ELSEVIER

Journal of Econometrics 74 (1996) 237–254

JOURNAL OF Econometrics

Bayesian estimation of an autoregressive model using Markov chain Monte Carlo

Glen Barnett, Robert Kohn*, Simon Sheather

Australian Graduate School of Management, University of New South Wales, Sydney, NSW 2052, Australia

(Received February 1993; final version received March 1995)

Abstract

We present a complete Bayesian treatment of autoregressive model estimation incorporating choice of autoregressive order, enforcement of stationarity, treatment of outliers, and allowance for missing values and multiplicative seasonality. The paper makes three distinct contributions. First, we enforce the stationarity conditions using a very efficient Metropolis-within-Gibbs algorithm to generate the partial autocorrelations. Second we show how to carry out the Gibbs sampler when the autoregressive order is unknown. Third, we show how to combine the various aspects of fitting an autoregressive model giving a more comprehensive and efficient treatment than previous work. We illustrate our methodology with a real example.

Key words: Gibbs sampler; Metropolis algorithm; Missing data; Order selection; Outliers
JEL classification: C11; C15; C22

1. Introduction

The paper presents a comprehensive approach to Bayesian modelling of autoregressive time series including robust estimation of parameters, identification of outliers, enforcement of stationarity, and the interpolation of missing observations. Most importantly, account is taken of the lack of knowledge of the correct model by averaging parameter estimates and forecasts over the set of permissible models. Posterior moments and densities are obtained by Markov chain Monte Carlo sampling. The methodology is applied to the autoregressive

* Corresponding author.

The work of Robert Kohn and Simon Sheather was partially supported by an Australian Research Grant.

SSDI 0304-4076(95)01744-X

modelling of arsenic concentration in sludge obtained from a Sydney Water Board treatment plant.

Autoregressive models are used extensively for a number of purposes including short- and long-term forecasting; understanding the dynamics of an underlying physical system by looking at the model order and parameter values and by simulating observations from the model; spectral density estimation of a stationary process by robustly fitting an autoregression to the observations, e.g., Kleiner, Martin, and Thompson (1979); including a stationary component in a bigger model, e.g., Jacquier, Polson, and Rossi (1994) who model (unobserved) stochastic volatility as a first-order autoregressive process; determining if there is a unit root in a macroeconomic time series, e.g., Phillips (1991). In all of these applications, robust estimation, enforcement of stationarity, and averaging parameter estimates and forecasts over the posterior probabilities of different models are often important.

There is an extensive literature on fitting autoregressions using both frequentist methods, e.g., Box and Jenkins (1976), and Bayesian methods, e.g., Zellner (1971, Ch. 7), Box and Jenkins (1976, p. 250), Monahan (1984), and Marriott and Smith (1992). Both maximum likelihood and Bayesian parameter estimates can be badly affected by outliers and level shifts. To overcome this nonrobustness a number of authors, e.g., Tsay (1988) and Chen and Liu (1993), propose an iterative approach to model fitting consisting of the following steps: (i) model identification using some model selection criterion; (ii) parameter estimation; (iii) detection of outliers; (iv) model checking. Most current approaches handle these steps one at a time, iterating (i) to (iv) until a satisfactory fit is obtained. Although useful, such an iterative approach remains somewhat unsatisfactory for two reasons. First, steps (i)–(iii) are highly dependent. Model selection and parameter estimation are often dramatically affected by outliers. Conversely, great care needs to be taken with detection of the outliers as these outliers are relative to the current model, which may be incorrect. A data point identified as an outlier under one model may be a perfectly acceptable point under another model. Second, it is often difficult to know when to stop when using the above steps and the significance levels used to guide the iterative analysis are usually very approximate.

Using Markov chain Monte Carlo sampling, parts (i) to (iii) of the model fitting procedure outlined above can be handled simultaneously, which makes the parameter estimates and the order selected robust to outliers. Unlike previous approaches, such as using the partial autocorrelations or Akaike's Information Criterion, the present approach also permits the standard errors of the parameter estimates and the widths of prediction intervals to incorporate the extra variation due to uncertainty about the correct model.

We briefly summarise the literature on Markov chain Monte Carlo for autoregressive models. A detailed comparison of our approach with previous work is given in Section 4. McCulloch and Tsay (1994) propose a Gibbs sampler for the Bayesian analysis of an autoregressive model which allows additive

outliers, level shifts, variance shifts, and missing observations. However, each of these is handled one at a time (together with the generation of the coefficients of the model) rather than simultaneously. They do not enforce stationarity, nor do they handle order selection. By generating the additive outliers and the outlier indicators simultaneously our approach improves on that of McCulloch and Tsay, and so is likely to converge faster. Chib (1993) proposes a Gibbs sampler for a regression model with autoregressive errors; his results are subsumed in Chib and Greenberg (1994) which is discussed below. Albert and Chib (1993) propose a Gibbs sampler for an autoregressive model with intercept and variance shifts governed by a Markov process with the variance shifts similar to our innovation outliers. We could in principle generalise the sampler in Section 3 so that the indicator variables are not independent, but have a Markov structure. Marriott, Ravishanker, Gelfand, and Pai (1994) and Chib and Greenberg (1994) propose Markov chain samplers for autoregressive moving average models whose aim is to generate the model parameters and enforce stationarity and invertibility. Section 4 shows that for autoregressive models our way of generating the parameters is more efficient than that of Marriott et al. (1994) and that their approach can be very inefficient in the presence of outliers. Chib and Greenberg (1994) generate all the autoregressive parameters as a block and enforce stationarity using a rejection step. Section 4 shows that it seems difficult to extend the Chib and Greenberg approach to handle model selection and model averaging. In addition, their approach can be very inefficient for seasonal models with roots close to the unit circle.

2. Model and prior assumptions

We consider the model

$$y_t = w_t + o_t, \qquad \phi(B)\Phi(B^s)[(1-B^s)^{d_s}(1-B)^d w_t - \mu] = e_t, \tag{2.1}$$

where y_t is the tth observation on the integrated autoregressive process w_t, with o_t an additive outlier. The backshift operator B is defined as $By_t = y_{t-1}$, s is the seasonal period, d is the order of regular differencing, and d_s is the order of seasonal differencing. The regular and seasonal autoregressive polynomials are

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p, \quad \Phi(B) = 1 - \Phi_1 B - \cdots - \Phi_q B^q, \tag{2.2}$$

with orders p and q respectively.

Assumption 1. Both $\phi(B)$ and $\Phi(B)$ have all their roots outside the unit circle.

