LIKELIHOOD ANALYSIS OF NON-GAUSSIAN PARAMETER-DRIVEN MODELS

NEIL SHEPHARD
Nuffield College, Oxford, OX1 1NF, UK
AND MICHAEL K PITT
Department of Statistics, University of Oxford, OX1 3TG and
Nuffield College, Oxford, OX1 1NF, UK

December 18, 1995

Abstract

In this paper we provide a unified likelihood treatment of a wide class of non-Gaussian time series, semi-parametric and non-parametric regression, index models, spectral density estimation and panel data problems. The methods rely on the use of Markov chain Monte Carlo to carry out non-Gaussian simulation smoothing and Bayesian posterior analysis of parameters, and on importance sampling to estimate the likelihood function for classical inference. We exploit the structure of the models to ensure that our methods are highly efficient, delivering MCMC simulations with little autocorrelation and highly effective estimates of the likelihood.

Some key words: Blocking; Durations; Exponential family; Importance sampling; Index models; Markov chain Monte Carlo; Non-parametric regression; Penalised likelihood; Simulation smoother; Spectral density estimation; Splines; Stochastic volatility.

1 Introduction

In this paper, we provide a unified likelihood treatment of a wide class of problems which allow dependence by letting functions of parameters evolve according to an unobserved stochastic process. In a review of the time series literature Cox (1981) called such time dependence models "parameter-driven". Here we show that a wide range of problems can be construed as "parameter-driven" providing likelihood techniques which link together panel data, periodogram smoothing and non and semi-parametric regression estimation, as well as time series. Elsewhere in econometrics and statistics these types of models are known under various labels, such as "latent" (Bartholomew (1987)), "multilevel" (Goldstein (1995)), "hierarchical" (Bernardo and Smith (1994, pp. 222-226)) and "state space" (Harvey (1989)). Some economic theory also uses parameter-driven models, relating observations to unobserved state variables. Examples of this include the Cox, Ingersoll, and Ross (1985) term structure model of interest rates and the Hull and White (1987) option pricing model under stochastic volatility (SV).

The basis of our work will be the assumption that the univariate observations y_{it} are distributed, conditionally on a univariate θ_{it}, according to $f(y_{it}|\theta_{it}), i = 1, ..., N;\ t = 1, ..., n$. Throughout $\log f$ is assumed to be twice continuously differentiable with respect to θ_{it}. The groups $(y_{i1}, ..., y_{in}, \theta_{i1}, ..., \theta_{in})'$ are assumed to be independent across i, that is, the y_{it} are dependent over the index t, but independent over the index i. A leading example of this is the exponential family, which is written as

$$\log f(y_{it} \mid \theta_{it}) = y_{it}\theta_{it} - b(\theta_{it}) + c(y_{it}),$$

where b and c are flexible but known functions. We will write $\mu_{it} = \partial b(\theta_{it})/\partial\theta_{it} = E(y_{it}|\theta_{it})$.

Here θ_{it} will be a stochastic process, whose form will be given by:

$$\begin{array}{ccccl} h(\theta_{it}) & = & d_{it} & = & z_{it}\alpha_{it} + x_{it}\beta \\ & & \alpha_{it+1} & = & W_{it}\beta + T_{it}\alpha_{it} + H_{it}u_{it}, \quad u_{it} \sim \mathsf{NID}(0, I), \\ & & \alpha_{i1} & \sim & \mathsf{N}(a_{i1|0}, p_{i1|0}). \end{array}$$

$h(.)$ is a known function which is continuously twice differentiable, while the α_{it} are sometimes called states. In the exponential family case $g(\mu_{it}) = d_{it}$ is called a link function. Typically T_{it}, z_{it}, H_{it} will be selection matrices indexed by a small number of unknown parameters. These parameters will be denoted by ψ. This framework, which constitutes a Markov Random Field, will be called a parameter-driven model. The substantive role of i and t will vary over the applications we discuss.

The role of linearity and Gaussianity in the autoregression for α_{it} is not crucial. Formally all the algorithms we develop in this paper will maintain their correctness when the assumptions are weakened, however some of the theoretical results on the log-concavity of the densities will no longer hold. We do not address all these technical details in this paper, leaving those areas to later work.

1.1 Literature

There is a large literature on finding approximate solutions to this type of natural parameter-driven extension of the linear model. This subsection briefly discusses the literature around each of the subfields featured in the paper.

1.1.1 Time Series

The considerable literature extending linear time series models is usually taken to start with West, Harrison, and Migon (1985). The more recent work of Fahrmeir (1992) and Durbin and Koopman (1992) developed algorithms to find the mode of $\alpha \mid y$ and performed approximate likelihood inference. Here $y = (y_1, ..., y_n)'$ and $\alpha = (\alpha_1', ..., \alpha_n')'$.

Papers closer in spirit to our own are Carlin, Polson, and Stoffer (1992) and Chan and Ledolter (1995) who used what we call a "single-move" Gibbs sampler on this model; similar techniques were used by Carlin, Gelfand, and Smith (1992) in a change point analysis of poisson time series and Carlin, Polson, and Stoffer (1992) in a more general non-linear time series setting.

More recently in econometrics there has been considerable interest in finding likelihood solutions to special cases of the above framework. A leading example is the SV model, analysed by Jacquier, Polson, and Rossi (1994) who used a particular MCMC algorithm to provide the likelihood solution to this problem.

