Biometrika (1996), **83**, 3, *pp.* 589–601
Printed in Great Britain

Markov chain Monte Carlo in conditionally Gaussian state space models

BY C. K. CARTER AND R. KOHN
Australian Graduate School of Management, University of New South Wales, Kensington, N.S.W., Australia, 2052

SUMMARY

A Bayesian analysis is given for a state space model with errors that are finite mixtures of normals and with coefficients that can assume a finite number of different values. A sequence of indicator variables determines which components the errors belong to and the values of the coefficients. The computation is carried out using Markov chain Monte Carlo, with the indicator variables generated without conditioning on the states. Previous approaches use the Gibbs sampler to generate the indicator variables conditional on the states. In many problems, however, there is a strong dependence between the indicator variables and the states causing the Gibbs sampler to converge unacceptably slowly, or even not to converge at all. The new sampler is implemented in $O(n)$ operations, where n is the sample size, permitting an exact Bayesian analysis of problems that previously had no computationally tractable solution. We show empirically that the new sampler can be much more efficient than previous approaches, and illustrate its applicability to robust nonparametric regression with discontinuities and to a time series change point problem.

Some key words: Change point problem; Kalman filter; Mixture of normals; Nonparametric regression; Outlier; Time series.

1. INTRODUCTION

Linear Gaussian state space models are used extensively, with unknown parameters usually estimated by maximum likelihood: Wecker & Ansley (1983), Harvey (1989, Ch. 3). However, many time series and nonparametric regression applications, such as change point problems, outlier detection and switching regression, require the full generality of the conditionally Gaussian model: Harrison & Stevens (1976), Shumway & Stoffer (1991), West & Harrison (1989, Ch. 12), Gordon & Smith (1990). The presence of a large number of indicator variables makes it difficult to estimate conditionally Gaussian models using maximum likelihood, and a Bayesian approach using Markov chain Monte Carlo appears more tractable. We propose a new sampler, which is used to estimate an unknown function nonparametrically when there are jumps in the function and outliers in the observations; it is also applied to a time series change point problem previously discussed by Gordon & Smith (1990). For the first example the Gibbs sampler works poorly, and for the second it does not work at all.

The conditionally Gaussian model can also be applied to more general error distributions by approximating them by a mixture of normals. For example, Shephard (1994) discusses a changing variance model with observations that have a log chi-squared distribution with one degree of freedom, which he approximates by a mixture of normals.

Section 2 describes the model and the new sampling scheme. Section 3 applies the sampler to robust nonparametric regression with discontinuities and § 4 to a time series change point problem. Section 5 shows how to generate the indicator variables in $O(n)$ operations. The basis for generating the indicator variables from reduced conditionals is a sampling scheme introduced in Appendix 1 which groups variables in a more flexible way than the Gibbs sampler and is useful for general Bayesian analysis. Appendix 2 shows that the new sampler converges to the correct posterior distribution.

2. Model, prior assumptions and sampling schemes

The model is

$$y_i = h_i' x_i + \gamma_i e_i, \quad x_i = F_i x_{i-1} + \Gamma_i u_i; \tag{2·1}$$

the observations y_i are scalar and the state vector x_i *is* $m \times 1$. The errors e_i are independent $N(0, \sigma^2)$ and the errors u_i are independent $N(0, \tau^2 I_{m'})$. The coefficients h_i, γ_i, F_i and Γ_i are determined by the discrete variable K_i. To define m' uniquely, we assume that Γ_i has full column rank for at least one value of K_i.

The following notation is used: $Y := (y_1, \ldots, y_n)'$ is the vector of observations, $X := (x_1', \ldots, x_n')'$ is the total state vector, and $K := (K_1, \ldots, K_n)$. Let $p(K)$ be the distribution of K and $\mathscr{K} := \{K : p(K) > 0\}$ the state space of K. We make the following three assumptions.

Assumption 1. A priori, σ^2, τ^2 and K are independent, with the priors for σ^2 and τ^2 inverse gamma. The prior distribution for K is a Markov chain with known transition probabilities.

Assumption 2. Given K_1 and τ^2, the distribution of x_1 is Gaussian.

In many applications the distribution of x_1 is partially diffuse. For this case, the sampler and algorithm are similar to those given below, except that the modified Kalman filter of Ansley & Kohn (1990) is used to carry out the computation, instead of the Kalman filter.

Assumption 3. The density $p(Y|K, \sigma^2, \tau^2) > 0$, for all $K \in \mathscr{K}$, $\sigma^2 > 0$ and $\tau^2 > 0$.

We propose the following sampler for generating X, K, σ^2 and τ^2. Unlike the Gibbs sampler, described below, the variable K_i is generated without conditioning on X. Appendix 1 explains that this is equivalent to generating X and K_i as a block, without the necessity actually to generate X. Let $g_i := h_i' x_i$ and $G := (g_1, \ldots, g_n)'$.

Sampling Scheme 2·1. Generate from the conditional distributions:

(i) $p(\tau^2 | Y, G, K, \sigma^2)$, which simplifies to $p(\tau^2 | G, K)$;
(ii) $p(K_i | Y, K_j, j \neq i, \sigma^2, \tau^2)$ for $i = 1, \ldots, n$;
(iii) $p(X | Y, K, \sigma^2, \tau^2)$;
(iv) $p(\sigma^2 | Y, X, K, \tau^2)$, which simplifies to $p(\sigma^2 | Y, G, K)$.

Steps (i) and (iv) are carried out as in Carter & Kohn (1994, § 3), who show that $p(\sigma^2 | G, K)$ and $p(\tau^2 | Y, G, K)$ are inverse gamma. The total state vector X is generated as a block as in Carter & Kohn (1994) and Frühwirth-Schnatter (1994); for some models the algorithm in de Jong & Shephard (1995) is more efficient. The variables K_i are generated from reduced conditionals as in § 5.

The next lemma gives necessary and sufficient conditions for Sampling Scheme 2·1 to

converge. To state the convergence conditions we need to define the following sampling scheme on K.

Sampling Scheme K. Generate from $p(K_i | K_j, j \neq i)$, for $i = 1, \ldots, n$.

LEMMA 2·1. *Sampling Scheme* 2·1 *has invariant distribution* $p(X, K, \sigma^2, \tau^2 | Y)$. *It is aperiodic if and only if Sampling Scheme K is aperiodic, and is irreducible if and only if Sampling Scheme K is irreducible.*

Proof. Invariance is shown in Appendix 1, and aperiodicity and irreducibility in Appendix 2. □

The block Gibbs sampler for (2·1) is described by the following sampling scheme.