Assumption 2. To allow for additive and innovation outliers, the errors o_t and e_t are modelled as finite mixtures of normals. Thus $o_t \sim \mathrm{N}(0, K_{1t}\sigma^2)$ and $e_t \sim \mathrm{N}(0, K_{2t}\sigma^2)$ and $K_t = (K_{1t}, K_{2t})$ has a bivariate multinomial distribution

with $K_{1t} \geq 0$ and $K_{2t} \geq 1$. If $K_{1t} > 0$, then there is an additive outlier at time t, and if $K_{2t} > 1$, then there is an innovation outlier at time t. The case of no additive outlier and no innovation outlier corresponds to $K_t = (0, 0)$ so that $o_t = 0$ and $e_t \sim \mathrm{N}(0, \sigma^2)$.

In practice, we usually only allow either an innovation outlier or an additive outlier at each time point, due to the difficulty of identifying both at the same time. This restriction is achieved by setting the prior probability of all combination outliers to zero and greatly simplifies the computation. For example, Table 2 shows the prior distribution for K_t used in the sludge example in Section 5. McCulloch and Tsay (1994) use a mixture of two components to model additive outliers, with one component having very large variance, but we have found that a multi-component mixture does better at picking up outliers of different sizes.

Let $\psi = (\psi_1, \ldots, \psi_p)'$ be the first p partial autocorrelations of a zero mean stationary autoregressive process with autoregressive polynomial $\phi(B)$ and let $\Psi = (\Psi_1, \ldots, \Psi_q)'$ be the first q partial autocorrelations of an autoregressive process with autoregressive polynomial $\Phi(B)$. By reparameterising ϕ and Φ in terms of ψ and Ψ as in Monahan (1984) the stationarity constraints become $-1 < \psi_i < 1$ for $i = 1, \ldots, p$ and $-1 < \Psi_i < 1$ for $i = 1, \ldots, q$. The orders p and q are chosen by the user as the maximal regular and seasonal orders. To allow some of the ψ_j and Ψ_j to be zero, let $J_{1j} = 0$ if $\psi_j = 0$ and let $J_{1j} = 1$ otherwise, $j = 1, \ldots, p$. Let J_{2j} be defined similarly with respect to the Ψ_j, for $j = 1, \ldots, q$.

Assumption 3. (i) The prior for $J_{1j} = 1$, $j = 1, \ldots, p$, is prescribed by the user. For example, Table 3 gives the prior for the sludge example with $p = 10$ and the probabilities of $J_{1j} = 1$ decline as j increases. The prior for J_{2j} is similarly prescribed by the user. (ii) For $i = 1, \ldots, p$ and $j = 1, \ldots, q$ the indicators J_{1j} and J_{2j} are independently distributed. (iii) Given the indicators I_{1j} and I_{2j}, ψ_j and Ψ_j are independently distributed of each other. If $J_{1j} = 1$, then ψ_j is uniformly distributed on $(-1, 1)$, and if $J_{2j} = 1$, then Ψ_j is uniformly distributed on $(-1, 1)$.

Assumption 4. The prior for μ is flat and the prior for σ^2 is $f(\sigma^2) \propto 1/\sigma^2$.

To allow for additive outliers, innovation outliers, and missing observations in the period $1 \leq t \leq n$ it is convenient to also generate the pre-period values $Y^- = (y_0, y_{-1}, \ldots, y_{1-D-Q})'$, where $D = d + sd_s$, $Q = p + sq$. Let $u_t = (1 - B)^d (1 - B^s)^{d_s} w_t$ so that $\phi(B)\Phi(B^s)(u_t - \mu) = e_t$, $u^- = (u_0, \ldots, u_{1-Q})'$, and $w^- = (w_{-Q}, w_{-Q-1}, \ldots, w_{1-Q-D})'$.

Assumption 5. (i) There are no innovation or additive outliers for $t \leq 0$. This means that u_t is stationary for $t \leq 0$ so that u^- has known mean and variance

given μ, ϕ, Φ, and σ^2. (ii) As in Kohn and Ansley (1986), we also assume that w^- is diffuse, i.e., $w^- \sim N(0, \kappa I_D)$ with $\kappa \to \infty$.

In dealing with diffuse initial conditions we adopt the convention that if X and Z are two random vectors and the conditional density $f(X|Z, \kappa)$ tends to a finite nonzero limit as $\kappa \to \infty$, then we write this limit as $f(X|Z)$.

The following notation is used in the rest of the paper. Let $K_t = (K_{1t}, K_{2t})$, $K = (K_1, \ldots, K_n)$, $O = (o_1, \ldots, o_n)$, $J_1 = (J_{11}, \ldots, J_{1n})$, $J_2 = (J_{21}, \ldots, J_{2n})$, and $Y = (y_1, \ldots, y_n)$. Let Y^M consist of all y_t, $1 < t < n$ that are missing and Y^O consist of all y_t that are observed.

Remark 2.1. Instead of modeling the errors o_t and e_t as heavy-tailed using a discrete mixture of normals it is possible to use a continuous mixture such as a t distribution. We prefer to use discrete mixtures because it allows K_{1t}, K_{2t}, and o_t to be generated jointly. This is not possible using continuous mixtures.

Remark 2.2. We assume a flat prior for ψ_j and assume that the ψ_j are a priori independent. It is easy to deal with other priors for ψ by generating ψ as we do now and then using the Metropolis algorithm. An example of a different prior is obtained by placing a flat prior on ϕ and then transforming to obtain the appropriate prior for ψ. We note that different priors on ψ imply different priors on the roots of the autoregressive polynomial and hence different priors on the dynamics of the autoregressive process. In particular, different priors give different probabilities that the autoregressive polynomial has complex roots. This point is extensively discussed by Hong (1989). Similar remarks apply to Ψ. The choice of appropriate priors for ψ and Ψ is not discussed in this paper and is the subject of future research.

3. Sampling scheme

The following sampling scheme is used to estimate the posterior distribution of ψ, J_1, Ψ, J_2, K, μ, and σ^2 and any missing data. Starting with some initial values of all the unknown parameters and observations, the sampler successively generates a parameter or group of parameters conditional on the observations and the other parameters as described below. The iterates of the sampler are divided into a warmup period and a sampling period. It is assumed that the sampler has converged to the correct posterior distribution at the end of the warmup period and estimates of the posterior moments and densities are based on the iterates in the sampling period. We now describe the sampler with implementation details given in the Appendix.

Sampling Scheme 1. Generate from

(i) $f(J_{1j}, \psi_j | Y, Y^-, \psi_{i \neq j}, J_{1,i \neq j}, \Psi, J_2, K, O, \sigma^2, \mu)$ for $j = 1, \ldots, p$.

(ii) $f(J_{2j}, \Psi_j | Y, Y^-, \psi, J_1, \Psi_{i \neq j}, J_{2,i \neq j}, K, O, \sigma^2, \mu)$ for $j = 1, \ldots, q$.

