Journal of Business & Economic Statistics, January 1993, Vol. 11, No. 1

Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts

James H. Albert
Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403

Siddhartha Chib
John M. Olin School of Business, Washington University, St. Louis, MO 63130

We examine autoregressive time series models that are subject to regime switching. These shifts are determined by the outcome of an unobserved two-state indicator variable that follows a Markov process with unknown transition probabilities. A Bayesian framework is developed in which the unobserved states, one for each time point, are treated as missing data and then analyzed via the simulation tool of Gibbs sampling. This method is expedient because the conditional posterior distribution of the parameters, given the states, and the conditional posterior distribution of the states, given the parameters, all have a form amenable to Monte Carlo sampling. The approach is straightforward and generates marginal posterior distributions for all parameters of interest. Posterior distributions of the states, future observations, and the residuals, averaged over the parameter space are also obtained. Several examples with real and artificial data sets and weak prior information illustrate the usefulness of the methodology.

KEY WORDS: Data augmentation; Hidden Markov models; Missing data; Mixture distribution; Monte Carlo simulation; Regime shifts.

An important problem in time series analysis, actually in virtually all of statistics, is the detection and modeling of abrupt changes in the model specified. Typically, interest centers on the behavior of the first few moments of the series—for example, the mean and the variance. The objective is to determine whether those moments are homogenous over time. Lack of homogeneity, if not captured, can severely affect conclusions drawn from the data. Recognition of this fact has led to an interest in such nonlinear models as the bilinear (Tong 1983) and other models that allow for regime shifts or parameter instability (Tsurumi 1988). In this article, we focus on the autoregressive (AR) model with switching introduced by Sclove (1983) and Hamilton (1989). Both the mean and the variance of the real-valued time series are parameterized in terms of an unobserved state variable that follows a two-state Markov process with unknown transition probabilities. In the engineering literature, related models, including the standard state-space model, are called hidden Markov models (Juang and Rabiner 1985) following the early canonical work of Baum and Eagon (1967). The Markov switching AR model has been recently applied with success to several economic and financial data sets. For example, Hamilton (1988, 1989) used the model to date the timing of recessions and booms with gross national product (GNP) data and to model the term structure of interest rates, Pagan and Schwert (1990) used it to model stock returns, and other references may be found in the work of Hamilton (1991).

Our objective here is to address, via Bayesian methods, inference issues that arise in the analysis of such models. The cornerstone of our approach is the idea that the unobserved states, one for each time point, can be treated as missing data and then analyzed, along with the other unknown parameters, via the simulation tool of Gibbs sampling (Gelfand, Hills, Racine-Poon, and Smith 1990; Gelfand and Smith 1990; Geman and Geman 1984; Tanner and Wong 1987; Tierney 1991). Applications of Gibbs sampling to time series include those of Carlin, Polson, and Stoffer (1992), Chib (in press), Chib and Greenberg (1992), and McCulloch and Tsay (1991). There are a number of attractive features of this approach. First, the messy calculations entailed in the direct calculation of the likelihood function are avoided. Second, posterior distributions of all unknown parameters and functions thereof are obtained by simulating standard distributions, such as the multivariate normal and inverted gamma—these posterior distributions convey much more information than the mode and curvature summaries that arise from the maximum likelihood (ML) framework. Third, the approach provides posterior distributions of the states, and of future observations, marginalized over all of the unknown parameters. This improves on "plug-in" approaches in which unknown parameters—for example, those ap-

pearing in the distribution of the states—are replaced by sample estimates. Finally, residual analysis proceeds in a straightforward fashion by using the distribution of generated states to compute the posterior distribution of the model residual.

The plan of the article is as follows. The model, the inference problems of interest, and Hamilton's ML approach are presented in Section 1. The method of Gibbs sampling that is used to implement our approach is outlined in Section 2. The conditional distributions that are the inputs into the Gibbs sampler are derived in Section 3 for standard prior families of distributions. The issues of residual analysis and forecasting future observations are taken up in Section 4, and the methodology is illustrated using diffuse priors on several data sets in Section 5. Concluding remarks and directions for future research are contained in Section 6.

1. MODEL

1.1 An Autoregressive Model With Markov Jumps

Consider the following Gaussian AR model in which the observation at time t, y_t is generated by

$$y_t = x_t'\beta + \gamma s_t + \phi_1(y_{t-1} - x_{t-1}'\beta - \gamma s_{t-1}) + \cdots + \phi_r(y_{t-r} - x_{t-r}'\beta - \gamma s_{t-r}) + v(s_t)^{1/2}u_t, \quad (1.1)$$

where $t = 1, \ldots, n$, $u_t \sim N(0, 1)$, $v(s_t)$ is the variance function defined later, x_t is a $k \times 1$ vector of covariates, β is the corresponding regression parameter, γ is a scalar, and $\{s_t \in \{0, 1\}: t = 1, 2, \ldots\}$ is a hidden stationary Bernoulli random variable following a two-state Markov process with transition probability matrix

	$s_t = 0$	$s_t = 1$
$s_{t-1} = 0$	$(1 - a)$	a
$s_{t-1} = 1$	b	$(1 - b)$,

where the probability parameters a and b are unknown. One can rewrite (1.1) more compactly as

$$\phi(L)(y_t - \mu(s_t, x_t)) = v(s_t)^{1/2}u_t, \quad (1.2)$$

where $\mu(s_t, x_t) = x_t'\beta + \gamma s_t$, $v(s_t) = \sigma^2 + \tau^2 s_t$, and $\phi(L) = (1 - \phi_1 L - \cdots - \phi_r L^r)$ is an rth order polynomial in the lag operator L, where $L^k z_t = z_{t-k}$ for $k \geq 0$. Here $\mu(s_t, x_t)$ is the mean function and $v(s_t)$ is the conditional variance, both of which depend on the outcome of the state s_t. In the sequel we will sometimes parameterize the variance function as $\sigma^2(1 + \omega s_t)$, with ω representing the proportionate increase in variance when $s_t = 1$. Note that Markov switching affects the intercept of the mean function, not the regression parameter vector β. The following assumptions are made:

A1. All of the roots of $\phi(L)$ lie outside the unit circle.
A2. a and b each lie in the open unit interval.
A3. The parameters γ and τ^2 are positive.

Assumption A1 imposes stationarity given the state sequence, whereas A2 ensures that neither state is transient and that the Markov chain converges to a stationary distribution. This assumption is also required for identification because, if $s_t = 0$ for all t or if $s_t = 1$ for all t, then neither γ not τ^2 are identified. The final assumption is necessary to identify state 1 as the high-level, high-variance state.

For future use we now consider (a) the joint density of n observations conditioned on S_n and the parameters and (b) the joint density of the observations and the states, given the parameters. Define $Y_t \equiv (y_1, \ldots, y_t)$ and $S_t \equiv (s_1, \ldots, s_t)$ for $t \geq 1$, and let $\eta = (\beta, \gamma, \phi, \sigma^2, \tau^2)$, where $\phi \equiv (\phi_1, \ldots, \phi_r)$. Then from the law of total probability, conditioned on (η, S_n), the density of the observations can be factored as

$$f(Y_n|S_n, \eta) = f(Y_r|S_r, \eta) \prod_{t=r+1}^{n} f(y_t|Y_{t-1}, S_t, \eta), \quad (1.3)$$

where $f(Y_r|S_r, \eta)$ is the density of the first r observations (and can be obtained by exploiting stationarity) and $f(y_t|Y_{t-1}, S_t, \eta)$ is the one-step-ahead conditional density of y_t that can be deduced from (1.1). Specifically, let $W_r = \text{diag}((1 + \omega s_1)^{1/2}, \ldots, (1 + \omega s_r)^{1/2})$ and let Σ_r satisfy the equation $\Sigma_r = \Phi\Sigma_r\Phi' + e_1e_1'$, where

$$\Phi = \begin{pmatrix} \phi_{-r} & \phi_r \\ I_{r-1} & 0 \end{pmatrix},$$

$\phi_{-r} = (\phi_1, \ldots, \phi_{r-1})'$, and $e_1 = (1, 0, \ldots, 0)'$ is an $r \times 1$ unit vector. Let $\Omega_r = W_r\Sigma_r W_r$. Then

$$f(Y_r|S_r, \eta) \propto \exp(-\sigma^{-2}(Y_r - X_r\beta - S_r\gamma)\Omega_r^{-1}(Y_r - X_r\beta - S_r\gamma)/2), \quad (1.4)$$

which is the density of an $N_r(X_r\beta + S_r\gamma, \sigma^2\Omega_r)$ distribution, where $N_r(.\ ,.)$ denotes the r-variate normal distribution. Now for $t \geq r + 1$, define the function $\hat{y}_{t|t-1} = (1 - \phi(L))y_t + \phi(L)(x_t'\beta + \gamma s_t)$. Then, from (1.1),

$$f(y_t|Y_{t-1}, S_t, \eta) \propto \exp(-v(s_t)^{-1}(y_t - \hat{y}_{t|t-1})^2/2), \quad t \geq r + 1, \quad (1.5)$$

which is the kernel of the normal density with mean $\hat{y}_{t|t-1}$ and variance $v(s_t)$.

Next, the full joint density of the observations *and* the states, now given $\theta = (\eta, a, b)$, is simply the product of (1.3) and the density of the states, namely,

$$f(Y_n, s_1, s_2, \ldots, s_n|\theta) = f(Y_n|S_n, \eta) \prod_{t=2}^{n} \Pr(s_t|s_{t-1})^{s_{t-1}s_t} \Pr(s_1), \quad (1.6)$$

where $\Pr(s_t|s_{t-1})$ is the transition probability and $\Pr(s_1)$ is the initial probability distribution of the chain; their dependence on a and b is suppressed for convenience. It is very important to bear in mind that *neither* (1.3) *nor* (1.6) constitute the likelihood function of θ. The likelihood of θ is only obtained after the states are integrated out from (1.6).

