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

3rd June 1996

Abstract

In this paper we obtain a closed form expression for the convergence rate of the Gibbs sampler applied to an AR(1) plus noise model in terms of the parameters of the model. We also 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; 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 Geman and Geman (1984). The methods were discussed for more widespread statistical implementation by Ripley (1987) and Gelfand and Smith (1990). Many examples of the current use of MCMC techniques are given in the booklength review by Gilks, Richardson and Spiegelhalter (1996).

Gibbs sampling is now increasingly used as a tool to fit non-Gaussian state space models. 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 state space models. More recently Fruhwirth-Schnatter (1994) used Gibbs sampling to draw from the posterior density of the parameters in state space models. 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).

In this paper we consider how the performance of the single move Gibbs sampler applied to the states of the AR(1) plus noise model varies with the parameters of the model. A closed form relationship is provided. We also consider the effect of reparameterisation when the parameters of a state space model are sampled as well as the states. In particular we consider strategies of "centering" which have been explored outside the time series literature.

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 in the case of the Stochastic Volatility model using data on exchange rate returns.

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

Roberts and Sahu (1995) examine convergence rates for different blocking strategies applied to multivariate Gaussian densities. Their general framework is outlined briefly below.

The rate of convergence, ρ, is defined as follows. Suppose that $\{x^{(i)}, i = 0, 1, ..\}$ is a Markov chain with stationary density function $h(x)$. Let f be a square h-integrable function of x and $h(f)$ denote the expectation of f under the target density h. Let ρ be the minimum number such that for all square h-integrable functions f, and for all $r < \rho$, we can find a function $V(.) \geq 1$ such that $h(V) < \infty$ and

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

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

We can calculate ρ as follows. Consider x subdivided into s blocks each of length r_i, $i = 1, ..., s$. We also partition Σ and $Q = \Sigma^{-1}$ according to the Gibbs sampler blocking scheme being considered, i.e.,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1s} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{s1} & \Sigma_{s2} & \cdots & \Sigma_{ss} \end{pmatrix}, Q = \begin{pmatrix} Q_{11} & Q_{12} & \dots & Q_{1s} \\ Q_{21} & Q_{22} & \cdots & Q_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{s1} & Q_{s2} & \dots & Q_{ss} \end{pmatrix},$$

where Σ_{ij} and Q_{ij} are matrices of order $r_i \times r_j$. Let $A = I - diag(Q_{11}^{-1}, ..., Q_{ss}^{-1})Q$ where I is the identity matrix.

Now 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 $B = (I - L)^{-1}U$. Hence the rate of convergence of the Markov chain to its stationary distribution is given by $\rho = \rho(B)$, the spectral radius (maximum modulus eigenvalue) of B. This provides the exact rate of convergence for a deterministically updated Gibbs sampling scheme. However, if Q is of block tridiagonal form then we can use Theorem 5 of Roberts and Sahu (1995) which states that $\rho(B) = \{\lambda(A)\}^2$, where $\lambda(X)$ denotes the largest eigenvalue of X.

For an AR(1) plus noise model, it is instructive to closely examine the relationship between the parameters of this model, σ_η^2 and ϕ, and the rate of convergence, ρ for the single updating strategy applied to the states α. The AR(1) plus noise model can be expressed as follows:

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

We have deliberately chosen the initial condition $\alpha_1 \sim \mathsf{N}(0, \sigma_\eta^2)$ rather than choosing the more usual condition $\alpha_1 \sim \mathsf{N}\{0, \sigma_\eta^2/(1 - \phi^2)\}$ for convenience in terms of the resulting matrices. In any case we are interested in the convergence rate as $n \longrightarrow \infty$ so the initial variance becomes irrelevant. Here Q, the inverse variance-covariance matrix for the states α which arises from the AR(1) plus noise model, is given by

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

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

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

where $a = \frac{\phi}{1+\phi^2+\sigma_\eta^2\sigma_\varepsilon^{-2}}$, $b = \frac{\phi}{1+\sigma_\eta^2\sigma_\varepsilon^{-2}}$. We obtain the following result, proved in the Appendix.

Theorem 1: *The spectral radius of B, ρ 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 ρ can be expressed in terms of the parameters of the AR(1) plus noise model as:

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

Theorem 1 indicates that the convergence rate always lies between 0 and 1 and as $\phi \longrightarrow 1, \sigma_\eta^2\sigma_\varepsilon^{-2} \longrightarrow 0$ then $\rho \longrightarrow 1$. Hence it can be seen that for reasonably persistent 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 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 ρ for $n = 50$ and that the convergence rate changes very little as n increases from 50 to ∞.

Figure 1 shows contours of the asymptotic convergence rate for different values of ϕ and the signal to noise ratio, $\sigma_\eta^2/\sigma_\varepsilon^2$.

