

# Analytic convergence rates and parameterisation issues for the Gibbs sampler applied to state space models

MICHAEL K PITT

*Department of Statistics, University of Oxford, OX1 3TG and  
Nuffield College, Oxford, OX1 1NF, UK*

AND NEIL SHEPHARD

*Nuffield College, Oxford, OX1 1NF, UK*

August 14, 1997

## Abstract

In this paper we obtain a closed form expression for the convergence rate of the Gibbs sampler applied to the unobserved states of an AR(1) plus noise model. The rate is expressed in terms of the parameters of the model, which are regarded as fixed. For the case where the unconditional mean of the states is a parameter of interest we provide evidence that a “centered” parameterisation of a state space model is preferable for the performance of the Gibbs sampler. These two results provide guidance when the Gaussianity or linearity of the state space form is lost. We illustrate this by examining the performance of a Markov Chain Monte Carlo sampler for the stochastic volatility model.

KEYWORDS: Blocking; Convergence rates; Gibbs sampling; Parameterisation; Markov Chain Monte Carlo; Simulation smoother; Stochastic volatility.

## 1 Introduction

Markov Chain Monte Carlo (MCMC) methods and more specifically Gibbs sampling have been used to tackle a wide variety of statistical problems. Early examples of their use in image analysis includes Ripley (1977) and Geman and Geman (1984). The methods were suggested for more widespread statistical implementation by Ripley (1987) and Gelfand and Smith (1990). The methods became widely used in research papers on Bayesian statistics following the influential work of Tanner and Wong (1987) and Gelfand and Smith (1990). Many examples of their current use are given in the booklength review by Gilks, Richardson, and Spiegelhalter (1996). The ability of MCMC methods to facilitate the analysis of unobserved, or latent, variable models has perhaps largely accounted for their recent popularity compared to more traditional classical approaches. Time series models which involve unobserved variables form the focus of this paper.

One of the most general and widely used time series structures is the Gaussian state space form (GSSF) which allows the observation at time  $t$ ,  $y_t$ , to be linearly related to the corresponding time varying autoregressive state,  $\alpha_t$ , but corrupted through the addition of Gaussian noise. Kalman (1960) devised a recursive technique, known as the Kalman filter, which calculates the

density of the current state,  $\alpha_t$ , given all observations up to time  $t$ . This filter also delivers the log-likelihood which in turn allows a classical analysis of unknown parameters (see, for example, Harvey (1989)). In addition, direct methods for efficiently computing the smoothing density, which delivers the density of all states conditional upon all the observations, are available (see, for example, de Jong (1989)).

We define non-Gaussian measurement state space models to be the same as GSSFs except that the logarithm of the measurement density is no longer a quadratic function of the corresponding state. Instead the log-density is considered to be an arbitrary, generally concave, known function of the state. These kinds of models arise frequently in engineering problems and in finance as well as more generally in statistics. In section 3.4, we consider the stochastic volatility (SV) model as an example. The analytic tractability which enables straightforward analysis of the GSSF is no longer available.

Recently, there has been research into using MCMC methods for smoothing and parameter estimation for these models. Gibbs sampling is now increasingly used as a tool to fit non-Gaussian state space models. Gibbs sampling has as its aim the drawing of random variables from  $x = \alpha, \theta|y$ . By summarising these draws we can conduct Bayesian inference on  $\theta$  through  $\theta|y$  and  $\alpha$  through  $\alpha|y$ . Here  $y$  denotes the data vector,  $\theta$  the unknown parameters in the model and  $\alpha$  the vector of all of the states at all time points.

To simplify notation write  $\psi = (\alpha', \theta)'$ , then for the problem of simulating from a multivariate density  $\pi(\psi|y)$ , the Gibbs sampler is defined by a blocking scheme which splits the vector  $\psi$  into  $d$  blocks. We write this as  $\psi = (\psi_1', \dots, \psi_d)'$ , while working with the associated full conditional distributions  $\psi_i|y, \psi_{\setminus i}$ , where  $\psi_{\setminus i}$  denotes  $\psi$  excluding the block  $\psi_i$ . The algorithm proceeds by sampling each block from the full conditional distributions where the most recent values of the conditioning blocks are used in the simulation. One cycle of the algorithm is called a sweep or a scan. Under regularity conditions, as the sampler is repeatedly swept, the draws from the sampler converge to draws from the target density at a geometric rate.

Some proposed MCMC methods make use of what is termed single-move Gibbs sampling. This means that each of the  $\dim(\psi)$  scalar elements of  $\psi$ ,  $\psi_i$ , are sampled one at a time from its complete conditional Bayesian posterior density,

$$\psi_i|y, \psi_{\setminus i}, \quad i = 1, \dots, \dim(\psi).$$

In the time series literature the earliest reference to the use of single move Gibbs sampling seems to be Carlin, Polson, and Stoffer (1992) who noted the conditional dependence structure of the non-Gaussian state space models.

Single-move Gibbs sampling can be problematic. It is well known that sampling separately

from the conditional density of variables which are highly correlated can lead to slow convergence. This problem is thought to be particularly severe for non-Gaussian SSF models when the states are sampled individually. Early attempts to sample blocks of states simultaneously, for a non-Gaussian or a partial Gaussian state space model, include Shephard (1994) and Carter and Kohn (1994). Further, Fruhwirth-Schnatter (1994) used Gibbs sampling to draw from the posterior density of the parameters in GSSF models by setting  $d = 2$  and by defining  $\psi_1 = \alpha$  and  $\psi_2 = \theta$ . Her Gibbs sampler drew from  $\alpha|y, \theta$  and  $\theta|y, \alpha$  by exploiting the Gaussian structure of the model. A more general approach is provided by Shephard and Pitt (1997) who use a Metropolis method to sample large blocks of states simultaneously. Shephard and Pitt (1997) demonstrate, by simulation, that this blocking method is far superior to single sampling.

The first concern of this paper, section 2, is to obtain an analytic convergence rate for the single move Gibbs sampler applied to the states of the first order autoregression (written AR(1)) plus noise Gaussian model. This rate is expressed in terms of the parameters of the model which are regarded as fixed. We obtain a formula for the upper and lower bound of the rate for finite  $n$  (the time dimension) and an asymptotically limiting rate as  $n \rightarrow \infty$ . The convergence rate informs us about the speed of convergence (how rapidly the starting values become irrelevant). It also enables consideration of the efficiency of the Gibbs method in equilibrium (relative to independent samples, say). We find that the single move Gibbs sampler applied to the states is very poor for persistent models both in terms of efficiency and convergence.

In section 3 we are concerned with the effect of reparameterisation. Issues of equivalent reparameterisations have been shown to be important outside the time series literature (see, for example, Gelfand, Sahu, and Carlin (1995)). We consider two alternative parameterisations for the unconditional mean of the states of the AR(1) plus noise model. The efficiency and convergence rates are considered for the Gibbs sampler applied to both the states and the unconditional mean. For this analysis, we consider all of the states being sampled simultaneously. The other elements of  $\theta$  are assumed fixed. We derive conditions under which the so called “centered” parameterisation is better than the “uncentered” alternative. We also provide tight upper and lower bounds on the relative efficiency of the two schemes considered.

The similarity between the GSSF and non-Gaussian SSF has been highlighted in the existing statistical literature. Harvey, Ruiz, and Shephard (1994) analyse the non-Gaussian SV model by approximating it by a GSSF. More recently, Shephard and Pitt (1997) note that a Laplace approximation to the conditional density of the unobserved states leads to a GSSF, enabling efficient proposals for a Metropolis method. The more persistent the state evolution, the better the approximation will be since the Gaussian VAR(1) evolution of the states will dominate over

the measurement density. Outside the time series domain, Roberts and Sahu (1996) consider convergence rates of the Gibbs sampler for non-Gaussian models via Gaussian approximation.

In this paper, it is hoped that the suggestions made for linear state space models, in particular the AR(1) plus noise model, will provide useful insights into the most effective procedure for Gibbs sampling for non-Gaussian state space models. Indeed, the suggestion of “centering” the parameterisation is explored, section 3.4, for the stochastic volatility model using data on exchange rate returns.

## 2 Analytic convergence rates for AR(1) plus noise model

Roberts and Sahu (1997) have shown that it is possible to compute convergence rates for the Gibbs sampler applied to multivariate Gaussian densities. In this section, we apply their methodology to obtain a closed form for the rate of convergence of the single move Gibbs sampler when used on the AR(1) plus noise model

$$\begin{aligned} y_t &= \mu + \alpha_t + \varepsilon_t, & \varepsilon_t &\sim \text{NID}(0, \sigma_\varepsilon^2), t = 1, \dots, n, \\ \alpha_t &= \phi\alpha_{t-1} + \eta_t, & \eta_t &\sim \text{NID}(0, \sigma_\eta^2), t = 1, \dots, n, & \alpha_1 &\sim \text{N}\{0, \sigma_\eta^2/(1 - \phi^2)\}. \end{aligned}$$

Throughout  $\varepsilon_t$  and  $\eta_s$  are independent for all  $t$  and  $s$  and both of these sets of random variables are independent from  $\alpha_1$ . We choose the mean and variance of the initial state to be the same as the unconditional mean and variance of the states. The AR(1) evolution is initially assumed to be stationary, so  $|\phi| < 1$ , although the results will go through when  $|\phi| = 1$ . At first we will assume that all the parameters of the model,  $\theta = (\sigma_\varepsilon^2, \sigma_\eta^2, \phi, \mu)'$ , are known.

The Gibbs sampling problem will revolve around setting  $\psi = \alpha$  and so drawing from  $x = \alpha|y, \theta$ . The sampler draws with replacement from

$$x_t \sim x_t|x_{\setminus t} = \alpha_t|\alpha_{\setminus t}, y, \theta = \alpha_t|\alpha_{t-1}, \alpha_{t+1}, y_t, \theta, \quad t = 1, \dots, n,$$

where the most recent values of the conditioning elements are used in the simulation. Having progressed all the way through the states we say we have carried out a single sweep or scan of the sampler. The results of the first sweep are denoted by the vector  $x^{(1)}$ , while the results from the  $i$  –  $th$  sweep will be written as  $x^{(i)}$ . We can compute the analytic convergence rate of this procedure, as it is repeatedly swept, using the general approach of Roberts and Sahu (1997).

