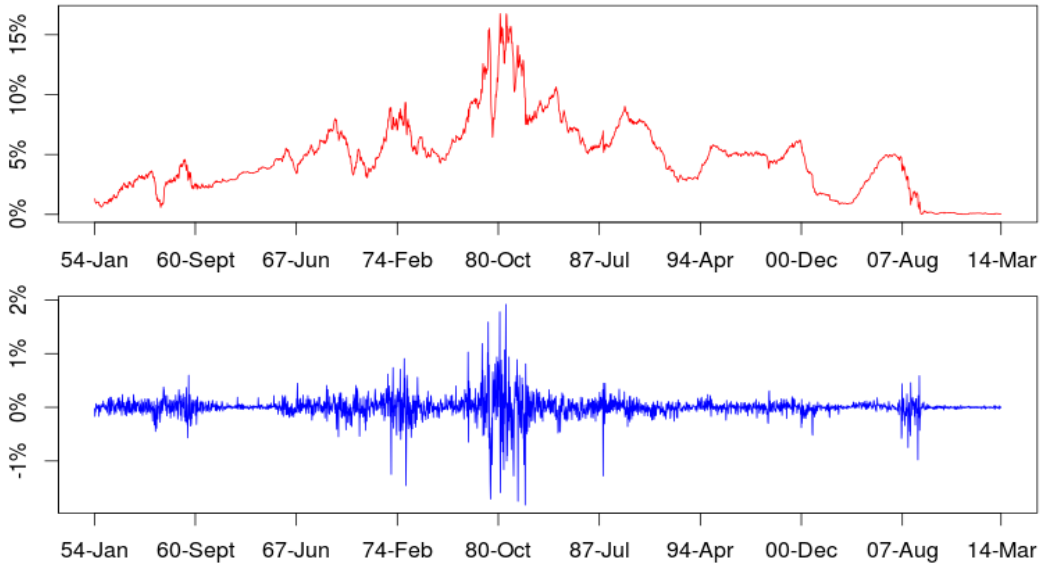

Figure 2: Histograms of States

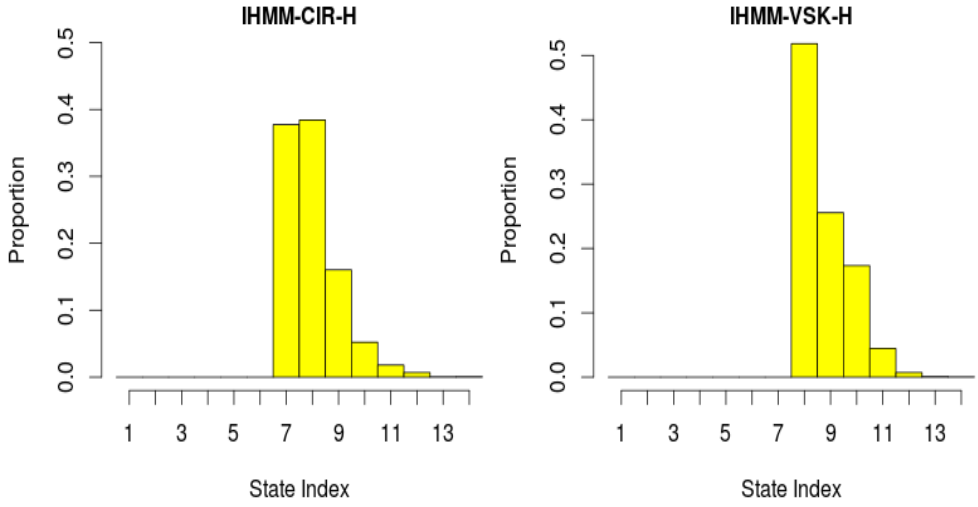

Figure 3: Posterior of iHMM-CIR-H

