

A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling

BRADLEY P. CARLIN, NICHOLAS G. POLSON, and DAVID S. STOFFER*

A solution to multivariate state-space modeling, forecasting, and smoothing is discussed. We allow for the possibilities of nonnormal errors and nonlinear functionals in the state equation, the observational equation, or both. An adaptive Monte Carlo integration technique known as the Gibbs sampler is proposed as a mechanism for implementing a conceptually and computationally simple solution in such a framework. The methodology is a general strategy for obtaining marginal posterior densities of coefficients in the model or of any of the unknown elements of the state space. Missing data problems (including the k -step ahead prediction problem) also are easily incorporated into this framework. We illustrate the broad applicability of our approach with two examples: a problem involving nonnormal error distributions in a linear model setting and a one-step ahead prediction problem in a situation where both the state and observational equations are nonlinear and involve unknown parameters.

KEY WORDS: Forecasting; Gibbs sampler; Kalman filter; Smoothing.

The state-space model has become a powerful tool for modeling and forecasting dynamic systems. Such models, in conjunction with the Kalman filter, have been used in a wide range of applications in many disciplines including biology, economics, and engineering and consequently have become of increasing interest to statisticians. It is well known, however, that the Kalman filter is optimal only in the case where the dynamic system is Gaussian. If the system is not Gaussian, the Kalman filter yields the best linear predictor (Brockwell and Davis 1987, sec. 12.1), but the difference between the optimal forecast and the best linear predictor can be quite substantial. Moreover, it also is well known that the Kalman filter under Gaussian assumptions is nonrobust (Meinhold and Singpurwalla 1989). Many authors have suggested modeling dynamic systems with state-space models in conjunction with various alternatives to the Kalman filter. For example, Kitagawa (1987) proposed recursive formulas based on piecewise linear approximations to the density functions for prediction, filtering, and state estimation of nonstationary time series via non-Gaussian state-space models. Meinhold and Singpurwalla (1989) suggested a robustification of the state-space model using approximate methods involving poly- t distributions and a recursive mechanism for implementing a multivariate t distribution based on the Kalman filter recursions. Other approaches and approximation techniques in a Bayesian framework were presented by Alspach and Sorenson (1972), Harrison and Stevens (1976), Smith and West (1983), West, Harrison, and Migon (1985), West (1986), and, most recently, Gordon and Smith (1990).

Consider the standard state-space model

$$\begin{aligned}x_t &= F_t x_{t-1} + u_t, \quad \text{and} \\y_t &= H_t x_t + v_t, \quad t = 1, \dots, n, \quad (1)\end{aligned}$$

where x_t is the $p \times 1$ state vector, y_t is the $q \times 1$ observation vector, F_t is a $p \times p$ matrix of constants, and H_t is a $q \times p$ matrix of constants. Let $\mathbf{y} = (y_1, \dots, y_n)$ denote the observed data, $\mathbf{x} = (x_1, \dots, x_n)$ the (unknown) elements of the state, and x_0 the initial state, where we assume $x_0 \sim N_p(\mu_0, \Sigma_0)$. Typically u_t and v_t are independent and identically distributed, with $u_t \sim N_p(0, \Sigma)$ and $v_t \sim N_q(0, \Upsilon)$, where N_p denotes the p -dimensional normal distribution. Also, the matrices F_t , H_t , Σ , and Υ generally are assumed to be known. These assumptions enable simple updating of estimates via the usual Kalman filter but in practice are frequently found to be too restrictive for realistic data analysis.

The first purpose of this article then is to develop methodology for modeling the nonnormality of the u_t , the v_t , or both. A second departure from the model specification (1) is to allow for unknown variances in the state or observational equation, as well as for unknown parameters in the transition matrices F_t and H_t . As a third generalization we allow for nonlinear model structures; that is,

$$\begin{aligned}x_t &= f_t(x_{t-1}) + u_t, \quad \text{and} \\y_t &= h_t(x_t) + v_t, \quad t = 1, \dots, n, \quad (2)\end{aligned}$$

where $f_t(\cdot)$ and $h_t(\cdot)$ are given, but perhaps also depend on some unknown parameters. The experimenter may wish to entertain a variety of error distributions. Our goal throughout the article is an analysis for general state-space models that does not resort to convenient assumptions at the expense of model adequacy.

Section 1 discusses the model specification, and provides a methodology for modeling nonnormality in the form of normal scale mixtures (Andrews and Mallows 1974). The methodology is developed with the implementation of the Gibbs sampler in mind. The latter is an adaptive Markovian updating scheme useful for obtaining marginal posterior distributions in cases where exact numerical results are unavailable and traditional numerical integration techniques are difficult or infeasible. Section 2 considers the problem

* Bradley P. Carlin is Assistant Professor, Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455. Nicholas G. Polson is Assistant Professor, Graduate School of Business, University of Chicago, Chicago, IL 60637. David S. Stoffer is Associate Professor, Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA 15260. The work of Carlin was supported in part by National Science Foundation Grant DMS 88-05676; the work of Stoffer was supported in part by National Science Foundation Grant DMS 90-00522 and by a grant from the Centers for Disease Control through a cooperative agreement with the Association of Schools of Public Health. This research was done while all three authors were visiting the Department of Statistics at Carnegie Mellon University.

of determining estimates of the marginal densities of the model parameters, as well as $p(x_{n+1} | \mathbf{y}, y_{n+1})$ (filtering) and $p(x_{n+1} | \mathbf{y})$ (one-step ahead prediction). In Section 3 we consider two examples of state-space models that allow for nonnormal error structure and nonstationarity in the state and observation spaces. We summarize our findings in Section 4.

1. MODEL SPECIFICATIONS AND THE GIBBS SAMPLER

1.1 Model Specification

In general the likelihood specification for our model, suppressing the conditioning on $(\mu_0, \Sigma_0, F_t, H_t)$, is given by

$$p(y_1, \dots, y_n, x_0, x_1, \dots, x_n | \Sigma, \Upsilon) = g_1(x_0 | \mu_0, \Sigma_0) \prod_{t=1}^n g_1(x_t | x_{t-1}, \Sigma) \prod_{t=1}^n g_2(y_t | x_t, \Upsilon) \quad (3)$$

for some densities $g_1(\cdot)$ and $g_2(\cdot)$. Specifically, we model g_1 and g_2 by letting

$$g_1(x_t | x_{t-1}, \Sigma) = \int_{\Lambda} p(x_t | x_{t-1}, \lambda_t, \Sigma) p_1(\lambda_t) d\lambda_t, \\ g_2(y_t | x_t, \Upsilon) = \int_{\Omega} p(y_t | x_t, \omega_t, \Upsilon) p_2(\omega_t) d\omega_t, \\ t = 1, \dots, n, \quad (4)$$

where, conditional on the nuisance parameters λ and ω ,

$$x_t | x_{t-1}, \lambda_t, \Sigma \sim N(f_t(x_{t-1}), \lambda_t \Sigma), \\ y_t | x_t, \omega_t, \Upsilon \sim N(h_t(x_t), \omega_t \Upsilon), \quad t = 1, \dots, n. \quad (5)$$

