A FLEXIBLE PRIOR DISTRIBUTION FOR MARKOV SWITCHING AUTOREGRESSIONS WITH STUDENT-T ERRORS

Philippe J. Deschamps

Université de Fribourg

Final revision, January 2005

Abstract. This paper proposes an empirical Bayes approach for Markov switching autoregressions that can constrain some of the state-dependent parameters (regression coefficients and error variances) to be approximately equal across regimes. By flexibly reducing the dimension of the parameter space, this can help to ensure regime separation and to detect the Markov switching nature of the data. The permutation sampler with a hierarchical prior is used for choosing the prior moments, the identification constraint, and the parameters governing prior state dependence. The empirical relevance of the methodology is illustrated with an application to quarterly and monthly real interest rate data.

Séminaire d'économétrie, Avenue de Beauregard 13, CH-1700 Fribourg, Switzerland. Telephone: +41-26-300-8252. Telefax: +41-26-481-3845.

E-mail address: philippe.deschamps@unifr.ch

JEL classification: C11;C15;C22;C52;E43.

Key words and phrases. Hidden Markov models; empirical Bayes prior; truncated inverted gamma; permutation sampler; real interest rate.

1. Introduction

Markov switching regression models have been proposed as appropriate explanations of the growth rate of real GNP (Hamilton, 1989; Chib, 1996; Frühwirth-Schnatter, 2001), of the growth rate of exchange rates (Engel and Hamilton, 1990), and have also been used to model real interest rates (Garcia and Perron, 1996; Hamilton, 1994b). In the recent literature, the practicality of Markov chain Monte Carlo (MCMC) simulation has led several authors to favor Bayesian methods for estimating these models.

The Bayesian estimation of Markov switching regression models precludes an improper prior on the state-specific parameters, since this would lead to an improper posterior (however, an improper prior on a state-invariant parameter might be used). On this topic, see, e.g., Frühwirth-Schnatter (2001, Section 2.2.1). A second issue is the identification of Markov switching models: without a prior inequality restriction on some Markov switching parameters, a multimodal posterior is obtained. A poor choice of this prior restriction will cause misspecification.

An elegant solution to the second problem is the permutation sampler, proposed by Frühwirth-Schnatter (2001). In a version of this method, each pass of a Gibbs or Metropolis-Hastings sampler is followed by a random permutation of the regime definitions. This version is known as the random permutation sampler. Scatter plots of the simulation output can then be used to suggest an appropriate identification constraint, such as $\theta_1 > \theta_2 > \cdots > \theta_K$, where θ_i is a particular population parameter in regime i. The model can then be reestimated under this constraint by enforcing the corresponding permutation of the regimes; this is known as the constrained permutation sampler (it is easy to check that implementing this sampler is equivalent to truncating the prior distribution on the support defined by the identification constraint). At this stage, a phenomenon known as label switching might indicate that the inequality restriction poorly characterizes the data, or that the assumed number K of regimes is too large.

The permutation sampler requires a prior that is invariant with respect to relabeling. Unfortunately, this may put a large prior mass near a subspace where the state-specific parameters are equal. When the likelihood does not strongly dominate the prior, this can totally mask regime separation and the Markov switching nature of the data (and make very difficult the task of selecting a proper identification constraint). Such a situation can occur with a fairly large sample when the number of observations in a particular regime is low.

As this paper will illustrate, this difficulty can often be avoided by allowing the prior dependence of some state-specific parameters. For example, if the prior correlation coefficient between any two

The author is indebted to the reviewers for valuable comments, including the suggestion that mixture priors might be preferable to the Bayes factor approach that was used in a previous version of this paper. Any remaining errors or shortcomings are the author's responsibility.

elements of $(\theta_1, \ldots, \theta_K)$ is close to 1, these parameters are constrained to be approximately equal across regimes; this may allow the likelihood to overrule the prior for those other parameters that are, in fact, subject to regime shifts. On the other hand, a negative prior correlation coefficient on regime-switching parameters might also help in identifying such shifts. The problem, of course, is to choose the prior dependence in a meaningful way. Ideally, any prior correlation coefficient should be treated as an unknown random parameter, with a probability distribution that can be mostly determined from the data.

This paper proposes a prior for the Markov switching autoregression model that fully addresses the issues mentioned in the two previous paragraphs. First, the prior distributions on the regression coefficients are mixtures of equicorrelated multivariate normals, with correlation coefficients that are uniformly distributed. Secondly, the state-specific error variances are modeled by introducing latent scale variables that can be interpreted as percentages of the total variance corresponding to the prevailing regime. A mixture of Dirichlet prior distributions on these scale variables is proposed. The conditional Dirichlet distributions can be uniform on the unit simplex, and can also concentrate the prior mass on a region where the error variances are approximately equal.

The model also incorporates the possibility of heavy-tailed disturbances. This is done by introducing other latent scale variables that have identical, independent, and state-invariant inverted gamma distributions; this formulation is similar to the one used by Geweke (1993). The equation errors then follow Student-t distributions with Markov switching variances and with a common unknown degrees of freedom parameter, which is restricted to be greater than an arbitrarily chosen prior lower bound.

The plan of the paper is as follows. In Section 2, the model and prior specification are presented. Section 3 discusses the algorithm for simulating the joint posterior, using the random permutation sampler with the mixture prior. Implementing this algorithm involves drawing candidates from a truncated inverted gamma density, which is generalized in the sense that negative degrees of freedom are allowed; a very efficient mixed rejection algorithm for drawing such candidates is described in the Appendix. Section 4 illustrates the potential importance of the new prior by comparing it with the usual independent prior, using the random permutation sampler with simulated data. Section 5 discusses the special issues that arise in an implementation of the constrained permutation sampler; an empirical Bayes approach is proposed where the prior parameters are chosen as point estimates obtained from the random permutation results. Sections 6 and 7 present applications to US quarterly and monthly real interest rate data. Section 8 concludes.

2. The model and the prior

A regression model allowing for structural change and departures from normality may be written as:

$$y_t = \sum_{j=1}^{p} x_{tj}\beta_{S_t}^j + \left(h_{S_t} w_t \sigma^2\right)^{\frac{1}{2}} u_t \quad \text{for } t = 1, \ldots, T \tag{2.1}$$

where S_t is a discrete random variable with $S_t \in \{1, \ldots, K\}$, and where w_t is a continuous random variable. h_{S_t} and $\beta_{S_t}^j$ take the values $h_{S_t} = h_i$ and $\beta_{S_t}^j = \beta_i^j$ if $S_t = i$. We assume that $h_i > 0$ for all i, and that $\sum_{i=1}^{K} h_i = 1$. Conditionally on the x_{tj}, on S_t, and on w_t, the disturbance u_t is standard normal. The covariates x_{tj} can include lagged values of y_t, constant terms, deterministic trends, and current and lagged values of exogenous variables.

If $w_t = 1$ for all t and if $S_1, \ldots, S_T$ follow a Markov process, this is a standard regression model with Markov switching coefficients $\beta_{S_t}^j$ and with Markov switching innovation variance $\sigma^2 h_{S_t}$. So, σ^2 is the sum of the state-specific innovation variances and $\sigma^2 h_{S_t}$ is a state-specific variance component.

If $\beta_1^j = \cdots = \beta_K^j = \beta^j$ for all j and $h_1 = \cdots = h_K = 1/K$, and if the w_t follow identical independent inverted gamma distributions with parameters $a = \nu/2$ and $b = \nu/2$ (see Bernardo and Smith, 2000, p. 119), equation (2.1) implies that the observations y_t have conditional Student-t distributions with expectation $\sum_{j=1}^{p} x_{tj}\beta^j$, constant scale parameter σ^2/K, and ν degrees of freedom; see Geweke (1993).

The parameterization of the equation variances in (2.1) is unusual, and is motivated by our objective of enabling the imposition of an approximate equality of these variances. There are other reasons for choosing this parameterization. Firstly, since the common variance factor σ^2 is state-invariant, an improper prior such as $p(\sigma^2) \propto \sigma^{-2}$ could be used. Secondly, since the support of $(h_1, \ldots, h_K)$ is bounded by construction, a uniform prior on these parameters (which many authors would use to represent complete ignorance) will be proper, so that the difficulties mentioned in the second paragraph of the Introduction do not arise.

We will now propose suitable priors on the parameters of the preceding model, which consist of:

(1) the time-specific latent variables, forming the vectors:

$$S = \begin{pmatrix} S_1 \\ \vdots \\ S_T \end{pmatrix} \text{ and } w = \begin{pmatrix} w_1 \\ \vdots \\ w_T \end{pmatrix};$$

(2) the regression coefficients β_i^j, forming the $K \times p$ matrix:

$$B = \begin{pmatrix} \beta_1^1 & \ldots & \beta_1^p \\ \vdots & \vdots & \vdots \\ \beta_K^1 & \ldots & \beta_K^p \end{pmatrix}; \tag{2.2}$$

(3) the state-specific variance factors, forming the vector $h = (h_1, \ldots, h_K)$;

(4) the matrix $\mathbf{P}$ of transition probabilities, with elements $P_{ij} = P[S_{t+1} = j \mid S_t = i]$;

(5) the common variance factor σ^2;

(6) the number ν of degrees of freedom, characterizing the distribution of w_t.

In so doing, we will follow three guiding principles. First, the *conditional* priors should, in a limiting sense, encompass simplifications of (2.1). These simplifications include: (a) a regression model without Markov switching in the state-specific parameters $\beta^j_{S_t}$ and h_{S_t}; (b) a regression model with normal errors. Secondly, the considerations in the fifth paragraph of the Introduction suggest hierarchical (mixture) marginal priors on B and h. Thirdly, for the reasons given by Frühwirth-Schnatter (2001), the prior should be invariant with respect to a relabeling of the states: for instance, $p(h)$ must be the same as $p(h^*)$, where h^* is an arbitrary permutation of h. These objectives are met by the following densities.

Conditionally on the transition probabilities, the prior on S is Markov:

$$P[S_1 = s_1, \ldots, S_T = s_T \mid \mathbf{P}] = \pi_1 \prod_{j=1}^{K} \prod_{i=1}^{K} P_{ij}^{N_{ij}} \tag{2.3}$$

where N_{ij} is the number of one-step transitions from i to j in $(s_1, \ldots, s_T)$, and π_1 (the probability of the initial state) is an element of the vector of ergodic probabilities, which is the sum of the columns of $(A'A)^{-1}$, where:

$$A = \begin{pmatrix} I_K - \mathbf{P}' \\ \imath'_K \end{pmatrix}$$

$\imath_K$ being the $K \times 1$ vector with all elements equal to unity; see Hamilton (1994a, p. 684).

For the sake of parsimony, we put independent prior distributions on the columns B^j of B in (2.2). Conditionally on the hyperparameters (v_j, r_j, μ_j), the prior distribution of B^j is multinormal with covariance matrix:

$$M_j = v_j \left[(1 - r_j) I_K + r_j \imath_K \imath'_K\right] \tag{2.4}$$

and expectation vector $(\mu_j, \ldots, \mu_j)$, so that:

$$p(B^j \mid v_j, r_j, \mu_j) \propto (\det M_j)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(B^j - \mu_j \imath_K)' M_j^{-1} (B^j - \mu_j \imath_K)\right]. \tag{2.5}$$

An approximate state-invariance of the jth regression coefficient is obtained when r_j tends to one.

The hyperprior on (v_j, r_j, μ_j) is now introduced. The eigenvalues of M_j are easily shown to be $\lambda_j = [1 + (K-1)r_j]v_j$ (with multiplicity 1) and $\psi_j = (1 - r_j)v_j$ (with multiplicity $K - 1$). For M_j to be positive definite, we must then have $v_j > 0$ and $-(K-1)^{-1} < r_j < 1$. We take the hyperprior for r_j to be uniform on this interval. For the conditional prior variances v_j, we take independent inverted gamma hyperpriors with parameters α_j and ϕ_j. Finally, we assume independent central normal hyperpriors on the μ_j with variances θ_j^2. This implies the following joint hyperpriors:

$$p(v_j, r_j, \mu_j) \propto v_j^{-\alpha_j - 1} \exp\left(-\frac{\phi_j}{v_j}\right) \exp\left(-\frac{\mu_j^2}{2\theta_j^2}\right) I(r_j) \tag{2.6}$$

where $I(r_j) = 1$ if $-(K-1)^{-1} < r_j < 1$, $I(r_j) = 0$ otherwise.

The vector h follows a Dirichlet distribution with all parameters equal to γ:

$$p(h_1, h_2, \ldots, h_{K-1} \mid \gamma) = \frac{\Gamma(K\gamma)}{[\Gamma(\gamma)]^K} \left[h_1 h_2 \ldots h_{K-1} \left(1 - \sum_{j=1}^{K-1} h_j \right) \right]^{(\gamma-1)} I_{\mathbb{S}}(h_1, \ldots, h_{K-1}) \tag{2.7}$$

where $\mathbb{S} = \{(h_1, \ldots, h_{K-1}) \mid h_i > 0, \sum_{j=1}^{K-1} h_j < 1\}$ and $I_{\mathbb{S}}(.)$ is an indicator function.

When $\gamma = 1$, this is a uniform prior on the unit simplex. When γ tends to infinity, h converges in probability to the barycentre where $h_1 = \cdots = h_K = 1/K$, and this imposes an approximate equality of the variances across states.

The prior on h is made flexible by specifying a translated exponential hyperprior for γ, which takes the form:

$$p(\gamma) = \xi \exp[-\xi(\gamma - \epsilon)] I_{[\epsilon,\infty)}(\gamma) \tag{2.8}$$

where $\xi > 0$ and $\epsilon \geq 0$. In (2.8), choosing $\epsilon < 1$ would lead to an unbounded and multimodal marginal density $p(h)$, which the author deems inadvisable. A large value of ξ would concentrate the prior mass of γ on a neighborhood of ϵ. In the empirical part of this paper, the author chose $\xi = 0.1$ and $\epsilon = 1$ since this implies confidence intervals consistent with his prior beliefs.

