ELSEVIER

Journal of Econometrics 64 (1994) 183–206

JOURNAL OF Econometrics

Bayes inference in regression models with ARMA (p, q) errors

Siddhartha Chib*[,a], Edward Greenberg[b]

[a]*John M. Olin School of Business, Washington University, St. Louis, MO 63130, USA*
[b]*Department of Economics, Washington University, St. Louis, MO 63130, USA*

(Received April 1993; final version received August 1993)

Abstract

We develop practical and exact methods of analyzing ARMA(p, q) regression error models in a Bayesian framework by using the Gibbs sampling and Metropolis–Hastings algorithms, and we prove that the kernel of the proposed Markov chain sampler converges to the true density. The procedures can be applied to pure ARMA time series models and to determine features of the likelihood function by choosing appropriate diffuse priors. Our results are unconditional on the initial observations. We also show how the algorithm can be further simplified for the important special cases of stationary AR(p) and invertible MA(q) models. Recursive transformations developed in this paper to diagonalize the covariance matrix of the errors should prove useful in frequentist estimation. Examples with simulated and actual economic data are presented.

Key words: Gibbs sampling; Metropolis–Hastings algorithm; Data augmentation; Time series; ARMA processes; Markov chain; Bayesian statistics
JEL classification: C11; C15; C22

1. Introduction

Regression models with correlated errors have been the focus of considerable attention in econometrics and statistics. Although textbook presentations

* Corresponding author.

We have benefited from the comments of William Bell, Stefan Mittnik, Wilhelm Neuefeind, an associate editor, and two anonymous referees. This is a revision of 'Bayes Inference via Gibbs Sampling in Regression Models with AR(p) and MA(q) Errors', April 2, 1992.

SSDI 0304407693015838

usually restrict attention to autoregressive (AR) and moving average (MA) models, the latter often of the first order, the mixed autoregressive and moving average (ARMA) model is clearly the most interesting case. Unfortunately, the unconditional likelihood function for the general stationary and invertible ARMA(p, q) error model is quite complicated and can present serious computational problems. Therefore, despite the approaches to maximum likelihood estimation developed in Newbold (1974), Pagan and Nicholls (1976), Box and Jenkins (1976), Ansley (1979), and Gardner et al. (1979), software packages are organized around the method of nonlinear least squares or its equivalent, the conditional maximum likelihood (Harvey, 1981). Another line of inquiry has been directed at feasible generalized least squares estimators, most notably in Otto et al. (1987) and Galbraith and Zinde-Walsh (1992).

Absent to a large extent from this literature is the Bayesian analysis of regression models with ARMA(p, q) errors. Although a Bayesian perspective for time series has been actively pursued, a full treatment for such models is not available. Much of the early work is concentrated on autoregressive models (see Zellner, 1971), while the later work on mixed ARMA models was spurred by the approach of Monahan (1983), which is most useful for low-order processes. Broemeling and Shaarway (1984) enlarge the scope of Bayesian time series analysis by conditioning on initial values of pre-sample errors and other simplifications that replace the unknown errors appearing in the likelihood function with estimates obtained by nonlinear least squares.

In recent years many of the perceived difficulties of implementing the Bayesian paradigm have effectively disappeared through the emergence of Markov chain Monte Carlo (MCMC) simulation methods such as the Gibbs sampler (see Tanner and Wong, 1987; Gelfand and Smith, 1990) and Metropolis–Hastings (MH) algorithms (see Metropolis et al., 1953; Hastings, 1970; and Tierney, 1993). These methods are powerful tools for simulating intractable joint distributions that rely on the convergence of a suitably constructed Markov chain to the joint distribution of interest. The output of the simulation is a sample of draws that can be used for various purposes, for example to compute posterior moments and quantiles. The value of these methods for operationalizing Bayesian inference for time series regression, especially with autoregressive processes conditioned on initial observations, was recognized early by Chib (1993), McCulloch and Tsay (1993), and Albert and Chib (1993). In this paper we continue this line of attack but focus on a more general class of models, namely, regression models, perhaps with lagged dependent variables, whose errors follow a stationary and invertible ARMA(p, q) process of any specified order. Furthermore, our results are unconditional on the initial observations.

To put our work in perspective, recall that the quest in Gibbs sampling is to express the joint posterior density of the parameters in a form that lends itself to simulation, usually over a block of parameters at a time, conditioned on the remaining blocks. Achieving this in the current context necessitates the use of

several related strategies and the development of several new results. First, we introduce a set of additional parameters into the simulation, an example of data augmentation. These variables are not the $p + q$ pre-sample errors that are used to define the conditional likelihood, but rather $m = \max(p, q + 1)$ functions of these errors obtained from the state space representation of the model. Second, we show that two transformations of the data can be separately used to diagonalize the covariance matrix of the error. From the transformed observations we obtain the full conditional distributions of the regression parameters, the autoregressive coefficients, and the error variance. Third, we combine Markov chain strategies, as has been done in prior work by Müller (1993), but with a different class of candidate-generating densities. Fourth, we obtain the full conditional distribution of the transformed pre-sample errors by Kalman smoothing. Fifth, we specialize the analysis for AR(p) and MA(q) models and show that much of the analysis can be simplified. Finally, we formally prove the convergence of the MCMC algorithm to the desired joint posterior distribution of the parameters. In the proof, we establish a result of independent interest that states that the set of parameters that lead to stationarity and invertibility is arc-connected.

In concurrent and independent work Marriott et al. (1992) develop a different approach to the estimation of ARMA models that is based on sampling functions of the partial autocorrelations. A virtue of their approach is that one-for-one draws of each partial autocorrelation can be obtained but at the cost of a more complex algorithm. For the most part, that paper focuses on the important data analytic issues related to forecasting, missing values, and model adequacy. In contrast, we explicitly allow for a regression structure, derive exact forms for complete conditional distributions, conduct the sampling with blocks of parameters to improve the convergence of the Markov chain, and verify formal convergence conditions for our proposed algorithm.

The plan of this paper is as follows. Section 2 presents the model and the prior distributions. Section 3 contains the transformations mentioned above, the full conditional distributions, the details of the MCMC algorithm, and a theorem that the algorithm converges. Section 4 takes up the AR(p) and MA(q) special cases. Several numerical examples based on simulated and actual data are presented in Section 5, while Section 6 contains concluding remarks. The Appendix contains a proof of Proposition 2.

2. Model and prior assumptions

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

$$y_t = \boldsymbol{x}_t'\beta + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{1}$$

where x_t is a $k \times 1$ vector of covariates, β is the $k \times 1$ vector of regression parameters, and ε_t is a random error. Suppose that ε_t follows an ARMA(p, q) process

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \cdots + \phi_p \varepsilon_{t-p} + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}, \tag{2}$$

which is expressed in terms of a polynomial in the backshift operator L as

$$\phi(L)\varepsilon_t = \theta(L)u_t, \tag{3}$$

where $\phi_p \neq 0, \theta_q \neq 0, u_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \sigma^2 > 0, \mathcal{N}$ denotes the normal distribution, $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$, and $\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$. Equivalently, the model in (1) and (2) can be expressed in state space form (see Harvey, 1981) as follows:

$$y_t = x_t'\beta + z'\alpha_t \tag{4}$$

$$\alpha_t = G\alpha_{t-1} + fu_t, \tag{5}$$

where $z = (1, 0, \ldots, 0)'$: $m \times 1$, $\alpha_t = (\alpha_{1t}, \ldots, \alpha_{mt})'$: $m \times 1$, $m = \max(p, q+1)$,

$$G = \begin{pmatrix} \phi_1 & \vdots & & & \\ \phi_2 & \vdots & & & \\ \phi_3 & \vdots & & I_{m-1} & \\ \vdots & \vdots & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \phi_m & \vdots & 0 & \cdots & 0 \end{pmatrix} : \quad m \times m,$$

and $f = (1, \theta_1, \ldots, \theta_m)'$. In writing G and f we employ the conventions that $\phi_s = 0$ for $s > p$, $\theta_r = 0$ for $r > q$, and $\theta_0 = 1$. We make the following assumptions:

Assumption M (Model): The data $y = (y_1, \ldots, y_n)$ are generated by (1) and (2), with p and q known.

Assumption S (Stationarity): All roots of $\phi(L)$ lie outside the unit circle.

Assumption I (Invertibility): All roots of $\theta(L)$ lie outside the unit circle.