Of course if $h_t(x_t) = H_t x_t$ and $f_t(x_{t-1}) = F_t x_{t-1}$, we have the linear model (1). Note that by varying $p_1(\lambda_t)$ and $p_2(\omega_t)$, the distributions g_1 and g_2 are scale mixtures of multivariate normals for each t , thus enabling a wide variety of nonnormal error densities to emerge in (3). For example, in the univariate case (where we denote Σ and Υ by σ and τ) the distributions $x_t | x_{t-1}$, σ and $y_t | x_t$, τ can be double exponential, logistic, exponential power, stable, or t densities (Andrews and Mallows 1974; Carlin and Polson 1991; Kanter 1975; West 1987). In the multivariate case a rich class of densities emerges including the r -dimensional hyperbolic distribution (Barndorff-Neilsen and Halgreen 1977). Note that we are assuming $p(\lambda, \omega) = \prod_{t=1}^n p_1(\lambda_t) p_2(\omega_t)$, so that the densities $x_t | x_{t-1}, \Sigma$ and $y_t | x_t, \Upsilon$ possibly are different scale mixtures of normals. Another easily incorporated extension is to allow for different densities as t varies, $t = 1, \dots, n$.

The key to the approach is the introduction of the (generally high dimensional) nuisance parameters λ and ω and the structure (5), which lends itself naturally to the Gibbs sampler, our computational tool.

1.2 Implementation of the Gibbs Sampler

A Monte Carlo integration method that proceeds by a Markovian updating scheme, the Gibbs sampler is essentially a modification of the Metropolis algorithm (Metropolis et al. 1953), developed formally by Geman and Geman (1984)

in the context of image restoration. In the statistical framework, Tanner and Wong (1987) used essentially this algorithm in their substitution sampling approach. Recently, Gelfand and Smith (1990) developed the Gibbs sampler for fairly general parametric settings; see that paper for a discussion of the method and its properties. To summarize the method briefly, suppose we have a collection of k (possibly vector-valued) random variables U_1, \dots, U_k with complete conditional distributions, denoted generically by $f(U_s | U_r, r \neq s)$, $s = 1, \dots, k$, available for sampling. Here, *available* means that samples may be generated by some method, given values of the appropriate conditioning random variables. Under mild conditions (Besag 1974), these complete conditional distributions uniquely determine the full joint distribution, $f(U_1, \dots, U_k)$, and hence all marginal distributions $f(U_s)$, $s = 1, \dots, k$. The Gibbs sampler generates samples from the joint distribution as follows: Given an arbitrary starting set of values $U_{1(0)}, \dots, U_{k(0)}$, we draw $U_{1(1)}$ from $f(U_1 | U_{2(0)}, \dots, U_{k(0)})$, then $U_{2(1)}$ from $f(U_2 | U_{1(1)}, U_{3(0)}, \dots, U_{k(0)})$, and so on up to $U_{k(1)}$ from $f(U_k | U_{1(1)}, \dots, U_{k-1(1)})$ to complete one iteration of the scheme. After l such iterations we obtain $(U_{1(l)}, \dots, U_{k(l)})$. Geman and Geman (1984) showed that under mild conditions this k -tuple converges in distribution to a random observation from $f(U_1, \dots, U_k)$ as $l \rightarrow \infty$. For this reason, in the sequel we suppress the (l) subscript, assuming that l is sufficiently large for the generated sample to be thought of as a realization from the joint distribution. Now replicating the entire process in parallel G times provides iid k -tuples $(U_1^{(g)}, \dots, U_k^{(g)})$, $g = 1, \dots, G$ from the joint distribution. These observations then can be used for estimating any of the marginal densities. In particular if $f(U_s | U_r, r \neq s)$ is available in closed form, then

$$\hat{f}(U_s) = \frac{1}{G} \sum_{g=1}^G f(U_s | U_r^{(g)}, r \neq s). \quad (6)$$

Due to the relatively recent appearance of Gibbs sampling methodology in the statistical literature, several important theoretical and practical issues in its general implementation remain under investigation. These issues include the diagnosis of convergence, modification of the sampling order (including random visitation orders), efficient estimation and generation from nonstandardized complete conditional densities, and the comparison of results obtained from sampling schemes that are sequential (where we employ only one stream of Gibbs iterates, perhaps keeping every m th one to better simulate a stream of independent samples) as opposed to parallel (as described previously). The articles by Gelfand et al. (1990) and Zeger and Karim (1991) offered useful guidance concerning many of these issues.

In the context of our state-space models, to implement the Gibbs sampler we require samples from the following complete conditional distributions:

- $x_t | x_{j \neq t}, \lambda, \omega, \Sigma, \Upsilon, \mathbf{y}, t = 0, \dots, n$
- $\omega_t | \omega_{j \neq t}, \lambda, \Sigma, \Upsilon, \mathbf{y}, \mathbf{x}, x_0 \sim \omega_t | \Upsilon, y_t, x_t, t = 1, \dots, n$
- $\lambda_t | \lambda_{j \neq t}, \omega, \Sigma, \Upsilon, \mathbf{y}, \mathbf{x}, x_0 \sim \lambda_t | \Sigma, x_t, x_{t-1}, t = 1, \dots, n$
- $\Sigma | \lambda, \omega, \Upsilon, \mathbf{y}, \mathbf{x}, x_0 \sim \Sigma | \lambda, \mathbf{y}, \mathbf{x}, x_0$
- $\Upsilon | \lambda, \omega, \Sigma, \mathbf{y}, \mathbf{x}, x_0 \sim \Upsilon | \omega, \mathbf{y}, \mathbf{x}$

We now consider the first two of these distributions. The third follows in a similar manner to the second; the last two, under conjugate priors, follow from standard normal and Wishart distribution theory due to the conditioning on λ and ω .

First, in the linear case, we prove a lemma that determines the set of conditionals, $x_t|x_{j \neq t}, \lambda, \omega, \Sigma, \Upsilon, \mathbf{y}, t = 1, \dots, n$. The nonlinear case (2) will be illustrated in Example 1.2.

Lemma. Under model (1) and using the multivariate normal scale mixture error assumption (5), the complete conditional distribution $x_t|x_{j \neq t}, \lambda, \omega, \Sigma, \Upsilon, \mathbf{y}$ is $N_p(B_t b_t, B_t)$, where

$$B_t^{-1} = \frac{\Sigma^{-1}}{\lambda_t} + \frac{H_t^T \Upsilon^{-1} H_t}{\omega_t} + \frac{F_{t+1}^T \Sigma^{-1} F_{t+1}}{\lambda_{t+1}}$$

$$b_t^T = \frac{x_{t-1}^T F_t^T \Sigma^{-1}}{\lambda_t} + \frac{y_t^T \Upsilon^{-1} H_t}{\omega_t} + \frac{x_{t+1}^T \Sigma^{-1} F_{t+1}}{\lambda_{t+1}}, \quad (7)$$

the T superscript denoting the transpose operation.

