

Identification and Inference in Markov-Switching Models with Unknown Error Distributions

by

Shih-Tang Hwu¹

California State Polytechnic University

and

Chang-Jin Kim

University of Washington

This version, November 2019

Abstract

This paper deals with identification and Bayesian inference of the Markov-switching model when the unknown error distribution is approximated by the mixture of normals. We first deal with the problem of label switching, with a focus on the dependence among the regime-specific parameters that results from the ordering constraint necessary for identification. We then deal with the problem of disentangling the latent Markov-switching variable(s) from the latent mixture indicator variable. We show that, once these identification issues are appropriately taken care of, the standard Markov Chain Monte Carlo (MCMC) procedure can be employed for inference of the model.

When applied to the log difference of postwar U.S. industrial production index [1947:M1-2017M1], the proposed identification schemes and the MCMC algorithm can effectively control for the irregular components or outliers in the error term. This results in sharp and accurate inferences on recession probabilities just like the dynamic factor models of the co-incident variables do. Furthermore, while the performance of the model with a normality assumption is very sensitive to the priors employed for the parameters, the performance of the proposed model is robust to them.

Key Words: Label Switching Problem, Ordering Constraint, Rejection Sampling, Permutation Sampling, Identification Condition, Unknown Error Distribution, MCMC.

¹ **Hwu:** Department of Economics, California State Polytechnic University, Pomona, CA 91768 [Email: shwu@cpp.edu]; **Kim:** (Corresponding author) Department of Economics, University of Washington, Seattle, WA. 98195 [Email: changjin@uw.edu]. Kim acknowledges financial support from the Bryan C. Cressey Professorship at the University of Washington.

1. Introduction

Since the seminal work of Hamilton (1989), the basic Markov-switching model has been extended in various ways. For example, Diebold et al. (1994) and Filardo (1994) extend the model to allow the transition probabilities governing the Markov process to be functions of exogenous or predetermined variables. Kim (1994) extends it to the case of the state-space model, which encompasses general dynamic models that includes autoregressive moving average processes, unobserved components models, dynamic factor models, etc. Chib (1998) introduces a structural break model with unknown multiple change-points, by constraining the transition probabilities of the Markov-switching model so that the latent state variable can either stay at the current value or jump to the next higher value.² More recently, Kaufman (2015) propose a general K – *state* models with time-varying transition probabilities by employing the multinomial Logit specification. Fox et al. (2011), Song (2014), and Bauwens et al. (2017) introduce infinite hidden Markov models and generalize the finite-state Markov switching model of Hamilton (1989) to the case of an infinite number of states. These models integrate the regime switching and structural break dynamics in a unified Bayesian framework.

Without an exception, however, estimation of the aforementioned models and the other Markov-switching models in the literature has relied upon parametric assumptions on the distribution of the error term. Most applications in the literature assume normally distributed errors, with rare exceptions like Dueker (1997) and Bulla (2011), who propose Markov-switching models of stock returns in which the innovations are assumed to be drawn from a Student-t distribution, and De Angelis and Viroli (2017), who assume the normal-inverse Gaussian distribution as conditional form of financial returns and model innovations.

In this paper, we deal with identification and Bayesian inference of the Markov-switching model when the unknown error distribution is approximated by the mixture of normals. Two important identification issues delivered in this paper include: i) an issue related to the label switching problem and ii) an issue related to disentangling the latent Markov-switching variable(s) from the latent mixture indicator variable. We show that, once these

² For surveys of earlier literature on Markov switching models, refer to Frühwirth-Schnatter (2006) and Hamilton (2016).

identification issues are appropriately taken care of, the standard Markov Chain Monte Carlo (MCMC) procedure can be employed for inference of the model.

A conventional approach employed to deal with the first identification issue is to impose an ordering constraint on the regime-specific parameters and to employ a rejection method to generate them jointly. But Stephens (2000) and Frühwirth-Schnatter (2001) provide illustrative examples in which this conventional approach may not solve the label switching problem. As an alternative, Frühwirth-Schnatter (2001) proposes to employ the permutation sampler, in order to reorder the labeling when the ordering constraint is violated at a particular MCMC run. She has shown that her permutation sampler considerably improves upon the conventional approach based on the rejection method. However, note that the ordering constraint results in dependence among the regime-specific parameters, which is ignored when specifying the priors of the regime-specifying parameters for the permutation sampler. This may have non-negligible effect on the performance of the permutation sampler when the sample size is small. We show that we can indirectly specify the prior dependence among the regime-specific parameters by employing appropriately transformed parameters. The resulting sampling procedure that we present can be very easily implemented even when the model involves more than one latent Markov-switching variables.

In order to deal with the second identification issue, we mathematically derive the identification conditions and show that they can be easily incorporated into the MCMC procedure by appropriately specifying the priors for the transition probabilities of the latent Markov-switching variable(s). The rest of this paper is organized as follows. In Section 2, we motivate our paper by performing a Monte Carlo experiment, which is designed to investigate the effect of maximizing a normal log-likelihood when the normality assumption is violated. In Section 3, we deliver two identification issues. In Section 4, we present how the identification schemes developed in Section 3 can be implemented to a more general model. We also present an MCMC algorithm for inference of the model. In Section 6, we apply the proposed identification schemes and the MCMC algorithm to the log difference of postwar U.S. industrial production index [1947:M1- 2017M1]. Section 7 concludes the paper.

2. Pitfalls in Ignoring Non-normality in Markov-switching Models and Maximizing a Normal Log Likelihood

In this section, we investigate the small-sample performance of the maximum likelihood estimation of Markov-switching models, when a normal log-likelihood is maximized but the normality assumption is violated. For this purpose, we consider the following model with Markov-switching mean:

$$\begin{aligned} y_t &= \beta_{S_t} + \sigma \varepsilon_t^*, \quad \varepsilon_t^* \sim i.i.d.(0, 1), \quad S_t = 1, 2, \\ t &= 1, 2, \dots, T, \end{aligned} \tag{1}$$

where S_t is a 2-state Markov-switching process with transition probabilities

$$Pr[S_t = 1 | S_{t-1} = 1] = p_{S,11}, \quad Pr[S_t = 2 | S_{t-1} = 2] = p_{S,22}. \tag{2}$$

We consider the following four alternative distributions for the error term ε_t , the first two of which are symmetric and the other two are asymmetric:

Case #1

$$\varepsilon_t^* \sim i.i.d. \quad N(0, 1)$$

Case #2

$$\varepsilon_t^* = \frac{u_t}{\sqrt{\nu/(\nu - 2)}}, \quad u_t \sim i.i.d. \quad t - \text{distribution with d.f.} = \nu$$

Case #3

$$\varepsilon_t^* = \frac{\ln(u_t^2) - E(\ln(u_t^2))}{\sqrt{\text{var}(\ln(u_t^2))}}, \quad u_t \sim i.i.d. \quad N(0, 1),$$

where $E(\ln u_t^2) = -1.2704$, $\text{var}(\ln u_t^2) = \pi^2/2$.

Case #4

$$\varepsilon_t^* | D_t \sim i.i.d. \quad N(\mu_{D_t}^*, h_{D_t}^{*2}), \quad D_t = 1, 2, 3,$$

$$Pr[D_t = 1] = p_{D,1}, \quad Pr[D_t = 2] = p_{D,2}, \quad Pr[D_t = 3] = p_{D,3}$$

For each of the above four cases, we generate 1,000 sets of data. For each data set generated, we estimate the model in equations (1) and (2) by maximizing a normal log-likelihood. While the normality assumption is satisfied for Case #1, it is violated for the

other three cases. We consider three alternative sample sizes: $T = 500$, $T = 5,000$ and $T = 5,0000$. The parameter values we assign are given below:

$$\beta_1 = -0.6, \quad \beta_2 = 0.7; \quad \sigma^2 = 1.1; \quad p_{S,11} = 0.9, \quad p_{S,22} = 0.95;$$

$$\nu = 5;$$

$$\mu_1^* = 1.05, \quad \mu_2^* = 0.1, \quad \mu_3^* = -1.35; \quad h_1^{*2} = 0.2, \quad h_2^{*2} = 0.05, \quad h_3^{*2} = 1.695;$$

$$p_{D,1} = 0.2, \quad p_{D,2} = 0.6, \quad p_{D,3} = 0.2$$

Table 1 reports the mean of the estimates for each parameter and for each case, as well as the root mean squared error (RMSE) for the estimates (in the parenthesis). For case #1, in which we have normally distributed error term, the mean parameter estimates are very close to their true values for all sample sizes, with the RMSE's getting smaller as the sample size increases. For case #2, in which error term follows a t-distribution, a deviation from normality does not seem to affect the estimation results a lot. When the sample size is 500, the biases are larger than those for Case #1. But they quickly disappear with smaller RMSE's as the sample size increases, even though the RMSE's remain larger than those for Case #1 for all sample sizes. For cases #3 and #4, in which the error distributions are asymmetric, however, we have considerably larger biases and RMSE's than for Cases #1 or 2. Biases and RMSE's decrease as the sample size is increased. However, they remain sizable even when the sample size is as large as 50,000, especially for Case #4.

In order to investigate how inferences on regime probabilities are affected by the violation of the normality assumption when the log normal likelihood is maximized, we conduct another simulation study. When generating data, we consider the same data generating processes as given above, except that we generate S_t , $t = 1, 2, \dots, T$, only once and fix them in repeated sampling. The sample size we consider is $T = 500$. For each data set generated in this way, we estimate the model in equations (1) and (2) by maximizing a normal log-likelihood and then calculate smoothed probabilities conditional on estimated parameters. Figure 1 plots the average smoothed probabilities of low-mean regime ($S_t = 1$) for each case. The shaded areas represent the true low-mean regime. Case #1 with the normal error term provides us with the sharpest regime inferences. However, as the distribution of the

error term deviates from normality, inferences about regime probabilities deteriorate a lot especially for Case #4, in which the degree of asymmetry in the error distribution is the largest.

The simulation study in this section clearly demonstrates the pitfalls of estimating Markov-switching models by maximizing a normal log-likelihood when the normality assumption is violated. For all the cases we consider, the maximum likelihood estimation seems to result in consistent parameter estimates, in the sense that both the biases and RMSE's decrease as sample size increases. When the normality assumption is violated, however, the maximizing a normal log likelihood results in poor small sample properties of the estimators and poor inferences on regime probabilities. In particular, in a situation like Case #4, in which the degree of asymmetry in the error distribution is very high, even a sample with as many as 50,000 observations may not be considered as a large sample, in the sense that a considerable degree of biases still remains in the parameter estimates along with large RMSE's.

3. Basic Model and Identification Issues

3.1. Identification Issue #1: The Label Switching problem

3.1.1. The Label Switching Problem and Its Solution

Consider the following model with an unknown error distribution: ³

$$y_t = \beta_{S_t} + \sigma \varepsilon_t^*, \quad \varepsilon_t^* \sim i.i.d.(0, 1), \quad S_t = 1, 2, \dots, K; \quad t = 1, 2, \dots, T, \quad (3)$$

where S_t is a first order Markov-switching process with the following transition probabilities:

$$Pr[S_t = j | S_{t-1} = i] = p_{S,ij}, \quad \sum_{j=1}^K p_{S,ij} = 1, \quad i, j = 1, 2, \dots, K, \quad (4)$$

The distribution of the error term ε_t^* is unknown and potentially non-normal. We can approximate the distribution of ε_t^* by the following mixture of normals: ⁴

$$\varepsilon_t^* | D_t \sim i.i.d. N(\mu_{D_t}^*, h_{D_t}^{*2}), \quad D_t = 1, 2, \dots, M, \quad (5)$$

³ Equation (3) will be fully generalized in Section 4. For simplicity of the discussion on identification issues involved, we focus on a simplified model with equation (3).

⁴ We allow for potential asymmetry in the distribution of ε_t^* . Note that in case $\mu_m^* = 0$ for all m , the distribution is ε_t^* is symmetric.

where M is given and finite in the case of finite mixture of normals and M is potentially infinity in the case of Dirichlet process of mixture of normals. The mixture indicator variable D_t is serially independent. Furthermore, as the unconditional expectation and variance of ε_t^* are 0 and 1, respectively, we have the following restrictions on the conditional means and variances of ε_t^* :

$$\sum_{m=1}^M \mu_m^* Pr[D_t = m] = 0; \quad \text{and} \quad \sum_{m=1}^M (h_m^{*2} + \mu_m^{*2}) Pr[D_t = m] = 1. \quad (6)$$

Here, a typical way of labeling the states for S_t and D_t is given below:

$$\begin{aligned} \beta_{S_t} &= \sum_{k=1}^K \beta_k S_{k,t}, \\ \mu_{D_t} &= \sum_{m=1}^M \mu_m^* D_{m,t} \quad \text{and} \quad h_{D_t}^{*2} = \sum_{m=1}^M h_m^{*2} D_{m,t}, \end{aligned} \quad (7)$$

such that

$$\begin{aligned} S_{k,t} &= \begin{cases} 1, & \text{if } S_t = k; \quad k = 1, 2, \dots, K \\ 0, & \text{otherwise,} \end{cases} \\ D_{m,t} &= \begin{cases} 1, & \text{if } D_t = m; \quad m = 1, 2, \dots, M \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \quad (8)$$

The above labeling is not unique and the unconstrained parameter spaces for β 's and μ 's (or $h^{2'}$ s) contain $K!$ and $M!$ subspaces, respectively, each corresponding to different way to label states. As discussed in Stephens (2000) and Frühwirth-Schnatter (2001), when sampling from the unconstrained posterior via MCMC methods, which component of the sampled parameter corresponds to which state is unknown due to potential label switching. Thus, as noted by Stephens (2000), summarizing joint posterior distributions by marginal distribution may lead to nonsensical answers due to the lack of identification.

The label switching problem is not an issue at all for the serially independent mixture indicator D_t , as we are not interested in the marginal distribution of μ 's or $h^{2'}$ s or in the inferences on D_t . Furthermore, the complete data likelihood $f(y_1, \dots, y_T | D_1, \dots, D_T; \cdot)$ and the prior for D_t is invariant to the relabeling of the states in D_t .

However, it is critical that we take care of the label switching problem for S_t during the MCMC procedure, as we want to obtain meaningful inferences on S_t and on the regime-

specific parameters from their marginal posterior distributions. This can be done by imposing the following identifiability constraints:

$$\beta_1 < \beta_2 < \dots < \beta_K, \tag{9}$$

which can be implemented in the MCMC sampler by using some truncation or rejection method after drawing $\{\beta_1 \beta_2 \dots \beta_K\}$ jointly. For example, if the ordering constraint is not satisfied for the generated β 's, they are drawn again until the constraint is satisfied. But Stephens (2000) and Frühwirth-Schnatter (2001) provide illustrative examples in which this conventional approach may not solve the label switching problem. As an alternative, Frühwirth-Schnatter (2001) proposes to employ the permutation sampler. For example, whenever the ordering constraint is violated at a particular MCMC run, a suitable permutation is applied.

Note that the ordering constraint in equation (9) results in non-zero correlations among $\beta_k, k = 1, 2, \dots, K$. Furthermore, the correlations among the β parameters are dependent upon the distances among them. For example, the smaller the distances among the β parameters, the higher will be the correlations among them. Imposing the ordering constraint using the rejection method suffers from two problems. First, it fails to incorporate these correlations when specifying the priors for the β parameters. Second, every time the β parameters are redrawn when the ordering constraint fails, we lose sample information about these correlations. Thus, only when the distances among the β parameters are large enough relative to the variance of the error term, will the approach based on the rejection method work. This is because the low correlations among the β parameters would leave us with little information to lose, whether it is non-sample information or sample information.

The permutation sampler proposed by Frühwirth-Schnatter (2001) successfully handles the second problem mentioned above. As a suitable permutation is employed without redrawing the β parameter, the MCMC output loses no sample information about the correlations among the β parameters. This is why the permutation sampler considerably improves upon the rejection method as illustrated by Frühwirth-Schnatter (2001). However, the permutation sampler may not be free from the first problem mentioned earlier, due to its difficulty in designing the joint prior of the β parameters that account for the correlation structure. Ad-

ditional drawback of the permutation sampler would be that we have to set all the marginal priors for β_k , $k = 1, 2, \dots, K$, to be identical.

In what follows, we provide an alternative procedure for dealing with the label switching problem, in which the correlations among the β parameters are fully considered indirectly. We do this by employing the following alternative specification for β_{S_t} in equations (7):

$$\beta_{S_t} = \beta_1 + \sum_{k=2}^{S_t} a_k, \quad S_t = 2, 3, \dots, K, \quad a_k > 0 \text{ for all } k. \quad (10)$$

$$(\text{or } \beta_k = \beta_{k-1} + a_k, \quad a_k > 0, \quad k = 2, \dots, K)$$

An advantage of the above specification is that we can indirectly and completely specify the prior dependence among β_k , $k = 1, 2, \dots, K$, by employing independent marginal priors for β_1 and a_k , $k = 2, 3, \dots, K$. We can then draw each of these parameters from an appropriate truncated marginal posterior distribution. For example, we draw α_k conditional on β_1 , $\tilde{\alpha}_{\neq k} = \{\alpha_2, \dots, \alpha_{k-1}, \alpha_k, \dots, \alpha_K\}$, all the other variates, and data, for $k = 2, 3, \dots, K$. We can then recover β_2, \dots, β_K parameters from equation (10). Here, an important issue to consider is that the likelihood function for α_k depends only on the observations for which $S_t = j$, $j = k, k+1, \dots, K$, while the likelihood function for β_1 depends on all the observations in the sample.