Assumption P (Prior distributions):

$$\begin{aligned}[\beta, \phi, \theta, \sigma^2, \alpha_0] &= [\beta][\phi][\theta][\sigma^2][\alpha_0|\beta, \phi, \theta, \sigma^2] \\ &= \mathcal{N}_k(\beta|\beta_0, B_0^{-1})\mathcal{N}_p(\phi|\phi_0, \Phi_0^{-1})I_{S_\phi}\mathcal{N}_q(\theta|\theta_0, \Theta_0^{-1})I_{S_\theta} \\ &\quad \times \mathcal{IG}(\sigma^2|v_0/2, \delta_0/2)[\alpha_0|\beta, \phi, \theta, \sigma^2],\end{aligned} \tag{6}$$

where $\phi = (\phi_1, \ldots, \phi_p)'$, $\theta = (\theta_1, \ldots, \theta_q)'$, $\mathcal{N}_s(\cdot)$ is the s-variate normal distribution, $\mathcal{IG}(\cdot)$ is the inverted gamma distribution, I_A is the indicator function of the set A, S_ϕ is the set of ϕ that satisfies Assumption S, and S_θ is the set of θ that satisfies Assumption I. The hyperparameters β_0, B_0, ϕ_0, Φ_0, θ_0, Θ_0, ν_0, and δ_0 are known.

A few comments about these assumptions are in order. Assuming stationarity does not limit in any important way the ability to model nonstationary data, since $\boldsymbol{x}_t$ may contain lagged values of y_t whose coefficients are unrestricted. In the absence of lagged y_t, the variables in $\boldsymbol{x}_t$ may be (unit-root) nonstationary in which case Assumption S amounts to an assertion that y and $\boldsymbol{x}$ are cointegrated. The long-run relation between a nonstationary y and its covariates would otherwise break down; y in effect would be a pure time series process, and there would be no interest in estimating β. Assumption I is introduced for identification purposes. With respect to Assumption P, it will be noted that the usual normal-inverted gamma distribution has been assumed for β and σ^2, while those for ϕ and θ are multivariate normal truncated to their stationary and invertible regions, respectively. Vague prior information can be entertained by centering these distributions at zero and setting each prior precision matrix equal to ε times an identity matrix, where ε is a small number. A highly informative prior (large precision) on a parameter can take the place of a constraint, thereby permitting the analysis of seasonal ARMA processes. For the initial state vector, the stationarity assumption implies that α_0, conditioned on β, ϕ, θ, and σ^2, follows a normal distribution with parameters $\mathrm{E}(\alpha_0) = 0$ and $\mathrm{E}(\alpha_0 \alpha_0') = \Omega$, where

$$\mathrm{vec}(\Omega) = \sigma^2 (I - G \otimes G)^{-1} \mathrm{vec}(\boldsymbol{f}\boldsymbol{f}'). \tag{7}$$

Finally, the normal and truncated normal priors that we assume are defensible on several grounds, primarily analytical tractability and flexibility. Nonetheless, if desired a different class of priors can be employed because it is possible to sample from nonstandard distributions (as we do below for θ) within a Markov chain algorithm by employing the Metropolis–Hastings algorithm. Equally important, output corresponding to different prior distributions, for example those not in the above class, can be obtained by a weighted bootstrap applied to the sampled draws. We remark further on these points below.

3. Main results

3.1. A preliminary result

The goal of the paper is to determine moments and other features of the posterior distribution of $\psi = (\beta, \phi, \theta, \sigma^2)$ under Assumptions M, S, I, and P. By

Bayes theorem, the posterior density is given by $f(\psi|\boldsymbol{y}) \propto \pi(\psi) f(\boldsymbol{y}|\psi)$, where $\pi(\psi)$ is the prior density and $f(\boldsymbol{y}|\psi)$ is the likelihood function. The direct calculation of the exact likelihood function is intractable. It is well known, however, that given the pre-sample errors $\lambda = (\varepsilon_0, \ldots, \varepsilon_{-p+1}, u_0, \ldots, u_{-q+1})$, the density of $\boldsymbol{y}$ given (ψ, λ) can be expressed as

$$f(\boldsymbol{y}|\psi, \lambda) = \prod_{t=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2} u_t^2\right]$$

$$= \prod_{t=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2} (y_t - \hat{y}_{t|t-1})^2\right], \tag{8}$$

where $\hat{y}_{t|t-1} = \boldsymbol{x}_t'\beta + (\phi(L) - 1)(y_t - \boldsymbol{x}_t'\beta) + (\theta(L) - 1)u_t$ is the one-step-ahead prediction of y_t given information up to time $t-1$.[1] We therefore develop an approach that relies on (8), but nevertheless provides the posterior density for the exact likelihood.

First, we show that the conditional likelihood can be expressed in terms of only m pre-sample variables, not all the $p+q$ elements in λ. This surprising result is actually a consequence of the state space form of the ARMA model and appears to have been overlooked in the literature. Our demonstration begins by considering the period $t = 1$. Then from (8),

$$\hat{y}_{1|0} = \boldsymbol{x}_1'\beta + \phi_1\varepsilon_0 + \cdots + \phi_p\varepsilon_{-p+1} + u_0 + \theta_1 u_0 + \cdots + \theta_q u_{-q+1}.$$

By solving the state space form for α_1 from the bottom up, we find that

$$\phi_1\varepsilon_0 + \cdots + \phi_p\varepsilon_{-p+1} + \theta_1 u_0 + \cdots + \theta_q u_{-q+1} = \phi_1\alpha_{10} + \alpha_{20}.$$

Therefore, $\hat{y}_{1|0} = \boldsymbol{x}_1'\beta + \phi_1\alpha_{10} + \alpha_{20}$; i.e., the elements of λ enter only through rows of α_0. This is true for all values of t as the following argument proves. For $1 \leqslant t \leqslant p$, define $\phi_t(L) = 1 - \phi_1 L - \cdots - \phi_{t-1} L^{t-1}$. Then

$$\begin{aligned}\phi_t(L)(y_t - \boldsymbol{x}_t'\beta) &= \phi_t(L)z'\alpha_t = z'\phi_t(L)\alpha_t \\ &= z'(\alpha_t - \phi_1\alpha_{t-1} - \cdots - \phi_{t-1}\alpha_1) \\ &= \alpha_{1t} - \phi_1\alpha_{1,t-1} - \cdots - \phi_{t-1}\alpha_{11}.\end{aligned} \tag{9}$$

By repeatedly using the recursion $\alpha_{rt} = \phi_r\alpha_{1,t-1} + \alpha_{r+1,t-1} + \theta_{r-1}u_t$, (9) can be rewritten as

$$\phi_t(L)(y_t - \boldsymbol{x}_t'\beta) = \phi_t\alpha_{10} + \alpha_{t+1,0} + u_t + \theta_1 u_{t-1} + \cdots + \theta_{t-1}u_1,$$

[1] There is no need to introduce the pre-sample errors if the model does not contain a moving average component. In that case a direct approach can be based on the assumption that the first p observations come from the stationary distribution. This important special case, and our treatment of it, is described in Section 4.1.

from which it follows that

$$\hat{y}_{t|t-1} = \boldsymbol{x}_t'\beta + \phi_1(y_{t-1} - \boldsymbol{x}_{t-1}'\beta) + \cdots + \phi_{t-1}(y_1 - \boldsymbol{x}_1'\beta) + \theta_1 u_{t-1} + \cdots + \theta_{t-1}u_1 + \phi_t\alpha_{10} + \alpha_{t+1,0}. \tag{10}$$

Thus y_t $(1 \leqslant t \leqslant p)$ depends on λ only through α_0. Upon multiplying by $\phi(L)$, the same type of argument shows that

$$\hat{y}_{t|t-1} = \boldsymbol{x}_t'\beta + \sum_{i=1}^{p} \phi_i(y_{t-i} - x_{t-i}'\beta) + \theta_1 u_{t-1} + \cdots + \theta_{t-1}u_1 + \alpha_{t+1,0}, \quad t = p+1, \ldots, n, \tag{11}$$

where we have used the conventions that $\theta_j = 0$ $(j > q)$ and $\alpha_{r0} = 0$ $(r > m)$. Since pre-sample errors enter $\hat{y}_{t|t-1}$ only through rows of α_0, we have established the result that $f(\boldsymbol{y}|\psi, \lambda) = f(\boldsymbol{y}|\psi, \alpha_0)$. Therefore we include β, ϕ, θ, σ^2, and α_0 as elements in our MCMC algorithm, simulating these parameters from the following conditional densities: $\pi(\beta|\boldsymbol{y}, \psi_{-\beta}, \alpha_0)$, $\pi(\phi|\boldsymbol{y}, \psi_{-\phi}, \alpha_0)$, $\pi(\theta|\boldsymbol{y}, \psi_{-\theta}, \alpha_0)$, $\pi(\sigma^2|\boldsymbol{y}, \psi_{-\sigma^2}, \alpha_0)$, and $\pi(\alpha_0|\boldsymbol{y}, \psi)$, where, e.g., $\psi_{-\beta}$ denotes all parameters in ψ other than β. To derive these densities, we proceed by noting that each is proportional to the joint posterior density for the augmented parameter vector (ψ, α_0) given by

$$\pi(\psi, \alpha_0|\boldsymbol{y}) \propto \pi(\psi)\pi(\alpha_0|\psi)f(\boldsymbol{y}|\psi, \alpha_0), \tag{12}$$

where $f(\boldsymbol{y}|\psi, \alpha_0)$ is the conditional density of $\boldsymbol{y}$ (see below) and the other densities are taken from Assumption P. Simplifying (12) is the next order of business.