It has been pointed out by two referees that the mixture (2.7)–(2.8) is not the only way in which an approximate equality constraint on the variances could be imposed. Equation (2.1) can also be written as:

$$\ln \left(y_t - \sum_{j=1}^{p} x_{tj} \beta_{S_t}^j \right)^2 - \ln w_t = \ln \sigma_{S_t}^2 + \ln u_t^2$$

with $\sigma_{S_t}^2 = h_{S_t} \sigma^2$. A prior similar to (2.5) could then be used on $(\ln \sigma_1^2, \ldots, \ln \sigma_K^2)$. Unfortunately, the non-normality of $\ln u_t^2$ necessitates the careful choice of a candidate-generating density for implementing a Metropolis-Hastings step; and the rejection probability for this step must be computed from the entire set of observations. Kim et al. (1998) find it necessary to use an auxiliary mixture model in similar circumstances. By contrast, as will be shown in the next section, (2.7)–(2.8) leads to full conditional posteriors that are proportional to densities that can be sampled directly; the factor of proportionality has a particularly simple form. The implementation of Metropolis-Hastings in this case becomes very easy, and does not require the choice of tuning parameters. In fact, Chib and Greenberg (1995, p. 330) seem to recommend this simple implementation of the Metropolis-Hastings sampler when it is available.

Choosing non-uniform conjugate priors on the transition probabilities would raise the question of prior parameter choice, and this would warrant the introduction of a three-level prior hierarchy for S. Since the proposed prior hierarchies for B and h include only two levels, the author sees little point in such an exercise. For this reason, the rows of $\mathbf{P}$ are taken to be uniformly and independently distributed

on the unit simplex:

$$p(P_{i1}, \ldots, P_{i,K-1}) = \Gamma(K) I_{\mathbb{S}}(P_{i1}, \ldots, P_{i,K-1}). \tag{2.9}$$

The prior on the common variance factor is the usual inverted gamma with parameters a and b:

$$p(\sigma^2) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp\left(-\frac{b}{\sigma^2}\right). \tag{2.10}$$

As previously mentioned, a and b can both tend to zero.

Following Geweke (1993), the w_t have independent identical inverted gamma distributions with parameters $a = b = \nu/2$:

$$p(w \mid \nu) = \left(\frac{\nu}{2}\right)^{\frac{T\nu}{2}} \left[\Gamma\left(\frac{\nu}{2}\right)\right]^{-T} \left(\prod_{t=1}^{T} w_t\right)^{-\frac{\nu}{2}-1} \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \frac{\nu}{w_t}\right]. \tag{2.11}$$

Finally, the prior on the degrees of freedom parameter is a translated exponential with parameters $\lambda > 0$ and $\delta \geq 0$:

$$p(\nu) = \lambda \exp\left[-\lambda(\nu - \delta)\right] I_{[\delta,\infty)}(\nu). \tag{2.12}$$

When λ becomes large, the prior mass becomes concentrated in the neighborhood of δ; a prior constraint on the degrees of freedom can be imposed in this fashion. Of course, approximate error normality is obtained for large values of δ. When $\delta = 0$, the prior in (2.12) is the same as the one proposed by Geweke (1993). The present (modest) generalization is useful for two reasons: firstly, it is potentially important, on numerical grounds, to bound ν away from zero; secondly, approximate normality can be imposed while maintaining a reasonably tight prior, and this can improve the convergence properties of the MCMC algorithm. In Sections 6 and 7, the author chose $\delta = 1$ and $\lambda = 0.01$, implying a prior expectation of $E(\nu) = 101$.

To summarize, the joint prior is the product of (2.3), (2.5)–(2.6) for $j = 1, \ldots, p$, and (2.7)–(2.12).

3. Simulating the joint posterior

The algorithm of this section is a special case of the random permutation sampler described in Frühwirth-Schnatter (2001); the special issues that arise in an implementation of a constrained permutation sampler will be discussed in Section 5. An initial draw is made from an arbitrary proper distribution, such as the joint prior described in Section 2. This draw is followed by N passes of the permutation sampler. Let μ, r, and v be vectors with elements μ_j, r_j, and v_j for $j = 1, \ldots, p$. A single pass first draws, in succession and using the most recent conditioning values, from the full conditional posteriors:

$$p(S \mid \mathbf{P}, \sigma^2, h, B, w, \text{data}) \tag{3.1}$$

$$p(\mathbf{P} \mid S) \tag{3.2}$$

$$p(B \mid S, \sigma^2, h, w, \mu, r, v, \text{data}) \tag{3.3}$$

$$p(\mu \mid B, r, v) \tag{3.4}$$

$$p(r \mid B, \mu, v) \tag{3.5}$$

$$p(v \mid B, \mu, r) \tag{3.6}$$

$$p(w \mid S, \nu, \sigma^2, h, B, \text{data}) \tag{3.7}$$

$$p(\nu \mid w) \tag{3.8}$$

$$p(\sigma^2 \mid S, h, B, w, \text{data}) \tag{3.9}$$

$$p(\gamma \mid h) \tag{3.10}$$

$$p(h_i \mid h_{-i}, S, \sigma^2, B, w, \gamma, \text{data}) \quad \text{for } i = 1, \ldots, K-1 \tag{3.11}$$

where h_{-i} is defined as the vector h without elements h_i and h_K. h_K is given by $1 - \sum_{j=1}^{K-1} h_j$.

Then, in order to generate a balanced sample from the posterior, a random permutation $\Pi = (\Pi_1, \ldots, \Pi_K)$ of $(1, 2, \ldots, K)$ is selected with probability $(K!)^{-1}$. For all i and j in $\{1, \ldots, K\}$, element (i, j) of $\mathbf{P}$ is replaced by the element with indices (Π_i, Π_j); the ith row of B is replaced by the row with index Π_i; and the ith element of h is replaced by the element with index Π_i. For $t = 1, \ldots, T$, S_t is replaced by Π_{S_t}.

Sampling S and $\mathbf{P}$. Simulating (3.1) and (3.2) does not involve any novel techniques; the results in Chib (1996) can be used without modifications.

Sampling B. Simulating (3.3) is also straightforward. Upon defining $R_{ti} = 1$ if $S_t = i$, $R_{ti} = 0$ otherwise,

$$X_S = \begin{pmatrix} x_{11}R_{11} & \ldots & x_{11}R_{1K} & \ldots & x_{1p}R_{11} & \ldots & x_{1p}R_{1K} \\ x_{21}R_{21} & \ldots & x_{21}R_{2K} & \ldots & x_{2p}R_{21} & \ldots & x_{2p}R_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{T1}R_{T1} & \ldots & x_{T1}R_{TK} & \ldots & x_{Tp}R_{T1} & \ldots & x_{Tp}R_{TK} \end{pmatrix}, \tag{3.12}$$

V_B as a block-diagonal matrix with jth diagonal block equal to M_j in (2.4), $\mu = (\mu_1 \dots \mu_p)'$, and Ω_S as a diagonal matrix with t-th diagonal element equal to $\sigma^2 w_t h_{S_t}$, the full conditional posterior of $\operatorname{vec} B$ is multinormal with parameters:

$$E(\operatorname{vec} B \mid S, \sigma^2, h, w, \mu, r, v, \text{data}) = \left[V_B^{-1} + (X_S' \Omega_S^{-1} X_S)\right]^{-1} \left[V_B^{-1}(\mu \otimes \imath_K) + X_S' \Omega_S^{-1} y\right] \tag{3.13}$$

$$V(\operatorname{vec} B \mid S, \sigma^2, h, w, \mu, r, v, \text{data}) = \left[V_B^{-1} + (X_S' \Omega_S^{-1} X_S)\right]^{-1}. \tag{3.14}$$

Note that an analytical inverse V_B^{-1} is available, since:

$$M_j^{-1} = \frac{1}{v_j(1-r_j)} \left(I_K - \frac{r_j}{[1+(K-1)r_j]} \imath_K \imath_K' \right).$$

Sampling μ, r, and v. We now discuss the simulation of (3.4) to (3.6). Note that only B provides posterior information on (μ, r, v). The method for simulating (3.4) and (3.5) relies on the observation that assumption (2.5) can be equivalently restated, using the spectral form of M_j, as:

$$B^j = \mu_j \imath_K + \left(\sqrt{\lambda_j} Q_1 + \sqrt{\psi_j} Q_2 \right) \eta \tag{3.15}$$

where $\eta \sim N(0, I_K)$, $Q_1 = K^{-1} \imath_K \imath_K'$, and $Q_2 = I_K - Q_1$. Note that $\imath_K' Q_1 = \imath_K'$ and that $\imath_K' Q_2$ is a null vector. Multiplying (3.15) by $\imath_K'/K$ yields:

$$\frac{\imath_K' B^j}{K} = \mu_j + u \sqrt{\frac{\lambda_j}{K}} \tag{3.16}$$

where u is standard normal. The full conditional posteriors $p(\mu_j \mid B^j, r_j, v_j) = p(\mu_j \mid B^j, \lambda_j)$ then follow from (2.6) and from (3.16), as normal distributions with expectations:

$$(\theta_j^{-2} + K\lambda_j^{-1})^{-1} \lambda_j^{-1} (\imath_K' B^j)$$

and variances $(\theta_j^{-2} + K\lambda_j^{-1})^{-1}$.

The method for drawing r relies on the fact that the r_j are conditionally independent, and that the full conditional posterior of r_j is proportional to a density that can be sampled. Indeed, equations (3.15) and (2.6) imply:

$$p(r_j \mid B^j, \mu_j, v_j) \propto f_j(r_j) \left[(1-r_j)^{-\frac{K-1}{2}} \exp\left(-\frac{y_j' Q_2 y_j}{2(1-r_j)} \right) I_{(0, \frac{K}{K-1})}(1-r_j) \right] \tag{3.17}$$

where $y_j = (B^j - \mu_j \imath_K)/\sqrt{v_j}$, and:

$$f_j(r_j) = [1+(K-1)r_j]^{-\frac{1}{2}} \exp\left(-\frac{y_j' Q_1 y_j}{2[1+(K-1)r_j]} \right).$$

The square bracket in (3.17) is a generalized (since $K \leq 3$ is not excluded) truncated inverted gamma kernel on $(1 - r_j)$ with parameters $n = K - 1$, $b = K/(K-1)$, and $s_j = y_j' Q_2 y_j$ (see the Appendix). A candidate $((1 - r_1), \ldots, (1 - r_p))$ is drawn from the product of the generalized truncated inverted gamma densities,[1] and is accepted with probability:

$$\min \left[1, \frac{f_1(r_1) \ldots f_p(r_p)}{f_1(r_1^{\text{old}}) \ldots f_p(r_p^{\text{old}})} \right]$$

where $(r_1^{\text{old}}, \ldots, r_p^{\text{old}}) = r^{\text{old}}$ is the most recently drawn vector. If the candidate is rejected, r^{old} is retained.[2]

Finally, if we define $\Theta_j = (1 - r_j) I_K + r_j \imath_K \imath_K'$, the full conditional posteriors $p(v_j \mid B^j, \mu_j, r_j)$ follow immediately from (2.5) and (2.6), as inverted gamma distributions with parameters $\alpha_j^* = \alpha_j + K/2$, and:

$$\phi_j^* = \phi_j + \frac{1}{2} (B^j - \mu_j \imath_K)' \Theta_j^{-1} (B^j - \mu_j \imath_K).$$

Sampling w. The full conditional posterior of w_t in (3.7) is obtained as:

$$p(w_t \mid S, \nu, \sigma^2, h, B, \text{data}) \propto w_t^{-\frac{\nu+3}{2}} \exp \left[-\frac{b_t}{w_t} \right] \tag{3.18}$$

with:

$$b_t = \frac{1}{2} \left[\frac{(y_t - \sum_{j=1}^p x_{tj} \beta_{S_t}^j)^2}{\sigma^2 h_{S_t}} + \nu \right]$$

which is an inverted gamma with parameters $(\nu + 1)/2$ and b_t.

Sampling ν. Draws from (3.8) are made by optimized rejection sampling from a translated exponential source density. The method is a straightforward modification of Geweke (1996, p. 749). The target density is:

$$p(\nu \mid w) \propto \left(\frac{\nu}{2} \right)^{\frac{T\nu}{2}} \left[\Gamma \left(\frac{\nu}{2} \right) \right]^{-T} \exp \left[-\varphi \nu \right] I_{[\delta, \infty)}(\nu) \tag{3.19}$$

with:

$$\varphi = \frac{1}{2} \sum_{t=1}^{T} [\ln w_t + w_t^{-1}] + \lambda. \tag{3.20}$$

A candidate ν is drawn from a translated exponential source density:

$$g(\nu; \alpha, \delta) = \alpha \exp \left[-\alpha(\nu - \delta) \right] I_{[\delta, \infty)}(\nu) \tag{3.21}$$

where α maximizes the acceptance probability. This choice of α is found by solving:

$$\frac{T}{2} \left[\ln \left(\frac{1 + \alpha\delta}{2\alpha} \right) + 1 - \Psi \left(\frac{1 + \alpha\delta}{2\alpha} \right) \right] + \alpha - \varphi = 0 \tag{3.22}$$

[1] In practice, $1 - r_j$ is restricted to $[\varepsilon^*, K/(K-1) - \varepsilon^*]$, where ε^* is a small positive number such as 10^{-6}.