The rate of convergence,  $\rho$ , is formally defined as follows. The Gibbs sampling sequence  $\{x^{(i)}, i = 0, 1, \dots\}$  forms a Markov chain which we assume has a stationary density function  $\pi(x)$ . Let  $f$  be a square  $\pi$ -integrable function of  $x$  and  $\pi(f)$  denote the expectation of  $f$  under the target density  $\pi$ . Let  $\rho$  be the minimum number such that for all square  $\pi$ -integrable functions

$f$ , and for all  $r < \rho$ , we can find a function  $V(\cdot) \geq 1$  such that  $\pi(V) < \infty$  and

$$\left| E \left\{ f(x^{(i)}) \right\} - \pi(f) \right| \leq V(x^{(0)})r^i. \quad (1)$$

If there exists a  $\rho < 1$  satisfying (1) then we say the Markov chain is geometrically ergodic and  $\rho$  is its rate of convergence.

As  $y, \alpha|\theta$  is Gaussian so  $x = \alpha|y, \theta$  must be Gaussian. Let us write  $x \sim \mathbf{N}(\mu\mathbf{1}, \Sigma)$ , where  $\mathbf{1}$  is a vector of ones. Then  $\Sigma$  will be positive definite so long as  $\sigma_\varepsilon^2 > 0$ . Under this condition we can calculate  $\rho$  by using  $Q = \Sigma^{-1}$ .

We will write the  $i, j$  - *th* element of  $Q$  as  $Q_{ij}$ . Let  $A = I - \text{diag}(Q_{11}^{-1}, \dots, Q_{nn}^{-1})Q$  where  $I$  is the identity matrix. If we define  $L$  to be the block lower triangular matrix with blocks in the lower triangle being identical to  $A$  and  $U = A - L$  we can calculate the Markov Chain induced by a deterministic Gibbs sampler (DUGS).

We have  $E(x^{(i+1)}|x^{(i)}) = Bx^{(i)} + b$ , where  $B = (I - L)^{-1}U$  and  $b = (I - B)^{-1}\mu$ . Hence the rate of convergence of the Markov chain to its stationary distribution is given by  $\rho = \rho(B)$ , the spectral radius of  $B$ . This provides the exact rate of convergence for a DUGS scheme. If  $Q$  is of tridiagonal form then we can use Theorem 5 of Roberts and Sahu (1997) which states that  $\rho(B) = \{\lambda(A)\}^2$ , where  $\lambda(X)$  denotes the largest eigenvalue of  $X$ .

The convergence rate essentially measures how quickly our initial starting values become irrelevant. If  $\rho$  is close to 1, we have very slow convergence. If we have independent samples then  $\rho = 0$ . The convergence rate can also be used as a guide to the burn-in period. If we require an accuracy of 0.001, meaning that the right hand side of (1) has to be less than  $V(x^{(0)}) \times 0.001$ , then we require a burn-in period of  $\log(0.001)/\log(\rho)$ . As  $\rho \rightarrow 1$ , it is clear that the ‘‘burn in’’ efficiency of the Gibbs sampler is proportional to  $1/(1 - \rho)$ . The relative ‘‘burn in’’ efficiency (how many times less samples we require for a given accuracy) of a Gibbs scheme with rate  $\rho_2$  to a scheme with rate  $\rho_1$  is  $\log \rho_2 / \log \rho_1$ . As  $\rho_1, \rho_2 \rightarrow 1$ , this becomes  $(1 - \rho_2)/(1 - \rho_1)$ .

Since we have a VAR(1) process in the DUGS scheme, we can also straightforwardly calculate the efficiency, after equilibrium, of a Gibbs sampler with rate  $\rho$ . The inefficiency, in equilibrium, of the sampler is defined as how many more samples, than an independent sampler, we require to estimate the expectation of a linear combination of the states to a given accuracy. For the Gibbs sampler applied to the Gaussian target density, this is simply  $(1 + \rho)/(1 - \rho)$ .

For an AR(1) plus noise model the  $Q$ , the inverse variance-covariance matrix for the states

$\alpha$  which arises from the AR(1) plus noise model, is given by

$$Q = \sigma_\varepsilon^{-2}I + \sigma_\eta^{-2} \begin{pmatrix} 1 & -\phi & 0 & \dots & 0 \\ -\phi & 1 + \phi^2 & -\phi & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & & -\phi & 1 + \phi^2 & -\phi \\ 0 & \dots & 0 & -\phi & 1 \end{pmatrix},$$

a tridiagonal matrix. Hence for the single move Gibbs sampler we obtain,

$$A = I - \text{diag}(Q_{11}^{-1}, \dots, Q_{nn}^{-1})Q = \begin{pmatrix} 0 & b & 0 & \dots & 0 \\ a & 0 & a & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & a & 0 & a \\ 0 & \dots & 0 & b & 0 \end{pmatrix},$$

where  $a = \phi / (1 + \phi^2 + \sigma_\eta^2 \sigma_\varepsilon^{-2})$  and  $b = \phi / (1 + \sigma_\eta^2 \sigma_\varepsilon^{-2})$ . Notice  $A$  depends only on  $\phi$  and the signal to noise ratio  $\sigma_\eta^2 \sigma_\varepsilon^{-2}$  and so the exact rate of convergence can only depend on these quantities. Further, we obtain

$$B = (I - L)^{-1}U = \begin{pmatrix} 0 & b & 0 & \dots & \dots & 0 \\ 0 & ab & a & & & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & a^{n-2}b & a^{n-2} & \dots & a^2 & a \\ 0 & a^{n-2}b^2 & a^{n-2}b & \dots & a^2b & ab \end{pmatrix}.$$

We obtain the following result, proved in the Appendix, noting that  $\rho(B) = \{\lambda(A)\}^2$ , since  $Q$  is tridiagonal.

**Theorem 1:** *The spectral radius of  $B$ ,  $\rho(B)$ , satisfies the following inequality, if  $\sigma_\eta^2, \sigma_\varepsilon^2 > 0$ ,*

$$4 \cos^2 \left( \frac{\pi}{n+1} \right) a^2 < \rho < 4a^2,$$

where  $a$  is defined above.

This provides a tight bound for reasonably large  $n$ . Hence the convergence rate  $\rho$  can be expressed in terms of the parameters of the AR(1) plus noise model as:

$$\lim_{n \rightarrow \infty} \rho = 4a^2 = 4 \frac{\phi^2}{(1 + \phi^2 + \sigma_\eta^2 \sigma_\varepsilon^{-2})^2}.$$

This formula holds even when  $\phi$  equals one, although now  $\alpha_1$  is initialized using a diffuse prior.

Theorem 1 indicates that the convergence rate always lies between 0 and 1 and as  $n \rightarrow \infty, |\phi| \rightarrow 1, \sigma_\eta^2 \sigma_\varepsilon^{-2} \rightarrow 0$  then  $\rho \rightarrow 1$ . Hence it can be seen that for reasonably persistent

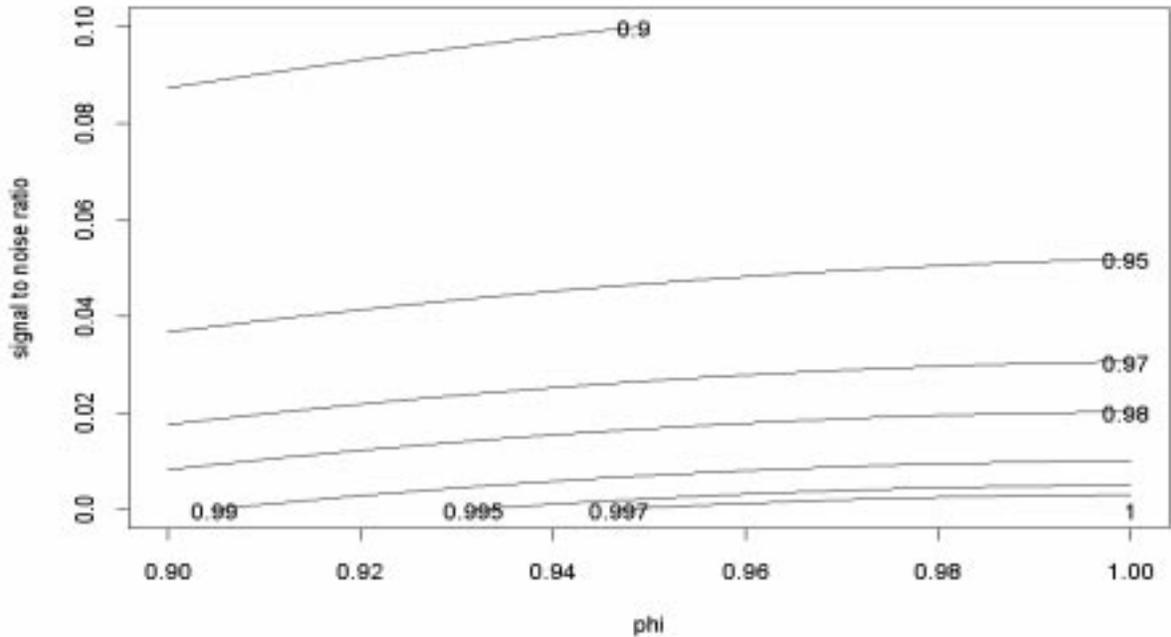


Figure 1: *Contour plot of levels of asymptotic rate plotted against  $\phi$  and signal to noise ratio  $\sigma_\eta^2\sigma_\varepsilon^{-2}$ .*

parameterisations (which frequently arise in many time series applications) the convergence rate will be close to 1, indicating slow convergence. The convergence rate tends towards a steady state constant, independent of  $n$ , as  $n$  increases. This is not surprising as the correlation between  $\alpha_t|y$  and  $\alpha_{t+k}|y$  decays at a geometric rate for large  $k$  as  $k$  increases, even for the random walk case ( $\phi = 1$ ), so long as  $\sigma_\eta^2\sigma_\varepsilon^{-2} \neq 0$ .

As an example let us consider a model for which the signal to noise ratio,  $\sigma_\eta^2\sigma_\varepsilon^{-2} = 0.1$  and  $\phi = 0.98$ . When  $n = \infty$  then  $\rho = 0.90491$  (this gives us the upper bound for  $n < \infty$ ). Now for  $n = 50$  we obtain a lower bound of 0.90161. It is clear that we have a very tight bound on  $\rho$  for  $n = 50$  and that the convergence rate changes very little as  $n$  increases from 50 to infinity.