3.2. *Full conditional distributions*

Two results are central to the analysis that now follows. We show that the density $f(\boldsymbol{y}|\psi, \alpha_0)$ can be diagonalized by recursive transformations of the data to produce a regression relationship for β and ϕ. These simple recursions for the general ARMA problem have not appeared elsewhere and may be useful for frequentist estimation.

Definition 1. Let the scalars $y_s = y_s^* = 0$ *and the vectors* $\boldsymbol{x}_s = \boldsymbol{x}_s^* = 0$, $s \leqslant 0$, *and let* $\alpha_{r0} = 0$, $r > m$. *For* $t = 1, \ldots, n$, *define*

$$y_t^* = y_t - \sum_{s=1}^{p} \phi_s y_{t-s} - \sum_{i=1}^{q} \theta_i y_{t-i}^* - \phi_t\alpha_{10} - \alpha_{t+1,0},$$

$$\boldsymbol{x}_t^* = \boldsymbol{x}_t - \sum_{s=1}^{p} \phi_s \boldsymbol{x}_{t-s} - \sum_{i=1}^{q} \theta_i \boldsymbol{x}_{t-i}^*.$$

This definition implies the following lemma:

Lemma 1. Let $\boldsymbol{y}^*$ *be the* $n \times 1$ *vector of the* y_t^* *and let* X^* *be the* $n \times k$ *matrix with* $\boldsymbol{x}_t^{*\prime}$ *as its* t*th row. Then*

$$f(\boldsymbol{y}^* | \psi, \alpha_0) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{y}^* - X^*\beta)'(\boldsymbol{y}^* - X^*\beta) \right].$$

Proof. Verify that $y_1^* - \boldsymbol{x}_t^{*\prime}\beta = u_1$ and proceed by induction, making use of (10) and (11). ■

From the definition of y_t^* and its appearance in $f(\boldsymbol{y}^* | \psi, \alpha_0)$ we see how α_0 enters the conditional density. Moreover, the regression relationship $\boldsymbol{y}^* = X^*\beta + u$, where $u \sim \mathcal{N}_n(0, \sigma^2 I_n)$, immediately yields the full conditional distribution of β and σ^2. We continue by introducing a transformation that allows us to determine the full conditional distribution of ϕ.

Definition 2. For $s \leqslant 0$, *let the scalars* $y_s = \bar{y}_s = \bar{x}_s = 0$ *and the vectors* $\boldsymbol{x}_s = 0$, *and let* $\alpha_{r0} = 0$, $r > m$. *For* $t = 1, \ldots, n$, *define*

$$\bar{y}_t = y_t - \boldsymbol{x}_t'\beta - \sum_{i=1}^{q} \theta_i \bar{y}_{t-i} - \alpha_{t+1,0},$$

$$\bar{\boldsymbol{x}}_t = y_t - \boldsymbol{x}_t'\beta - \sum_{i=1}^{q} \theta_i \bar{\boldsymbol{x}}_{t-i}.$$

With this definition we can prove the following lemma:

Lemma 2. Let $\bar{\boldsymbol{y}}$ *be the* $n \times 1$ *column vector of the* $\bar{y}_t$ *and let* $\bar{X}$: $n \times p$ *be given by*

$$\bar{X} = \begin{pmatrix} \alpha_{10} & 0 & \cdots & \cdots & 0 \\ \bar{x}_1 & \alpha_{10} & 0 & \cdots & 0 \\ \bar{x}_2 & \bar{x}_1 & \alpha_{10} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{p-1} & \bar{x}_{p-2} & \cdots & \cdots & \alpha_{10} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{n-1} & \bar{x}_{n-2} & \bar{x}_{n-3} & \cdots & \bar{x}_{n-p} \end{pmatrix}.$$

Then

$$f(\bar{\boldsymbol{y}} | \psi, \alpha_0) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} (\bar{\boldsymbol{y}} - \bar{X}\phi)'(\bar{\boldsymbol{y}} - \bar{X}\phi) \right].$$

Proof. Verify that $\bar{y}_1 - \bar{X}'_1\beta = u_1$, where $\bar{X}'_1$ is the first row of $\bar{X}$, and proceed by induction, making use of (10) and (11). ■

A corollary of this result is that $\bar{y} = \bar{X}\phi + u$, where $u \sim \mathcal{N}_n(0, \sigma^2 I_n)$.

At this point, we introduce notation for Proposition 1, which is presented below. We let $B_n = B_0 + \sigma^{-2}X^{*\prime}X^*$, $\Phi_n = \Phi_0 + \sigma^{-2}\bar{X}'\bar{X}$, and define the function $p(\phi, \theta, \sigma^2) = (\sigma^2)^{-m/2}|\Omega(\phi, \theta)|^{-1/2}\exp[-(1/2\sigma^2)\alpha_0'\Omega(\phi, \theta)^{-1}\alpha_0]$, which is the prior density $\pi(\alpha_0|\beta, \phi, \theta, \sigma^2)$. For a given value of (θ, σ^2), the latter function is denoted as $p_1(\phi)$, and for a given value of (ϕ, σ^2), it is denoted as $p_2(\theta)$. Also let $d_1 = \|y^* - X^*\beta\|^2$ and $d_2 = \alpha_0'\Omega(\phi, \theta)^{-1}\alpha_0$. Finally, $\hat{\alpha}_{0|n}$ and $R_{0|n}$ are the mean and covariance of the full conditional distribution of α_0, which are obtained from the recursions (see Harvey, 1981) $\hat{\alpha}_{t|n} = \hat{\alpha}_{t|t} + B_t(\hat{\alpha}_{t+1|n} - G\hat{\alpha}_{t|t})$, $R_{t|n} = R_{t|t} + B_t(R_{t+1|n} - R_{t+1|t})B_t'$, $t = n-1, n-2, \ldots, 0$, and $B_t = R_{t|t}GR_{t+1|t}^-$, $0 \leqslant t \leqslant n-1$, where $\hat{\alpha}_{t|s}$ and $R_{t|s}$ for $s \leqslant t$ are the forward filter estimates and $R_{t+1|t}^-$ is the Moore–Penrose inverse.[2]

We are now in a position to present the full conditional distributions that are used in the simulation for the regression model with ARMA(p, q) errors.

Proposition 1. Under Assumptions M, S, I, and P, the full conditional distributions for $\beta, \phi, \sigma^2, \alpha_0$, and θ are given by

(*i*) $\beta|y, \psi_{-\beta}, \alpha_0 \sim \mathcal{N}_k(B_n^{-1}(B_0\beta_0 + \sigma^{-2}X^{*\prime}y^*), B_n^{-1})$,

(*ii*) $\phi|y, \psi_{-\phi}, \alpha_0 \propto p_1(\phi) \times \mathcal{N}_p(\Phi_n^{-1}(\Phi_0\phi_0 + \sigma^{-2}\bar{X}'\bar{y}), \Phi_n^{-1})I_{S_\phi}$,

(*iii*) $\sigma^2|y, \psi_{-\sigma^2}, \alpha_0 \sim \mathcal{IG}((v_0 + n + m)/2, (\delta_0 + d_1 + d_2)/2)$,

(*iv*) $\alpha_0|y, \psi \sim \mathcal{N}_m(\hat{\alpha}_{0|n}, R_{0|n})$,

(*v*) $$\pi(\theta|y, \psi_{-\theta}, \alpha_0) \propto p_2(\theta) \times \prod_{t=1}^{n} \exp[-(1/2\sigma^2)u_t(\theta)^2] \times \exp[-\tfrac{1}{2}(\theta - \theta_0)'\Theta_0(\theta - \theta_0)]\, I_{S_\theta}.$$

Proof. (*i*) and (*iii*) follow from Assumption P and Lemma 1; (*ii*) follows from Assumption P and Lemma 2; (*iv*) follows from Assumption P and the definition of the Kalman smoothing recursions; and (*v*) follows from the definition of the full conditional distribution. ■

---

[2] The prior distribution of α_0 enters this expression through $R_{0|0} = \mathrm{cov}(\alpha_0)$, where the covariance is that of the prior distribution. The Moore–Penrose inverse is required because $R_{t+1|t}$ becomes singular for large t. Moreover, since $\alpha_{1t} = y_t - x_t'\beta$, $R_{t|t}$ is always singular. But of more importance is that $R_{t|t} \to 0$ as $t \to \infty$. This implies that not all of the n observations contain information about α_0 so that the filter can be terminated for large enough t.

In passing we mention that the terms $p_1(\phi)$, $p_2(\theta)$, and m and d_2 (in the inverted gamma distribution) arise from the prior on α_0 through its dependence on the parameters (ϕ, θ, σ^2).