Proof. By Bayes's theorem, the required exponent is a sum of three terms; that is, modulo a normalizing constant, $-2 \log p(x_t|x_{j \neq t}, \lambda, \omega, \Sigma, \Upsilon, \mathbf{y})$ is

$$\frac{1}{\lambda_t} (x_t - F_t x_{t-1})^T \Sigma^{-1} (x_t - F_t x_{t-1})$$

$$+ \frac{1}{\omega_t} (y_t - H_t x_t)^T \Upsilon^{-1} (y_t - H_t x_t)$$

$$+ \frac{1}{\lambda_{t+1}} (x_{t+1} - F_{t+1} x_t)^T \Sigma^{-1} (x_{t+1} - F_{t+1} x_t),$$

which on manipulation gives the desired result.

Note that adjustments will need to be made to formula (7) for the cases $t = 0$ and $t = n$ due to slight modifications and deletions in the likelihood for these "endpoint" cases. We illustrate these modifications in Example 1.1.

Now consider the determination of $\omega_t|x_{j \neq t}, \lambda, \Sigma, \Upsilon, \mathbf{y}, \mathbf{x} \sim \omega_t|\Upsilon, y_t, x_t, t = 1, \dots, n$. By Bayes's theorem, $\omega_t|\Upsilon, y_t, x_t \propto p(y_t|x_t, \omega_t, \Upsilon) p_2(\omega_t)$. But by (4), the normalization constant is known and is given by $g_2(y_t|x_t, \Upsilon)$. Hence the complete conditional for ω_t is of known functional form. Generation of the required samples may be done directly if this form is a standard density; otherwise, a carefully selected rejection method may be used.

Example 1.1: Univariate Linear Model. For illustration, consider model (1) with $p = q = 1$ and $\Sigma_0 = \sigma_0^2$, $\Sigma = \sigma^2$, $\Upsilon = \tau^2$, $H_t = H$ and $F_t = F$. Using the previous lemma and taking care with the endpoint cases we have $x_t|x_{j \neq t}, \lambda, \omega, \sigma, \tau, \mathbf{y} \sim N(B_t b_t, B_t)$, where

$$B_t^{-1} = \frac{1}{\sigma_0^2} + \frac{F^2}{\sigma^2 \lambda_1}, \quad t = 0$$

$$= \frac{1}{\sigma^2} \left(\frac{1}{\lambda_t} + \frac{F^2}{\lambda_{t+1}} \right) + \frac{H^2}{\tau^2 \omega_t}, \quad t = 1, \dots, n-1$$

$$= \frac{1}{\sigma^2 \lambda_n} + \frac{H^2}{\tau^2 \omega_n}, \quad t = n,$$

and

$$b_t = \frac{\mu_0}{\sigma_0^2} + \frac{F x_1}{\sigma^2 \lambda_1}, \quad t = 0$$

$$= \frac{F}{\sigma^2} \left(\frac{x_{t-1}}{\lambda_t} + \frac{x_{t+1}}{\lambda_{t+1}} \right) + \frac{H y_t}{\tau^2 \omega_t}, \quad t = 1, \dots, n-1$$

$$= \frac{F x_{n-1}}{\sigma^2 \lambda_n} + \frac{H y_n}{\tau^2 \omega_n}, \quad t = n.$$

The complete conditionals for σ^2 and τ^2 are obtained as follows. Assuming the independent *a priori* specifications $\sigma^2 \sim \text{IG}(a_0, b_0)$ and $\tau^2 \sim \text{IG}(c_0, d_0)$, where IG denotes the inverse (reciprocal) gamma distribution, then

$$\sigma^2 | \lambda, \mathbf{y}, \mathbf{x}, x_0$$

$$\sim \text{IG} \left(a_0 + \frac{n}{2}, \left\{ \frac{1}{b_0} + \frac{1}{2} \sum_{t=1}^n (x_t - F x_{t-1})^2 / \lambda_t \right\}^{-1} \right)$$

$\tau^2 | \omega, \mathbf{y}, \mathbf{x}$

$$\sim \text{IG} \left(c_0 + \frac{n}{2}, \left\{ \frac{1}{d_0} + \frac{1}{2} \sum_{t=1}^n (y_t - H x_t)^2 / \omega_t \right\}^{-1} \right). \quad (8)$$

For the ω complete conditionals, suppose we wish to model $\mathbf{y}|\mathbf{x}, \tau$ as a product of double exponentials. The necessary *a priori* specification for ω_t is then $\omega_t \sim \exp(2)$, the exponential distribution having mean 2. Because $y_t|x_t, \omega_t, \tau \sim N(H x_t, \omega_t \tau^2)$, the complete conditional for ω_t is then

$$\omega_t | \tau, \mathbf{y}, \mathbf{x} \propto \omega_t^{-1/2} \exp \left[-\frac{1}{2} \left(\omega_t + \frac{(y_t - H_t x_t)^2}{\omega_t \tau^2} \right) \right]; \quad (9)$$

that is, $\omega_t | \tau, \mathbf{y}, \mathbf{x} \sim \text{GIG}[\frac{1}{2}, 1, (y_t - H_t x_t) / \tau^2]$, where GIG denotes the generalized inverse Gaussian distribution (Devroye 1986, p. 478). To sample from this density, we note that it is the reciprocal of an inverse Gaussian ($|\tau / (y_t - H_t x_t)|, 1$), a density from which we may sample easily.

A similar approach to the one just described could be used to model nonnormality in the state equation via the λ complete conditionals. Finally, if F or H are thought of as unknown parameters (as often is the case in practice), then their complete conditional distributions also will be required to implement the Gibbs sampler (see Example 3.1).

Example 1.2: Nonlinear Model. We now determine the distributions $x_t|x_{j \neq t}, \lambda, \omega, \Sigma, \Upsilon, \mathbf{y}$ for model (2), the nonlinearity presenting no further complications in the remaining complete conditional distributions. We consider separately the three cases where nonlinearity occurs in the state equation, the observation equation, or both.

First, suppose that $h_t(x_t) = H_t x_t$ but that the state equation is nonlinear. Then $x_t|x_{j \neq t}, \lambda, \omega, \Sigma, \Upsilon, \mathbf{y} \propto w_1(x_t) N_p(B_t b_t, B_t)$, where

$$B_{1t}^{-1} = \frac{\Sigma^{-1}}{\lambda_t} + \frac{H_t^T \Upsilon^{-1} H_t}{\omega_t},$$

$$b_{1t}^T = \frac{f_t(x_{t-1})^T \Sigma^{-1}}{\lambda_t} + \frac{y_t^T \Upsilon^{-1} H_t}{\omega_t}, \quad (10)$$

and $w_1(x_t) = \exp\{-(1/2\lambda_{t+1})[x_{t+1} - f_t(x_t)]^T \Sigma^{-1}[x_{t+1} - f_t(x_t)]\}$. But clearly $0 \leq w_1(x_t) \leq 1$ for all x_t , and so the distribution from which we want to sample is dominated by the $N(B_{1t}b_{1t}, B_{1t})$ density. Hence we may use rejection sampling (Devroye 1986, sec. II.3) to obtain a random observation from the required complete conditional. That is, we sample an observation x_t from a $N(B_{1t}b_{1t}, B_{1t})$ density and subsequently accept it with probability $w_1(x_t)$.