Sampling Scheme 2·2 (*Gibbs sampler*). Generate from:
(i) $p(X | Y, K, \sigma^2, \tau^2)$;
(ii) $p(K | Y, X, \sigma^2, \tau^2)$;
(iii) $p(\sigma^2, \tau^2, | Y, X, K)$, which simplifies to $p(\sigma^2 | Y, X, K)p(\tau^2 | X, K)$.

Carter & Kohn (1994) propose the following modification of the Gibbs sampler which generates τ^2 more efficiently by conditioning τ^2 on G and K instead of X and K and requires minor extra computation. Because of its extra efficiency, we use it instead of the Gibbs sampler in the empirical comparisons with Sampling Scheme 2·1.

Sampling Scheme 2·3. Generate from the conditional distributions:
(i) $p(\tau^2 | Y, G, K)$, which simplifies to $p(\tau^2 | G, K)$;
(ii) $p(X | Y, K, \sigma^2, \tau^2)$;
(iii) $p(\sigma^2 | Y, X, K, \tau^2)$, which simplifies to $p(\sigma^2 | Y, G, K)$.

The next lemma gives conditions for Sampling Schemes 2·2 and 2·3 to converge. Its proof is similar to that of Lemma 2·1 and is omitted.

LEMMA 2·2. *Sampling Schemes* 2·2 *and* 2·3 *are invariant to the distribution* $p(X, K, \sigma^2, \tau^2 | Y)$. *They are irreducible and aperiodic if and only if* $\gamma_i^2 > 0$ *and* Γ_i *is of full column rank for all values of* K_i, *and all i.*

Remark 2·1. Suppose the K_i are a priori independent. Then they are also independent conditionally on the states, which means that generating the K_i simultaneously from $p(K | Y, X, \sigma^2, \tau^2)$ is equivalent to generating them one at a time using $p(K_i | Y, X, K_{j \neq i}, \sigma^2, \tau^2)$. Thus, if the K_i are independent, the theoretical results in Liu, Kong & Wong (1994) suggest that Sampling Scheme 2·1 is likely to converge faster than Sampling Schemes 2·2 and 2·3, as the K_i are conditioned on less information.

We also note that if the K_i are a priori independent then Sampling Scheme 2·1 is irreducible if Sampling Schemes 2·2 and 2·3 are irreducible.

Remark 2·2. Sampling Scheme 2·1 can be extended in a straightforward way to allow the coefficients h_i, γ_i, F_i and Γ_i to depend on an unknown parameter vector which is also generated.

3. ROBUST NONPARAMETRIC REGRESSION WITH DISCONTINUITIES

A robust nonparametric Bayesian approach is now presented for estimating a regression function, assumed smooth except for a small number of discontinuities in either the func-

tion or its first derivative; the points of discontinuity are allowed to be unknown. Our approach has the following properties:

(i) it provides a good estimate of the smooth part of the regression function;
(ii) it detects the jump points and obtains the posterior probability of a jump at any given point, enabling discrimination between a real and spurious jump at any given point;
(iii) it allows for outliers in the observations so as not to confound outliers with jumps in the regression function.

Non-Bayesian approaches to nonparametric regression with discontinuities, together with motivating examples, are given by McDonald & Owen (1986) and Müller (1992). Their approaches satisfy (i), but not (ii) and (iii).

Consider observations generated by the regression model

$$y_i = f(t_i) + K_{1i}^{1/2} e_i \quad (i = 1, \ldots, n),$$

where $f(.)$ is the unknown regression function. We take, without loss of generality, $0 = t_0 \leqslant t_1 \leqslant t_2 \leqslant \ldots \leqslant t_n$, and let $\delta_i = t_i - t_{i-1}$. The errors e_i are assumed independent $N(0, \sigma^2)$. For ordinary observations $K_{1i} = 1$, whereas for outliers K_{1i} is taken large. Wahba (1978) gives a prior for a smooth regression function; this prior can be expressed in state space form as in Carter & Kohn (1994) with state vector $x_i = (f(t_i), f^{(1)}(t_i))'$. Using Wahba's prior, the posterior mean of the regression function is a cubic smoothing spline. We adapt Wahba's prior to allow for jumps in the function and its first derivative by expressing it in the state space form (2·1) with $h_i = (1, 0)'$, $\Gamma_i \Gamma_i' = K_{2i} U_i$,

$$F_i = \begin{pmatrix} 1 & \delta_i \\ 0 & 1 \end{pmatrix}, \quad U_i = \begin{pmatrix} \delta_i^3/3 & \delta_i^2/2 \\ \delta_i^2/2 & \delta_i \end{pmatrix}.$$

The errors in the state transition equation are $u_i \sim N(0, \tau^2 I_2)$. The variable $K_{2i} = 1$ when there is no jump at t_i, and K_{2i} is large if there is a jump. As in Wahba (1978), a diffuse prior is placed on the initial conditions, that is $x_1 = (f(t_1), f^{(1)}(t_1))' \sim N(0, cI_2)$, with $c \to \infty$.

Let $K_i = (K_{1i}, K_{2i})$ with the prior for K_i given in Table 1. The K_i are assumed to be independent and, to simplify the computation, we impose the restriction that an outlier and a jump cannot occur simultaneously. To complete the Bayesian specification of the model, σ^2 and τ^2 have the improper priors $p(\sigma^2) \propto 1/\sigma^2 \exp(-\beta_\sigma/\sigma^2)$, with $\beta_\sigma = 10^{-10}$, and $p(\tau^2) \propto 1/\tau^2$. It is readily checked that all posterior distributions are proper.

Table 1. *Distribution of* K_i

(j, k)	$(1, 1)$	$(10, 1)$	$(10^2, 1)$	$(1, 10)$	$(1, 10^2)$	$(1, 10^3)$	$(1, 10^4)$	$(1, 10^5)$	$(1, 10^6)$
$\mathrm{pr}\{K_i = (j, k)'\}$	0·95	0·00625	0·00625	0·00625	0·00625	0·00625	0·00625	0·00625	0·00625

Lemmas 2·1 and 2·2 imply that Sampling Schemes 2·1, 2·2 and 2·3 are irreducible and aperiodic and hence converge to the posterior distribution. Sampling Schemes 2·1 and 2·3 are now empirically compared for data generated from the model $y_i = f(t_i) + e_i$, with the e_i independent $N(0, 0{\cdot}15^2)$. The regression function $f(.)$ is piecewise constant with $f(t) = 0$ for $0 \leqslant t \leqslant 0{\cdot}5$ and $f(t) = 1$ for $0{\cdot}5 < t \leqslant 1$. The sample size is $n = 100$ and the t_i are equally spaced. Three large outliers are added to the data, and are displayed in Fig. 1(a) and Fig. 2(a). The results below show that Sampling Scheme 2·1 converges quickly, whereas Sampling Scheme 2·3, and hence also Sampling Scheme 2·2, converges so slowly as to be impractical.