Figure 1 shows contours of the asymptotic convergence rate for different values of  $\phi$  and the signal to noise ratio,  $\sigma_\eta^2/\sigma_\varepsilon^2$ . It is clear that the convergence will be very slow for persistent time series ( $\phi$  close to 1,  $\sigma_\eta^2$  close to 0). For time series with low amounts of autocorrelation the simple single move sampler is likely to be very successful however.

These results ignore the fact that we are generally interested in the parameters of the model as well as the states. Issues relating to the choice of parameterisation are covered in the following section.

### 3 Reparameterisation for linear state space models

#### 3.1 Background

We shall now focus on the following models,

$$y|\alpha \sim \mathbf{N}(\mu\mathbf{1} + \alpha, \sigma_\varepsilon^2 I_n), \text{ where } \alpha \sim \mathbf{N}(0, D), \quad (2)$$

and

$$y|\omega \sim \mathbf{N}(\omega, \sigma_\varepsilon^2 I_n), \text{ where } \omega \sim \mathbf{N}(\mu\mathbf{1}, D), \quad (3)$$

where  $y = (y_1, \dots, y_n)'$  represents the  $n \times 1$  vector of observations,  $\mathbf{1}$  represents an  $n \times 1$  vector consisting of 1's and  $\omega, \alpha$  represent unobserved  $n \times 1$  vectors. The scalar  $\sigma_\varepsilon^2$  and positive definite matrix  $D$  are regarded as being known, whilst the scalar  $\mu$  is regarded as unknown. It can be seen that the marginal distribution of  $y$  is unaffected by which parameterisation we choose. This framework provides a general structure for examining the special case in which  $\alpha, \omega$  are of AR(1) form. For this case the first setup is,

$$\begin{aligned} y_t &= \mu + \alpha_t + \varepsilon_t, & \varepsilon_t &\sim \text{NID}(0, \sigma_\varepsilon^2), & t &= 1, \dots, n, \\ \alpha_t &= \phi\alpha_{t-1} + \eta_t, & \eta_t &\sim \text{NID}(0, \sigma_\eta^2), & \alpha_1 &\sim \mathbf{N}\left\{0, \sigma_\eta^2/(1 - \phi^2)\right\}, \end{aligned}$$

and the second parameterisation is,

$$\begin{aligned} y_t &= \omega_t + \varepsilon_t, & \omega_t &= \mu + \phi(\omega_{t-1} - \mu) + \eta_t, \text{ and} \\ \omega_1 &\sim \mathbf{N}\left\{\mu, \sigma_\eta^2/(1 - \phi^2)\right\}, & t &= 1, \dots, n. \end{aligned}$$

These two models are equivalent since the density  $f(y|\mu)$  is the same for both. The posterior density  $f(\mu|y) \propto f(y|\mu)f(\mu)$  is therefore also invariant under the two parameterisations. The form of the posterior for a flat prior on  $\mu$ ,  $f(\mu) \propto \text{const}$ , is given by (15) in section 6.2. The assumption of a flat prior on  $\mu$ , whilst the initial state is assumed to arise from the unconditional distribution of the AR(1) process, is standard. It is advocated in the near to unit root autoregressive literature by a number of econometricians. See, for example, the review paper by Schotman (1994, pp 590–591). Of course if  $|\phi| = 1$  the parameter  $\mu$  will be unidentified and so we confine our attention to the stationary,  $|\phi| < 1$ , case.

The aim in this section is to consider the effect of Gibbs sampling for each of the two models. Our approach is similar to that of Gelfand, Sahu, and Carlin (1995) who consider reparameterisation for hierarchical linear normal models rather than time series models. We shall consider the simple Gibbs strategy sampling  $\mu$  from its complete conditional density  $f(\mu|y, \alpha)$ ,

for the first parameterisation, and from  $f(\mu|y, \omega)$  for the second configuration. Throughout  $\sigma_\varepsilon^2$ ,  $\sigma_\eta^2$  and  $\phi$  shall be assumed known.

The marginal and conditional distributions for the (2) and (3) models are given in the Appendix, 6.2. In particular it can be seen that the conditional distributions for  $\alpha|y$  will have a larger variance than the ‘‘centered’’ alternative of  $\omega|y$  if  $VD^{-1}\mathbf{1}\mathbf{1}'$  is small compared to  $\mathbf{1}\mathbf{1}'$ , where  $V = (\sigma_\varepsilon^{-2}I + D^{-1})^{-1}$ .

Now clearly for Gibbs sampling schemes it is the correlation structure which is particularly interesting. We are interested in the relative performance of the ‘‘uncentered’’ sampler which updates  $\mu|y, \alpha$  and  $\alpha|y, \mu$  and the ‘‘centered’’ sampler which updates  $\mu|y, \omega$  and  $\omega|y, \mu$ . Note that in both cases we consider all the states being updated simultaneously.

### 3.2 Autocorrelations

Since we have a two fold updating scheme (e.g. sampling  $\mu|y, \omega$  then  $\omega|y, \mu$ ) this gives rise to an AR(1) generation process in  $\mu$  obtained by integrating the states out. Hence it is the autocorrelation at lag 1 which is of interest in determining the efficiency of the two parameterisations. We shall denote the autocorrelation at lag 1 for the uncentered sampler by  $\rho_\mu(1; \alpha)$  and similarly use  $\rho_\mu(1; \omega)$  to denote the autocorrelation for the centered alternative.

Thus we obtain

$$\rho_\mu(1; \alpha) = \frac{1}{n}\mathbf{1}'(I - VD^{-1})\mathbf{1} = 1 - \frac{1}{n}\mathbf{1}'VD^{-1}\mathbf{1}. \quad (4)$$

Clearly if  $VD^{-1}$  were to consist entirely of 0's then  $\rho_\mu(1; \alpha) = 1$ . In practice, for persistent models  $\rho_\mu(1; \alpha)$  is close to 1. Note that  $VD^{-1} = (\sigma_\varepsilon^{-2}D + I)^{-1} = \sigma_\varepsilon^2 D^{-1}(\sigma_\varepsilon^2 D^{-1} + I)^{-1}$ .

For the centered parameterisation we have,

$$\rho_\mu(1; \omega) = \frac{\mathbf{1}'D^{-1}VD^{-1}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}} = 1 - \sigma_\varepsilon^{-2} \frac{\mathbf{1}'VD^{-1}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}}, \quad (5)$$

due to lemma 2 given in the Appendix.

Hence  $\rho_\mu(1; \omega) < \rho_\mu(1; \alpha)$  provided  $\mathbf{1}'D^{-1}\mathbf{1}\sigma_\varepsilon^2 < n$ . For the AR(1) plus noise model this reduces to, using lemma 3 in the Appendix,

$$\frac{\sigma_\varepsilon^2 \sigma_\eta^{-2}}{n} \left\{ (n-2)(1 + \phi^2) + 2 - 2(n-1)\phi \right\} < 1$$

and

$$\lim_{n \rightarrow \infty} \frac{\sigma_\varepsilon^2 \sigma_\eta^{-2}}{n} \left\{ (n-2)(1 + \phi^2) + 2 - 2(n-1)\phi \right\} = \sigma_\varepsilon^2 \sigma_\eta^{-2} (1 - \phi)^2.$$

We define the relative efficiency of the centered sampler to the uncentered alternative as the number of samples from the uncentered sampler we would require to estimate  $E(\mu)$  to a given

precision divided by the number of samples from the centered algorithm we would require for the same precision. For the AR(1) parameterisations the relative efficiency,  $e$ , is given by,

$$\begin{aligned}
e &= \frac{\{1 - \rho_\mu(1; \omega)\} \{1 + \rho_\mu(1; \alpha)\}}{\{1 - \rho_\mu(1; \alpha)\} \{1 + \rho_\mu(1; \omega)\}} \\
&= \frac{\{1 + \rho_\mu(1; \alpha)\}}{\{1 + \rho_\mu(1; \omega)\}} \frac{n}{\sigma_\varepsilon^2 \sigma_\eta^{-2} \{(n-2)(1 + \phi^2) + 2 - 2(n-1)\phi\}} \\
&\rightarrow \frac{\{1 + \rho_\mu^*(\alpha)\}}{\{1 + \rho_\mu^*(\omega)\}} \frac{\sigma_\eta^2}{\sigma_\varepsilon^2 (1 - \phi)^2}, \text{ as } n \rightarrow \infty,
\end{aligned} \tag{6}$$

where  $\rho_\mu(1; \alpha) \rightarrow \rho_\mu^*(\alpha)$ ,  $\rho_\mu(1; \omega) \rightarrow \rho_\mu^*(\omega)$  as  $n \rightarrow \infty$ . We have immediately that

$$\frac{1}{2} \frac{\sigma_\eta^2 n}{\sigma_\varepsilon^2 \{(n-2)(1 + \phi^2) + 2 - 2(n-1)\phi\}} < e < 2 \frac{\sigma_\eta^2 n}{\sigma_\varepsilon^2 \{(n-2)(1 + \phi^2) + 2 - 2(n-1)\phi\}} \tag{7}$$

and as  $n \rightarrow \infty$ ,

$$\frac{1}{2} \frac{\sigma_\eta^2}{\sigma_\varepsilon^2 (1 - \phi)^2} < e < 2 \frac{\sigma_\eta^2}{\sigma_\varepsilon^2 (1 - \phi)^2}.$$

As  $\phi \rightarrow 1$  then we have  $\rho_\mu(1; \omega), \rho_\mu^*(\omega) \rightarrow 0$  and  $\rho_\mu(1; \alpha), \rho_\mu^*(\alpha) \rightarrow 1$ , a result stated in lemma 4, Appendix. The upper bound of (7) will therefore become increasingly tight as  $\phi$  increases. The efficiency bounds depend on the parameters  $\sigma_\varepsilon^2$  and  $\sigma_\eta^2$  only through the signal to noise ratio  $\sigma_\varepsilon^2 \sigma_\eta^{-2}$ . A plot of these bounds together with the actual efficiency, of (6), is given for two different signal to noise ratios in Figures 2 and 3 for the case  $n = 50$ . The efficiency is plotted against values of  $\phi$  from 0.7 to 1. The values of  $\sigma_\eta^2$  are 0.02, while  $\sigma_\varepsilon^2$  varies from 1.0, Figure 2, to 0.1, Figure 3. It is clear from both graphs that the upper bound is close to the true efficiency, becoming increasingly accurate as  $\phi \rightarrow 1$  and as  $\sigma_\varepsilon^2 \sigma_\eta^{-2}$  becomes smaller. It is also clear that substantial gains in efficiency are achieved for moderately persistent cases.