[2] Another implementation would draw the correlation coefficients r_j one at a time. This would make monitoring the rejection rates more cumbersome, but may be preferable when the number of regressors is large.

where $\Psi(.)$ is the digamma function. The candidate is accepted with probability:

$$p = \frac{k(\nu)}{s(\alpha,\delta)g(\nu;\alpha,\delta)} \tag{3.23}$$

where $k(\nu)$ is the kernel in (3.19), and:

$$\begin{aligned} s(\alpha,\delta) &= \frac{k\left(\frac{1+\alpha\delta}{\alpha}\right)}{g\left(\frac{1+\alpha\delta}{\alpha};\alpha,\delta\right)} \\ &= \left[\frac{1+\alpha\delta}{2\alpha}\right]^{\frac{T(1+\alpha\delta)}{2\alpha}} \left[\Gamma\left(\frac{1+\alpha\delta}{2\alpha}\right)\right]^{-T} \alpha^{-1} \exp\left[1-\frac{\varphi(1+\alpha\delta)}{\alpha}\right]. \end{aligned} \tag{3.24}$$

Substituting for $k(\nu)$, $s(\alpha,\delta)$, and $g(\nu;\alpha,\delta)$ in (3.23) yields:

$$p = \left[\frac{\Gamma\left(\frac{1+\alpha\delta}{2\alpha}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right]^{T} \left[\frac{\nu}{2}\right]^{\frac{T\nu}{2}} \left[\frac{1+\alpha\delta}{2\alpha}\right]^{\frac{-T(1+\alpha\delta)}{2\alpha}} \exp\left[(\nu-\delta)(\alpha-\varphi)+\frac{\varphi}{\alpha}-1\right]. \tag{3.25}$$

Sampling σ^2. The density in (3.9) is an inverted gamma:

$$p(\sigma^2 \mid S,h,B,w,\text{data}) \propto (\sigma^2)^{-a^*-1} \exp\left(-\frac{b^*}{\sigma^2}\right) \tag{3.26}$$

where $a^* = a + \frac{T}{2}$, and:

$$b^* = b + \frac{1}{2}\left[\sum_{t=1}^{T} \frac{(y_t - \sum_{j=1}^{p} x_{tj}\beta^j_{S_t})^2}{w_t h_{S_t}}\right].$$

Sampling γ. According to (2.7) and (2.8), the kernel of the density in (3.10) is:

$$\frac{\Gamma(K\gamma)}{[\Gamma(\gamma)]^K} \exp\left[\left(\sum_{i=1}^{K} \ln h_i - \xi\right)\gamma\right] I_{[\epsilon,\infty)}(\gamma). \tag{3.27}$$

The method is similar to the one used for drawing ν. We choose the source density:

$$g(\gamma;\alpha,\epsilon) = \alpha \exp\left[-\alpha(\gamma-\epsilon)\right] I_{[\epsilon,\infty)}(\gamma) \tag{3.28}$$

where α maximizes the acceptance probability. A solution of the saddlepoint problem in Geweke (1996, p. 749) implies in this case:

$$\Psi\left(K\left[\frac{1+\alpha\epsilon}{\alpha}\right]\right) - \Psi\left(\frac{1+\alpha\epsilon}{\alpha}\right) + \frac{\sum_{i=1}^{K} \ln h_i - \xi + \alpha}{K} = 0. \tag{3.29}$$

The left-hand side of (3.29) is a strictly increasing function of α. Upon using an asymptotic expansion of the digamma function (Abramowitz and Stegun, 1972, Section 6.3.18) and the fact that $\sum_{i=1}^{K} \ln h_i/K \leq -\ln K$, it is easy to show that this left-hand side has a strictly negative limit as $\alpha \to 0$ and is unbounded as $\alpha \to +\infty$. Hence, a unique solution to (3.29) always exists. The conditional acceptance probability can be calculated, along the lines of (3.23)–(3.25), as:

$$p = \frac{\Gamma(K\gamma)}{\Gamma\left(K\left[\frac{1+\alpha\epsilon}{\alpha}\right]\right)} \left[\frac{\Gamma\left(\frac{1+\alpha\epsilon}{\alpha}\right)}{\Gamma(\gamma)}\right]^{K} \exp\left[\left(\sum_{i=1}^{K} \ln h_i - \xi\right)\left(\gamma-\epsilon-\alpha^{-1}\right) + \alpha(\gamma-\epsilon) - 1\right]. \tag{3.30}$$

Sampling h. The full conditional posterior of h_i in (3.11) is, for $i = 1, \ldots, K-1$:

$$p(h_i \mid h_{-i}, S, \sigma^2, B, w, \gamma, \text{data}) \propto \left[h_i \left(1 - \sum_{j=1}^{K-1} h_j \right) \right]^{\gamma - 1} h_i^{-\frac{N_i}{2}} \exp\left[-\frac{1}{2h_i} \sum_{t=1}^{T} \frac{(y_t - \sum_{j=1}^{p} x_{tj}\beta_i^j)^2 \mathcal{J}_t}{\sigma^2 w_t} \right] I_{(0, 1-\sum_{j\neq i}^{K-1} h_j)}(h_i) \quad (3.31)$$

where $\mathcal{J}_t = 1$ if $S_t = i$, $\mathcal{J}_t = 0$ otherwise, and where $N_i = \#\{S_t = i\}$. Note that h_{-i} is defined as the vector h without the elements h_i and h_K; if $K = 2$, the sum appearing in the indicator function is zero.

This is a non-standard density, which can be characterized by a truncated inverted gamma kernel (if $N_i > 2$) a generalized truncated inverted gamma kernel (if $N_i = 1$ or $N_i = 2$), or a uniform kernel (if $N_i = 0$) multiplied by:

$$f_h(h_i, h_{-i}) = \left[h_i \left(1 - \sum_{j=1}^{K-1} h_j \right) \right]^{\gamma - 1}.$$

A candidate is drawn from this uniform, or, using the algorithm in the Appendix, from the density:

$$p(h_i; N_i, b_i, s_i) \propto h_i^{-\frac{N_i}{2}} \exp\left(-\frac{s_i}{2h_i} \right) I_{(0,b_i)}(h_i) \quad (3.32)$$

where b_i and s_i can be inferred from (3.31). This candidate is accepted with probability:

$$\min\left[\frac{f_h(h_i, h_{-i})}{f_h(h_i^{\text{old}}, h_{-i})}, 1 \right] \quad (3.33)$$

where h_i^{old} is the most recently drawn value of h_i. If the candidate is rejected, h_i^{old} is retained.

The last element h_K is drawn as $1 - \sum_{j=1}^{K-1} h_j$. Since, for any $s \in \{1, \ldots, K\}$, we have $h_s = 1 - \sum_{j\neq s}^{K} h_j$, the last element of h may be associated with state K without loss of generality: the degenerate conditional distribution of h will remain the same if state K is redefined as state s, state s as state K, h_s and h_K are interchanged, and rows s and K of B are interchanged.

The validity of the algorithm and the correctness of the computer program were checked by the following method, suggested by Geweke (2004). If, at each pass of the permutation sampler, the dependent variables y_t are simulated from the likelihood, a sample from the joint density $f(y, \theta)$ is obtained. The generated θ values should follow the prior distribution; this can be verified by means of empirical distribution tests. This method is quite powerful, and is applicable to the constrained and unconstrained versions of the permutation sampler.

4. Random permutation with simulated data

In this section, we will illustrate how the unconstrained (or random) permutation sampler, used in conjunction with the mixture prior of Section 2, can serve to elicit an appropriate identification constraint. A comparison with results obtained with independent priors on the regression parameters will also be made. Two simulated samples of $T = 200$ observations were generated from $y_t = \beta^1_{S_t} + \sigma_{S_t} u_t$ with the following data generating processes:

DGP1 (switching intercept, constant variance).

$$\begin{aligned} y_t &= 0.25 + \sqrt{0.5}u_t \quad \text{if } S_t = 1 \\ &= 1 + \sqrt{0.5}u_t \quad \text{if } S_t = 2 \end{aligned}$$

DGP2 (constant intercept, switching variance).

$$\begin{aligned} y_t &= 3 + \sqrt{0.5}u_t \quad \text{if } S_t = 1 \\ &= 3 + \sqrt{0.05}u_t \quad \text{if } S_t = 2 \end{aligned}$$

where S_t follows a first-order Markov process with $P_{11} = P_{22} = 0.9$, and where u_t is standard normal.

For both samples, the parameters of the mixture prior were chosen as $a = b = 1$, $\epsilon = 1$, $\xi = 0.1$, $\delta = 100$, $\lambda = 0.1$, $\theta_1 = 1$, $\alpha_1 = 11$, and $\phi_1 = 5$. So, the prior on σ^2 is proper, but has no finite moments; approximate error normality is imposed by constraining ν to be greater than 100; and the hyperprior for v_1 has an expectation of $\phi_1/(\alpha_1 - 1) = 0.5$ and a variance of $\phi_1^2/[(\alpha_1 - 1)^2(\alpha_1 - 2)] = 25/900$.

In Figure 1, histograms of the simulated regression parameters β^1_1, $h_1 = \sigma^2_1/(\sigma^2_1+\sigma^2_2)$ and $\sigma^2 = \sigma^2_1+\sigma^2_2$ (based on 10000 replications) are presented. Note that, since the random permutation sampler is used, the joint posterior is multimodal by construction, and that a histogram of β^1_2 would contain exactly the same information as the histogram of β^1_1. The modes of the six histograms in Figure 1 correspond quite closely to the DGP values, and the proper identification constraints are clearly suggested by the top left panel and by the middle right panel.

In Figure 2, the histograms of the hyperparameters μ_1, v_1, r_1, and γ illustrate a point suggested in the Introduction: for DGP1, the Dirichlet parameter γ can take fairly large values, thus constraining the error variances to be approximately equal across states; this may help to achieve the bimodality that was apparent in the top left panel of Figure 1. On the other hand, the prior expectation of the intercept is close to the average of the DGP values, and this may also help in achieving regime separation. For DGP2, the prior correlation coefficient r_1 has a clear mode of unity; this constrains the intercept to be approximately equal in both states, and may help to achieve the bimodality that was apparent in the middle right panel of Figure 1. Again for DGP2, most of the mass of the Dirichlet parameter γ is now concentrated near a value of one, which would correspond to a uniform prior on h_1.

It is interesting to compare the results in Figure 1 with those obtained with independent priors. For this purpose, the random permutation sampler was run on the previous series, but this time with independent $N(0, 100)$ prior distributions on β_1^1 and β_2^1, and independent inverted gamma priors on σ_1^2 and σ_2^2 with $a = b = 1$. In this case, exact error normality was assumed. The results are presented in Figure 3, where the extreme outliers in the left-hand panels illustrate the dangers involved in using relatively non-informative priors on state-specific parameters.

In the right-hand panels of Figure 3, two things are apparent. Firstly, regime separation in the variances is much sharper than in Figure 1. This is the cost of flexibility: by construction, the mixture prior used in Figure 1 puts substantial mass near the subspace where the variances are equal. If the parameter ξ were increased from 0.1 to 100, for instance, this phenomenon would become much less noticeable. Secondly, the histogram of σ^2 shows some upward bias (compare with the bottom right panel of Figure 1; the DGP value is 0.55). A possible explanation is that an invariant prior has been used, so that the priors on σ_1^2 and σ_2^2 must be the same; their common mode is equal to the DGP value of σ_1^2; but σ_2^2 is much lower than σ_1^2. This may cause an upward bias in their sum. A nearly improper prior on the variances would not work: when a and b are set equal to 10^{-6}, the algorithm fails for numerical reasons. This does not happen when the parameterization of equation (2.1) is used.

We conclude this section by providing some details on the implementation and on the performance of the random permutation sampler with the mixture prior. Convergence checking was done by means of heteroscedasticity and autocorrelation consistent Wald tests of the equality of the expectation vectors from two independent Markov chains, differing by their numbers of burn-in passes. The covariance matrices of the vectors of means were estimated by the method of Andrews (1991), using a Parzen kernel and VAR(1) prewhitening (Andrews and Monahan, 1992). This ensured easy, optimal, and automatic bandwidth selection. The autocorrelation properties of the chains are quite acceptable. Table 1 presents the relative numerical efficiencies (RNE), defined as the ratio between the sample variance and 2π times the estimated spectral density at zero. A relative numerical efficiency close to one indicates negligible autocorrelation. The RNE are always larger than 0.19; the value of 0.19 occurs for σ^2 in DGP1, the first two partial autocorrelation coefficients being 0.679 and 0.186. By contrast, the corresponding coefficients for h_1 in DGP1 were -0.067 and 0.001. The rejection rates in the Metropolis-Hastings step for h_1 in DGP1 and DGP2 were respectively 0.11 and 0.17. The corresponding values for the correlation coefficient r_1 were 0.20 and 0.25. These low values indicate that the candidate densities are close to the exact conditional posteriors. Finally, the optimized rejection methods for drawing ν and γ appear to be quite effective. In DGP1, the acceptance rate for ν was never less than 0.03, with an average of 0.86; the corresponding figures for DGP2 are 0.06 and 0.86. The performance of the method for drawing γ is even more favorable. The acceptance rate was never lower than 0.16, with an average of 0.91 for DGP1 and 0.95 for DGP2.

5. Implementing the constrained permutation sampler

There would, in principle, be no difficulties in imposing identification restrictions with the prior of Section 2. However, it should be noted that if the identifying restriction is an inequality constraint on the error variances, this constraint should be imposed by changing the support of the full conditionals in (3.31): for instance, when $K = 2$, $h_1 < h_2$ is imposed by restricting h_1 to the interval $(0, 0.5)$.

Nevertheless, as is clear from Section 4, the prior in (2.5)–(2.8) will often put substantial mass on regions where the parameters are positively correlated across states. When a parameter is affected by an identification constraint, this may interfere with the permutation sampler and cause spurious label switching. In such a case, independent priors on those parameters that are used for model identification might be preferable.

Other arguments also plead in favor of using fixed prior parameters with the constrained permutation sampler. When identification is imposed, misspecification diagnostics that are not available at the random permutation stage can be used. One such method is based on an analysis of predictive probabilities, called "p-scores" by Kaufmann and Frühwirth-Schnatter (2002). A simple variant of this method conditions on point estimates (posterior averages) $\hat{\theta}_i$ of the regression parameters and transition probabilities. Upon defining y^{t-1} as the vector of observations on y_s up to $t-1$, the conditional p-scores are defined as:

$$p_t = \sum_{j=1}^{K} P[Y_t \leq y_t \mid y^{t-1}, \hat{\theta}, S_t = j] P[S_t = j \mid y^{t-1}, \hat{\theta}]. \tag{5.1}$$

The conditional probability that $Y_t \leq y_t$ is easily computed from the Student-t integral, and the probabilities of the regimes can be computed by one pass of the discrete filter described in the Appendix of Kaufmann and Frühwirth-Schnatter (2002). If the model is correct, the p-scores should have independent uniform distributions asymptotically. The transformed p-scores $u_t = \Phi^{-1}(p_t)$, where $\Phi(.)$ is the normal integral, can also be used and should be independent standard normal. Also, sensitivity analysis can easily be done by estimating Bayes factors in favor of more diffuse priors, using the approach in Geweke (1998).

The previous paragraphs suggest an empirical Bayes approach, where the random permutation sampler (with the mixture prior of Section 2) is run for the elicitation of fixed prior parameters, which are used *at the constrained permutation stage*. The prior parameters μ_j, v_j, r_j, and γ can be set equal to point estimates (e.g. posterior medians) obtained from the hyperprior estimated by random permutation. If random permutation suggests identification from the variances, the author recommends that (2.7) and (2.10) be replaced by independent identical inverted gamma priors on $\sigma_1^2, \ldots, \sigma_K^2$, with parameters that are suggested by the random permutation results. For instance, choosing a first parameter of $a = 1$ will guarantee stochastically bounded variances; since the prior mode is $b/(a+1)$, the

second parameter b can then be chosen as $(a+1)\hat{\sigma}^2/K$, where $\hat{\sigma}^2$ is the point estimate of σ^2 obtained by random permutation.