3.1.2. Simulation Study: Performance of Alternative Sampling Procedures

In order to compare the performances of the above-mentioned three sampling procedures designed to deal with the label switching, we perform a simulation study in what follows. We consider a simple data generating process with $K = 2$ and normally distributed errors, as given below:

$$\begin{aligned} y_t &= \beta_1 S_{1,t} + \beta_2 S_{2,t} + \sigma \varepsilon_t, \\ &= \beta_1 + \alpha S_{2,t} + \sigma \varepsilon_t, \quad \varepsilon_t \sim i.i.d.N(0, 1), \quad t = 1, 2, \dots, T, \\ &\quad \beta_1 < \beta_2 \text{ or } \alpha > 0 \end{aligned}$$

$$T = 500; \quad \beta_1 = -1, \quad \beta_2 = 1; \quad \sigma^2 = 1.35; \quad p_{S,11} = p_{S,22} = 0.65.$$

The priors we employ for each sampling procedure are summarized below:

Priors for Rejection Sampling

$$\beta_1 \sim N(-0.5, 2); \beta_2 \sim N(0.5, 2); \sigma^2 \sim IG(4, 4); p_{jj} \sim \text{Beta}(0.6, 0.4), j = 1, 2,$$

Priors for Permutation Sampling

$$\beta_1 \sim N(0, 2); \beta_2 \sim N(0, 2); \sigma^2 \sim IG(4, 4); p_{jj} \sim \text{Beta}(0.6, 0.4), j = 1, 2,$$

Priors for Proposed Sampler

$$\beta_1 \sim N(-0.5, 2); \alpha_2 \sim N(1, 2); \sigma^2 \sim IG(4, 4); p_{jj} \sim \text{Beta}(0.6, 0.4), j = 1, 2,$$

where, in the cases of the rejection sampling and the proposed sampling, the implied prior means for β_1 and β_2 are the same. In the case of the permutation sampling, the choice of the priors is limited, i.e., the priors for β_1 and β_2 should be identical. We set the posterior mean of β_1 and β_2 to be the average of their true values.

Our simulation study is based on an arbitrary sample generated from the data generating process described above. For each of the cases, we obtain 55,000 MCMC draws and discard the first 5,000 draws. Figure 2 depicts the posterior marginal distributions of β_1 and β_2 obtained by employing different sampling procedures. The first graph illustrates the typical label switching problem that arises when the ordering constraint is not imposed. The marginal posterior distributions of β_1 and β_2 are bimodal. The second graph shows that the rejection sampler somewhat relieves the problem. The third graph shows that the permutation sampler considerably improves upon the rejection sampler. However, the problem still seems to remain, even though it is minor. The last graph shows that the proposed sampler improves upon the permutation sampler, even though the improvement is only marginal. For the proposed sampler, the marginal posterior distributions are unimodal and centered around the true values.

A note is in order. When the distance between the true values of β_1 and β_2 is very large relative to the true value of σ^2 , all the three sampling procedures satisfactorily handle the label switching. In the other extreme case, when the distance between the true values of β_1

and β_2 gets small enough relative to the true value of σ^2 , none of the three samplers may work.

3.2. Identification Issue #2: Disentangling the Markov-Switching Variable (S_t) Against the Mixture Indicator Variable (D_t)

There is an additional identification issue that needs to be delivered other than the problem of label switching. In this section, we consider the identification of the latent Markov-switching variable S_t in equation (3) against the latent and serially independent mixture indicator variable D_t in equation (5). For this purpose, we substitute equations (10) into equations (3) and rearrange the terms in order to obtain

$$y_t = \beta_1 + a_2 \sum_{k=1}^K S_{k,t} + a_3 \sum_{k=2}^K S_{k,t} + \dots + a_{K-1} \sum_{k=K-1}^K S_{k,t} + a_K S_{K,t} + \sigma \varepsilon_t^*, \quad (11)$$

$$\varepsilon_t^* \sim i.i.d.(0, 1),$$

which can be rewritten as:

Model with Transformed Parameters

$$y_t = a_2 \sum_{k=2}^K S_{k,t} + a_3 \sum_{k=3}^K S_{k,t} + \dots + a_{K-1} \sum_{k=K-1}^K S_{k,t} + a_K S_{K,t} + \varepsilon_t, \quad (12)$$

$$\varepsilon_t \sim i.i.d.(\beta_1, \sigma^2),$$

where $\varepsilon_t = \beta_1 + \sigma \varepsilon_t^*$.

We can then approximate the unknown distribution of e_t by the following mixture of normals:

$$\varepsilon_t | D_t \sim i.i.d. N(\mu_{D_t}, h_{D_t}^2), \quad D_t = 1, 2, \dots, M. \quad (13)$$

An additional advantage of employing equations (12) and (13) in place of equations (3) and (5) is that, the parameters in equation (13) are restriction-free unlike those in equation (5). Once D_t , μ_{D_t} and $h_{D_t}^2$, $t = 1, 2, \dots, T$, are drawn, we can recover β_1 and σ_1^2 by:

$$\beta_1 = \sum_{m=1}^M \mu_m Pr[D_t = m]; \quad \text{and} \quad \sigma^2 = \sum_{m=1}^M (h_m^2 + \mu_m^2) Pr[D_t = m], \quad (14)$$

where $Pr(D_t = m)$ is the mixture probability.

Note that we can rewrite equation (12) in the following way:

$$y_t = \bar{\beta}^* \bar{S}_t + \varepsilon_t, \quad \varepsilon_t | D_t \sim i.i.d.N(\mu_{D_t}, h_{D_t}^2), \quad D_t = 1, 2, \dots, M, \quad (15)$$

where $\bar{\beta}^* = [\beta_2^* \quad \beta_3^* \quad \dots \quad \beta_K^*]'$, with $\beta_k^* = \sum_{j=2}^k \alpha_j$; and $\bar{S}_t = [S_{2,t} \quad \dots \quad S_{K,t}]'$, with $S_{k,t}$, $k = 2, 3, \dots, K$, being defined in equation (8). Also, it is easy to show that the dynamics of S_t with the transition probabilities in equation (4) can be represented by the following VAR process for \bar{S}_t :⁵

$$\bar{S}_t = Q_0 + Q_1 \bar{S}_{t-1} + \bar{\nu}_t, \quad (16)$$

where the elements of the $(K-1) \times 1$ vector Q_0 and the $(K-1) \times (K-1)$ matrix Q_1 are functions of the transition probabilities.

By defining C as a collection of the eigenvectors for Q_1 , we can rewrite equation (16) as:

$$\bar{S}_t^* = \Lambda_0 + \Lambda_1 \bar{S}_{t-1}^* + \bar{\nu}_t^*, \quad (17)$$

where $\bar{S}_t^* = C^{-1} \bar{S}_t$; $\Lambda_0 = C^{-1} Q_0$; $\bar{\nu}_t^* = C^{-1} \bar{\nu}_t$; $\Lambda_1 = C^{-1} Q_1 C$ is a diagonal matrix that consists of eigenvalues of Q_1 ; and rows of ν_t^* are independent of one another. Furthermore, by multiplying both sides of equation (15) by $C^{-1}(\bar{\beta}^* \bar{\beta}^{*'})^{-1} \bar{\beta}^*$, we get:⁶

$$\bar{y}_t^* = \bar{S}_t^* + \bar{\varepsilon}_t^*, \quad \bar{\varepsilon}_t^* | D_t \sim i.i.d.N(B_{D_t}, \Omega_{D_t}), \quad D_t = 1, 2, \dots, M, \quad (18)$$

where $\bar{y}_t^* = C^{-1}(\bar{\beta}^* \bar{\beta}^{*'})^{-1} \bar{\beta}^* y_t$; $\bar{\varepsilon}_t^* = C^{-1}(\bar{\beta}^* \bar{\beta}^{*'})^{-1} \bar{\beta}^* \varepsilon_t$; rows of $\bar{\varepsilon}_t^*$ are independent of one another; and Ω_{D_t} is diagonal.

Then, particular rows from equations (17) and (18) can be represented by the following univariate unobserved components model:

$$\begin{aligned} \bar{y}_{k,t}^* &= \bar{S}_{k,t}^* + \bar{\varepsilon}_{k,t}^*, \quad \bar{\varepsilon}_{k,t}^* | D_t \sim i.i.d.N(b_{k,D_t}, \omega_{k,D_t}^2), \quad D_t = 1, 2, \dots, M, \\ \bar{S}_{k,t}^* &= \Lambda_{k,0} + \lambda_k \bar{S}_{k,t-1}^* + \bar{\nu}_{k,t}^*, \\ k &= 2, 3, \dots, K, \end{aligned} \quad (19)$$

⁵ For a derivation, readers are referred to Appendix A. For a two-state Markov-switching process for S_t (i.e., $K = 2$), for example, $\bar{S}_t = S_{2,t}$, $Q_0 = 1 - p_{S,11}$, and $Q_1 = p_{S,11} + p_{S,22} - 1$.

⁶ Given the constraints that we adopt (i.e., $\alpha_2 < \alpha_3 < \dots < \alpha_K$) in order to avoid non-identification due to label switching, the inverse of $\bar{\beta}^* \bar{\beta}^{*'}$ exists.

where λ_k , $k = 2, 3, \dots, K$, are the eigenvalues of Q_1 ; b_{k,D_t} is the k -th row of B_{D_t} and ω_{k,D_t}^2 is the k -th diagonal element of Ω_{D_t} . As the reduced-form model for the unobserved-components model in equation (19) is an ARMA(1,1) model, it is easy to show that identification of the model in equation (19) can be achieved if the following assumption hold:

Assumption 1: $\bar{\varepsilon}_{k,t}^*$ is independent of $\bar{\nu}_{i,t}^*$ for $k = 2, 3, \dots, K$.

Assumption 2: $\lambda_k \neq 0$ for $k = 2, 3, \dots, K$.

Note that Assumption 1 implies that S_t and D_t are independent of each other. Assumption 2 requires all the eigenvalues or the real parts of the eigenvalues for the Q_1 matrix to be non-zero. For economic data, however, a negative correlation between current and past regimes does not seem to make a lot of sense. We thus assume that $\lambda_k > 0$ for $k = 2, 3, \dots, K$. It is easy to show that a sufficient condition for this to hold is the following:

$$p_{S,kk} > 0.5, \quad k = 1, 2, \dots, K, \quad (20)$$

which can be imposed through priors.

In case only Assumption 2 holds, the number of parameters for the structural unobserved component model in equation (19) is larger than that for its reduced-form ARMA(1,1) model, leading to under-identification. In case only Assumption 1 holds, it is also easy to show that the model specified in equations (12) and (13) is not identified. If $\lambda_2 = 0$ and $\lambda_k \neq 0$, $k = 3, 4, \dots, K$, for example, we can alternatively derive a model with a $(K - 1)$ -state Markov-switching process for S_t and a mixture of $M + 1$ normals for ε_t . This alternative model has exactly the same likelihood value as that for model in equations (12) and (13) with a K -state Markov-switching process for S_t and a mixture of M normals for ε_t .

4. General Model Specification and Approximating the Unknown Error Distribution

4.1. A General Model

Consider the following generalized model: ⁷

⁷ The first equation in (21) can be further generalized to the following regression equation

$$\begin{aligned}
y_t &= \beta_{S_t} + u_t, \quad S_t = 1, 2, \dots, K, \\
\phi(L)u_t &= \sigma_{W_t}\varepsilon_t^*, \quad \varepsilon_t^* \sim i.i.d.(0, 1), \quad W_t = 1, 2, \dots, N, \\
\beta_1 &< \beta_2 < \dots < \beta_K; \quad \sigma_1^2 < \sigma_2^2 < \dots < \sigma_N^2,
\end{aligned} \tag{21}$$

where $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p$ is a polynomial equation in the lag operator; all roots of $\phi(L) = 0$ lie outside the complex unit circle; the transitional dynamics of S_t is specified in equation (4). We assume that W_t is independent of S_t and follows an N -state, first-order Markov-switching process with the following transition probabilities:

$$Pr[W_t = j | W_{t-1} = i] = p_{W,ij}, \quad \sum_{j=1}^N p_{W,ij} = 1, \quad i, j = 1, 2, \dots, N. \tag{22}$$

In order to avoid the non-identification resulting from the problem of label switching, we follow Section 3.1 in employing the following specifications for the β_{S_t} and $\sigma_{W_t}^2$ parameters:

$$\begin{aligned}
\beta_{S_t} &= \beta_1 + \sum_{k=2}^{S_t} a_k, \quad a_k > 0 \text{ for all } k, \quad S_t = 2, 3, \dots, K, \\
&\text{(or } \beta_k = \beta_{k-1} + a_k, \quad a_k > 0, \quad k = 2, \dots, K, \text{)} \\
\sigma_{W_t}^2 &= \sigma_1^2 \prod_{n=1}^{W_t} (1 + b_n), \quad 1 + b_n > 1 \text{ for all } n, \quad W_t = 2, 3, \dots, N, \\
&\text{(or } \sigma_n^2 = \sigma_{n-1}^2 (1 + b_n), \quad 1 + b_n > 1 \quad n = 2, \dots, N, \text{)}
\end{aligned} \tag{23}$$

which allow us to employ independent priors for $\{\beta_1, \alpha_k, k = 2, 3, \dots, K\}$ and for $\{\sigma_1^2, (1 + b_n), n = 2, 3, \dots, N\}$. This way, we can also fully account for dependence among $\beta_k, k = 1, 2, \dots, K$ and that among $\sigma_n^2, n = 1, 2, \dots, N$, which result from the ordering constraints ($\beta_1 < \beta_2 < \dots < \beta_K$ and $\sigma_1^2 < \sigma_2^2 < \dots < \sigma_N^2$).

By substituting equation (23) into equation (21) and rearranging terms, we obtain:

$$\begin{aligned}
y_t &= \beta_1 + a_2 \sum_{k=1}^K S_{k,t} + a_3 \sum_{k=2}^K S_{k,t} + \dots + a_{K-1} \sum_{k=K-1}^K S_{k,t} + a_K S_{K,t} + u_t, \\
\phi(L)u_t &= g_{W_t} u_t^*, \quad u_t^* \sim i.i.d.(0, \sigma_1^2),
\end{aligned} \tag{24}$$

where

with a vector of covariates x_t :

$$y_t = \beta_{S_t} + \Gamma'_{S_t} x_t + u_t, \tag{21'}$$

where x_t is a vector of exogenous variables. For simplicity of expositions, we focus on the model in equation (21).

$$g_{W_t}^2 = \frac{\sigma_{W_t}^2}{\sigma_1^2} = \prod_{n=1}^{W_t} (1 + b_n), \quad g_1^2 = 1, \quad W_t = 2, 3, \dots, N. \quad (25)$$

$$\left(\text{or } g_n^2 = g_{n-1}^2(1 + b_n), \quad g_1^2 = 1, \quad n = 2, 3, \dots, N \right)$$

Then, by defining $e_t = \beta_1 + u_t$ and $\varepsilon_t = \phi(1)\beta_1/g_{W_t} + u_t^*$, equation (24) can be rewritten as:

8

Model with Transformed Parameters

$$y_t = a_2 \sum_{k=2}^K S_{k,t} + a_3 \sum_{k=3}^K S_{k,t} + \dots + a_{K-1} \sum_{k=K-1}^K S_{k,t} + a_K S_{K,t} + e_t, \quad (26)$$

$$\phi(L)e_t = g_{W_t} \varepsilon_t, \quad \varepsilon_t | W_t \sim i.i.d. \left(\frac{1}{g_{W_t}} \phi(1)\beta_1, \sigma_1^2 \right),$$

where the unknown distribution of $\varepsilon_t | W_t$ can be approximated by the following mixture of normals:⁹

$$\varepsilon_t | W_t, D_t \sim i.i.d. \quad N \left(\frac{1}{g_{W_t}} \mu_{D_t}, h_{D_t}^2 \right), \quad D_t = 1, 2, \dots, M. \quad (27)$$

To complete the model other than the specification of the unknown error distribution, we employ the following priors for the parameters except those associated with the mixture of normals in equation (27):

Priors

⁸ For a Markov-switching model with covariates, the first equation in (26) can be extended to:

$$y_t = a_2 \sum_{k=2}^K S_{k,t} + a_3 \sum_{k=3}^K S_{k,t} + \dots + a_{K-1} \sum_{k=K-1}^K S_{k,t} + a_K S_{K,t} + \sum_{k=1}^K S_{k,t} \Gamma'_k x_t + e_t.$$

⁹ Note that σ_1^2 and β_1 can be easily recovered from

$$\beta_1 = \frac{1}{\phi(1)} \sum_{m=1}^M \mu_m p_{D,m}; \quad \sigma_1^2 = \left(\frac{1}{g^2} \right) \sum_{m=1}^M (\mu_m - \bar{\mu})^2 p_{D,m} + \bar{h}^2,$$

where $\left(\frac{1}{g^2} \right) = \frac{1}{T} \sum_{t=1}^T \frac{1}{g_{W_t}^2}$, $\bar{\mu} = \sum_{m=1}^M \mu_m p_{D,m}$, and $\bar{h}^2 = \sum_{m=1}^M \sigma_m^2 p_{D,m}$.

$$\begin{aligned}
\tilde{\phi} &= [\phi_1 \quad \phi_2 \quad \dots \quad \phi_p]' \sim N(A_{\tilde{\phi}}, \Sigma_{\tilde{\phi}})1_{[S_{\tilde{\phi}}]}, \\
a_k &\sim N(A_{a,k}, \sigma_{a,k}^2)1_{[a_k > 0]}, \quad k = 1, 2, \dots, K, \\
(1 + b_n) &\sim IG\left(\frac{\nu_{n,0}}{2}, \frac{\delta_{n,0}}{2}\right)1_{[(1+b_n) > 1]}, \quad n = 2, 3, \dots, N, \\
[p_{S,k1} \quad p_{S,k2} \quad \dots \quad p_{S,kK}]' &\sim Dir(\alpha_{S,i1}, \alpha_{S,i2}, \dots, \alpha_{S,iK})1_{[p_{S,kk} > 0.5]}, \quad k = 1, 2, \dots, K, \\
[p_{W,m1} \quad p_{W,m2} \quad \dots \quad p_{W,mM}]' &\sim Dir(\alpha_{W,m1}, \alpha_{W,m2}, \dots, \alpha_{W,mM})1_{[p_{W,mm} > 0.5]}, \\
&\quad m = 1, 2, \dots, M,
\end{aligned} \tag{28}$$

where $1[\cdot]$ is the indicator function; $S_{\tilde{\phi}}$ refers to the stationary region of $\tilde{\phi}$; $IG(\cdot)$ refers to the inverted Gamma distribution; and $Dir(\cdot)$ refers to the Dirichlet distribution. Note that we impose the restrictions $p_{S,kk} > 0.5$, $k = 1, 2, \dots, K$, and $p_{W,nn} > 0.5$, $n = 1, 2, \dots, N$, in order to identify the Markov-switching processes S_t and W_t against the mixture indicator variable D_t . Priors associated with the mixture of normals for ε_t in equation (27) are discussed in the next section.