1.1.2 Non-parametric regression

The standard parametric analysis of Green and Silverman (1994) is to use the exponential family measurement model

$$\log f\left\{y_{[t]}|\theta_{[t]}=g(\tau_t)\right\}=y_{[t]}g(\tau_t)-b\left\{g(\tau_t)\right\}+c(\tau_t)$$

where τ_j denote the value of the ordered (in ascending size, so $\tau_j \leq \tau_{t+1}$) explanatory variables and $y_{[t]}$ is the corresponding dependent variable. Green and Silverman (1994) find the mode of the density of $g(\tau_1),...,g(\tau_n)$, given the observations. Such a fitting mechanism is called "penalized likelihood". Here the function g is given a prior whose density is the "penalty" $-\sigma_\eta^2\int g''(\tau)d\tau/2$. For small values of σ_η^2 the prior requires the g function to be very smooth, while as σ_η^2 increases the fitting process becomes more unconstrained. We will estimate this "smoothness" parameter, which corresponds to the bandwidth in kernel regression, from the data using likelihood methods.

In the penalized likelihood literature, the mode is found numerically by using a Fisher scoring algorithm (Green and Silverman (1994, pp. 99-100)), exploiting the band structures in the covariance matrix to ensure the computations are carried out in $O(n)$ operations. This work precisely corresponds to the inference method for time series suggested by Fahrmeir (1992) and Durbin and Koopman (1992), although the computational methods differ.

Unlike the time series literature, this literature does provide a consistent way to estimate the parameter, or roughness penalty, of the model. Green and Silverman (1994, pp. 107–110) use cross-validation or generalized cross-validation.

1.1.3 Periodogram smoothing

Wahba (1980) suggested that the periodogram could be estimated using a smoothing spline, by taking a log transformation of the periodogram. This transformation linearises the periodogram into the spectrum and $\log\chi^2$ measurement errors. This can then be smoothed using a spline. The presence of $\log\chi^2$ distributions can cause inefficient estimation. Work by Pawitan and O'Sullivan (1994) and Carter and Kohn (1995) has gone a considerable way toward overcoming this problem.

1.2 Methodology

The modelling framework given above specifies a complete parametric model. Our approach is to perform either Bayesian or classical inference on the parameters which index this model, while reporting features of the distribution of the states given the observations. Although we present new methodology for carrying out this task, we emphasize that this is simply a difficult computational problem, requiring the numerical solution of integrals whose dimension are positive integer multiples of the size of the data.

This paper presents some new methodology for the analysis of parameter-driven models. It will be based on the estimation by simulation of particular features of the model. This has the advantage that the estimates can be made as accurate as needed and their degree of accuracy can be assessed statistically. These simulation methods replace, in the Gaussian case, the role of the Kalman filter and smoother which usually estimate the states α from the observations y. In a simple example, if it were possible to simulate from $\alpha \mid y, \psi$ easily, then any interesting function of α can be straightforwardly estimated from the draws. Likewise, if a Bayesian paradigm is adopted for the parameters which index the model, then the parameters can be estimated by recording appropriate summaries of the simulations from $(\alpha', \psi')' \mid y$. If a classical view is taken then an approximate Gaussian simulation smoother can be used as an input into an importance sample estimator of the likelihood.

Although this methodology is straightforward to state, it is by no means immediately obvious how to directly simulate from $\alpha \mid y, \psi$, which is a highly multivariate random variable. The Gaussian case is tackled in a paper by de Jong and Shephard (1995), but outside that special case this problem is not satisfactorily solved. Instead of drawing immediately from these multivariate densities, it is possible to use Markov Chain Monte Carlo (MCMC) methods. These methods break the vector of the variables into pieces and then draw with replacement (and rejection) from these pieces, iterating and converging towards a draw from the correct multivariate density. This approach leaves open the question of how to do the breaking and how this affects the speed of convergence. This is the question addressed in this paper.

The use of MCMC methods to simulate complex multivariate distributions has become increasingly widespread among statisticians. The Gibbs sampling algorithm is the best known of these methods, see for example Gelfand and Smith (1990). Increasingly, however, statisticians use the more general Metropolis algorithm. This technique was originally developed by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and subsequently generalised by Hastings (1970). These techniques were not widely used amongst the statistical fraternity until recently, following papers by Smith and Roberts (1993) and Tierney (1994) among others. A recent review of the use of MCMC in econometrics is given in Geweke (1996).

1.2.1 Gibbs sampling

Suppose we wish to simulate from the joint distribution of $x_1, x_2, .., x_m$, denoted f, where the x_i stands for a posterior on a group of states or parameters in the model, given the data. The Gibbs sampler generates a Markov chain whose equilibrium distribution is the joint distribution of $x_1, x_2, .., x_m$. The Markov chain is generated by the following algorithm: each variable is visited in turn and sampled with replacement from the distribution of that variable conditional on all the other variables. So for the $(k+1)-th$ iteration of the chain, $x_1^{(k+1)}$ is drawn from $f(x_1 \mid x_{\backslash 1}^{(k)})$, $x_2^{(k+1)}$ is drawn from $f(x_2 \mid x_{\backslash 2}^{(k)})$, and so on. Here $x_{\backslash i}^{(k)} = \left\{x_1^{(k+1)}, ..., x_{i-1}^{(k+1)}, x_{i+1}^{(k)}, ..., x_m^{(k)}\right\}'$. The new state, $x^{(k+1)}$, of the Markov chain is that resulting from an entire sweep through the variables.

Sufficient conditions for this sampler to converge are that $f(x_i \mid x_{\backslash i}^{(k)})$ is everywhere positive (Tierney (1994)). Formally these conditions will not hold for the states of parameter-driven models due to the constraints which are usually present in the vector autoregressive representation of the dynamics of the model. These constraints will mean that H_{it} may have deficient row rank. However, the Gibbs sampler is valid when performed indirectly on the disturbances, u_{it}, of the model. This condition always holds, by construction of the model, and implies a satisfactory equilibrium density for the implied simulated states (which are just weighted sums of the disturbances). This point was first raised by de Jong and Shephard (1995).