Using this empirical Bayes prior on the two samples of Section 4 turns out to be very satisfactory. Figure 4 presents the results of an application of the constrained permutation sampler to these series, using priors with parameter values given by $\lambda = 0.1$, $\delta = 100$, and by the medians of the estimated distributions in Figure 2 ($\mu_1 = 0.52$, $v_1 = 0.46$, $r_1 = -0.06$, $\gamma = 9.4$, and (2.10) with $a = b = 1$ for DGP1; $\mu_1 = 2.25$, $v_1 = 0.482$, $r_1 = 0.606$, and independent inverted gamma priors on σ_1^2 and σ_2^2 with $a = 1$ and $b = 0.552$ for DGP2). No significant bias is present, in spite of some label switching for DGP1: after the initial burn-in period, 206 changes in the chosen permutation occurred. No such label switching occurred for DGP2.

6. An application to quarterly real interest rate data

Garcia and Perron (1996) used the following model to describe the ex post real interest rate:

$$F(L)(y_t - \mu_{S_t}) = \sigma_{S_t} u_t \tag{6.1}$$

where $F(L)$ is a polynomial in the lag operator L, S_t is a discrete latent variable following a Markov process, u_t is $N(0,1)$, and y_t is the difference between the nominal interest rate and the inflation rate. The authors estimated two- and three-state models with quarterly and monthly data spanning the years 1961 to 1986. Hamilton (1994b, p. 3071) estimated a simplified version of this model with $F(L) = 1$ and $S_t \in \{1, 2, 3\}$, using an extended quarterly data set spanning the years 1960 to 1992. In both contributions, the estimates were maximum likelihood.

In this section, we will show:

(1) that a Bayesian estimation of the model in this paper, which has simpler dynamics than (6.1) when $F(L) \neq 1$, can provide an adequate description of quarterly real interest rate data spanning the period 1953:1 to 2002:3. The estimates are very close to the ones obtained by Hamilton for the smaller sample. However, a simple switching AR(0) model appears to be slightly misspecified;

(2) that the new prior proposed in this paper, and in particular equation (2.7), turns out to be important for a clear separation of the regimes.

The dependent variable is $y_t = i_t - \pi_t$, where i_t is the nominal annual interest rate on 3-month US treasury bills for the third month of quarter t (as given by the series FYGM3 in the DRI-WEFA economics database) and π_t is 400 times the quarterly change in the logarithm of the consumer price index (given by the series PZRNEW in the same database).

6.1. Random permutation.

Table 2 presents summary statistics of the state-invariant parameters obtained by running the random permutation sampler on the following autoregression:

$$y_t = \beta^1_{S_t} + \beta^2_{S_t} y_{t-1} + \sum_{j=1}^{3} \beta^{j+2}_{S_t} \Delta y_{t-j} + (h_{S_t} w_t \sigma^2)^{\frac{1}{2}} u_t \tag{6.2}$$

with $S_t \in \{1, 2, 3\}$, and with the following prior parameters:

$$a = b = 10^{-6}, \lambda = 0.01, \delta = 1; \tag{6.3}$$

$$\xi = 0.1, \epsilon = 1, \theta^2_1 = 10, \theta^2_2 = \cdots = \theta^2_5 = 0.01, \alpha_1 = \cdots = \alpha_5 = 11, \phi_1 = 100, \phi_2 = \cdots = \phi_5 = 0.1. \tag{6.4}$$

So, the prior on σ^2 is almost improper; the distribution of the errors is not constrained to be normal; and the hyperprior parameters on the regression coefficients correspond to prior beliefs that the intercept is probably not much larger than 6 in absolute value, and that the DGP is stationary. On the other hand, the relatively high values of the α_j imply fairly informative hyperpriors on the variances v_j. The reason for an informative choice will be discussed shortly.

The rejection rates were somewhat higher than those reported in Section 4 (an average of 0.24 for the h_i; 0.53 for the r_j; and averages of 0.76 for ν and 0.14 for γ). The results are based on 70000 passes of the algorithm in Section 3, of which the first 60000 were discarded. Convergence was tested by the method described in Section 4.[3]

An examination of the medians of the prior correlation coefficients in Table 2 suggests that the *autoregression* coefficients might not exhibit much variation across states. Indeed, the histograms for r_2, r_4, and r_5 have clear modes of unity, whereas r_3 is almost uniform. By contrast, the median of the prior correlation coefficient of the intercept is close to zero, so that a substantial posterior mass is concentrated on negative values. The data do not appear to be very informative on the prior variances v_j, since their posterior moments are quite close to the prior ones (this was also the case in the simulation experiments of Section 4 and in other experiments made by the author, which is one reason why high values of the α_j were chosen). The estimated posterior expectation of γ is close to the prior value of 11; this suggests that the error variances might not be very different across states.

An examination of the histograms of the state-specific parameters confirmed that an identification restriction on the intercept is very probably the most appropriate. In the top panel of Figure 5, the kernel density estimate for β^1_1 shows three clear modes. The corresponding graphs for all the other state-specific parameters (coefficients, variance ratios, and transition probabilities) were clearly unimodal.

In the two bottom panels of Figure 5, we report (for comparison purposes only) the results of the random permutation sampler with two other priors. The first one is the empirical Bayes prior suggested

[3]The number of burn-in passes is larger than necessary; this conservative number was chosen in order to minimize the chances of having to run the convergence test more than once, since the statistic is costly to compute.

in Section 5: the priors (2.5) and (2.7) with parameters $(\mu_j, v_j, r_j, \gamma)$ given by the relevant medians in Table 2 were used. The other parameters were as in (6.3). The corresponding kernel density estimate is given in the middle panel, where state separation is sharper than in the top panel, as can be expected. The bottom panel of Figure 5 shows the result of replacing the value of 9.23 for γ by a value of 1.01 (a uniform prior gave draws of h_i less than 10^{-16}, leading to numerical problems). In this case, the multimodality becomes barely noticeable. This suggests that an approximate equality constraint on the variances may, indeed, help to identify the model.

6.2. Constrained permutation.

We now discuss applications of the constrained sampler to the same data. The left-hand panels of Figure 6 present kernel posterior density estimates of the intercept in each state, using the identification constraint $\beta_1^1 < \beta_2^1 < \beta_3^1$, and the empirical Bayes prior where $(\mu_j, v_j, r_j, \gamma)$ are equal to the medians in Table 2. The values in (6.3) are used for the other parameters. Label switching did not occur after the initial burn-in passes.[4]

The right-hand panels of Figure 6 are obtained with the identification constraint $\sigma_1^2 < \sigma_2^2 < \sigma_3^2$, identical independent inverted gamma priors on the variances with $a = 1$ and $b = 7.365$, and the previous empirical Bayes prior on the regression coefficients. In this case, label switching occurred for 4332 out of the 10000 postburn-in passes. A very clear cross-contamination of the states occurs in the right-hand panels, further confirming that identification from the variances is not appropriate here.

Table 3 presents the summary statistics obtained with the empirical Bayes prior, when the model is identified by constraining the intercept. An examination of the estimates in Table 3 reveals that the middle regime ($S_t = 2$) has a lower innovation variance and a higher persistence than the other two. The disturbances are close to normality, with a point estimate ($\bar{\theta}$) of the degrees of freedom parameter equal to 118.8. The relative numerical efficiencies are quite high, except for the degrees of freedom. This may be due to the relatively low value of $\lambda = 0.01$ (compare with the corresponding values in Table 1, where $\lambda = 0.1$). The last rows of Table 3 correspond to the equilibrium expectations $E[y_t \mid S_t]$, given by $e_{S_t} = \beta_{S_t}^1/(1 - \beta_{S_t}^2)$. The results are quite close to the ones reported by Hamilton (1994b), in spite of the longer sample period used in this paper and of the differences in the estimation methods. Indeed, Hamilton reports estimates of $\hat{e}_1 = -1.58$, $\hat{e}_2 = 1.58$, and $\hat{e}_3 = 5.69$.

Figure 7 presents the smoothed probabilities $P[S_t = i \mid \text{data}]$, estimated as the percentage of replications of S_t corresponding to regime i. Since 95% confidence bands are practically indistinguishable from the point estimates, they are not reported in the figure. The years 1953 to 1972 can be clearly associated with the intermediate regime, with the exception of a short spell (1956–1958) during which the data are less informative. The years 1973 to 1980 are clearly associated with the low-mean state. Thereafter,

[4]Note that the ordering $\beta_1^1 < \beta_2^1 < \beta_3^1$ only applies to each individual replication, so that the three left-hand panels in Fig. 6 can overlap.

the process switches to the high-mean state until 1987, when the middle regime again prevails. At the end of the sample, a transition to the low-mean state is suggested.

We will now attempt to justify the estimates in Table 3 by reporting the results of misspecification tests. Table 4 presents the results of autoregressions on the transformed p-scores u_t, defined in Section 5, and on their squares u_t^2. Kolmogorov-Smirnov p-values for testing the hypothesis $u_t \sim N(0,1)$ are also presented. No evidence of misspecification is present when identification is done by constraining the intercept: the F-statistic for testing the joint nullity of all autoregression coefficients for u_t has a p-value of 0.0394, none of the individual coefficients being significant. For u_t^2, the p-value of the F-statistic is 0.4435. When identification is done by constraining the variances, the p-value of the F-statistic for u_t falls to 0.0000196.

It is interesting to note that this method also provides evidence of misspecification when all the lagged endogenous variables are excluded from (6.2): in this case, the fourth partial autocorrelation coefficient of the transformed p-scores becomes strongly significant. The significance of the lag polynomial appears to be mainly due to seasonal effects. Indeed, when (6.2) is replaced by an AR(0) regression equation with seasonal dummies, results very similar to those reported in Tables 3 and 4 and in Figure 7 are obtained.

7. An application to monthly real interest rate data

We will now discuss an application of the Markov switching regression model to the same data as in Section 6, but sampled at a monthly rather than quarterly frequency.

The monthly data are much noisier than the quarterly ones. A simple switching AR(12) model appeared unable to fully capture seasonality. Significant residual ARCH effects were also present in such a formulation. We will limit our presentation to a single reasonably parsimonious specification that appears to be congruent with the data. It is the following:

$$y_t = \beta^1_{S_t} + \beta^2_{S_t} D_{4t} + \beta^3_{S_t} D_{6t} + \beta^4_{S_t} D_{11,t} + \beta^5_{S_t} D_{12,t} + \beta^6_{S_t} y_{t-1} + \beta^7_{S_t} y^2_{t-1} + (h_{S_t} w_t \sigma^2)^{\frac{1}{2}} u_t \qquad (7.1)$$

where $S_t \in \{1,2,3\}$, and D_{4t}, D_{6t}, $D_{11,t}$, $D_{12,t}$ are seasonal dummies for the months of April, June, November, and December, respectively. The dependent variable is $y_t = i_t - \pi_t$, where i_t is the nominal annual interest rate on 3-month US treasury bills and π_t is 1200 times the monthly change in the logarithm of the consumer price index. The sample period is 1952:2 to 2002:9.

The inclusion of the squared lag y^2_{t-1} in (7.1) could be justified by a first-order Taylor approximation of an LSTAR model with transition variable y_{t-1} (see van Dijk et al., 2002, p.11) or by a second-order Taylor expansion of an arbitrary non-linear conditional mean function; it was suggested by significant first-order autocorrelations in the squared p-scores when this variable was omitted. Posterior 95% confidence intervals on the coefficients of other seasonal dummies contained the zero value in all regimes, suggesting their exclusion from (7.1).

Table 5 presents the results of an application of the random permutation sampler to (7.1), with the following prior parameters:

$$a = b = 10^{-6}, \lambda = 0.01, \delta = 1 \tag{7.2}$$

$$\xi = 0.1, \epsilon = 1, \alpha_1 = \cdots = \alpha_7 = 11;$$

$$\theta_1^2 = \cdots = \theta_5^2 = 2, \theta_6^2 = 0.01, \theta_7^2 = 0.0001;$$

$$\phi_1 = \cdots = \phi_5 = 20, \phi_6 = 0.1, \phi_7 = 0.001. \tag{7.3}$$

A choice of larger values for the θ_j^2 and ϕ_j in (7.3) led to poor sampler convergence.

An examination of Table 5 suggests conclusions similar to the previous ones: the prior variances v_j are essentially identified by the hyperprior; on the other hand, the data are reasonably informative on the expectations μ_j and the correlation coefficients r_j. In particular, the histogram of r_1 (corresponding to the intercept) showed a clear mode at the lower bound of -0.5. As before, a kernel density plot of β_1^1 had three clear modes, whereas unimodality was observed for the other state-specific parameters. The rejection rates were similar to those reported in Section 6 (an average of 0.16 for the h_i; 0.60 for the r_j; and averages of 0.83 for ν and 0.15 for γ).

Table 6 reports the results of the *constrained* permutation sampler, with the identification restriction $\beta_1^1 < \beta_2^1 < \beta_3^1$, and with the empirical Bayes prior based on (7.2) and on the medians in Table 5. Label switching occurred in only 4 of the 10000 postburn-in passes. The diagnostics are presented in Table 7, and show no evidence of misspecification.

In Table 6, the middle state is again characterized by a higher persistence and a lower innovation variance. The autoregression coefficient $\beta_{S_t}^6$ is close to zero when $S_t = 2$ and has an approximate value of 0.10 in the other two regimes. The squared lag coefficient $\beta_{S_t}^7$ is approximately the same in all regimes, and is indeed significant. Seasonal effects appear to differ somewhat across regimes, suggesting that a naive deseasonalization procedure would not be appropriate.

A noticeable difference between the monthly and quarterly estimates concerns the degrees of freedom. In Tables 5 and 6 (contrary to Tables 2 and 3) the estimated posterior expectation of ν is considerably lower than the prior value of 101. Indeed, when the Bayes factor for $\lambda = 0.1$ versus $\lambda = 0.01$ is estimated by the method in Geweke (1998), a value of 2.11 is obtained; it is significantly greater than unity, with a numerical standard error of 0.09. The corresponding estimate for the quarterly sample of Section 6 was 0.54, with a numerical standard error of 0.07. This suggests that accounting for leptokurtic errors is important for explaining the behavior of the monthly data.

Finally, the estimated state probabilities are given in Figure 8. They agree quite closely with the corresponding quarterly estimates in Figure 7. Note, however, that the evidence in favor of the low regime for the years 1956 to 1958 is now much clearer. The close agreement between Figure 7 and Figure 8 gives further credence to the claim that acceptable models have been identified. Sampling

the data at a quarterly rather than monthly frequency leads to simple linear conditional models, and appears to mask a non-linear STAR-type effect that occurs in the monthly data.

8. Discussion and conclusions

This paper has formulated a Markov switching regression model which includes, as special limiting cases, models where some or all of the state-dependent parameters are equal across regimes. Using a hyperprior allows the parameters governing state dependence to be estimated from the data. The possibility of heavy-tailed disturbances was taken into account, and this was found to be necessary for an adequate modeling of monthly real interest rate data.