3.3. *Implementation notes*

We have now shown in Proposition 1 that the full conditional distributions of β, σ^2, α_0 are straightforward to compute, belong to standard families of distributions, and are readily simulated. Evidently, the situation with ϕ and θ is more intricate and, therefore, a short digression is in order.

Metropolis–Hastings (MH) Algorithm: Suppose $p(z)$ is a density function of a multi-dimensional z that is to be simulated. An MCMC algorithm that produces a sample of draws from $p(\cdot)$ proceeds as follows. Suppose that $Z^{(i)}$ is the current draw in the chain. To obtain the next draw $Z^{(i+1)}$, first draw a candidate Z' from a suitable density $q(Z^{(i)}, z)$, which is called the candidate-generating density. The candidate draw is now subjected to a further randomization and is accepted with probability

$$\alpha(Z^{(i)}, Z') = \min\left\{\frac{p(Z')/q(Z^{(i)}, Z')}{p(Z^{(i)})/q(Z', Z^{(i)})}, 1\right\}.$$

If Z' is rejected, $Z^{(i+1)}$ is set equal to $Z^{(i)}$. This process is iterated. It should be noted that this procedure does not require the normalizing constant of $p(z)$. For more details see Tierney (1993).

Clearly, successful implementation of the MH algorithm, with a high acceptance rate of candidate draws, requires a suitable candidate-generating density. Fortunately, such densities are available for both ϕ and θ. Suppose we let $q(\phi^{(i)}, \phi)$ be the density of $\mathcal{N}_p(\Phi_n^{-1}(\Phi_0\phi_0 + \sigma^{-2}\bar{X}'\bar{y}), \Phi_n^{-1})I_{S_\phi}$, and let ϕ' be a draw from this distribution.[3] Then the Metropolis–Hastings step amounts to an acceptance–rejection of ϕ' with probability

$$\alpha(\phi^{(i)}, \phi') = \min\left\{\frac{p_1(\phi')}{p_1(\phi^{(i)})}, 1\right\}.$$

For θ, a suitable $q(\cdot, \cdot)$ density is the truncated normal approximation to $\pi(\theta|y, \psi_{-\theta}, \alpha_0)$ given by

$$q(\theta|y, \beta, \phi, \sigma^2, \theta^\dagger)$$
$$\equiv q(\theta) \propto \exp\left[-\tfrac{1}{2}[\theta - m(\theta^\dagger)]'V(\theta^\dagger)^{-1}[\theta - m(\theta^\dagger)]\right] I_{S_\theta}, \tag{13}$$

[3] A convenient strategy is to sample the untruncated normal and retain the drawing if it lies in S_ϕ. This strategy may be inefficient if the mass of the posterior is *not* concentrated over the stationary region, which may also indicate that the model is misspecified.

where $\theta^{\dagger}$ denotes the nonlinear least squares estimate of θ,

$$m(\theta^{\dagger}) = V(\theta^{\dagger})[\Theta_0\theta_0 + \sigma^{-2}W(\theta^{\dagger})'(u(\theta^{\dagger}) + W(\theta^{\dagger})\theta^{\dagger})],$$

$$V(\theta^{\dagger}) = [\Theta_0 + \sigma^{-2}W(\theta^{\dagger})'W(\theta^{\dagger})]^{-1},$$

$$u(\theta^{\dagger}) = y^*(\theta^{\dagger}) - X^*(\theta^{\dagger})\beta,$$

$$W(\theta^{\dagger}) = (\partial u(\theta)/\partial\theta')|_{\theta=\theta^{\dagger}},$$

the elements of which can be computed from the recursion (see also Fuller, 1976, p. 358)

$$W_{it} = \begin{cases} 0, & t \leqslant 0, \\ u_{t-i}(\theta^{\dagger}) - \sum_{j=1}^{q} \theta_j^{\dagger} W_{i,t-j}(\theta^{\dagger}), & t = 1, \ldots, n, \quad i = 1, \ldots, q. \end{cases}$$

We have suppressed the dependence of these expressions on y, β, ϕ, and σ^2 and defined $u_t(\cdot)$ as the tth row of $u(\cdot)$. Note that the density in (13) is obtained by expanding $u_t(\theta)$ around $\theta^{\dagger}$ as $u_t(\theta) \approx u_t(\theta^{\dagger}) - w_t'(\theta - \theta^{\dagger})$ (where w_t' is the tth row of W), substituting into (v), and combining terms. The MH acceptance–rejection probability is now easily defined.[4]

With these results, the sampling process can be run to obtain any desired number of draws. In this process, given the ith draw on $(\psi^{(i)}, \alpha_0^{(i)})$ the next draw is obtained by simulating β from $\beta|y, \phi^{(i)}, \theta^{(i)}, \sigma^{2(i)}, \alpha_0^{(i)}$; ϕ from $\phi|y, \beta^{(i+1)}, \theta^{(i)}, \sigma^{2(i)}, \alpha_0^{(i)}$; θ from $\theta|y, \beta^{(i+1)}, \phi^{(i+1)}, \sigma^{2(i)}, \alpha_0^{(i)}$; σ^2 from $\sigma^2|y, \beta^{(i+1)}, \phi^{(i+1)}, \theta^{(i+1)}, \alpha_0^{(i)}$; and α_0 from $\alpha_0|y, \beta^{(i+1)}, \phi^{(i+1)}, \theta^{(i+1)}, \sigma^{2(i+1)}$. Although the Markov chain generated by this process will converge to the target posterior distribution, as the next section demonstrates, practical monitoring of the sampling process, for example, via the methods of Gelman and Rubin (1993), Ritter and Tanner (1992), and Zellner and Min (1992), will be useful.

3.4. Convergence results

In this section we show that the Gibbs sampler presented above defines a Markov chain that converges as $M \to \infty$ to $\pi(\psi, \alpha_0 \mid y)$ in the L^1 norm and that sample moments of integrable functions of (ψ, α_0) converge almost surely to their expectations under the target density. To prove this result we utilize Theorem 2 of Roberts and Smith (1992) and Proposition 4 of Tierney (1993), which provide sufficient conditions for these results. Before proceeding further we define $T_\phi = S_\phi \cup S_\phi'$ and $T_\theta = S_\theta \cup S_\theta'$, where $S_\phi' = \{\phi\colon \phi_p = 0, z \neq 0$ and

[4] The tactic of combining Markov chain strategies has been successfully employed by, among others, Müller (1993) and Chib and Greenberg (1993a). Jacquier et al. (1992) sample all full conditionals with the MH algorithm in stochastic volatility models.

$\phi(z) = 0 \Rightarrow |z| > 1\}$ and $S'_\theta = \{\theta\colon \theta_p = 0,\ z \neq 0$ and $\theta(z) = 0 \Rightarrow |z| > 1\}$. We can now establish the following result:

Proposition 2. The posterior density of $(\psi, \alpha_0) = (\beta, \phi, \theta, \sigma^2, \alpha_0)$ *defined on the product set* $D = \Re^k \times T_\phi \times T_\theta \times \Re^+ \times \Re^m$ *satisfies the following properties*:

(*i*) $f(\psi, \alpha_0|\boldsymbol{y})$ *is lower semicontinuous at* 0, *i.e., it has the property that if* $f(\psi', \alpha'_0|\boldsymbol{y}) > 0$, *there exists an open neighborhood* $N_{(\psi', \alpha'_0)} \ni (\psi', \alpha'_0)$ *and* $\varepsilon > 0$ *such that, for all* $(\psi, \alpha_0) \in N_{(\psi', \alpha'_0)}$, $f(\psi, \alpha_0|\boldsymbol{y}) \geqslant \varepsilon > 0$.

(*ii*) $\int f(\psi, \alpha_0|\boldsymbol{y})\mathrm{d}\rho$ *is locally bounded, where* ρ *is any of the parameter vectors included in the Gibbs sampler.*

(*iii*) *The support* D *of* $f(\psi, \alpha_0|\boldsymbol{y})$ *is arc-connected.*

Proof. See Appendix.

Proposition 2 immediately implies the following:

Proposition 3. Let K *denote the transition density of the Markov chain defined by the Gibbs–MH algorithm and let* $K^{(N)}$ *denote the* N*th iterate of the kernel. Then for all* (ψ, α_0) *in* D *as* $N \to \infty$:

(*i*) $|K^{(N)} - \pi(\psi, \alpha_0|\boldsymbol{y})| \to 0$.

(*ii*) *For real-valued,* π*-integrable functions* g,

$$N^{-1} \sum_{i=1}^{N} g(\psi^{(i)}, \alpha_0^{(i)}) \to \int g(\psi, \alpha_0)\pi(\psi, \alpha_0|\boldsymbol{y})\mathrm{d}(\psi, \alpha_0).$$

This result follows from Theorem 2 of Roberts and Smith (1992) since the conditions of their theorem were verified in Proposition 2 and from Tierney (1993) since $p(\cdot)/q(\cdot, \cdot)$ in the MH step is bounded.