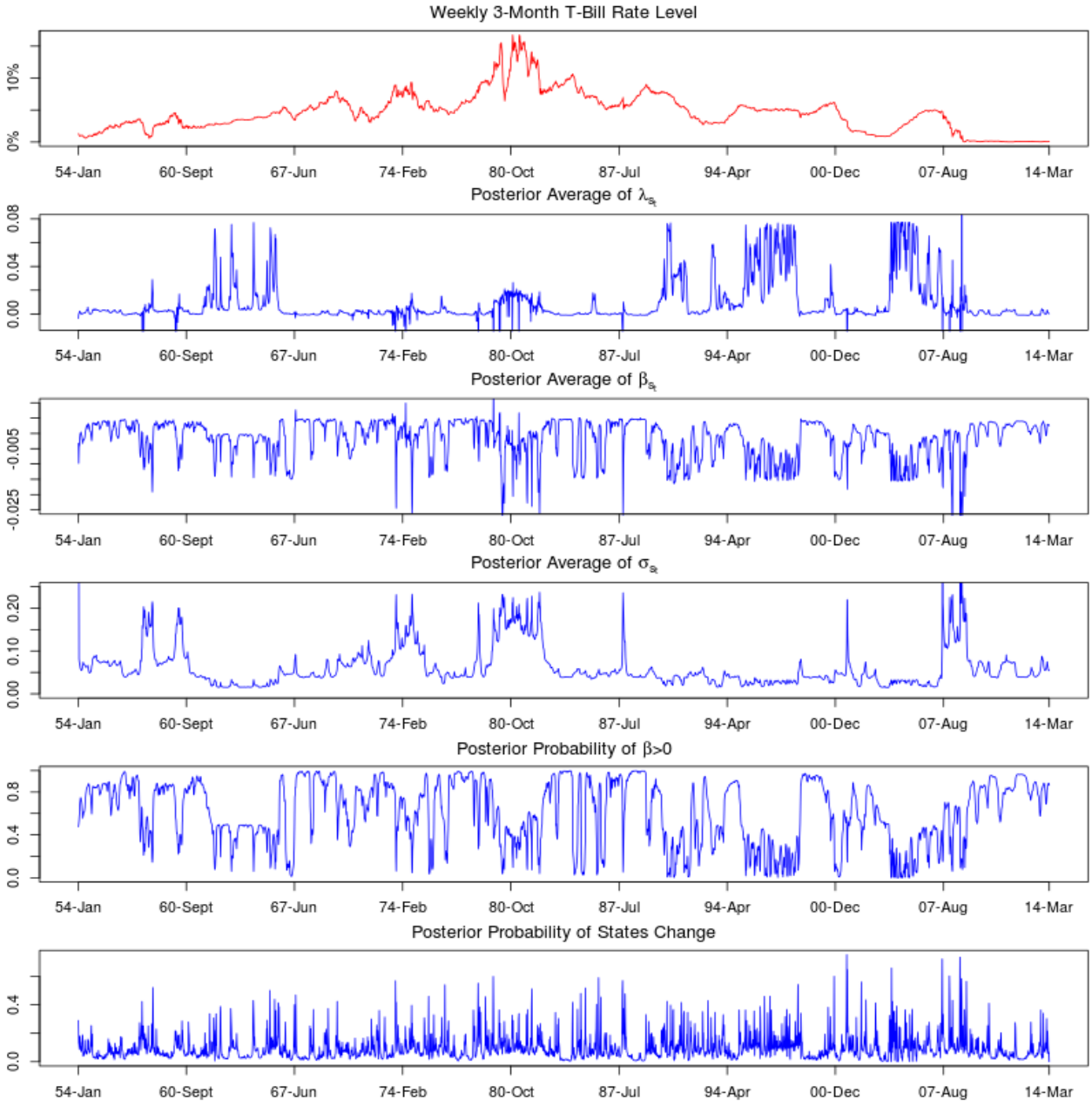

Figure 4: Posterior of iHMM-VSK-H

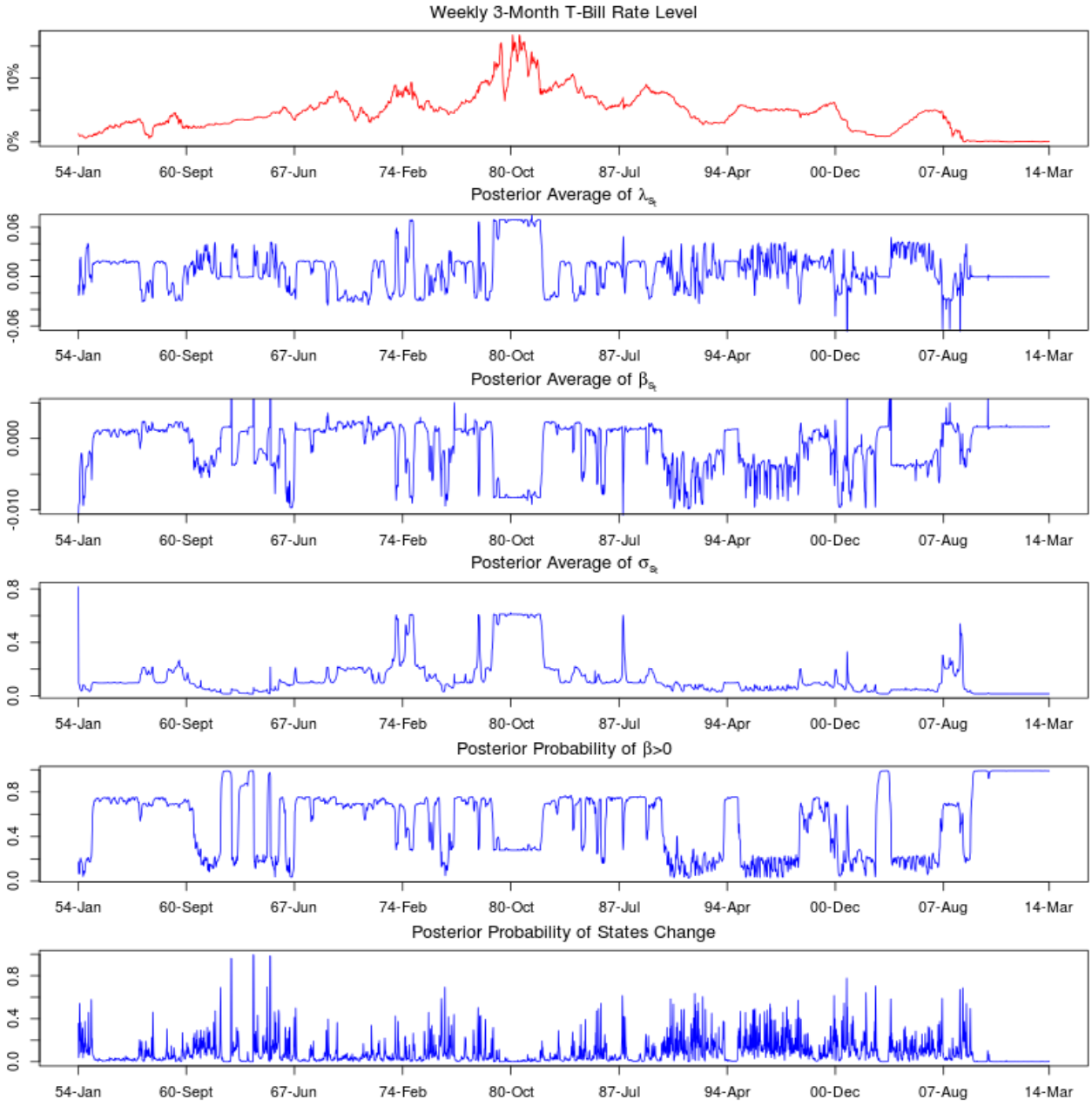

Figure 5: Heap Map of states for IHMM-CIR-H