Of course this algorithm may be rather inefficient if the $w_1(x_t)$ are close to 0; in such cases more sophisticated envelope functions may be needed. Such envelope functions often are normal or t densities chosen to be as similar as possible to the desired complete conditional, thus enabling more efficient rejection sampling (see Carlin and Polson 1991 for an example). The experimenter needs to take care that such an envelope function does in fact blanket the complete conditional distribution for all x_t . Gilks and Wild (1992) overcame this problem for log-concave densities using a piecewise exponential envelope. Along similar lines is the technique presented in Wakefield, Gelfand, and Smith (1991), which improved on the traditional ratio-of-uniforms method of rejection sampling. Still, the substantial computational overhead involved makes all of these approaches less attractive unless the naive method described in the previous paragraph is prohibitively slow.

Second, suppose that $f_t(x_{t-1}) = F_t x_{t-1}$ but that now the observational equation is nonlinear. The $x_t | x_{j \neq t}, \lambda, \omega, \Sigma, \Upsilon, \mathbf{y} \propto w_2(x_t) N_p(B_{2t}b_{2t}, B_{2t})$, where

$$B_{2t}^{-1} = \frac{\Sigma^{-1}}{\lambda_t} + \frac{F_{t+1}^T \Sigma^{-1} F_{t+1}}{\lambda_{t+1}},$$

$$b_{2t}^T = \frac{x_{t-1}^T F_t^T \Sigma^{-1}}{\lambda_t} + \frac{x_{t+1}^T \Sigma^{-1} F_{t+1}}{\lambda_{t+1}},$$

and $w_2(x_t) = \exp\{-(1/2\omega_t)[y_t - h_t(x_t)]^T \Upsilon^{-1}[y_t - h_t(x_t)]\}$, and again rejection may be employed. Finally, when both components are nonlinear $x_t | x_{j \neq t}, \lambda, \omega, \Sigma, \Upsilon, \mathbf{y} \propto w_1(x_t) w_2(x_t) N_p(f_t(x_{t-1}), \lambda_t \Sigma)$. Thus we sample a $N_p(f_t(x_{t-1}), \lambda_t \Sigma)$ random variable and accept it with probability $w_1(x_t) w_2(x_t)$.

2. ESTIMATED MARGINAL POSTERIOR DENSITIES

With all the complete conditionals available for sampling, it now remains to show how to estimate the marginal posterior densities of the quantities of interest using the generated Gibbs samples. If we denote this collection by $\{(x_t^{(g)}, \lambda_t^{(g)}, \omega_t^{(g)}, t = 1, \dots, n), x_0^{(g)}, \Sigma^{(g)}, \Upsilon^{(g)}, g = 1, \dots, G\}$, then we may use (6) obtain

$$\hat{p}(x_t | \mathbf{y}) = \frac{1}{G} \sum_{g=1}^G p(x_t | x_{t-1}^{(g)}, x_{t+1}^{(g)}, \lambda_t^{(g)}, \lambda_{t+1}^{(g)}, \omega_t^{(g)}, \Sigma^{(g)}, \Upsilon^{(g)}, y_t). \tag{11}$$

Note that this of course assumes that the x_t complete conditional distribution is available in closed form. If this is not the case (as in Example 1.2 above), an alternative would be to simply compute a kernel density using the $\{x_t^{(g)}\}$ samples themselves. Another approach would be to obtain the G standardizing constants necessary in equation (11) by uni-

variate numerical integration, perhaps a simple trapezoidal rule. While a bit more work, this latter approach generally produces a better density estimate, as it does not discard the functional form used to obtain the $\{x_t^{(g)}\}$ iterates. This approach is illustrated in Example 3.2.

We note that equation (11) could be used to obtain a marginal posterior density estimate for x_{n+1} provided y_{n+1} was available, offering a solution to the so-called filtering problem. If y_{n+1} is not yet available, the problem becomes one of one-step ahead prediction and can be solved by a slight modification of the Gibbs algorithm. In fact the k -step ahead prediction problem can be easily handled as follows: Suppose we desire an estimate of $p(x_{n+k} | \mathbf{y})$ where again $\mathbf{y} = (y_1, \dots, y_n)$ and y_{n+1}, \dots, y_{n+k} have not yet been observed. We simply add $\{x_{n+1}, \dots, x_{n+k}, y_{n+1}, \dots, y_{n+k}, \lambda_{n+1}, \dots, \lambda_{n+k}, \omega_{n+1}, \dots, \omega_{n+k}\}$ to the Gibbs sampler as $4k$ additional unknown parameters. The complete conditional distributions for the new x 's are again obtained using the lemma in Section 1, where now of course the upper endpoint condition pertains to x_{n+k} instead of x_n . Similarly, the complete conditionals for the new λ 's and ω 's arise in a manner analogous to that described in Section 1. Finally, the complete conditional distributions for the new y variables come directly from the model specification, namely

$$y_{n+t} | \{x_i, \lambda_i, \omega_i\}_{i=1}^{n+k}, x_0, \Sigma, \Upsilon, \mathbf{y} \sim y_{n+t} | x_{n+t}, \omega_{n+t}, \Upsilon \\ \sim N(h_{n+t}(x_{n+t}), \omega_{n+t} \Upsilon), \quad t = 1, \dots, k.$$

We now simply run the Gibbs sampler as usual, obtaining for any $i \in \{1, \dots, k\}$ the slightly modified version of (11),

$$\hat{p}(x_{n+i} | \mathbf{y}) = \frac{1}{G} \sum_{g=1}^G p(x_{n+i} | x_{n+i-1}^{(g)}, x_{n+i+1}^{(g)}, \lambda_{n+i}^{(g)}, \omega_{n+i}^{(g)}, \Sigma^{(g)}, \Upsilon^{(g)}, y_{n+i}^{(g)}), \tag{12}$$

the primary difference being the dependence on the generated values $\{y_{n+i}^{(g)}, g = 1, \dots, G\}$ rather than on an observed data value y_{n+i} . Of course as these future y_{n+i} values become available, we simply use these values in lieu of sampled values $y_{n+i}^{(g)}$ and rerun the algorithm—a computationally simple solution to the filtering problem. Example 3.2 illustrates this process.

3. NUMERICAL EXAMPLES

Example 3.1: Univariate Linear Model. Consider again the model presented in Example 1.1. We apply this model to the data displayed in Table 1, which gives estimated total physician expenditures by year as measured by the Social Security Administration. This data set was included in a state-space analysis by Shumway and Stoffer (1982) using a maximum likelihood procedure via the EM algorithm.