4. Special cases

Let us now consider how models with AR(p) or MA(q) errors could be estimated. One straightforward possibility is to specialize the ARMA(p, q) algorithm by directly imposing the restrictions, as we do in Section 5. For example, to fit an AR(p) model, we simply set $\theta_i = 0$ in Definitions 1 and 2, and apply Proposition 1 to simulate β, ϕ, σ^2, and α_0. Another possibility is to use specific algorithms that are optimized for these special cases. In fact, for the AR(p) error model it is not necessary to introduce α_0 at all, while in the case of MA(q) errors, $\lambda = (u_0, \ldots, u_{-q+1})$, instead of α_0, suffices. In both cases, the algorithms are simple enough that for the sake of completeness and importance of these models, it is worthwhile to present the details.

4.1. Regression with AR(p) errors

As mentioned above, there is no need to introduce α_0 if the error process does not contain any moving average components. Accordingly, suppose that the error follows a stationary AR(p) process $\varepsilon_t = \phi_1 \varepsilon_{t-1} + \cdots + \phi_p \varepsilon_{t-p} + u_t$, $u_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, and that the first p data points $\boldsymbol{y}_1 = (y_1, \ldots, y_p)'$ are drawn from the stationary distribution

$$\boldsymbol{y}_1 | \beta, \sigma^2, \phi \sim \mathcal{N}_p(X_1 \beta, \sigma^2 \Sigma_p), \tag{14}$$

where $\Sigma_p = \Phi \Sigma_p \Phi' + e_1(p) e_1(p)'$,

$$\Phi = \begin{bmatrix} \phi'_{-p} & \phi_p \\ I_{p-1} & \mathbf{0} \end{bmatrix},$$

$e_1(p) = (1, 0, \ldots, 0)'$ is the $p \times 1$ unit vector, and $\phi_{-p} = (\phi_1, \ldots, \phi_{p-1})'$. Define the following quantities (which are local to this subsection): $\boldsymbol{y}_1^* = Q^{-1} \boldsymbol{y}_1$ and $X_1^* = Q^{-1} X_1$, where X_1 contains the first p rows of X and Q satisfies the equation $QQ' = \Sigma_p$. Also, for $t \geq p+1$, define an $n-p$ vector $\boldsymbol{y}_2^*$ with tth element given by $\phi(L) y_t$ and an $n-p \times k$ matrix X_2^* with tth row $\phi(L) \boldsymbol{x}_t'$. In stacked form let $\boldsymbol{y}^* = (\boldsymbol{y}_1^{*\prime}, \boldsymbol{y}_2^{*\prime})'$ and likewise for X^*. Finally, let $e = (e_{p+1}, \ldots, e_n)'$ and let E denote the $n-p \times p$ matrix with tth row given by $(e_{t-1}, \ldots, e_{t-p})$, where $e_t = y_t - \boldsymbol{x}_t' \beta$, $t \geq p+1$.

Then it can be shown that the full conditional distributions are given by

$$\beta | \boldsymbol{y}, \psi_{-\beta} \sim \mathcal{N}_k(B_n^{-1}(B_0 \beta_0 + \sigma^{-2} X^{*\prime} \boldsymbol{y}^*), B_n^{-1}),$$

$$\phi | \boldsymbol{y}, \psi_{-\phi} \propto \Psi(\phi) \times \mathcal{N}_p(\hat{\phi}, \Phi_n^{-1}) I_{S_\phi},$$

$$\sigma^2 | \boldsymbol{y}, \psi_{-\sigma^2} \sim \mathcal{IG}((\nu_0 + n)/2, (\delta_0 + d_1)/2),$$

where $\hat{\phi} = \Phi_n^{-1}(\Phi_0 \phi_0 + \sigma^{-2} E' e)$, $\Phi_n = (\Phi_0 \phi_0 + \sigma^{-2} E' E)$, $d_1 = \| \boldsymbol{y}^* - X^* \beta \|^2$, and

$$\Psi(\phi) = |\Sigma_p(\phi)|^{-1/2} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{y}_1 - X_1 \beta)' \Sigma_p^{-1}(\phi) (\boldsymbol{y}_1 - X_1 \beta) \right].$$

It is easy to simulate from the conditional distributions for β and σ^2. To simulate ϕ we can apply a Metropolis–Hastings step since a natural candidate-generating density is available in $\mathcal{N}_p(\phi | \hat{\phi}, \hat{\Phi}^{-1}) I_{S_\phi}$. Therefore, after taking a draw ϕ' from the latter distribution we accept it as the next sample value with probability min $(\Psi(\phi')/\Psi(\phi^{(i)}), 1)$. If the candidate value is rejected, we stay at $\phi^{(i)}$. This concludes the MCMC algorithm for the regression model with stationary AR(p) errors.

4.2. Regression with MA(q) errors

Estimation of the moving average model can also be simplified. First, instead of α_0 we now only require the q elements in $\lambda = (u_0, u_{-1}, \ldots, u_{-q+1})'$. Second, the simulation of λ does *not* require the smoothing recursions employed in the general ARMA case, since the transformations defined in this section allow us to write the elements of λ as regression coefficients, after which it is easy to derive the full conditional distribution of β, λ, and σ^2.

Define the following transformation (the notation is local to this subsection):

$$y_t^* = y_t - \sum_{i=1}^{q} \theta_i y_{t-i}^*, \qquad \boldsymbol{x}_t^* = \boldsymbol{x}_t - \sum_{i=1}^{q} \theta_i \boldsymbol{x}_{t-i}^*,$$

where for $t \leqslant 0$ the scalars $y_t^* = 0$ and the vectors $\boldsymbol{x}_t^* = 0$. For $j = 1, \ldots, q$ and for $t = 1, \ldots, n$, let

$$v_{tj} = -\sum_{i=1}^{q} \theta_i v_{t-i,j} + \theta_{t+j-1},$$

where $v_{sj} = 0$ for $s < 0$, and set $\theta_r = 0$ for $r > q$. With this transformation, we show below that $y_t^* = \boldsymbol{x}_t^{*\prime}\beta + \sum_{j=0}^{q-1} v_{tj} u_{-j} + u_t$, or, in vector-matrix form,

$$\boldsymbol{y}^* = X^*\beta + V\lambda + u, \tag{15}$$

where $V = (v_{tj})$: $n \times q$. The full conditional distributions $[\beta, \lambda | \boldsymbol{y}, \sigma^2, \theta]$ and $[\sigma^2 | \boldsymbol{y}, \beta, \lambda, \theta]$ can be immediately derived. Finally, the simulation of θ can be achieved, as in Proposition 1, through an MH step.

We now prove (15). For $t > q$, by using

$$y_t^* = \boldsymbol{x}_t^{*\prime}\beta + u_t + v_{t1}u_0 + \cdots + v_{tq}u_{-q+1},$$

we may rewrite $y_t = \boldsymbol{x}_t'\beta + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q+1}$ as

$$\begin{aligned} y_t = \boldsymbol{x}_t'\beta + u_t &+ \theta_1[y_{t-1}^* - \boldsymbol{x}_{t-1}^{*\prime}\beta - v_{t-1,1}u_0 - \cdots - v_{t-1,q}u_{-q+1}] \\ &+ \theta_2[y_{t-2}^* - \boldsymbol{x}_{t-2}^{*\prime}\beta - v_{t-2,1}u_0 - \cdots - v_{t-2,q}u_{-q+1}] \\ &+ \cdots + \theta_q[y_{t-q}^* - \boldsymbol{x}_{t-q}^{*\prime}\beta - v_{t-q,1}u_0 - \cdots - v_{t-q,q}u_{-q+1}]. \end{aligned}$$

By rearranging and collecting terms in the u_j, $j = 0, -1, \ldots, -q+1$, the last $n - q$ rows of (15) are obtained. For $t \leqslant q$, write

$$y_t = \boldsymbol{x}_t'\beta + u_t + \theta_1 u_{t-1} + \cdots + \theta_{t-1}u_1 + \theta_t u_0 + \cdots + \theta_q u_{t-q},$$

and then substitute as above for $u_{t-1}, \ldots, u_1$. Collecting terms in the pre-sample errors verifies the first q rows of (15).

5. Examples

5.1. Simulated data

In this subsection we present three examples with simulated data designed to illustrate the efficacy of our proposed methodology. For comparison we provide from the MICROTSP program approximate maximum likelihood (AML) results that incorporate a backcasting step. In all examples, the variable x_t is generated from the autoregression $x_t = 0.8x_{t-1} + v_t$, $v_t \sim \mathcal{N}(0, 8)$. The examples are now described.

Example 1: AR(3) errors

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t, \qquad n = 100,$$

with $\beta = (1, 1)'$, $\phi = (1.2, -0.2, -0.2)'$, $\sigma^2 = 1$.