(iii) $f(K_t, o_t | Y, Y^-, \psi, J_1, \Psi, J_2, K_{s \neq t}, o_{s \neq t}, \sigma^2, \mu)$ for $t = 1, \ldots, n$. An improved step is described below.

(iv) $f(\sigma^2 | Y, Y^-, \psi, J_1, \Psi, J_2, K, O, \mu)$.

(v) $f(\mu | Y, Y^-, \psi, J_1, \Psi, J_2, K, O, \sigma^2)$.

(vi) $f(Y^-, Y^M | Y^O, \psi, J_1, \Psi, J_2, K, O, \sigma^2, \mu)$. □

Sampling Scheme 1 is invariant to the posterior distribution by a similar argument to that showing that the Gibbs and Metropolis samplers are invariant; it is irreducible and aperiodic because, with positive probability, it is possible to get in one step from any point in the generated variable space to any other point. Therefore Sampling Scheme 1 converges to the correct posterior distribution by Tierney (1994).

In step (i), ψ_j and J_{1j} are generated jointly by first generating J_{1j} from a binomial distribution with ψ_j integrated out; next ψ_j is generated conditional on J_{1j}, using a Metropolis–Hastings step with $g(\psi_j) \propto f(Y | Y^-, \psi_{i \neq j}, J_1, \Psi, J_2, K, O, \sigma^2, \mu)$ as the proposal density. This is computationally very efficient, as g is Gaussian in ψ_j because ϕ is linear in ψ_j, if the other $\psi_{i \neq j}$ are held constant; see Appendix. The terms J_{2j} and Ψ_j are generated similarly.

In step (iii), o_t and K_t are generated jointly by first integrating o_t out and generating K_t from a multinomial distribution. It is straightforward to integrate out o_t and to generate it as the likelihood is Gaussian in o_t. The variables o_t and K_t are generated jointly because there is likely to be a high posterior dependence between K and O, which may result in slower convergence if K_t was generated conditional on o_t and o_t was generated conditional on K_t. The following simple modification to step (iii) is even more effective in reducing this dependence.

(iii′) Generate $f(K_t, o_{t-1}, o_t, o_{t+1} | Y, Y^-, J_1, \Psi, J_2, K_{s \neq t}, o_s, s \neq t-1, t, t+1, \sigma^2, \mu)$ by first generating K_t with o_{t-1}, o_t and o_{t+1} integrated out as in the Appendix and then generate o_{t-1}; there is no need to generate o_t and o_{t+1} as they are integrated out when generating K_{t+1}.

The parameters σ^2 and μ are generated from inverse gamma and normal distributions respectively using a similar derivation to that of McCulloch and Tsay (1994). In principle, it is straightforward to generate u^-, w^-, and Y^M simultaneously as their joint conditional distribution is Gaussian. In practice, if (u^-, w^-, Y^M) is high-dimensional, then it may be more convenient to generate the elements of this vector one at a time or in small groups.

The first and second posterior moments of ψ and Ψ and the posterior distributions of σ^2, μ, Y^-, Y^M, J_1, J_2, and K_t, $t = 1, \ldots, n$, can be efficiently estimated using mixture estimates as in Gelfand and Smith (1992, Sec. 2.6).

4. Comparison with previous work

We now compare Sampling Scheme 1 to previous work on autoregressive models and variable selection using Markov chain Monte Carlo. McCulloch and Tsay (1994) consider an autoregressive model of fixed order p and allow additive outliers, missing observations, and level shifts. They assume that the first p observations are available and condition on them. As well as giving a more comprehensive approach than McCulloch and Tsay, Sampling Scheme 1 improves on the samplers in McCulloch and Tsay in handling initial conditions and additive and innovation outliers. First, conditioning on initial values is wasteful if p is large, as in seasonal models, and there are missing observations at the beginning of the series. Suppose, for example, that $p = 12$, the data is stationary, and $n = 100$. By conditioning on the first 12 observations McCulloch and Tsay effectively lose about 10% of their data. If, in addition, y_3, y_{10}, and y_{20} are missing, then they would have to take observation 21 as the beginning of their data, which means effectively losing about a third of their data. To overcome this starting value problem a prior on the initial values such as the one given in Section 1 appears necessary.

Second, McCulloch and Tsay model additive outliers as $o_t = K_{1t}a_t$ with $a_t \sim \mathrm{N}(0, \sigma^2)$ and their sampling scheme generates a_t conditional on K_{1t} and K_{1t} conditional on a_t. Step (iii) of Sampling Scheme 1 is equivalent to generating a_t and K_{1t} simultaneously at no extra cost. The results in Liu, Wong, and Kong (1994) suggest that Sampling Scheme 1 is likely to produce faster convergence of the Markov chain than the sampler in McCulloch and Tsay (1994) as a_t and K_{1t} are likely to be dependent given the data. The improved step (iii') above performs even better. The extra efficiency from simultaneous generation becomes even more pronounced when innovation outliers and level shifts are dealt with simultaneously with additive outliers. Gelman and Rubin (1992) argue that it is difficult to judge convergence of samplers from single runs and propose multiple runs with starting values near the modes of the posterior distribution. For the autoregressive model with innovation and additive outliers there are $2n$ indicator variables, so it is difficult to determine the posterior modes without running the Markov chain, and hence it is important to use as efficient a sampler as possible. Finally, we note that McCulloch and Tsay do not enforce stationarity, nor do they carry out order selection and these are important in some applications as mentioned in the introduction.

Marriott et al. (1996) propose a Bayesian analysis of a stationary autoregressive moving average model. If we specialise their results to a stationary