We assume that the estimates y_t in the data are unbiased for the true annual physician expenditures x_t ; thus set $H = 1$. Further, a plot of the data and the analysis of Shumway and Stoffer suggest that the simple exponential growth model given by $F_t = F$ is not unreasonable; however, we wish to treat F as an unknown parameter. This can be easily incorporated into the Gibbs framework developed in Section 1 by assuming that $F \sim N(\mu_F, \sigma_F^2)$ and noting that the com-

Table 1. Estimated Physician Expenditures (Millions of Dollars)

Year (t)	y_t	Year (t)	y_t	Year (t)	y_t	Year (t)	y_t	Year (t)	y_t
1949	2,633	1954	3,574	1959	5,481	1964	8,065	1969	12,629
1950	2,747	1955	3,689	1960	5,684	1965	8,745	1970	14,306
1951	2,868	1956	4,067	1961	5,895	1966	9,156	1971	15,835
1952	3,042	1957	4,419	1962	6,498	1967	10,287	1972	16,916
1953	3,278	1958	4,910	1963	6,891	1968	11,099	1973	18,200

plete conditional is given by $F|\lambda, \omega, \sigma, \tau, \mathbf{x}, \mathbf{y} \sim N(B_F b_F, B_F)$, where

$$B_F^{-1} = \frac{1}{\sigma^2} \sum_{t=1}^n \frac{x_{t-1}^2}{\lambda_t} + \frac{1}{\sigma_F^2} \quad \text{and} \quad b_F = \frac{1}{\sigma^2} \sum_{t=1}^n \frac{x_t x_{t-1}}{\lambda_t} + \frac{\mu_F}{\sigma_F^2}.$$

As in Section 1, we place independent inverse gamma priors with parameters (a_0, b_0) and (c_0, d_0) on σ^2 and τ^2 , so that their complete conditional distributions are again given by (8). Similarly, the complete conditionals for $x_t, t = 0, \dots, n$ are the same as those in Example 1.1 (recall that we use a $N(\mu_0, \sigma_0^2)$ prior on x_0). The densities $p(x_t|\mathbf{y})$ may be estimated using equation (11) with the argument $F^{(t)}$ added to the list of conditioning arguments, because F is no longer known but instead is a component of the sampler.

To demonstrate the approach to nonnormal error distributions, consider the two models \mathcal{M}_1 and \mathcal{M}_2 given by

$$\mathcal{M}_1: u_t \sim N(0, \sigma^2), v_t \sim N(0, \tau^2), \quad \text{and}$$

$$\mathcal{M}_2: u_t \sim \text{DE}(0, \sigma), v_t \sim \text{DE}(0, \tau).$$

For \mathcal{M}_1 we take $\lambda_t = \omega_t = 1$ with probability 1 for all $t = 1, \dots, n$, leading to complete conditional distributions for λ_t and ω_t that also are degenerate at the value 1. For \mathcal{M}_2 we take both the λ_t and ω_t to be independently distributed *a priori* as $\text{Exp}(2)$ random variables, leading to the complete conditionals

$$\lambda_t \sim \text{GIG}\left[\frac{1}{2}, 1, \left(\frac{x_t - Fx_{t-1}}{\sigma}\right)^2\right], \quad \text{and}$$

$$\omega_t \sim \text{GIG}\left[\frac{1}{2}, 1, \left(\frac{y_t - x_t}{\tau}\right)^2\right],$$

in a manner similar to that surrounding equation (9). We complete the specification of the prior on x_0 by setting $\mu_0 = 2,500$ and $\sigma_0 = 100$ and place vague priors on σ^2 and τ^2 , both having prior mean and prior standard deviation equal to 100,000 (i.e., $a_0 = c_0 = 3, b_0 = d_0 = 5 \times 10^{-6}$). Finally, we set $\mu_F = 1.1$ and $\sigma_F = .1$, implying a rough *a priori* belief in a 10% annual growth rate.

For our analysis we ran the Gibbs sampler for $l = 50$ iterations on each model separately, obtaining the two model-specific density estimates $\hat{p}(F|\mathbf{y}, \mathcal{M}_i)$ shown in Figure 1. In each case our algorithm used $G = 2,500$ parallel replications per iteration, and convergence was judged both by monitoring sample moments of the Gibbs values themselves and by plotting successive density estimates for the inflation constant F . We see that the normal errors assumption produces a posterior distribution for F that is slightly less variable and centered around a slightly higher mean inflation rate. In ei-

ther case a point estimate of just over 9% for F is suggested. A fully Bayesian approach would involve obtaining estimates of the posterior probabilities $p(\mathcal{M}_i|\mathbf{y}), i = 1, 2$, leading to a Bayes factor between the two models; a Gibbs sampling approach useful in choosing among competing error distributions was discussed by Carlin and Polson (1991). Overall, the preliminary results obtained here indicate that the assumption of normal errors is not a grossly misleading one.

As a computational remark we note that although we have used a rather large value of G and included a fairly large number of parameters ($3n + 4 = 79$), the fact that all generation is one-for-one (no rejection algorithms are needed) means that a typical run takes no more than 10 minutes using FORTRAN on a DECStation 3100.

Example 3.2: Univariate Nonstationary Growth Model. The y and x values displayed as solid lines in Figure 2 were generated according to the model

$$\begin{aligned} x_t &= \alpha x_{t-1} + \beta x_{t-1} / (1 + x_{t-1}^2) \\ &\quad + \gamma \cos(1.2(t - 1)) + u_t \\ y_t &= x_t^2 / 20 + v_t, \quad t = 1, \dots, 100, \end{aligned} \quad (13)$$

where $x_0 \equiv 0$, the u_t are independent random variables having a t -distribution with $\nu = 10$ degrees of freedom, mean 0, and variance 10, and the v_t are distributed as $N(0, 1)$ random variables independent of the $u_t, t = 1, \dots, 100$. In the rejoinder to his paper, Kitagawa (1987) fit a non-Gaussian filter and smoother to data generated from this model, where the u_t and v_t were both Gaussian white noise sequences with these same means and variances and the values $\alpha = .5, \beta = 25$, and $\gamma = 8$ assumed known. We shall use these values for α, β , and γ in our study but will assume they are unknown to the experimenter and obtain marginal posterior densities

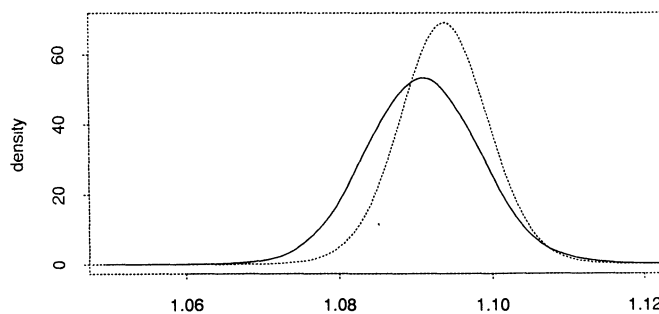


Figure 1. Estimated Marginal Posteriors for F , Example 3.1. Density estimates, for DE errors (solid line) and normal errors (dashed line), $G = 2500$; modes are 1.091 and 1.094, respectively.