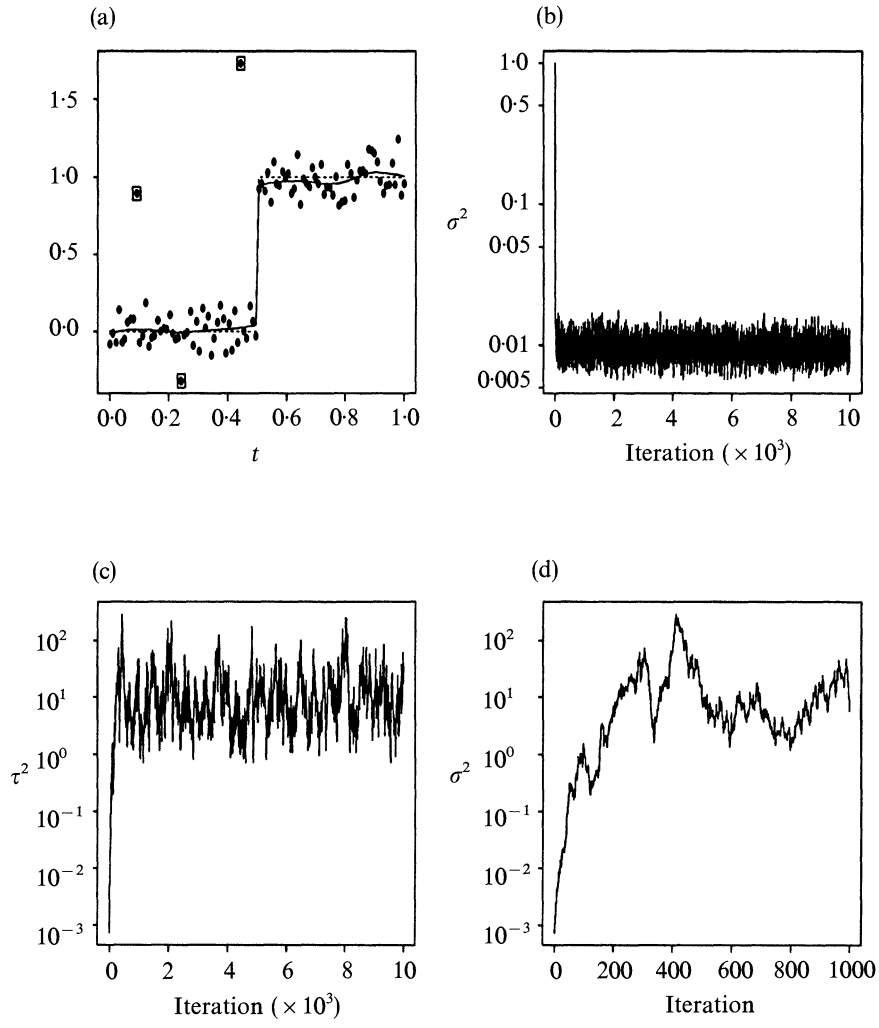

Fig. 1. Sampling Scheme 2·1. (a) shows the signal $f(t)$ with the generated data and the Markov chain Monte Carlo estimate of $f(t)$; outliers are indicated by squares. (b) shows the iterates of σ^2. (c) shows the iterates of τ^2. (d) shows the first 1000 iterates of τ^2.

Sampling Schemes 2·1 and 2·3 were run for a variety of starting values for σ^2, X and K; a starting value is not required for τ^2 as it is the first variable generated. Figure 1 shows the results for a particular run of Sampling Scheme 2·1 using a warm-up period of 5000 iterations followed by a sampling period of 5000 iterations. The starting values are $\sigma^2 = 1$, $K = ((1, 1)' \ldots, (1, 1)')'$ and $x_i = E(x_i | Y, K, \sigma^2, \tau^2 = 1)$ for $i = 1, \ldots, n$. Figure 1 shows the function estimates for the sampling period, and the iterates of σ^2 and τ^2, on a log scale, for both the warm-up and the sampling periods. Figure 2 shows the corresponding results for a particular run of Sampling Scheme 2·3. The warm-up and sampling periods, as well as the starting values, are the same as for Fig. 1.

From Fig. 1, Sampling Scheme 2·1 appears to converge after about 200 iterations. Similar results were obtained for other arbitrary starting values, suggesting that the results shown in Fig. 1 represent the whole posterior distribution and not just a local mode. The function estimates in Figs. 1 and 2 are quite different, showing that Sampling Scheme 2·3 did not