The rate of convergence to the equilibrium distribution is a crucial feature of MCMC methods. In general it is difficult to develop computable analytic convergence rates. However, Roberts and Sahu (1995) examine analytic rates for multivariate Gaussian densities, $x \sim \mathsf{N}(\mu, \Sigma)$. They define the rate of convergence, ρ, as follows. Suppose that $\{x^{(k)}, k = 0, 1, ..\}$ is a Markov chain with stationary density function $f(x)$. Let h be a square f-integrable function of x and $f(h)$ denote the expectation of h under the target density f. Let ρ be the minimum number such that for all square f-integrable functions h, and for all $r < \rho$, we can find a function $V(.) \geq 1$ such that $f(V) < \infty$ and

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

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

Roberts and Sahu (1995) partition Σ and $Q = \Sigma^{-1}$ according to the Gibbs sampler scheme, so $Cov(x_i, x_j) = \Sigma_{ij}$, and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1m} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{m1} & \Sigma_{m2} & \cdots & \Sigma_{mm} \end{pmatrix}, Q = \begin{pmatrix} Q_{11} & Q_{12} & \cdots & Q_{1m} \\ Q_{21} & Q_{22} & \cdots & Q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{m1} & Q_{m2} & \cdots & Q_{mm} \end{pmatrix}.$$

Let $A = I - diag(Q_{11}^{-1}, ..., Q_{ss}^{-1})Q$ where I is the identity matrix, define L to be the block lower triangular matrix with blocks in the lower triangle being identical to A and $U = A - L$. Then we have $E(x^{(k+1)}|x^{(k)}) = Bx^{(k)} + b$, 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 the Gibbs sampler in Gaussian cases. In non-Gaussian cases, this could serve as an approximation to the true convergence rates.

1.2.2 Metropolis algorithm

The difference between the Gibbs sampler and the Metropolis algorithm, whose equilibrium distribution is identical to that of the Gibbs sampler, is that all the states are not necessarily updated with each sweep. This may be particularly helpful in cases where it is not straightforward to sample from $f(x_i|x_{\setminus i})$. The sampling scheme becomes, for our setup:

1. Initialize $x^{(0)}$. Set $k = 0$.

2. Generate a new candidate value z based on the current value $x_{\setminus i}^{(k)}$ by using a *transition density*, $q(z, x_{\setminus i}^{(k)})$ and write $x_i^{(k)} = z$ with probability

$$\min\left(\frac{f(z|x_{\setminus i}^{(k)})q(x_i^{(k)}, x_{\setminus i}^{(k)})}{f(x_i^{(k)}|x_{\setminus i}^{(k)})q(z, x_{\setminus i}^{(k)})}, 1\right).$$

   This step is called the Metropolis rejection. Repeat for $i = 1, ..., m$.

3. Write $x^{(k+1)} = x^{(k)}$ and then $k = k + 1$, goto 2.

Sufficient conditions for this sampler to converge at a geometric rate are that $f(z \mid x_{\setminus i}^{(k)})$ and $q(z, x_{\setminus i}^{(k)})$ are everywhere positive (Tierney (1994)).

Tierney (1994) has drawn attention to a special case of this setup which integrates acceptance/rejection sampling (Ripley (1987, pp. 60-61)) within this framework. Basic rejection sampling can be used in Gibbs sampling for it allows draws to be made from $f(x_i|x_{\setminus i})$ when it is possible to find a density $h(z)$ from which it is easy to sample and there is a known constant c such that $f(z|x_{\setminus i}) \leq ch(z)$ for all z. In this case we say that h dominates or bounds f. Then to obtain a sample from $f(x|x_{\setminus i})$,

1. Generate a candidate value z from $h(\cdot)$ and a value u from $\mathsf{U}(0, 1)$.

2. If $u \leq f(z|x_{\setminus i})/ch(z)$ return z, otherwise goto 1.

As the expected number of iterations of steps 1 and 2 above is $1/c$, the optimal value for c is, $c = \sup_z f(z)/h(z)$.

Unfortunately for some parameter-driven models it is not possible to find a density which is both bounding and easy to sample. In a response to general problems of this type, Tierney (1994) proposed dropping the bounding condition. By repeating the formal steps 1 and 2 in the previous algorithm, even though $f(z|x_{\setminus i})/ch(z)$ is not now bounded from above by one, the resulting sampled density is proportional to min $\left[f(z|x_{\setminus i}^{(k)}), ch(z)\right]$. This is incorrect for the Gibbs sampling unless h is truly dominating. However, by writing $q(z, x_{\setminus i}^{(k)}) \propto \min\left[f(z|x_{\setminus i}^{(k)}), ch(z)\right]$ it is possible to produce a valid Metropolis algorithm. If h truly dominates there are no Metropolis rejections. This technique, called pseudo-dominating rejection, is adopted by Jacquier, Polson, and Rossi (1994) in their analysis of the SV model.

1.2.3 Multi-move samplers

The sampling techniques described above suggest taking x_i to stand for individual states or parameters (see Carlin, Polson, and Stoffer (1992) who used this method on some complicated non-linear models). Then the MCMC algorithm updates a single variable at a time, in order over the index i. An advantage of the single variable updating approach is that the conditional densities are usually not complicated. The main disadvantage of the single updating approach is the correlation between variables over successive sweeps of the Metropolis sample. Carter and Kohn (1994) and Shephard (1994) point out that this problem is particularly severe in (Gaussian or conditionally Gaussian) parameter-driven modelling. If the variables are extremely correlated then the sampling will be very slow in exploring the whole of the posterior distribution, and statistics derived from early segments of the resulting Markov chain may be entirely misleading.