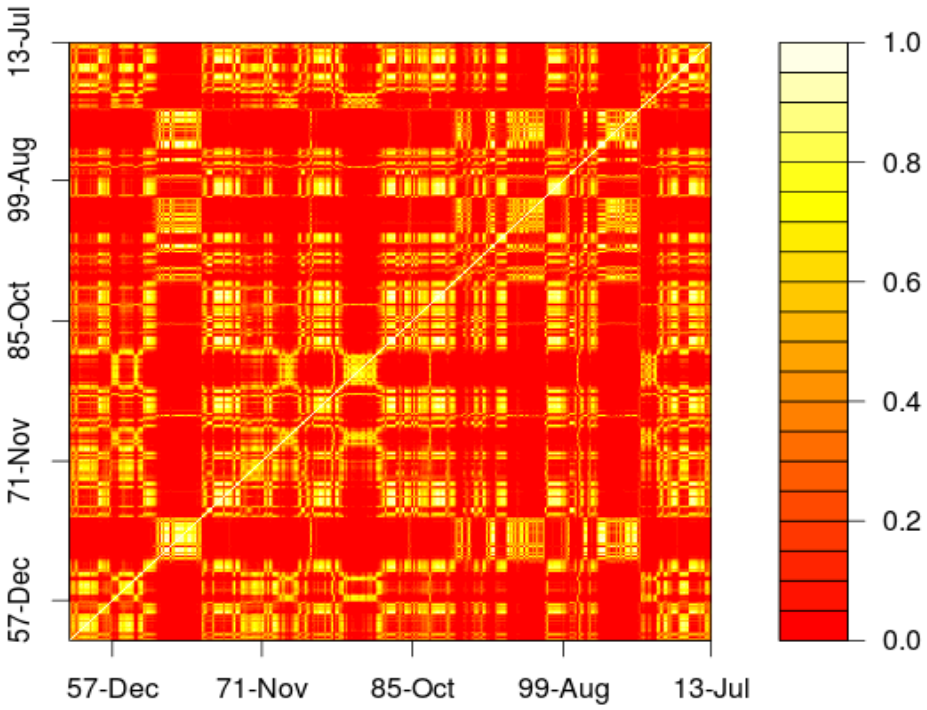

Figure 6: Heap Map of states for IHMM-VSK-H

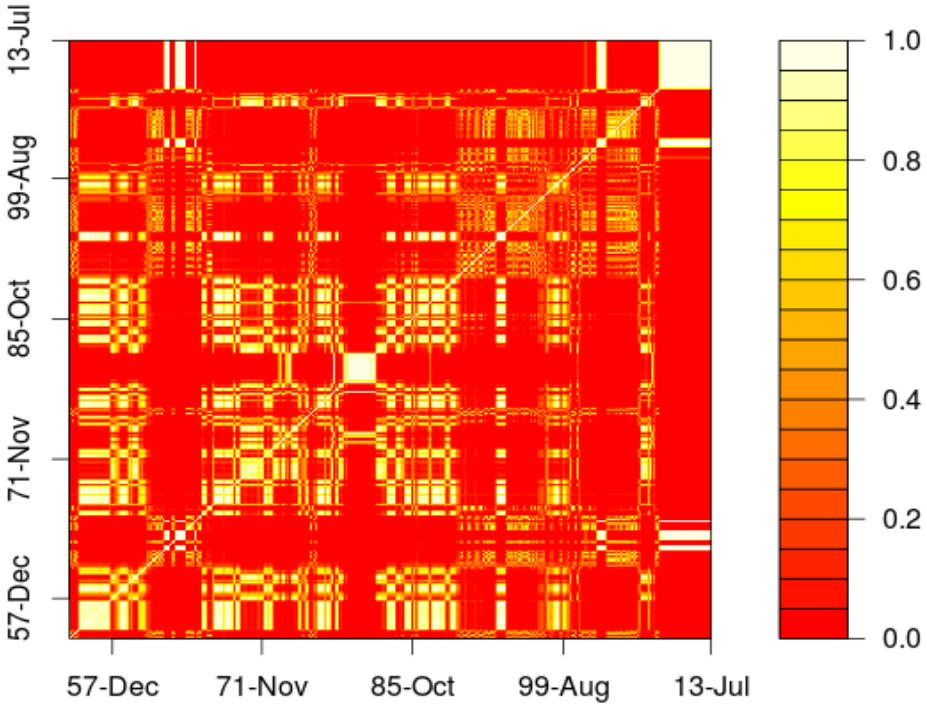