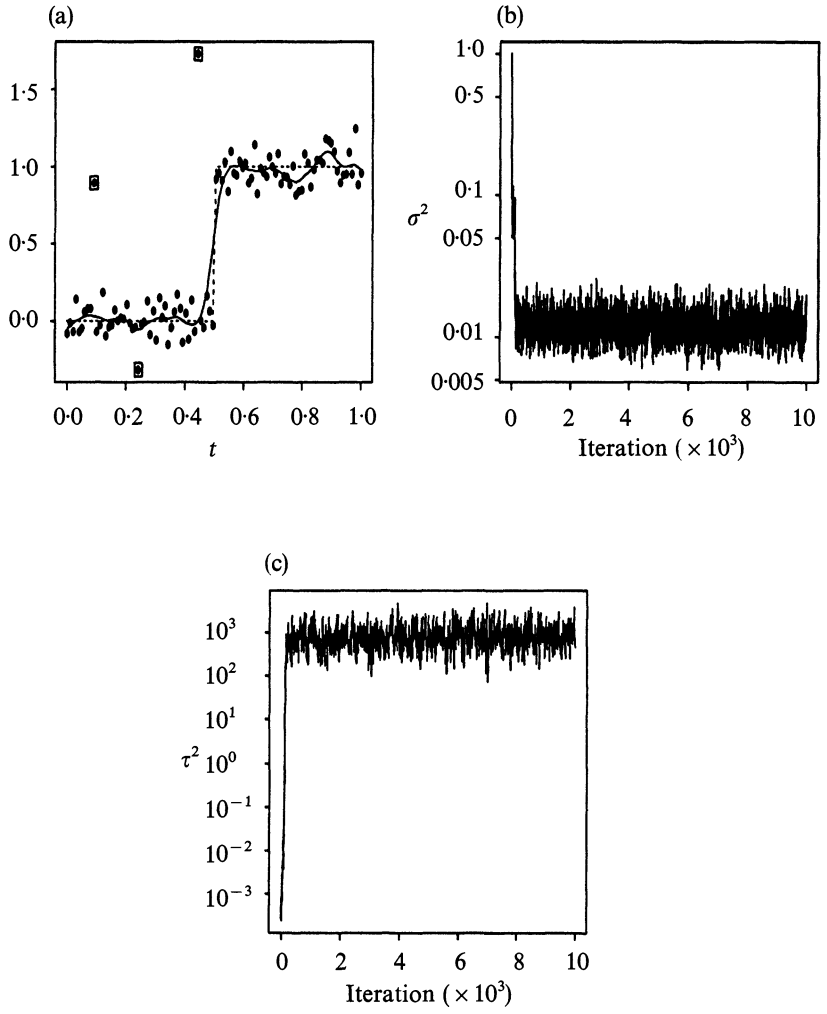

Fig. 2. Sampling Scheme 2·3. (a) shows the signal $f(t)$ with the generated data and the Markov chain Monte Carlo estimate of $f(t)$; outliers are indicated by squares. (b) shows the iterates of σ^2. (c) shows the iterates of τ^2.

converge even after 10 000 iterations. We found that for other arbitrary choices of starting values Sampling Scheme 2·3 did not converge within a reasonable number of iterations.

Remark 3·1. A version of Sampling Scheme 2·3 with t-distributed errors for both the observation and state equations was also run and difficulties with convergence similar to those reported above were again encountered.

4. Change point problems in time series

4·1. *Introduction*

Sampling Scheme 2·1 is now applied to a biomedical change point model discussed by Gordon & Smith (1990). For this model Sampling Schemes 2·2 and 2·3 are reducible. There are many other change point problems to which Sampling Scheme 2·1 applies, including the models discussed by Harrison & Stevens (1976), Shumway & Stoffer (1991),

Hamilton (1989), and Chapters 11 and 12 in West & Harrison (1989). Our results can also be extended to handle outliers and level shifts for autoregressive moving-average models.

4·2. *Piecewise linear signal with change points*

Gordon & Smith (1990) model kidney function in patients who had recently undergone kidney transplant. The level of kidney function is indicated by the rate at which chemical substances are cleared from the blood, and the rate can be inferred indirectly from measurements on weight-adjusted serum creatinine. Gordon & Smith argue on physiological grounds that if kidney function is stable then the response series varies about a constant level; if the kidney function is improving with constant growth then the response series should decay roughly linearly, and the reverse is true if the kidney function decays at a constant rate. About 5% of the observations are subject to error due to mistakes in data transcription, equipment malfunction or blood contamination. The series also experiences jumps in its level due to dialysis treatment. To capture all these effects Gordon & Smith model weight-adjusted reciprocal serum creatinine concentration, y_i, by

$$y_i = \mu_i + K_0^{1/2} e_i, \quad \mu_i = \mu_{i-1} + \beta_i + K_{1i}^{1/2} a_{1i}, \quad \beta_i = \beta_{i-1} + K_{2i}^{1/2} a_{2i}, \tag{4·1}$$

with μ_i the level, and β_i the slope, at time i. A rejection episode is indicated by $\beta_{i-1} \leqslant 0$ and $\beta_i > 0$. The errors e_i, a_{1i} and a_{2i} are all independent $N(0, \sigma^2)$. Gordon & Smith allow the indicator vector (K_{0i}, K_{1i}, K_{2i}) to take the four values (1, 0, 0), (1, 90, 0), (1, 0, 60) and (100, 0, 0), with prior probabilities 0·85, 0·06, 0·07 and 0·02 respectively. The prior for σ^2 is the same as in § 3. The indicator variable K_{0i} takes the values 1 and 100 with the second value representing an observation outlier; the indicator variable K_{1i} takes the values 0 and 90, with the second values representing a jump in the level; the indicator variable K_{2i} takes the values 0 and 60, with the second values representing a jump in the slope. Model (4·1) can be expressed in state space form (2·1) with state vector $x_i = (\mu_i, \beta_i)'$. We assume that x_1 is diffuse and the K_i are independent. The error u_i in the state transition equation is $N(0, \sigma^2 I_2)$ and

$$\Gamma_i = \begin{pmatrix} K_{1i}^{1/2} & K_{2i}^{1/2} \\ 0 & K_{2i}^{1/2} \end{pmatrix}.$$

Sampling Scheme 2·1 is irreducible. However, $K_{2i} = 0$ means that $\beta_i = \beta_{i-1}$, and $\beta_i = \beta_{i-1}$ means that $K_{2i} = 0$ almost surely, so that Sampling Schemes 2·2 and 2·3 do not converge. Sampling Scheme 2·1 is now applied to the data used by Gordon & Smith. Figure 3 shows the results for a particular run with both warm-up and sampling periods of length 1000. The starting values are $\sigma^2 = 1$, $K = ((1, 0, 0), \ldots, (1, 0, 0))'$ and $x_i = E(x_i | Y, K, \sigma^2)$. Figure 3(a) plots the data and the estimate of the posterior mean of the level μ_i. Figure 3(b) plots the iterates of σ^2 showing that the sampler appears to converge after about 200 iterations. Figures 3(c) and 3(d) plot the posterior probabilities of a jump in the level and slope, respectively, and show there is a jump in the slope at about time 10 and at times 110 and 112. Other starting values give similar results, suggesting that the sampler converged to the whole of the posterior distribution and not just a local mode.