The parameterization of the model in (1.1) and (1.2) is quite rich, and in particular instances we might be content to work with the following special cases: Stationary AR(r) models without covariates or Markov switching, the switching regression models of Goldfeld and Quandt (1973) and Lindgren (1978), and the bivariate normal mixture models discussed by Everitt and Hand (1981) and Titterington, Smith, and Makov (1985).

1.2 Inference Problems

The fundamental inference problem is to use the available data, Y_n, and any nonsample information, to learn about θ and S_n. Specifically, one may be interested in τ^2, the extent of the increase in variance, given the data and everything else treated as nuisance variables. More fundamentally, we will be interested in making marginal inferences about regime shifts, given the data. Related inference problems include analyzing functions of the Markov probabilities—for example, the quantities $\pi_0 = b(a + b)^{-1}$ and a^{-1}, which are, respectively, the limiting probabilities of being in state 0 and the expected duration of being in state 0 given that the current state is 0. Other important concerns involve the issues of residual analysis and forecasting out-of-sample observations. Of course, in the Bayesian approach we adopt, all such inferential questions will be answered from the marginal posteriors of the quantity in question. While deriving the posteriors, we will incorporate, at the very least, the restrictions on the parameter space contained in assumptions A1–A3. In the framework used here, these restrictions are conveniently imposed by taking the usual prior and multiplying them by indicator functions that take the value 1 when the restriction is satisfied and the value 0 otherwise.

1.3 Maximum Likelihood Fitting

Hamilton (1989) developed an innovative procedure to compute the likelihood function because brute force marginalization of (1.6) involves 2^n summations over all possible state sequences of s_t that can comprise S_n. The estimates for (θ, S_n) are obtained via a two-step procedure in which θ is first estimated from the likelihood function, and then inference about S_n is based on the estimated θ. The likelihood function of θ is defined as follows: For a given value of t and knowledge of the conditional probability $\Pr(s_{t-1}, \ldots, s_{t-r}|Y_t, \theta)$, one computes the conditional probability $\Pr(s_t, \ldots, s_{t-r+1}|Y_t, \theta) = \Sigma_{s_{t-r}=0}^{1} \Pr(s_t, \ldots, s_{t-r}|Y_{t-1}, \theta)$, where

$$\Pr(s_t, \ldots, s_{t-r}|Y_t, \theta) \propto \Pr(s_t|s_{t-1})\Pr(s_{t-1}, \ldots, s_{t-r}|Y_{t-1}, \theta) f(y_t|Y_{t-1}, S_t, \theta).$$

The normalizing constant of the latter probability is the conditional likelihood of y,

$$f(y_t|Y_{t-1}, \theta) = \sum_{s_t=0}^{1} \cdots \sum_{s_{t-r}=0}^{1} f(y_t, s_t, \ldots, s_{t-r}|Y_{t-1}, \theta). \quad (1.7)$$

After (1.7) is computed for $t = r + 1, \ldots, n$, the sample conditional log-likelihood is given by

$$\log f(y_{r+1}, \ldots, y_n|Y_r, \theta) = \sum_{t=r+1}^{n} \log f(y_t|Y_{t-1}, \theta). \quad (1.8)$$

The maximum likelihood estimator (MLE) is found iteratively via the Newton–Raphson method as the root of the score function, where the score is computed by numerical differentiation. Once the estimate $\hat{\theta}$ is found, inferences about s_t are based on either $\Pr(s_t|Y_t, \hat{\theta})$, or on the r lag smoother $\Pr(s_{t-r}|Y_t, \hat{\theta})$. Note that the uncertainty of θ is not incorporated in the latter calculations. In addition, this approach does not provide a complete description of the likelihood. As we show in the examples, information about the shape of the likelihood, such as bimodality or nonsymmetry, is obtained through our method.

2. GIBBS SAMPLING

Suppose that the objective is to simulate from the posterior distribution of a parameter vector ψ, partitioned perhaps with vector components as $(\psi_1, \ldots, \psi_k)$. In many cases, the full k-dimensional distribution is complicated and cannot be simulated directly. Traditional methods to deal with this problem have relied on the method of Monte Carlo with importance sampling. This typically requires determining a suitable importance function and also the evaluation of the likelihood function for each draw. Both limitations of the importance sampling can be overcome if the full conditional distributions of the parameters—that is, the distribution of ψ_j, given all the other parameters—can be easily simulated. It is then possible to use a correlated sampling scheme referred to as the Gibbs sampler. The main idea is to construct a Markov chain on a general state space such that the limiting distribution of the chain is the joint distribution of interest.

Let [.] and [.|.] denote marginal and conditional distributions, respectively, and suppose that for suitable choices for the ψ_j the complete conditional distributions, $[\psi_1|\psi_2, \ldots, \psi_k]$, $[\psi_2|\psi_1, \psi_3, \ldots, \psi_k]$, up to $[\psi_k|\psi_1, \ldots, \psi_{k-1}]$, where the conditioning on Y_n is suppressed, have a form that lends itself to Monte Carlo sampling. Then the Gibbs algorithm for obtaining a draw $(\psi, \ldots, \psi_k)$ from $[\psi_1, \ldots, \psi_k]$ proceeds as follows:

Step 1. Specify arbitrary initial values, $(\psi_1^{(0)}, \ldots, \psi_k^{(0)})$, and set $i = 1$.

Step 2. Cycle through the full conditionals drawing

(a) $\psi_1^{(i)}$ from $[\psi_1|\psi_2^{(i-1)}, \ldots, \psi_k^{(i-1)}]$

(b) $\psi_2^{(i)}$ from $[\psi_2|\psi_1^{(i)}, \psi_3^{(i-1)} \ldots, \psi_k^{(i-1)}]$

$\vdots$

(k) $\psi_k^{(i)}$ from $[\psi_k|\psi_1^{(i)}, \ldots, \psi_{k-1}^{(i)}]$. (2.1)

Step 3. Set $i = i + 1$, and go to step 2.

After iterating on this cycle T times, the sample value $\psi^{(T)} = (\psi_1^{(T)}, \ldots, \psi_k^{(T)})$ is obtained. Under regularity conditions (e.g., Tierney 1991), the distribution of $\psi^{(T)}$, as T approaches infinity, converges to the distribution of ψ. Convergence to the desired distribution can be informally checked via quantile-quantile plots, as suggested by Gelfand and Smith (1990). Thus, for a suitable choice of i, say M, the simulated values $(\psi_1^{(i)}, \ldots, \psi_k^{(i)})$ $(i = M + 1, \ldots, M + N)$ can be regarded as an approximate simulated sample from $[\psi_1, \psi_2, \ldots, \psi_k]$.

Once this simulated sample has been obtained, any posterior moment of interest or any marginal density can be easily estimated. Specifically, the posterior expectation of a function of the parameters, $g(\psi)$, can be estimated by the sample average

$$E(g(\psi)) \simeq \frac{1}{N} \sum_{i=M+1}^{M+N} g(\psi^{(i)}) \tag{2.2}$$

or as an average of the conditional mean of $g(\psi)$ if the latter is available. To compute the posterior density of any component, say ψ_1, we can average its full conditional distribution,

$$[\hat{\psi}_1] = \frac{1}{N} \sum_{i=M+1}^{M+N} [\psi_1 | \psi_2^{(i)}, \ldots, \psi_k^{(i)}], \tag{2.3}$$

or apply a nonparametric density estimation procedure to the simulated values. Since both (2.2) and (2.3) are sums of correlated observations, the usual "standard deviation divided by the square root of sample size" formula cannot be used as a numerical standard error for these estimates. One can, however, use the well-known batch-means method (e.g., Ripley 1987, chap. 6) or time series methods such as the spectral approach of Geweke (1991) to obtain a measure of numerical accuracy for the estimates. The batch-means method is illustrated in later examples.

3. A BAYESIAN APPROACH

Direct Bayesian inference about θ based on its posterior distribution, $\pi(\theta|Y_n) \propto \pi(\theta) f(Y_n|\theta)$, where $\pi(\theta)$ is the prior and $f(Y_n|\theta)$ is the likelihood function, is not an attractive option here, since it entails the computation of the complicated likelihood. We therefore propose treating the states $\{s_t\}$ as additional unknown parameters and then analyzing them jointly with θ by Monte Carlo methods. The crucial point is that the joint posterior distribution of (θ, S_n) is proportional to (1.6) and does not invoke the likelihood function of θ. As we show in this section, the joint posterior distribution of (θ, S_n) leads to a very tractable conditional structure. Given the states, the posterior distribution for the parameters can be derived, while, given the parameters, the conditional distribution of the states can again be found, though the derivations are a bit more involved. In particular, the conditional distributions that form the basis of the simulation are given by

- $[s_t|Y_n, S_{-t}, \theta], \quad t = 1, \ldots, n$
- $[\beta, \gamma|Y_n, S_n, \theta_{-(\beta,\gamma)}]$
- $[\phi|Y_n, S_n, \theta_{-\phi}]$
- $[\sigma^2|Y_n, S_n, \theta_{-\sigma^2}]$
- $[\tau^2|Y_n, S_n, \theta_{-\tau^2}]$
- $[a, b|Y_n, S_n, \theta_{-(a,b)}]$,

where $S_{-t} \equiv \{s_j: 1 \le j \le n, \text{ and } j \ne t\}$, and, for example, $\theta_{-(\beta,\gamma)}$ denotes all the parameters in θ excluding β and γ. Each of these complete conditionals can be simulated, thus leading, via the Gibbs sampler, to a posterior sample from the joint distribution of the parameters and the states.

3.1 Full Conditional of $\{s_t, t = 1, \ldots, n\}$

We begin by deriving the distribution of s_t conditional on (Y_n, S_{-t}, θ). Since s_t is a binary random variable, the distribution of interest, $\Pr(s_t|Y_n, S_{-t}, \theta)$, is a two-point distribution of the probabilities that s_t is 0 or 1. First, consider the Markov chain in isolation. In that case, it is easy to see that $\Pr(s_t|S_{-t}) \propto \Pr(s_t|s_{t-1})\Pr(s_{t+1}|s_t)$ due to the Markov property. Thus the distribution depends only on the value of the state at two neighboring points. In other words, the Markov chain is Gibbsian of order 1. In the general case, due to the autocorrelation in y, the distribution of s_t is Gibbsian with order r; that is, the complete conditional distribution depends on the states at times $(t - 1, \ldots, t - r)$ from the past and times $(t + 1, \ldots, t + r)$ into the future.