Figure 7: Log-predictive Likelihood Comparison of CIR Based Models

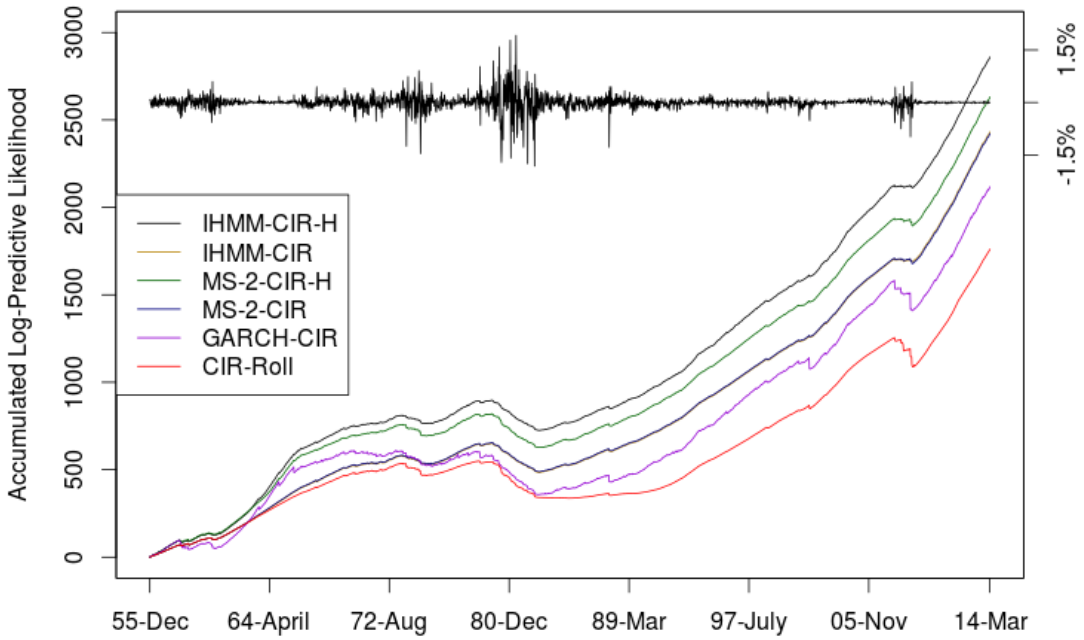

Figure 8: Log-predictive Likelihood