5. Generating the indicator variables

5·1. *An expression for* $p(Y | K)$

This section shows how to generate efficiently all the indicator variables in Sampling Scheme 2·1 in $O(n)$ operations. For notational convenience, dependence on σ^2 and τ^2 is

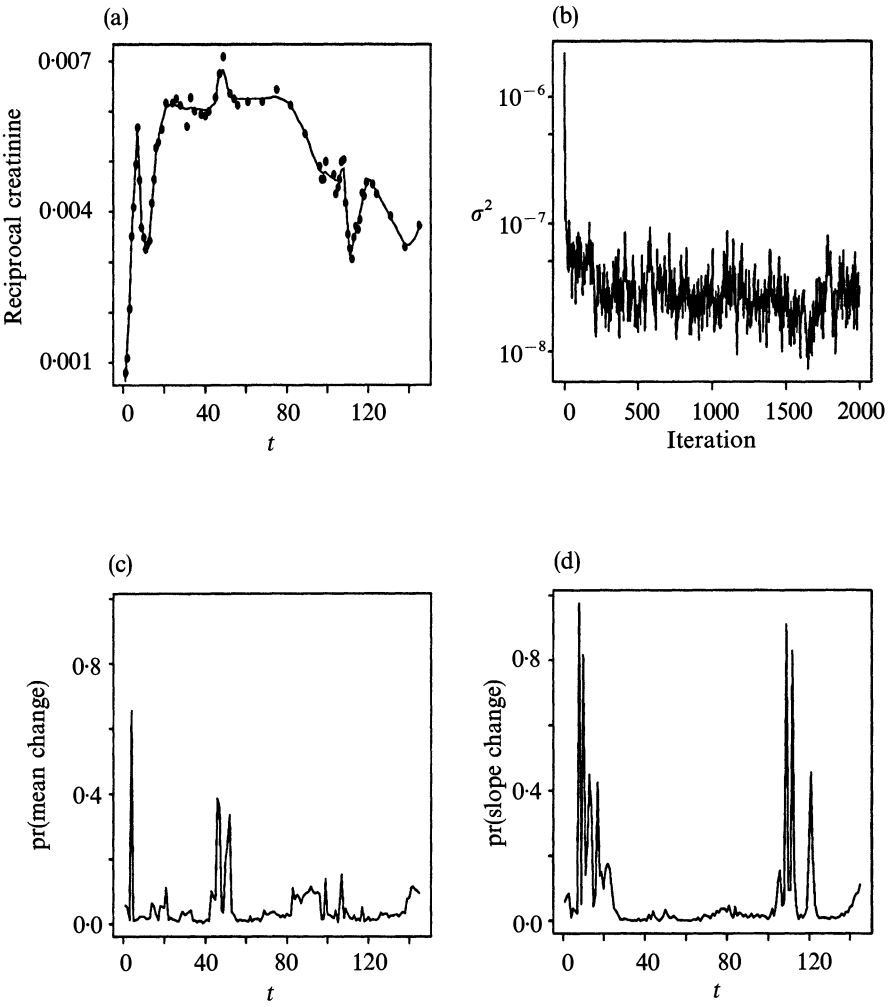

Fig. 3. Biomedical data using Sampling Scheme 2·1. (a) shows reciprocal creatinine level together with the Markov chain Monte Carlo estimate. (b) shows the iterates of σ^2. (c) shows the estimated posterior probabilities of a mean change. (d) shows the estimated posterior probabilities of a slope change.

omitted. A normal random vector is called nonsingular if it has a proper density and superscripts are used to denote subvectors of Y, X and K, for example $Y^{i,j} := (y_i, \ldots, y_j)'$. The distributions of $x_i | Y^{1,i}, K^{1,i}$ and $x_i | Y^{i+1,n}, K$ are assumed nonsingular. The next result expresses $p(K_i | y, K_j, j \neq i)$ for $i = 1, \ldots, n$ in a form suitable for computation.

LEMMA 5·1. *Suppose that for all* $K \in \mathscr{K}$, *the distributions* $Y | K$, $x_i | Y^{1,i}$ *and* $x_i | Y^{i+1,n}, K$ *are nonsingular. Then*

$$p(K_i | Y, K_{j \neq i}) \propto p(Y | K) p(K_i | K_{j \neq i}).$$

For $i = 1, \ldots, n-1$,

$$p(Y | K) \propto p(y_i | Y^{1,i-1}, K^{1,i}) p(Y^{i+1,n} | K) \int \frac{p(x_i | Y^{1,i}, K^{1,i}) p(x_i | Y^{i+1,n}, K)}{p(x_i | K^{1,i})} dx_i; \quad (5{\cdot}1)$$

$$p(K_i | K_{j \neq i}) \propto p(K_{i+1} | K_i) p(K_i | K_{i-1}). \quad (5{\cdot}2)$$

For $i = n$,

$$p(Y|K) \propto p(y_n|Y^{1,n-1}, K), \quad p(K_n|K_j, j<n) = p(K_n|K_{n-1}). \tag{5·3}$$

Proof. To obtain (5·1),

$$\begin{aligned} p(Y|x_i, K) &= p(Y^{1,i}|x_i, K)p(Y^{i+1,n}|Y^{1,i}, x_i, K) \\ &= p(Y^{1,i}|x_i, K)p(Y^{i+1,n}|x_i, K) \\ &= \frac{p(x_i|Y^{1,i}, K)p(Y^{1,i}|K)p(x_i|Y^{i+1,n}, K)p(Y^{i+1,n}|K)}{p(x_i|K^{1,i})^2}, \end{aligned}$$

so

$$\begin{aligned} p(x_i|Y, K) &= \frac{p(Y|x_i, K)p(x_i|K)}{p(Y|K)} \\ &= \frac{p(x_i|Y^{1,i}, K)p(Y^{1,i}|K)p(x_i|Y^{i+1,n}, K)p(Y^{i+1,n}|K)}{p(Y|K)p(x_i|K^{1,i})}. \end{aligned}$$

Using this result,

$$\begin{aligned} 1 &= \int p(x_i|Y, K)\, dx_i \\ &\propto \frac{p(Y^{1,i}|K^{1,i})p(Y^{i+1,n}|K)}{p(Y|K)} \int \frac{p(x_i|Y^{1,i}, K^{1,i})p(x_i|Y^{i+1,n}, K)}{p(x_i|K^{1,i})}\, dx_i. \end{aligned} \tag{5·4}$$

Equation (5·1) follows from (5·4) on noting that $p(Y^{1,i}|K^{1,i}) \propto p(y_i|Y^{1,i-1}, K^{1,i})$. It is straightforward to obtain equations (5·2) and (5·3).