4.2. Bayesian Modeling of the Mixture of Normals for the Error Term: Review

In the literature, there are two alternative ways of modeling the mixture of normals for ε_t in equation (27), as surveyed in Marin et al. (2005). One is the finite mixture normals model in which the total number of mixtures is fixed and given, and the other is the Dirichlet process mixture normals model in which the total number of mixtures is potentially infinity and treated as a random variable. Kim et al. (1998) and Omori et al. (2007) demonstrate the usefulness of the finite mixture of normals in approximating the log chi-square distribution in stochastic volatility models; and Alexander and Lazar (2006) employ it to approximate the unknown error distribution in a GARCH model. More recently, Jensen and Maheu (2013) apply the Dirichlet process mixture of normals to a multivariate GARCH model; Jensen and Maheu (2010, 2014) apply it to deal with unknown error distributions in stochastic volatility models; and Jin and Maheu (2016) apply it for Bayesian semi-parametric modeling of realized covariance matrices.

We employ the Dirichlet process mixture of normals in this paper. In what follows, we first provide a brief review of the finite mixture of normals and the Dirichlet process mixture

of normals with a focus on their differences. When the total number of mixtures, M , is fixed and pre-specified, we have the following specification for finite mixture of normals:

Finite Mixture of Normals

$$\begin{aligned}
\varepsilon_t | W_t, D_t &\sim i.i.d. N\left(\frac{1}{g_{W_t}} \mu_{D_t}, h_{D_t}^2\right), \quad D_t = 1, 2, \dots, M, \\
(p_{D,1}, p_{D,2}, \dots, p_{D,M}) &\sim \text{Dirichlet}\left(\frac{\alpha}{M}, \dots, \frac{\alpha}{M}\right), \\
(\mu_m, h_m^2) &\sim G_0, \quad m = 1, 2, \dots, M, \\
G_0 &\equiv N(\lambda_0, \psi_0 h_m^2) IG\left(\frac{\delta_h}{2}, \frac{\nu_h}{2}\right),
\end{aligned} \tag{29}$$

where $p_{D,m}$ is the mixture probability. G_0 or the joint prior distribution of (μ_m, σ_m^2) is assumed to be Normal-Inverse-Gamma. The α parameter can be either fixed or random.

For the above finite mixture of normals, the prior probability of D_t conditional on $\tilde{D}_{\neq t}$ can be derived as: ¹⁰

$$\begin{aligned}
Pr[D_t = m | \tilde{D}_{\neq t}, \alpha] &= \frac{T_{\neq t, m} + \frac{\alpha}{M}}{T - 1 + \alpha}, \quad m = 1, 2, \dots, M, \\
(\text{with } \sum_{m=1}^M Pr[D_t = m | \tilde{D}_{\neq t}, \alpha] &= 1)
\end{aligned} \tag{30}$$

where $\tilde{D}_{\neq t} = [D_1 \ \dots \ D_{t-1} \ D_{t+1} \ \dots \ D_T]'$ is the collection of mixture indicators excluding D_t ; and $T_{\neq t, m}$ is the total number of observations that belong to class m in a sample that excludes period t . An important thing to note is that the above probabilities always add up to 1. With this background, we are now ready to discuss the Dirichlet process mixture of normals and its properties.

As suggested by Neal (2000), Gorur and Rasmussen (2010), and others, the limit of the model in equation (29) as $M \rightarrow \infty$ is equivalent to the Dirichlet process mixture of normals. A formal specification for the Dirichlet process mixture of normals is given below:

Dirichlet Process Mixture of Normals

¹⁰ Proof of equation (30) is given in Appendix B.

$$\begin{aligned}
\varepsilon_t | W_t, D_t &\sim i.i.d. N\left(\frac{1}{g_{W_t}} \mu_{D_t}, h_{D_t}^2\right), \quad D_t = 1, 2, 3, \dots, \\
(\mu_m, h_m^2) &\sim G, \quad m = 1, 2, 3, \dots, \\
G | G_0, \alpha &\sim DP(\alpha, G_0), \\
G_0 &\equiv N(\lambda_0, \psi_0 h_m^2) IG\left(\frac{\delta_h}{2}, \frac{\nu_h}{2}\right),
\end{aligned} \tag{31}$$

where $DP(\cdot, \cdot)$ refers to the Dirichlet process; G_0 and α are referred to as the base distribution and the concentration parameter, respectively. Note that in the case of the finite mixture of normals, the joint distribution of (μ_m, σ_m^2) is given by G_0 , and thus, $G \equiv G_0$.

In the case of the Dirichlet process mixture of normals, the joint distribution of (μ_m, σ_m^2) is a random distribution generated by a Dirichlet process with based distribution G_0 and the concentration parameter α .¹¹ The prior probability of D_t conditional on $\tilde{D}_{\neq t}$ can be obtained by taking the limit $M \rightarrow \infty$ for equation (30), as given below:

$$\begin{aligned}
Pr[D_t = m | \tilde{D}_{\neq t}, \alpha] &= \frac{T_{\neq t, m}}{T - 1 + \alpha}, \quad m = 1, 2, \dots, M_{\neq t}^*, \\
(\text{with } \sum_{m=1}^M Pr[D_t = m | \tilde{D}_{\neq t}, \alpha] &< 1)
\end{aligned} \tag{32}$$

where $T_{m, \neq t}$ is defined earlier and $M_{\neq t}^*$ is the total number of distinctive classes (or mixtures) realized in the sample that excludes period t .

Unlike the case of the finite mixture of normals in equation (30), the above probabilities do not add up to 1, suggesting that there always exists a non-zero probability that an observation at period t belongs to a new class that does not belong to the existing $M_{\neq t}^*$ classes. This probability is given below:

$$\begin{aligned}
Pr[D_t = M_{\neq t}^* + 1 | \tilde{D}_{\neq t}, \alpha] &= 1 - \sum_{m=1}^{M_{\neq t}^*} Pr[D_t = m | \tilde{D}_{\neq t}, M_{\neq t}^*] \\
&= \frac{\alpha}{T - 1 + \alpha},
\end{aligned} \tag{33}$$

which suggests that, if α is larger, the prior mean for the total number of mixture is higher with less concentrated distribution for G in equation (31). The α parameter can be either fixed or random. In case the α parameter is random, it is common to employ a beta prior.

¹¹ That is, the Dirichlet process provides a random distribution over distributions on infinite sample spaces. The hierarchical models in which the Dirichlet process is used as a prior over the distribution of the parameters are referred to as the Dirichlet process mixture model.

5. Estimation of the Model and Simulation Study

In order to illustrate the MCMC procedure for estimation of the model that consists of equations (26)-(28) and (31), we define the corresponding vectors of parameters and latent variables in the following way:

Variates associated with Markov-switching Model in equation (26)

$$\begin{aligned}\tilde{a} &= [a_2 \quad a_3 \quad \dots \quad a_K]', \quad \tilde{g}^2 = [g_2^2 \quad g_3^2 \quad \dots \quad g_N^2]', \\ \tilde{S}_T &= [S_1 \quad S_2 \quad \dots \quad S_T]', \quad \tilde{W}_T = [W_1 \quad W_2 \quad \dots \quad W_T]', \\ \tilde{p}_S &= [p_{S,11} \quad p_{S,12} \quad \dots]', \quad \tilde{p}_W = [p_{W,11} \quad p_{W,12} \quad \dots]',\end{aligned}$$

where the transition probabilities of S_t and W_t are represented by vectors \tilde{p}_S and \tilde{p}_W , respectively.

Variates associated with Mixture of Normals in equation (31)

$$\tilde{\mu} = [\mu_1 \quad \dots \quad \mu_M]'; \quad \tilde{h}^2 = [h_1^2 \quad \dots \quad h_M^2]'; \quad \tilde{D}_T = [D_1 \quad \dots \quad D_T]'; \quad \alpha,$$

where α is the concentration parameter for the Dirichlet process. Then, the following two steps can be iterated until the convergence is achieved:

Step 1: Draw \tilde{a} , \tilde{g}^2 , \tilde{S}_T , \tilde{W}_T , \tilde{p}_S , and \tilde{p}_W conditional on $\tilde{\mu}$, \tilde{h}^2 , and \tilde{D}_T , and data $\tilde{Y}_T = [y_1 \quad y_2 \quad \dots \quad y_T]'$. In this step, the concentration parameter α is irrelevant once \tilde{D}_T is given.

Step 2: Draw $\tilde{\mu}$, \tilde{h}^2 , \tilde{D}_T , and α conditional on \tilde{a} , \tilde{g}^2 , \tilde{S}_T , \tilde{W}_T and data \tilde{Y}_T . This step is equivalent to drawing $\tilde{\mu}$, \tilde{h}^2 , \tilde{D}_T , α conditional on $\varepsilon_T = [\varepsilon_1 \quad \varepsilon_2 \quad \dots \quad \varepsilon_T]$ and \tilde{g}^2 . In this step, \tilde{p}_S and \tilde{p}_W are irrelevant once \tilde{S}_T and \tilde{W}_T are given.

Note that equation (27) implies

$$\varepsilon_t = \frac{1}{g_{W_t}} \mu_{D_t} + h_{D_t} v_t, \quad v_t \sim i.i.d.N(0, 1). \quad (34)$$

By multiplying both sides of the first equation in (26) by $\phi(L)$ and by substituting equation (34) in the resulting equation, we obtain

$$\begin{aligned} \phi(L)y_t &= \phi(L)(a_2 \sum_{k=2}^K S_{k,t} + a_3 \sum_{k=3}^K S_{k,t} + \dots + a_{K-1} \sum_{k=K-1}^K S_{k,t} + a_K S_{K,t}) + \mu_{D_t} + g_{W_t} h_{D_t} v_t, \\ v_t &\sim i.i.d.N(0, 1), \end{aligned} \tag{35}$$

$$(or \quad \phi(L)y_t = \phi(L)z_t' \tilde{a} + \mu_{D_t} + g_{W_t} h_{D_t} v_t, \quad v_t \sim i.i.d.N(0, 1),)$$

where $z_t = [\sum_{k=2}^K S_{k,t} \quad \sum_{k=3}^K S_{k,t} \quad \dots \quad \sum_{k=K-1}^K S_{k,t} \quad S_{K,t}]'$ and $\tilde{a} = [a_2 \quad a_3 \quad \dots \quad a_K]'$. Based on equation (35), we explain procedures for drawing the variates in Steps 1 and 2 from their appropriate full conditional distributions in the next two sub-sections (Sections 5.1 and 5.2). Section 5.3 provides a simulation study.

5.1. Drawing Variates Associated with Markov-switching Regression Equation Conditional on the Mixture of Normals and Data

Drawing \tilde{a} Conditional on $\tilde{\phi}, \tilde{g}^2, \tilde{S}_T, \tilde{W}_T, \tilde{\mu}, \tilde{h}^2, \tilde{D}_T$ and Data

Rearranging equation (35), we obtain

$$y_{1t} = a_2 z_{2t}^\dagger + a_3 z_{3t}^\dagger + \dots + a_K z_{Kt}^\dagger + v_t, \quad v_t \sim i.i.d.N(0, 1) \tag{36}$$

where $y_{1t} = \frac{\phi(L)y_t - \mu_{D_t}}{g_{W_t} h_{D_t}}$ and $z_{jt}^\dagger = \frac{\sum_{k=j}^K \phi(L)S_{k,t}}{g_{W_t} h_{D_t}}$, $j = 2, 3, \dots, K$. Then, for given y_{1t} and z_{jt}^\dagger , $t = p + 1, 2, \dots, T$, $j = 2, 3, \dots, K$, we can generate a_2, a_3, \dots, a_K directly from the following truncated normal distributions, without resorting to the rejection sampling:¹²

1) Draw a_2 from

$$a_2 \mid a_3, a_4, \dots, a_K \sim N(c_{a,2}, \omega_{a,2}^2)_{1[a_2 > 0]}$$

2) Draw a_3 from

$$a_3 \mid a_2, a_4, \dots, a_K \sim N(c_{a,3}, \omega_{a,3}^2)_{1[a_3 > 0]}$$

¹² Note that we do not employ the rejection method to avoid the labeling problem, as discussed in Section 3.1.

.....
.....

K-1) Draw a_K from

$$a_K \mid a_2, a_3, \dots, a_{K-1} \sim N(c_{a,K}, \omega_{a,K}^2)_{1[a_K > 0]},$$

where $c_{a,j}$ and $\omega_{a,j}^2$ refer to the posterior mean and the posterior variance of the truncated full conditional distribution of a_j , $j = 2, 3, \dots, K$.

Here, as discussed in Section 3.1, the derivation of the conditional posterior distribution for a_k should be based on the observations for which $S_t = j$, $j = k, k + 1, \dots, K$. This is because a_k is a common element only in β_{S_t} , $S_t = k, k + 1, \dots, K$.

Drawing $\tilde{\phi}$ Conditional on \tilde{a} , \tilde{g}^2 , \tilde{S}_T , \tilde{W}_T , $\tilde{\mu}$, \tilde{h}^2 , \tilde{D}_T , and Data

Rearranging equation (35), we obtain

$$y_{2t} = z_t^{*'} \tilde{\phi} + v_t, \quad v_t \sim i.i.d.N(0, 1), \quad (37)$$

where $y_{2t} = \frac{y_t - z_t' \tilde{a} - \mu_{D_t}}{g_{W_t} h_{D_t}}$ and $z_t^* = \left[\frac{y_{t-1} - z_{t-1}' \tilde{a}}{g_{W_{t-1}} h_{D_{t-1}}} \quad \frac{y_{t-2} - z_{t-2}' \tilde{a}}{g_{W_{t-2}} h_{D_{t-2}}} \quad \dots \quad \frac{y_{t-p} - z_{t-p}' \tilde{a}}{g_{W_{t-p}} h_{D_{t-p}}} \right]'$. Based on equation (37), we can draw $\tilde{\phi}$ from an appropriate posterior distribution.

Drawing \tilde{g}^2 Conditional on \tilde{a} , $\tilde{\phi}$, \tilde{S}_T , \tilde{W}_T , $\tilde{\mu}$, \tilde{h}^2 , \tilde{D}_T , and Data

By defining $\zeta_t = g_{W_t} v_t$ in equation (35), we can calculate ζ_t by

$$\zeta_t = \frac{\phi(L)(y_t - z_t' \tilde{a}) - \mu_{D_t}}{h_{D_t}}. \quad (38)$$

Furthermore, we know that

$$\zeta_t \mid W_t = n \sim N(0, g_n^2) \equiv g_{n-1} N(0, (1 + b_n)), \quad (39)$$

as we have

$$g_n^2 = g_{n-1}^2 (1 + b_n), \quad g_1^2 = 1, \quad n = 2, 3, \dots, N. \quad (40)$$

Here, we want to draw $(1 + b_n)$ conditional on $g_{n-1}^2, (1 + b_{n+1}), \dots, (1 + b_N)$ for $n = 2, 3, \dots, N$, and then we calculate g_n^2 , $n = 2, 3, \dots, N$, based on equation (40). What's important here is that the likelihood function for $(1 + b_n)$ depends on the values of ζ_t for which

$W_t = n, n+1, \dots, N$, as $(1+b_n)$ is a common element only in $g_{W_t}^2$, $W_t = n, n+1, \dots, N$. Thus, if we define

$$\zeta_{n,t}^* = \frac{\zeta_t}{g_{n-1} \sqrt{\prod_{i=n+1}^N (1+b_i W_{i,t})}}, \quad (41)$$

we have

$$\zeta_{n,t}^* \mid g_{n-1}, (1+b_{n+1}), \dots, (1+b_N) \sim N(0, (1+b_n)), \quad (42)$$

for $T_n = \{t : W_t = n, n+1, \dots, N\}$.

Then, given the prior for $(1+b_n)$ in equation (28), we can draw $(1+b_n)$ from the following truncated inverse Gamma distribution:

$$(1+b_n) \mid g_{n-1}^2, (1+b_{n+1}), \dots, (1+b_N), \tilde{u}_T \sim IG\left(\frac{\nu_{n,1}}{2}, \frac{\delta_{n,1}}{2}\right)_{1[1+b_n > 1]}, \quad (43)$$

where

$$\begin{aligned} \delta_{n,1} &= \delta_{n,0} + \sum_{T_n} \zeta_{n,t}^* \\ \nu_{n,1} &= \nu_{n,0} + c_n, \end{aligned} \quad (44)$$

with c_n referring to the cardinality of T_n . When drawing $(1+b_n)$ from equation (43), we draw $(1+b_n)$ directly from the truncated Inverse Gamma distribution.