We derive the complete conditional distributions for the following three cases:

1. $[s_t|Y_n, S_{-t}, \theta]$, $t \le r$.
2. $[s_t|Y_n, S_{-t}, \theta]$, $r + 1 \le t \le n - r + 1$.
3. $[s_t|Y_n, S_{-t}, \theta]$, $n - r \le t \le n$.

For simplicity, suppress the conditioning on θ, and consider the conditionals in 2. Applying the Bayes theorem, we have that

$$\Pr(s_t|Y_n, S_{-t}) = \frac{\Pr(s_t|Y_t, S_{-t}) f(y_{t+1}, \ldots, y_n|Y_t, S_{-t}, s_t)}{f(y_{t+1}, \ldots, y_n|Y_t, S_{-t})}. \tag{3.1}$$

The second term in the numerator cancels with the denominator term of (3.1) if the observations are independent because then $(y_{t+1}, \ldots, y_n)$ is independent of s_t, given S_{-t}. Otherwise, $(y_{t+r+1}, \ldots, y_n)$ is independent of s_t, given S_{-t}. Therefore, we can write

$$\Pr(s_t|Y_n, S_{-t}) \propto \Pr(s_t|Y_t, S_{-t}) f(y_{t+1}, \ldots, y_{t+r}|Y_t, s_{-t}, S_t). \tag{3.2}$$

The first term of (3.2) is now simplified via the Bayes theorem as

$$\begin{aligned}\Pr(s_t|Y_t, S_{-t}) &\propto \Pr(s_t|Y_{t-1}, S_{t-1}) f(y_t, s_{t+1}, \ldots, s_n|Y_{t-1}, S_t) \\ &\propto \Pr(s_t|s_{t-1}) f(y_t|Y_{t-1}, S_t) \Pr(s_{t+1}|Y_t, S_t) \\ &\quad \times \Pr(s_{t+2}, \ldots, s_n|Y_t, S_{t+1}) \\ &\propto \Pr(s_t|s_{t-1}) f(y_t|Y_{t-1}, S_t) p(s_{t+1}|s_t),\end{aligned} \tag{3.3}$$

where in the third line the independence of $(s_{t+2}, \ldots, s_n)$ from s_t given s_{t+1} is used. Application of the product law to the second term of (3.2) gives

$$f(y_{t+1}, \ldots, y_{t+r}|Y_t, S_n) \propto f(y_{t+1}|Y_t, S_n) \ldots f(y_{t+r}|Y_{t+r-1}, S_n). \quad (3.4)$$

The Bayes theorem applied to each term in (3.4)—for example, to the first—yields $f(y_{t+1}|Y_t, S_{t+1})\Pr(s_{t+2}, \ldots, s_n|Y_{t+1}, S_{t+1})$, where the latter term is independent of s_t. One concludes by combining the resulting terms with (3.3) and inserting in (3.2) to obtain the result

$$\Pr(s_t|Y_n, S_{-t}) \propto \Pr(s_t|s_{t-1})\Pr(s_{t+1}|s_t) \prod_{k=t}^{t+r} f(y_k|Y_{k-1}, S_k), \quad (3.5)$$

where the conditional density of y is given in (1.5) and the constant of proportionality is the sum of the two numbers that emerges from (3.5) for $s_t = 0$ and $s_t = 1$.

Next, for cases 1 and 3 the same kind of reasoning leads to the results

$$\Pr(s_t|Y_n, S_{-t}) \propto \Pr(s_t|s_{t-1})p(s_{t+1}|s_t)f(y_t, \ldots, y_r|Y_{t-1}, S_r) \times \prod_{k=r+1}^{t+r} f(y_k|Y_{k-1}, S_k), \quad (3.6)$$

and

$$\Pr(s_t|Y_n, S_{-t}) \propto \Pr(s_t|s_{t+1})\Pr(s_{t+1}|s_t) \prod_{k=t}^{n} f(y_k|Y_{k-1}, S_k), \quad (3.7)$$

respectively, where the third term of (3.6) is obtained from (1.4) and $\Pr(s_1|s_0)$ is the stationary distribution of the Markov chain.

The preceding expressions are quite easy to program. A convenient strategy is to take the most recent value of S_n and proceed with the recursions backwards from time n. At time t, the tth element of S_n is replaced with the simulation based on the probabilities calculated via (3.5)–(3.7). Note that the first two terms in those equations, excluding $t = 1$ and n, can only take one of four possible values depending on the two neighboring states at the current point in the iterations. For example, if $s_{t-1} = 0$, $s_t = 0$, and $s_{t+1} = 0$, then those two terms are both equal to $(1 - a)$. Note that identification arguments mentioned in Section 1 dictate that a particular sample S_n be accepted only if it contains at least one draw of each state.

3.2 Full Conditionals of β, γ, σ^2

Once S_n has been simulated, two simplifications occur. First, the complete conditional distribution of $(\beta, \gamma, \sigma^2)$, given S_n, becomes independent of (a, b), and second, the model becomes linear in those parameters. The framework developed by Chib (1991) and Chib and Greenberg (1992) can be used with minor modifications to derive the remaining conditional distributions. Although we could use other priors, we work with the prior specification in which the parameters are mutually independent and

$$\beta, \gamma, \sigma^2 \propto N_k(\beta|\beta_0, B_0^{-1})N(\gamma|\gamma_0, G_0^{-1})I_{\gamma>0}\mathrm{IG}\left(\sigma^2\middle|\frac{\nu_0}{2}, \frac{\delta_0}{2}\right),$$

where I is the indicator function on $[\gamma > 0]$, IG denotes the inverse gamma distribution, and the hyperparameters $(\beta_0, B_0, \gamma_0, G_0, \nu_0, \delta_0)$ are known. The choice of these hyperparameters will, of course, be motivated by the problem at hand. Modest information on β may be incorporated by allowing that $\beta_0 = 0$ and $B_0 = \varepsilon I_{k-1}$, where ε is a small number. Because the prior of γ is restricted to the positive real line, γ_0 should be a positive number. The quantity ν_0 reflects the strength of the prior of σ^2 and can be assessed in terms of the number of pre-sample observations that are used to form the prior information.

It is now convenient to transform the variables so that the transformed y's are independent and possess a scalar covariance. For each of (Y_r, X_r, S_r), apply the mapping $Z \to Z^* \equiv Q^{-1}Z$, where $QQ' = \Omega_r$. For the rest, define $Z \to Z^* \equiv (1 + \omega s_t)^{-1/2}\phi(L)Z$. On collecting all of the n-transformed variables as (Y_n^*, X^*, S_n^*), it is clear that $Y_n^*|S_n, \eta \sim N(W^*\alpha, \sigma^2 I_n)$, where $W^* = (X^*, S_n^*)$ and $\alpha = (\beta', \gamma)'$. From this we can conclude that

$$\alpha|Y_n, S_n, \theta_{-(\beta,\gamma)} \sim N_{k+1}(\alpha|\hat{\alpha}, A^{-1})I_{[\gamma>0]}, \quad (3.8)$$

where $A = (A_0 + \sigma^{-2}W^{*\prime}W^*)$, $\hat{\alpha} = A^{-1}(A_0\alpha_0 + \sigma^{-2}W^{*\prime}Y_n^*)$, $\alpha_0 = (\beta_0', \gamma_0)'$, and $A_0 = \mathrm{diag}(B_0, G_0)$. In this distribution, the support of γ is restricted to the positive real line. An easy way to draw (β, γ) from (3.8) is to draw β from its marginal normal distribution and then draw γ from the truncated conditional of γ given β; each of these distributions can be derived from (3.8). The draw of γ from the truncated normal distribution is obtained by the method of inversion. For example, to simulate from an $N(\mu, \sigma^2)I_{[a,b]}$, we first simulate a uniform random variable U and then obtain the required draw as $\mu + \sigma\Phi^{-1}(U(p_2 - p_1) + p_1)$, where Φ^{-1} is the inverse cdf of the normal distribution, $p_1 = \Phi((a - \mu)/\sigma)$, and $p_2 = \Phi((b - \mu)/\sigma)$. Finally, the complete conditional distribution of σ^2 is

$$\sigma^2|Y_n, S_n, \theta_{-\sigma^2} \sim \mathrm{IG}\left(\frac{\nu_0 + n}{2}, \frac{\delta_0 + \|Y_n^* - X^*\beta - S_n^*\gamma\|^2}{2}\right), \quad (3.9)$$

where $\|.\|$ denotes the Euclidean norm.

3.3 Full Conditional of ω

Note that, given S_n, ω is independent of (a, b) and that the likelihood function of ω depends only on the

Table 1. Simulated Data From AR(4) Homoscedastic Markov Switching Model

	Prior		Posterior			
Parameter (true)	*Mean*	*Std. dev.*	*Mean*	*Std. dev.*	*95% interval*	*Lag1*
β (−.368)	0	5	−.360	.170	(−.740, −.07)	.70
γ (1.522)	.3	5	1.543	.173	(1.210, 1.90)	.52
ϕ_1 (.014)	0	5	.012	.136	(−.241, .297)	.60
ϕ_2 (−.058)	0	5	−.060	.124	(−.298, .185)	.54
ϕ_3 (−.247)	0	5	−.269	.112	(−.486, −.056)	.45
ϕ_4 (−.213)	0	5	−.282	.111	(−.497, −.056)	.41
σ^2 (.591)	—	—	.555	.097	(.401, .780)	.50
a (.245)	.20	.16	.243	.086	(.103, .437)	.45
b (.095)	.20	.16	.110	.039	(.046, .198)	.36

NOTE: Lag1 denotes the first-order correlation of the Gibbs run. Prior distribution of σ^2 is improper. Estimated model: AR(4) Markov switching; $n = 135$; $M = 200$; $N = 6{,}000$.

observations for which $s_t = 1$. For convenience, consider the observations between $r + 1$ and n. Letting $T \equiv \{t{:}s_t = 1\}$, transforming $\omega \to \bar{\omega} \equiv (1 + \omega)$, and choosing the prior $\mathrm{IG}(\bar{\omega}|\omega_0/2, \eta_0/2)I_{\bar{\omega}>1}$, one can show that the complete conditional distribution of $\bar{\omega}$ is

$$\bar{\omega}|Y_n, S_n, \theta_{-\omega} \propto \mathrm{IG}\left(\frac{n_1 + \omega_0}{2}, \frac{\Sigma\{(\bar{y}_t - \bar{x}_t'\beta - \gamma\bar{s}_t)/\sigma\}^2 + \eta_0}{2}\right)I_{\bar{\omega}>1}, \quad (3.10)$$

where $\bar{y}_t$, $\bar{x}_t$, and $\bar{s}_t$ are obtained by multiplying the corresponding starred variables by $(1 + \omega s_t)^{1/2}$, n_1 is the cardinality of T, and the sum is over the elements of T. Sampling this distribution is somewhat nonstandard because of the truncation from below at 1. An obvious procedure is to take the reciprocal of a draw from a gamma distribution with the same parameters as in (3.10) and then discard the draws that fall in the interval (0, 1].