In addition it can be seen that as  $\phi$  becomes closer to 1 we obtain autocorrelations close to 1 for the uncentered sampler and autocorrelation functions close to 0 for the centered parametrization. In cases where  $\phi \geq 0.9$  the autocorrelations are clearly far lower for the centered sampler.

### 3.3 Example 1: a Gaussian illustration

To observe the effect of the different parameterisations on the performance of the Gibbs sampler in practise, the AR(1) plus noise model described above was simulated for  $n = 100$  with  $\phi = 0.98$ ,  $\sigma_\eta^2 = 0.02$ ,  $\mu = 3.0$  and  $\sigma_\varepsilon^2 = 0.1$ . Two Gibbs samplers were set up for both the uncentered and centered samplers. The sampling was performed in the manner analysed i.e. sampling  $\mu$  given all the states and observations, then all the states given  $\mu$  and the observations. The simultaneous sampling of the states in this manner, from the densities  $f(\alpha|y, \mu)$  and  $f(\omega|y, \mu)$  for the uncentered and centered samplers respectively is performed by using the simulation

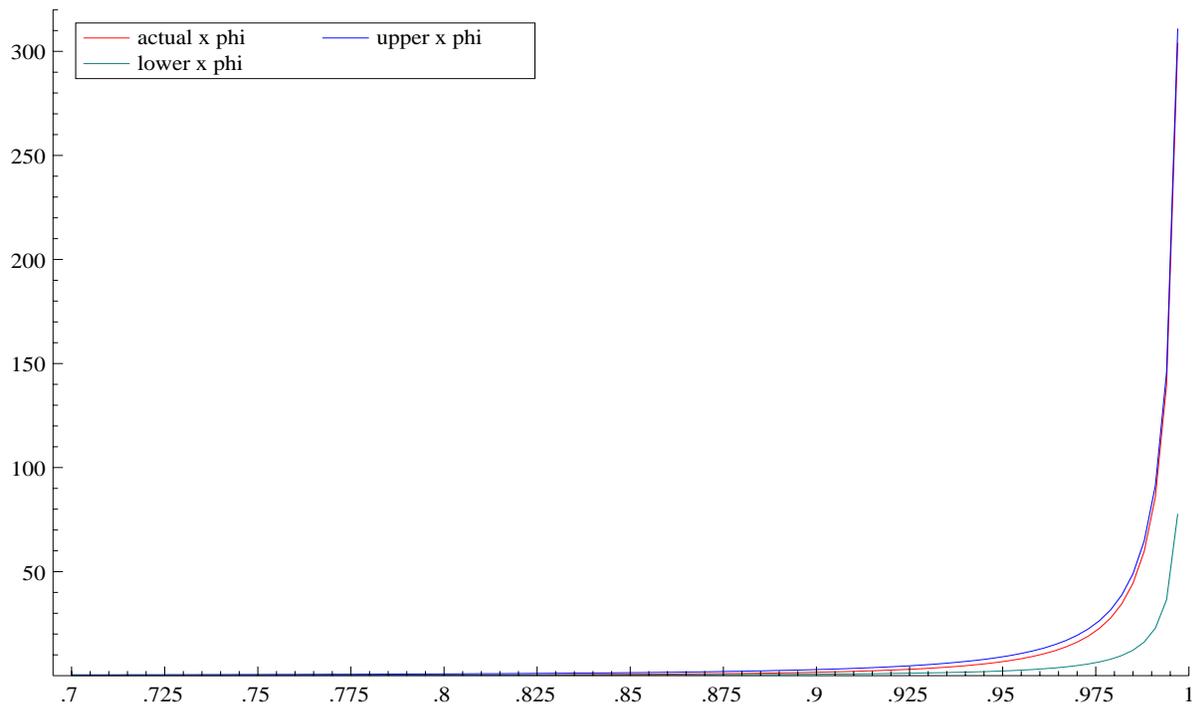


Figure 2: *Plot of relative efficiency of centred parameterisation to uncentred parameterisation for  $\sigma_\eta^2 = 0.02$ ,  $n = 50$  plotted against  $\phi$ . For the case when  $\sigma_\varepsilon^2 = 1.0$*

smoother of de Jong and Shephard (1995). The true relative efficiency, using (6), is 494.28, whilst the upper bound is 505.05.

Figure 4 shows the samples (20,000 in each case) of  $\mu$  resulting from the two parameterisations together with their autocorrelations. The samplers of  $\mu$  both result in samples which vary around the true value ( $\mu = 3.0$ ) with the same mean and variance, as we would expect. The autocorrelations are strikingly different however, with the uncentered parametrization leading to autocorrelation beyond lags of 500 and the centered parametrization leading to near independence.

In practice we use the mean of the simulations from the Gibbs sampler to estimate  $E(\mu)$ . The autocorrelation present in these simulations affects the variance of the sample mean. In order to assess the inefficiency of a Gibbs sampling scheme relative to an independent sampler we examine how much the variance is increased due to the autocorrelation of the Gibbs sampler. If, for example, the estimated variance of the sample mean Gibbs simulations is double that of the corresponding variance arising from independent samples then the inefficiency of the Gibbs sampler relative to the independent scheme is estimated as 2. In fact to estimate the ratio of the variance of the sample mean arising from  $M$  samples,  $\alpha^j$   $j = 1, \dots, M$ , from the Gibbs sampler to that arising from independent samples we use a Parzen window, see Priestley (1981, Chapter

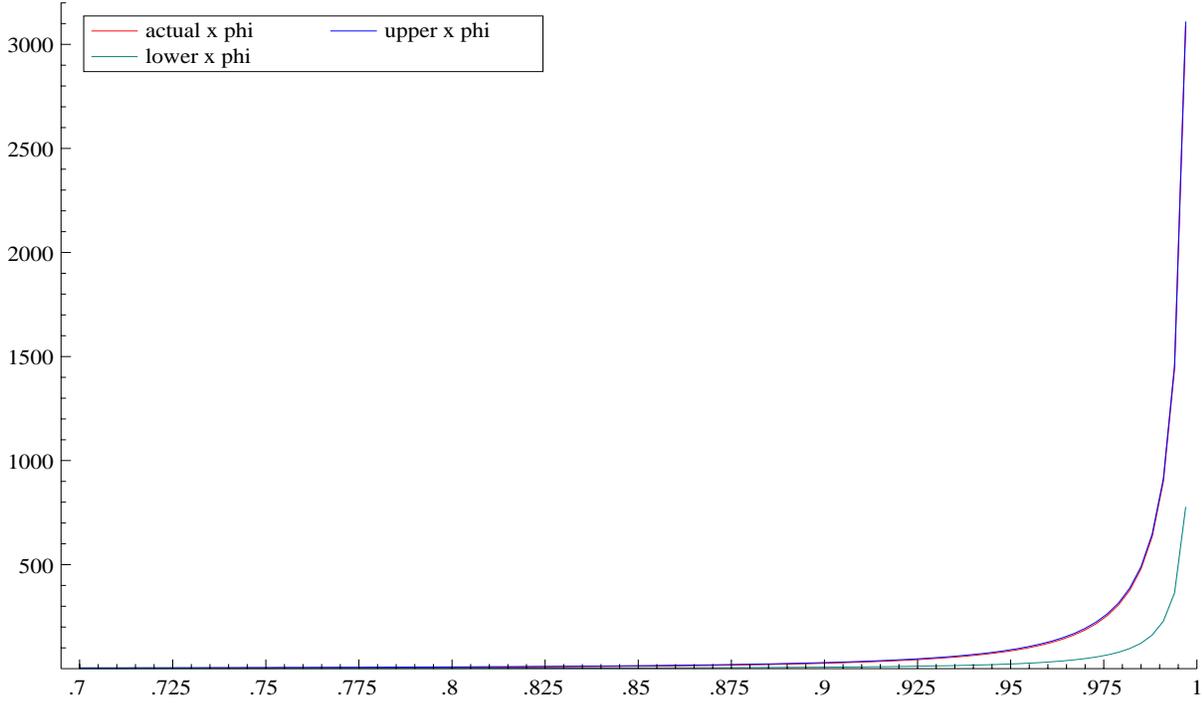


Figure 3: *Plot of relative efficiency of centred parameterisation to uncentred parameterisation for  $\sigma_\eta^2 = 0.02$ ,  $n = 50$  plotted against  $\phi$ . For the case when  $\sigma_\varepsilon^2 = 0.1$*

6). The ratio is estimated as follows,

$$\hat{R}_M = 1 + \frac{2M}{M-1} \sum_{i=1}^{B_M} K\left(\frac{i}{B_M}\right) \hat{\rho}(i), \text{ where } \hat{\rho}(i) = \frac{\hat{\Gamma}(i)}{\hat{\Gamma}(0)}$$

$$\hat{\Gamma}(i) = \frac{1}{M} \sum_{j=i+1}^M (\alpha^j - \bar{\alpha})(\alpha^{j-i} - \bar{\alpha}), \quad \bar{\alpha} = \frac{1}{M} \sum_{j=1}^M \alpha^j,$$

where the Parzen kernel is

$$\begin{aligned} K(x) &= 1 - 6x^2 + 6x^3, & x \in [0, \frac{1}{2}], \\ &= 2(1-x)^3, & x \in [\frac{1}{2}, 1], \\ &= 0, & \text{elsewhere,} \end{aligned}$$

and  $B_M$  represents band length. Here  $\hat{\rho}(i)$  is an estimate of the autocorrelation at lag  $i$ .

For this example the Parzen estimator gave estimates of 516.424 (with  $B_M = 2000$ ) and 1.016 ( $B_M = 5$ ) for the uncentered and centered sampler respectively. This indicates that the centered parametrization is about 500 times more efficient than the uncentered alternative, as anticipated.