Example 2: MA(4) errors

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t, \qquad n = 100,$$

with $\beta = (1, 1)'$, $\theta = (1.6, 0.5, -0.4, -0.2)'$, $\sigma^2 = 0.50$.

Example 3: Unit root autoregression with ARMA(2, 3) errors

$$y_t = \beta_1 + \beta_2 x_t + y_{t-1} + \varepsilon_t, \qquad n = 150,$$

where $\beta = (1, 1)'$, $\phi = (1, -0.2)'$, $\theta = (0.5, 0.2, 0.2)'$, $\sigma^2 = 1$.

No specific prior information about any of the parameters is incorporated: proper priors centered at zero with large variances are used for β, ϕ, and θ. For simplicity, a diffuse normal prior was also adopted for α_0, and the results were not found to change very much if (7) was used for the prior covariance since the sample sizes are relatively large. With α_0 assumed to be independent of the remaining parameters, the terms $p_1(\phi)$, $p_2(\theta)$, m, and d_2 disappear from the equations of Proposition 1. For σ^2 the prior is specified through $\nu_0 = \delta_0 = 0$. In all cases, the results that are reported are obtained from the ARMA(p, q) algorithm of Section 3. For models 1 and 2, virtually identical results were also obtained from the specialized algorithms given in Section 4. The third example presents an interesting problem: The coefficient of the intercept is not identified in the presence of a unit root, because the expected value of y_t does not exist in that case. As a result, the constant term is not well estimated, which is revealed in large variance and instability in the algorithm for this parameter. For simulation cycles when the coefficient is not equal to 1, of course, the intercept is identified.

Our implementation and monitoring of the MCMC algorithm is straightforward. The iterations are started from the least squares values, the first 200

Table 1
AR(3) model
$y_t = 1 + x_t + e_t$, $e_t = 1.2e_{t-1} - 0.2e_{t-2} - 0.2e_{t-3} + u_t$, $\sigma^2 = 1$

		Posterior distribution					
Parameter	AML estimate	Mean	Std. dev.	Median	Lower 95% limit	Upper 95% limit	Corr.
β_1	1.524 (0.435)	1.521 (0.004)	0.307	1.518	0.932	2.128	0.088
β_2	1.027 (0.079)	1.095 (0.001)	0.089	1.095	0.923	1.274	0.026
ϕ_1	1.379 (0.105)	1.351 (0.002)	0.113	1.352	1.133	1.576	0.078
ϕ_2	−0.550 (0.169)	−0.511 (0.002)	0.176	−0.511	−0.866	−0.171	0.073
ϕ_3	−0.052 (0.105)	−0.070 (0.001)	0.108	−0.070	−0.276	0.149	0.026
σ^2	0.952 —	0.992 (0.002)	0.147	0.981	0.744	1.320	0.041

Numerical standard error of posterior mean is in parentheses. Correlation denotes the first-order correlation of the Gibbs run. For AML, standard error is in parentheses. Sample size is 100; 6000 simulations.

draws are discarded, and the next 6000 are retained. Different starting values were seen to produce estimates in a range that is consistent with the numerical standard errors. Admittedly, the practical convergence of the chain can be monitored, as mentioned in Section 3.3, but we were satisfied with our simple sampling scheme for the purpose of these illustrations. Moreover, we found that the serial correlation of the run was generally negligible and tended to dissipate quickly for Example 1, and by the fifth to tenth lag in Examples 2 and 3. All posterior moments are computed as sample averages, which is justified by Proposition 3. To compute marginal density functions, however, samples are taken as fixed intervals to produce an approximately independent sample. Numerical standard errors are computed for the posterior mean by the batch means method described in Ripley (1987). In particular, the 6000 simulated values were placed into v batches of $6000/v$ observations. The batch size is increased until the lag 1 correlation of the batch means is less than 0.05. The numerical standard errors are estimated as $s/\sqrt{v}$, where s is the standard deviation of the batch means. Our experience in this and other problems suggests that this method is quite adequate. Alternatively, the spectral approach of Geweke (1992) can be used to compute the numerical standard errors.

Table 1 presents our results for Example 1 and results from an AML regression. For this model and data set, both AML and our Bayes procedures yield very similar results. A 95% confidence interval using the 0.025 and 0.975 percentiles of the simulated draws includes every true parameter value.

Table 2
MA(4) model
$y_t = 1 + x_t + e_t$, $e_t = u_t + 1.60u_{t-1} + 0.50u_{t-2} - 0.40u_{t-3} - 0.20u_{t-4}$, $\sigma^2 = 0.50$

		Posterior distribution					
Parameter	AML estimate	Mean	Std. dev.	Median	Lower 95% limit	Upper 95% limit	Corr.
β_1	0.682 (0.078)	0.720 (0.004)	0.188	0.720	0.344	1.108	0.259
β_2	1.034 (0.031)	1.026 (0.001)	0.026	1.025	0.978	1.080	0.261
θ_1	1.394 (0.071)	1.506 (0.004)	0.117	1.508	1.263	1.730	0.778
θ_2	0.197 (0.089)	0.404 (0.005)	0.169	0.399	0.076	0.761	0.614
θ_3	− 0.405 (0.076)	− 0.382 (0.004)	0.157	− 0.384	− 0.681	− 0.058	0.496
θ_4	− 0.156 (0.065)	− 0.189 (0.003)	0.107	− 0.192	− 0.393	0.021	0.661
σ^2	0.758 —	0.596 (0.001)	0.089	0.588	0.447	0.790	0.076

Numerical standard error of posterior mean is in parentheses. Correlation denotes the first-order correlation of the Gibbs run. For AML, standard error is in parentheses. Sample size is 100; 6000 simulations.

Table 2 presents AML and Bayes results for Example 2. As in Example 1, the 95% confidence interval traps the true parameter every time. For making hypothesis tests it is interesting to note that the standard deviations for the θ_i are larger than the standard errors reported by AML. This suggests that the normal approximation employed by AML understates the variability.

Results for Example 3 are contained in Table 3. It is noteworthy that the Bayes approach has no difficulty in finding a coefficient close to unity for the lagged dependent variable, while AML reports a much lower value. Note again that the standard deviations of the posterior distributions are larger for the ARMA parameters than the corresponding AML standard errors. Interestingly, the opposite is true for β_2 and β_3. True parameter values are contained in the 95% Bayesian confidence intervals in all cases.

Of course, the above results cannot be used to demonstrate the superiority of either AML or posterior means for estimation. They are based on only one simulated data set, for example, and only a small number of models. They do reveal, however, that the Bayes approach is practical and does not merely reproduce the AML results. In particular, differences in standard deviations are of potential importance. The frequentist sampling properties of our Bayes estimator is an important issue that will be taken up in future work.

Table 3
ARMA(2,3) with lagged dependent variable model
$y_t = 1 + x_t + y_{t-1} + e_t$, $e_t = e_{t-1} - 0.20e_{t-2} + u_t + 0.50u_{t-1} + 0.20u_{t-2} + 0.20u_{t-3}$, $\sigma^2 = 1$

		Posterior distribution					
Parameter	AML estimate	Mean	Std. dev.	Median	Lower 95% limit	Upper 95% limit	Corr.
β_1	—	3.893 (0.109)	2.578	3.738	− 1.205	9.539	0.856
β_2	0.963 (0.080)	1.068 (0.001)	0.061	1.067	0.948	1.192	0.075
β_3	0.777 (0.057)	0.988 (0.001)	0.014	0.989	0.955	1.014	0.750
ϕ_1	1.269 (0.165)	1.221 (0.015)	0.239	1.250	0.664	1.624	0.895
ϕ_2	− 0.283 (0.161)	− 0.409 (0.014)	0.221	− 0.446	− 0.747	0.134	0.873
θ_1	0.573 (0.145)	0.369 (0.016)	0.257	0.336	− 0.063	0.955	0.928
θ_2	0.279 (0.172)	0.148 (0.012)	0.196	0.138	− 0.218	0.567	0.850
θ_3	0.170 (0.115)	0.101 (0.005)	0.116	0.103	− 0.140	0.324	0.629
σ^2	1.025 —	1.180 (0.013)	0.267	1.123	0.866	1.909	0.661

Numerical standard error of posterior mean is in parentheses. Correlation denotes the first-order correlation of the Gibbs run. For AML, standard error is in parentheses. Sample size is 149; 6000 simulations.

5.2. *GNP data*

Our last example examines U.S. real GNP data from 1951.2 to 1988.4, taken from *Business Conditions Digest*, September 1989. A large literature on the behavior of this and related series is concerned with the possible existence of a unit root. To investigate this question we estimate the model

$$\ln(GNP_t) = \beta_1 + \beta_2 t + \beta_3 \ln(GNP_{t-1}) + \varepsilon_t.$$

Since several of the specifications that we examined led to approximately the same inferences about β_2 and β_3, we assumed an MA(2) process for the errors. Models similar to this have been investigated extensively in the emerging Bayesian literature on unit-root tests but with an uncorrelated error term.