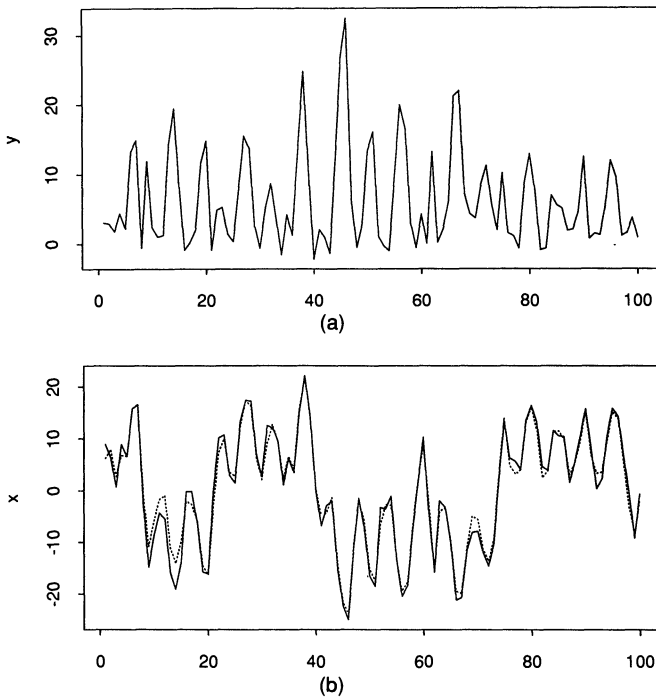


Figure 2. Data and Estimates, Example 3.2: (a) Observed y Values, (b) True x Values (solid line) and Point Estimates (dashed line), y_{101} Unknown.

for all three. In addition, we shall obtain an estimate of $p(x_{101} | \mathbf{y})$, the density of the one-step ahead predicted state.

To implement the Gibbs sampler we follow the model outlined in Example 1.2, where $p = q = 1$. We again assume $\sigma^2 \sim \text{IG}(a_0, b_0)$ and $\tau^2 \sim \text{IG}(c_0, d_0)$, which again leads to inverse gamma complete conditionals of a form similar to that given in equation (8). Next, by letting $\nu/\lambda_t \sim \chi^2_\nu$ we get that marginally, $u_t | \sigma \sim t(0, \sigma, \nu)$ as required, leading to the complete conditional $\lambda_t | \sigma, \alpha, \beta, \gamma, \mathbf{y}, \mathbf{x}, x_0$ being distributed as

$$\text{IG}\left(\frac{\nu + 1}{2}, 2\left\{\frac{[x_t - \alpha x_{t-1} - \beta x_{t-1}/(1 + x_{t-1}^2) - \gamma \cos(1.2(t-1))]^2}{\sigma^2} + \nu\right\}^{-1}\right),$$

$$t = 1, \dots, 101.$$

Because we are assuming the observation noise to be Gaussian, we may take $\omega_t \equiv 1, t = 1, \dots, 101$. Turning to the x_t complete conditionals and again making the prior assumption $x_0 \sim N(\mu_0, \sigma_0^2)$, we note that the nonlinear structure in both the state and observational equations precludes closed-form complete conditionals, but we may use the rejection algorithm discussed in Example 1.2 to generate the necessary samples. That is, we generate x_t from a $N(\alpha x_{t-1} + \beta x_{t-1}/(1 + x_{t-1}^2) + \gamma \cos(1.2(t-1)), \lambda_t \sigma^2)$ distribution and accept it with probability $w_1(x_t)w_2(x_t)$, where

$$w_1(x_t) = \exp\left\{-\frac{1}{2\lambda_{t+1}\sigma^2}(x_{t+1} - \alpha x_t + \beta x_t/(1 + x_t^2) + \gamma \cos(1.2t))^2\right\},$$

and

$$w_2(x_t) = \exp\left\{-\frac{1}{2\omega_t\tau^2}(y_t - x_t^2/20)^2\right\}, \quad t = 1, \dots, 100.$$

For $t = 0$ we generate $x_t \sim N(\mu_0, \sigma_0^2)$ and accept with probability $w_1(x_t)$; for $t = 101$ we generate x_t as usual but accept with probability $w_2(x_t)$. Note that this last complete conditional depends on y_{101} , a latent data value that is not observed but instead is generated according to its complete conditional distribution, which of course is $N(x_{101}^2/20, \omega_t\tau^2)$.

Finally, for the prior on the state equation model parameters we suppose that $(\alpha, \beta, \gamma)^T \sim N_3((\mu_\alpha, \mu_\beta, \mu_\gamma)^T, V)$, where $V = \text{Diag}(\sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2)$. This enables complete conditionals of the form $N(Bb, B)$, where for α

$$B^{-1} = \frac{1}{\sigma_\alpha^2} + \frac{1}{\sigma^2} \sum_{t=1}^{101} \frac{x_{t-1}^2}{\lambda_t} \quad \text{and} \quad b = \frac{\mu_\alpha}{\sigma_\alpha^2} + \frac{1}{\sigma^2} \sum_{t=1}^{101} \frac{x_{t-1}}{\lambda_t} \left(x_t - \beta \frac{x_{t-1}}{1 + x_{t-1}^2} - \gamma \cos(1.2(t-1))\right)$$

and for β

$$B^{-1} = \frac{1}{\sigma_\beta^2} + \frac{1}{\sigma^2} \sum_{t=1}^{101} \frac{x_{t-1}^2}{\lambda_t(1 + x_{t-1}^2)^2} \quad \text{and} \quad b = \frac{\mu_\beta}{\sigma_\beta^2} + \frac{1}{\sigma^2} \sum_{t=1}^{101} \frac{x_{t-1}}{\lambda_t(1 + x_{t-1}^2)} [x_t - \alpha x_{t-1} - \gamma \cos(1.2(t-1))],$$

and finally for γ

$$B^{-1} = \frac{1}{\sigma_\gamma^2} + \frac{1}{\sigma^2} \sum_{t=1}^{101} \frac{\cos^2(1.2(t-1))}{\lambda_t}, \quad \text{and} \quad b = \frac{\mu_\gamma}{\sigma_\gamma^2} + \frac{1}{\sigma^2} \sum_{t=1}^{101} \frac{\cos(1.2(t-1))}{\lambda_t} \left(x_t - \alpha x_{t-1} - \beta \frac{x_{t-1}}{1 + x_{t-1}^2}\right).$$