Multi-move, or block, samplers work by updating several variables, sub-vectors of the vector of interest, at the same time. When carried out sensibly, block samplers tend to be very effective. Liu, Wong, and Kong (1994) produce results which show that "grouping random components in a Gibbs sampler with two or more components usually results in more efficient sampling schemes", while Amit and Grenander (1991) argue, perhaps too strongly, that "the larger the blocks that are updated simultaneously — the faster the convergence".

Block samplers raise a number of issues. How should the blocks be constructed so that they are easy to sample? Should the blocks be the same for each sweep (a deterministically updated MCMC) or should they be randomly selected (a random block MCMC)? A natural way of assessing solutions to these problems would be to measure their convergence rates, either by simulation or by analytic work. In the Gaussian case we can use the analysis of Roberts and Sahu (1995) to help us. We have already given their results for the deterministically updated Gibbs sampler (DUGS). The results for random blocks are less precise, only giving linear convergence rates ρ_L — that is convergence rates for linear functions of the x. It is clear that

$\rho_L \leq \rho$, but Roberts and Sahu indicate that "the two rates are usually equal".

The Roberts and Sahu analysis of random blocks writes $E\left\{x^{(k+1)}|x^{(k)}\right\} = C_i x^{(k)} + c_i, i = 1, ..., b$, where C_i is a valid Gibbs sampling transition matrix obtained by appropriate ordering and blocking of the x vector. The subscript i is selected randomly with probability ω_i. Then the rate of convergence is the maximum modulus eigenvalue of $\sum_i \omega_i C_i$.

1.2.4 Variance estimation

One of the key features in estimating characteristics of the density of $\alpha|y$ from simulations is the fact that the simulations will be correlated. Hence, although we can estimate $\mathsf{E}g(\alpha)$ by $M^{-1}\sum g(\alpha^t)$, where α^j is the $j-th$ simulated value after reaching equilibrium, the variance of this estimator is less straightforward. In this paper we will work with the Parzen kernel using an automatic bandwidth selection procedure. Our discussion here will follow Andrews (1991) and we will suppose for simplicity $g(\alpha) = \alpha$. Further, $\widehat{J}_M$ will denote the estimation of the variance of the $\sqrt{M}$ scaled sample mean about the population mean. The suggested $\widehat{J}_M$ is then

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

where the kernel is

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

The term B_M, the bandwidth, is selected empirically and will be stated in all our calculations.

The formula for $\widehat{J}_M$ is informative for MCMC methods. Typically if the MCMC sequence is highly positively correlated, then $\widehat{J}_M$ will be large and so the information content of each additional iteration will be small. That means for a given degree of accuracy the method will have to be run for longer for high correlation cases. Further, and as important, it is well known that the precision of $\widehat{J}_M$ reduces as the correlation increases for "All the estimators considered perform very poorly in an absolute sense when the amount of correlation is large". (Andrews (1991, (pp. 849))). This gives a second reason for designing low correlation MCMC methods: they provide a better basis on which to make inferences.

1.3 Structure of the paper

The structure of the paper is as follows. In section 2 we argue that simple single-move MCMC algorithms are unreliable empirical tools for the analysis of real data. The core of our contribution comes in Section 3, which discusses the design of our methods to maximise efficiency for parameter-driven models. Section 4 looks at importance sampling. Section 5 concludes, while an Appendix details some key algorithms.

2 Examples

2.1 Motivation

In this paper the focus will be on the dependence between observations, rather than the independence across subscript i for that generalisation is typically straightforward. Hence we will ignore the subscript i in our notation. In this subsection the first three examples will be about time series, while the third will be a non-parametric regression problem. However, first we consider some general principles for MCMC methods. Here we focus on single move algorithms.

2.2 General results for acceptance/rejection

We first show that reasonably generally it is possible to produce a Gibbs sampler for these problems, for we can sample from $\alpha_t|\alpha_{t-1}, \alpha_{t+1}, y_t$ straightforwardly. Part of the analysis will be based around the jackknife density: $\alpha_t|\alpha_{\backslash t} \sim \mathsf{N}(\mu_t, \Sigma_t)$, where

$$\alpha_{\backslash t} = (\alpha_1', ..., \alpha_{t-1}', \alpha_{t+1}', ..., \alpha_n')'.$$

We will write $\log f(y_t|\theta_t) = l(\alpha_t)$, suppressing the dependence on y_t for compactness of notation. Finally let

$$\frac{\partial l(\alpha_t)}{\partial \alpha_t} = l'(\alpha_t) \quad \text{and} \quad \frac{\partial^2 l(\alpha_t)}{\partial \alpha_t \partial \alpha_t'} = l''(\alpha_t).$$

Theorem 1.

We suppose Σ_t is non-singular and that $l''(\alpha_t)$ is negative semi-definite for all values of α_t. Then the following two results hold.

1. We can sample from $\alpha_t|\alpha_{t-1}, \alpha_{t+1}, y_t$ by making suggestions α_t from

$$\mathsf{N}\left[\mu_t + \Sigma_t l'(m_t), \Sigma_t\right], \tag{2}$$

which are accepted with probability

$$\exp\left[l(\alpha_t) - l(m_t) - l'(m_t)'(\alpha_t - m_t)\right], \tag{3}$$

whatever the value of m_t.