3.4 Full Conditional of ϕ

To derive this distribution, we transform the model so that it has autocorrelated errors. Premultiply both sides of (1.2) by $\phi(L)^{-1}$, and define the new error as $\varepsilon_t = \phi(L)^{-1}v(s_t)^{1/2}u_t$. The transformed model is given by $y_t = x_t'\beta + \gamma s_t + \varepsilon_t$, $\phi(L)\varepsilon_t = v(s_t)^{1/2}u_t$. Conditioned on the parameters and the states, the errors $\varepsilon_t \equiv y_t - x_t'\beta - \gamma s_t$ are degenerate. Thus the desired conditional distribution may be computed from the model $\varepsilon_t = \phi_1\varepsilon_{t-1} + \cdots + \phi_r\varepsilon_r + v(s_t)^{1/2}u_t$. Form the vector e^* with tth element $(1 + \omega s_t)^{-1/2}\varepsilon_t$, form the matrix E^*: $n - r \times r$ with tth row given by $(\varepsilon_{t-1}, \ldots, \varepsilon_{t-r})$, and let the prior of ϕ be $N_r(\phi|\phi_0, \Phi_0^{-1})$ on the region S_ϕ, where the roots of $\phi(L)$ lie outside the unit circle. Then the complete conditional distribution is given by

$$\phi|Y_n, S_n, \theta_{-\phi} \propto \psi(\phi)N(\hat{\phi}, \Phi_n^{-1})I_{S_\phi}, \quad (3.11)$$

where $\psi(\phi) = |\Omega_r|^{-1/2} \exp(-(1/2\sigma^2)(Y_r - X_r\beta)'\Omega_r^{-1}(Y_r - X_r\beta))$, $\hat{\phi} = \Phi_n^{-1}(\Phi_0\phi_0 + \sigma^{-2}E'\varepsilon)$, and $\Phi_n = (\Phi_0 + \sigma^{-2}E'E)$. There are two ways of proceeding. One is by ignoring $\psi(\phi)$, as in the work of Chib (in press) and the other is by using rejection sampling as in the work of Chib and Greenberg (1992). The former procedure, which amounts to conditioning on the first r observations, leads to easier simulations and is used in this article; one makes a draw of ϕ from $N(\hat{\phi}, \Phi_n^{-1})$, accepting it if it lies in S_ϕ. As long as most of the mass of the posterior is over the stationary region, this procedure will be quite efficient.

3.5 Full Conditionals of (*a*, *b*)

Since (a, b), given $s_1, \ldots, s_n$, is independent of $(Y_n, \beta, \gamma, \sigma^2)$, we need to consider only the conditional

Table 2. Simulated Data From AR(0) Homoscedastic Markov Switching Model

	Prior		Posterior			
Parameter (true)	*Mean*	*Std. dev.*	*Mean*	*Std. dev.*	*95% interval*	*Lag1*
β (−.368)	0	5	−.486	.191	(−.882, −.127)	.65
γ (1.522)	.3	5	1.586	.193	(1.186, 1.96)	.52
ϕ_1 (0)	0	5	−.053	.135	(−.304, .244)	.57
ϕ_2 (0)	0	5	−.037	.129	(−.305, .204)	.54
ϕ_3 (0)	0	5	−.130	.119	(−.357, .118)	.41
ϕ_4 (0)	0	5	−.016	.207	(−.238, .207)	.41
σ^2 (.591)	—	—	.579	.106	(.401, .823)	.53
a (.245)	.20	.16	.262	.089	(.112, .457)	.44
b (.095)	.20	.16	.114	.051	(.040, .236)	.60

NOTE: Lag1 denotes the first-order correlation of the Gibbs run. Prior distribution of σ^2 is improper. Estimated model: AR(4) Markov switching; $n = 135$; $M = 200$; $N = 6{,}000$.

Table 3. Simulated Data From Constant Mean–Constant Variance Model [see (5.2)]

	Prior		Posterior			
Parameter (true)	*Mean*	*Std. dev.*	*Mean*	*Std. dev.*	*95% interval*	*Lag1*
β (1.154)	0	5	.765	.251	(.136, 1.10)	.65
γ (0)	.3	5	.429	.264	(.029, 1.04)	.64
ϕ_1 (0)	0	3	−.087	.106	(−.306, .111)	.30
ϕ_2 (0)	0	3	−.156	.105	(−.374, .039)	.26
ϕ_3 (0)	0	3	−.081	.101	(−.285, .112)	.22
ϕ_4 (0)	0	3	−.181	.095	(−.370, .002)	.17
σ^2 (.591)	—	—	.543	.081	(.391, .713)	.31
a (0)	.20	.16	.251	.164	(.026, .644)	.74
b (0)	.20	.16	.100	.091	(.006, .351)	.88

NOTE: Lag1 denotes the first-order correlation of the Gibbs run. Prior distribution of σ^2 is improper. Estimated model: AR(4) Markov switching; n = 135; M = 200; N = 6,000.

distribution $a, b|S_n$, which can be obtained from standard Bayesian results on Markov chains. Given the data S_n, the sufficient statistics for a and b are the transitions, n_{ij} from state i to j. The likelihood function, conditioned on the initial state, is given by

$$L(a, b) = (1 - a)^{n_{00}}a^{n_{01}}b^{n_{10}}(1 - b)^{n_{11}}. \quad (3.12)$$

From the form of the likelihood, it is clear that the beta family of distributions is a conjugate prior for each of the transition probabilities. Therefore, let (a, b) be distributed as $\pi(a, b) \propto (1 - a)^{u_{00}}a^{u_{01}}b^{u_{10}}(1 - b)^{u_{11}}$, where the u_{ij} are the hyperparameters of the prior. If it is believed that the shifts between states occur occasionally, these hyperparameters can be chosen such that the bulk of the prior mass on a and b is in the interval (0, .5). Combining with (3.12), the desired posterior is also a product of independent beta distributions

$$a|s_1, \ldots, s_n \sim \text{beta}(u_{01} + n_{01}, u_{00} + n_{00}) \quad (3.13)$$

and

$$b|s_1, \ldots, s_n \sim \text{beta}(u_{10} + n_{10}, u_{11} + n_{11}). \quad (3.14)$$

3.6 Initialization of the Gibbs Sampler

The Gibbs sampler may now be run by cycling through the full conditionals of ϕ, β, γ, σ^2, ω, s_t, and a, b, in

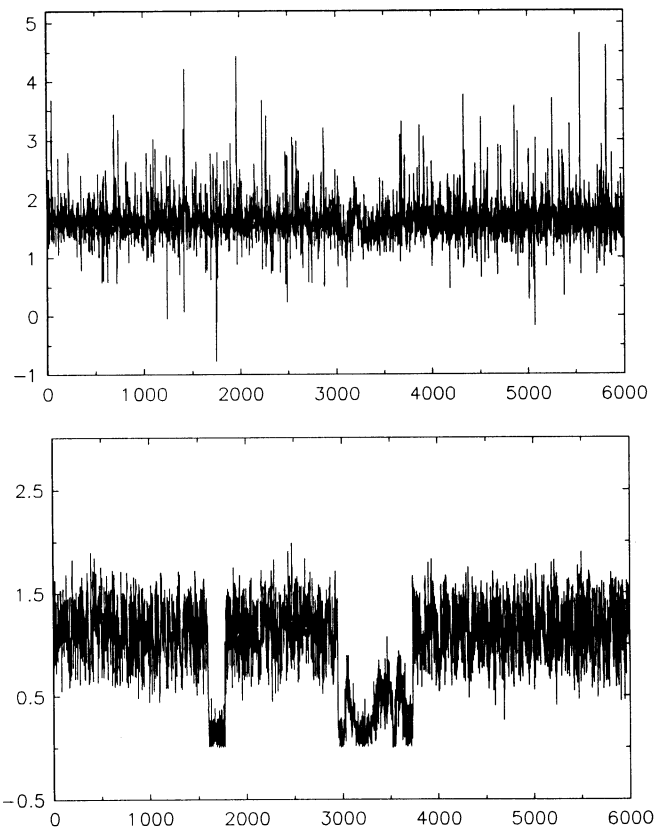

Figure 1. Gibbs Run for β and γ: Interest-Rate Data Set.