Drawing \tilde{S}_T , \tilde{p}_S , \tilde{W}_T , and \tilde{p}_W Conditional on \tilde{a} , $\tilde{\phi}$, \tilde{g}^2 , $\tilde{\mu}$, \tilde{h}^2 , \tilde{D}_T , and Data

For this step, we can rewrite equation (35) in the following way:

$$\phi(L)(y_t - \beta_{S_t}^*) = \mu_{D_t} + g_{W_t} h_{D_t} v_t, \quad v_t \sim i.i.d. N(0, 1), \quad (45)$$

where $\beta_{S_t}^* = \sum_{j=2}^{S_t} a_j$ with $\beta_1^* = 0$.

When drawing \tilde{S}_T conditional on all the other variates, equation (45) serves as a usual model with Markov-switching a latent variable S_t , while D_t and W_t serve as dummy variables. Furthermore, drawing \tilde{W}_T conditional on all the other variates, equation (45) serves as a usual model with a Markov-switching latent variable W_t , while D_t and S_t serve as dummy variables. Thus, drawing \tilde{S}_T and \tilde{W}_T is a standard procedure. Once \tilde{S}_T and \tilde{W}_T are drawn, we can draw \tilde{p}_S conditional on \tilde{S}_T and \tilde{p}_W conditional on \tilde{W}_T from the full conditional distributions derived by employing the Dirichlet distributions in equation (28) as priors.

5.2. Drawing Variates Associated with the Mixture of Normals Conditional on

$$\tilde{\varepsilon}_T = [\varepsilon_1 \quad \varepsilon_2 \quad \dots \quad \varepsilon_T]'$$

Conditional on $\tilde{\alpha}$, $\tilde{\phi}$, $\tilde{\sigma}^2$, \tilde{S}_T , \tilde{W}_T and data, we can calculate the error term ε_t from the equations in (26) as follows:

$$\varepsilon_t = \phi(L)(y_t - a_2 \sum_{k=2}^K S_{k,t} + a_3 \sum_{k=3}^K S_{k,t} + \dots + a_{K-1} \sum_{k=K-1}^K S_{k,t} + a_K S_{K,t}) \frac{1}{gW_t}. \quad (46)$$

Then, based on equation (34), we can draw the variates associated with mixture of normals (i.e., $\tilde{\mu}$, \tilde{h}^2 , \tilde{D}_T and α) conditional on $\tilde{\varepsilon}_T = [\varepsilon_1 \quad \varepsilon_2 \quad \dots \quad \varepsilon_T]'$. As discussed in Section 3.1, we are not interested in the marginal distribution of μ 's or h^2 's or in the inferences on D_t . Thus, the label switching problem for D_t is not an issue in here. We therefore draw $\tilde{\mu}$ or \tilde{h}^2 without any identifiability constraints. We follow the following procedure in drawing $\tilde{\mu}$, \tilde{h}^2 , \tilde{D}_T and α :

- i) Draw $\tilde{\mu}$ based on equation (34), conditional on \tilde{g} , \tilde{h}^2 , \tilde{D}_T , and $\tilde{\varepsilon}_T$.
- ii) Draw \tilde{h}^2 based on equation (34), conditional on $\tilde{\mu}$, \tilde{g}^2 , \tilde{D}_T , M , and $\tilde{\varepsilon}_T$.
- iii) Draw \tilde{D}_T and α for the Dirichlet process mixture of normals specified in equation (31), conditional on $\tilde{\mu}$, \tilde{h}^2 , \tilde{g}^2 , and $\tilde{\varepsilon}_T$. The total number of mixtures (M^*) realized at a particular MCMC iteration is obtained as a byproduct of drawing \tilde{D}_T .

Drawing $\tilde{\mu}$ and $\tilde{\sigma}^2$ from their full conditional distributions derived based on equation (34) is standard. The procedure for drawing \tilde{D}_T and α is based on West et al. (1994), Escobar and West (1995), and Neal (2000). Details are explained in Appendix C.

5.3. Simulation Study: Performance of the Proposed MCMC Algorithm

In Section 2, we observed considerable biases in the parameter estimates for case #4 with a sample size of 500, when a normal log-likelihood is maximized but the error distribution is given by a mixture of normals.

In this section, we perform a simulation study in order to show that the proposed model-identification schemes and the proposed algorithm work properly. For this purpose, we generate 100 sets of samples based on the following data generating process, which is the

same as Case #4 of Section 2 (with $K = 2$, $M = 1$, and $\phi(L) = 1$ for the model presented in Section 4):

Data Generating Process

$$\begin{aligned}
y_t &= \beta_{S_t} + \sigma \varepsilon_t^*, \quad \varepsilon_t \sim i.i.d(0, 1), \quad S_t = 1, 2; \quad t = 1, 2, \dots, T, \\
& \quad (y_t = \beta_1 + \alpha_1 S_{2,t} + \sigma \varepsilon_t^*, \quad \alpha_1 = \beta_2 - \beta_1,) \\
\varepsilon_t^* | D_t &\sim i.i.d. \quad N(\mu_{D_t}^*, h_{D_t}^{*2}), \quad D_t = 1, 2, 3, \\
T &= 500; \quad \beta_1 = -0.6, \quad \beta_2 = 0.7; \quad \sigma^2 = 1.1; \quad p_{11} = 0.9, \quad p_{22} = 0.95,
\end{aligned}$$

where $S_{2,t} = 1$ if $S_t = 2$ and $S_{2,t} = 0$, otherwise; S_t and D_t are independent of each other and $p_{ij} = Pr[S_t = j | S_{t-1} = i]$. The parameter values associated with the mixture of normals for ε_t^* are also the same as those for Case #4 in Section 2.

Based on the discussions on the identification issues in Section 3, we consider the following representation of the model for estimation:

$$y_t = a_2 S_{2,t} + \varepsilon_t, \quad \varepsilon_t \sim i.i.d.N(\beta_1, \sigma^2), \quad a_2 > 0,$$

where we approximate the distribution of e_t by the Dirichlet process mixture of normals in equation (31) with the following specification for the base distribution and the concentration parameter α :¹³

$$G_0 \equiv N(-0.6, 5h_m^2)IG\left(\frac{130}{2}, \frac{30}{2}\right); \quad \alpha \sim Gamma(10, 3),$$

and the priors for a_2 and the transition probabilities of S_t are specified as:

$$a_1 \sim N(1.3, 0.45)_{1[a_2 > 0]};$$

$$[p_{S,11} \quad p_{S,12}]' \sim Dirichlet(9, 1)_{1[p_{S,11} > 0.5]}; \quad [p_{S,21} \quad p_{S,22}]' \sim Dirichlet(0.5, 9.5)_{1[p_{S,22} > 0.5]}.$$

We apply the above modified model and algorithm to each generated data set. At each iteration of the MCMC procedure, we draw \tilde{S}_T , p_{11} , p_{22} , a_2 ; \tilde{D}_T and the resulting M^* ; μ_m and h_m^2 , $m = 1, 2, \dots, M^*$; and α . Then, we calculate β_1 , β_2 and σ^2 in the following way:

$$\beta_1 = \sum_{m=1}^{M^*} \mu_m^* Pr[D_t = m]; \quad \beta_2 = \beta_1 + a_2; \quad \text{and} \quad \sigma^2 = \sum_{m=1}^{M^*} (h_m^{*2} + \mu_m^{*2}) Pr[D_t = m],$$

¹³ And the prior distribution of the concentration parameter α implies that the prior mean for the number of mixture is 3.32 when sample size equal to 500.

where $Pr[D_t = m]$ is the probability of mixture for given M^* .

When we estimate the model under a normality assumption for the error term, we employ the following priors for β_1 and σ^2 :

$$\beta_1 \sim N(-0.6, 0.45); \quad \sigma^2 \sim IG(3.4, 2.7),$$

which are the same as the unconditional distributions for β_1 and σ^2 implied by our specification of the based distribution G_0 for the Dirichlet process mixture of normals. The priors for a_2 and the transition probabilities of S_t are the same as in the proposed model.

We obtain the posterior mean of each parameter conditional on each of 100 generated samples. We then calculate the mean and the standard deviation of 100 posterior means for each parameter obtained from these 100 samples. This is equivalent to investigating the sampling moments of the posterior mean for each parameter. The third column of Table 2 reports the sample mean and standard deviation (in the parenthesis) of the posterior means when the distribution of the error term is erroneously assumed to be normal. The results are almost the same as those based on the maximum likelihood approach as shown in the 6th column of Table 1 for $T = 500$. We have large biases in the parameter estimates. However, the fourth column of Table 2 shows that, when the non-normality of the error distribution is appropriately taken care of as outlined in Sections 5.1 and 5.2, these biases almost disappear.

6. An Application to the Growth of Postwar U.S. Industrial Production Index [1947M1-2017M1]

6.1. Specification for an Empirical Model

We consider the following univariate Markov-switching model for the growth of industrial production index (Δy_t), with a two-state Markov-switching mean ($S_t = 1, 2$) and a three-

state Markov-switching variance ($W_t = 1, 2, 3$): ¹⁴

$$\begin{aligned}\Delta y_t &= \beta_{1,C_t} + a_{2,C_t} S_{2,t} + u_t, \quad C_t = 1, 2, 3, \\ a_{2,C_t} &> 0, \quad \forall t, \\ u_t &= \phi u_{t-1} + g_{W_t} u_t^*, \quad u_t^* \sim i.i.d.(0, \sigma_1^2), \quad |\phi| < 1,\end{aligned}\tag{47}$$

where $S_{2,t} = 1$ if $S_t = 2$ and $S_{2,t} = 0$, otherwise; β_{1,C_t} is the mean growth rate during recession and $\beta_{1,C_t} + a_{2,C_t}$ is the mean growth rate during boom; $g_{W_t}^2$ is specified in equation (25) with $N = 3$; S_t and W_t are independent. The transitional dynamics of S_t and W_t are specified in equation (4) with $K = 2$ and in equation (22) with $N=3$, respectively.

Kim and Nelson (1999) show empirical evidence of a narrowing gap between growth rates of real GDP during recessions and booms. They argue that this narrowing gap is as important as the reduction in the volatility of the shocks as a feature of the Great Moderation. More recently, by specifying the regime-specific mean growth rates of real GDP to follow random walks, Eo and Kim (2016) also show that the mean growth rate during boom have been steadily decreasing along with the long-run mean growth rate since 1947. In order to incorporate these particular features of the business cycle discussed in Kim and Nelson (1999) and Eo and Kim (2016), we incorporate two structural breaks with unknown break points in the mean growth rates for boom and recession. For this purpose, we specify β_{1,C_t} and a_{2,C_t} in the following way: ¹⁵

$$\begin{aligned}\beta_{1,C_t} &= \gamma_1 + \gamma_2 C_{2,t} + (\gamma_2 + \gamma_3) C_{3,t}, \\ a_{2,C_t} &= (\eta_1 + \eta_2 + \eta_3) C_{1,t} + (\eta_2 + \eta_3) C_{2,t} + \eta_3 C_{3,t}, \\ \gamma_2 &> 0, \quad \gamma_3 > 0; \quad \eta_1 > 0, \quad \eta_2 > 0, \quad \eta_3 > 0,\end{aligned}\tag{48}$$

where

$$C_{k,t} = \begin{cases} 1, & \text{if } C_t = k; \quad k = 1, 2, 3 \\ 0, & \text{otherwise,} \end{cases}\tag{49}$$

¹⁴ We allow for a 3-state Markov-switching process for the variance of the shocks in order to capture the unusually high volatility during the Financial Crisis period. To avoid the identification problem associated with label switching, we specify the Markov switching variance of the shocks as in the second equation in (22).

¹⁵ Incorporating structural breaks in the mean growth rates for booms or recessions such that their gap is narrowing is based on the prior belief that the Great Moderation is not over with the onset of the Financial Crisis. In his recent study on whether the Great Moderation is over, Clark (2009) concludes that, over time, macroeconomic volatility will likely undergo occasional shifts between high and low levels with low volatility being the norm, suggesting that the Great Moderation is not over. Gadea-Rivas et al. (2014) also provide empirical evidence suggesting that output volatility remains subdued despite the turmoil created by the Financial Crisis of 2008.

and C_t follows a three-state Markov-switching process with absorbing states, as specified below:

$$\begin{aligned} p_{C,11} > 0.5; p_{C,12} = 1 - p_{C,11}; p_{C,13} = 0, p_{C,21} = 0, p_{C,22} > 0.5; p_{C,23} = 1 - p_{C,22}; \\ p_{C,31} = 0, p_{C,32} = 0, p_{C,33} = 1, \end{aligned} \quad (50)$$

where $p_{C,ij} = Pr[C_t = j | C_{t-1} = i]$.

Note that the existence of the absorbing states in C_t allows us to identify C_t from the Markov-switching process S_t in our model. Furthermore, the specification of β_{1,C_t} and a_{2,C_t} in equation (48) allows us to handle the label switching problem discussed in Section 3.1. The ordering constraints in the last line of equation (48) guarantees a narrowing gap between mean growth rates for booms and recessions, at the same time guaranteeing that $a_{2,C_t} > 0, \forall t$, allowing us to identify regime 2 (i.e., $S_t = 2$) as boom. A graphical illustration of the resulting implied priors for the mean growth rates during recessions and booms is depicted in Figure 4.

By substituting the first two equations in (48) into the first equation in (47), we obtain

$$\begin{aligned} \Delta y_t = \gamma_1 + \gamma_2 C_{2,t} + (\gamma_2 + \gamma_3) C_{3,t} \\ + ((\eta_1 + \eta_2 + \eta_3) C_{1,t} + (\eta_2 + \eta_3) C_{2,t} + \eta_3 C_{3,t}) S_{2,t} + u_t. \end{aligned} \quad (51)$$

Then, by defining $e_t = \gamma_1 + u_t$ and $\varepsilon_t = (1 - \phi)\gamma_1/g_{W_t} + u_t^*$ and rearranging terms, equation (51) can be rewritten as:

Empirical Model with Transformed Parameters

$$\begin{aligned} \Delta y_t = \gamma_2 \sum_{j=2}^3 C_{j,t} + \gamma_3 C_{3,t} + (\eta_1 C_{1,t} + \eta_2 \sum_{j=1}^2 C_{j,t} + \eta_3 \sum_{j=1}^3 C_{j,t}) S_{2,t} + e_t, \\ e_t = \phi e_{t-1} + g_{W_t} \varepsilon_t, \quad |\phi| < 1, \quad \varepsilon_t | W_t \sim i.i.d. \left(\frac{1}{g_{W_t}} (1 - \phi) \gamma_1, \sigma_1^2 \right), \\ g_n^2 = g_{n-1}^2 (1 + b_n), \quad g_1^2 = 1, \quad b_n > 0, \quad n = 2, 3, \end{aligned} \quad (52)$$

where the unknown distribution of the error term ε_t conditional on W_t is approximated by the Dirichlet Process mixture of normals specified in equation (31). We note that the independence of C_t from S_t or W_t and the existence of the absorbing states for C_t allow us to identify C_t from S_t or W_t . Given the truncated normal prior, each of the γ and η parameters can be sequentially drawn from appropriate truncated normal distributions as explained in Section 3.1, without resorting to the rejection sampling.

Lastly, note that allowing for structural breaks in the mean growth rates for booms and recessions results in structural breaks in the long-run mean growth rate. Based on equation (51), this time-varying long-run mean growth rate (τ_t) at each iteration of the MCMC can be obtained by:

$$\begin{aligned}
\tau_t &= \\
&= \gamma_1 + \gamma_2 Pr[C_t = 2|I_T] + (\gamma_2 + \gamma_3) Pr[C_t = 3|I_T] \\
&+ ((\eta_1 + \eta_2 + \eta_3) Pr[C_t = 1|I_T] + (\eta_2 + \eta_3) Pr[C_t = 2|I_T] + \eta_3 Pr[C_t = 3|I_T]) Pr[S_t = 2],
\end{aligned} \tag{53}$$

where γ_1 can be recovered as in equation (14), with M in equation (14) referring to realized number of mixtures at a particular iteration of the MCMC; I_T refers to information up to T ; and $Pr[S_t = 2]$ refers to the steady-state probability that $S_t = 2$, which is given by $Pr[S_t = 2] = (1 - p_{S,11}) / (2 - p_{S,11} - p_{S,22})$.