Constraining some state-specific parameters to be approximately equal across regimes turns out to be useful for the identification of the Markov switching nature of the model, and for the selection of an appropriate identification constraint. Indeed, imposing state-invariance amounts to reducing the dimension of the parameter space, and this can lead to a model that is better identified.

The hyperprior on the regression coefficients must be fairly informative; choosing its parameters necessitates prior judgments on the order of magnitude of these coefficients. However, the prior on the state-specific innovation variances can be quite diffuse when the model is parameterized in the variance ratios. The potential importance of this fact was illustrated in a simple example involving simulated data.

This paper has not investigated the problem of selecting the appropriate number K of regimes. It is possible that some of the methods for solving this problem, such as reversible jump MCMC (Green, 1995; Robert et al., 2000), are affected by the prior sensitivity illustrated in this paper. This is an interesting topic for further research.

Appendix. Efficient simulation from a generalized truncated inverted gamma distribution

A.1. Density and distribution. Our purpose is to generate draws from a density with the following kernel:

$$k(x;n,b,s) = x^{-\frac{n}{2}} \exp\left(-\frac{s}{2x}\right) I_{(0,b)}(x)$$

where $n \geq 1$ is an integer, $s > 0$, and $b > 0$. It can be checked that:

$$\begin{aligned}\int_0^b x^{-\frac{n}{2}} \exp\left(-\frac{s}{2x}\right) dx &= \left(\frac{s}{2}\right)^{1-\frac{n}{2}} \Gamma\left(\frac{n-2}{2}, \frac{s}{2b}\right) \quad \text{if } n \geq 2; \\ &= \exp\left(-\frac{s}{2b}\right)\left[2\sqrt{b} - \sqrt{2\pi s}\exp\left(\frac{s}{2b}\right)\operatorname{Erfc}\left(\sqrt{\frac{s}{2b}}\right)\right] \quad \text{if } n = 1;\end{aligned} \tag{A.1}$$

where $\Gamma(a,z) = \int_z^\infty e^{-t}t^{a-1}dt$ is the incomplete gamma integral and where $\operatorname{Erfc}(z) = \frac{2}{\sqrt{\pi}}\int_z^\infty e^{-t^2}dt$ is the complementary error function. Note that $\Gamma(\frac{1}{2}, z^2) = \sqrt{\pi}\operatorname{Erfc}(z)$ (Abramowitz and Stegun, 1972, p. 262), and that $\Gamma(0,z)$ is the exponential integral $E_1(z)$, for which accurate rational approximations are available (ibid., p. 231).

It follows that the normalized density and cumulative distribution function (cdf) are given by:

$$\begin{aligned}f(x;n,b,s) &= \frac{\left(\frac{s}{2}\right)^{\frac{n}{2}-1} x^{-\frac{n}{2}} \exp\left(-\frac{s}{2x}\right) I_{(0,b)}(x)}{\Gamma\left(\frac{n-2}{2}, \frac{s}{2b}\right)} \quad \text{if } n \geq 2; \\ &= \frac{\exp\left(\frac{s}{2b}\right) x^{-\frac{1}{2}} \exp\left(-\frac{s}{2x}\right) I_{(0,b)}(x)}{\left[2\sqrt{b} - \sqrt{2\pi s}\exp\left(\frac{s}{2b}\right)\operatorname{Erfc}\left(\sqrt{\frac{s}{2b}}\right)\right]} \quad \text{if } n = 1;\end{aligned} \tag{A.2}$$

$$\begin{aligned}F(x;n,b,s) &= \frac{\Gamma\left(\frac{n-2}{2}, \frac{s}{2x}\right)}{\Gamma\left(\frac{n-2}{2}, \frac{s}{2b}\right)} \quad \text{if } n \geq 2; \\ &= \frac{\left[2\sqrt{x}\exp\left(-\frac{s}{2x}\right) - \sqrt{2\pi s}\operatorname{Erfc}\left(\sqrt{\frac{s}{2x}}\right)\right]}{\left[2\sqrt{b}\exp\left(-\frac{s}{2b}\right) - \sqrt{2\pi s}\operatorname{Erfc}\left(\sqrt{\frac{s}{2b}}\right)\right]} \quad \text{if } n = 1.\end{aligned} \tag{A.3}$$

When $n > 2$, x follows an inverted gamma distribution with parameters $(n-2)/2$ and $s/2$, truncated on the interval $(0,b)$. When $n \leq 2$, the density is non-standard. It is readily checked that $f(x;n,b,s)$ attains a single mode at the point $w = \min(b, s/n)$. The next sections will discuss four complementary methods for simulating $f(x;n,b,s)$.

A.2. Gamma source density. When $n > 2$, $f(x; n, b, s)$ can be simulated by the obvious method:

Step a. Draw $x \sim \text{Ga}\left(\frac{n-2}{2}, \frac{s}{2}\right)$;

Step b. Return x^{-1} if $x^{-1} < b$, otherwise go to Step a;

provided that the acceptance probability remains high. This probability is given by:

$$p_{ga} = \frac{\Gamma\left(\frac{n-2}{2}, \frac{s}{2b}\right)}{\Gamma\left(\frac{n-2}{2}\right)} \tag{A.4}$$

which is the probability that a chi-square variate with $n-2$ degrees of freedom be greater than s/b. Clearly this will become unacceptably low for high values of s/b.

A.3. CDF inversion. Another straightforward method is:

Step a. Draw $u \sim U(0, 1)$;

Step b. Return a numerical approximation to the solution of $u = F(x; n, b, s)$, where $F(x; n, b, s)$ is given by (A.3).

This method will only be reliable if (A.3) can be evaluated accurately. Unfortunately, when s/b is very large, the numerical approximation error in the numerator of (A.3) is divided by an arbitrarily small number, and meaningless values can be generated.

A.4. Uniform source density. This method is based on the source density:

$$p_1(x; b) = \frac{1}{b} I_{(0,b)}(x).$$

A candidate z is drawn from p_1, and is accepted with conditional probability:

$$\frac{f(z; n, b, s)}{a p_1(z; b)}, \quad \text{where} \quad a = \sup_{0<x<b} \frac{f(x; n, b, s)}{p_1(x; b)}.$$

By iterated expectations, the unconditional acceptance probability is then $p_{un} = a^{-1}$.

This probability is easy to evaluate. Since the maximum of $f(x; n, b, s)$ is attained when $x = w = \min(b, s/n)$, we have $a = b f(w; n, b, s)$, or, using equation (A.2):

$$\begin{aligned} p_{un} &= \left(\frac{s}{2}\right)^{1-\frac{n}{2}} \Gamma\left(\frac{n-2}{2}, \frac{s}{2b}\right) b^{-1} w^{\frac{n}{2}} \exp\left(\frac{s}{2w}\right) \quad \text{if } n \geq 2 \\ &= \left[2\sqrt{b} - \sqrt{2\pi s} \exp\left(\frac{s}{2b}\right) \text{Erfc}\left(\sqrt{\frac{s}{2b}}\right)\right] w^{\frac{1}{2}} b^{-1} \exp\left(\frac{s}{2w} - \frac{s}{2b}\right) \quad \text{if } n = 1. \end{aligned} \tag{A.5}$$

Suppose that $b < s/n$. We then have $w = b$, and, upon letting $a = (n-2)/2$ and $z = s/(2b)$:

$$\begin{aligned} p_{un} &= \Gamma(a, z) z^{-a} \exp(z) \quad \text{if } n \geq 2; \\ &= 2\left[1 - \sqrt{\pi}\sqrt{z} \exp(z) \,\text{Erfc}(\sqrt{z})\right] \quad \text{if } n = 1. \end{aligned} \tag{A.5a}$$

As z tends to infinity, both these expressions tend to zero (Abramowitz and Stegun, 1972, pages 231, 263, and 298). We conclude that for given n, the uniform source density will be impractical when s/b is very large. However, for given $n > 2$, the convergence to zero of p_{un} is much slower than that of p_{ga}, due to the presence of the exponential term in p_{un}.

Suppose now that $b \geq s/n$, so that $w = s/n$. Equation (A.5) becomes:

$$\begin{aligned} p_{un} &= \left(\frac{s}{2b}\right) \Gamma\left(\frac{n-2}{2}, \frac{s}{2b}\right) \left(\frac{n}{2}\right)^{-\frac{n}{2}} \exp\left(\frac{n}{2}\right) \quad \text{if } n \geq 2; \\ &= 2\left[1 - \sqrt{\pi}\sqrt{\frac{s}{2b}} \exp\left(\frac{s}{2b}\right) \text{Erfc}\left(\sqrt{\frac{s}{2b}}\right)\right] \left(\frac{s}{b}\right)^{\frac{1}{2}} \exp\left(\frac{b-s}{2b}\right) \quad \text{if } n = 1. \end{aligned} \tag{A.5b}$$

We conclude that for given n, the uniform source density is a poor choice when s/b is close to zero. However, in this case, p_{ga} in (A.4) is close to one, so that the gamma candidate can be used provided that $n > 2$. If $n \leq 2$, then $s/(2b) \leq 1$, and the numerical difficulties mentioned at the end of Section A.3 do not arise, so that cdf inversion is reliable.

For the sake of completeness, we state the details of the uniform rejection method.

Step a. Let $w = \min(b, \frac{s}{n})$;

Step b. Draw $u \sim U(0, 1)$;

Step c. Draw $z \sim U(0, b)$;

Step d. If $u > \left(\frac{z}{w}\right)^{-\frac{n}{2}} \exp\left[-\frac{s}{2}\left(z^{-1} - w^{-1}\right)\right]$ then go to Step b;

Step e. Return z.

A.5. Exponential source density. The considerations in Section A.4 suggest that one of the three preceding methods will always work, provided that $b \geq s/n$.

If b is much less than s/n, all three methods will ultimately fail. However, in this case, a good approximation to $f(x; n, b, s)$ can be provided by the density of an exponential candidate that has been truncated, translated, and reflected, that is:

$$z = b - y \quad \text{where} \quad p(y) \propto e^{-\beta y} I_{(0,b)}(y).$$

The density of z is given by:

$$p_2(z; \beta, b) = \frac{\beta \exp(\beta z)}{\exp(\beta b) - 1} I_{(0,b)}(z).$$

We now address the issue of the choice of β. Again, we must first evaluate the supremum of the density ratio, which is the inverse of the acceptance probability. Upon letting:

$$h(z) = \frac{f(z; n, b, s)}{p_2(z; \beta, b)},$$

we see that:

$$\ln h(z) = -\frac{n}{2}\ln z - \frac{s}{2z} - \beta z + k.$$

This function is strictly concave if $z < 2s/n$, which will be the case if $z < b < s/n$. Equating its first derivative to zero yields the admissible root:

$$z_0 = \frac{s}{\frac{n}{2} + \sqrt{\frac{n^2}{4} + 2\beta s}} > 0.$$

For z_0 to be less than b, we must have:

$$\beta > \frac{1}{2b}\left(\frac{s}{b} - n\right) \tag{A.6}$$

and β will indeed be an admissible parameter of the exponential source density if $b < s/n$ (when this is not satisfied, the method of this section is not applicable). The lower bound in (A.6) turns out to be an appropriate choice. In this case, $h(z)$ is maximized for $z = b$, and the acceptance probability of a candidate z having the density p_2 is:

$$\begin{aligned} p^*_{ex} &= \frac{p_2(b;\beta,b)}{f(b;n,b,s)} \\ &= \left(\frac{s}{2b}\right)^{1-\frac{n}{2}} \left(\frac{s}{2b} - \frac{n}{2}\right) \Gamma\left(\frac{n-2}{2}, \frac{s}{2b}\right) \exp\left(\frac{s}{b} - \frac{n}{2}\right) \left[\exp\left(\frac{s}{2b} - \frac{n}{2}\right) - 1\right]^{-1} \quad \text{if } n \geq 2; \\ &= \left(\frac{s}{b} - 1\right) \exp\left(\frac{s}{2b} - \frac{1}{2}\right) \left[1 - \sqrt{\pi}\sqrt{\frac{s}{2b}} \exp\left(\frac{s}{2b}\right) \operatorname{Erfc}\left(\sqrt{\frac{s}{2b}}\right)\right] \left[\exp\left(\frac{s}{2b} - \frac{1}{2}\right) - 1\right]^{-1} \quad \text{if } n = 1. \end{aligned}$$

In order to obtain the unconditional probability that a (non-truncated) exponential candidate y be ultimately accepted, we must multiply p^*_{ex} by:

$$P(y < b) = 1 - e^{-\beta b} = 1 - \exp\left(\frac{n}{2} - \frac{s}{2b}\right).$$

Since $(1 - e^{-x})(e^x - 1)^{-1} = e^{-x}$, this product is:

$$\begin{aligned} p_{ex} &= \left(\frac{s}{2b}\right)^{1-\frac{n}{2}} \left(\frac{s}{2b} - \frac{n}{2}\right) \Gamma\left(\frac{n-2}{2}, \frac{s}{2b}\right) \exp\left(\frac{s}{2b}\right) \quad \text{if } n \geq 2; \\ &= \left(\frac{s}{b} - 1\right) \left[1 - \sqrt{\pi}\sqrt{\frac{s}{2b}} \exp\left(\frac{s}{2b}\right) \operatorname{Erfc}\left(\sqrt{\frac{s}{2b}}\right)\right] \quad \text{if } n = 1. \end{aligned} \tag{A.7}$$

Using well-known asymptotic expansions (see Abramowitz and Stegun, 1972, Sections 5.1.51, 6.5.32, and 7.1.23), we can show that:

$$\lim_{z\to\infty} \Gamma(a,z) e^z z^{1-a} = 1$$

$$\lim_{z\to\infty} (2z)[1 - \sqrt{\pi}\sqrt{z} e^z \operatorname{Erfc}(\sqrt{z})] = 1.$$

Upon letting $z = s/(2b)$ and $a = (n-2)/2$, we then see that $\lim p_{ex} = 1$ as $s/(2b) \to \infty$ for given n. Since s/b is bounded below by n, p_{ex} is bounded away from zero in all cases.

We now give the details of the exponential rejection method.

Step a. Draw $u_1 \sim U(0,1)$;

Step b. Draw $u_2 \sim U(0,1)$;

Step c. Compute $y = \dfrac{2b^2}{nb-s} \ln u_2$;

Step d. If $y \geq b$ go to Step b;

Step e. Compute $z = b - y$;

Step f. If $u_1 > \left(\dfrac{b}{z}\right)^{\frac{n}{2}} \exp\left[-\dfrac{s}{2z} - \dfrac{z}{2b}\left(\dfrac{s}{b} - n\right) + \dfrac{s}{b} - \dfrac{n}{2}\right]$ then go to Step a;

Step g. Return z.