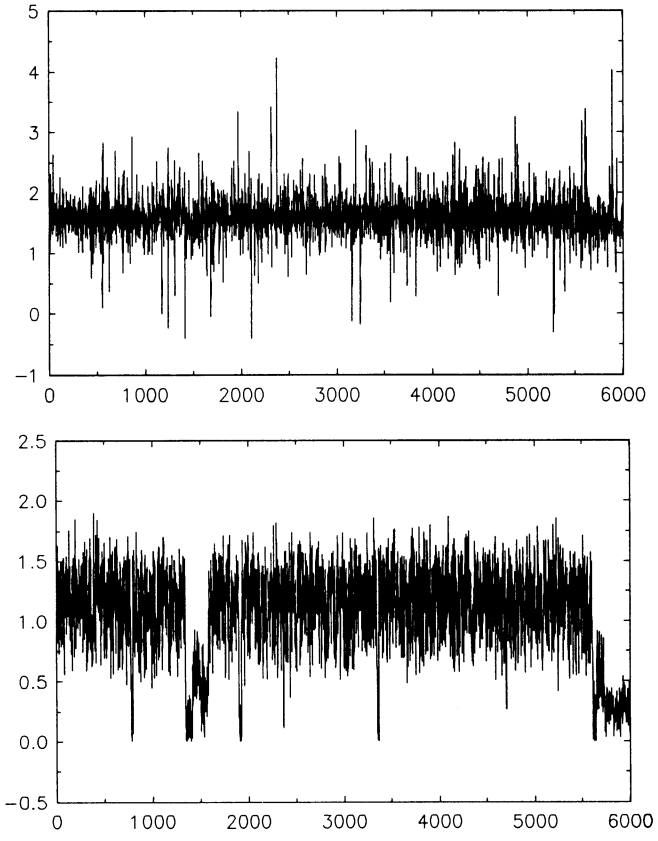

Figure 2. Gibbs Run for β and γ: Tight Prior.

that order. To initialize the Gibbs sampler, values of a and b are specified. Next, S_n can be simulated via the Markov chain, the estimates of $(\beta, \gamma, \sigma^2)$ are then set to least squares values with the constraint that γ is positive, and ω is set equal to 0. In our work, the initial values of a and b are the posterior mean values from a simulation run of $N = 500$ in the reduced model wherein $\phi(L) = 1$. To see if results are sensitive to the start-up values, different starting values can be tried.

4. EXTENSIONS

Based on the draws from the Gibbs sampling procedure it is straightforward to compute the posterior distributions of the error and of the out-of-sample observations. We briefly consider each of these problems.

4.1 Realized Error Analysis

The objective is to compute the posterior distribution of the error u_t that appears in (1.1). Because the states are simulated in our procedure, it is an easy matter to compute the residual for each time point and for every draw of $(\eta^{(i)}, S_n^{(i)})$ in the Gibbs cycle. We define, using obvious notation, the tth residual at the ith point in the Gibbs run as follows:

$$U_r^{(i)} = \Omega_r^{-1/2}(Y_r - X_r\beta^{(i)} - S_r^{(i)}\gamma^{(i)})$$

$$u_t^{(i)} = v(s_t^{(i)})^{-1/2}(1 - \phi_1^{(i)}L - \cdots - \phi_r^{(i)}L^r) \times (y_t - x_t\beta^{(i)} - s_t^{(i)}\gamma^{(i)}), \quad (4.1)$$

where the first equation gives the residual for the first r observations and the second equation gives the residual for the remaining observations. Posterior moments of the residual can be computed using (2.2); moreover, given that the conditional distribution of the residual is standard normal, the posterior distribution can be estimated using (2.3).

4.2 Prediction Density

We now consider the issue of forecasting future observations, an important issue with time series data. Consider the first out-of-sample observation, y_{n+1}. Let $S_t^k = (s_t, \ldots, s_k)$ denote the states starting at time t and ending at time k, and assume that the covariates are known. The objective is to determine the Bayes prediction density, $f(y_{n+1}|Y_n)$, which can be simplified as

$$\begin{aligned} f(y_{n+1}|Y_n) &= \int f(y_{n+1}|Y_n, s_{n+1}, S_{n+1-r}^n, \theta)d[s_{n+1}, S_{n+1-r}^n, \theta|Y_n] \\ &= \int f(y_{n+1}|Y_n, s_{n+1}, S_{n+1-r}^n, \theta)d[s_{n+1}|s_n, \theta]d[S_{n+1-r}^n, \theta|Y_n], \end{aligned} \quad (4.2)$$

where the conditional density of y_{n+1} is obtained from (1.5) by setting $t = n + 1$. Samples from the Bayes prediction density can be obtained by applying the method of composition to (4.2). For each draw of (S_{n+1-r}^n, θ) made available via the Gibbs sampler, we sample

$$\text{(a) } s_{n+1} \text{ from } \Pr(s_{n+1}|s_n, \theta)$$

$$\text{(b) } y_{n+1} \text{ from (2.2).} \quad (4.3)$$

These two steps can obviously be implemented comfortably along with the regular Gibbs cycle. Thus at the end of the algorithm one obtains samples from both the parameter posterior and the prediction posterior. It is easy to see how this process may be iterated to obtain a draw from the prediction density of any future ob-

Table 4. AR(4) Heteroscedastic Markov Switching Model for Interest-Rate Data

	Prior		Posterior				
Parameter	Mean	Std. dev.	Mean	Std. dev.	95% interval	Lag1	MLE
β	0	10	1.637 (.006)	.330	(1.143, 2.357)	.23	1.633 (.208)
γ	.3	10	1.035 (.021)	.393	(.065, 1.575)	.78	1.188 (.23)
ϕ_1	0	5	.838 (.0032)	.117	(.604, 1.053)	.22	.878 (.103)
ϕ_2	0	5	.109 (.0050)	.184	(−.219, .467)	.44	.068 (.15)
ϕ_3	0	5	.103 (.0044)	.173	(−.232, .452)	.43	.169 (.152)
ϕ_4	0	5	−.159 (.0022)	.109	(−.375, .061)	.21	−.211 (.095)
σ^2	—	—	.033 (.0003)	.007	(.020, .047)	.49	.0307
τ^2	—	—	.549 (.0083)	.362	(.139, 1.44)	.23	.5060
a	.167	.141	.038 (.0007)	.024	(.008, .099)	.28	.0396 (.01)
b	.20	.214	.180 (.0024)	.101	(.032, .419)	.19	.0901 (.082)

NOTE: Numerical standard error of posterior mean is in parentheses. Lag1 denotes the first-order correlation of the Gibbs run. Prior distributions of σ^2 and τ^2 are improper. For MLE, standard error is in parentheses. $n = 103$; $M = 200$; $N = 6{,}000$.

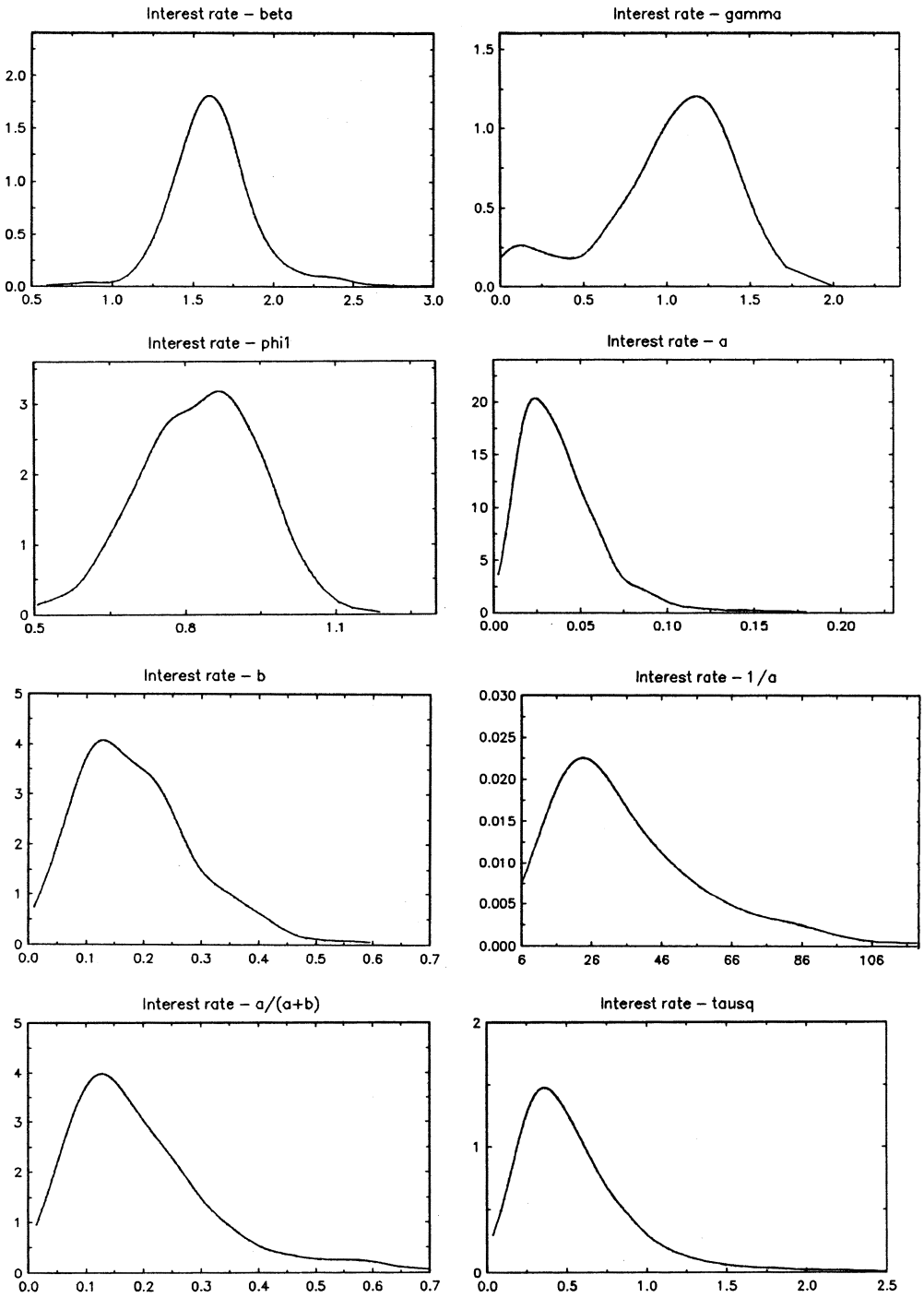

Figure 3. Marginal Posterior Densities: Interest Rate Data Set.

servation. For specificity consider y_{n+2}. Then, the steps

(a) s_{n+2} from $\Pr(s_{n+2}|s_{n+1}, \theta)$,

using s_{n+1} drawn in (a) of (4.3)

(b) y_{n+2} from (1.5),

using y_{n+1} drawn in (b) of (4.3) (4.4)

provide the desired sample. On the basis of the generated sample, prediction standard errors and prediction densities can be calculated. Unlike prediction using classical approaches, the method embodied in (4.3) and (4.4) yields predictive inferences that incorporate both parameter uncertainty and state uncertainty.