6.2. Empirical Results

Data employed is the seasonally-adjusted postwar U.S. industrial production index, which is obtained from the Federal Reserve Bank of St. Louis economic database (FRED), and the sample covers the period 1947M1-2017M1. Figure 3 depicts the data. We estimate both the proposed model and the model with a normality assumption for the error term. We obtain 200,000 MCMC draws and discard the first 50,000 to guarantee the convergence of the sampler and to avoid the effect of the initial values. All the inferences are based on the remaining 150,000 draws. We first consider the following tight priors:

Priors #1: Tight Priors

$$\begin{aligned}
\gamma_2 &\sim N(0.1, 0.1)_{[\gamma_2>0]}, & \gamma_3 &\sim N(0.2, 0.2)_{[\gamma_3>0]}, \\
\eta_1 &\sim N(1.5, 0.1)_{[\eta_1>0]}, & \eta_2 &\sim N(0.5, 0.2)_{[\eta_2>0]}, & \eta_3 &\sim N(0.2, 0.5)_{[\eta_3>0]}, \\
\phi &\sim N(0.5, 0.5)_{[|\phi|<1]}, & (1+b_2) &\sim IG(4, 4), & (1+b_3) &\sim IG(4, 8), \\
[P_{S,11}, P_{S,12}]' &\sim Dir(0.45, 0.05)_{P_{S,11}>0.5}, & [P_{S,21}, P_{S,22}]' &\sim Dir(0.05, 0.45)_{P_{S,22}>0.5}, \\
[P_{W,11}, P_{W,12}, P_{W,13}]' &\sim Dir(0.9, 0.05, 0.05)_{P_{W,11}>0.5}, \\
[P_{W,21}, P_{W,22}, P_{W,23}]' &\sim Dir(0.05, 0.9, 0.05)_{P_{W,22}>0.5}, \\
[P_{W,31}, P_{W,32}, P_{W,33}]' &\sim Dir(0.05, 0.05, 0.9)_{P_{W,33}>0.5}, \\
P_{C,11} &\sim Dir(9.9, 0.1)_{P_{C,11}>0.5}, & P_{C,22} &\sim Dir(9.9, 0.1)_{P_{C,22}>0.5}, \\
(\mu_m, h_m^2) &\sim G_0 \equiv N(-0.5, 3h_m^2)IG(17, 4)
\end{aligned}$$

where the base distribution specified in the last line implies the following unconditional distribution for γ_1 and σ_1^2 :

$$\gamma_1 \sim N(-0.5, 0.2) \quad \text{and} \quad \sigma_1^2 \sim IG(4.2),$$

which are used as the priors for γ_1 and σ_1^2 in the model with a normality assumption.

Table 3.1 reports the posterior moments of the parameters obtained under tight priors for a model with a normality assumption. When we performed a normality test for the posterior means of the standardized errors (i.e., u_t^* , $t = 1, 2, \dots, T$, in the last line of equation (47)), however, the null was rejected at a 5% significance level. This provides a justification for employing the proposed model, in which we approximate the unknown error term with the Dirichlet process mixture of normals. Table 3.2 reports the corresponding posterior moments for the proposed model. For most of the parameters, the posterior standard deviations are larger for the model with the normality assumption than for the proposed model. The posterior mean for the total number of mixture is slightly higher than 3, and the null hypothesis of normality not rejected for the posterior means of the standardized errors.¹⁶ These results suggest that the Dirichlet process mixture normals model reasonably well approximates the unknown distribution of the error term. Furthermore, a Bayesian model selection criterion

¹⁶ To calculate the Jarque-Bera test statistic for the normality test, we use the posterior mean of the standardized errors ($v_t^* = \frac{1}{h_{D_t}}(\varepsilon_t - \frac{1}{g_{W_t}}\mu_{D_t})$) obtained based on equation (34), for $t = 1, 2, \dots, T$.

(Watanabe-Akaike information criterion or WAIC by Watanabe (2010)) strongly prefers the proposed model.

Figures 5.A and 5.B depict the posterior probabilities of recession for the two models under the tight priors. The shaded areas represent the NBER recessions. Estimates of recession probabilities from the proposed model are much sharper and agree much more closely with the NBER reference cycles than those from a model with a normality assumption for the error term.

In order to examine the robustness of the results to the priors employed, we next consider the following loose priors for some of the parameters, by keeping the priors for the rest of the parameters unchanged:

Prior #2: Loose Priors

$$\begin{aligned} \gamma_2 &\sim N(0.1, 1)_{[\gamma_2 > 0]}, & \gamma_3 &\sim N(0.2, 2)_{[\gamma_3 > 0]}, \\ \eta_1 &\sim N(1.5, 1)_{[\eta_1 > 0]}, & \eta_2 &\sim N(0.5, 2)_{[\eta_2 > 0]}, & \eta_3 &\sim N(0.2, 4)_{[\eta_3 > 0]}, \\ [P_{S,11}, P_{S,12}]' &\sim Dir(0.09, 0.01)_{P_{S,11} > 0.5}, & [P_{S,21}, P_{S,22}]' &\sim Dir(0.01, 0.09)_{P_{S,22} > 0.5}, \\ P_{C,11} &\sim Dir(0.99, 0.01)_{P_{C,11} > 0.5}, & P_{C,22} &\sim Dir(0.99, 0.01)_{P_{C,22} > 0.5} \end{aligned}$$

Figure 4.2 compares the posterior probabilities of recession for the two competing models under the loose priors. For the model with a normality assumption in the error term, the inference on the recession probabilities deteriorates considerably with the loose priors. For the proposed model, however, the regime probabilities under the loose priors are almost the same as those under the tight priors, and we continue to have sharp inferences on the recession probabilities. That is, that the proposed model is robust to the priors employed, while the model with a normality assumption is very sensitive to the priors.

Lastly, Figure 6 depicts the posterior means for the time-varying volatility of the errors and those for the long-run mean growth of the IP series obtained based on equation (53), for the proposed model under the tight priors.¹⁷ The high and the medium volatility regimes are mostly focused on the period prior to the mid 1980s. In most of the post-1984 period, the low volatility regime dominates except for a few episodes of medium or high volatility

¹⁷ The results were almost the same as in the case of the loss priors.

that include the Great recession. The second panel of Figure 6 demonstrates a pattern for a steadily decreasing long-run mean growth rate, which is consistent with Stock and Watson (2012) and Eo and Kim (2016).

7. Concluding Remarks

In their dynamic factor models of business cycle, Kim and Yoo (1996), Chauvet (1998), and Kim and Nelson (1998) assume that each individual coincident variable consists of an idiosyncratic component and a common factor component that is subject to Markov-switching mean. They estimate their models either by the maximum likelihood method or by the Bayesian method, under the assumption of normally distributed errors. They all show that their estimates of turning points are much sharper and agree much more closely with the NBER reference cycles than the estimates from a univariate Markov switching model do. The intuition is that the idiosyncratic components that are not related to the business cycle or the outliers in the individual series are averaged out cross sectionally.

Within our univariate Markov-switching framework, approximating the error distribution by the mixture of normals allows us to effectively control for the irregular components or the outliers in the error term. This leads to sharp and accurate inferences on the regime probabilities just like the dynamic factor models do. It also allows the performance of the proposed model to be robust to the priors employed for the parameters of the model. However, the performance of the model with the normality assumption is very sensitive to the priors. For this model, one may not be able to obtain any reasonable inferences on the parameters or the regime probabilities without employing very informative priors.

Appendix A. Derivation of Equation (16)

Equation (8) and the transition probabilities in (4) allow to write the dynamics of the vector $[S_{1,t} \ S_{2,t} \ \dots \ S_{K,t}]'$ as the following VAR form:

$$\begin{bmatrix} S_{1,t} \\ S_{2,t} \\ \vdots \\ S_{K,t} \end{bmatrix} = \begin{bmatrix} p_{S,11} & p_{S,21} & \dots & p_{S,K1} \\ p_{S,12} & p_{S,22} & \dots & p_{S,K2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{S,1K} & p_{S,2K} & \dots & p_{S,KK} \end{bmatrix} \begin{bmatrix} S_{1,t-1} \\ S_{2,t-1} \\ \vdots \\ S_{K,t-1} \end{bmatrix} + \begin{bmatrix} \nu_{1,t} \\ \nu_{2,t} \\ \vdots \\ \nu_{K,t} \end{bmatrix}, \quad (A1)$$

where $[\nu_1 \ \nu_2 \ \dots \ \nu_K]$ is a vector of martingale difference sequences.

As $\sum_{j=1}^K p_{S,ij} = 1$ and $\sum_{j=1}^K S_{jt} = 1$, the first rows in equation (B1) do not carry additional information beyond that contained in the rest of the rows. Thus, by imposing the constraint $S_{1,t-1} = 1 - \sum_{j=2}^K S_{j,t-1}$ on the second through $K - th$ rows of equation (B1), we obtain the following dynamics for $[S_{2,t} \ S_{3,t} \ \dots \ S_{K,t}]'$:

$$\begin{bmatrix} S_{2,t} \\ \vdots \\ S_{K,t} \end{bmatrix} = \begin{bmatrix} p_{S,12} \\ \vdots \\ p_{S,1K} \end{bmatrix} + \begin{bmatrix} (p_{S,22} - p_{S,12}) & \dots & (p_{S,K2} - p_{S,12}) \\ \vdots & \ddots & \vdots \\ (p_{S,2K} - p_{S,1K}) & \dots & (p_{S,KK} - p_{S,1K}) \end{bmatrix} \begin{bmatrix} S_{2,t-1} \\ \vdots \\ S_{K,t-1} \end{bmatrix} + \begin{bmatrix} \nu_{2,t} \\ \vdots \\ \nu_{K,t} \end{bmatrix}, \quad (A2)$$

$$\left(\bar{S}_t = Q_0 + Q_1 \bar{S}_{t-1} + \bar{\nu}_t, \right)$$

which is given in equation (16).

Appendix B. Derivation of Equation (30)

Given the prior for $(p_{D,1}, p_{D,2}, \dots, p_{D,M})$ in equation (29), the marginal distribution of $p_{D,m}$ can be derived as:

$$p_{D,m} \sim \text{Beta}\left(\frac{\alpha}{M}, \frac{\alpha}{M}(M-1)\right), \quad m = 1, \dots, M, \quad (B1)$$

with the following density function:

$$f(p_{D,m}) \propto p_{D,m}^{\frac{\alpha}{M}-1} (1 - p_{D,m})^{\frac{\alpha}{M}(M-1)-1}. \quad (B2)$$

The likelihood of $\tilde{D}_{\neq t}$ given w_m can be expressed as:

$$f(\tilde{D}_{\neq t}|p_{D,m}) \propto p_{D,m}^{T_{m,\neq t}} (1 - p_{D,m})^{T-1-T_{\neq t,m}}, \quad (B3)$$

where $T_{\neq t,m}$ denotes the total number of observations that belong to class m in a sample that excludes period t .

By combining equations (B2) and (B3), we have:

$$\begin{aligned} f(p_{D,m}|\tilde{D}_{\neq t}) &\propto f(p_{D,m})Pr(\tilde{D}_{\neq t}|p_{D,m}) \\ &= p_{D,m}^{T_{m,\neq t} + \frac{\alpha}{M}-1} (1 - p_{D,m})^{T-1-T_{m,\neq t} + \frac{\alpha}{M}(M-1)-1}, \end{aligned} \quad (B4)$$

which suggests that

$$p_{D,m}|\tilde{D}_{\neq t} \sim \text{Beta}(T_{\neq t,m} + \frac{\alpha}{M}, T - 1 - T_{m,\neq t} + \frac{\alpha}{M}(M - 1)). \quad (B5)$$

from which we can derive the following probability of interest in equation (30):

$$\begin{aligned} Pr[D_t = m|\tilde{D}_{\neq t}] &= E(p_{D,m}|\tilde{D}_{\neq t}) \\ &= \frac{T_{m,\neq t} + \frac{\alpha}{M}}{T - 1 + \alpha}. \end{aligned} \quad (B6)$$

Appendix D. Details on Drawing \tilde{D}_T and the Concentration Parameter α conditional on $\tilde{\mu}$, \tilde{h}^2 , and \tilde{e}_T ¹⁸

D.1. Drawing \tilde{D}_T Conditional on α

If the total number of mixtures, M , were fixed as in the case of the finite mixture of normals, it would be straightforward to draw D_t based on the following full conditional distribution of D_t :

¹⁸ This section is based on West et al. (1994), Escobar and West (1995), and Neal (2000).

$$f(D_t|\tilde{\mu}, \tilde{\sigma}^2, \tilde{D}_{\neq t}, \varepsilon_t) \propto f(D_t|\tilde{D}_{\neq t}, \alpha)f(\varepsilon_t|\tilde{\mu}, \tilde{\sigma}^2, D_t), \quad D_t = 1, 2, \dots, M, \quad (C1)$$

where $\tilde{D}_{\neq t}$ is the collection of mixture indicators in the sample excluding D_t ; $f(D_t|\tilde{D}_{\neq t}, \alpha)$ is the prior probability in equation (30); and $f(\varepsilon_t|\tilde{\mu}, \tilde{\sigma}^2, D_t = m) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left[-\frac{(\varepsilon_t - \mu_m)^2}{2\sigma_m^2}\right]$. That is, we could draw D_t based on the following probabilities:

$$Pr[D_t = m|\varepsilon_t, \tilde{\mu}, \tilde{\sigma}^2, \tilde{D}_{\neq t}] = \frac{Pr[D_t = m|\tilde{D}_{\neq t}]f(\varepsilon_t|\tilde{\mu}, \tilde{\sigma}^2, D_t = m)}{\sum_{m=1}^M Pr[D_t = m|\tilde{D}_{\neq t}]f(\varepsilon_t|\mu_m, \sigma_m^2, D_t = m)}, \quad m = 1, 2, \dots, M. \quad (C2)$$

For the Dirichlet process mixture of normals, in which M is a random variable, Neal (2000) suggests that equation (D1) should be replaced by:

$$f(D_t|\tilde{\mu}, \tilde{\sigma}^2, D_{\neq t}, \alpha, \varepsilon_t) \propto f(D_t|\alpha, \tilde{D}_{\neq t})f(\varepsilon_t|\tilde{\mu}, \tilde{\sigma}^2, D_t), \quad D_t = 1, \dots, M_{\neq t}^*, M_{\neq t}^* + 1, \quad (C3)$$

where $M_{\neq t}^*$ is the number of distinctive classes (or mixtures) in the sample that exclude period t ; and $f(D_t|\tilde{D}_{\neq t}, \alpha)$ is the prior probability given in equation (32). Here, when $D_t = M_{\neq t}^* + 1$, it means that period t belongs to a new class that does not exist in $\tilde{D}_{\neq t}$. Given equation (D3), we can then draw D_t using the following probabilities:

$$Pr[D_t = m|\tilde{\mu}, \tilde{\sigma}^2, D_{\neq t}, \alpha, \varepsilon_t] = \frac{Pr[D_t = m|\tilde{D}_{\neq t}, \alpha]f(\varepsilon_t|\tilde{\mu}, \tilde{\sigma}^2, D_t)}{\sum_{m=1}^{M_{\neq t}^*+1} Pr[D_t = m|\tilde{D}_{\neq t}, \alpha]f(\varepsilon_t|\tilde{\mu}, \tilde{\sigma}^2, D_t)}, \quad (C4)$$

$$m = 1, 2, \dots, M_{\neq t}^*, M_{\neq t}^* + 1.$$

Depending on whether D_t belongs to the existing class ($m = 1, 2, \dots$, or $M_{\neq t}^*$) or a new class ($m = M_{\neq t}^* + 1$), we have the following two conditional densities for ε_t :

$$f(\varepsilon_t|\tilde{\mu}, \tilde{\sigma}^2, D_t = m) = f_N(\varepsilon_t|\mu_m, \sigma_m^2), \quad \text{for } m = 1, 2, \dots, M_{\neq t}^*; \quad (C5)$$

$$f(\varepsilon_t|\tilde{\mu}, \tilde{\sigma}^2, D_t = M_{\neq t}^* + 1) = \int f_N(\varepsilon_t|\mu_{M_{\neq t}^*+1}, \sigma_{M_{\neq t}^*+1}^2)dG_0(\mu_{M_{\neq t}^*+1}, \sigma_{M_{\neq t}^*+1}^2), \quad (C6)$$

where $f_N(\cdot|\mu_j, \sigma_j^2)$ refers to a normal density function with mean μ_j and variance σ_j^2 . The intuition for the integral in equation (D6) is that, when period t belongs to a new class of normal with unknown mean and variance, we evaluate the density of ε_t by taking average of the densities for all possible values of mean and variance drawn from the base distribution G_0 . This integral can be evaluated by Monte Carlo simulation as suggested by West et al. (1994).¹⁹

By denoting \tilde{D}_T as a collection of the mixture indicators (or class indicators) drawn from the previous iteration of the MCMC, we can draw D_t by repeating the following steps sequentially for $t = 1, 2, \dots, T$, starting with $t = 1$:

- i) Count the total number of distinctive classes in $\tilde{D}_{\neq t}$ and set it as $M_{\neq t}^*$.
- ii) Draw D_t according to the probabilities in equation (D4), and replace the t -th element of \tilde{D}_T with the drawn D_t .
- iii) If D_t is generated to be $M_{\neq t}^* + 1$, it means that period t belongs to a new class that does not exist in $\tilde{D}_{\neq t}$. In this case, we have to generate intermediate values for the mean ($\mu_{M_{\neq t}^*+1}$) and variance ($\sigma_{M_{\neq t}^*+1}^2$) that are associated with this new class. They can be generated from the following posterior distributions:

$$\sigma_{M_{\neq t}^*+1}^2 | \varepsilon_t \sim IG \left(\frac{1 + d_0}{2}, \frac{v_0 + (\varepsilon_t - \lambda_0/g_{W_t})^2 / (1 + \psi_0/g_{W_t}^2)}{2} \right), \quad (C7)$$

$$\mu_{M_{\neq t}^*+1} | \sigma_{M_{\neq t}^*+1}^2, \varepsilon_t \sim N \left(\frac{\lambda_0 + \psi_0 \varepsilon_t / g_{W_t}}{1 + \psi_0 / g_{W_t}^2}, \frac{\psi_0}{1 + \psi_0 / g_{W_t}^2} \sigma_{M_{\neq t}^*+1}^2 \right), \quad (C8)$$

which can be easily derived given the joint prior G_0 for $(\mu_{M_{\neq t}^*+1}, \sigma_{M_{\neq t}^*+1}^2)$ in equation (31) and a single observation ε_t .

- iv) Set $t=t+1$, and go to i).