Results appear in Table 4 and Figs. 1 and 2. The marginal and joint posterior densities are computed from a Gaussian kernel applied to every tenth draw in the Gibbs sampler, which achieves an approximately independent sample. For the reasons mentioned above, the intercept is not well estimated in this model. The mean value of 0.9191 for the coefficient of the lagged dependent variable suggests considerable persistence, although it is in the stationary region. Fig. 1

Table 4
U.S. GNP data: MA(2), lagged dependent variable
$\ln(GNP_t) = \beta_1 + \beta_2 t + \beta_3 \ln(GNP_{t-1}) + e_t$, $e_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2}$

		Posterior distribution					
Parameter	AML estimate	Mean	Std. dev.	Median	Lower 95% limit	Upper 95% limit	Corr.
β_1	0.619 (0.155)	0.6337 (0.0073)	0.5406	0.6066	−0.2828	1.7240	0.1125
β_2	0.0006 (0.0002)	0.0006 (0.0000)	0.0005	0.0006	−0.0002	0.0016	0.1202
β_3	0.9149 (0.0215)	0.9191 (0.0009)	0.0697	0.9225	0.7789	1.0373	0.1129
θ_1	0.355 (0.080)	0.363 (0.003)	0.123	0.358	0.117	0.618	0.572
θ_2	0.254 (0.081)	0.261 (0.002)	0.102	0.256	0.076	0.485	0.292
σ^2	0.0001	0.0002 (0.000)	0.0005	0.0001	0.0001	0.0010	0.3331

Numerical standard error of posterior mean is in parentheses. Correlation denotes the first-order correlation of the Gibbs run. For AML, standard error is in parentheses. Sample size is 151; 6000 simulations; the Metropolis acceptance rate is 0.8204.

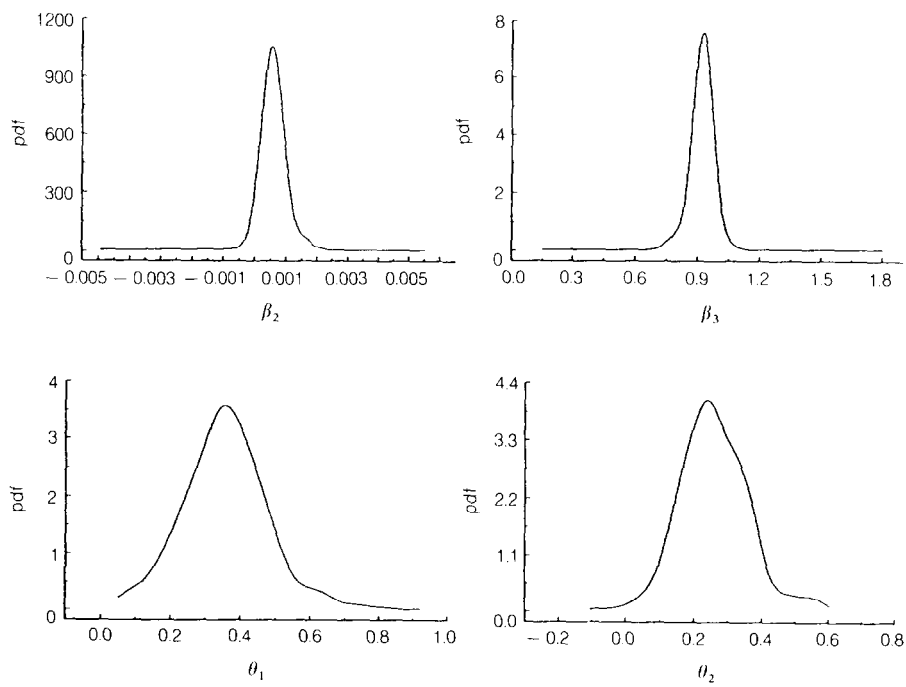

Fig. 1. GNP example: Selected posterior distributions.

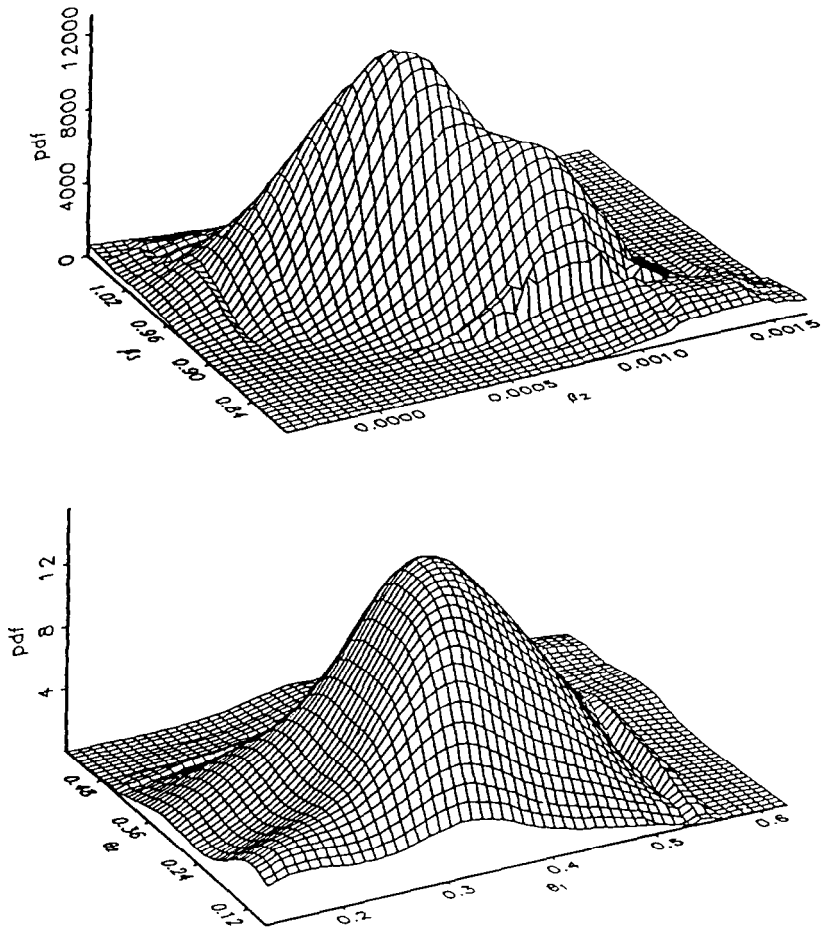

Fig. 2. GNP example: Selected joint posterior distributions.

reveals a distribution of β_3 that is quite compact around its mean, but with some probability of being greater than 1. The posterior probability that β_3 is greater than 0.95 is approximately 0.25. The mean of the time trend coefficient is very small and is also distributed closely about its mean. Table 4 and Fig. 1 indicate that θ_1 and θ_2 are clearly not zero, and the former displays a pronounced positive skewness. Fig. 2 demonstrates the ability of our procedure to generate exact joint posterior distributions. In particular, it appears that an approximate bivariate normal distribution would be highly misleading for β_2 and β_3.

6. Concluding remarks

We believe that the sampling-based approach taken in this paper is an important alternative to existing methods for the exploration of a very rich set of models. We have shown that a full Bayesian analysis of regression models with ARMA(p, q) errors is possible without special assumptions about pre-sample errors and without the evaluation of high-dimensional integrals. Since our approach yields the complete posterior distribution of the parameters, their behavior can be studied over their entire range rather than only around the mode. Our experience accumulated over many different problems is that the procedures discussed and illustrated above work very well. We have found that the usual reason for the failure of sample values to converge is the presence of common factors or nearly common factors in $\phi(L)$ and $\theta(L)$. Some coefficients are then not identified, and their posterior distributions display large standard errors.

Another feature of our approach is that its implementation is straightforward. Our transformations and the Kalman equations involve recursive calculations, simulation is largely from standard distributions, and no large matrix inversion, determinant computation, or numerical integration is required. Moreover, the analysis can easily be modified for classes of prior distributions other than those employed above. By reweighting (resampling) it is possible to transform a posterior sample based on one prior to a sample based on another, without additional simulations. See Smith and Gelfand (1992) for this technique.

We have also shown how the approach can be further simplified for the important special cases of stationary AR(p) and invertible MA(q) models. In the former case, there is no need to introduce α_0 and in the latter case only λ is required, whose simulation does not require the smoothing recursions employed in the general ARMA(p, q) model. The recursive transformations that we have developed in this paper to diagonalize the covariance matrix of the errors should also prove useful in frequentist estimation.