Comparison of VSK Based Models

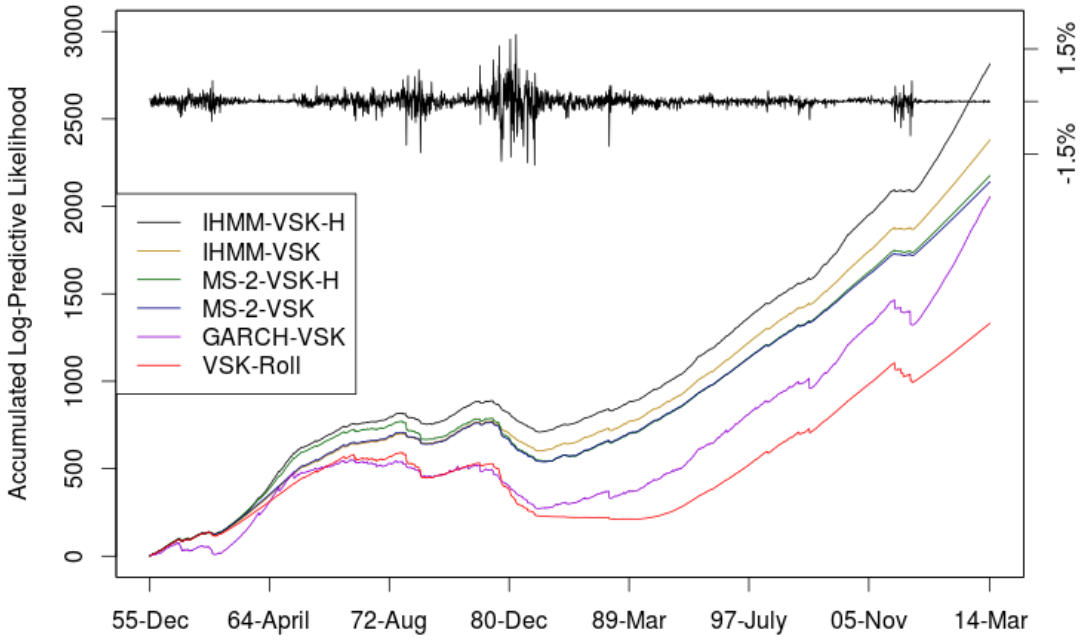

Figure 9: Regime Evolution of IHMM-CIR-H

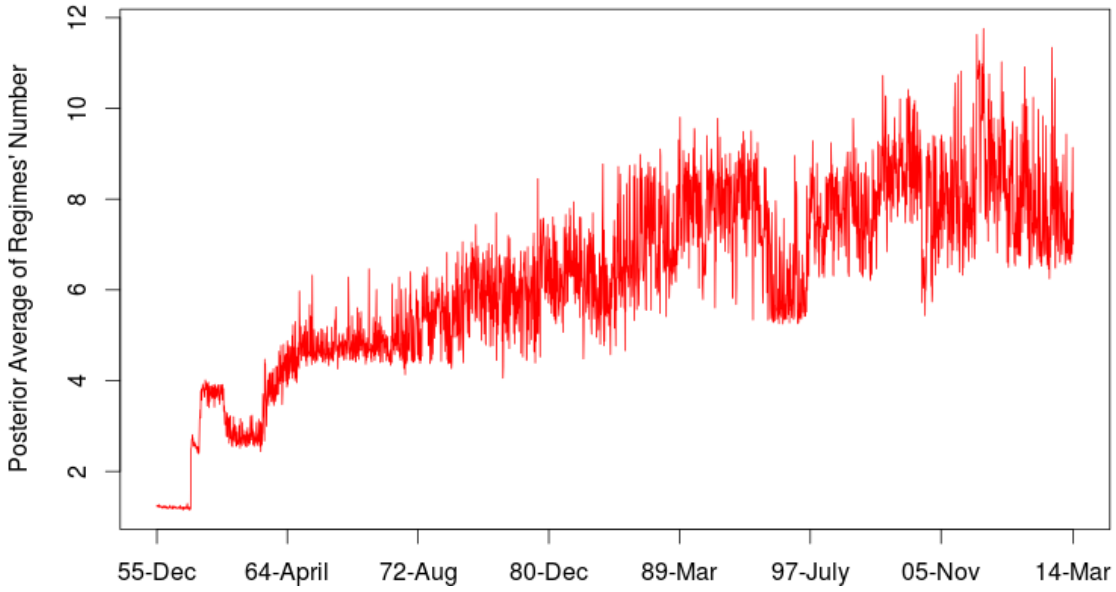