¹⁹ The integral in equation (D6) can be approximated by

$$\int f_N(\varepsilon_t | \mu_{M_{\neq t}^*+1}, \sigma_{M_{\neq t}^*+1}^2) dG_0(\mu_{M_{\neq t}^*+1}, \sigma_{M_{\neq t}^*+1}^2) \approx \frac{1}{R} \sum_{i=1}^R f_N(\varepsilon_t | \mu_i, \sigma_i^2),$$

where μ_i and σ_i^2 are drawn from the base distribution G_0 in equation (22) and R is large enough. Alternatively, Escobar and West (1995) analytically derive that this integral results in a density function for a scaled and shifted Student's t-distribution.

At the end of the iteration, we have a new set of \tilde{D}_T . The number of distinctive classes in \tilde{D}_T is the realized M or the realized total number of mixtures.

D.2. Drawing α conditional on \tilde{D}_T

In case α is treated as random, its conjugate prior is the Gamma distribution,

$$\alpha \sim \text{Gamma}(a, b) \tag{C9}$$

Drawing α conditional on \tilde{D}_T is equivalent to drawing α conditional on M , the total number of mixtures or classes in the sample.²⁰ In this section, we explain an algorithm for generating α as proposed by Escobar and West (1995).

Given the prior distribution of α in equation (D9), the prior density is:

$$f(\alpha) \propto \alpha^{a-1} \exp(-ab), \tag{C10}$$

and as derived by Antoniak (1974), the likelihood for M is

$$f(M|\alpha) \propto \alpha^M \frac{\Gamma(\alpha)}{\Gamma(\alpha + T)}, \tag{C11}$$

where $\Gamma(\cdot)$ refers to the Gamma function and T is the sample size. Thus, Escobar and West (1995) derive the posterior density of α as:²¹

$$\begin{aligned} f(\alpha|M) &\propto f(\alpha)f(M|\alpha) \\ &\propto \alpha^{a+M-2} \exp(-ab)(\alpha + T) \int_0^1 x^\alpha (1-x)^{T-1} dx, \end{aligned} \tag{C12}$$

²⁰ Note that the posterior distribution of α depends only on M , for given \tilde{D}_T .

²¹ Note that gamma functions in equation (D11) can be written as

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha + T)} = \frac{(\alpha + T)\beta(\alpha + 1, T)}{\alpha\Gamma(T)},$$

where $\beta(\cdot, \cdot)$ refers to the beta function, and

$$\beta(\alpha + 1, T) = \int_0^1 x^\alpha (1-x)^{T-1} dx$$

which implies that the posterior distribution of α is the marginal distribution obtained from a joint distribution of α and a continuous quantity η such that

$$f(\alpha, \eta|M) \propto \alpha^{a+M-1} \exp(-\alpha b)(\alpha + T)\eta^\alpha(1 - \eta)^{T-1}, \quad 0 < \eta < 1. \quad (C13)$$

As shown in their Appendix B, Escobar and West (1995) further derive the conditional posterior densities $f(\eta|\alpha, M)$ and $f(\alpha|\eta, M)$, and show that

$$\eta|\alpha, M \sim \text{Beta}(\alpha + 1, T) \quad (C14)$$

and

$$\alpha|\eta, M \sim r_\eta G(a + M, b - \ln(\eta)) + (1 - r_\eta)G(a + M - 1, b - \ln(\eta)), \quad (C15)$$

where the latter is a mixture of two Gamma distributions with $r_\eta/(1 - r_\eta) = (a + M - 1)/\{T[b - \ln(\eta)]\}$.

Thus, the following two-step algorithm can be employed to draw α :

- i) Conditional on α generated in the previous iteration of the Gibbs sampling, draw an intermediate random variable η from the distribution given in equation (D14).
- ii) Conditional on η and the realized number of mixture, M , draw α from the distribution given in equation (D15).

References

- [1] Albrt, J. and Chib, S., 2001, "Sequential Ordinal Modeling with Application to Survival data," *Biomtrics*, 57, 829-836.
- [2] Alexander, C., E. Lazar, 2006, "Normal Mixture GARCH(1,1): Applications to Exchange Rate Modeling," *Journal of Applied Econometrics*, 21, 307-336.
- [3] Antoniak, C. E., 1974, "Mixtures of Dirichlet Processes With Applications to Nonparametric Problems," *The Annals of Statistics*, 2, 1152-1174.
- [4] Bauwens, L., J.-F. Carpentier, A. Dufays, 2017, "Autoregressive Moving Average Infinite Hidden Markov-Switching Models," *Journal of Business and Economic Statistics*, 35, 162-182.
- [5] Bulla, Jan, 2011, "Hidden Markov Models with t Components. Increased Persistence and Other Aspects," *Quantitative Finance*, 11, 459475.
- [6] Chauvet, M., 1998, "An Econometric Characterization of Business Cycle Dynamics with Factor Structure and Regime Switches," *International Economic Review*, 39, 969-996.
- [7] Chib, S., 1998, "Estimation and Comparison of Multiple Change-Point Models," *Journal of Econometrics*, 86, 221-241.
- [8] Clark, Todd E., 2009, "Is the Great Moderation Over? An Empirical Analysis," *Economic Review*, Fourth Quarter 2009, Federal Reserve Bank of Kansas City, 5-42.
- [9] De Angelis, Luca and Viroli, Cinzia, 2017, "A Markov-Switching Regression Model with non-Gaussian Innovations: Estimation and Testing," *Studies in Nonlinear Dynamics & Econometrics*, Vol. 21, Issue 2, 1081-1826,
- [10] Diebold, F. X., J.-H. Lee, and G. C. Weinbach, 1994, "Regime Switching with Time-Varying Transition Probabilities," In C. Hargreaves, Ed.: *Nonstationary Time Series Analysis and Cointegration*, 283-302. Oxford University Press.
- [11] Dueker, M., 1997, "Markov Switching in GARCH Processes and Mean-Reverting Stock-

- Market Volatility,” *Journal of Business and Economic Statistics*, 15, 26-34.
- [12] Eo, Y. and C-J. Kim, 2016, “Markov-Switching Models with Evolving Regime-Specific Parameters: Are Postwar Booms and Recessions All Alike?,” *Review of Economics and Statistics*, 98, 940-949.
- [13] Escobar, M. D., and M. West, 1995, “Bayesian Density Estimation and Inference Using Mixtures,” *Journal of the American Statistical Association*, 90, 577-588.
- [14] Filardo, A. J., 1994, “Business Cycle Phases and Their Transitional Dynamics,” *Journal of Business and Economic Statistics*, 12, 299-308.
- [15] Fox, E., E. Sudderth, M. Jordan, and A. Willsky, 2011, “A Sticky HDP-HMM with Application to Speaker Diarization,” *Annals of Applied Statistics*, 5, 1020-1056.
- [16] Frühwirth-Schnatter, S., 2001, “Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models,” *Journal of the American Statistical Association*, Vol. 96, No. 453, 194-209.
- [17] Frühwirth-Schnatter, S., 2006, “Finite Mixture and Markov Switching Models. Springer Series in Statistics,” Springer, New York.
- [18] Gadea-Rivas, Maria D., Ana Gomez-Lscos, and Gabriel Perez-Quiros, 2014, “The Two Greatest Recession vs. Great Moderation,” Banco de Espana Working Paper No. 1423.
- [19] Gorur, D., and C. E. Rasmussen, 2010, “Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution,” *Journal of Computer Science and Technology*, 25, 615-626.
- [20] Hamilton, J. D., 1989, “A New Approach to the Economic Analysis of Nonstationary Time Series and The Business Cycle,” *Econometrica*, 57, 357-384.
- [21] Hamilton, J. D., 2016, “Macroeconomic Regimes and Regime Shifts,” *Handbook of Macroeconomics*, Vol. 2, 163-201.
- [22] Jensen, M. and J. Maheu, 2010, “Bayesian Semiparametric Stochastic Volatility Modeling,” *Journal of Econometrics*, 157, 306-316.
- [23] Jensen, M. and J. Maheu, 2013, “Bayesian Semiparametric Multivariate GARCH Mod-

- eling,” *Journal of Econometrics*, 176, 3-17.
- [24] Jensen, M, and J. Maheu, 2014, “Estimating a Semiparametric Asymmetric Stochastic Volatility Model with a Dirichlet Process Mixture,” *Journal of Econometrics*, 178, 523-538.
- [25] Jin, X. and J. Maheu, 2016, “Bayesian Semiparametric Modeling of Realized Covariance Matrices,” *Journal of Econometrics*, 192, 19-39.
- [26] Kaufmann, S., 2015, “K-state switching models with time-varying transition distributions - Does loan growth signal stronger effects of variables on Inflation?” *Journal of Econometrics*, 187, 82-94.
- [27] Kim, C-J., 1994, “Dynamic Linear Models with Markov Switching,” *Journal of Econometrics*, 60, 1-22.
- [28] Kim, C-J. and C. Nelson, 1998, “Business Cycle Turning Points, A New Coincident Index, and Tests of Duration Dependence Based on A Dynamic Factor Model with Regime-Switching,” *Review of Economics and Statistics*, 80, 188-201.
- [29] Kim, C-J. and C. Nelson, 1999, “Has the U.S. Economy Become More Stable? A Bayesian Approach Based on a Markov-Switching Model of the Business Cycle,” *Review of Economics and Statistics*, 81, 608-616.
- [30] Kim, S., N. Shephard, and S. Chib, 1998, “Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models,” *Review of Economic Studies*, 65, 361-393.
- [31] Kim, M.-J. and J.-S. Yoo, 1995, “New Index of Coincident Indicators: A Multivariate Markov Switching Factor Model Approach,” *Journal of Monetary Economics*, 36, 607-630.
- [32] Marin, J.-M., K. Mengersen, C.P. Robert, 2005, “Bayesian Modeling and Inference on Mixtures of Distributions,” *Handbook of Statistics*, 25, 459-507.
- [33] Neal, R. M., 2000, “Markov Chain Sampling Methods for Dirichlet Process Mixture Models,” *Journal of Computational and Graphical Statistics*, 9, 249-265.
- [34] Omori, Y., S. Chib, N. Shephard, and J. Nakajima, 2007, “Stochastic Volatility with

- Leverage: Fast and Efficient Likelihood Inference,” *Journal of Econometrics*, 140, 425-449.
- [35] Stephens, M., 2000, “Dealing with Label Switching in Mixture Models,” *Journal of the Royal Statistical Society, Series B*, 62, Part 4, 795-809.
- [36] Song, Y., 2014, “Modeling Regime Switching and Structural Breaks with an Infinite Hidden Markov Model,” *Journal of Applied Econometrics*, 29, 825-842.
- [37] Stock, J. and M. Watson, 2012, “Disentangling the Channels of the 2007-09 Recession,” *Brookings Papers on Economic Activity*, 81-156.
- [38] Watanabe, S., 2010, “Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory,” *Journal of Machine Learning Research*, 11, 3571-3594.
- [39] West, M., P. Muller, and M. D. Escobar, 1994, “Hierarchical Priors and Mixture Models, With Application in Regression and Density Estimation,” in *Aspects of Uncertainty: A Tribute to D.V. Lindley*, eds. P. R. Freeman and A. F. M. Smith, New York: Wiley, 363-386.
- [40] White, H., 1994, *Estimation, Inference, and Specification Analysis*, New York: Cambridge University Press.

Table 1. Maximizing the Normal Log Likelihood Function When the Error Distribution Is Potentially Non-normal: Monte Carlo Experiment

$T = 500$					
	True	Case #1	Case #2	Case #3	Case #4
β_1	-0.6	-0.609 (0.130)	-0.656 (0.555)	-0.983 (0.780)	-1.121 (0.892)
β_2	0.7	0.708 (0.088)	0.707 (0.083)	0.713 (0.117)	0.7213 (0.127)
σ^2	1.1	1.087 (0.090)	1.074 (0.140)	0.967 (0.210)	0.907 (0.264)
p_{11}	0.9	0.892 (0.049)	0.892 (0.056)	0.788 (0.203)	0.754 (0.228)
p_{22}	0.95	0.942 (0.029)	0.946 (0.023)	0.928 (0.038)	0.920 (0.043)
$T = 5,000$					
	True	Case #1	Case #2	Case #3	Case #4
β_1	-0.6	-0.602 (0.039)	-0.610 (0.041)	-0.742 (0.246)	-0.959 (0.629)
β_2	0.7	0.701 (0.025)	0.696 (0.024)	0.730 (0.046)	0.735 (0.088)
σ^2	1.1	1.100 (0.027)	1.100 (0.049)	1.013 (0.107)	0.944 (0.176)
p_{11}	0.9	0.899 (0.012)	0.902 (0.012)	0.848 (0.098)	0.788 (0.169)
p_{22}	0.95	0.949 (0.007)	0.951 (0.006)	0.934 (0.020)	0.921 (0.033)
$T = 50,000$					
	True	Case #1	Case #2	Case #3	Case #4
β_1	-0.6	-0.602 (0.013)	-0.610 (0.025)	-0.694 (0.096)	-0.758 (0.169)
β_2	0.7	0.700 (0.009)	0.695 (0.024)	0.734 (0.035)	0.764 (0.065)
σ^2	1.1	1.100 (0.008)	1.098 (0.038)	1.028 (0.074)	0.966 (0.137)
p_{11}	0.9	0.900 (0.004)	0.901 (0.029)	0.868 (0.033)	0.835 (0.069)
p_{22}	0.95	0.950 (0.002)	0.951 (0.030)	0.936 (0.015)	0.920 (0.031)

Note:

1. This table reports quasi maximum likelihood estimation results under different error distributions. Each cell contains the average of the 1,000 point estimates for each parameter and the root mean squared error of the estimates from the true value (in parentheses).
2. Case #1: normal distribution; Case #2: t-distribution; Case #3: χ^2 distribution; Case #4: mixture of 3 normals.

Table 2. Performance of the Proposed Algorithm [Simulation Study]

Data Generating Process

$$y_t = \beta_{S_t} + \sigma \varepsilon_t, \quad \varepsilon_t \sim i.i.d.(0, 1), \quad S_t = 1, 2; \quad t = 1, 2, \dots, T,$$

$$Pr(S_t = j | S_{t-1} = i) = p_{s,ij}; \quad i, j = 1, 2,$$

$$\varepsilon_t | D_t \sim i.i.d.N(\mu_{D_t}^*, h_{D_t}^{*2}), \quad D_t = 1, 2, 3,$$

$$\mu_1^* = 1.05, \quad \mu_2^* = 0.1, \quad \mu_3^* = -1.35; \quad h_1^{*2} = 0.2, \quad h_2^{*2} = 0.05, \quad h_3^{*2} = 1.695;$$

$$p_{D,1} = 0.2, \quad p_{D,2} = 0.6, \quad p_{D,3} = 0.2$$

<u>Parameters</u>	<u>True Value</u>	<u>Average Posterior Mean and RMSE</u>	
		<u>MCMC(1)</u>	<u>MCMC(2)</u>
β_1	-0.6	-1.113 (0.856)	-0.601 (0.088)
β_2	0.7	0.682 (0.132)	0.703 (0.069)
σ_1^2	1.1	0.923 (0.231)	1.156 (0.168)
$p_{s,11}$	0.9	0.774 (0.172)	0.897 (0.032)
$p_{s,22}$	0.95	0.924 (0.034)	0.944 (0.018)

Note:

1. MCMC(1) denotes the MCMC estimation result based on an erroneous assumption that the error term is normally distributed. MCMC(2) denotes the MCMC estimation result that accounts the non-normality in the error term. Each cell contains the average of the 100 posterior means and root mean squared errors (in parentheses).
2. Each MCMC posterior mean is calculated obtained based on 100,000 MCMC draws after 50,000 burn-in draws.

Table 3.1. Bayesian Inference of the Model with Normality Assumption [Log Difference of the U.S. Industrial Production Index, 1947M1-2017M1]: Tight Priors

Parameter	Prior		Posterior		90% HPDI	
	Mean	SD	Mean	SD		
γ_1			-0.745	0.346	-0.767	[-1.280,-0.131]
γ_2			0.393	0.261	0.365	[0.033,0.847]
γ_3			0.272	0.216	0.221	[0.212,0.694]
η_1			1.178	0.431	1.143	[0.545,1.903]
η_2			0.461	0.234	0.430	[0.117,0.892]
η_3			0.365	0.201	0.364	[0.044,0.687]
ϕ			0.205	0.066	0.202	[0.102,0.318]
σ_1^2			0.505	0.027	0.504	[0.463,0.551]
σ_2^2			0.847	0.112	0.834	[0.687,1.049]
σ_3^2			2.026	0.273	1.991	[1.655,2.516]
$p_{s,00}$			0.872	0.076	0.885	[0.730,0.965]
$p_{s,11}$			0.961	0.034	0.967	[0.917,0.997]
$p_{w,11}$			0.973	0.083	0.907	[0.706,0.966]
$p_{w,12}$			0.009	0.022	0.000	[0.000,0.056]
$p_{w,21}$			0.084	0.079	0.064	[0.000,0.244]
$p_{w,22}$			0.882	0.083	0.907	[0.706,0.966]
$p_{w,32}$			0.210	0.096	0.202	[0.062,0.385]
$p_{w,33}$			0.764	0.087	0.775	[0.599,0.889]
$p_{c,11}$			0.954	0.062	0.982	[0.823,0.999]
$p_{c,22}$			0.995	0.014	0.998	[0.982,1.000]

Table 3.1 (Continued).

WAIC	1807
JB	7.96 (0.019)

Note:

1. Out of 150,000 MCMC draws, the first 50,000 are discarded and inferences are based on the remaining 100,000 draws.
2. SD refers to standard deviation.
3. HPDI refers to a highest posterior density interval.
4. WAIC refers to the Watanabe-Akaike Information Criterion.
5. JB refers to the Jarque-Bera test statistic for a normality test. In the parenthesis is the p-value.