For this example we took $\mu_0 = 0$ and $\sigma_0^2 = 10, a_0 = 3$ and $b_0 = .05$ (so that the prior on σ^2 has mean and standard deviation equal to 10), and $c_0 = 3$ and $d_0 = .5$ (so that the prior on τ^2 has mean and standard deviation equal to 1). We also chose $\mu_\alpha = .5, \mu_\beta = 25, \mu_\gamma = 8, \sigma_\alpha = .25, \sigma_\beta = 10$, and $\sigma_\gamma = 4$. We then ran the Gibbs sampler for $l = 50$ iterations, using $G = 500$ parallel replications per iteration. The generation cycle in this case involves updating $3(101) + 7 = 310$ parameters per iteration, 102 of which (the x 's) must be sampled via rejection, thus substantially adding to the computational burden; however, programming effort is still quite minimal. Figure 3 shows the resulting marginal posterior density estimates of the form given in (6) for α, β , and γ . Note that this estimation is quite unambiguous, the posteriors being centered nearly at the true parameter values and fairly tightly concentrated. To compute the marginal posterior density of x_{101} , we could use equation (12) with $n = 100$ and $i = 1$; but the nonappearance of y_{101} and x_{102} in the likelihood implies that we may take advantage of the simplified conditional density given in (5), obtaining the estimated density as

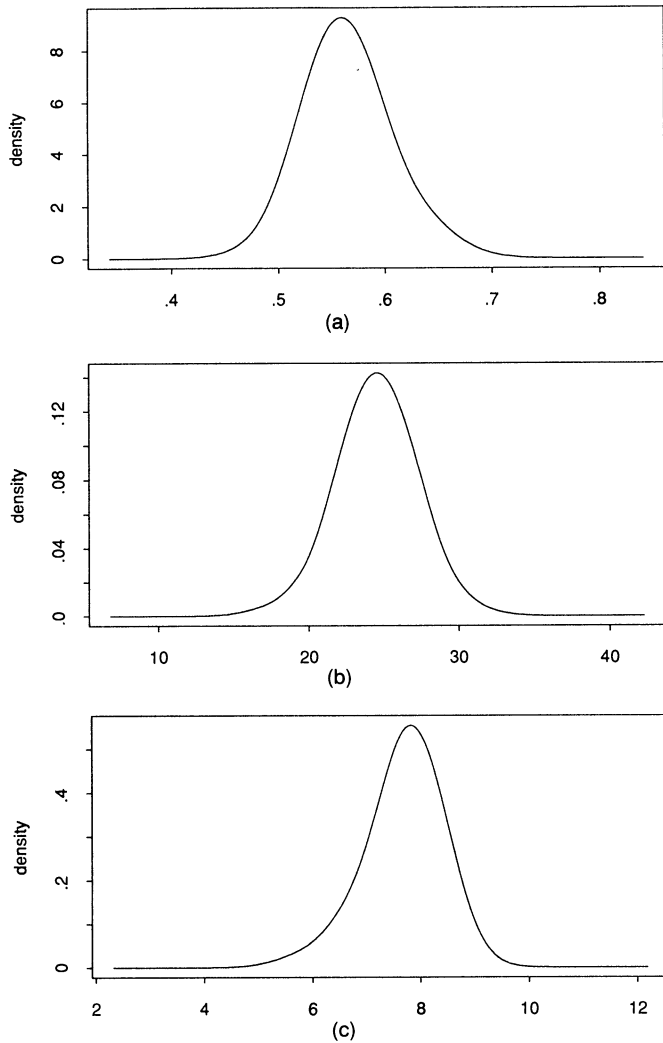


Figure 3. Estimated Marginal Posteriors, Example 3.2: (a) Marginal Posterior for Alpha, $G = 500$; Mode = .56, (b) Marginal Posterior for Beta, $G = 500$; Mode = 24.641, (c) Marginal Posterior for Gamma, $G = 500$; Mode = 7.826.

$$\hat{p}(x_{101} | \mathbf{y}) = \frac{1}{G} \sum_{g=1}^G N(x_{101} | \alpha^{(g)}x_{100}^{(g)} + \beta^{(g)}x_{100}^{(g)} / (1 + (x_{100}^{(g)})^2) + \gamma^{(g)} \cos(120), \lambda_{101} (\sigma^{(g)})^2), \quad (14)$$

where $N(\cdot | a, b)$ denotes a normal density with mean a and variance b . This estimate is given in Figure 4. To check the validity of the bimodal shape of this posterior, we constructed a histogram of the actual Gibbs values $\{x_{101}^{(g)}, g = 1, \dots, G\}$, shown in Figure 4, which also supports a bimodal shape. If we look again at the pattern of the true x values in Figure 2, the reason for the bimodality becomes apparent: the system is currently near the zero point and is likely to drop back down into the negative realm, as it has done most recently. However, there is a substantial probability that the system will now return to the positive realm, explaining the second mode.

To investigate the effect that knowledge of y_{101} would have on the posterior for x_{101} , we repeated the above analysis

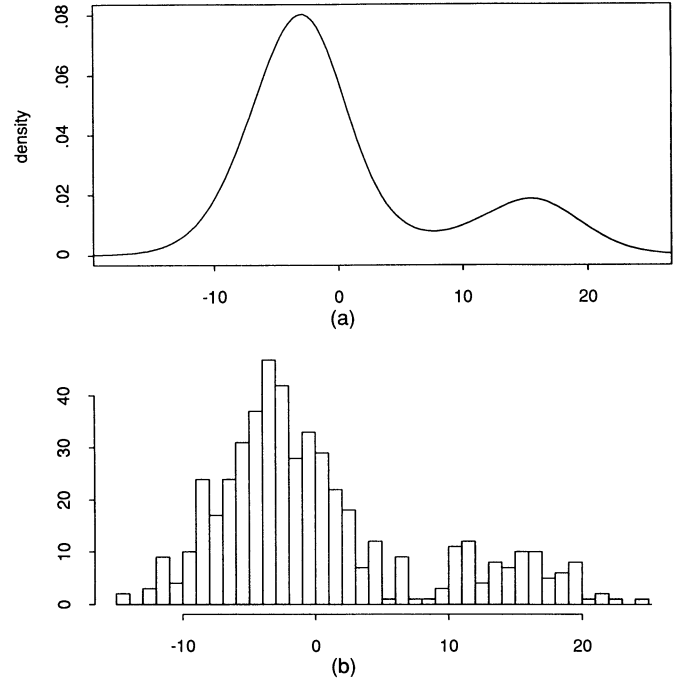


Figure 4. One-Step-Ahead Prediction, Example 3.2: (a) Marginal Posterior for $x_{101}^{(g)}$, $G = 500$; Mode = -2.829 , (b) Histogram of Gibbs $x_{101}^{(g)}$ Values, $G = 500$.

using the observed value $y_{101} = 4.55$. In computing the marginal posterior for x_{101} , we are now solving the filtering problem. The addition of y_{101} to the likelihood means obtaining this marginal posterior by simple mixing, as in equation (14), is no longer available and we must resort to mixing the full posteriors, as in equation (12). The normalization constants