The following notation is used below. Let

$$\mu_i := E(x_i|K^{1,i}), \quad S_i := \mathrm{var}(x_i|K^{1,i}), \quad \mu_{i|j,k} := E(x_i|Y^{j,k}, K^{1,k}), \quad S_{i|j,k} := \mathrm{var}(x_i|Y^{j,k}, K^{1,k}).$$

To generate K_i, it is necessary to evaluate (5·1) for each value of K_i. This requires computing $p(y_i|Y^{1,i-1}, K^{1,i})$, $p(Y^{i+1,n}|K)$, μ_i, S_i, $\mu_{i|1,i}$, $S_{i|1,i}$, $\mu_{i|i+1,n}$ and $S_{i|i+1,n}$ for each value of i. The terms μ_i and S_i are obtained from (2·1). The terms $p(y_i|Y^{1,i-1}, K^{1,i})$, $\mu_{i|1,i}$ and $S_{i|1,i}$ are obtained from $\mu_{i-1|1,i-1}$ and $S_{i-1|1,i-1}$ using one step of the Kalman filter, e.g. Anderson & Moore (1979, p. 105). It is more difficult to obtain efficiently the terms $p(Y^{i+1,n}|K)$, $\mu_{i|i+1,n}$ and $S_{i|i+1,n}$. We now outline how these terms may be obtained and what the difficulties are. Conditionally on K, the joint distribution of $x_1, \ldots, x_n$ is Gaussian and x_i is Markov, so that $p(x_i|X^{i+1,n}, K) = p(x_i|x_{i+1}, K)$. Thus, conditionally on K, $x_i = \tilde{F}_{i+1}x_{i+1} + \tilde{u}_{i+1}$, where

$$\tilde{F}_{i+1} = \mathrm{cov}(x_i, x_{i+1}|K)\,\mathrm{var}(x_{i+1}|K)^{-1}, \quad \tilde{U}_{i+1} := \mathrm{var}(\tilde{u}_{i+1}|K) = \mathrm{var}(x_i|x_{i+1}, K)$$

and $E(\tilde{u}_{i+1}|K) = \mu_i - \tilde{F}_{i+1}\mu_{i+1}$, so that $\tilde{F}_{i+1}$, $\tilde{U}_{i+1}$ and $E(\tilde{u}_{i+1}|K)$ depend on $K^{1,i+1}$. Thus, for each value of K_i, it is necessary in general to run the Kalman filter backwards for $j = n, \ldots, i$ to compute $p(Y^{i+1,n}|K)$, $\mu_{i|i+1,n}$ and $S_{i|i+1,n}$. This results in an $O(n^2)$ algorithm for generating the K_i, which is impractical for Markov chain Monte Carlo because the indicator variables are generated many times.

Section 5·2 shows how to evaluate (5·1) for all values of K_i and $i = 1, \ldots, n$ in $O(n)$ operations when the F_i matrices are nonsingular. Further details of the algorithm, and a

discussion of the case when both the transition matrices and the distribution of $x_i|Y^{1,i}, K^{1,i}$ are singular, are given in a technical report available from the authors.

5·2. *Backward state space model and algorithm*

To evaluate the right side of (5·1), the terms $p(y_i|Y^{1,i-1}, K^{1,i}, \mu_{i|1,i})$ and $S_{i|1,i}$ in (5·1) are obtained for each value of K_i from $\mu_{i-1|1,i-1}$ and $S_{i-1|1,i-1}$ using one step of the Kalman filter. To evaluate the composite term $p(x_i, Y^{i+1,n}|K)/p(x_i|K^{1,i})$ a 'backward' state space model is constructed whose joint distribution is similar to that of the original state space model (2·1). Given the indicator variables K, the backward model is defined by

$$y_{Bi} = h_i' x_{Bi} + e_{Bi}, \quad x_{B,i-1} = F_{Bi} x_{Bi} + u_{Bi},$$

where $F_{Bi} = F_i^{-1}$. The sequence e_{Bi} $(i = 1, \ldots, n)$ is independent normal with $E(e_{Bi}) = E(e_i)$ and $\text{var}(e_{Bi}) = \text{var}(e_i)$. The sequence u_{Bi} $(i = 1, \ldots, n)$ of normal random vectors is independent with $E(u_{Bi}) = -F_i^{-1}E(u_i)$ and $\text{var}(u_{Bi}) = F_i^{-1}\,\text{var}(u_i)(F_i')^{-1}$. The distribution of x_{Bn} initialising the backward recursions can be specified arbitrarily by the user; however, we take x_{Bn} diffuse, that is $x_{Bn} \sim N(0, cI_m)$ with $c \to \infty$, as this choice greatly simplifies the computation. Below, densities of the forward state space model (2·1) are written as $p(.)$ and densities of the backward state space model are written as $p_B(.)$. The shorthand notation $p_B(x_i)$ and $p_B(X^{i,n}, Y^{i+1,n})$ is used to mean the density $p_B(x_{Bi})$ evaluated at $x_{Bi} = x_i$ and the density $p_B(X_B^{i,n}, Y_B^{i+1,n})$ evaluated at $X_B^{i,n} = X^{i,n}$ and $Y_B^{i+1,n} = Y^{i+1,n}$ respectively. Whenever there is ambiguity in the notation, we write $p_B(x_{Bi} = x_i)$ and $p_B(X_B^{i,n} = X^{i,n}, Y_B^{i+1,n} = Y^{i+1,n})$.

It follows from Ansley & Kohn (1985) that

$$\bar{p}_B(x_{Bi}, Y_B^{i+1,n}|K) := \lim_{c\to\infty} c^{-\frac{1}{2}m} p_B(x_{Bi}, Y_B^{i+1,n}|K; c)$$

is finite and positive. By the construction of the backward state space model, $\bar{p}_B(x_{Bi}, Y_N^{i+1,n}|K)$ is independent of $K^{1,i}$. With some algebra it can be shown that

$$p(x_i, Y^{i+1,n}|K)/p(x_i|K^{1,i}) \propto \bar{p}_B(x_{Bi} = x_i, Y_B^{i+1,n} = Y^{i+1,n}|K) \tag{5·5}$$

$$\propto e^{-\frac{1}{2}(x_i-\alpha_i)'\Sigma_i(x_i-\alpha_i)}, \tag{5·6}$$

where

$$\alpha_i := \lim_{c\to\infty} E(x_{Bi}|Y_B^{i+1,n} = Y^{i+1,n}, K; c), \quad \Sigma_i := \lim_{c\to\infty} \{\text{var}(x_{Bi}|Y_B^{i+1,n} = Y^{i+1,n}, K; c)^{-1}\}.$$

The constants of proportionality in (5·5) and (5·6) are independent of $K^{1,i}$ and, by construction, so are α_i and Σ_i. For $i = 1, \ldots, n$, the terms α_i and Σ_i are obtained by running the modified Kalman filter backward. The right side of (5·1) is proportional to

$$p(y_i|Y^{1,i-1}, K^{1,i}) \times \int p(x_i|Y^{1,i}, K^{1,i}) e^{-\frac{1}{2}(x_i-\alpha_i)'\Sigma_i(x_i-\alpha_i)}\, dx_i,$$

and the integral can be evaluated by completing the square in the exponent of the integrand.