Table 3.2. Bayesian Inference of the Proposed Model [Log Difference of the U.S. Industrial Production Index, 1947M1-2017M1]: Tight Priors

Parameter	Prior		Posterior		90% HPDI	
	Mean	SD	Mean	SD		
γ_1			-0.696	0.195	-0.701	[-1.004,-0.380]
γ_2			0.298	0.181	0.286	[0.032,0.614]
γ_3			0.143	0.117	0.116	[0.010,0.371]
η_1			0.473	0.220	0.464	[0.129,0.847]
η_2			0.309	0.148	0.292	[0.098,0.586]
η_3			0.601	0.124	0.610	[0.383,0.786]
ϕ			0.123	0.046	0.123	[0.050,0.199]
σ_1^2			0.480	0.025	0.478	[0.441,0.523]
σ_2^2			0.790	0.110	0.782	[0.626,0.980]
σ_3^2			1.894	0.227	1.874	[1.557,2.296]
$p_{s,00}$			0.877	0.037	0.881	[0.811,0.931]
$p_{s,11}$			0.967	0.016	0.969	[0.944,0.983]
$p_{w,11}$			0.971	0.022	0.975	[0.932,0.996]
$p_{w,12}$			0.008	0.022	0.000	[0.000,0.056]
$p_{w,21}$			0.084	0.088	0.062	[0.000,0.273]
$p_{w,22}$			0.888	0.088	0.916	[0.690,0.970]
$p_{w,32}$			0.188	0.094	0.184	[0.014,0.351]
$p_{w,33}$			0.780	0.080	0.790	[0.632,0.894]
$p_{c,11}$			0.991	0.020	0.996	[0.971,1.000]
$p_{c,22}$			0.997	0.003	0.998	[0.992,1.000]

Table 3.2 (Continued).

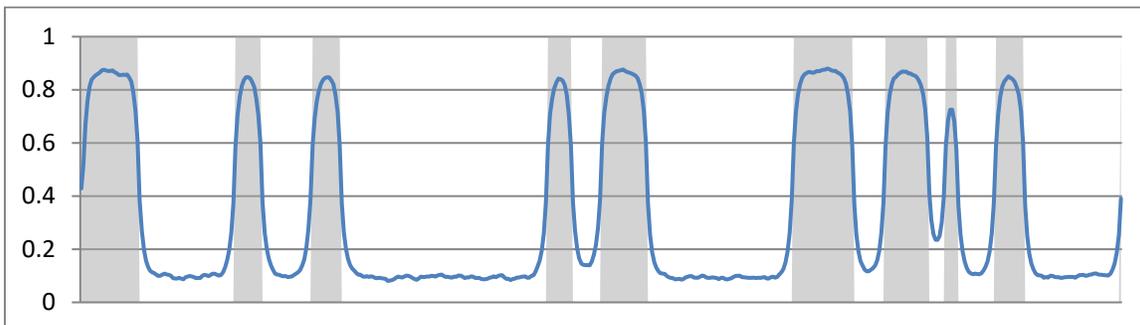
M^*	3.343 (1.273)
WAIC	1717
JB	2.614 (0.503)

Note:

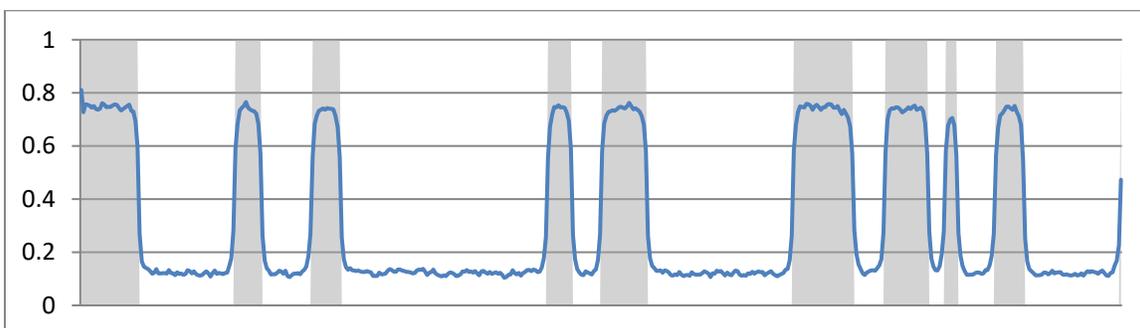
1. Out of 150,000 MCMC draws, the first 50,000 are discarded and inferences are based on the remaining 100,000 draws.
2. SD refers to standard deviation.
3. HPDI refers to a highest posterior density interval.
4. M^* refers to the posterior average number of non-empty mixtures, standard deviation is reported in parenthesis.
5. WAIC refers to the Watanabe-Akaike Information Criterion.
6. JB refers to the Jarque-Bera test statistic for a normality test. P-value is reported in the parenthesis.

Figure 1. Smoothed Probabilities of Regime 1 based on Maximum Likelihood Estimation under Normality Assumption for the Error Term [T=500]

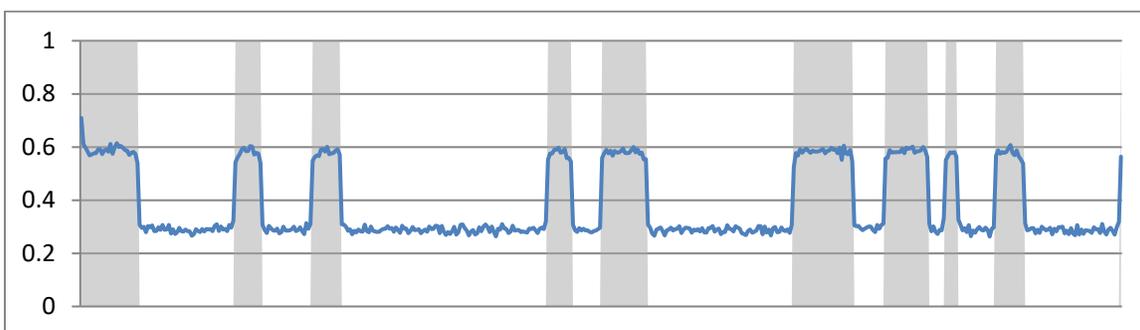
(i) DGP #1: Normal Distribution



(ii) DGP #2: Student's t Distribution



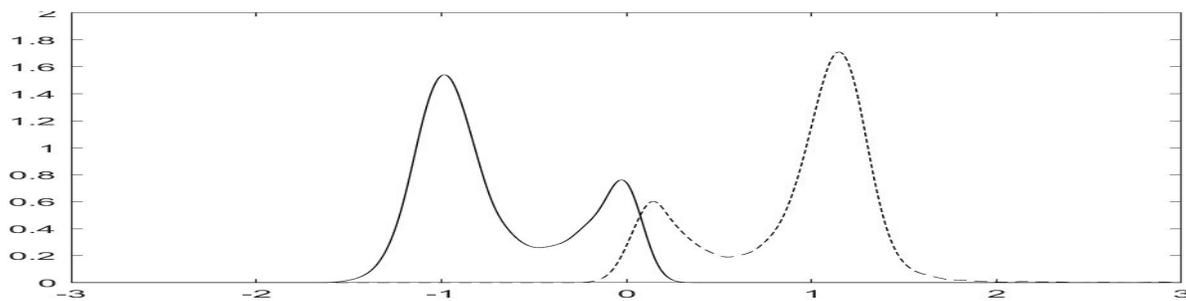
(iii) DGP #3: Log Chi-Square Distribution



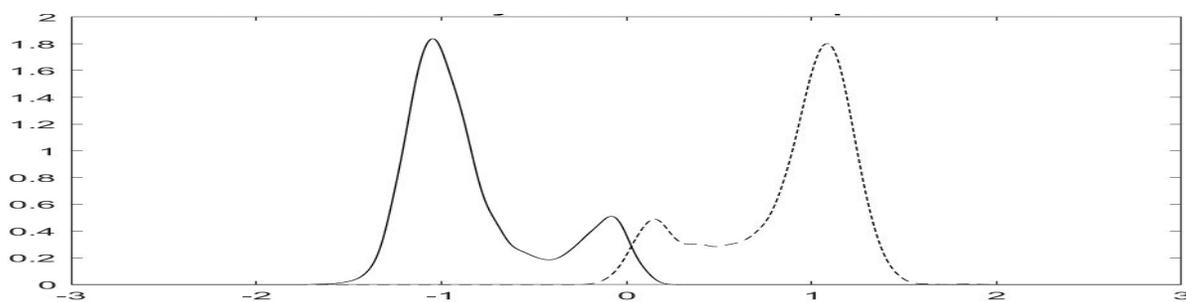
(iv) DGP #4: Mixture of Normals

Note: The shaded area denotes the data periods associated with regime 1.

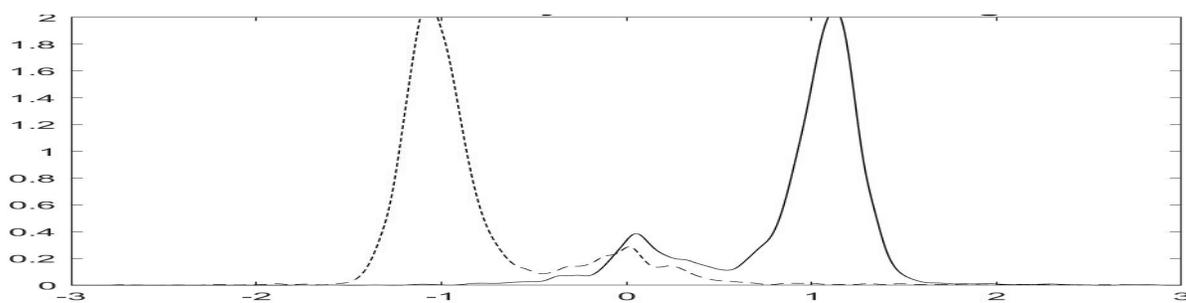
Figure 2. Dealing with the Label Switching Problem: Posterior distributions for β_1 and β_2 based on different sampling methods. [Solid Line: for β_1 ; Broken Line: for β_2]



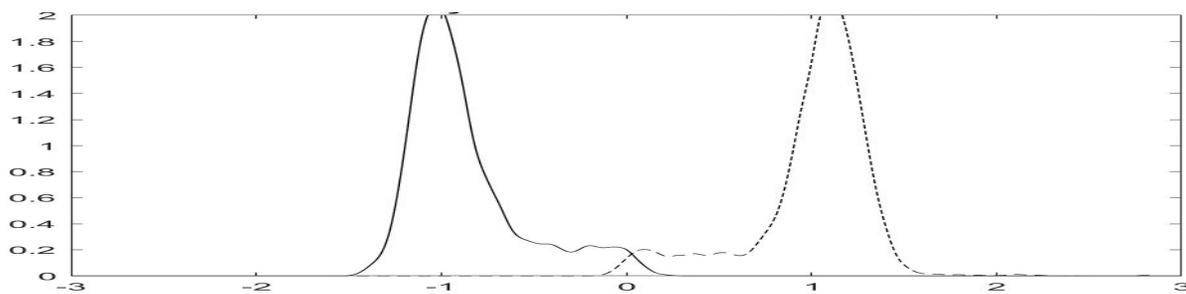
(i) Sampling without Ordering Constraint



(ii) Sampling with Ordering Restriction: Rejection Sampling

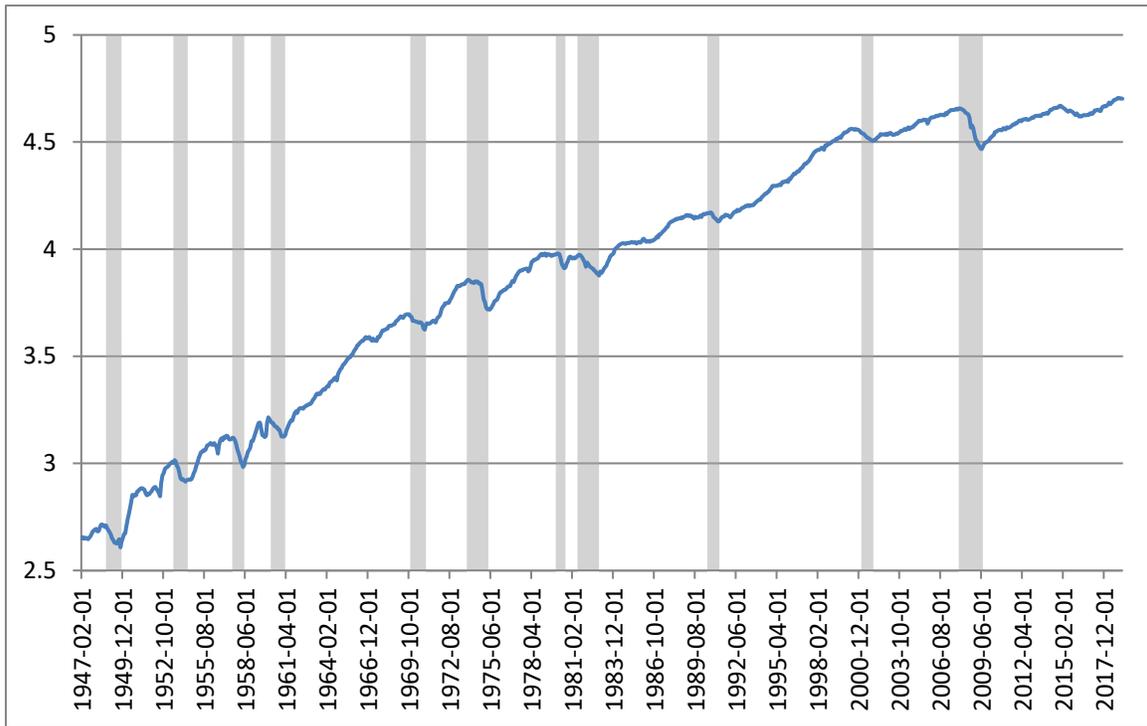


(iii) Sampling with Ordering Constraint: Permutation Sampling

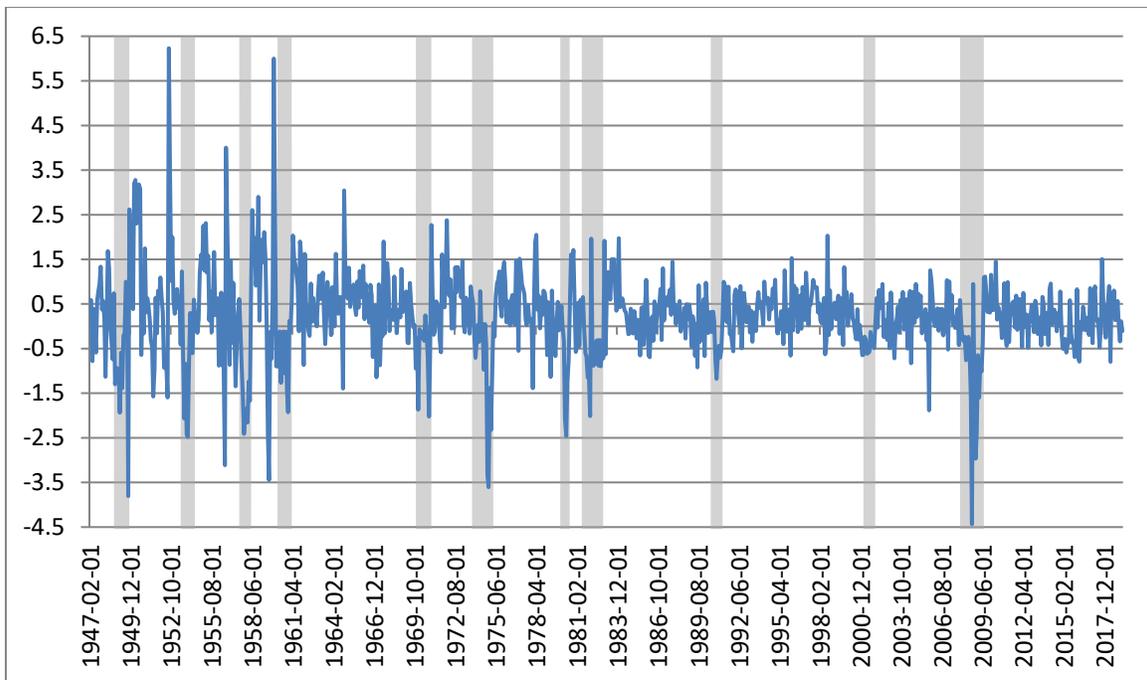


(a) Sampling with Ordering Restriction: Proposed Sampling

Figure 3. U.S. Industrial Production (IP) Index and Its Growth Rate [1947M1-2019M3]



(i) Logarithm of Seasonally Adjusted Industrial Production (IP) Index

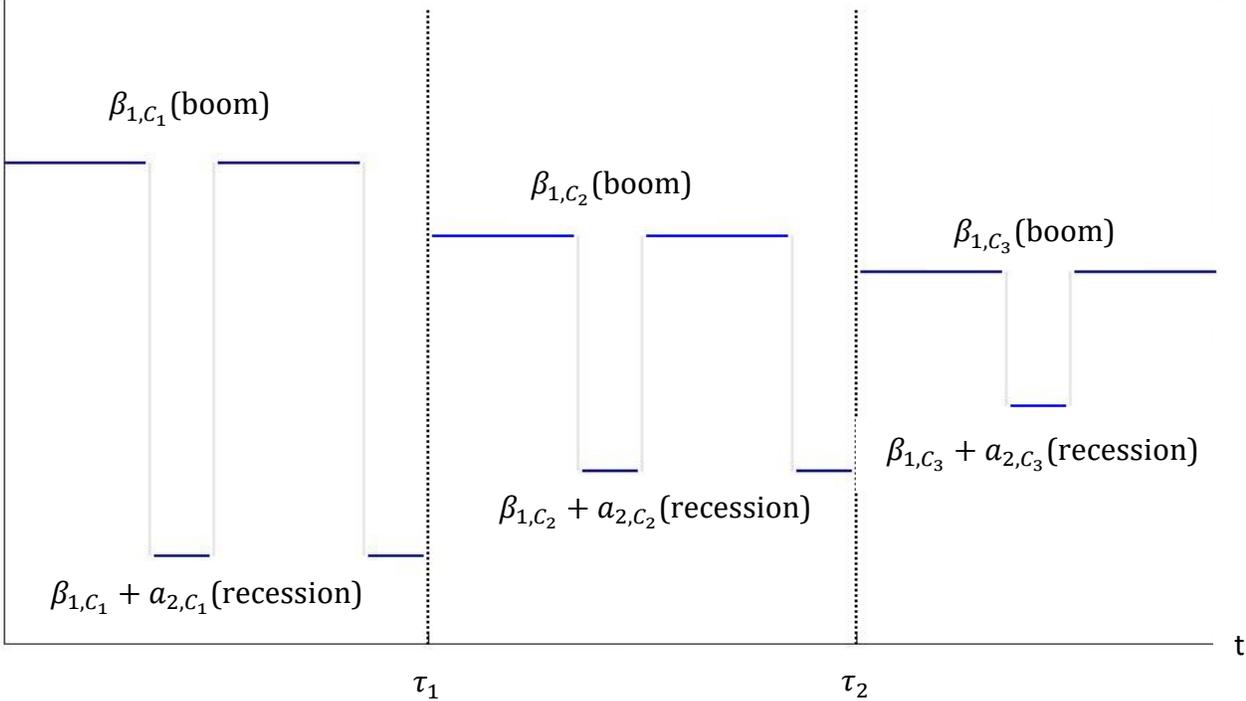


(ii) IP Index growth

Note: The shaded area denotes the NBER recession date.