In Figure 5 the samples of a typical state (the 50<sup>th</sup>) are shown for each of the two parameterisations. The autocorrelations are similar to those for  $\mu$ , unsurprisingly, but the variance of the uncentered state is higher than that of the centered state (this is anticipated — see the

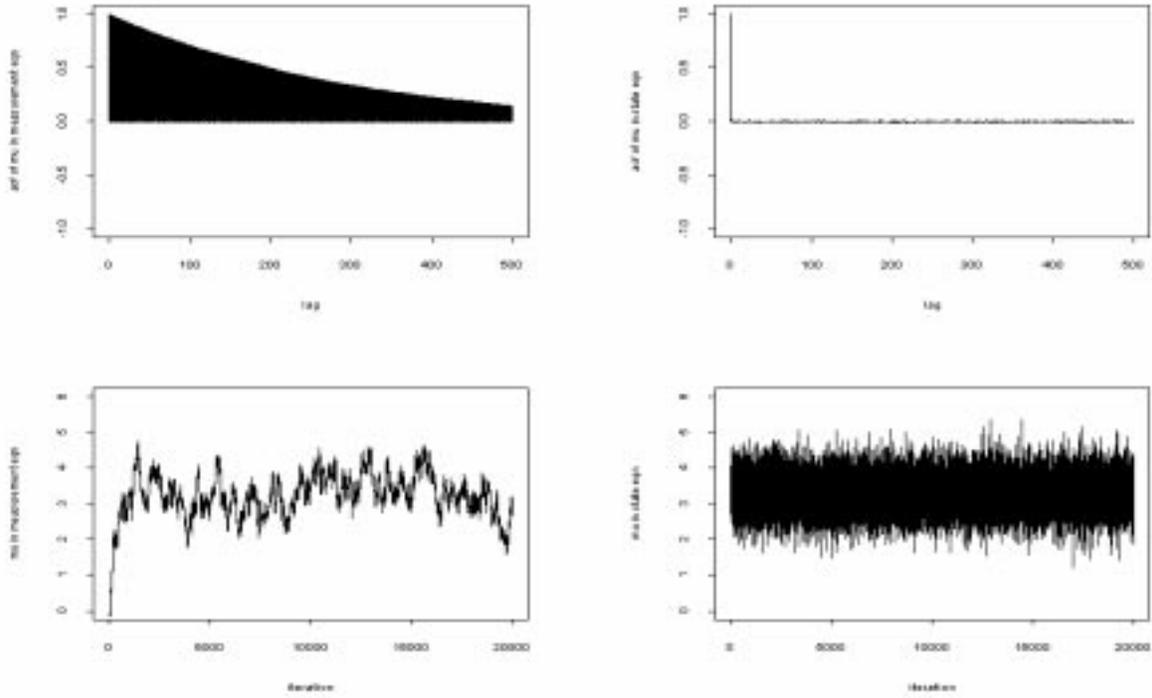


Figure 4: *Top plot are of the correlograms. Bottom plots are the corresponding sample paths for 20000 samples of  $\mu$  (true value 3),  $\phi$  and  $\sigma^2$  fixed (at 0.98 and 0.02). The left plots show the case of  $\mu$  in the measurement equation. The right plots show the corresponding plots for  $\mu$  in the state equation*

Appendix, section 6.2). The centered state varies around 3 while the uncentered state varies around 0.

### 3.4 Example 2: a stochastic volatility model

We would not necessarily wish to perform Gibbs sampling upon a GSSF model. This is because in this case we can integrate the unobserved states,  $\alpha$ , out to obtain the likelihood for any unknown set of parameters  $\theta$  via the Kalman filter (KF) obtaining

$$f(y|\theta) = \int f(y|\alpha, \theta) f(\alpha|\theta) d\alpha.$$

This enables straightforward classical or Bayesian inference. However, even for GSSF models, this is not always the easiest way to proceed from a Bayesian viewpoint. The KF delivers the log-likelihood allowing the posterior  $f(\theta|y)$  to be evaluated up to a normalizing constant. The resulting density, however, may not be easy to sample from. Further each evaluation of the KF which is necessary to compute the posterior density can be expensive. On the other hand the posterior densities for the parameters, or subsets of the parameters, conditional upon the states and observations,  $f(\theta|y, \alpha)$ , may be of closed form and simple to simulate from. Therefore

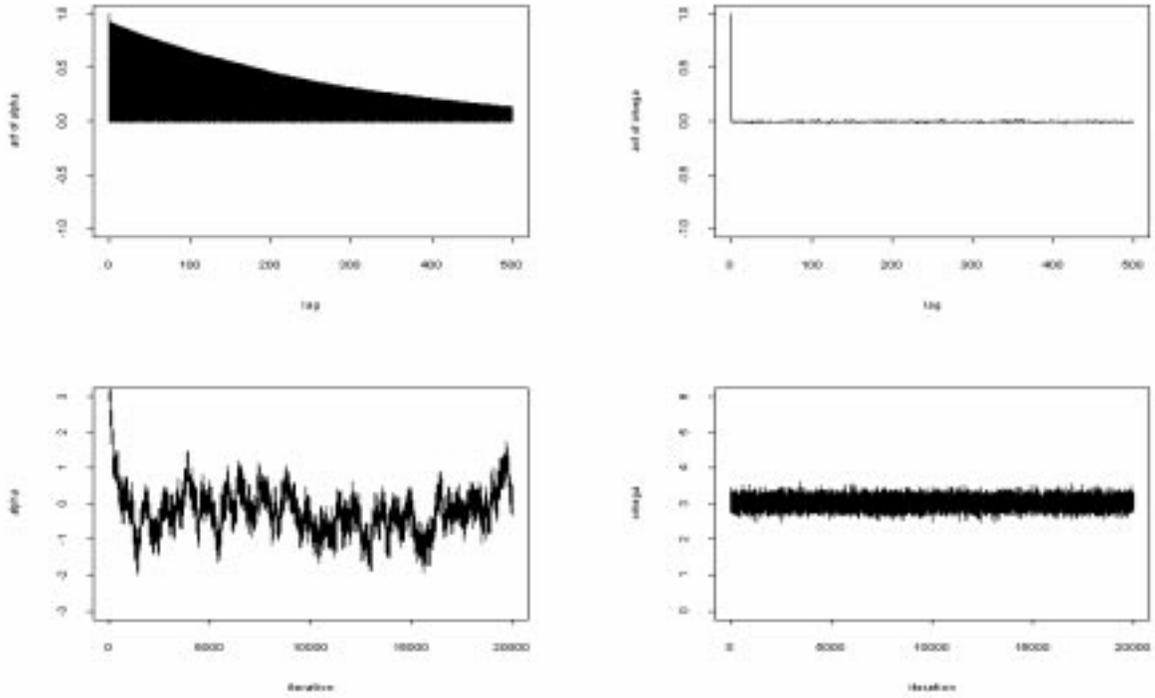


Figure 5: *Top plot are of the correlograms. Bottom plots are the corresponding sample paths for 20000 samples of  $\alpha_{50}|y; \theta$  and  $\omega_{50}|y; \theta$  respectively, with  $\phi$  and  $\sigma^2$  fixed (at 0.98 and 0.02). The left plots show the case of  $\alpha$  ( $\mu$  in the measurement equation). The right plots show the corresponding plots for  $\omega$  ( $\mu$  in the state equation)*

Gibbs sampling strategies (or Metropolis variations of them) may be extremely useful for GSSF models. Indeed, even for the AR(1) plus noise model above for inference about  $\mu$ , the centered Gibbs sampling scheme would probably be as efficient as a scheme based upon using the KF and would certainly be simpler to implement. For more GSSF models with more unknown parameters the Gibbs sampling scheme, conditioning upon the states, has been successfully applied by Fruhwirth-Schnatter (1994).

When the log of the measurement density is not quadratic in the states we have a non-Gaussian state space model. For this case it is impossible to obtain the likelihood  $f(y|\theta)$  since integration over the states cannot be performed directly. In this case the strategy of sampling the states conditional upon the parameters and then the parameters conditional upon the states is largely unavoidable. The issue of whether to use a centered or uncentered samplers is very similar to the issue for GSSF models. We now consider the issue of reparameterisation for the stochastic volatility model, a non-Gaussian state space model.

We examine two alternative parameterisations for the stochastic volatility (SV) model. This model has been considered extensively in the financial econometrics literature as it allows model-

ling the time-varying variance of returns on assets. Its major properties are reviewed by Shephard (1996). The standard SV model is given by

$$\begin{aligned} y_t &= \epsilon_t \beta \exp(\alpha_t/2), & \alpha_t &= \phi \alpha_{t-1} + \eta_t, \\ \eta_t &\sim \mathbf{N}(0, \sigma_\eta^2), & \alpha_1 &\sim \mathbf{N}\left\{0, \sigma_\eta^2/(1 - \phi^2)\right\}, \end{aligned} \quad (8)$$

and can be equivalently reparameterised as

$$\begin{aligned} y_t &= \epsilon_t \exp(\omega_t/2), & \omega_t &= \mu + \phi(\omega_{t-1} - \mu) + \eta_t, \\ \eta_t &\sim \mathbf{N}(0, \sigma_\eta^2), & \omega_1 &\sim \mathbf{N}\left\{\mu, \sigma_\eta^2/(1 - \phi^2)\right\}, \end{aligned} \quad (9)$$

without affecting the marginal distribution of  $y$  provided  $\mu = \log(\beta^2)$ . It is important to note that a flat prior on  $\mu$  is equivalent to a flat prior on  $\log(\beta)$ . In each case we are interested in sampling the states and  $\beta$  (or  $\mu$ ) keeping the other parameters fixed for the purposes of this example. Since the conditional distribution of  $\alpha|y, \beta$  and  $\omega|y, \mu$  is of non-standard highly multivariate form we use a Metropolis method, sampling large blocks of states (not all the states) simultaneously. The exact method will not be detailed here but the interested reader is referred to Shephard and Pitt (1997). Although this is a Metropolis method, rather than direct sampling, on the conditional density of the states, the acceptance rate (switching probability) is over 90%. If the rate were guaranteed to be 100% we would have a Gibbs method as we would be sampling directly from the relevant conditional density. However, since the rate is so high the correlation structure which the Gibbs sampling analysis of this paper focuses on is far more important than the slight retardation in efficiency due to the small number of Metropolis rejections.

It is not immediately clear how to obtain any insight into which of these parameterisations is better for the SV model. However, Harvey, Ruiz, and Shephard (1994) show that the measurement equations (8) and (9) can both be linearised (at the expense of Gaussianity) as

$$\log y_t^2 = \log(\beta^2) + \alpha_t + \log \epsilon_t^2, \quad \text{and} \quad \log y_t^2 = \omega_t + \log \epsilon_t^2, \quad (10)$$

so the similarity between the non-linear models and the GSSF is apparent. More general models are not necessarily amenable to direct linearisation in the manner shown above. However, expansion methods in these cases (see Gilks and Roberts (1996)) still enable useful comparisons with the GSSF. For example, (8) implies

$$E(y_t^2 | \alpha_t) = \beta^2 \exp(\alpha_t) \simeq \beta^2(1 + \alpha_t) = \beta^2 + \beta^2 \alpha_t.$$

In this case  $\beta^2$  represents the mean shift of the measurement equation and also a multiplicative term for  $\alpha_t$ . Since linearisations (in terms of a fixed function of  $\beta$  and in terms of the

states) can be performed on the measurement equation fairly straightforwardly this implies that comparison with the GSSF is informative. This is certainly true for the simple SV model as the linearisation can be performed directly.