2. The probability of rejecting the suggestion made in (2) is minimised by selecting m_t as the mode of $f(\alpha_t|\alpha_{t-1}, \alpha_{t+1}, y_t)$.

This result has a number of attractions. Evaluating the rejection rate (3) does not require computing the jackknife density. The suggestion (2) requires only the computation of l'.

There is also complete flexibility over the selection of m_t, although a choice near the mode has advantages.

Of course the mode of f is relatively straightforward to find numerically due to the assumption that $l''(\alpha_t)$ is negative semi-definite for all α_t. Indeed a recursion of the form

$$m_t^{(i+1)} = \mu_t + \Sigma_t l' \left\{ m_t^{(i)} \right\}, \quad \text{starting at} \quad m_t^{(0)} = \mu_t,$$

is guaranteed to converge to the mode. The one-step algorithm of taking $m_t = \mu_t + \Sigma_t l' \{\mu_t\}$, is likely to be very successful in most situations as Σ_t is usually quite small. Hence it is unlikely that there will be much gain from iterating the procedure until convergence, which after all does take some computer time.

Although this result looks quite constrained, requiring $l'' \leq 0$, many interesting models (such as the SV, poisson and binomial) have this property. A more serious worry is the requirement that $\Sigma_t > 0$, which may not hold for some models due to the presence of constraints in the transition equation. This problem will be overcome in later sections.

2.3 Stochastic volatility

Our first model will be the SV model. This model has attracted much recent attention in the econometrics literature; see, for instance, Hull and White (1987), Harvey, Ruiz, and Shephard (1994) and Jacquier, Polson, and Rossi (1994). The basic SV model has

$$y_t = \epsilon_t \beta \exp(\alpha_t/2), \quad \alpha_{t+1} = \phi \alpha_t + \eta_t, \tag{4}$$

where ϵ_t and η_t are independent Gaussian processes with variances of 1 and σ_η^2 respectively. MCMC methods have been used on this model by, for instance, Jacquier, Polson, and Rossi (1994).

In this paper we will analyse the daily returns on the Pound Sterling/US Dollar exchange rate from 1/10/81 to 28/6/85. This dataset has been previously analysed using quasi-likelihood methods in Harvey, Ruiz, and Shephard (1994). It is graphed in Figure 1, together with $E(\alpha_t|y)$ (computed using an efficient MCMC method), the correlogram and the correlogram for the squares of the data. The graph indicates no linear predictability, but some volatility clustering (low correlations, but quite persistent). Hence the potential usefulness of model (4).

2.3.1 Gibbs sampler

The results of Theorem 1 apply immediately as, ignoring constants,

$$l(\alpha_t) = \log f(y_t|\alpha_t) = -\frac{\alpha_t}{2} - \frac{y_t^2}{2\beta^2}\exp(-\alpha_t) \quad \text{implying} \quad \frac{\partial^2 \log f(y_t|\alpha_t)}{\partial \alpha_t^2} = -\frac{y_t^2}{2\beta^2}\exp(-\alpha_t) < 0.$$

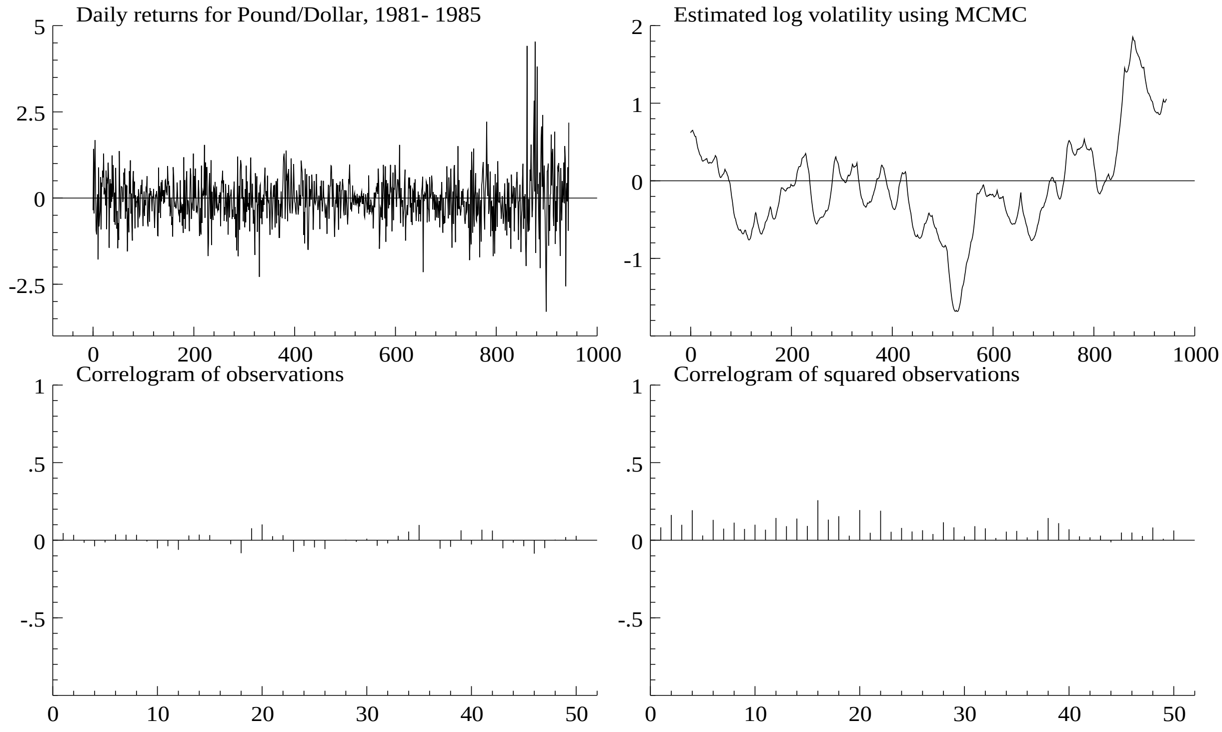

Figure 1: *Daily returns for the Pound against the Dollar: 1 October 1981 to 28 June 1985,* $n = 945$. *Estimated log volatility, correlogram of raw data and corresponding squares.*