3 Reparameterisation for linear State Space Models

We shall now focus on the following model parameterisations,

$$y \sim \mathsf{N}(\mu\boldsymbol{1} + \alpha, \sigma_\varepsilon^2 I_n), \text{ where } \alpha \sim \mathsf{N}(0, D), \tag{2}$$

and

$$y \sim \mathsf{N}(\omega, \sigma_\varepsilon^2 I_n), \text{ where } \omega \sim \mathsf{N}(\mu\boldsymbol{1}, D), \tag{3}$$

where $\boldsymbol{1}$ represents a vector of length n consisting of 1's and ω_t, α_t ($t = 1, .., n$) are of state space form. It can immediately be seen that the marginal distribution of y is unaffected by

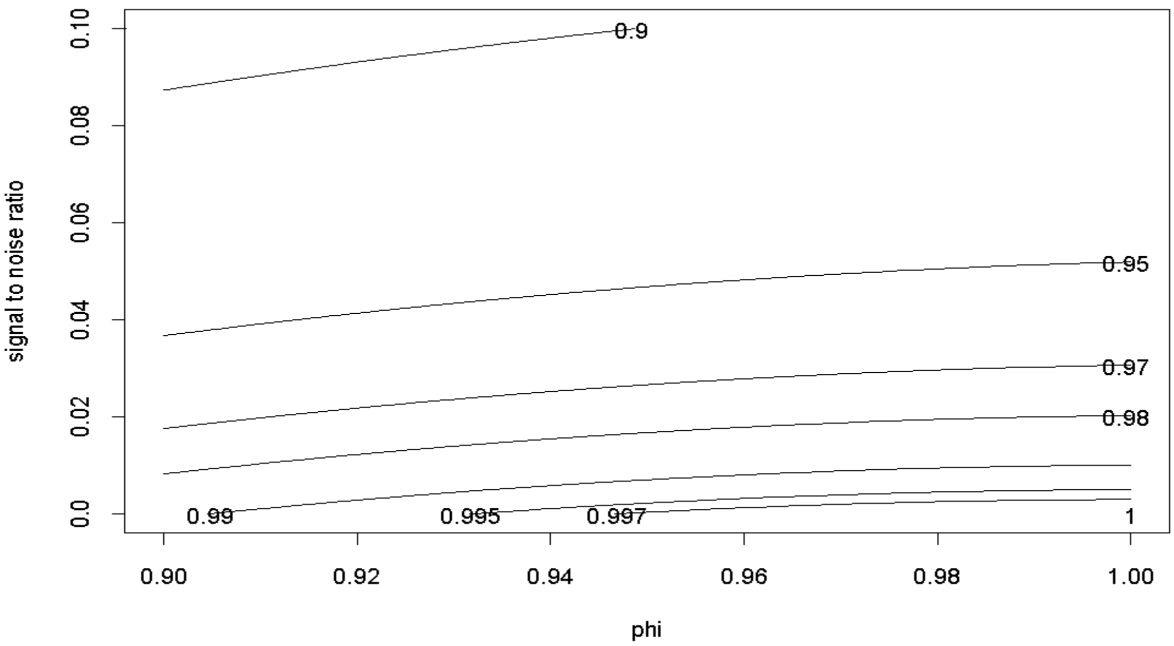

Figure 1. *Contour plot of levels of asymptotic rate plotted against ϕ and signal to noise ratio.*

which parameterisation we choose. In particular if we consider the special case in which α, ω are of AR(1) form then the first parameterisation is,

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

and the second parameterisation is,

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

The approach of Gelfand, Sahu and Carlin (1995) will be exploited at this stage in order to consider the effect of Gibbs sampling for each of the two general state space models. The models examined by Gelfand *et al.* (1995) are hierarchical linear normal models rather than time series models. They also have μ as a vector rather than the scalar which we are considering for the models above. We shall consider the simple case of updating $\mu|y, \alpha$ and $\mu|y, \omega$. The parameters σ_ε^2, σ_η^2 and ϕ shall be assumed fixed.

The marginal and conditional distributions for the two alternative parameterisations of (2) and (3) are given in the Appendix. In particular it can be seen that the conditional distributions for $\alpha|y$ will have a larger variance than the "centered" parameterisation of $\omega|y$ if $VD^{-1}\boldsymbol{1}\boldsymbol{1}'$ is small compared to $\boldsymbol{1}\boldsymbol{1}'$, 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. We shall attach flat priors for μ for both parameterisations.

3.1 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 μ 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 parameterisation by $\rho_\mu(1;\alpha)$ and similarly use $\rho_\mu(1;\omega)$ to denote the autocorrelation for the centered parameterisation.

Thus we obtain

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

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,

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

due to lemma 1 given in the Appendix.