Finally, the framework described above can be extended to accommodate a wide variety of inferential objectives or assumptions. For example, prediction densities of future observations can be obtained by the method of composition to generate a posterior sample of future y. Detailed calculations of this type are reported in Albert and Chib (1993), where prediction densities up to four steps ahead are found for Markov switching autoregressive models. Unlike frequentist calculations, these prediction densities fully incorporate both parameter and error term uncertainty. As another example, the Gaussian assumption can be relaxed in the direction of the Student-t family or, more generally, in the direction of scale-mixtures of normals. Finally, our results can be extended to vector ARMA(p, q) processes. Although this extension remains an object for future research, some progress toward that goal is made in Chib and Greenberg (1993b), where SUR models with low-order correlated vector error processes are shown to be amenable to Markov chain sampling.

Appendix: Proof of Proposition 2

A.1. Preliminaries

We provide a proof of Proposition 2 as it pertains to T_ϕ; the proof for T_θ is analogous. Note that Roberts and Smith (1992) assert that connectedness is required, but their proof utilizes the stronger condition of arc connectedness. It is easy to check the conditions for β, σ^2, and α_0 since they live on $\Re^k$, $\Re^+$, and $\Re^m$, which are open connected sets, with well-behaved multivariate normal, inverted gamma, and multivariate normal densities, respectively. The stationarity and invertibility conditions imposed on ϕ and θ, however, lead to domains that require further analysis. For the one- and two-dimensional cases, pictured in many textbooks, the structures of S_ϕ and S_θ are well known, but this does not appear to be the case for larger values of p. Our result may therefore be of independent interest. The set S_ϕ is extended to T_ϕ, because S_ϕ is not connected even for $p = 1$ since $0 < |\phi_1| < 1$. To ensure connectedness, we need to include $\phi_1 = 0$. But since both S'_ϕ and S'_θ are sets of measure zero, draws will be made from S_ϕ or S_θ a.s., and the simulations will therefore not be affected.

It is convenient to work with the polynomial

$$f(z) = z^p - \phi_1 z^{p-1} - \cdots - \phi_p, \tag{16}$$

which is obtained from $\phi(z)$ by multiplying by z^{-p} and interpreting the result as a polynomial in z^{-1}. In this form, stationarity implies that all roots lie inside the unit circle. Note that $\phi(z)$ and $f(z)$ are real polynomials.

A.2. Proof

First consider property (i). Baumol (1970, p. 247) quotes a version of the Routh–Hurwitz theorem that states that the roots of the polynomial (16) are all less than unity in absolute value if and only if the values of certain determinants, the elements of which are the ϕ_i, are positive. Since determinants are continuous functions of their elements, we have the result that T_ϕ is an open set. As noted above, the parameters are defined on the product set of the elements of (ψ, α_0), all of which are open, and continuous densities have been placed on each of these sets; property (i) is therefore verified. This result also shows that S'_ϕ and S'_θ have measure 0.

Next consider property (ii). As may be seen from (12) and Assumption P, the joint posterior density can be written in the form

$$\pi(\psi, \alpha_0 | \boldsymbol{y}) \propto \sigma^{-(n+\nu_0)} \exp[-\tfrac{1}{2}[\nu_0 \delta_0 + Q(\psi, \alpha_0)]], \tag{17}$$

where $Q(\psi, \alpha_0)$ is a quadratic function of its arguments. Since the exponential term is dominated by unity, integrating out any of β, ϕ, θ, or α_0 results in an upper bound of $KV\sigma^{-(n+\nu_0)}$, where K is the proportionality constant and V is

the volume of the region of integration. This expression is clearly bounded. To integrate out σ over the interval $(\sigma' - \mu, \sigma' + \mu)$, again dominate the exponential term by unity. Then we obtain as an upper bound $\int_{\sigma'-\mu}^{\sigma'+\mu} \sigma^{-n+\nu_0} d\sigma$, which is finite for sufficiently small μ. This completes the verification of property (ii).

Finally, we verify property (iii). We first show that the set $Z = \{z: z = (z_1, \ldots, z_p) \in \mathscr{C}^p, |z_i| < 1, f(z_i) = 0 \ (i = 1, \ldots, p)\}$ is arc connected, where $f(\cdot)$ is a real polynomial. It is well known that the roots and coefficients of a polynomial are related through $\phi_i = (-1)^i \sigma_i \ (i = 1, \ldots, p)$, where the σ_i are the elementary symmetric functions. Each σ_i is the sum of cross-products of the roots taken i at a time; i.e., σ_1 is the sum of the roots, σ_2 is the sum of all products of roots taken two at a time, and finally, σ_p is the product of all roots. If $z \in Z$, then $\lambda z \in Z$ for $0 \leqslant \lambda \leqslant 1$: Clearly $\lambda z \in \mathscr{C}^p$, $|\lambda z_i| = \lambda |z_i| < 1$, and the λz_i are the roots of a polynomial with coefficients $(-1)^i \sum \lambda^i \sigma_i$, which are real for real σ_i. Now consider any $u, v \in Z$. The result that $\lambda z \in Z$ proves that the straight lines from u to 0 and from 0 to v are entirely in Z and constitute a path from u to v. This proves that Z is arc connected. The verification of property (iii) is then completed by noting that the coefficients in T_ϕ are continuous functions of their roots and that continuous images of arc-connected sets are arc-connected. ■

References

Albert, J. and S. Chib, 1993, Bayesian inference for autoregressive time series with mean and variance subject to Markov jumps, Journal of Business and Economic Statistics 11, 1–15.

Ansley, C.F., 1979, An algorithm for the exact likelihood of a mixed autoregressive–moving average process, Biometrika 66, 59–65.

Baumol, W.J., 1970, Economic dynamics (Macmillan, New York, NY).

Broemeling, L.D. and S. Shaarway, 1984, Bayesian inferences and forecasts with moving average processes, Communications in Statistics, Theory and Methods 13, 1871–1888.

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

Chib, S., 1993, Bayes regression with autoregressive errors: A Gibbs sampling approach, Journal of Econometrics 58, 275–294.

Chib, S. and E. Greenberg, 1993a, Estimating nonlinear latent variable models using Markov chain Monte Carlo, Manuscript.

Chib, S. and E. Greenberg, 1993b, Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models, Journal of Econometrics, forthcoming.

Fuller, W., 1976, The statistical analysis of time series (Wiley, New York, NY).

Galbraith, J.W. and V. Zinde-Walsh, 1992, The GLS transformation matrix and a semi-recursive estimator for the linear regression model with ARMA errors, Econometric Theory 8, 95–111.

Gardner, G.A., A.C. Harvey, and G.D.A. Phillips, 1979, Algorithm AS154: An algorithm for the exact maximum likelihood estimation of autoregressive–moving average models by means of Kalman filters, Applied Statistics 19, 311–322.

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.B. Rubin, 1992, Inference from iterative simulation using multiple sequences (with discussion), Statistical Science 7, 457–511.

Geweke, J., 1992, Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, in: J.M. Bernardo et al., eds., Bayesian statistics 4 (Oxford University, New York, NY) 169–193.

Harvey, A., 1981, Time series models (Philip Allan, London).

Hastings, W.K., 1970, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57, 97–109.

Jacquier, E., N.G. Polson, and P.E. Rossi, 1992, Bayesian analysis of stochastic volatility models, Manuscript.

Marriott, J., N. Ravishanker, A. Gelfand, and J. Pai, 1992, Bayesian analysis of ARMA processes: Complete sampling based inference under full likelihoods, Manuscript.

McCulloch, R.E. and R.S. Tsay, 1991, Bayesian analysis of autoregressive time series via the Gibbs sampler, Manuscript.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, 1953, Equations of state calculations by fast computing machines, Journal of Chemical Physics 21, 1087–1092.

Monahan, J.F., 1983, Fully Bayesian analysis of ARMA time series models, Journal of Econometrics 21, 307–331.

Müller, P., 1993, A generic approach to posterior integration and Gibbs sampling, Journal of the American Statistical Association, forthcoming.

Otto, M.C., W.R. Bell, and J.P. Burman, 1987, An iterative GLS approach to maximum likelihood estimation of regression models with ARIMA errors, Manuscript.

Pagan, A.R. and D.F. Nicholls, 1976, Exact maximum likelihood estimation of regression models with finite order moving average errors, Review of Economic Studies 43, 383–388.

Ripley, B., 1987, Stochastic simulation (Wiley, New York, NY).

Ritter, C. and M. Tanner, 1992, The Gibbs stopper and the griddy Gibbs sampler, Journal of the American Statistical Association 87, 861–868.

Roberts, G.O. and A.F.M. Smith, 1992, Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms, Manuscript.

Smith, A.F.M. and A.E. Gelfand, 1992, Bayesian statistics without tears: A sampling–resampling perspective, The American Statistician 46, 84–88.

Tanner, M. and W.H. Wong, 1987, The calculation of posterior distributions by data augmentation, Journal of the American Statistical Association 82, 528–549.

Tierney, L., 1993, Markov chains for exploring posterior distributions, Annals of Statistics, forthcoming.

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

Zellner, A. and C. Min, 1992, Gibbs sampler convergence criteria (GSC^2), Technical report (Graduate School of Business, University of Chicago, Chicago, IL).