autoregressive model of order p, Marriott et al. (1996) use the parameterisation $\eta_j = \log[(1+\psi_j)/(1-\psi_j]$, $j = 1, \ldots, p$, so that $\eta = (\eta_1, \ldots, \eta_p)'$ are unconstrained and are a one-to-one transformation of ψ. Let $\hat{\eta}_1^{ML} = (\hat{\eta}_1^{ML}, \ldots, \hat{\eta}_p^{ML})'$ be the maximum likelihood estimate of η with Ω_η its asymptotic variance matrix. As in Marriott et al. (1996), outliers are not taken into account in obtaining $\hat{\eta}^{ML}$. Let $g(\eta)$ be the multivariate normal density $N(\hat{\eta}^{ML}, \Omega_\eta)$ with $g(\eta_j|\eta_{i\neq j})$ the corresponding conditional density. Marriott et al. generate η_j from $g(\eta_j|\eta_{i\neq j})$ and use the Metropolis algorithm in order to draw from the correct distribution $f(\eta_j|Y, Y^-, \sigma^2, \mu, \eta_{i\neq j})$. Our Metropolis proposal will always be computationally faster and have fewer rejections than theirs as it requires the ratio of two densities of Y^- at two different values of ψ_j (see Appendix), whereas they require the ratio of the joint density of Y and Y^-. Furthermore, when there are outliers in the data, $\hat{\eta}^{ML}$ may be a poor estimate of η, so using $g(\eta)$ as the Metropolis proposal in step (i) of Sampling Scheme 1 often results in a high rejection rate. We performed a small simulation study to check this. One hundred observations were generated from the first-order autoregression $y_t = 0.3y_{t-1} + e_t$, with $e_t \sim N(0, 1)$, and the maximum likelihood estimate of ψ and its asymptotic variance were computed. Sampling Scheme 1 was applied to the data with the order fixed at $p = 1$, with the variable selection turned off, and the Metropolis step in part (i) replaced by that of Marriott et al., except that instead of working with η we work with ψ. With outlier detection turned off, the rejection rate for the Marriott et al. Metropolis step was 4.4%. If outlier detection was turned on, the rejection rate was 26%. Next, we repeated the experiment when an additive outlier of size 10 was placed at $n = 50$. With outlier detection turned on, the rejection rate for Marriott et al. was 56%. The rejection rate for our algorithm was 2%. The results for a single outlier suggest that for multiple outliers the rejection rate for the Marriott et al. Metropolis proposal could become very high. George and McCulloch (1993, 1994) use the Gibbs sampler to select variables in a linear regression with indicator variables determining whether a coefficient is 'practically' zero. George and McCulloch (1993) propose a sampler in which the indicator variables are generated conditional on the regression coefficients and the regression coefficients are generated conditional on the indicator variables. George and McCulloch (1994) introduce a second sampler which generates the indicator variables with the regression coefficients integrated out. Sampling Schemes 2 and 3 stated below adapt the George and McCulloch (1993, 1994) approaches to autoregressive models and combine them with the other steps in Sampling Scheme 1. For simplicity we consider $d = 0$ and $s = 1$. Suppose that $\phi_j \sim N(0, J_{3j}\sigma^2)$ for $j = 1, \ldots, p$ where J_{3j} takes two values; J_{3j} small means that ϕ_j is not of practical significance and J_{3j} large means that the jth lag of the autoregression is important. A prior is prescribed for J_{3j}, $j = 1, \ldots, p$. Let $J_3 = (J_{31}, \ldots, J_{3p})$. The next sampling scheme is based on George and McCulloch (1993) and generates the indicators conditional on ϕ.

Sampling Scheme 2. Steps (iii) to (v) are the same as for Sampling Scheme 1. Let $R = J_{3,i \neq j}, \mu, \sigma^2, K, O$. Step (i) is replaced by generating from: $(i_a)\, f(J_{3j}|Y, Y^-, \phi, R) = f(J_{3j}|\phi_j)$ for $j = 1, \ldots, p$; $(i_b) f(\phi|Y, Y^-, J_3, R)$. Step (i_a) is straightforward and step (i_b) is done as in Chib and Greenberg (1994). □

The following sampling scheme is based on George and McCulloch (1994) and integrates out ϕ.

Sampling Scheme 3. Steps (iii) to (v) are the same as Sampling Scheme 1. Let R be the same as above. Step (i) is replaced by generating from: $(i_a) f(J_{3j}|Y, Y^-, J_{3,i \neq j}, R)$ for $j = 1, \ldots, p$; $(i_b) f(\phi|Y, Y^-, J_3, R)$. □

To generate J_{3j} in Sampling Scheme 3, note that

$$f(J_{3j}|Y, Y^-, J_{3,i \neq j}, R) \propto \int f(Y|Y^-, J_3, R, \phi) f(Y^-|J_3, R, \phi) f(\phi)\, d\phi. \quad (4.1)$$

This integral is in general intractable if ϕ is constrained to the stationarity region. If $f(Y^-|J_3, R, \phi)$ is independent of ϕ and $f(\phi)$ is Gaussian or uniform, then it is straightforward to integrate (4.1). We carried out several simulation experiments to compare how well Sampling Schemes 1 to 3 select the order of the autoregression and report on one of these. One hundred observations were generated from the sixth-order autoregressive model $y_t = \sum_{i=1}^{6} \phi_j y_{t-j} + e_t$ with $e_t \sim N(0, 1)$ and $\phi = (-0.09, 0.9, 0.0, -0.45, -0.045, 0.5)$ which corresponds to $\psi = (-0.9, 0.9, 0.0, 0.0, 0.0, 0.5)$. For simplicity, additive and innovation outlier detection was turned off. We took $p = 10$ as the maximal order of the autoregression and for Sampling Schemes 2 and 3 we took the small and large values of J_{3j} as 0.1 and 100. The priors for J_{1j} and J_{3j} are $f(J_{1j} = 1) = f(J_{3j} = 100) = 0.9^j$, $j = 1, \ldots, 10$. One hundred observations were generated from the autoregressive model and all sampling schemes were run for a warmup period of 50 iterations and a sampling period of 200 iterations. For each iteration in the sampling period the order of the autoregression was recorded and the modal model order determined over the 200 iterations. Stationarity was not enforced in Sampling Schemes 2 and 3. This procedure of generating the data and running the samplers was repeated 500 times. The results are presented in Table 1 which shows the frequency of the modal orders over the 500 runs. Table 1 shows that both Sampling Schemes 1 and 3 perform well, whereas Sampling Scheme 2 repeatedly selects an autoregression of order 2 and appears not to have converged. To check the convergence of Sampling Scheme 2 we ran it for 20,000 iterations with similar results to those in Table 1. We tried other values of J_{3j} and also experienced difficulties, e.g., the sampler appeared not to converge when the small value of J_{3j} was 0.01 and the large value was 1,000. Sampling Schemes 1 and 3 gave the same results from different starting conditions indicating that they had converged to the whole posterior distribution.

Table 1
Frequency of modal order selected

Sampling scheme	Order										
	0	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	416	19	25	24	16
2	1	0	483	2	0	0	0	14	0	0	0
3	0	0	7	1	1	2	445	6	14	14	10

These results were typical of other runs with Sampling Schemes 2 and 3. In general the speed of convergence of Sampling Scheme 2 was very sensitive to the small value taken by J_{3j}, whereas Sampling Scheme 3 (without enforcing the stationarity conditions) performed similarly to Sampling Scheme 1.

Chib and Greenberg (1994) propose a Bayesian analysis of a regression model with stationary autoregressive moving average errors. To compare their approach to ours, we specialise their approach to a stationary autoregressive model of order p and suppose that p is fixed, there are no outliers, and the prior $f(\phi)$ is Gaussian. Then $f(Y \mid Y^{-}, \phi, \mu, \sigma^2) f(\phi)$ is Gaussian in ϕ with some mean $\hat{\phi}$ and variance Ω. Chib and Greenberg (1994) generate ϕ as a block from $N(\hat{\phi}, \Omega)$ and test if the generated ϕ lies in the stationarity region. This is repeated until a ϕ in the stationarity region is obtained. A Metropolis step similar to step (i) in Sampling Scheme 1 is then performed. It seems difficult to generalise the Chib and Greenberg (1994) approach to handle model selection or model averaging because it seems difficult to enforce the stationarity conditions in Sampling Scheme 3.