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

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

and

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

We define the relative efficiency of the centered parameterisation to the uncentered parameterisation as the number of samples from the uncentered parameterisation we would require to estimate $E(\mu)$ to a given precision divided by the number of samples from the centered parameterisation we would require for the same precision. For the AR(1) parameterisations the relative efficiency, e, is given by,

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

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

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

and as $n \longrightarrow \infty$,

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

As $\phi \longrightarrow 1$ then we have $\rho_\mu(1;\omega), \rho_\mu^*(\omega) \longrightarrow 0$ and $\rho_\mu(1;\alpha), \rho_\mu^*(\alpha) \longrightarrow 1$, a result stated in lemma 3. The lower bound will therfore become increasingly tight. The efficiency bounds depend on the parameters σ_ε^2 and σ_η^2 only through the signal to noise ratio $\sigma_\varepsilon^2\sigma_\eta^{-2}$. A plot of these bounds together with the actual efficiency is given for two different signal to noise ratios in Figure 2 for the case $n = 50$. The efficiency is plotted against values of ϕ from 0.7 to 1. The values of σ_η^2 are 0.02, while σ_ε^2 varies from 1.0 (top plot) to 0.1 (bottom plot). It is clear from all the figures that the lower bound is close to the true efficiency, becoming increasingly accurate as $\phi \longrightarrow 1$ 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 ϕ becomes closer to 1 we obtain acfs close to 1 for the uncentered parameterisation and acfs close to 0 for the centered parameterisation. In cases where $\phi \geq 0.9$ the acfs are clearly far lower for the centered parameterisation.

3.2 Example 1

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 parameterisations. The sampling was performed in the manner analysed i.e. sampling μ given all the states then all the states given μ.

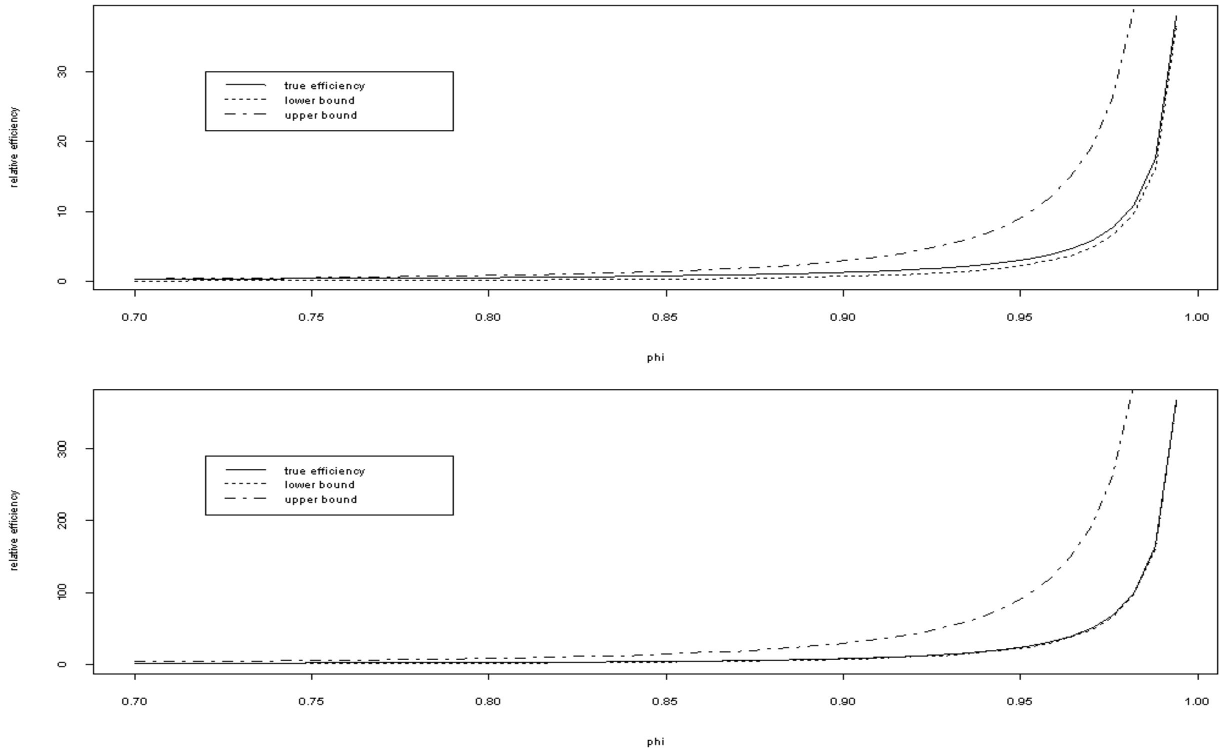

Figure 2. *Plot of relative efficiency of "centred" parameterisation to "uncentred" parameterisation for $\sigma_\eta^2 = 0.02$, $n = 50$ plotted for range of values of ϕ. The top graph shows the case when σ_ε^2 = 1.0, the bottom graph shows the case when σ_ε^2 = 0.1.*