5. EXAMPLES

In this section, we illustrate the proposed methodology using simulated and real data sets, focusing on inferences about θ, S_n, and future observations. We also provide evidence on how the method works when an incorrect model is fit to the data. For instance, the data might be generated by a Markov switching model with

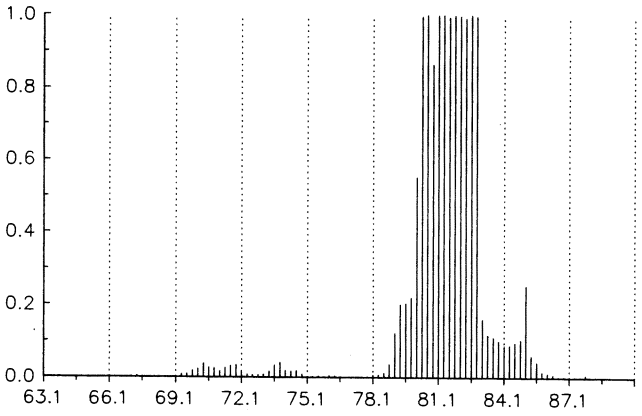

Figure 4. $Pr(s_t = 1|Y_n)$: Interest-Rate Data Set.

uncorrelated errors, and we might fit a Markov switching AR model to it. Another case might be that the data come from a pure AR model with no Markov switching, but we estimate a Markov switching model. This investigation clearly reveals that the approach reproduces the model that generated the data. We also consider two data sets that were analyzed previously by Hamilton (1988, 1989) by the ML method. Our results are broadly similar to those obtained by him, although in one case (the GNP data set) we find support for Markov switching but not for any AR components.

One of the virtues of the Gibbs-sampling approach is that it provides a sample from the joint posterior distribution of all the parameters and the states. For our purposes, we find it convenient to summarize the information obtained in terms of prior and posterior moments, 95% intervals based on the 2.5th and the 97.5th percentiles of the simulated draws, lag 1 correlation of the Gibbs run, and marginal posterior densities. The prior distributions of a, b are specified such that the means are in the interval (0, .5) but with large standard deviations. The other priors are also specified as being quite uninformative. The results generally are not dependent on the priors selected, and therefore a sensitivity analysis with respect to the prior inputs is not provided.

In our examples, the posterior moments are computed by averaging the simulated draws; the method of conditioning was not found to reduce the accuracy of the estimates and was therefore not used. Densities are computed by Gaussian kernel smoothing, although in many cases it is possible to use (2.3). The Gibbs simulation is run such that the first 200 draws are discarded and then the next 6,000 are recorded. Thus, using the notation of Section 2, $M = 200$ and $N = 6{,}000$. The numerical accuracy of the posterior mean estimates is obtained by the batch-means method (see Ripley 1987). The 6,000 simulated values were sectioned in v batches of size $6{,}000/v$. The size of each batch was increased until the lag correlation of the batch means is under 5%. The numerical standard error is estimated by $s/\sqrt{(v)}$, where s is the standard deviation of the batch means.

5.1 Simulated Data

We first consider data generated by an AR(4) Markov switching model with $x_t = 1$, for all t. The parameter values used to generate the data of $n = 135$ observations are given by

$$\beta = -.368; \quad \gamma = 1.522;$$
$$\phi = (.014, -.058, -.247, -.247);$$
$$\sigma^2 = .591; \quad a = .245; \quad b = .095. \qquad (5.1)$$

Table 5. Pure AR(4) Model for Percentage Change in U.S. GNP

	Prior		Posterior				
Parameter	Mean	Std. dev.	Mean	Std. dev.	95% interval	Lag1	MLE
β	0	5	.744 (.0016)	.122	(.512, .993)	.02	.711 (.129)
ϕ_1	0	5	.315 (.001)	.091	(.137, .496)	.02	.312 (.089)
ϕ_2	0	5	.129 (.001)	.093	(−.057, .315)	.01	.122 (.093)
ϕ_3	0	5	−.115 (.001)	.091	(−.288, .067)	−.01	−.116 (.092)
ϕ_4	0	5	−.083 (.001)	.093	(−.262, .098)	.03	−.081 (.089)
σ^2	—	—	1.028 (.002)	.128	(.807, 1.309)	.03	1.001
y_{n+1} (1.1894)			.304 (.014)	1.055	(−1.763, 2.41)	.02	
y_{n+2} (.6047)			.518 (.0141)	1.090	(−1.611, 2.65)	.01	
y_{n+3} (1.012)			.695 (.0145)	1.126	(−1.480, 2.89)	−.02	
y_{n+4} (.7289)			.782 (.015)	1.127	(−1.531, 2.95)	−.02	

NOTE: Numerical standard error of posterior mean is in parentheses. Lag1 denotes the first-order correlation of the Gibbs run. Prior distribution of σ^2 is improper. Actual value of future observation is in parentheses. For MLE, standard error is in parentheses. $n = 135$; $M = 200$; $N = 6{,}000$.

Table 6. AR(4) Markov Switching Model for Percentage Change in U.S. GNP

	Prior		Posterior				
Parameter	*Mean*	*Std. dev.*	*Mean*	*Std. dev.*	*95% interval*	*Lag1*	*MLE*
β	0	5	−.376 (.026)	.424	(−1.19, .519)	.82	−.368 (.265)
γ	.5	5	1.444 (.026)	.413	(.416, 2.21)	.77	1.522 (.264)
ϕ_1	0	5	.184 (.008)	.148	(−.110, .452)	.65	.014 (.120)
ϕ_2	0	5	.067 (.007)	.138	(−.216, .328)	.61	−.058 (.137)
ϕ_3	0	5	−.160 (.005)	.120	(−.396, .081)	.46	−.247 (.107)
ϕ_4	0	5	−.146 (.004)	.110	(−.362, .067)	.37	−.213 (.110)
σ^2	—	—	.701 (.008)	.146	(.477, 1.066)	.64	.591
a	.20	.16	.302 (.006)	.131	(.089, .583)	.65	.245 (.097)
b	.20	.16	.109 (.006)	.069	(.027, .377)	.80	.095 (.034)
y_{n+1} (1.1894)			.409 (.013)	1.013	(−1.68, 2.27)	.05	
y_{n+2} (.6047)			.715 (.014)	1.013	(−1.494, 2.68)	.05	
y_{n+3} (1.012)			.875 (.014)	1.102	(−1.38, 2.84)	.07	
y_{n+4} (.7289)			.897 (.014)	1.090	(−1.38, 2.90)	.03	

NOTE: Numerical standard error of posterior mean is in parentheses. Lag1 denotes the first-order correlation of the Gibbs run. Prior distribution of σ^2 is improper. Actual value of future observation is in parentheses. For MLE, standard error is in parentheses. $n = 135$; $M = 200$; $N = 6{,}000$.

Thus in this model the mean is specified by a constant plus the state switching variable, and the parameters ϕ_1 and ϕ_2 are close to 0. (These parameter values are actually the ML estimates in the GNP example considered later.) Our results are summarized in Table 1 (Tables 1–7 are on pages 6–11). Numerical standard errors are not provided because they are all small and are not central to our illustration. We can clearly see that the posterior means are generally close to the true values that generated the data. The posterior standard deviations and the 95% posterior intervals indicate that all of the posterior distributions are concentrated around the true values. The posterior correlation matrix of the parameters (not included here) displays some positive

Table 7. AR(0) Markov Switching Model for Percentage Change in U.S. GNP

	Prior		Posterior			
Parameter	*Mean*	*Std. dev.*	*Mean*	*Std. dev.*	*95% interval*	*Lag1*
β	0	5	−.411 (.017)	.337	(−1.16, .156)	.79
γ	.5	5	1.538 (.017)	.286	(.998, 2.159)	.61
σ^2	—	—	.736 (.0034)	.122	(.535, 1.011)	.42
a	.20	.16	.276 (.004)	.104	(.110, .507)	.50
b	.20	.16	.108 (.002)	.053	(.032, .239)	.66
y_{n+1} (1.1894)			.746 (.014)	1.084	(−1.49, 2.739)	.01
y_{n+2} (.6047)			.725 (.0141)	1.076	(−1.546, 2.71)	.02
y_{n+3} (1.012)			.718 (.014)	1.089	(−1.567, 2.66)	.01
y_{n+4} (.7289)			.728 (.014)	1.096	(−1.599, 2.72)	.03

NOTE: Numerical standard error of posterior mean is in parentheses. Lag1 denotes the first-order correlation of the Gibbs run. Prior of σ^2 is improper. Actual value of future observation is in parentheses. $n = 135$; $M = 200$; $N = 6{,}000$.