Figure 10: Regime Evolution of IHMM-VSK-H

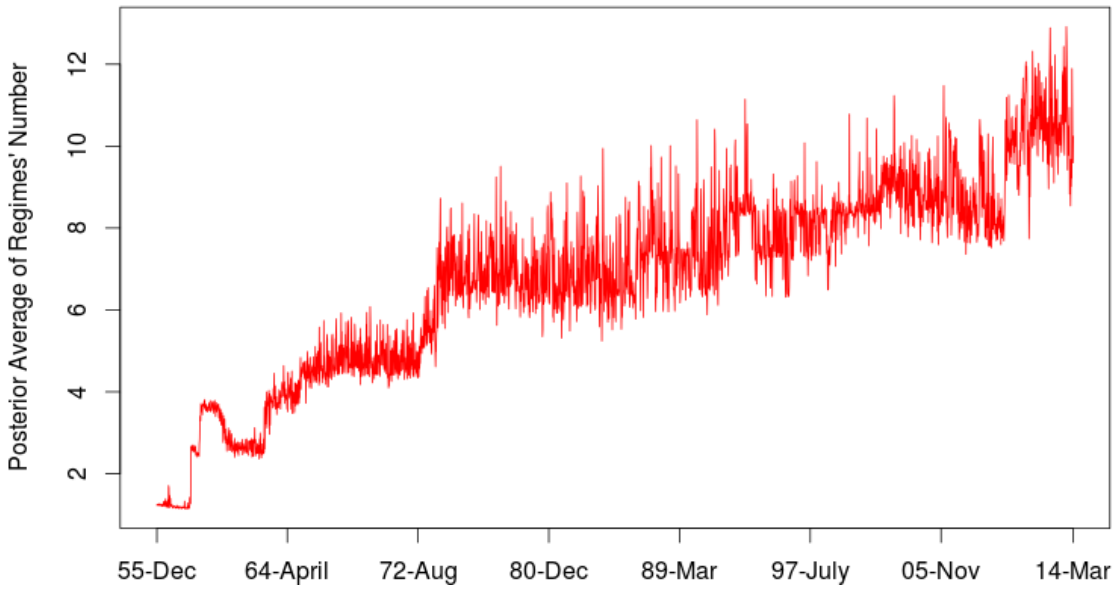

Figure 11: Predictive, Log-Predictive Densities