Figure 3 shows the samples of μ resulting from the two parameterisations together with their autocorrelations. The samplers of μ both result in samples which vary around the true value with similar variances (although the variances need not be the same). The autocorrelations are strikingly different however, with the uncentered parameterisation leading to autocorrelation beyond lags of 500 and the centered parameterisation 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 from the Gibbs sampler to that arising from independent samples we use a Parzen window, see Priestley (1981, Chapter

6). The ratio is estimated as follows,

$$
\begin{aligned}
\widehat{R}_M &= 1 + \frac{2M}{M-1} \sum_{i=1}^{B_M} K\left(\frac{i}{B_M}\right) \widehat{\rho}(i), \text{ where } \widehat{\rho}(i) = \frac{\widehat{\Gamma}(i)}{\widehat{\Gamma}(0)} \\
\widehat{\Gamma}(i) &= \frac{1}{M} \sum_{j=i+1}^{M} (\alpha^j - \overline{\alpha})(\alpha^{j-i} - \overline{\alpha}), \ \overline{\alpha} = \frac{1}{M} \sum_{j=1}^{M} \alpha^j,
\end{aligned}
$$

where the Parzen kernel is

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

and B_M represents band length. Here $\widehat{\rho}(i)$ 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 parameterisations respectively. This indicates that the centered parameterisation is about 500 times more efficient than the uncentered parameterisation.

In Figure 4 the samples of the third state are shown for each of the two parameterisations. The autocorrelations are similar to those for μ, unsuprisingly, but the variance of the uncentered state is higher than that of the centered state (this is not surprising — see the Appendix). The centered state varies around 3 while the uncentered state varies around 0.

3.3 Example 2

We now examine two alternative parameterisations for the Stochastic Volatility (SV) model. This model has been considered extensively in the financial econometrics literature as it allows modelling the time-varying variance of returns on assets. It major properties are reviewed by Shephard (1996). The standard SV model is given by

$$
\begin{aligned}
y_t &= \epsilon_t \beta \exp(\alpha_t/2), \quad && \alpha_t = \phi \alpha_{t-1} + \eta_t, \\
\eta_t &\sim \mathsf{N}(0, \sigma_\eta^2), && \alpha_1 \sim \mathsf{N}\left\{0, \sigma_\eta^2/(1-\phi^2)\right\},
\end{aligned}
\tag{6}
$$

and can be reparameterised as

$$
\begin{aligned}
y_t &= \epsilon_t \exp(\omega_t/2), \quad && \omega_t = \mu + \phi(\omega_{t-1} - \mu) + \eta_t, \\
\eta_t &\sim \mathsf{N}(0, \sigma_\eta^2), && \omega_1 \sim \mathsf{N}\left\{\mu, \sigma_\eta^2/(1-\phi^2)\right\},
\end{aligned}
\tag{7}
$$

without affecting the marginal distribution of y provided $\mu = \log(\beta^2)$. In each case we are interested in sampling the states and β (or μ) keeping the other parameters fixed for the purposes