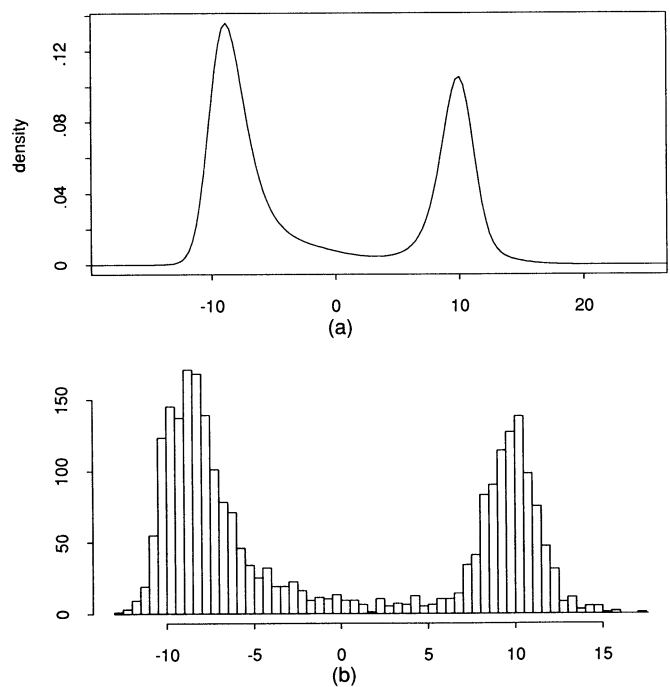


Figure 5. Filtering, Example 3.2: (a) Marginal Posterior for $x_{101}^{(g)}$, $G = 2,500$; Mode = -8.859 , (b) Histogram of Gibbs $x_{101}^{(g)}$ values, $G = 2,500$.

needed for each term of this sum were computed using a trapezoidal approximation. Figure 5 shows the resulting estimated posterior and actual Gibbs samples from running $l = 50$ iterations of $G = 2,500$ replications each (the larger G being required to obtain the same level of accuracy with the more complicated density estimation procedure). We see that the bimodal shape observed in Figure 4 has become more exaggerated, the additional information provided by y_{101} leading to a tighter distribution for both modes. The peaks also have shifted to the left by roughly 5 units; interestingly, the true value $x_{101} = -9.05$ is very close to the location of the first mode ($x = -8.86$). The ability to effectively handle bimodalities is a feature of Monte Carlo integration methods like the Gibbs sampler; analytic approximations such as Laplace's method (Tierney and Kadane 1986) generally are not recommended for use in such situations.

We note that calculations similar to those undertaken for x_{101} also could be performed for all of the remaining x_t states. In particular point estimates and credible sets for each x_t could be computed easily from the resulting estimated marginal posteriors. However, rough point and interval estimates for any parameter θ may be obtained simply by taking appropriate functions or quantiles of the $\{\theta^{(g)}, g = 1, \dots, G\}$ iterates themselves. For example, point estimates of x_t are given by $\sum_{g=1}^G x_t^{(g)}/G$. These estimates are plotted as the dashed lines in Figure 2. They perform surprisingly well and on the whole seem quite competitive with those obtained by Kitagawa (1987, p. 1062), especially given our assumption of nonnormal errors in the state-space and that $\alpha, \beta, \gamma, \sigma$, and τ were all unknown.

4. CONCLUSION

In this article we have discussed some computational aspects of state-space modeling, forecasting, and smoothing. The Gibbs sampler was shown to provide a convenient mechanism for implementing our methods. The experimenter can account for nonnormality in either the observation or state-space via the notion of scale mixtures of normals. This process adds many more parameters to the model, but the conditioning feature of the Gibbs algorithm and the ready availability of the required distributions causes no increase in programming complexity and offers a much broader class of possible distributional assumptions than was previously available. Further, complicated nonlinear model structures involving unknown parameters also are tractable using this approach.

[Received July 1990. Revised May 1991.]

REFERENCES

- Alspach, D. L., and Sorenson, H. W. (1972), "Nonlinear Bayesian Estimation Using Gaussian Sum Approximations," *IEEE Transactions on Automatic Control*, 17, 439-447.
- Andrews, D. F., and Mallows, C. L. (1974), "Scale Mixtures of Normality," *Journal of the Royal Statistical Society, Ser. B*, 36, 99-102.
- Barndorff-Neilsen, O. E., and Halgreen, C. (1977), "Infinite Divisibility of the Hyperbolic and Generalized Inverse Gaussian Distributions," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 38, 309-311.
- Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 36, 192-236.
- Brockwell, P. J., and Davis, R. A. (1987), *Time Series: Theory and Methods*, New York: Springer-Verlag.
- Carlin, B. P., and Polson, N. G. (1991), "Inference for Nonconjugate Bayesian Models Using the Gibbs Sampler," *Canadian Journal of Statistics*, 19, 399-405.
- Devroye, L. (1986), *Non-Uniform Random Variate Generation*. New York: Springer-Verlag.
- Gelfand, A. E., Hills, S. E., 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.
- Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Journal of the Royal Statistical Society, Ser. C*, 41, 337-348.
- Gordon, K., and Smith, A. F. M. (1990), "Modeling and Monitoring Biomedical Time Series," *Journal of the American Statistical Association*, 85, 328-337.
- Harrison, P. J., and Stevens, C. F. (1976), "Bayesian Forecasting" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 38, 205-247.
- Kanter, M. (1975), "Stable Densities Under Change of Scale and Total Variation Inequalities," *The Annals of Probability*, 3, 697-707.
- Kitagawa, G. (1987), "Non-Gaussian State-Space Modeling of Nonstationary Time Series" (with discussion), *Journal of the American Statistical Association*, 82, 1032-1063.
- Meinhold, R. J., and Singpurwalla, N. D. (1989), "Robustification of Kalman Filter Models," *Journal of the American Statistical Association*, 84, 479-486.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087-1091.
- Shumway, R. H., and Stoffer, D. S. (1982), "An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm," *Journal of Time Series Analysis*, 3, 253-264.
- Smith, A. F. M., and West, M. (1983), "Monitoring Renal Transplants: An Application of the Multiprocess Kalman Filter," *Biometrics*, 39, 867-878.
- Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528-550.
- Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82-86.
- Wakefield, J. C., Gelfand, A. E., and Smith, A. F. M. (1991), "Efficient Computation of Random Variates Via the Ratio-of-Uniforms Method," *Statistics and Computing*, 1, 129-134.
- West, M. (1986), "Bayesian Model Monitoring," *Journal of the Royal Statistical Society, Ser. B*, 48, 70-78.
- (1987), "On Scale Mixtures of Normality," *Biometrika*, 74, 694-697.
- West, M., Harrison, P. J., and Migon, H. S. (1985), "Dynamic Generalised Linear Models and Bayesian Forecasting" (with discussion), *Journal of the American Statistical Association*, 80, 73-97.
- Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models with Random Effects; A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79-86.