A.6. The full simulation algorithm. A rule for choosing the most suitable method must now be formulated. It is convenient to work with the ratios of the acceptance probabilities, since this obviates the need to compute incomplete gamma integrals and error functions. From (A.4), (A.5), and (A.7), and recalling that $w = \min(b, s/n)$, we see that:

$$\alpha = \ln \frac{p_{un}}{p_{ga}} = \ln \Gamma\left(\frac{n-2}{2}\right) + \ln \frac{s}{2b} - \frac{n}{2} \ln \frac{s}{2w} + \frac{s}{2w} \quad (\text{if } n > 2) \tag{A.8}$$

$$\beta = \ln \frac{p_{ex}}{p_{ga}} = \ln \Gamma\left(\frac{n-2}{2}\right) + \left(1 - \frac{n}{2}\right) \ln \frac{s}{2b} + \ln\left(\frac{s}{2b} - \frac{n}{2}\right) + \frac{s}{2b} \quad (\text{if } n > 2 \text{ and } b < \frac{s}{n}) \tag{A.9}$$

$$\frac{p_{ex}}{p_{un}} = \frac{s}{2b} - \frac{n}{2} \quad (\text{if } b < \frac{s}{n}) \tag{A.10}$$

so that $p_{un} < p_{ex}$ whenever:

$$\frac{s}{b} > n + 2.$$

The full simulation algorithm can now be stated.

a. If $n > 2$:

- If $s/b > n + 2$: Compute β from (A.9). If $\beta > 0$ use the exponential candidate, otherwise use the gamma candidate.
- If $s/b \leq n + 2$: Compute α from (A.8). If $\alpha > 0$ use the uniform candidate, otherwise use the gamma candidate.

b. If $n \leq 2$:

- If $s/b > n + 2$, use the exponential candidate.
- If $s/b \leq n + 2$, compute p_{un} from (A.5). If $p_{un} > 0.001$ (say) use the uniform candidate. Otherwise use the cdf inversion method.

The arguments in Sections A.2 to A.5 show that for given n, the acceptance probability of any candidate will always be bounded away from zero, and that the inversion method will give reliable results under its conditions of use. In order to assess the overall practicality of the method, it remains to be

shown that the maximum of p_{un}, p_{ga}, and p_{ex} cannot become intolerably low. Plots of $\max(p_{un}, p_{ga}, p_{ex})$ against s/b for various values of n reveal that when $3 \leq n \leq 20$, the global minimum of this function is attained for $s/b = n + 2$ (that is, when $p_{un} = p_{ex}$). In this case, it is a decreasing function of n, ranging from 0.346 (when $n = 3$) to 0.238 (when $n = 20$). When $n > 20$, the uniform rejection method becomes dominated by the other two over the entire range of s/b, and the minimax occurs when $p_{ex} = p_{ga}$. In this case, it becomes a slowly increasing function of n, ranging from 0.236 (when $n = 21$) to 0.328 (when $n = 500$).

When $n = 1$, the minimax between p_{un} and p_{ex} is 0.380 under the constraint that $s/(2b) > 4.6 \times 10^{-8}$ (below this threshold, $p_{un} < 0.001$, and cdf inversion is used). When $n = 2$, the corresponding values are 0.361 and $s/(2b) > 0.000038$.

These values show that the method of this Appendix will indeed remain practical in all cases.

References

Abramowitz, M., Stegun, I.A., 1972. Handbook of Mathematical Functions (Dover Publications, New York).

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.

Andrews, D.W.K., Monahan, J.C., 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. Econometrica 60, 953–966.

Bernardo, J.M., Smith, A.F.M., 2000. Bayesian Theory (paperback edition, Wiley, Chichester).

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

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. The American Statistician 49, 327–335.

Engel, C., Hamilton, J.D., 1990. Long swings in the dollar: are they in the data and do markets know it? American Economic Review 80, 689–713.

Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96, 194–209.

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

Geweke, J., 1993. Bayesian treatment of the independent Student-t linear model. Journal of Applied Econometrics 8 (supplement), S19–S40.

Geweke, J., 1996. Monte Carlo simulation and numerical integration. In: Amman, H.M., Kendrick, D.A., Rust, J. (eds.), Handbook of Computational Economics, vol. 1 (Elsevier, Amsterdam), 731–800.

Geweke, J., 1998. Simulation methods for model criticism and robustness analysis. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.), Bayesian Statistics 6 (Oxford University Press, Oxford), 275–299.

Geweke, J., 2004. Getting it right: joint distribution tests of posterior simulators. Journal of the American Statistical Association 99, 799–804.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

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

Hamilton, J.D., 1994a. Time Series Analysis (Princeton University Press, Princeton).

Hamilton, J.D., 1994b. State-space models. In: Engle, R.F., McFadden, D.L. (eds.), Handbook of Econometrics, vol.4 (Elsevier, Amsterdam), 3039–3080.

Kaufmann, S., Frühwirth-Schnatter, S., 2002. Bayesian analysis of switching ARCH models. Journal of Time Series Analysis 23, 425–458.

Kim, S., Shephard, N., Chib, S., 1998. Stochastic volatility: likelihood inference and comparison with ARCH models. Review of Economic Studies 65, 361–393.

Robert, C.P., Rydén, T., Titterington, D.M., 2000. Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. Journal of the Royal Statistical Society (series B) 62, 57–75.

Van Dijk, D., Teräsvirta, T., Franses, P.H., 2002. Smooth transition autoregressive models – a survey of recent developments. Econometric Reviews 21, 1–47.

TABLE 1. *Relative numerical efficiencies (random permutation sampler with mixture prior, simulated data)*

Parameter	DGP1	DGP2
β_1^1	0.78	0.34
β_2^1	0.75	0.40
h_1	1.14	1.01
h_2	1.14	1.01
P_{11}	0.27	0.28
P_{12}	0.27	0.28
P_{21}	0.26	0.29
P_{22}	0.26	0.29
σ^2	0.19	0.27
ν	0.35	0.32
μ_1	0.76	0.60
v_1	0.90	0.75
r_1	0.43	0.40
γ	0.83	0.36

TABLE 2. *State-invariant parameters (random permutation sampler with mixture prior, quarterly real interest data)*

θ	$\theta_{0.025}$	$\theta_{0.5}$	$\theta_{0.975}$	$\bar{\theta}$	min.	max.	NSE $\times 1000$	RNE
μ_1	−2.386	1.360	4.495	1.227	−6.955	9.217	17.911	0.883
v_1	5.547	9.445	18.044	10.061	4.082	37.585	34.622	0.878
r_1	−0.475	0.068	0.788	0.094	−0.499	1.000	7.669	0.226
μ_2	−0.076	0.088	0.287	0.092	−0.224	0.440	1.821	0.248
v_2	0.005	0.009	0.019	0.010	0.003	0.062	0.042	0.716
r_2	−0.444	0.363	0.965	0.332	−0.496	1.000	10.468	0.167
μ_3	−0.210	−0.049	0.102	−0.051	−0.372	0.228	1.408	0.309
v_3	0.005	0.009	0.018	0.010	0.003	0.046	0.037	0.822
r_3	−0.458	0.257	0.957	0.253	−0.499	0.999	10.142	0.178
μ_4	−0.273	−0.128	0.027	−0.127	−0.401	0.196	1.157	0.433
v_4	0.005	0.009	0.018	0.010	0.003	0.038	0.038	0.791
r_4	−0.430	0.474	0.980	0.407	−0.497	1.000	10.157	0.169
μ_5	−0.232	−0.096	0.046	−0.095	−0.368	0.248	1.029	0.483
v_5	0.005	0.009	0.018	0.010	0.003	0.034	0.036	0.826
r_5	−0.421	0.410	0.978	0.366	−0.496	1.000	10.089	0.177
γ	1.426	9.235	39.372	12.033	1.006	98.006	131.988	0.585
σ^2	8.002	11.048	16.554	11.357	5.096	30.916	41.656	0.270
ν	9.995	100.936	385.926	126.698	3.911	1376.088	7646.400	0.019

Based on the prior parameters in (6.3)–(6.4). θ_α: estimated posterior quantile at probability α. $\bar{\theta}$: estimate of posterior expectation. NSE: numerical standard error. RNE: relative numerical efficiency.

TABLE 3. *Posterior simulation summaries (constrained permutation sampler with empirical Bayes prior, quarterly real interest data)*

θ	$\theta_{0.025}$	$\theta_{0.5}$	$\theta_{0.975}$	$\bar{\theta}$	min.	max.	NSE $\times 1000$	RNE
β_1^1	−2.207	−1.239	−0.335	−1.246	−4.496	0.401	7.563	0.389
β_2^1	0.975	1.414	1.868	1.416	0.557	2.403	3.356	0.469
β_3^1	3.363	4.706	5.875	4.687	1.774	7.142	8.755	0.528
β_1^2	−0.039	0.129	0.297	0.129	−0.200	0.493	0.942	0.830
β_2^2	−0.022	0.133	0.287	0.133	−0.205	0.450	0.953	0.666
β_3^2	−0.048	0.110	0.269	0.110	−0.214	0.424	0.886	0.821
β_1^3	−0.189	−0.030	0.129	−0.030	−0.383	0.316	0.893	0.826
β_2^3	−0.247	−0.117	0.008	−0.117	−0.387	0.143	0.735	0.806
β_3^3	−0.206	−0.059	0.091	−0.058	−0.308	0.204	0.778	0.919
β_1^4	−0.343	−0.202	−0.058	−0.201	−0.464	0.072	0.757	0.938
β_2^4	−0.304	−0.185	−0.064	−0.185	−0.426	0.019	0.666	0.838
β_3^4	−0.350	−0.213	−0.077	−0.213	−0.480	0.071	0.725	0.931
β_1^5	−0.314	−0.166	−0.015	−0.166	−0.462	0.129	0.824	0.861
β_2^5	−0.272	−0.156	−0.038	−0.155	−0.416	0.071	0.705	0.722
β_3^5	−0.260	−0.126	0.013	−0.125	−0.384	0.160	0.720	0.949
h_1	0.248	0.371	0.513	0.373	0.153	0.647	0.991	0.464
h_2	0.165	0.248	0.349	0.250	0.099	0.459	1.123	0.176
h_3	0.225	0.375	0.537	0.377	0.110	0.648	1.323	0.373
P_{11}	0.686	0.894	0.974	0.878	0.025	0.996	1.445	0.280
P_{12}	0.002	0.048	0.233	0.065	0.000	0.632	1.328	0.233
P_{13}	0.006	0.047	0.158	0.057	0.000	0.778	0.534	0.631
P_{21}	0.003	0.024	0.079	0.028	0.000	0.201	0.355	0.324
P_{22}	0.898	0.965	0.992	0.960	0.773	0.999	0.453	0.294
P_{23}	0.000	0.008	0.047	0.012	0.000	0.190	0.222	0.334
P_{31}	0.001	0.036	0.173	0.050	0.000	0.367	0.608	0.575
P_{32}	0.004	0.064	0.216	0.076	0.000	0.381	0.765	0.549
P_{33}	0.717	0.885	0.972	0.874	0.491	0.996	0.809	0.691
σ^2	8.380	11.207	15.597	11.397	6.146	21.929	37.344	0.246
ν	11.086	80.864	444.251	118.793	4.557	823.197	7666.532	0.023
e_1	−2.544	−1.417	−0.401	−1.432	−6.756	0.570	8.688	0.384
e_2	1.213	1.632	2.036	1.632	0.651	2.609	3.117	0.454
e_3	4.076	5.283	6.314	5.266	2.395	7.711	7.885	0.508

Based on equation (6.2) with 3 regimes and on the prior parameters in (6.3) for σ^2 and ν; the other prior parameters are equal to the medians in Table 2. Inequality constraint on the intercepts. θ_α: estimated posterior quantile at probability α. $\bar{\theta}$: estimate of posterior expectation. NSE: numerical standard error. RNE: relative numerical efficiency. e_i: equilibrium expectations $E[y_t \mid S_t = i]$.

TABLE 4. *P-values of misspecification diagnostics (constrained permutation sampler with empirical Bayes prior, quarterly real interest data)*

Identification	Dependent variable	F-stat.	KS	BJ	AR(5)	ARCH(4)	White
$\beta_1^1 < \beta_2^1 < \beta_3^1$	u_t	0.0394	0.7257	0.7361	0.8324	0.2393	0.5382
	u_t^2	0.4435	NA	NA	0.4130	0.7934	0.5492
$\sigma_1^2 < \sigma_2^2 < \sigma_3^2$	u_t	0.0000	0.8070	0.4069	0.7960	0.2515	0.5643
	u_t^2	0.1926	NA	NA	0.8126	0.6118	0.4511

Based on 12-lag autoregressions on the transformed p-scores u_t, defined as the inverse normal integrals of p_t in (5.1), and on their squares. F-stat: F-statistic for joint nullity of all autoregression coefficients. KS: Kolmogorov-Smirnov statistic. BJ: Bera-Jarque statistic. AR(5): Breusch-Godfrey statistic. ARCH(4): LM statistic for residual autoregressive conditional heteroscedasticity. White: White's statistic for residual heteroscedasticity (no cross-terms). NA: not applicable.