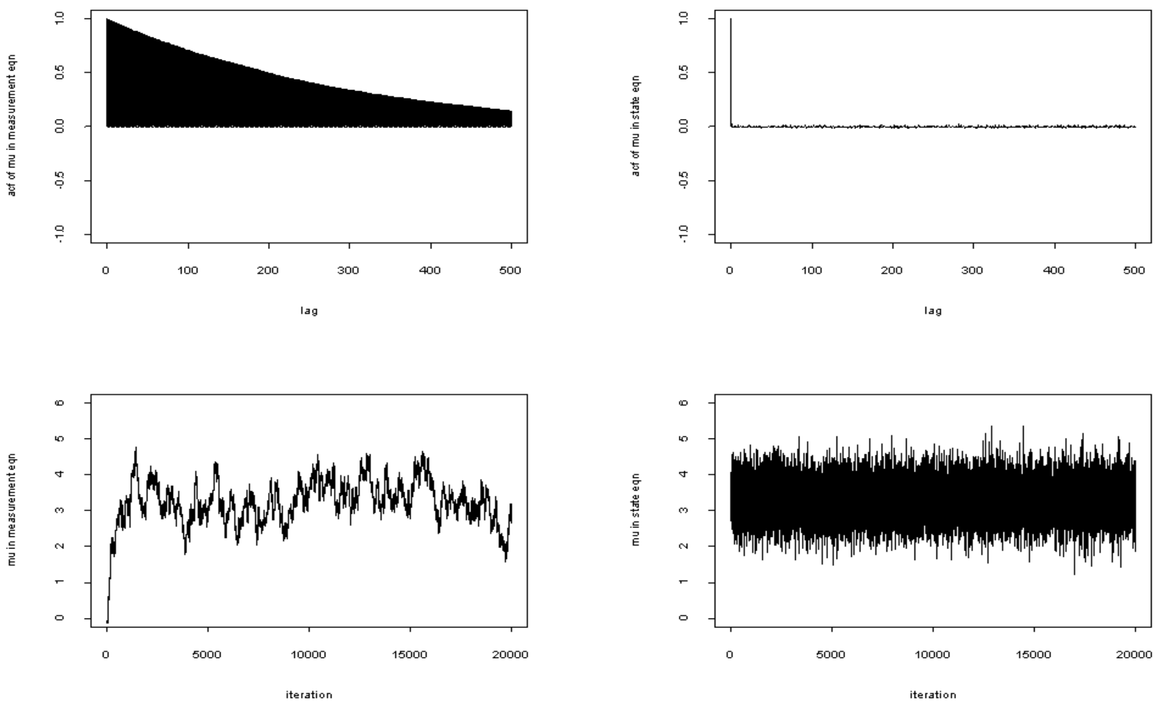

Figure 3. *Plot of acfs together with sample path for 20000 samples of μ (true value 3), ϕ and σ^2 fixed (at 0.98 and 0.02). The left plots show the case of μ in the measurement equation. The right plots show the corresponding plots for μ in the state equation.*

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 (1995).

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 (6) and (7) can both be linearised (at the expense of Gaussianity) as

$$\log y_t^2 = \log(\beta^2) + \alpha_t + \log \epsilon_t^2, \text{ and } \log y_t^2 = \omega_t + \log \epsilon_t^2, \tag{8}$$

so the similarity between the non-linear models and the Gaussian state space form (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, (6) 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 β^2 represents the mean shift of the measurement equation and also a multi-

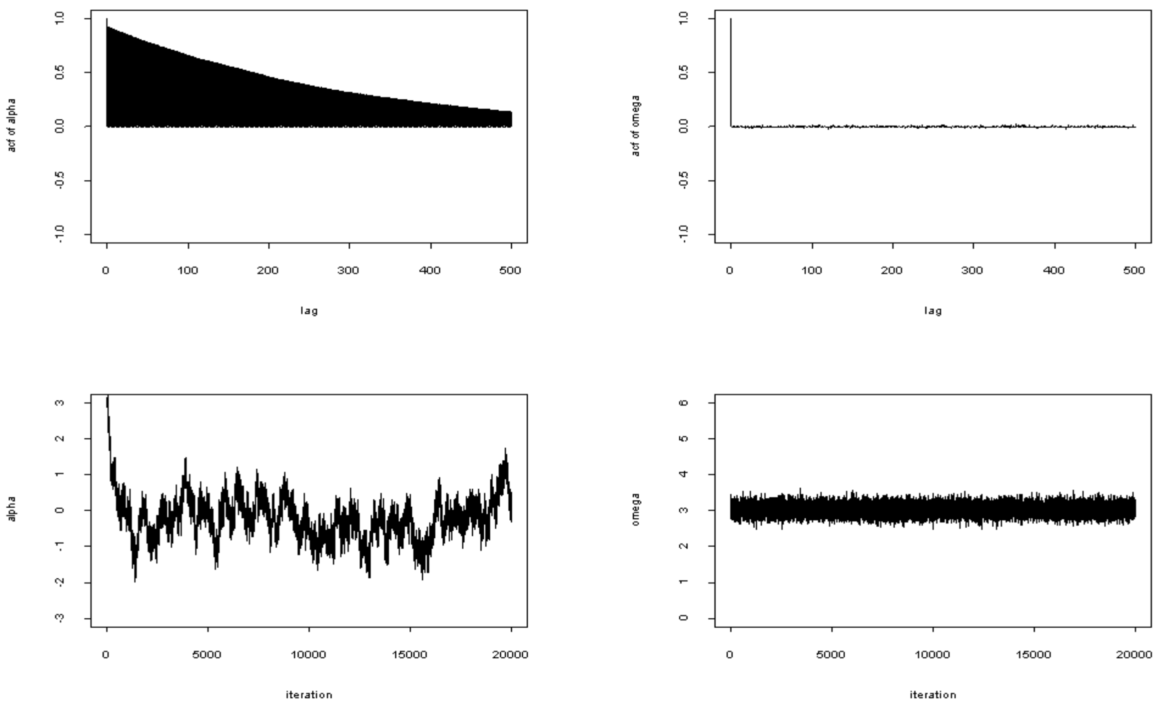

Figure 4. *Plot of acfs together with sample path for 20000 samples of the second states , ϕ and σ^2 fixed (at 0.98 and 0.02). The left plots show the case of α (μ in the measurement equation). The right plots show the corresponding plots for ω (μ in the state equation).*

plicative term for α_t. Since linearisations (in terms of a fixed function of β and in terms of the states) can be performed on the measurement equation fairly straightforwardly this implies that comparison with the Gaussian state space form (GSSF) is informative. This is certainly true for the simple SV model as the linearisation can be performed directly.