For a fixed model, the Chib and Greenberg approach, which generates all the autoregressive parameters at once, can be substantially more efficient than our approach, which generates the partials one at a time. However, in high-order autoregressive models with roots close to the unit circle, their sampler can have very high rejection rates both when testing for stationarity and in the Metropolis step. We ran a small simulation to study this. One hundred observations were generated from the twelfthorder autoregression $y_t = 0.9y_{t-12} + e_t$ with $e_t \sim N(0, 1)$. The least squares estimate of $\phi = (\phi_1, \ldots, \phi_{12})$ was computed for the model

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots \phi_{12} y_{t-12} + e_t, \tag{4.2}$$

with the elements $\phi_1, \ldots, \phi_{12}$ unconstrained. This was repeated 1000 times. We found that 54% of the least squares estimates were outside the stationarity region. Next, 100 data sets were generated from the seasonal model (4.2), each consisting of 100 observations, and the Chib and Greenberg sampler was run with warmup and sampling periods of 1000 iterations each. For those data sets

for which the least squares estimate was outside the stationarity region, 71% of the iterates were generated outside the stationarity region. For those data sets for which the least squares estimate was inside the stationarity region, 49% of the iterates were generated outside the stationarity region. Once an iterate of ϕ is within the stationarity region, the rejection rate in the Metropolis step was 69%. This compares with a rejection rate of 6.5% for Sampling Scheme 1.

5. Example

The Bayesian approach to estimating an autoregressive model robustly is illustrated by an application to arsenic concentration in sewage sludge from one of the Sydney Water Board's treatment plants. Log concentrations are used rather than concentrations since it is natural to think of percentage changes in concentration rather than absolute changes. A second justification for taking logs is that concentrations are proportions which in this data set are extremely small. Thus log concentration is essentially the logit of a proportion. As a result, taking logs of the concentration stabilizes the variance and makes the data much more symmetric. The data are based on measurements of samples collected daily over a 488-day period. During this period there were four days on which it was not possible to collect samples of sludge, namely days 69, 214, 266, and 440. Thus the data is made up of 484 log concentrations and four missing values.

Fig. 1(a) plots the data and shows the presence of outliers. The values on the y axis are omitted to preserve confidentiality. A plot of the partial autocorrelations suggested fitting an autoregression of order 6. Sampling Scheme 1 in Section 3 is applied to a stationary autoregressive model with a maximal order $p = 10$ and no seasonal component. The prior distribution for K_t is given in Table 2, with additive and innovation outliers not allowed to occur simultaneously. Table 3 gives the prior probability that $J_{1j} = 1$ for $j = 1, \ldots, 10$. For the results reported below, Sampling Scheme 1 was run for a warmup period of 1250 iterations with a sampling period of 2500 iterations. The sampler was initialized with μ as the sample mean, σ^2 as four times the sample variance of the observations, $\psi = 0$, all elements of Y^- set to y_1, $J_1 = 0$, $K = 0$, and a missing observation set to the previous observation. In part (i) of the sampler, which involves a Metropolis rejection step, the rejection rate was 6%, suggesting the algorithm is very efficient.

Fig. 1(b) plots the data as ordinary observations (O), additive outliers (A), and innovation outliers (I). An observation is classified as an additive outlier if its posterior probability of being an additive outlier is greater than its posterior probability of not being an outlier or an innovation outlier, with a similar classification for an innovation outlier. Figs. 2(a) and 2(b) are the estimated

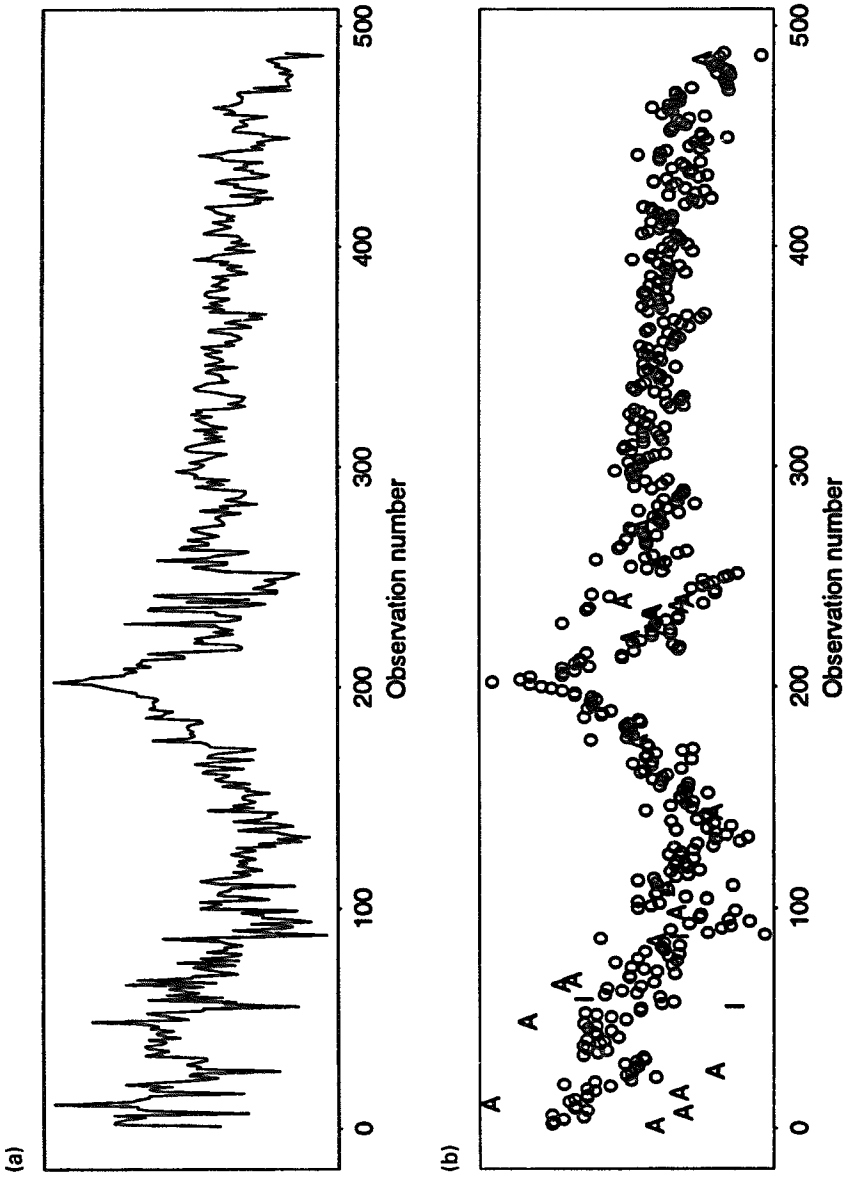

Fig. 1. (a) Log arsenic concentration; (b) plot of observations by selected type of outlier.