The following algorithm generates $K_1, \ldots, K_n$ in $O(n)$ operations.

Algorithm 5·1

Step 1. Run the modified Kalman filter backwards for $i = n, \ldots, 1$ and calculate α_i and Σ_i.

Step 2. For $i = 1, \ldots, n$:

Step 2·1: For each value of K_i:

Step 2·1·1: Run the Kalman filter one step forwards to calculate $p(y_i | Y^{1,i-1}, K^{1,i})$, $\mu_{i|1,i}$ and $S_{i|1,i}$ from $\mu_{i-1|1,i-1}$ and $S_{i-1|1,i-1}$.

Step 2·1·2: Evaluate (5·1) and (5·3) using the output of Step 1 and Step 2·1·1.

Step 2·2. Calculate the normalising constant and generate K_i.

Step 2·3. Run the Kalman filter one step forwards to calculate $\mu_{i|1,i}$ and $S_{i|1,i}$.

Acknowledgement

The research was partially supported by an Australian Research Council grant. We thank the editor for improving the presentation of the paper.

Appendix 1

Sampling from reduced conditionals

Sampling Scheme A, described below, shows how to construct Markov chain Monte Carlo schemes that group variables together in a more flexible way than the Gibbs sampler. It is used to show that Sampling Scheme 2·1 is invariant.

Suppose we wish to sample from the distribution $\pi(Z) = \pi(Z_1, \ldots, Z_l)$. We assume, for simplicity, that $Z \in \Re^l$, but the result applies to more general settings. We use π to denote corresponding marginal and conditional distributions. The Markov chain Monte Carlo approach is to construct a Markov chain $Z^{[0]}, Z^{[1]}, Z^{[2]}, \ldots$ that is π-invariant, π-irreducible and aperiodic. By Tierney (1994), the distribution of $Z^{[t]}$ converges to π as $t \to \infty$. Let

$$\{f_1(Z), f_{-1}(Z)\}, \{f_2(Z), f_{-2}(Z)\}, \ldots, \{f_k(Z), f_{-k}(Z)\}$$

define k different partitions of the variables $Z_1, \ldots, Z_l$ into 2 subsets such that each variable Z_j appears in at least one of $f_1(Z), \ldots, f_k(Z)$ and it is possible to generate from the conditional distributions $\pi\{Z | f_{-i}(Z)\}$. We consider Markov chains that generate $Z^{[t+1]}$, given $Z^{[t]}$, from Sampling Scheme A which is now described.

Sampling Scheme A.

Step 1. Set $z^0 = Z^{[t]}$.

Step 2. For $i = 1, \ldots, k$, generate z^i from $\pi\{Z | f_{-i}(Z) = f_{-i}(z^{i-1})\}$.

Step 3. Set $Z^{[t+1]} = z^k$.

Let $Q(Z^{[0]}, A) = \mathrm{pr}(Z^{[1]} \in A | Z^{[0]})$ denote the transition kernel of the Markov chain resulting from Sampling Scheme A.

Lemma A1·1. *The transition kernel Q is π-invariant; that is $\pi(A) = \int Q(Z, A)\pi(dZ)$ for all measurable sets A.*

The proof is straightforward.

Sampling Scheme A includes the Gibbs sampler and the substitution sampling algorithm of Gelfand & Smith (1990) as special cases. The Gibbs sampler corresponds to the case where each variable Z_j appears in exactly one of $f_1(Z), \ldots, f_k(Z)$; the substitution sampling algorithm corresponds to the case where each variable Z_j appears in exactly one of $f_{-1}(Z), \ldots, f_{-k}(Z)$. Our motivation for considering Sampling Scheme A is to try to use reduced conditional distributions whenever possible. Gelfand & Smith (1990, p. 401) derive a different sampling scheme that uses reduced conditional distributions. It can be shown, however, that their sampling scheme does not ensure π-invariance.

Lemma A1·1 is now used to show that Sampling Scheme 2·1 is invariant to $p(X, K, \sigma^2, \tau^2 | Y)$. For a given value of the indicator vector K, it follows from the definition of G that $G = H_1 X$, where the matrix H_1 is $n \times mn$ and is of rank n. To simplify notation, the dependence of H_1 on K is not shown. There exists a matrix H_2 such that $(H_1', H_2')'$ is nonsingular; let $G_2 := H_2 X$. Then Sampling Scheme 2·1 is equivalent to the following sampling scheme.

Generate from:

(i) $p(G_2, \tau^2 | Y, G, K, \sigma^2)$,

(ii) $p(X, K_i | Y, K_j, j \neq i, \sigma^2, \tau^2)$ for $i = 1, \ldots, n$,

(iii) $p(X | Y, K, \sigma^2, \tau^2)$,

(iv) $p(\sigma^2 | Y, X, K, \tau^2)$.

In (i) it is unnecessary to generate G_2, and in (ii) it is unnecessary to generate X. Invariance follows from Lemma A1·1.