The conditional distributions for β and μ 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 parameterisation we obtain,

$$\mu | y, \omega \sim \mathsf{N}(b/a, \sigma_\eta^2/a), \text{ where}$$

$a = (n-1)(1-\phi)^2 + (1-\phi^2); s = \sum_{2}^{n}(\omega_t - \phi\omega_{t-1}); b = (1-\phi)s + \omega_1(1-\phi^2)$. These parameterisations are equivalent since we assume a flat prior for μ and $\log \beta$.

We shall now compare the two Gibbs sampling 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 involv-

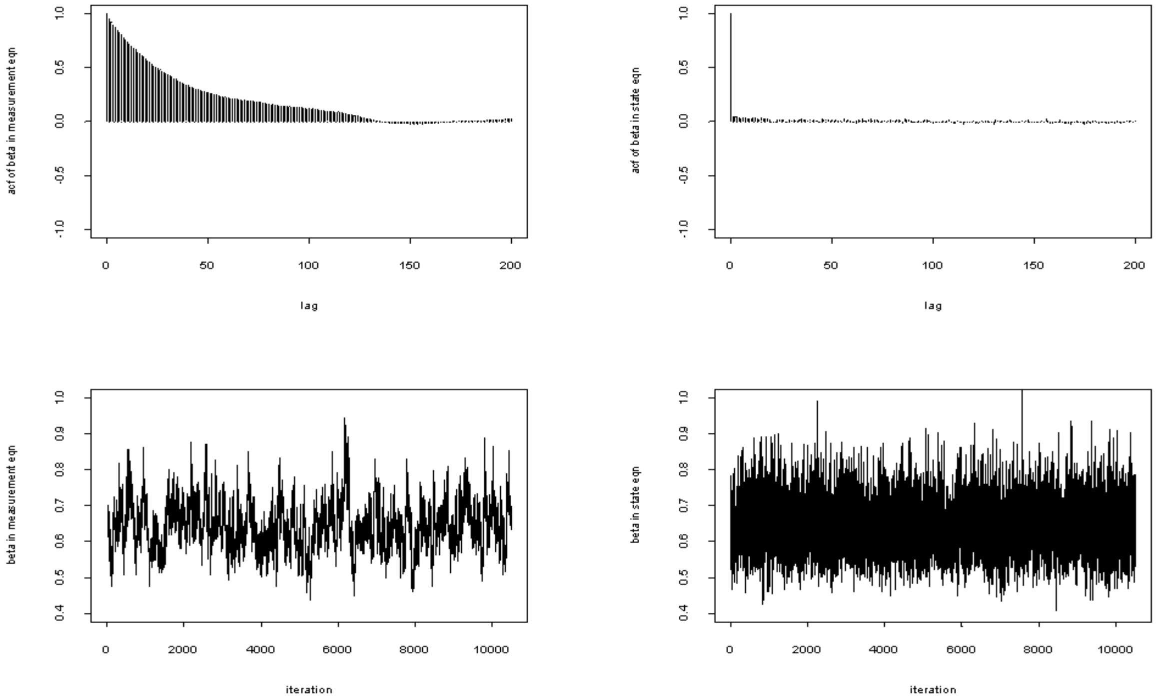

Figure 5. *Plot of acfs together with sample path for 10200 samples of β, ϕ and σ^2 fixed (at 0.98 and 0.02). The left plots show the case of β in the measurement equation. The right plots show the corresponding plots for β ($exp(\mu/2)$) in the state equation.*

ing, 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 ϕ, σ_η^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 5. The samples of μ 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 parameterisation yields substantial efficiency gains.

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 (ϕ close to 1, σ_η^2 close to 0) leads to extremely bad convergence. Clearly we would not actually wish to perform Gibbs sampling to estimate the posterior density of the states for a Gaussian linear state space model.

However, in many interesting time series problems the Gaussianity or linearity of the state space model breaks down although the dependency structure is identical. 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 (1995).

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

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.

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.