Table 2
Prior probabilities for additive and innovation outliers

$K_t = (K_{1t}, K_{2t})$	(0,1)	(3.3,1)	(10,1)	(32,1)	(0,3.3)	(0,10)	(0,32)
$P(K_t)$	0.90	0.04	0.009	0.001	0.04	0.009	0.001

Table 3
Prior probability that $\psi_j \neq 0$ for $j = 1, \ldots, 10$

j	1	2	3	4	5	6	7	8	9	10
$P(\psi_j \neq 0)$	0.90	0.85	0.80	0.75	0.70	0.65	0.60	0.55	0.50	0.45

posterior probabilities of additive and innovation outliers at each time point. The two most frequently chosen models are: (i) ψ_1, ψ_2 and ψ_4 nonzero and the rest of the ψ_j equal to zero, which is selected 20% of the time; (ii) $\psi_1, \ldots, \psi_4$ nonzero and the rest equal to zero, which is selected 14% of the time. An examination of the posterior means and standard deviations of $\psi_5, \ldots, \psi_{10}$ suggested that these coefficients are not significant and that an autoregressive model of order 4 is sufficient. Table 4 gives the posterior means and standard deviations of μ, $\phi_1, \ldots, \phi_4$, and σ, and compares them to maximum likelihood estimates both for the original data and data that was 'cleaned' of outliers. The results show that the maximum likelihood parameter estimates for the 'cleaned' data are much closer to the Bayesian estimates than for the original data. To obtain the non-Bayesian estimates we first fitted an autoregression of order 4 to the log arsenic data using maximum likelihood. The parameter estimates are given in column 2 of Table 4 with standard errors in parentheses. Next, we 'cleaned' the data by replacing each data point that was visually identified as an outlier by the average of its two nearest nonoutlier neighbours, one on each side of the point. The sample partial autocorrelations of the 'cleaned' data suggested fitting an autoregression of order 4, which is the same order suggested by the Bayesian analysis and differs from the sixth-order autoregression suggested by the autocorrelations of the original data. We fitted a fourth-order autoregression to the 'cleaned' data with the maximum likelihood estimates given in column 3 of Table 4. Finally, column 4 of Table 4 gives the Bayesian estimates. Table 4 shows that the maximum likelihood estimates of the 'cleaned' data are closer to the estimated posterior means than the parameter estimates of the original data. In particular, the maximum likelihood estimates of σ for the 'cleaned' data is virtually the same as the posterior mean of σ. Several other runs with different starting values gave similar results suggesting that we have converged to the whole posterior distribution and not just a local mode.

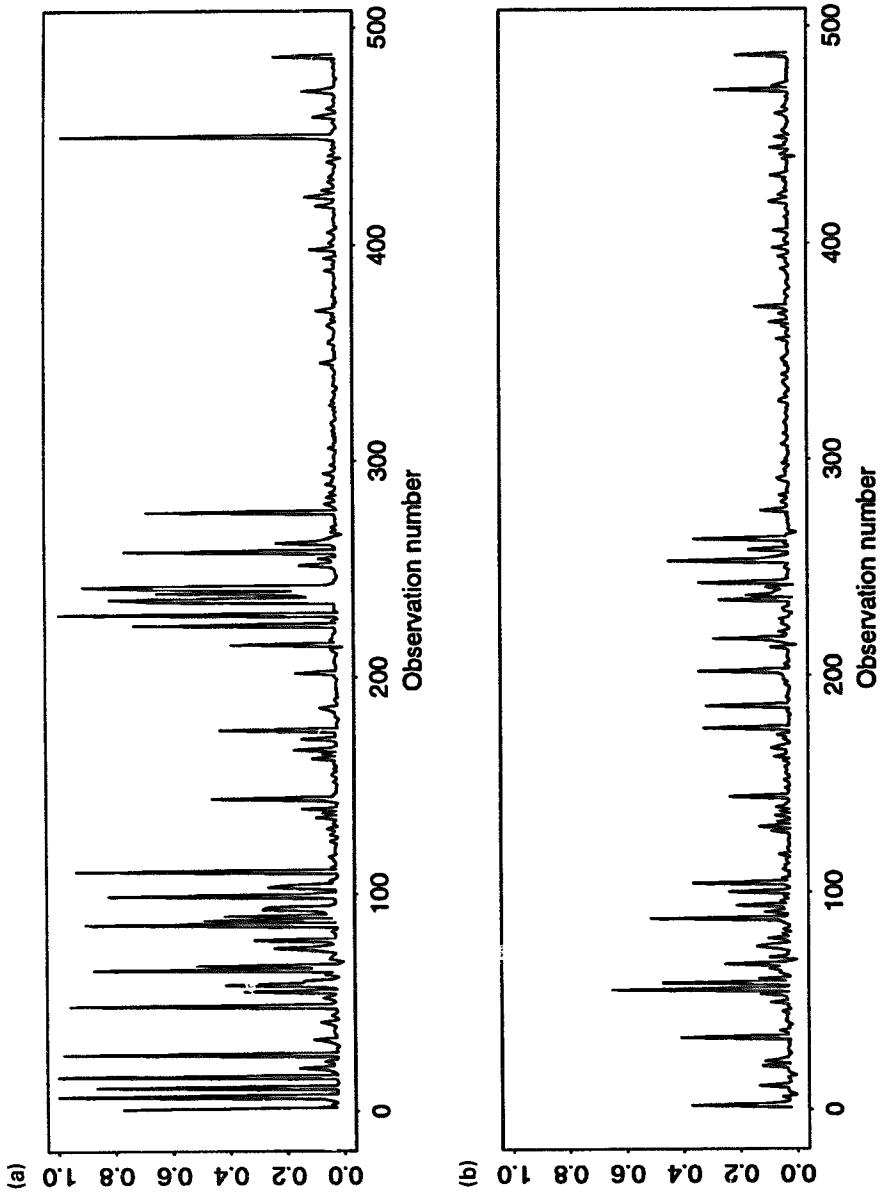

Fig. 2. (a) Posterior probability of an additive outlier; (b) posterior probability of an innovation outlier.