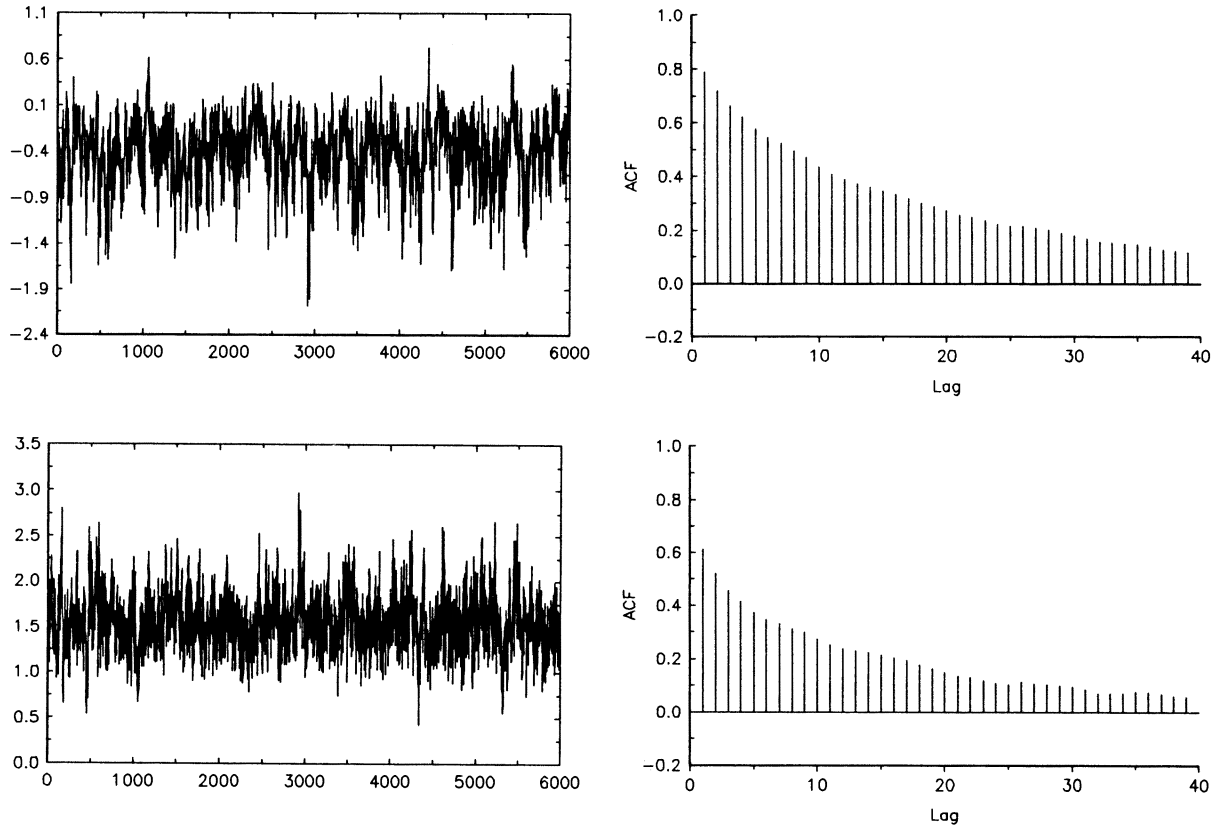

Figure 5. Gibbs Run and Autocorrelation Functions for β and γ: GNP Data Set.

correlation among the parameters, but it is high only for β and γ and to a lesser extent for (β, γ) and (a, b). Note also that serial correlation in the Gibbs run is not a problem and that the autocorrelation functions for each parameter (which are not reported to conserve space) damp down to 0 by the 10th to 15th lag.

In the preceding example, two of the parameters were close to but not exactly equal to 0. What happens if the model is overfit? We generate data in which all of the ϕ's are 0 [the AR(0) model] and the remaining parameters are as in (5.1). An AR(4) model is estimated again on 135 data points. The results are provided in Table 2. We find that the point estimates of ϕ_1, ϕ_2, and ϕ_4 are close to 0. Even though the point estimate of ϕ_3 is $-.130$, its 95% posterior interval contains 0, allowing us to conclude that our method accurately reproduces the correct model and the true parameter values.

We also study the case in which the true data is generated from a process with no Markov switching and a Markov switching AR(4) model is estimated. The true data is generated from the model

$$y_t = 1.154 + u_t,\ u_t \sim \text{iid}N(0, .591), \qquad (5.2)$$

which we refer to as the AR(0) constant mean model. The results are reported in Table 3. The estimate of σ^2 agrees closely with the true value, while the sum of the posterior mean of β and γ is 1.189; this is equal to the unconditional mean up to the first decimal place. An important observation is that if the mean of y is positive *both* β and γ end up positive. The behavior of the Gibbs sampler is very interesting. Because γ is constrained to be positive, the sampler assigns the value 1 to most of the states leading to a low posterior mean of b. The unconditional mean value of y is then split up between β and γ. By observing that all of the ϕ's are close to 0 and the posteriors of γ, a, and b have high variability, however, it is possible to conclude that there is no Markov switching in the data.

5.2 Interest-Rate Data

We next analyze the data set considered by Hamilton (1988). The dependent variable is the yield to maturity (multiplied by 100) on three-month Treasury bills (quarterly rate) for the period 1962.1 to 1987.3. The model is specified as

$$(1 - \phi_1 L - \cdots - \phi_4 L^4)(y_t - \beta - \gamma s_t) = \sigma(1 + \omega s_t)^{1/2} u_t, \qquad (5.3)$$

that is, as a fourth-order process with heteroscedastic variances. Figure 1 plots the 6,000 simulated values for the parameters β and γ. For β, the simulated values quickly settle down in the range 1–2.5. The simulated values of γ behave similarly with the exception of occasional visits to values close to 0. This behavior suggests that the posterior density for γ is bimodal. To investigate whether this behavior is an indication that γ is not identified, we also try a tighter prior on γ with a mean of 1 and standard deviation of 2, leaving all

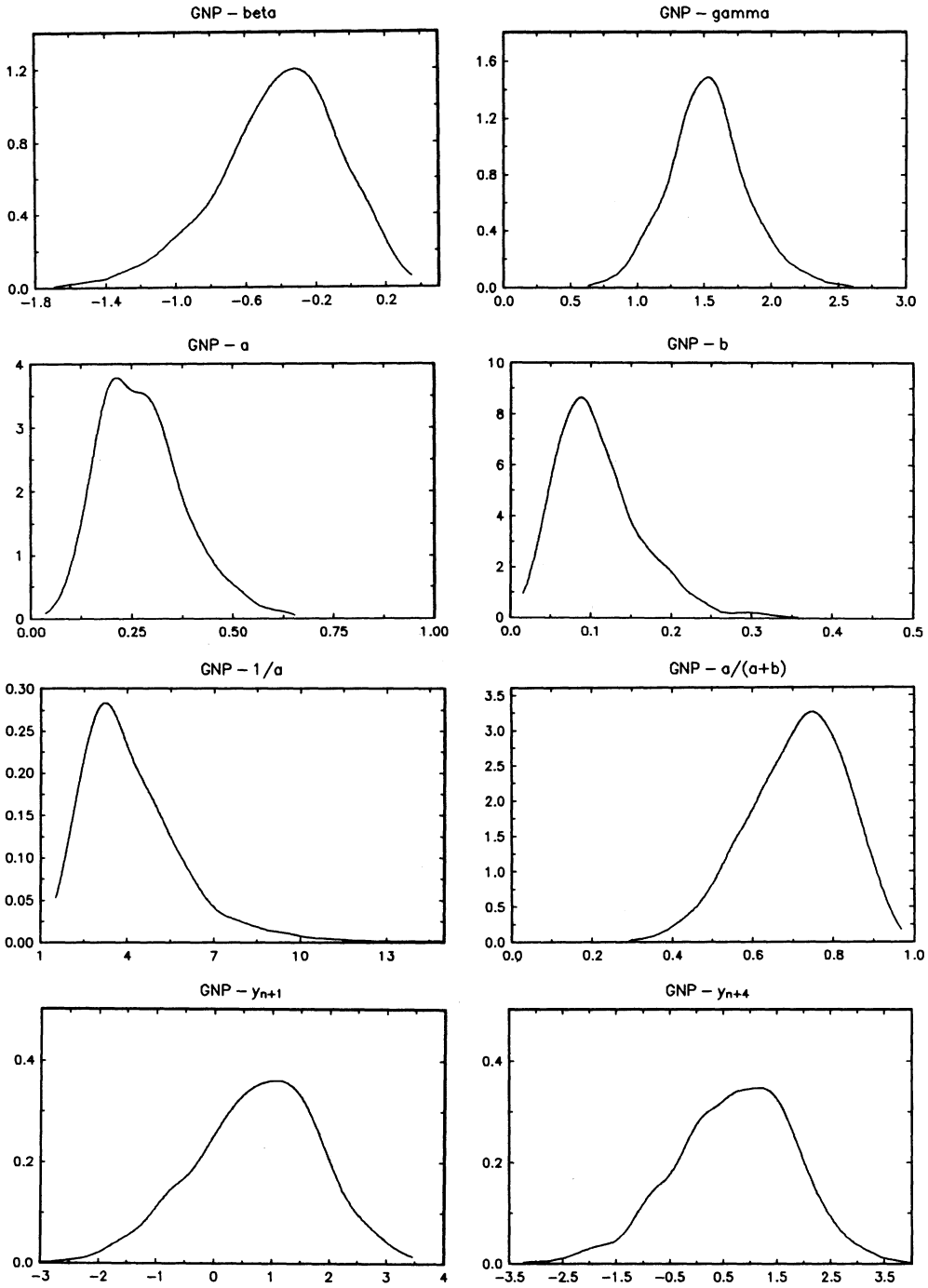

Figure 6. Marginal Posterior Densities: GNP Data Set.

other priors unchanged. The posterior moments are almost identical to those with the more diffuse prior, and the Gibbs run for γ, which is reported in Figure 2, displays the same tendency to visit 0, although the timing is now different. Other investigations of the same type lead us to conclude that the bimodality of the posterior of γ is a real feature of this data set. (Figs. 1–6 are on pp. 7–13.)

In Table 4, the Gibbs posterior mean estimates for the parameters β, γ, ϕ, σ^2, τ^2, a, and b. The posterior correlation matrix is not reported but the finding is that $\text{corr}(\phi_1, \phi_2) = .659$ and $\text{corr}(\phi_3, \phi_4) = -.614$, while the other correlations are negligible, even that between β and γ. It is interesting to compare the Bayes results with the ML estimates computed by Hamilton (1988). Generally, the posterior means and the ML values are in close agreement. Possible exceptions to this agreement are the ϕ_3 and ϕ_4 parameters and the Markov-state probability b. Even in these cases, the ML estimates are within one posterior standard deviation of the posterior mean.