TABLE 5. *State-invariant parameters (random permutation sampler with mixture prior, monthly real interest data)*

θ	$\theta_{0.025}$	$\theta_{0.5}$	$\theta_{0.975}$	$\bar{\theta}$	min.	max.	NSE $\times 1000$	RNE
μ_1	−0.617	0.953	2.253	0.906	−2.303	3.651	9.011	0.632
v_1	1.250	2.167	4.137	2.312	0.842	8.732	8.464	0.802
r_1	−0.489	−0.163	0.524	−0.119	−0.500	0.962	6.330	0.193
μ_2	−2.423	−0.737	1.095	−0.713	−4.312	3.800	10.339	0.733
v_2	1.052	1.796	3.394	1.909	0.688	8.084	6.508	0.883
r_2	−0.428	0.478	0.979	0.395	−0.498	1.000	10.618	0.166
μ_3	−2.483	−0.841	1.015	−0.805	−4.040	3.276	10.753	0.688
v_3	1.059	1.803	3.433	1.915	0.636	6.551	6.454	0.913
r_3	−0.453	0.413	0.978	0.356	−0.495	0.999	10.655	0.165
μ_4	−0.976	0.966	2.665	0.927	−2.572	4.535	10.962	0.700
v_4	1.069	1.797	3.447	1.915	0.719	6.846	6.589	0.895
r_4	−0.423	0.467	0.979	0.396	−0.498	1.000	10.464	0.163
μ_5	−0.627	1.293	2.937	1.252	−2.637	4.506	11.043	0.676
v_5	1.060	1.810	3.385	1.918	0.706	8.422	6.595	0.867
r_5	−0.450	0.473	0.976	0.395	−0.499	0.999	10.746	0.163
μ_6	−0.072	0.066	0.192	0.064	−0.233	0.315	1.010	0.437
v_6	0.005	0.009	0.018	0.010	0.003	0.036	0.038	0.779
r_6	−0.466	0.174	0.946	0.194	−0.499	1.000	10.830	0.151
μ_7	−0.003	0.012	0.026	0.012	−0.017	0.039	0.116	0.405
v_7	0.0001	0.0001	0.0002	0.0001	0.0000	0.0003	0.0004	0.792
r_7	−0.456	0.367	0.975	0.329	−0.499	1.000	12.154	0.129
γ	1.569	9.560	38.658	12.264	1.001	140.640	124.308	0.642
σ^2	15.624	20.692	26.723	20.804	12.252	32.979	67.358	0.171
ν	6.143	14.234	43.530	17.150	3.663	76.601	524.977	0.037

Based on the prior parameters in (7.2)–(7.3). θ_α: estimated posterior quantile at probability α. $\bar{\theta}$: estimate of posterior expectation. NSE: numerical standard error. RNE: relative numerical efficiency.

TABLE 6. *Posterior simulation summaries (constrained permutation sampler with empirical Bayes prior, monthly real interest data)*

θ	$\theta_{0.025}$	$\theta_{0.5}$	$\theta_{0.975}$	$\bar{\theta}$	min.	max.	NSE $\times 1000$	RNE
β_1^1	−2.147	−1.273	−0.298	−1.253	−3.658	0.299	9.194	0.261
β_2^1	1.040	1.426	1.785	1.423	0.604	2.215	2.526	0.568
β_3^1	2.282	3.567	4.727	3.550	1.338	5.790	10.573	0.342
β_1^2	−2.393	−0.909	0.621	−0.904	−4.068	1.754	8.495	0.813
β_2^2	−2.139	−1.276	−0.362	−1.273	−2.988	1.450	5.360	0.705
β_3^2	−2.763	−1.082	0.600	−1.078	−4.577	1.951	9.370	0.829
β_1^3	−3.095	−1.559	−0.067	−1.561	−4.469	2.239	8.489	0.812
β_2^3	−1.776	−0.906	−0.065	−0.910	−2.728	0.747	5.041	0.750
β_3^3	−2.906	−1.125	0.618	−1.130	−4.435	2.411	10.089	0.774
β_1^4	0.145	1.676	3.259	1.683	−1.216	4.648	8.663	0.825
β_2^4	0.249	1.104	1.934	1.100	−0.596	2.953	5.098	0.723
β_3^4	−0.200	1.505	3.174	1.493	−2.451	4.704	9.543	0.818
β_1^5	0.324	1.914	3.495	1.915	−1.121	4.910	9.329	0.740
β_2^5	1.212	2.102	2.977	2.099	0.176	3.772	5.569	0.642
β_3^5	−0.024	1.673	3.308	1.662	−1.703	4.725	9.114	0.858
β_1^6	−0.040	0.098	0.234	0.097	−0.199	0.418	0.893	0.608
β_2^6	−0.082	0.025	0.129	0.025	−0.170	0.263	0.723	0.545
β_3^6	−0.021	0.133	0.287	0.133	−0.190	0.457	0.870	0.805
β_1^7	0.002	0.017	0.032	0.017	−0.017	0.045	0.085	0.830
β_2^7	0.002	0.017	0.032	0.017	−0.010	0.050	0.085	0.808
β_3^7	0.006	0.020	0.034	0.020	−0.005	0.049	0.074	0.900
h_1	0.300	0.409	0.529	0.410	0.225	0.631	1.032	0.311
h_2	0.163	0.232	0.313	0.233	0.131	0.526	1.180	0.106
h_3	0.227	0.356	0.487	0.357	0.116	0.614	1.284	0.268
P_{11}	0.890	0.958	0.988	0.954	0.734	0.998	0.465	0.305
P_{12}	0.002	0.024	0.088	0.030	0.000	0.213	0.466	0.245
P_{13}	0.001	0.013	0.053	0.017	0.000	0.117	0.203	0.463
P_{21}	0.002	0.012	0.034	0.013	0.000	0.091	0.146	0.319
P_{22}	0.953	0.982	0.995	0.980	0.894	0.999	0.212	0.268
P_{23}	0.000	0.005	0.026	0.007	0.000	0.067	0.151	0.221
P_{31}	0.001	0.015	0.073	0.021	0.000	0.329	0.314	0.404
P_{32}	0.003	0.030	0.094	0.035	0.000	0.276	0.343	0.497
P_{33}	0.872	0.949	0.987	0.944	0.644	0.999	0.426	0.498
σ^2	17.158	22.762	28.853	22.819	13.291	37.123	84.979	0.126
ν	6.929	19.886	124.008	32.282	4.133	222.386	1426.770	0.049

Based on equation (7.1) with 3 regimes and on the prior parameters in (7.2) for σ^2 and ν; the other prior parameters are equal to the medians in Table 5. Inequality constraint on the intercepts. θ_α: estimated posterior quantile at probability α. $\bar{\theta}$: estimate of posterior expectation. NSE: numerical standard error. RNE: relative numerical efficiency.

TABLE 7. *P-values of misspecification diagnostics (constrained permutation sampler with empirical Bayes prior, monthly real interest data)*

Identification	Dependent variable	F-stat.	KS	BJ	AR(7)	ARCH(7)	White
$\beta_1^1 < \beta_2^1 < \beta_3^1$	u_t	0.0578	0.4606	0.0457	0.6702	0.0889	0.5408
	u_t^2	0.3715	NA	NA	0.5325	0.9964	0.9990
$\sigma_1^2 < \sigma_2^2 < \sigma_3^2$	u_t	0.0000	0.4162	0.0058	0.6354	0.1659	0.7602
	u_t^2	0.6237	NA	NA	0.5581	0.9486	0.9981

Based on 36-lag autoregressions on the transformed p-scores u_t, defined as the inverse normal integrals of p_t in (5.1), and on their squares. F-stat: F-statistic for joint nullity of all autoregression coefficients. KS: Kolmogorov-Smirnov statistic. BJ: Bera-Jarque statistic. AR(7): Breusch-Godfrey statistic. ARCH(7): LM statistic for residual autoregressive conditional heteroscedasticity. White: White's statistic for residual heteroscedasticity (no cross-terms). NA: not applicable.

Figure 1. Regression parameters, unconstrained sampler with mixture prior, simulated data

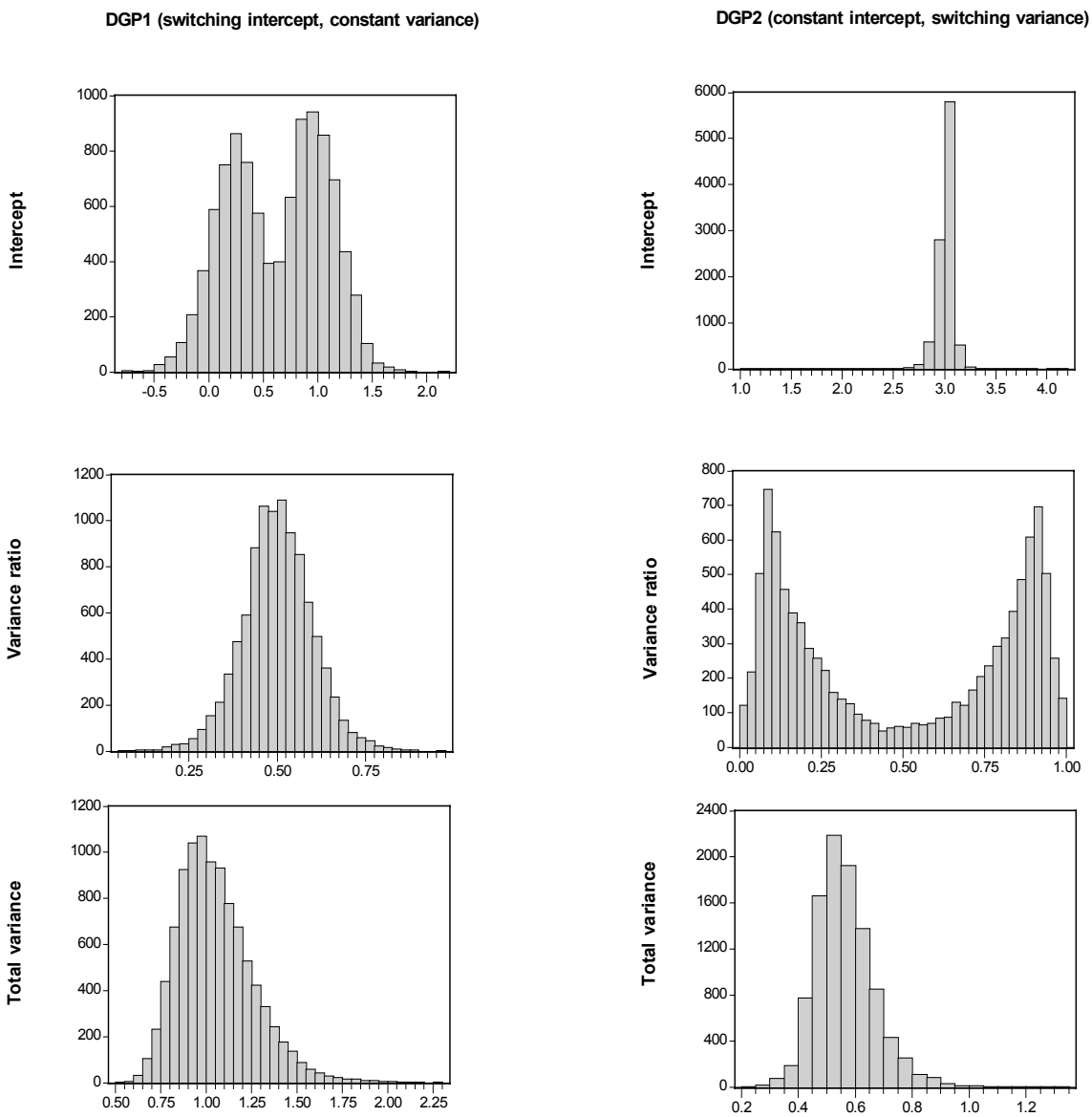


Figure 2. Hyperparameters, unconstrained sampler with mixture prior, simulated data

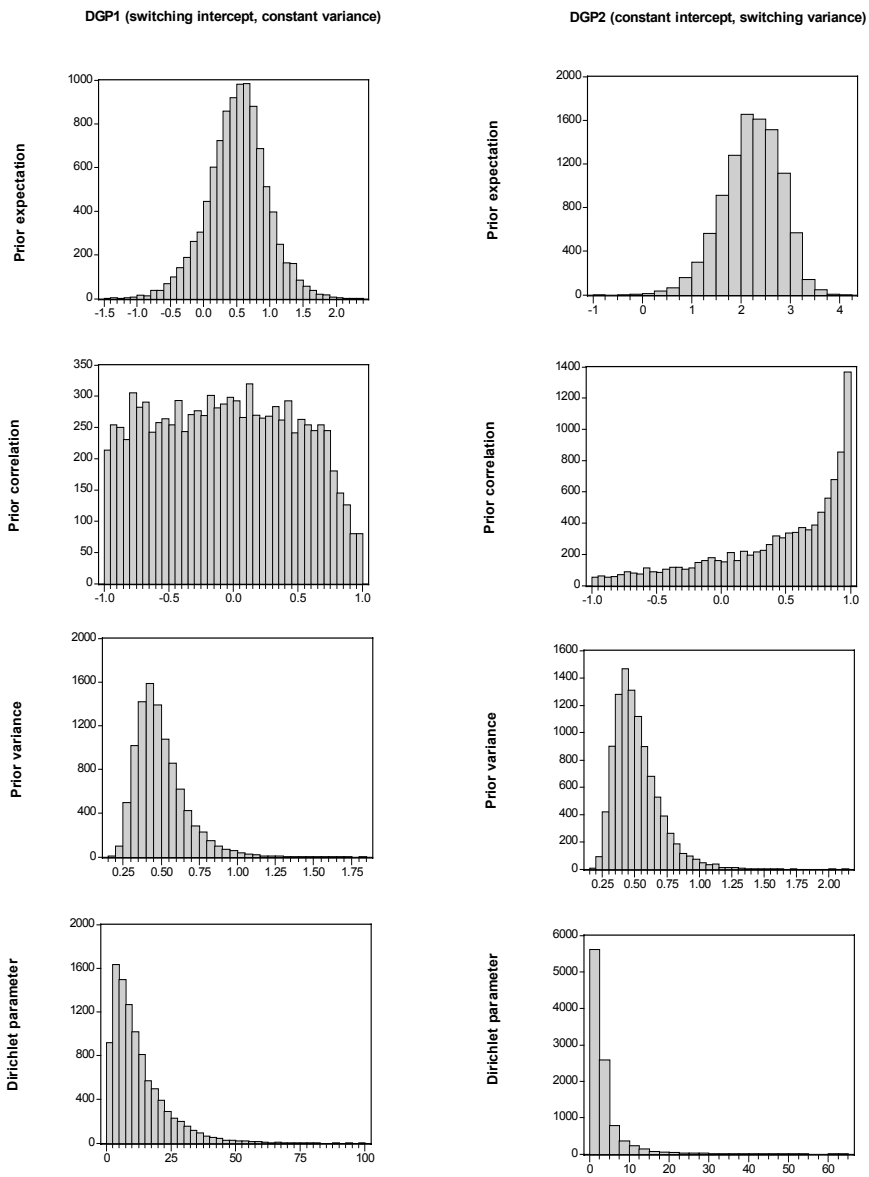


Figure 3. Regression parameters, unconstrained sampler with independent prior, simulated data