If we write the jackknife density as $\alpha_t|\alpha_{\setminus t} \sim \mathsf{N}(\mu_t, \sigma_t^2)$, then we can draw from $\alpha_t|y_t, \alpha_{\setminus t}$ by sampling from [1]

$$\mathsf{N}\left[\mu_t + \frac{\sigma_t^2}{2}\left\{\frac{y_t^2}{\beta^2}\exp\left(-\mu_t\right) - 1\right\}, \sigma_t^2\right],$$

and rejecting with probability (3). For typical problems this algorithm accepts with probability of approximately $1 - \frac{y_t^2}{4\beta^2}\exp(-\mu_t)\sigma_t^2$; a derivation of this result is given in the Appendix. The acceptance probability is typically over 99% for most financial datasets and seems very robust. On a P5/133 computer with $n = 500$, this sampler carries out approximately 100 complete sweeps in about one second. Notice that this approach allows the end conditions to be easily dealt with as the only thing that changes is the states' jackknife density. To compute $\alpha_1|\alpha_2$ we need to make a prior assumption about α_1. Here we assume the α_t process is stationary, although the same algorithm would work if it is given a diffuse prior.

2.3.2 Pseudo-dominating Metropolis sampler

In Section 3 we will use pseudo-dominating Metropolis samplers in high dimensional multi-move samplers. Here we will develop this method in the single-move case to introduce some

[1] In different contexts, first order Taylor expansions of the likelihood are used in the work of Chib and Greenberg (1994) and Chib and Greenberg (1995) to generate Metropolis suggestions.

of the ideas. We exploit a second order expansion [2] of the log-density

$$
\begin{aligned}
\log f(\alpha_t|y_t, \alpha_{\backslash t}) &= \log f(\alpha_t|\alpha_{t-1}, \alpha_{t+1}) + \log f(y_t|\alpha_t) \\
&= -\tfrac{(\alpha_t-\mu_t)^2}{2\sigma_t^2} - \tfrac{\alpha_t}{2} - \tfrac{y_t^2}{2\beta^2}\exp(-\alpha_t) \\
&\approx -\tfrac{(\alpha_t-\mu_t)^2}{2\sigma_t^2} - \tfrac{\alpha_t}{2} - \tfrac{y_t^2}{2\beta^2}\exp(-\mu_t)\left\{1 - (\alpha_t - \mu_t) + \tfrac{1}{2}(\alpha_t - \mu_t)^2\right\} \\
&= \log g.
\end{aligned}
$$

The quadratic term in $\log g$ means that it does not bound $\log f$. This delivers a pseudo-dominating suggestion, with suggested draws z for $\alpha_t|y_t, \alpha_{\backslash t}$ being made from[3]

$$
\mathsf{N}\left[\frac{\sigma_t^{*2}}{\sigma_t^2}\mu_t + \frac{\sigma_t^{*2}}{2}\left\{\frac{y_t^2}{\beta^2}\exp(-\mu_t)(1+\mu_t) - 1\right\}, \sigma_t^{*2}\right], \quad \text{where} \quad \sigma_t^{*-2} = \sigma_t^{-2} + \frac{y_t^2}{2\beta^2\exp(\mu_t)}.
$$

Notice that if $y_t = 0$, then $\sigma_t^{*2} = \sigma_t^2$ and f is truly normally distributed and equals g. The precision of σ_t^{*-2} increases with y_t^2.

In the accept/reject part of the algorithm, these suggestions are accepted with probability $\min(f/g, 1)$, while the Metropolis probability of accepting z is

$$
\min\left(\frac{f(z \mid y_t, \alpha_{\backslash t}^{(k)}) \min\left[f(\alpha_t^{(k)} \mid y_t, \alpha_{\backslash t}^{(k)}), g(\alpha_t^{(k)})\right]}{f(\alpha_t^{(k)} \mid y_t, \alpha_{\backslash t}^{(k)}) \min\left[f(z \mid y_t, \alpha_{\backslash t}^{(k)}), g(z)\right]}, 1\right),
$$

and now has to be calculated and employed when $f(z)/g(z) > 1$ or $f(\alpha_t^{(k)})/g(\alpha_t^{(k)}) > 1$.

A slightly less direct way of thinking about this analysis is that we are using a Gaussian approximation to the log-likelihood, $\log f(y_t|\alpha_t)$, which is then added to the then-conjugate Gaussian prior to deliver a Gaussian posterior. Notice that as y_t goes to zero this likelihood becomes uninformative, although the posterior is perfectly well-behaved. This way of thinking about this problem is easily extended to updating more than a single state at a time.