APPENDIX 2

Proof of Lemma 2·1

Appendix 1 shows that Sampling Scheme 2·1 has invariant distribution $p(X, K, \sigma^2, \tau^2 | Y)$. The aperiodicity and irreducibility results in Lemma 2·1 follow from Lemma A2·1 below. For any $\bar{K}$, $\tilde{K} \in \mathscr{K}$, let

$$Q(\bar{K}, \tilde{K}) := \prod_{i=1}^{n} p(\tilde{K}_i | \tilde{K}_j, j < i, \bar{K}_j, j > i) \tag{A2·1}$$

be the one-step transition kernel for Sampling Scheme K. Let

$$Q(\bar{K}, \tilde{K} | Y, \sigma^2, \tau^2) := \prod_{i=1}^{n} p(\tilde{K}_i | \tilde{K}_j, j < i, \bar{K}_j, j > i, Y, \sigma^2, \tau^2). \tag{A2·2}$$

Let $Z := (X, \sigma^2, \tau^2)$ and denote the iterates of Sampling Scheme 2·1 by

$$(Z^{[j]}, K^{[j]}) = (X^{[j]}, (\sigma^2)^{[j]}, (\tau^2)^{[j]}, K^{[j]}).$$

Let $A_Z \subseteq \Omega_Z = \Re^{m \times n} \times \Re^+ \times \Re^+$ and define the one-step transition kernel for Sampling Scheme 2·1 by

$$Q\{(\bar{Z}, \bar{K}), (A_Z, \tilde{K})\} := \mathrm{pr}(Z^{[j+1]} \in A_Z, K^{[j+1]} = \tilde{K} | Y, Z^{[j]} = \bar{Z}, K^{[j]} = \bar{K}).$$

For any one-step transition kernel Q, let Q^r be the r step transition kernel.

LEMMA A2·1. *Suppose that*

$$\mathrm{pr}(Z \in A_Z, \tilde{K} | Y) > 0. \tag{A2·3}$$

Then $Q^r(\bar{K}, \tilde{K}) > 0$ *if and only if* $Q^r\{(\bar{Z}, \bar{K}), (A_Z, \tilde{K})\} > 0$.

Proof. Showing that $Q^r(\bar{K}, \tilde{K}) > 0$ if $Q^r\{(\bar{Z}, \bar{K}), (A_Z, \tilde{K})\} > 0$ is straightforward. We now prove that $Q^r\{(\bar{Z}, \bar{K}), (A_Z, \tilde{K})\} > 0$ if $Q^r(\bar{K}, \tilde{K}) > 0$.

Without loss of generality, we can consider sets A_Z of the form $A_Z = A_X \times A_{\sigma^2} \times A_{\tau^2}$, where $A_X \subseteq \Re^{m \times n}$, $A_{\sigma^2} \subseteq \Re^+$ and $A_{\tau^2} \subseteq \Re^+$. Suppose that $Q(\bar{K}, \tilde{K}) > 0$.

The result is first obtained for $r = 1$. We can show that

$$Q\{(\bar{Z}, \bar{K}), (A_Z, \tilde{K})\} = \int_{A_{\tau^2}} p(d\tau^2 | Y, \bar{G}, \bar{K}, \bar{\sigma}^2) \times \left[Q(\bar{K}, \tilde{K} | Y, \bar{\sigma}^2, \tau^2) \int_{A_X} p(dX | Y, \tilde{K}, \bar{\sigma}^2, \tau^2) \left\{ \int_{A_{\sigma^2}} p(d\sigma^2 | Y, X, \tilde{K}, \tau^2) \right\} \right]. \tag{A2·4}$$

The inequality $Q(\bar{K}, \tilde{K} \mid Y, \sigma^2, \tau^2) > 0$ is deduced from (A2·1), (A2·2), and Assumption 2 in § 2. The inequalities

$$\text{pr}(\tau^2 \in A_{\tau^2} \mid Y, G, \bar{K}, \sigma^2) > 0, \quad \text{pr}(X \in A_X \mid Y, \tilde{K}, \sigma^2, \tau^2) > 0, \quad \text{pr}(\sigma^2 \in A_{\sigma^2} \mid Y, X, \tilde{K}, \tau^2) > 0$$

are obtained from (A2·3) and Assumptions 1, 2 and 3 in § 2. Substituting into (A2·4) gives $Q\{(\bar{Z}, \bar{K}), (A_Z, \tilde{K})\} > 0$ as required.

Now suppose the result is true for $i = 1, \ldots, r$ and $Q^{r+1}(\bar{K}, \tilde{K}) > 0$. This implies that there exists $K^r \in \mathscr{K}$ such that $Q^r(\bar{K}, K^r) > 0$ and $Q(K^r, \tilde{K}) > 0$. Note that $\text{pr}(Z \in \Omega_Z, K^r \mid Y) = p(K^r \mid Y) > 0$. Applying the inductive hypothesis gives

$$Q^r\{(\bar{Z}, \bar{K}), (\Omega_Z, K^r)\} > 0, \quad Q\{(Z, K_r), (A_Z, \tilde{K})\} > 0.$$

Thus,

$$Q^{r+1}\{(\bar{Z}, \bar{K}), (A_Z, \tilde{K})\} \geqslant \int_{\Omega_Z} Q^r\{(\bar{Z}, \bar{K}), (dZ, K_r)\} Q\{(Z, K_r), (A_Z, \tilde{K})\} > 0,$$

as required. □

References

Anderson, B. D. O. & Moore, J. B. (1979). *Optimal Filtering*. Englewood Cliffs, NJ: Prentice Hall.

Ansley, C. F. & Kohn, R. (1985). Estimation, filtering and smoothing in state space models with partially diffuse initial conditions. *Ann. Statist.* **13**, 1286–316.

Ansley, C. F. & Kohn, R. (1990). Filtering and smoothing in state space models with partially diffuse initial conditions. *J. Time Ser. Anal.* **11**, 277–93.

Carter, C. K. & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–53.

De Jong, P. & Shephard, N. (1995). Efficient sampling from the smoothing density in time series models. *Biometrika* **82**, 339–50.

Frühwirth-Schnatter, S. (1993). Data augmentation and dynamic linear models. *J. Time Ser. Anal.* **15**, 183–202.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398–409.

Gordon, K. & Smith, A. F. M. (1990). Monitoring and modeling biomedical time series. *J. Am. Statist. Assoc.* **85**, 328–37.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–84.

Harrison, P. J. & Stevens, C. F. (1976). Bayesian forecasting (with Discussion). *J. R. Statist. Soc.* B **38**, 205–47.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Liu, J., Wong, W. H. & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

McDonald, J. A. & Owen, A. B. (1986). Smoothing with split linear fits. *Technometrics* **28**, 195–208.

Müller, H. G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.* **20**, 737–61.

Shephard, N. (1994). Partial non-Gaussian state space models. *Biometrika* **81**, 115–32.

Shumway, R. H. & Stoffer, D. S. (1991). Dynamic linear models with switching. *J. Am. Statist. Assoc.* **86**, 763–9.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701–62.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc.* B **40**, 133–50.

Wecker, W. E. & Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Am. Statist. Assoc.* **78**, 81–9.

West, M. & Harrison, J. (1989). *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

[*Received March* 1994. *Revised June* 1995]