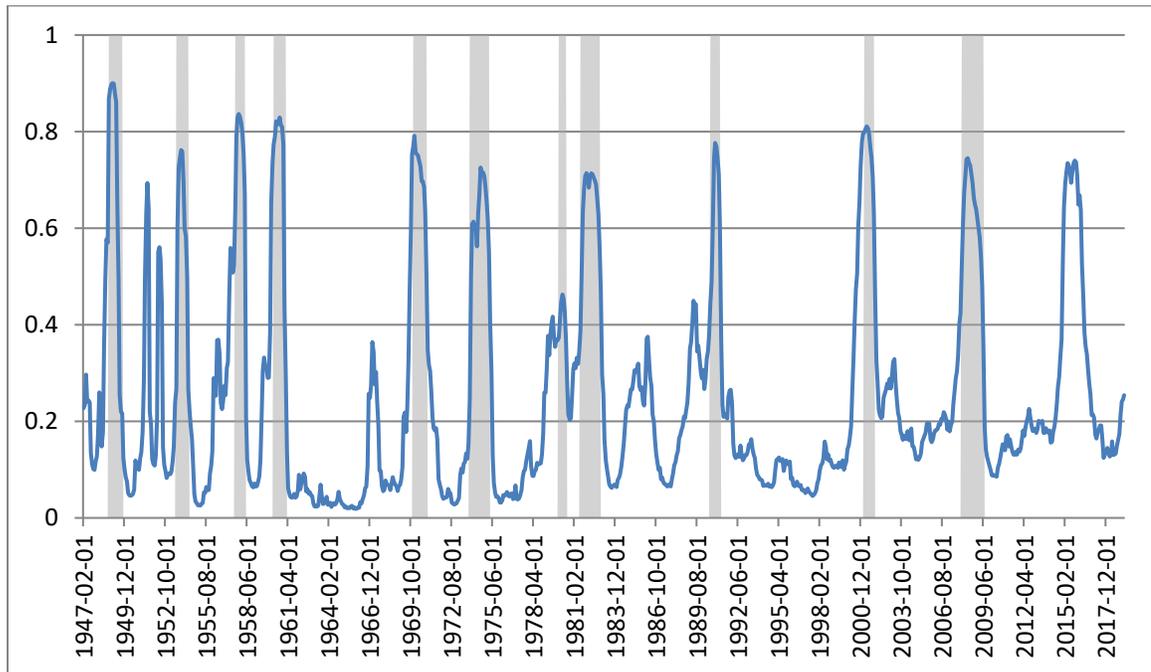
Figure 4. Graphical Illustration of the Priors for the Narrowing Gap between Mean Growth rates During Boom and Recession

Mean Growth Rates

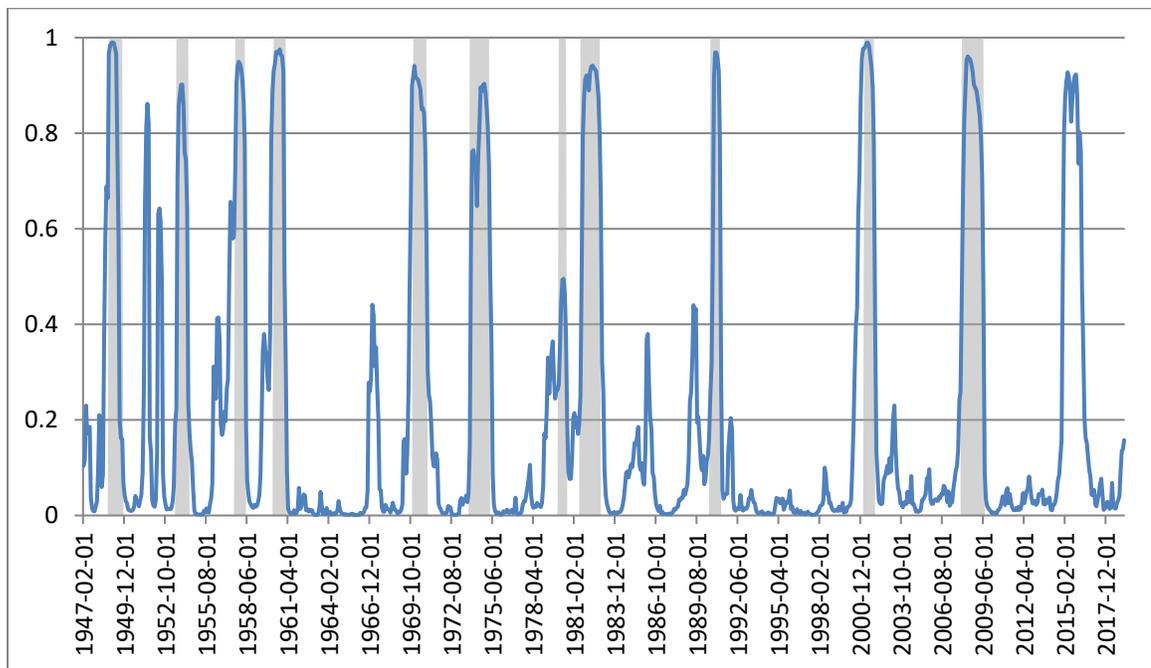


First Structural Break Point Second Structural Break Point
 (Unknown) (Unknown)

Figure 5.A. Posterior Probabilities of Recession based on the Two Competing Models : Tight Priors



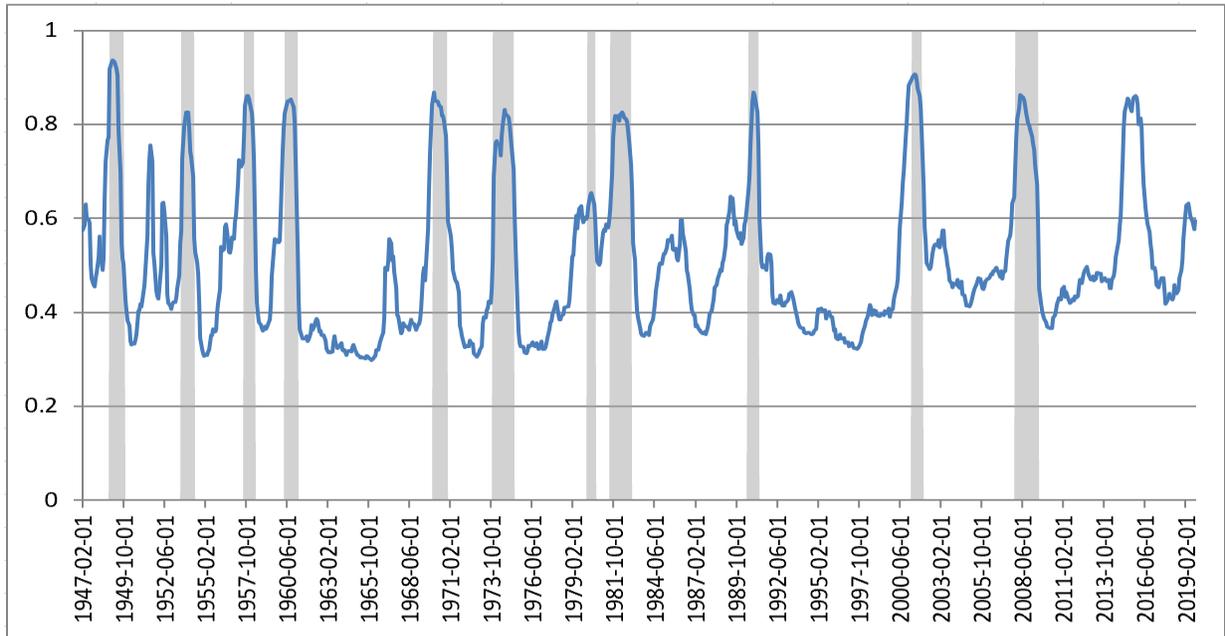
(i) Model with Normally Distributed Errors



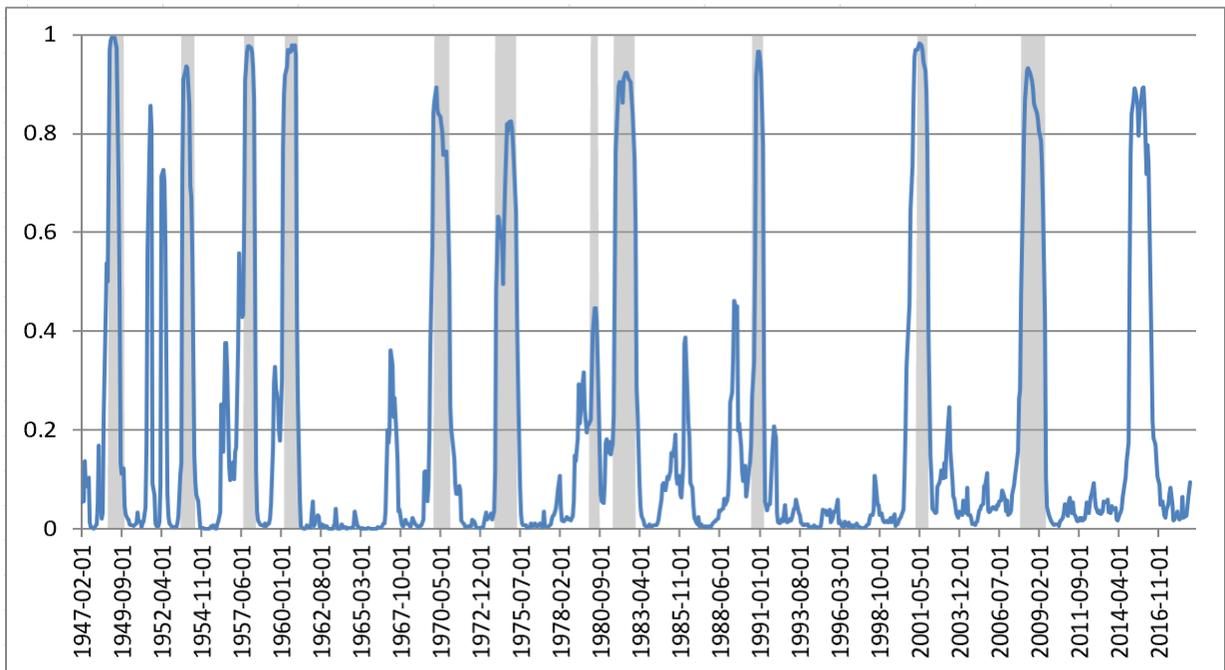
(ii) Model with Mixture of Normals for Error Terms: Proposed Model

Note: The shaded area denotes the NBER recession date.

Figure 5.B. Posterior Probabilities of Recession on the Two Competing Models: Loose Priors



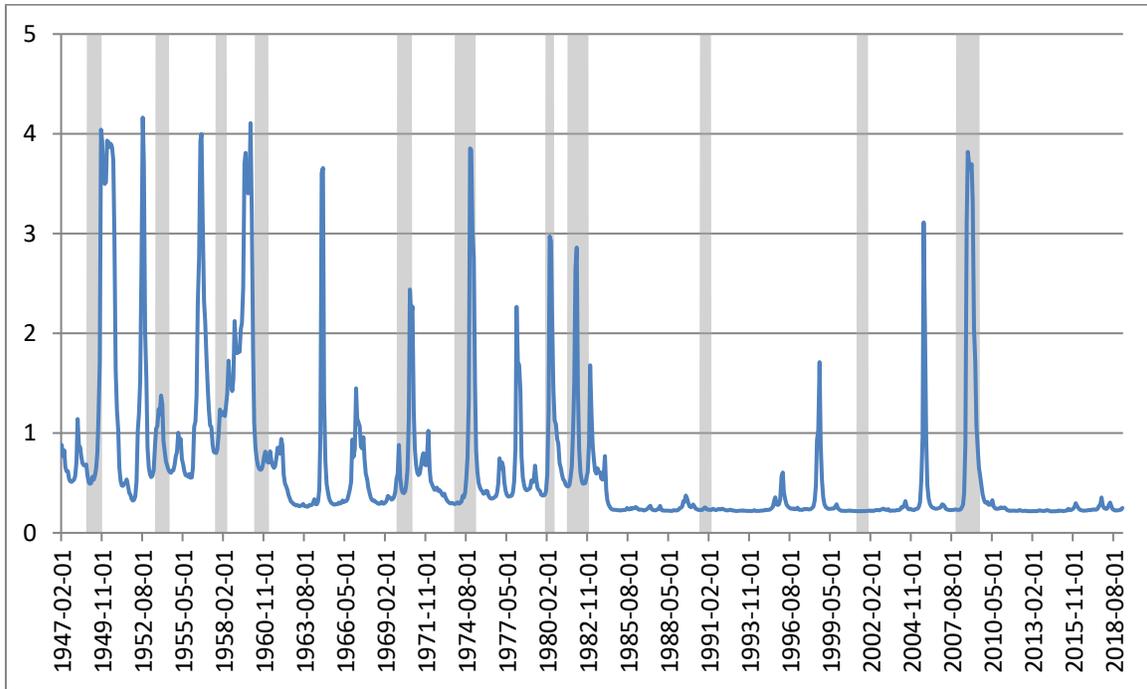
(i) Model with Normally Distributed Errors



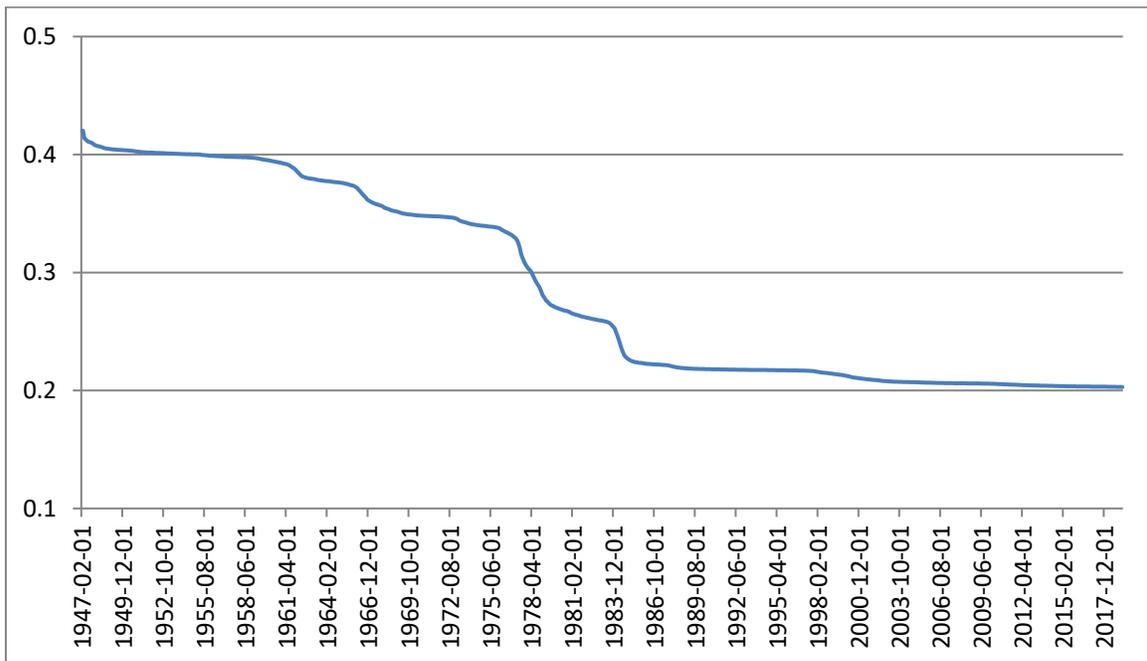
(ii) Model with Mixture of Normals for Error Terms: Proposed Model

Note: The shaded area denotes the NBER recession date.

Figure 6. Time-Varying Volatility and Long-Run Mean Growth Rate of IP: Proposed Model [Tight Priors]



(i) Volatility of Error Term



(ii) Long-Run Mean Growth Rate