Table 4
Comparison of maximum likelihood and Bayesian parameter estimates for the arsenic data

	Maximum likelihood estimate (standard error)		Bayesian estimate (standard error)
	Original data	Clean data	
μ	0.8756	0.8687	0.8825 (0.0224)
ϕ_1	0.4295 (0.0452)	0.6109 (0.0449)	0.5653 (0.0598)
ϕ_2	0.2166 (0.0491)	0.1206 (0.0527)	(0.1308 (0.0615)
ϕ_3	0.0879 (0.0491)	0.0335 (0.0527)	− 0.0782 (0.0766)
ϕ_4	0.1347 (0.0452)	0.1620 (0.0449)	0.2096 (0.0711)
σ	0.045	0.032	0.032

Appendix: Implementing Sampling Scheme 1

Details are now given on how to implement Sampling Scheme 1 in Section 3. Part (a) of the lemma below shows how to obtain ϕ from ψ and is given by Barndorff-Nielsen and Schou (1973). Part (b) expresses ϕ_i as a linear function of ψ_k given $\psi_{j\neq k}$ and seems to be new. Its proof follows directly from part (a) and is omitted.

Lemma. (a) *The autoregressive coefficients ϕ are obtained from the partials ψ as follows: for $i = 1,\ldots,p$, let $\phi^{ij} = \phi^{i-1,j} - \psi_i\phi^{i-1,i-j}$ for $i > 1$ and $j = 1,\ldots,i-1$, $\phi^{ii} = \psi_i$. Then $\phi_i = \phi^{pi}$ for $i = 1,\ldots,p$.*

(b) *For given k, $\phi_1,\ldots,\phi_p$ are linear functions of ψ_k, i.e., $\phi_i = a_i + \psi_k b_i$ with a_i and b_i functionally independent of ψ_k. The coefficients a_i and b_i are obtained as follows: if $i < k$, let $a^{ii} = \psi_i$, $b^{ii} = 0$, $a^{ij} = a^{i-1,j} - \psi_i a^{i-1,i-j}$, and $b^{ij} = 0$ for $j = 1,\ldots,i-1$; if $i = k$, let $a^{ii} = 0$, $b^{ii} = 1$, $a^{i,j} = a^{i-1,j}$ and $b^{ij} = -a^{i-1,i-j}$; if $i > k$, let $a^{ii} = \psi_i$, $b^{ii} = 0$, $a^{ij} = a^{i-1,j} - \psi_i a^{i-1,i-j}$, and $b^{ij} = b^{i-1,j} - \psi_i b^{i-1,i-j}$ for $j = 1,\ldots,i-1$. Then, $a_i = a^{pi}$ and $b_i = b^{pi}$ for $i = 1,\ldots,p$.* □

Generating J_1, ψ, J_2, and Ψ. It is sufficient to show how to generate ψ_j and J_{1j} as Ψ_j and J_{2j} are generated similarly. Let $R = \{\psi_{i\neq j}, J_{1,i\neq j}, \Psi, J_2, K, O, \sigma^2, \mu\}$ and suppose that $J_{1j} = 1$. To show that

$$f(J_{1j},\psi_j|Y,Y^-,R) \propto \tfrac{1}{2} f(J_{1j}) f(Y|Y^-,R,J_{1j},\psi_j) f(u^-|R,J_{1j},\psi_j), \tag{A.1}$$

we write

$$f(J_{1j},\psi_j|Y,Y^-,R;\kappa)=f(Y|Y^-,R,J_{1j},\psi_j;\kappa)f(Y^-|R,J_{1j},\psi_j;\kappa)$$
$$\times\tfrac{1}{2}f(J_{1j})/f(Y,Y^-R;\kappa), \tag{A.2}$$

and $f(Y^-|R,J_{1j},\psi_j;\kappa)=f(u^-|R,J_{1j},\psi_j)f(w^-|u^-,R,J_{1j},\psi_j)$. By Kohn and Ansley (1986), $\kappa^{D/2}f(Y|Y^-,R;\kappa)$ and $\kappa^{D/2}f(w^-|u^-,R,\psi_j,J_{1j};\kappa)$ tend to finite nonzero limits as $\kappa\to\infty$ with the limits independent of ψ_j, J_{1j}. Eq. (A.1) is obtained by taking the limit as $\kappa\to\infty$ in (A.2).

Let $\varepsilon_t=\Phi(B^s)[(1-B)^d(1-B^s)^{d_s}w_t-\mu]$ and put $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_n)$ and $\varepsilon^-=(\varepsilon_0,\ldots,\varepsilon_{1-p})$. Then

$$f(Y|Y^-,R,J_{1j},\psi_j)\propto f(\varepsilon|\varepsilon^-,\psi,J_1,K)\propto e^{-S(\psi_j)/2\sigma^2},$$

where $S(\psi_j)=\sum_{t=1}^{n}(\varepsilon_t-\phi_1\varepsilon_{t-1}-\cdots-\phi_p\varepsilon_{t-p})^2/K_{2t}$. From above, $\phi_i=a_i+\psi_jb_i$ for all i. Let $v_t=\varepsilon_t-\sum_i a_i\varepsilon_{t-i}$, $z_t=\sum_i b_i\varepsilon_{t-i}$, $A=\sum_t z_t^2/K_{2t}$, $B=\sum_t v_tz_t/K_{2t}$, and $C=\sum_t v_t^2/K_{2t}$. Then $S(\psi_j)=A(\psi_j-B/A)^2+C-B^2/A$, so that from (A.2)

$$P(J_{1j}=1|R)\propto\tfrac{1}{2}f(J_{1j}=1)e^{-(C-B^2/A)/2\sigma^2}$$
$$\times\int_{-1}^{1}e^{A(\psi_j-B^2/A)^2/2\sigma^2}f(u^-|R,J_{1j},\psi_j)d\psi_j.$$

The integral is quickly and accurately evaluated using 11-point Gauss–Legendre integration. Similarly,

$$f(J_{1j}=0|R)\propto f(J_{1j}=0)e^{-S(\psi_j=0)/2\sigma^2},$$

so that $f(J_{1j}=1|R)$ can be obtained and J_{1j} generated. If $J_{1j}=0$, then $\psi_j=0$. If $J_{1j}=1$, then

$$f(\psi_j|Y,Y^-,R,J_{1j}=1)\propto e^{A(\psi_j-B^2/A)^2/2\sigma^2}f(u^-|R,\psi_jJ_{1j}).$$

The parameter ψ_j is generated using the Meteropolis-Hastings algorithm (Hastings, 1970) by first generating ψ_j from the normal distribution with mean B^2/A and variance σ^2/A, with ψ_j constrained to $(-1,1)$ as in Devroye (1986, p. 38, Problem 10). Let ψ_j^c and ψ_j^n be the current and new values of ψ_j and put $\alpha=\min(1,f(u^-|R,J_{1j},\psi_j^n)/f(u^-|R,J_{1j},\psi_j^c))$. Then ψ_j^n is accepted with a probability of α and ψ_j^c is kept otherwise. The Metropolis–Hastings step is computationally very efficient as the dimension of u^- is $p+sq$, which is independent of the sample size n. The low dimension also suggests that the rejection rate will be low and we have found this to be the case in practice.