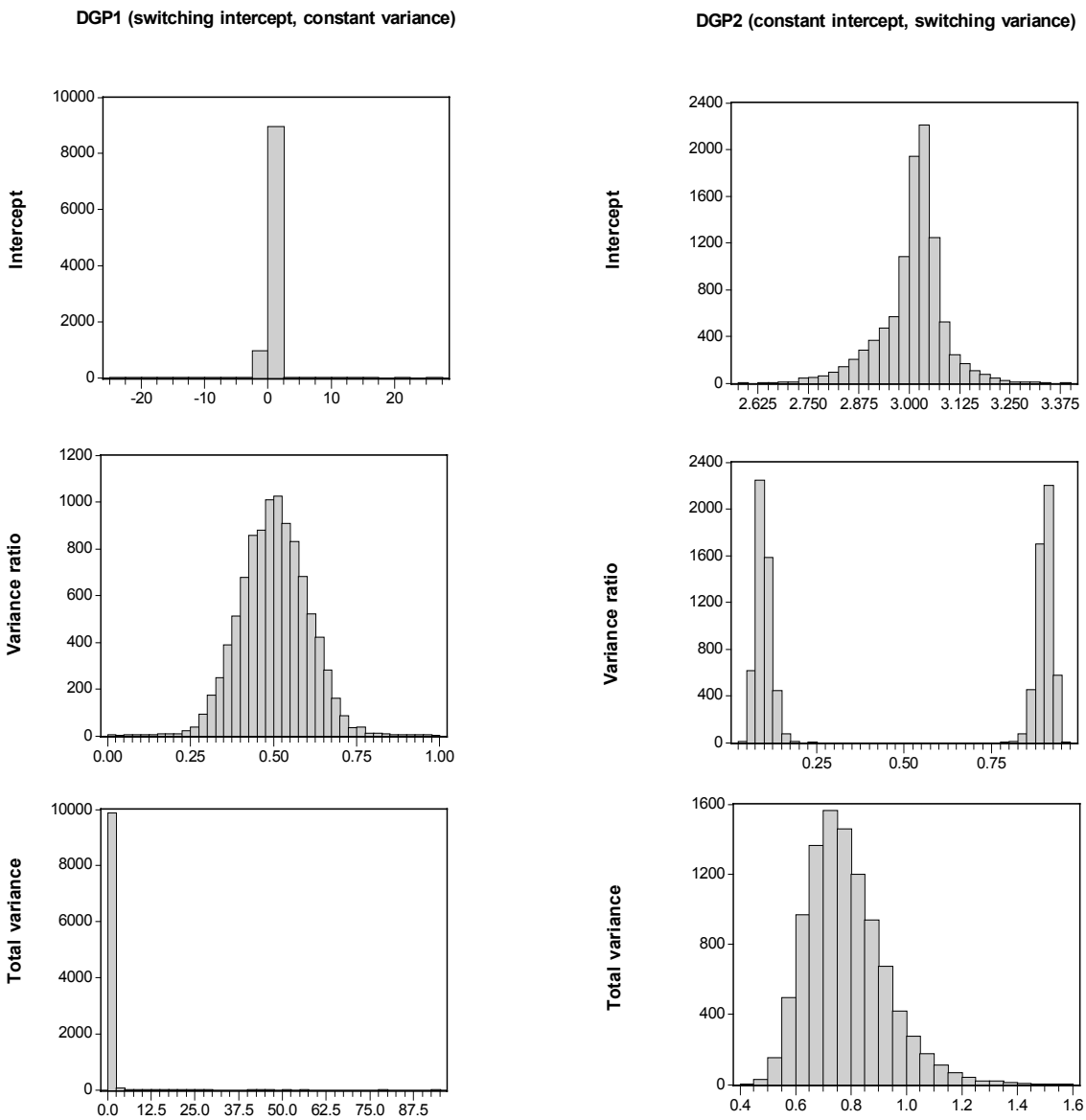


Figure 4. Regression parameters, constrained sampler with empirical Bayes prior, simulated data

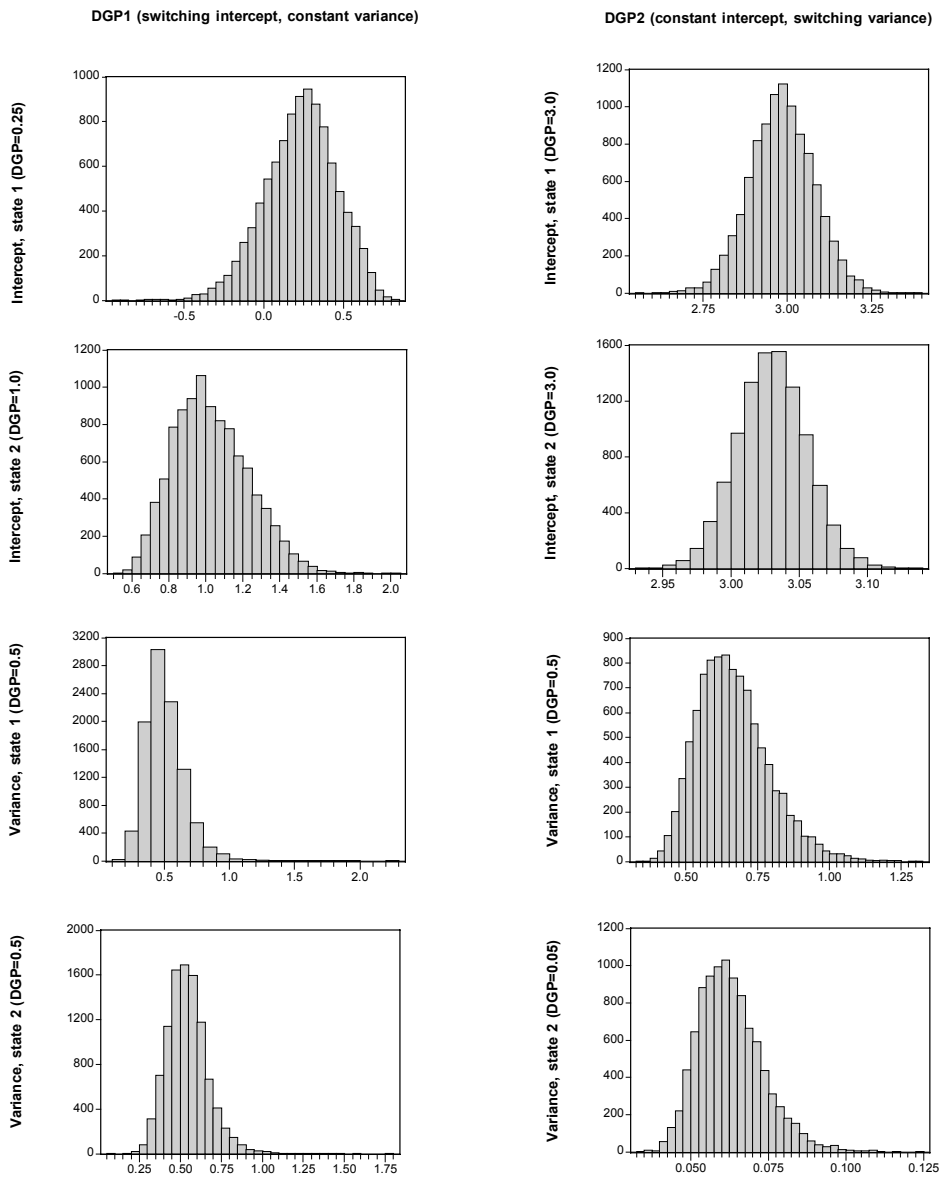
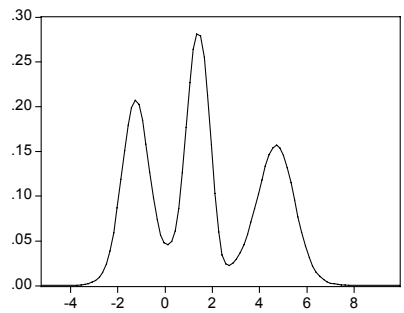
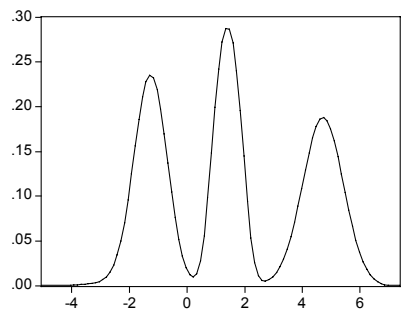


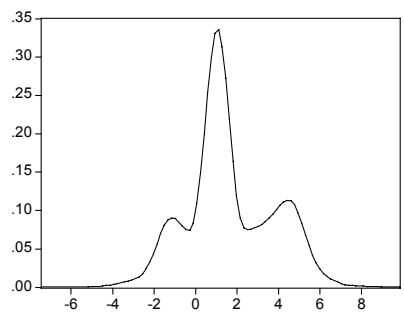
Figure 5. Kernel posterior density estimates of the intercept in state 1 (unconstrained sampler, quarterly interest rate data)



Mixture prior

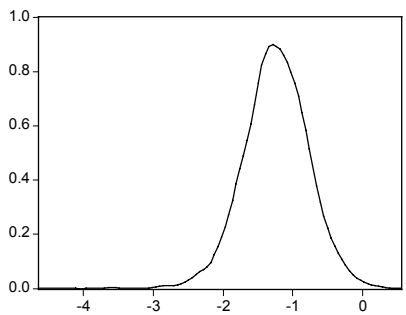


Empirical Bayes prior on coefficients and variance ratios

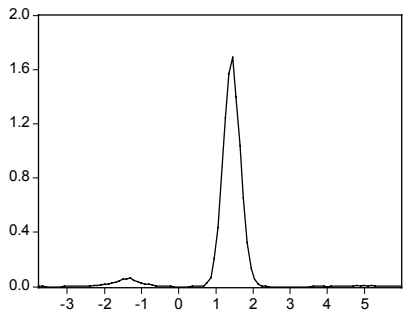


Empirical Bayes prior on coefficients, diffuse on variance ratios

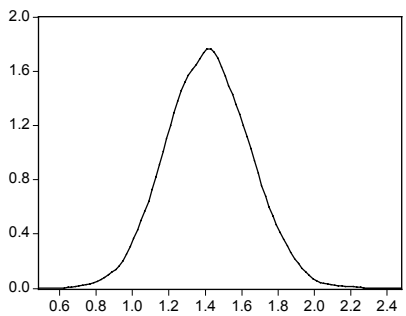
Figure 6. Kernel posterior density estimates of the intercept (constrained sampler, quarterly interest rate data)



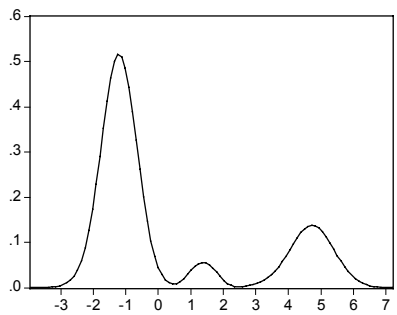
State 1, identification from intercept



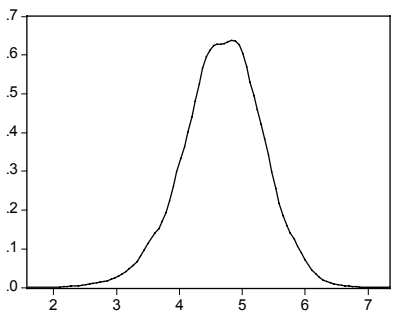
State 1, identification from variances



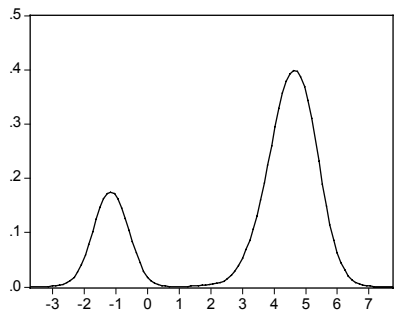
State 2, identification from intercept



State 2, identification from variances



State 3, identification from intercept



State 3, identification from variances

Figure 7. State probabilities (quarterly interest rate data)

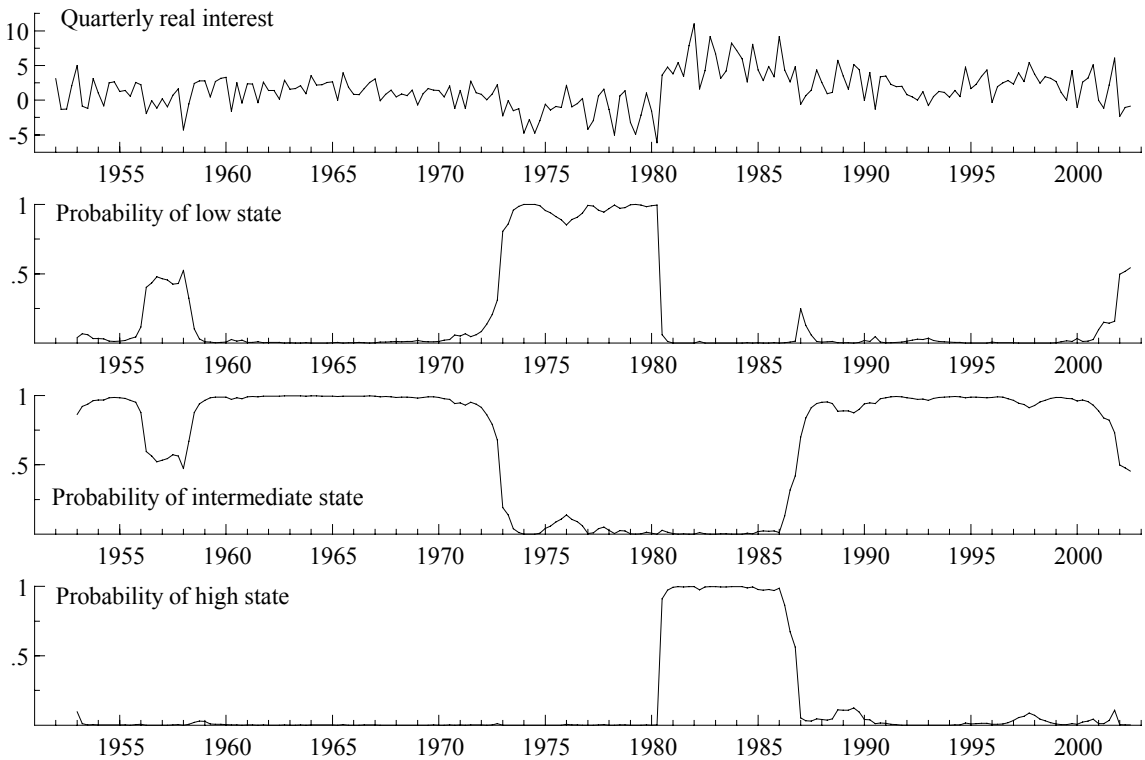


Figure 8. State probabilities (monthly interest rate data)