The conditional distributions for  $\beta$  and  $\mu$  are given as follows. By assuming a flat prior for  $\log \beta$  we achieve the posterior

$$\beta^2 | y, \alpha \sim \chi_n^{-2} \sum_{t=1}^n y_t^2 \exp(-\alpha_t).$$

Similarly for the centered parameterization we obtain,  $\mu | y, \omega \sim \mathbf{N}(b/a, \sigma_\eta^2/a)$ , where

$$a = (n-1)(1-\phi)^2 + (1-\phi^2); \quad s = \sum_{t=2}^n (\omega_t - \phi\omega_{t-1}); \quad b = (1-\phi)s + \omega_1(1-\phi^2).$$

These setups are equivalent since we assume a flat prior for  $\mu$  and  $\log \beta$ .

We shall now compare the two MCMC methods using a dataset consisting of the daily returns on the Pound Sterling/US Dollar exchange rate from 1/10/81 to 28/6/85 involving, in total, 944 observations. The number of “knots” (states which remain fixed for a particular sweep of the MCMC method) is set to 10 in each case. The parameters  $\phi, \sigma_\eta^2$  are set to be fixed at 0.98 and 0.02, values which we have found to be reasonable (from sampling these parameters in other MCMC simulations). The results of 10,500 samples are shown in Figure 6. The samples of  $\mu$  have been transformed to yield  $\beta = e^{\mu/2}$ . From the plot of the samples, and the resulting correlogram, it is clear that the centered setup yields substantial efficiency gains.

In further experiments we also found that similar efficiency gains are encountered even when the parameters  $\phi$  and  $\sigma_\eta^2$  are sampled rather than remaining fixed as they have been for this analysis.

## 4 Conclusion

The second section of this paper provides an analytic expression for the relationship between the convergence rate of the Gibbs sampler and the parameters of the AR(1) plus noise model. This gives us considerable insight into the performance of the single move Gibbs sampling scheme. It can be seen that a high degree of persistence for this model ( $\phi$  close to 1,  $\sigma_\eta^2$  close to 0) leads to extremely bad convergence. Effective blocking is therefore essential for these models in order to achieve reasonable convergence rates. Blocking strategies for state space forms have been explored by Carter and Kohn (1994), de Jong and Shephard (1995) and more recently by Shephard and Pitt (1997).

The third section of this paper indicates that a centered parameterization is far more effective than an uncentered parameterization in terms of the efficiency of the Gibbs sampler for a

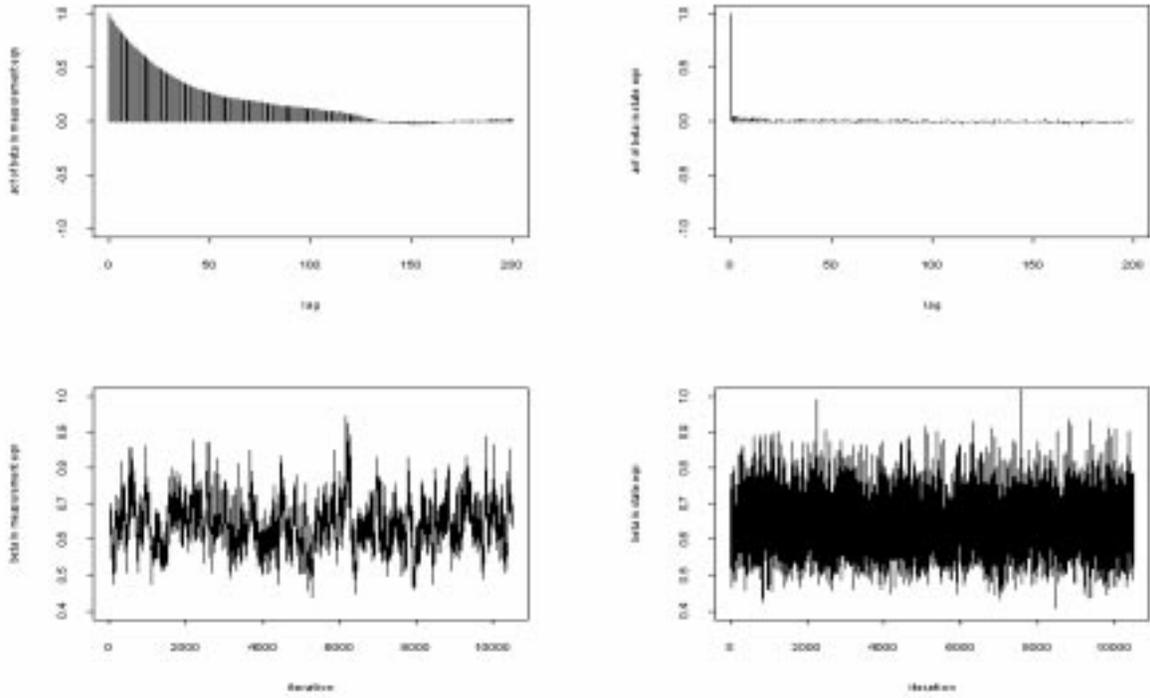


Figure 6: *Stochastic volatility case. Top plot are correlograms, bottom graphs show the sample path for 10200 samples of  $\beta, \phi$  and  $\sigma^2$  fixed (at 0.98 and 0.02). The left plots show the case of  $\beta$  in the measurement equation. The right plots show the corresponding plots for  $\beta$  ( $\exp(\mu/2)$ ) in the state equation*

reasonably persistent model. Intuitively, we expect similar results for non-Gaussian state space models and this is borne out by considering the SV model.

The Gibbs sampler is now accepted as a standard tool for a Bayesian statistician. However, these results confirm that careful consideration of the properties of the time series model can lead to very significant improvements in the performance of these methods.

## 5 Acknowledgments

We would like to thank the ESRC for their financial support for this research under the grant “Estimation via simulation in econometrics”. We also thank David Handscomb and Brian Stewart for various suggestions, and a referee for helpful comments on a previous draft.

## 6 Appendix

In this Appendix we present the proofs of theorems and lemmas stated in the paper. We also provide the marginal and conditional distributions for the centered and uncentered parameterisations.

## 6.1 Analytic convergence rate

### Proof of Theorem 1:

Using Theorem 5 of Roberts and Sahu (1997) we have that  $\rho(B) = \{\lambda(A)\}^2$ , since  $Q$  is tridiagonal. From section 2, we have

$$A = I - \text{diag}(Q_{11}^{-1}, \dots, Q_{nn}^{-1})Q = \begin{pmatrix} 0 & b & 0 & \dots & 0 \\ a & 0 & a & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & a & 0 & a \\ 0 & \dots & 0 & b & 0 \end{pmatrix},$$

where the dimension of  $A$  is  $n \times n$  and  $a = \phi / (1 + \phi^2 + \sigma_\eta^2 \sigma_\varepsilon^{-2})$ ,  $b = \phi / (1 + \sigma_\eta^2 \sigma_\varepsilon^{-2})$ . So

$$D_n \equiv \det(A - \lambda I) = \begin{vmatrix} -\lambda & b & 0 & \dots & 0 \\ a & -\lambda & a & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & a & -\lambda & a \\ 0 & \dots & 0 & b & -\lambda \end{vmatrix}.$$

Let us examine two simplifications of the above determinant. Firstly let us define

$$\Lambda_n \equiv \begin{vmatrix} -\lambda & a & 0 & \dots & 0 \\ a & -\lambda & a & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & a & -\lambda & a \\ 0 & \dots & 0 & a & -\lambda \end{vmatrix}.$$

Secondly let

$$\Lambda_n^* \equiv \begin{vmatrix} -\lambda & b & 0 & \dots & 0 \\ a & -\lambda & a & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & a & -\lambda & a \\ 0 & \dots & 0 & a & -\lambda \end{vmatrix}.$$

It can straightforwardly be seen that the two determinants above satisfy the following difference equations

$$\Lambda_{n+2} + \lambda \Lambda_{n+1} + a^2 \Lambda_n = 0, \quad \text{and} \quad \Lambda_{n+2}^* + \lambda \Lambda_{n+1}^* + a^2 \Lambda_n^* = 0.$$

Let us first consider solving the first difference equation to yield an expression for the characteristic polynomial for any  $n$ . Then we have initial conditions  $\Lambda_0 = 1$ ,  $\Lambda_1 = -\lambda$ . This equation can be solved via Chebychev polynomials. Let us define  $A(t) = \sum_{n=0}^{\infty} \Lambda_n t^n$  and sum the recursion equation as follows

$$\sum_{n=0}^{\infty} t^{n+2} \Lambda_{n+2} + \lambda \sum_{n=0}^{\infty} t^{n+2} \Lambda_{n+1} + a^2 \sum_{n=0}^{\infty} t^{n+2} \Lambda_n = 0. \quad (11)$$

Then we obtain

$$A(t)(1 + \lambda t + a^2 t^2) = \Lambda_0 + t\Lambda_1 + \lambda t\Lambda_0 = 1,$$

implying  $A(t) = 1/(1 + \lambda t + a^2 t^2)$ . Now the generating equation for Chebychev polynomials is

$$\frac{1}{(1 - 2Tx + T^2)} = \sum_{n=0}^{\infty} T^n U_n(x),$$

where  $U_s(x)$  is the  $s^{\text{th}}$  Chebychev polynomial of the second kind defined as

$$U_s(x) = \sin \left\{ (s+1) \cos^{-1}(x) \right\} / \sin \left\{ \cos^{-1}(x) \right\}.$$

Hence, we have to find values of  $T$  and  $x$  such that  $1 + \lambda t + a^2 t^2 = 1 - 2Tx + T^2$ . Values which satisfy this are  $T = -ta$ , and  $x = \lambda/2a$ . Hence we obtain

$$A(t) = \sum_{n=0}^{\infty} (-1)^n a^n t^n U_n \left( \frac{\lambda}{2a} \right).$$

By comparing coefficients of  $t^n$  to those of  $A(t)$ , we obtain  $\Lambda_n = (-1)^n a^n U_n \{ \lambda / (2a) \}$ . The  $n$  solutions of this characteristic equation satisfy  $\sin \{ (n+1) \cos^{-1} \{ \lambda / (2a) \} \} = 0$ . Hence the  $n$  distinct eigenvalues are given by

$$\lambda = 2a \cos \left( \frac{k\pi}{n+1} \right), \text{ where } k = 1, 2, \dots, n.$$

Clearly the largest eigenvalue is  $\lambda_1 = 2a \cos \{ \pi / (n+1) \}$ .