Generating K and O. Define the polynomial $\gamma(B)=\phi(B)\Phi(B^s)\times(1-B)^d(1-B^s)^{d_s}$ with coefficients $\gamma_1,\ldots,\gamma_{D+Q}$. We show how to generate K_t

and O_t as a block. Consider first the case $K_{1t}>0$. Let $R=\{o_{s\neq t}, K_{s\neq t},\psi,J_1,\Psi,J_2,\sigma^2,\mu\}$. Then

$$f(o_t,K_t|Y,Y^-,R)\propto f(Y|Y^-,R,K_t,o_t)f(o_t|K_{1t})f(K_t)$$

$$\propto (K_{1t}K_{2t})^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\frac{o_t^2}{K_{1t}}+S(o_t)\right)\right\}, \tag{A.3}$$

where

$$S(o_t)=\sum_{j=t}^{\min(n,t+Q+D)}\left(y_j-o_j-\sum_{i=1}^{Q+D}\gamma_i(y_{j-i}-o_{j-i})\right)^2\Big/K_{2j}.$$

Let $v_t=y_t-\sum_i\gamma_i w_{t-i}$, $z_t=1$; let $v_j=w_j-\sum_{i\neq j-t}\gamma_i w_{j-i}-\gamma_{j-t}y_t$ and $z_j=-\gamma_{j-t}$ for $j=t+1,\ldots,\min(n,t+Q+D)$; let A,B, and C be defined with respect to v_j and z_j as above. Then $S(o_t)=Ao_t^2-2Bo_t+C$ and $o_t^2/K_{1t}+S(o_t)=A'(o_t-B^2/A')^2+C-B^2/A'$, where $A'=A+1/K_{1t}$. Integrating (A.3) with respect to o_t we obtain

$$f(K_t|Y,Y^-,R)\propto (K_{1t}K_{2t})^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left(C-\frac{B^2}{A'}\right)\right\}(2\pi A'\sigma^2)^{-1/2}.$$

Similarly, if $K_{1t}=0$, then

$$f(K_t|Y,Y^-,R)\propto K_{2t}^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}S(o_t=0)\right\},$$

and so we can evaluate $f(K_t|Y,Y^-,R)$ for all values of K_t. It is now straightforward to generate from $f(K_t|Y,Y^-,R)$ because it is multinomial. The error o_t is generated from $f(o_t|Y,Y^-,R,K_t)$ which is normal with mean B^2/A' and variance σ^2/A'.

It is straightforward to generate σ^2 and μ from inverse gamma and normal distributions respectively and we omit details.

Generating u^-, w^-, and Y^M. Let $R=\{\psi,J_1,\Psi,J_2,K,O,\sigma^2,\mu\}$. Then

$$f(u^-,w^-,Y^M|Y^O,R;\kappa)\propto f(Y|u^-,w^-,R)f(u^-|R)f(w^-|u^-,R;\kappa)/f(Y^O|R;\kappa).$$

Letting $\kappa\to\infty$ we obtain $f(u^-,w^-,Y^M|Y^O,R)\propto f(Y|u^-,w^-,R)f(u^-|R)$, which is Gaussian in u^-, w^-, and Y^M.

References

Albert, J.H. and S. Chib, 1993, Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts, Journal of Business and Economic Statistics 11, 1–15.

Barndorff-Nielson, O.E. and G. Schou, 1973, On the reparameterization of autoregressive models by partial autocorrelations, Journal of Multivariate Analysis 3, 408–419.

Box, G.E.P. and G.M. Jenkins, 1976, Time series: Analysis, forecasting and control (Holden Day, San Francisco, CA).

Chen, C. and L. Liu, 1993, Joint estimation of model parameters and outlier effects in time series, Journal of the American Statistical Association 88, 284–297.

Chib, S., 1993, Bayes regression with autoregressive errors, Journal of Econometrics 58, 275–294.

Chib, S. and E. Greenberg, 1994, Bayes inference in regression models with ARMA(p,q) errors, Journal of Econometrics 64, 183–206.

Devroye, L., 1986, Non-uniform random variate generation (Springer-Verlag, New York, NY).

Gelfand, A.E. and A.F.M. Smith, 1990, Sampling-based approaches to calculating marginal densities, Journal of the American Statistical Association 85, 398–409.

Gelman, A. and D. Rubin, 1992, Inference from iterative simulation using multiple sequences, Statistical Science 7, 457–511.

George, E.I. and R.E. McCulloch, 1993, Variable selection via Gibbs sampling, Journal of the American Statistical Association 88, 881–889.

George, E.I. and R.E. McCulloch, 1994, Fast Bayes variable selection, Preprint.

Hastings, W.K., 1970, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57, 97–119.

Hong, C., 1989, Forecasting real output growth rates and cyclical properties of models: A Bayesian approach, Unpublished Ph.D. thesis (University of Chicago, Chicago, IL).

Jacquier, E., N.G. Polson, and P.E. Rossi, 1994, Bayesian analysis of stochastic volatility models, Journal of Business and Economic Statistics 12, 371–412.

Kleiner, B., R.D. Martin, and D.J. Thompson, 1979, Robust estimation of power spectra, Journal of the Royal Statistical Society B 41, 313–351.

Kohn, R. and C.F. Ansley, 1986, Estimation, prediction and interpolation for ARIMA models with missing data. Journal of the American Statistical Association 81, 751–761.

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

Marriott, J. and A.F.M. Smith, 1992, Reparameterization aspects of numerical Bayesian methods for autoregressive moving average models, Journal of Time Series Analysis 13, 327–343.

Marriott, J., N. Ravishanker, A. Gelfand, and J. Pai, 1996, Bayesian analysis of ARMA processes: Complete sampling based inference under full likelihoods, in: D. Berry, K. Chaloner and J. Geweke, eds., Bayesian statistics and econometrics: Essays in honor of Arnold Zellner (Wiley, New York, NY).

McCulloch, R.E. and R.S. Tsay, 1994, Bayesian analysis of autoregressive time series via the Gibbs sampler, Journal of Time Series Analysis 15, 235–250.

Monahan, J., 1984, Full Bayesian analysis of ARMA time series models, Journal of Econometrics 21, 307–331.

Phillips, P.C.B., 1991, To criticize the critics: An objective Bayesian analysis of stochastic trends, Journal of Applied Econometrics 6, 333–364.

Tierney, L., 1994, Markov chains for exploring posterior distributions, Annals of Statistics 22, 1701–1762.

Tsay, R.S., 1988, Outliers, level shifts and variance changes in time series, Journal of Forecasting 7, 1–20.

Zellner, A., 1971, An introduction to Bayesian inference in econometrics (Wiley, New York, NY).