Figure 3 gives marginal posterior density estimates for some of the parameters of interest. These plots were constructed using normal kernel density estimates from

every 5th draw of the Gibbs run; values are skipped to achieve an approximately independent sample. Some general comments can be made from viewing these plots. First, the shapes of the densities are distinctly skewed; the usual ML assumption of normality appears to be inaccurate for these parameters. The posterior density of τ^2 is concentrated away from 0, indicating that the variance in the high state is larger than the variance in the low state.

Figure 4 plots the estimated probability that the state variable s_t is equal to 1 given the *entire* sample information and marginalizing over all other parameters. Note that this estimated probability is close to 0 except for a short period from 1979:4 through 1982:3, which parallels the results obtained by Hamilton (1988). This period reflects the effects of the change in the Federal Reserve's operating procedures in October 1979.

5.3 GNP Data

We also study the U.S. GNP data set analyzed by Hamilton (1989). The variable of interest is the percentage change (multiplied by 100) in the postwar real GNP covering the period 1951.2 to 1984.4. Hamilton interpreted the state $s_t = 1$ as corresponding to booms in the economy and the state $s_t = 0$ to recessions.

Since Hamilton estimated an AR(4) Markov switching model, we first estimate a pure stationary AR(4), $(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_4 L^4)(y_t - \beta) = \sigma\varepsilon_t$. This model is estimated by dropping s_t, γ, and τ^2 from (1.1) and setting $x_t = 1$. To check the predictive capabilities of the model we forecast out-of-sample values of the dependent variable for the four quarters (of 1985) succeeding the last sample observation. The actual values for the forecast period are calculated from the *Business Conditions Digest*, September 1989, p. 101, series 50. The results are reported in Table 5. In summary, we note that only the first autoregressive coefficient is positive and significant; the 95% equal-tailed posterior interval includes 0 for ϕ_2, ϕ_3, and ϕ_4; the prediction standard deviations at all four lead times is large. The large prediction intervals reflect the combined influences of the parameter uncertainty and residual error uncertainty; the latter is significant for this model, as can be seen from the distribution of σ^2. In fact, the mean value of σ is almost as large as the within-sample mean value of y_t.

Next, in Table 6 we report results for the same model as in (5.3) but with a homoscedastic variance. We can observe that for most parameters the Bayes posterior means are close to the ML estimates, but there are some important differences. In particular, note the difference in the estimate of ϕ_1 (the posterior mean is much larger) and between the standard errors of a and b and their posterior standard deviations (the latter indicating more variation). The Bayes posterior distribution for ϕ tends to suggest that the model is overparameterized and that the ϕ's could be dropped from the model. Other than the high correlation in the Gibbs run for β and γ (their autocorrelation damps to 0 only around the 55th lag), the estimated results display the same features as those in Table 2. It may be mentioned that the correlation problem is made worse if improper priors on β and γ are used; proper priors with diffuse hyperparameters are useful in this problem.

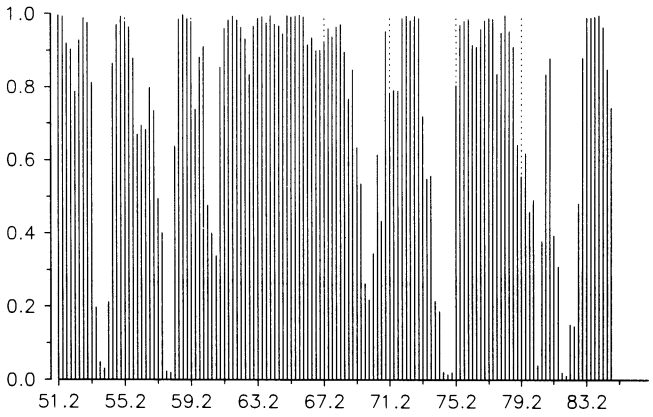

Figure 7. $Pr(s_t = 1|Y_n)$: GNP Data Set.

We therefore estimate the Markov switching model under the assumption that the errors are uncorrelated. The results are reported in Table 7. We make the following observations: The constant β is not different from 0; the variance of the error is slightly increased but the point forecasts (the mean of the posterior) are closer to the true values than the point forecasts from the AR switching model; the correlation problem in the Gibbs run for β and γ is mitigated as can be seen in Figure 5 by the plots of their simulated values and their autocorrelation functions. In this connection, it may be mentioned that the autocorrelation function of the mean in the high state $(\beta + \gamma)$ looks quite like the autocorrelation function of γ (with correlation at lag 1 of .61 and at lag 40 of .11). In addition, the posterior mean and standard deviation of $(\beta + \gamma)$ are 1.128 and .142, respectively. Due to negative posterior correlation between β and γ, the sum is more precisely estimated than either β or γ. We conclude that the AR(0) Markov switching model is a useful description for the GNP data set. Next, in Figure 6 we give the marginal posterior densities for six parameters of interest and the prediction densities at leads times of 1 and 4. We note that except for γ, all the pdf's display some skewness. Figure 7 plots the estimated values $\Pr(s_t = 1)$ against time t. On the basis of these probabilities, it is not difficult to classify the observations into one or the other state. In fact, the timing of booms and recessions implied by Figure 7 is consistent with those reported by Hamilton (1989).

6. CONCLUSION

We have used Gibbs sampling to summarize the joint posterior distributions of parameters, the unknown states, and the future observations for an autoregressive time series model with Markov jumps. The goal was to pre-

sent the sampling algorithm and illustrate the use of the simulation output on specific estimation problems. This methodology has some important advantages over the ML fitting approach of Hamilton (1988). First, this simulation algorithm is relatively easy to implement. The Gibbs sampler involves simulation from a number of conditional posterior distributions, all of which are of standard functional forms. Second, the simulation output gives much more information about the parameters than the ML approach. Since draws from the joint posterior distribution of the entire set of parameters are generated, it is easy to estimate the marginal posterior density for any function of the parameters. Posterior density plots such as those presented in Figures 3, 6 indicate that the densities may be significantly skewed or display multiple modes. In contrast, in the ML approach, it can be difficult to remove nuisance parameters. For example, inferences about s_t are based on the sampling distribution conditional on the estimated value of θ. This is in contrast to the probabilities of s_t presented in Figures 4 and 7 that represent *marginal* posterior probabilities of the high state. Likewise, predictive inferences are based on marginalization over both parameters and states. Sampling distributions in the MLE approach are typically assumed to be of Gaussian shape, but Figures 1–7 indicate that this may be a poor approximation for some of the parameters.

In conclusion, we emphasize that the Markov switching model and the inference approach developed for it in this article can be readily applied to a variety of other problems. The method is appropriate for the standard state-space model with discrete jumps in the model (see Shumway and Stoffer [1991] for switches that occur according to a time-independent random process). Another interesting use is in discrete response data models (see Albert and Chib [in press] for the Gibbs formulation without Markov switching), particularly when they are applied to time series data or panel data. Poisson regression models with Markov switching can also be considered. Applications to these other models is currently under study and will be reported elsewhere.

ACKNOWLEDGMENTS

Versions of this article have been presented at Indiana University, University of Missouri, and the University of Rochester. We thank Ed Greenberg, Bruce Hansen, Adrian Pagan, and two anonymous referees for their comments.

[*Received October 1991. Revised July 1992.*]

REFERENCES

Albert J., and Chib, S. (in press), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88.

Baum, L. E., and Eagon, J. A. (1967), "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, 37, 1554–1563.

Carlin, B., Polson, N., and Stoffer, D. (1992), "A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling," *Journal of the American Statistical Association*, 87, 493–500.

Chib, S. (in press), "Bayes Regression With Autocorrelated Errors: A Gibbs Sampling Approach," *Journal of Econometrics*, 47.

Chib, S., and Greenberg, E. (1992), "Bayes Inference via Gibbs Sampling of Regression Models With AR(p) and MA(q) Errors," unpublished manuscript.

Everitt, B. S., and Hand, D. J. (1981), *Finite Mixture Distributions*, New York: Chapman and Hall.

Gelfand, A. E., Hills, S. I., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association*, 85, 972–985.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geweke, J. (1991), "Evaluating the Accuracy of Sampling Based Approaches to the Calculations of Posterior Moments," unpublished manuscript.

Goldfeld, S. M., and Quandt, R. E. (1973), "A Markov Model for Switching Regressions," *Journal of Econometrics*, 1, 3–16.

Hamilton, J. D. (1988), "Rational Expectations Econometric Analysis of Changes in Regime: An Investigation of the Term Structure of Interest Rates," *Journal of Economic Dynamics and Control*, 12, 385–423.

——— (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357–384.

——— (1991), "Estimation, Inference, and Forecasting of Time Series Subject to Changes in Regime," unpublished manuscript.

Juang, B. H., and Rabiner, L. R. (1985), "Mixture Autoregressive Hidden Markov Models for Speech Signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 6, 1404–1413.

Lindgren, G. (1978), "Markov Regime Models for Mixed Distributions and Switching Regressions," *Scandinavian Journal of Statistics*, 5, 81–91.

McCulloch, R. E., and Tsay, R. E. (1991), "Bayesian Analysis of Autoregressive Time Series via the Gibbs Sampler," unpublished manuscript.

Pagan, A. R., and Schwert, G. W. (1990), "Alternative Models for Conditional Stock Volatility," *Journal of Econometrics*, 45, 267–290.

Ripley, B. D. (1987), *Stochastic Simulation*, New York: John Wiley.

Sclove, S. L. (1983), "Time Series Segmentation: A Model and a Method," *Information Sciences*, 29, 7–25.

Shumway, R. H., and Stoffer, D. S. (1991), "Dynamic Linear Models With Switching," *Journal of the American Statistical Association*, 86, 763–769.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–549.

Tierney, L. (1991), "Markov Chains for Exploring Posterior Distributions," unpublished manuscript.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, Chichester, U.K.: John Wiley.

Tong, H. (1983), *Threshold Models in Non-linear Time Series Analysis*, New York: Springer-Verlag.

Tsurumi, H. (1988), "Survey of Bayesian and Non-Bayesian Testing of Model Stability in Econometrics," in *Bayesian Analysis of Time Series and Dynamic Linear Models*, ed. J. C. Spall, New York: Marcel Dekker, pp. 75–100.