We have that  $b = (1 + \varepsilon)a$  where we define  $\varepsilon = \phi^2 / (1 + \sigma_\eta^2 \sigma_\varepsilon^{-2})$ . Now let  $A^*(t) = \sum_{n=0}^{\infty} \Lambda_n^* t^n$ , the generating function for  $\Lambda_n^*(\lambda)$ . Our initial conditions are now different, changing to  $\Lambda_0^* = b/a = 1 + \varepsilon$ ,  $\Lambda_1^* = -\lambda$ . Summing as for (11) we now have

$$A^*(t)(1 + \lambda t + a^2 t^2) = \Lambda_0^* + t\Lambda_1^* + \lambda t\Lambda_0^* = (1 + \varepsilon) - \lambda t + \lambda t(1 + \varepsilon) = (1 + \varepsilon) + \varepsilon \lambda t.$$

Therefore

$$A^*(t) = \frac{(1 + \varepsilon) + \varepsilon \lambda t}{(1 + \lambda t + a^2 t^2)} = \{(1 + \varepsilon) + \varepsilon \lambda t\} A(t) = \{(1 + \varepsilon) + \varepsilon \lambda t\} \sum_{n=0}^{\infty} \Lambda_n t^n.$$

Hence we obtain

$$\Lambda_0^* = (1 + \varepsilon)\Lambda_0 \quad \text{and} \quad \Lambda_n^* = (1 + \varepsilon)\Lambda_n + \varepsilon \lambda \Lambda_{n-1}, \quad n > 0.$$

But it can also be seen that the characteristic equation of interest,  $D_n = -\lambda \Lambda_{n-1}^* - ab \Lambda_{n-2}^*$ . Using the difference equation for  $\Lambda_n^*$  we obtain  $D_n = \Lambda_n^* - a^2 \varepsilon \Lambda_{n-2}^*$ . Hence, expressed in terms of  $\Lambda_n$  we have

$$D_n = (1 + \varepsilon)\Lambda_n + \varepsilon \lambda \Lambda_{n-1} - a^2 \varepsilon \{ (1 + \varepsilon)\Lambda_{n-2} + \varepsilon \lambda \Lambda_{n-3} \}.$$

Again noting from the difference equation for  $\Lambda_n$ , that

$$\lambda\Lambda_{n-1} = -a^2\Lambda_{n-2} - \Lambda_n \quad \text{and} \quad \lambda\Lambda_{n-3} = -a^2\Lambda_{n-4} - \Lambda_{n-2},$$

we obtain, after rearranging,

$$D_n = \Lambda_n - 2a^2\epsilon\Lambda_{n-2} + a^4\epsilon^2\Lambda_{n-4}.$$

Hence we have that the solution of  $D_n(\lambda) = 0$  satisfies

$$f(\theta) = \sin(n+1)\theta - 2y\sin(n-1)\theta + y^2\sin(n-3)\theta = 0, \quad (12)$$

where  $\theta = \cos^{-1}(\lambda/2a)$ ,  $y = a^2\epsilon$  and  $0 < \theta < \pi$ . It can be seen that the largest eigenvalue  $\lambda_1$ , say, corresponds to the smallest solution  $\theta_1$  of (12). We can show that  $\theta_1$  lies in the interval  $[0, \pi/(n+1)]$ . When  $\theta = 0$ ,  $f(0) = 0$  and

$$\begin{aligned} f'(\theta) &= (n+1)\cos(n+1)\theta - 2y(n-1)\cos(n-1)\theta + y^2(n-3)\cos(n-3)\theta \\ \text{so } f'(0) &= (n+1) - 2y(n-1) + y^2(n-3) \\ \text{so } f'(0) &> 0 \text{ provided } y < 1 \text{ or } y > (n+1)/(n-3). \end{aligned}$$

Now

$$y = a^2\epsilon = \frac{\phi^2}{(1 + \phi^2 + \sigma_\eta^2\sigma_\epsilon^{-2})^2} \times \frac{\phi^2}{1 + \sigma_\eta^2\sigma_\epsilon^{-2}}.$$

So it is certainly the case that  $y < 1$  when  $\phi < 1$ . In fact  $y < \frac{1}{4}$ . Hence  $\forall \delta > 0 \exists 0 < \epsilon < \delta$  s.t.  $f(\epsilon) > 0$ . This demonstrates that  $f(\theta)$  is positive immediately after  $\theta = 0$ . Now at  $\theta = \pi/(n+1)$  we require that

$$f(\theta) = -2y\sin\left\{\frac{(n-1)\pi}{n+1}\right\} + y^2\sin\left\{\frac{(n-3)\pi}{n+1}\right\} < 0$$

or

$$\sin\left\{\frac{(n-1)\pi}{n+1}\right\} / \sin\left\{\frac{(n-3)\pi}{n+1}\right\} > \frac{y}{2}$$

From lemma 1, we have that the *LHS*  $> 0.5$  which verifies the inequality since  $y < 0.5$  provided  $\phi < 1$ . So the smallest root  $\theta_1$ , say, of (12) satisfies  $0 < \theta_1 < \pi/(n+1)$ . Since the largest root of  $D_n$  is  $\lambda_1 = 2a\cos\theta_1$  we obtain

$$2a\cos\left(\frac{\pi}{n+1}\right) < \lambda_1 < 2a.$$

Hence  $\rho(B) = \lambda_1^2$  satisfies

$$4a^2\cos^2\left(\frac{\pi}{n+1}\right) < \rho(B) < 4a^2.$$

Hence as  $n \rightarrow \infty$ ,

$$\rho(B) \rightarrow 4a^2 = 4\frac{\phi^2}{(1 + \phi^2 + \sigma_\eta^2\sigma_\epsilon^{-2})^2}.$$

Also as  $\phi \rightarrow 1$ ,  $n \rightarrow \infty$ ,  $\sigma_{\eta}^2 \sigma_{\varepsilon}^{-2} \rightarrow 0$  then  $\rho(B) \rightarrow 1$ .

**Lemma 1:** We establish the result that,

$$\begin{aligned} \sin \left\{ \frac{(n-1)\pi}{n+1} \right\} / \sin \left\{ \frac{(n-3)\pi}{n+1} \right\} &> \frac{1}{2} \text{ and} \\ \lim_{n \rightarrow \infty} \sin \left\{ \frac{(n-1)\pi}{n+1} \right\} / \sin \left\{ \frac{(n-3)\pi}{n+1} \right\} &= \frac{1}{2}. \end{aligned}$$

**Proof of Lemma 1:**

We shall establish the limiting result first. Using l'Hopital's theorem, we have

$$\lim_{n \rightarrow \infty} \sin \left\{ \frac{(n-1)\pi}{n+1} \right\} / \sin \left\{ \frac{(n-3)\pi}{n+1} \right\} = \lim_{n \rightarrow \infty} \frac{\cos \left\{ \frac{(n-1)\pi}{n+1} \right\} \frac{2\pi}{(n+1)^2}}{\cos \left\{ \frac{(n-3)\pi}{n+1} \right\} \frac{4\pi}{(n+1)^2}} = \frac{1}{2}.$$

To prove the first assertion we now show that the derivative of the expression with respect to  $n$  is always negative. Now

$$\frac{d}{dn} \left[ \frac{\sin \left\{ \frac{(n-1)\pi}{n+1} \right\}}{\sin \left\{ \frac{(n-3)\pi}{n+1} \right\}} \right] = \frac{\sin \left\{ \frac{(n-3)\pi}{n+1} \right\} \cos \left\{ \frac{(n-1)\pi}{n+1} \right\} \frac{2\pi}{(n+1)^2} - \sin \left\{ \frac{(n-1)\pi}{n+1} \right\} \cos \left\{ \frac{(n-3)\pi}{n+1} \right\} \frac{4\pi}{(n+1)^2}}{\sin^2 \left\{ \frac{(n-3)\pi}{n+1} \right\}}. \quad (13)$$

The denominator will clearly always be positive while the numerator,  $Num$ , may be simplified as

$$Num = \frac{2\pi}{(n+1)^2} (\sin a \cos b - 2 \sin b \cos a),$$

where  $a = (n-3)\pi/(n+1)$  and  $b = (n-1)\pi/(n+1)$ . Now

$$\begin{aligned} &\sin a \cos b - 2 \sin b \cos a \\ &= \sin \left( b - \frac{2\pi}{n+1} \right) \cos b - 2 \sin b \cos \left( b - \frac{2\pi}{n+1} \right) \\ &= \left\{ \sin b \cos \left( \frac{2\pi}{n+1} \right) - \cos b \sin \left( \frac{2\pi}{n+1} \right) \right\} \cos b - 2 \left\{ \cos b \cos \left( \frac{2\pi}{n+1} \right) + \sin b \sin \left( \frac{2\pi}{n+1} \right) \right\} \sin b. \end{aligned}$$

Noting that  $\sin \{2\pi/(n+1)\} = \sin b$  and  $\cos \{2\pi/(n+1)\} = -\cos b$  the expression simplifies to  $-2 \sin b \sin^2 b$ , which is negative for all  $n$ . The denominator of (13),  $Den$ , simplified to

$$Den = \sin^2 a = \left\{ \sin b \cos \left( \frac{2\pi}{n+1} \right) - \cos b \sin \left( \frac{2\pi}{n+1} \right) \right\}^2 = 4 \sin^2 b \cos^2 b.$$

Hence

$$\frac{d}{dn} \left[ \sin \left\{ \frac{(n-1)\pi}{n+1} \right\} / \sin \left\{ \frac{(n-3)\pi}{n+1} \right\} \right] = -\frac{1}{2} \frac{\sin b}{\cos^2 b}.$$

This will clearly be negative for all  $n$  and will tend towards 0 as  $n \rightarrow \infty$ . The negative derivative together with the limiting result establishes the inequality

$$\sin \left\{ \frac{(n-1)\pi}{n+1} \right\} / \sin \left\{ \frac{(n-3)\pi}{n+1} \right\} > \frac{1}{2}.$$