2.3.3 Samplers for the parameters

In this section we will pursue a Bayesian analysis of the parameters in the model. As we will use this setup three different times in this paper, it is useful to spell out our assumptions in a some detail here. Given the states, sampling from the parameters of the model is straightforward for β and σ_η^2. First assuming a flat prior for $\log\beta$ we achieve the posterior $\beta^2|y, \alpha \sim \chi_n^{-2}\sum y_t \exp(-\alpha_t)$, while assuming a prior of $\chi_p^{-2}S_0$ for $\sigma_\eta^2|\phi$ we have

$$
\sigma_\eta^2|\phi, \alpha \sim \chi_{n+p}^{-2}\left\{\sum_{t=2}^{n}(\alpha_t - \phi\alpha_{t-1})^2 + \alpha_1^2(1-\phi^2) + S_0\right\}.
$$

In our work we assume for daily data $p = 10$ and $S_0 = p \times 0.01$, while for weekly data S_0 is taken to be $p \times 0.05$. Our prior on the persistence parameter ϕ will be designed to ensure that

[2]A similar quadratic expansion appears in Green and Han (1990) in their analysis of Poisson images with a Gaussian prior.

[3]By analogy with Theorem 1, it is possible to carry out a quadratic Taylor expansion of $\exp(-\alpha_t)$ at any point and still produce a Gaussian proposal density. A higher rate can be obtained by iterating the second order expansion. The result would be a Laplace approximation suggestion density.

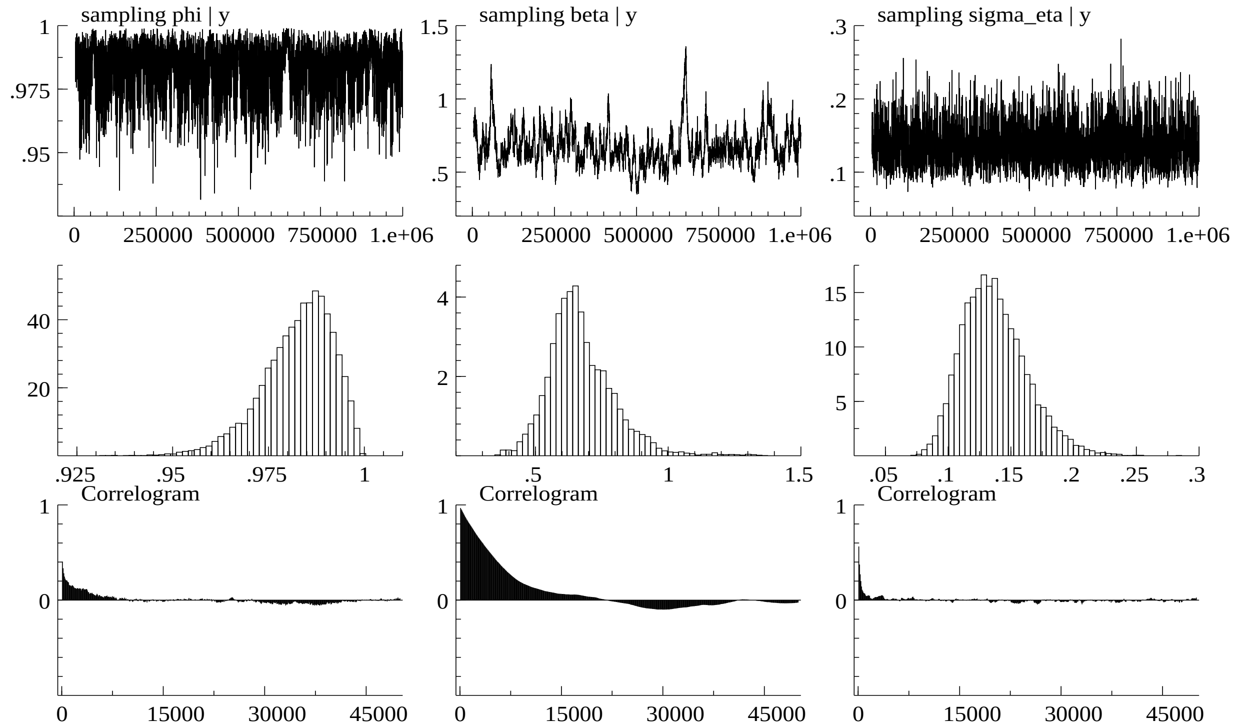

Figure 2: *Daily returns for the Pound against the Dollar. Top graphs: the simulation against iteration number. Middle graphs: histograms of the resulting marginal distribution. Bottom graphs: the corresponding correlogram for the iterations.*

the log volatility process is stationary. To carry this out we employ a beta prior on $(\phi + 1)/2$, with $\mathsf{E}(\phi) = \{2\delta/(\gamma+\delta)\} - 1$. In the analysis we use below, we set $\delta = 20$ and $\gamma = 1.5$ so the prior mean is 0.86 and standard deviation is 0.11. It could be argued that our prior should be closer to a unit root. As a result of our prior choice, the posterior is non-conjugate and we sample from $\phi|\sigma_\eta^2, \alpha$ using an accept/reject algorithm (using a variant of Theorem 1).

2.3.4 Empirical effectiveness

Our Gibbs sampler was initialised by setting all the log-volatilities to zero and $\phi = 0.95$, $\sigma_\eta^2 = 0.02$ and $\beta = 1$. We iterated the Gibbs sampler on the states for 1,000 iterations and then the parameters and states for 50,000 more iterations, before recording any answers. The next 100,000 iterations are graphed in Figure 2 and summarised in Table 1.

The correlogram shows significant autocorrelations for ϕ at 10,000 lags, for β at 25,000 lags and for σ_η at 5,000 lags. This is not an unfamiliar pattern for Gibbs samplers in high dimensional problems (see Ripley and Kirkland (1990) for a good example in spatial statistics).