Proof of Theorem 1*:*

We have

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

where the dimension of A is $n \times n$. So

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

Let Λ_n denote the determinant D_n when $b = a$. Then we have, in general, that

$$D_{n+2} + \lambda\Lambda_{n+1} + ab\Lambda_n = 0.$$

Also Λ_n satisfies the following difference equation,

$$\Lambda_{n+2} + \lambda\Lambda_{n+1} + a^2\Lambda_n = 0.$$

subject to the 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

$$\sum_{n=0}^{\infty} t^{n+2}\Lambda_{n+2} + \lambda \sum_{n=0}^{\infty} t^{n+2}\Lambda_{n+1} + ab \sum_{n=0}^{\infty} t^{n+2}\Lambda_n = 0.$$

Then we obtain

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

implying

$$A(t) = \frac{1}{(1 + \lambda t + a^2t^2)}.$$

Now we want to find values of T and x such that $1 + \lambda t + a^2t^2 = 1 - 2Tx + T^2$. The values which satify this are $T = -ta$, and $x = \lambda/2a$. Now

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

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

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

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

$$\Lambda_n = (-1)^n a^n U_n\left(\frac{\lambda}{2a}\right).$$

The n solutions of this characteristic equation are satisfy $\sin\left\{(n+1)\cos^{-1}(\frac{\lambda}{2a})\right\} = 0$. Hence the n distinct eigenvalues are given by

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

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

Since $D_n = -\lambda\Lambda_{n-1} - ab\Lambda_{n-2} = \Lambda_n - a^2\varepsilon\Lambda_{n-2}$, where $b = (1+\varepsilon)a$, $\varepsilon = \frac{\phi^2}{1+\sigma_\eta^2\sigma_\varepsilon^2}$ we have that the solution of $D_n(\lambda) = 0$ satisfies

$$f(\theta) = \sin(n+1)\theta - \varepsilon\sin(n-1)\theta = 0 \tag{9}$$

where $\theta = \cos^{-1}(\lambda/2a)$ and $0 < \theta < \pi$. It can be seen that the largest eigenvalue λ_1, say, corresponds to the smallest solution θ_1 of (9). We now wish to show that this root lies in the interval $[0, \frac{\pi}{n+1}]$. When $\theta = 0$, $f(0) = 0$ and

$$\begin{aligned} f'(\theta) &= (n+1)\cos(n+1)\theta - \varepsilon(n-1)\cos(n-1)\theta \\ \therefore\ f'(0) &> 0 \text{ if } \varepsilon < \frac{n+2}{n}. \end{aligned}$$

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 $\theta = \frac{\pi}{n+1}$,

$$f(\theta) = -\varepsilon\sin\left(\frac{(n-1)\pi}{n+1}\right) < 0 \qquad \text{since } \varepsilon > 0.$$

So the smallest root θ_1, say, of (9) satisfies

$$0 < \theta_1 < \frac{\pi}{n+1}.$$

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 $\rho(B) = \lambda_1^2$ satifies

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

Hence the as $n \longrightarrow \infty$,

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

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

Lemma 1: *We note the result that,*

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

Proof of Lemma 1:

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

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

Hence for the special case above we have

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

so that, as required,

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

Lemma 2: *For the AR(1) plus noise model*

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

Proof of Lemma 2: 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 $\boldsymbol{1}'D^{-1}\boldsymbol{1}$ sums all the elements of the matrix we obtain the required result.

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

$$\begin{aligned} \rho_\mu(1;\alpha) &\longrightarrow 1 \text{ as } \phi \longrightarrow 1 \\ \rho_\mu(1;\omega) &\longrightarrow 0 \text{ as } \phi \longrightarrow 1 \end{aligned} \qquad (10)$$

Proof of Lemma 3:

We have

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

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

Similarly we have
$$\begin{aligned}\rho_\mu(1;\omega) &= \frac{\mathbf{1}'D^{-1}VD^{-1}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}} = \frac{\mathbf{1}'D^{-1}(D^{-1}+\sigma_\varepsilon^{-2}I)^{-1}D^{-1}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}} \\ &\le \frac{\sigma_\varepsilon^2\mathbf{1}'D^{-2}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}}.\end{aligned}$$
We have
$$D^{-2} = \sigma_\eta^{-4}\begin{pmatrix} 1+\phi^2 & -\phi(2+\phi^2) & \phi^2 & 0 & \cdots & 0 \\ -\phi(2+\phi^2) & 1+4\phi^2+\phi^4 & -2\phi(1+\phi^2) & \phi^2 & \cdots & \vdots \\ \phi^2 & -2\phi(2+\phi^2) & 1+4\phi+\phi^4 & -2\phi(1+\phi^2) & & \vdots \\ \vdots & & & & \ddots & -\phi(2+\phi^2) \\ 0 & \cdots & \cdots & \phi^2 & -\phi(2+\phi^2) & 1+\phi^2 \end{pmatrix},$$
implying
$$\frac{\sigma_\varepsilon^2\mathbf{1}'D^{-2}\mathbf{1}}{\mathbf{1}'D^{-1}\mathbf{1}} = \frac{\sigma_\varepsilon^2}{\sigma_\eta^2}\frac{\{(n-2)\phi^4-4(n-2)\phi^3+6(n-2)\phi^2+2\phi^2-4(n-1)\phi+n\}}{\{(n-2)(1+\phi^2)+2-2(n-1)\phi\}}.$$
Using L'Hopital's rule we have that
$$\begin{aligned}&\lim_{\phi\to 1}\frac{\{(n-2)\phi^4-4(n-2)\phi^3+6(n-2)\phi^2+2\phi^2-4(n-1)\phi+n\}}{\{(n-2)(1+\phi^2)+2-2(n-1)\phi\}} \\ = \;& \frac{4(n-1)-12(n-2)+12(n-2)+4-4(n-1)}{2(n-2)-2(n-1)} = \frac{0}{-2} = 0,\end{aligned}$$
yielding the required result.