**Lemma 2:** We note the result that,

$$D^{-1}VD^{-1} = D^{-1}(\sigma_\varepsilon^{-2}D + I)^{-1} = D^{-1} - (\sigma_\varepsilon^2I + D)^{-1} = D^{-1} - \sigma_\varepsilon^{-2}VD^{-1}.$$

**Proof of Lemma 2:**

We use the result (Harvey (1993, p 104)) that if  $M = (A + BCB')^{-1}$  where  $A, B$  and  $C$  are non-singular conformable matrices, then

$$M = A^{-1} - A^{-1}B(C^{-1} + B'A^{-1}B)^{-1}B'A^{-1}.$$

Hence, using a special case,

$$(I + \sigma_\varepsilon^{-2}D)^{-1} = I - (\sigma_\varepsilon^2D^{-1} + I)^{-1},$$

so that, as required,

$$D^{-1}(\sigma_\varepsilon^{-2}D + I)^{-1} = D^{-1} - (\sigma_\varepsilon^2I + D)^{-1}.$$

**Lemma 3:** For the AR(1) plus noise model

$$\frac{1}{n}\mathbf{1}'D^{-1}\mathbf{1}\sigma_\varepsilon^2 = \frac{1}{n}\sigma_\varepsilon^2\sigma_\eta^{-2} \left\{ (n-2)(1+\phi^2) + 2 - 2(n-1)\phi \right\}.$$

**Proof of Lemma 3:** This result can easily be verified by examining the  $n \times n$  matrix  $D^{-1}$ ,

$$D^{-1} = 1/\sigma_\eta^2 \begin{pmatrix} 1 & -\phi & 0 & \cdots & 0 \\ -\phi & 1+\phi^2 & -\phi & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\phi & 1+\phi^2 & -\phi \\ 0 & \cdots & 0 & -\phi & 1 \end{pmatrix}.$$

Since the expression  $\mathbf{1}'D^{-1}\mathbf{1}$  sums all the elements of the matrix we obtain the required result.

**Lemma 4:** For the AR(1) plus noise model

$$\rho_\mu(1; \alpha) \longrightarrow 1 \text{ as } \phi \longrightarrow 1 \quad \text{and} \quad \rho_\mu(1; \omega) \longrightarrow 0 \text{ as } \phi \longrightarrow 1. \quad (14)$$

**Proof of Lemma 4:**

We have

$$\rho_\mu(1; \alpha) = \frac{1}{n}\mathbf{1}'[I - VD^{-1}]\mathbf{1} = 1 - \frac{1}{n}\mathbf{1}'VD^{-1}\mathbf{1}.$$

But

$$\frac{1}{n}\mathbf{1}'VD^{-1}\mathbf{1} = \frac{\sigma_\varepsilon^2}{n}\mathbf{1}'D^{-1}(\sigma_\varepsilon^2D^{-1} + I)^{-1}\mathbf{1} \leq \frac{\sigma_\varepsilon^2}{n}\mathbf{1}'D^{-1}\mathbf{1}.$$

Since from lemma 4,  $\mathbf{1}'D^{-1}\mathbf{1} = \sigma_\eta^{-2} \left\{ (n-2)(1+\phi^2) + 2 - 2(n-1)\phi \right\} \longrightarrow 0$  as  $\phi \longrightarrow 1$  we obtain the required result for  $\rho_\mu(1; \alpha)$ .

Similarly

$$\rho_\mu(1; \omega) = \frac{\mathbf{1}'D^{-1}VD^{-1}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}} = \frac{\mathbf{1}'D^{-1}(D^{-1} + \sigma_\varepsilon^{-2}I)^{-1}D^{-1}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}} \leq \frac{\sigma_\varepsilon^2\mathbf{1}'D^{-2}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}}.$$

We have

$$D^{-2} = \sigma_\eta^{-4} \begin{pmatrix} 1 + \phi^2 & -\phi(2 + \phi^2) & \phi^2 & 0 & \dots & 0 \\ -\phi(2 + \phi^2) & 1 + 4\phi^2 + \phi^4 & -2\phi(1 + \phi^2) & \phi^2 & \dots & \vdots \\ \phi^2 & -2\phi(1 + \phi^2) & 1 + 4\phi + \phi^4 & -2\phi(1 + \phi^2) & & \\ \vdots & & & & \ddots & -\phi(2 + \phi^2) \\ 0 & \dots & \dots & \phi^2 & -\phi(2 + \phi^2) & 1 + \phi^2 \end{pmatrix},$$

implying

$$\frac{\sigma_\varepsilon^2\mathbf{1}'D^{-2}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}} = \frac{\sigma_\varepsilon^2 \{(n-2)\phi^4 - 4(n-2)\phi^3 + 6(n-2)\phi^2 + 2\phi^2 - 4(n-1)\phi + n\}}{\sigma_\eta^2 \{(n-2)(1 + \phi^2) + 2 - 2(n-1)\phi\}}.$$

Using l'Hopital's rule results in

$$\begin{aligned} & \lim_{\phi \rightarrow 1} \frac{\{(n-2)\phi^4 - 4(n-2)\phi^3 + 6(n-2)\phi^2 + 2\phi^2 - 4(n-1)\phi + n\}}{\{(n-2)(1 + \phi^2) + 2 - 2(n-1)\phi\}} \\ &= \frac{4(n-2) - 12(n-2) + 12(n-2) + 4 - 4(n-1)}{2(n-2) - 2(n-1)} = \frac{0}{-2} = 0, \end{aligned}$$

yielding the required result.

## 6.2 Marginal and conditional distributions for different parameterisations

The marginal distribution of  $\mu|y$  for both models is

$$\mu|y \sim \mathbf{N} \left\{ (\mathbf{1}'\Sigma^{-1}y)(\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}, (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1} \right\}, \quad \text{where} \quad \Sigma = \sigma_\varepsilon^2 I_n + D. \quad (15)$$

So  $\mu\mathbf{1}|y \sim \mathbf{N}(\mathbf{1}\hat{\mu}, \sigma_\mu^2\mathbf{1}\mathbf{1}')$ , where for convenience I shall now denote  $\hat{\mu} = (\mathbf{1}'\Sigma^{-1}y)(\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}$  and  $\sigma_\mu^2 = (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}$ . Assuming a flat prior on  $\mu$ , the conditional distribution for  $\mu|y, \alpha$  for the uncentered first model is given by  $\mu|y, \alpha \sim \mathbf{N}(\bar{y} - \bar{\alpha}, \sigma_\varepsilon^2/n)$ . The conditional distribution for  $\mu|y, \omega$  is given by

$$\mu|y, \omega \sim \mathbf{N} \left\{ (\mathbf{1}'D^{-1}\omega)(\mathbf{1}'D^{-1}\mathbf{1})^{-1}, (\mathbf{1}'D^{-1}\mathbf{1})^{-1} \right\}.$$

In the simple AR(1) case this reduces simply to,  $\mu|y, \omega \sim \mathbf{N}(q/p, \sigma_\eta^2/p)$ , where

$$p = (n-1)(1 - \phi)^2 + (1 - \phi^2) \quad \text{and} \quad q = \omega_1(1 - \phi^2) + (1 - \phi) \sum_{i=2}^n (\omega_i - \phi\omega_{i-1}).$$

The conditional distribution of  $\omega|y, \mu$  is given by  $\omega|y, \mu \sim \mathbf{N}(Vb, V)$ , where

$$V = (\sigma_\varepsilon^{-2}I_n + D^{-1})^{-1} = \sigma_\varepsilon^2\Sigma^{-1}D \quad \text{and} \quad b = \sigma_\varepsilon^{-2}y + D^{-1}\mathbf{1}\mu.$$

The marginal distribution  $\omega|y$  is

$$\omega|y \sim N(V\hat{b}, V + \sigma_\mu^2 VD^{-1}\mathbf{1}\mathbf{1}'D^{-1}V) \quad \text{and} \quad \hat{b} = \sigma_\epsilon^{-2}y + D^{-1}\mathbf{1}\hat{\mu}.$$

Similarly the conditional distribution of  $\alpha|y, \mu$  is given by  $\alpha|y, \mu \sim N(Vb - \mathbf{1}\mu, V)$ . Hence

$$\alpha|y \sim N\{V\hat{b} - \mathbf{1}\hat{\mu}, V + \sigma_\mu^2(VD^{-1} - I)\mathbf{1}\mathbf{1}'(VD^{-1} - I)'\}.$$

These results are analogous to those results of Gelfand, Sahu, and Carlin (1995) for the hierarchical linear model.

## References

- Carlin, B. P., N. G. Polson, and D. Stoffer (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modelling. *J. Am. Statist. Assoc.* 87, 493–500.
- Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika* 81, 541–53.
- de Jong, P. (1989). Smoothing and interpolation with the state space model. *J. Am. Statist. Assoc.* 84, 1085–8.
- de Jong, P. and N. Shephard (1995). The simulation smoother for time series models. *Biometrika* 82, 339–50.
- Fruhwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Analysis* 15, 183–202.
- Gelfand, A. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* 85, 398–409.
- Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* 82, 479–488.
- Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions, PAMI* 6, 721–41.
- Gilks, W. K., S. Richardson, and D. J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
- Gilks, W. K. and G. O. Roberts (1996). Strategies for improving MCMC. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.
- Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

- Harvey, A. C. (1993). *Time Series Models* (2nd ed.). Hemel Hempstead: Harvester Wheatsheaf.
- Harvey, A. C., E. Ruiz, and N. Shephard (1994). Multivariate stochastic variance models. *Rev. Economic Studies* 61, 247–64.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, Transactions ASMA, Series D* 82, 35–45.
- Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.
- Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *J. R. Statist. Soc. B* 39, 172–212.
- Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.
- Roberts, G. O. and S. K. Sahu (1996). Rate of convergence of the Gibbs sampler by Gaussian approximation. Technical report 96-21, Statistical Lab, University of Cambridge.
- Roberts, G. O. and S. K. Sahu (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. R. Statist. Soc. B* 59, 291–317.
- Schotman, P. C. (1994). Priors for the AR(1) model: parameterisation issues and time series considerations. *Econometric Theory* 10, 579–595.
- Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika* 81, 115–31.
- Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In D. R. Cox, O. E. Barndorff-Nielsen, and D. V. Hinkley (Eds.), *Time Series Models in Econometrics, Finance and Other Fields*, pp. 1–67. London: Chapman & Hall.
- Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* 84. Forthcoming.
- Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Assoc.* 82, 528–50.