6.1 Marginal and conditional distributions for different parameterisations

The marginal distribution of $\mu|y$ for both models is
$$\mu|y \sim \mathsf{N}\left\{(\mathbf{1}'\Sigma^{-1}y)(\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}, (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}\right\},$$
where $\Sigma = \sigma_\varepsilon^2 I_n + D$. So
$$\mu\mathbf{1}|y \sim \mathsf{N}(\mathbf{1}\widehat{\mu}, \sigma_\mu^2\mathbf{1}\mathbf{1}'),$$
where for convenience we shall now denote $\widehat{\mu} = (\mathbf{1}'\Sigma^{-1}y)(\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}$ and $\sigma_\mu^2 = (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}$. The conditional distribution for $\mu|y,\alpha$ for the first model is given by
$$\mu|y,\alpha \sim \mathsf{N}(\overline{y}-\overline{\alpha}, \sigma_\varepsilon^2/n).$$

The conditional distribution for $\mu|y,\omega$ is given by

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

In the simple AR(1) case this reduces simply to,

$$\begin{aligned}
\mu|y,\omega &\sim \mathsf{N}(q/p, \sigma_\eta^2/p), \quad \text{where} \\
p &= (n-1)(1-\phi)^2 + (1-\phi^2), \quad \text{and} \\
q &= \omega_1(1-\phi^2) + (1-\phi)\sum_{i=2}^{n}(\omega_i - \phi\omega_{i-1}).
\end{aligned}$$

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

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

The marginal distribution $\omega|y$ is

$$\omega|y \sim \mathsf{N}(V\widehat{b}, V + \sigma_\mu^2 V D^{-1}\boldsymbol{1}\boldsymbol{1}'D^{-1}V) \quad \text{and} \quad \widehat{b} = \sigma_\varepsilon^{-2} y + D^{-1}\boldsymbol{1}\,\widehat{\mu}.$$

Similarly the conditional distribution of $\alpha|y,\mu$ is given by

$$\alpha|y,\mu \sim N(V\widehat{b} - \boldsymbol{1}\,\mu, V).$$

Hence

$$\alpha|y \sim \mathsf{N}(V\widehat{b} - \boldsymbol{1}\,\widehat{\mu}, V + \sigma_\mu^2 V D^{-1}\boldsymbol{1}\boldsymbol{1}'D^{-1}V + \sigma_\mu^2\boldsymbol{1}\boldsymbol{1}' - 2\sigma_\mu^2 V D^{-1}\boldsymbol{1}\boldsymbol{1}').$$

These results are analogous to those results of Gelfand *et al.* (1995) for the hierarchical linear model.

REFERENCES

Carlin, B. P., Polson, N. G. and Stoffer, D. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modelling, *J. Am. Statist. Assoc.*, **87**, 493–500.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models, *Biometrika*, **81**, 541–553.

de Jong, P. and Shephard, N. (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. E., Sahu, S. K. and Carlin, B. P. (1995). Efficient parameterisations for normal linear mixed models, *Biometrika*, **82**, 479–488.

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

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images, *IEEE Transactions, PAMI*, **6**, 721–741.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving MCMC, In W R Gilks, S. R. and Spiegelhalter, D. J. (eds.), *Markov Chain Monte carlo in Practice*. London: Chapman & Hall.

Harvey, A. C. (1993). *Time Series Models* 2nd edition. Hemel Hempstead: Harvester Wheatsheaf.

Harvey, A. C., Ruiz, E. and Shephard, N. (1994). Multivariate stochastic variance models, *Rev. Economic Studies*, **61**, 247–264. Reprinted as pp. 256–276 in Engle, R.F.(ed) (1995), *ARCH: Selected Readings*, Oxford: Oxford University Press.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

Roberts, G. and Sahu, S. (1995). Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler, Unpublished, Statistical Laboratory, Cambridge University.

Shephard, N. (1994). Partial non-Gaussian state space, *Biometrika*, **81**, 115–131.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility, In Cox, D. R., Barndorff-Nielson, O. E. and Hinkley, D. V. (eds.), *Time Series Models in Econometrics, Finance and Other Fields*, pp. 1–67. London: Chapman & Hall. Forthcoming.

Shephard, N. and Pitt, M. (1995). Parameter-driven exponential family models, Unpublished working paper: Nuffield College, Oxford.