2.4 Duration time series

Our second example will be a model of consecutive durations. Observation-driven autoregressive models of this type have been suggested in work by Wold (1948), Cox (1955) and,

	Mean	Monte Carlo S.E.	Covariance & *Correlation* of Posterior		
$\phi\|y$	0.9821	0.000277	8.343×10^{-5}	*-0.629*	*0.217*
$\sigma_\eta\|y$	0.1382	0.000562	-0.0001479	0.0006631	*-0.156*
$\beta\|y$	0.6594	0.0121	0.0002089	-0.0004228	0.01209
Computer time	233,303				

Table 1: *Daily returns for the Pound against the Dollar. Summaries of Figure 2. The S.E. of simulation is computed using 100,000 lags. Italics are correlations rather than covariances. Computer time is seconds on a P5/133.*

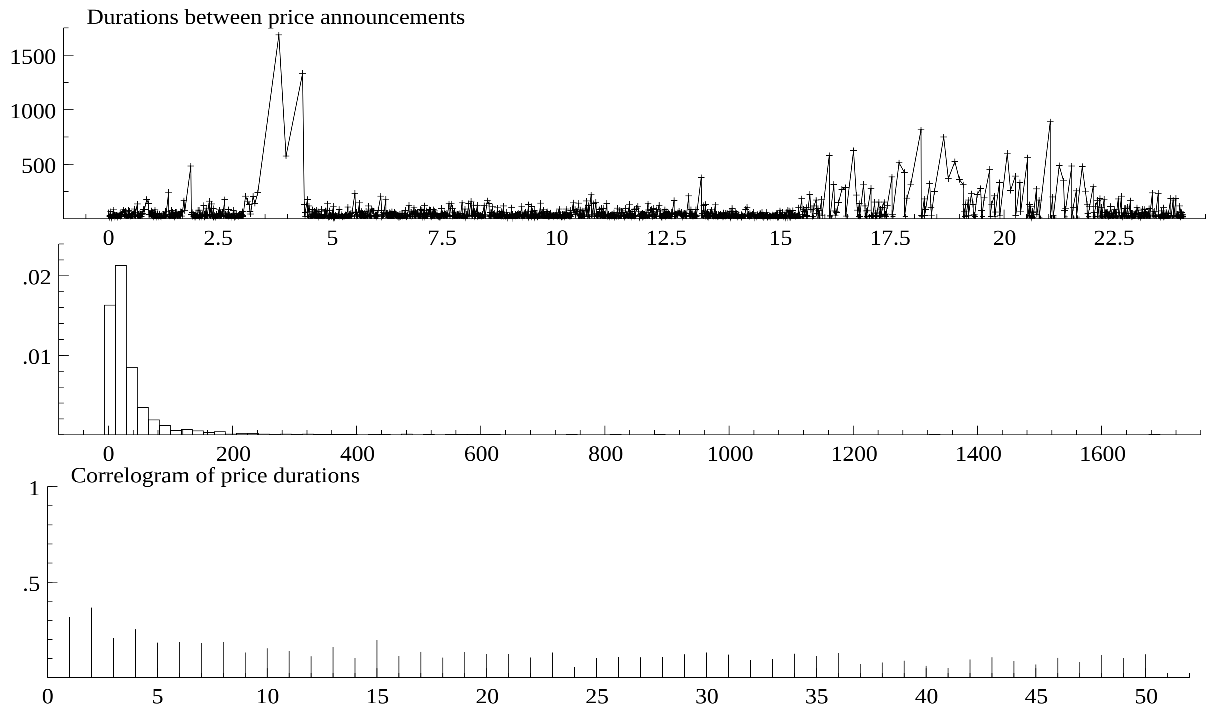

Figure 3: *Time between price announcements in the Yen/Dollar market on 5 April 1993. Histogram of the unconditional distribution and correlogram of the consecutive duration times.*

more recently, Engle and Russell (1994). Parameter-driven models have been analysed by Gamerman (1992).

Here we work with a very simple exponential distribution model to summarise a day of durations in seconds between price announcements (bids and asks) in the Japanese Yen against the US Dollar exchange rate recorded in the Olsen and Associates tape for 5th April 1993. Summary information for this series is given in Figure 3. If y_t denotes the time between price announcements, then the basic model will be the same as the SV model (4) but with the normality assumption on ε_t replaced by an exponential density with mean one.

An obvious model for this type of data would be a doubly stochastic Poisson point process (a Cox process) with a log-link intensity model which evolves according to an Ornstein-Uhlenbeck process. Here we use a discrete time approximation to this model, with exponentially distributed time intervals

$$\begin{aligned} f(y_t|\lambda_t) &= \lambda_t \exp(-\lambda_t y_t), \\ \lambda_t &= \beta \exp(-\alpha_t), \\ \alpha_{t+1} &= \phi\alpha_t + \eta_t, \eta_t \sim \mathsf{NID}(0, \sigma_\eta^2). \end{aligned}$$

This type of discrete time approximation to the continuous time model was used previously by Gamerman (1992) and Carlin, Gelfand, and Smith (1992).

The time series behaviour of this model can be summarised by

$$\mathsf{E}(y_t) = \mathsf{E}\left(\lambda_t^{-1}\right) = \beta^{-1}\exp(\sigma_\alpha^2/2), \quad \mathsf{Var}(y_t) = \beta^{-2}\exp(\sigma_\alpha^2)\left\{2\exp(\sigma_\alpha^2) - 1\right\},$$

and

$$\begin{aligned} \mathsf{Cov}(y_t, y_{t+s}) &= \mathsf{E}\left(\lambda_t^{-1}\lambda_{t+s}^{-1}\right) + \mathsf{Cov}(\lambda_t^{-1}\lambda_{t+s}^{-1}), \\ &= \beta^{-2}\exp(\sigma_\alpha^2)\left[2\exp\left\{\sigma_\alpha^2(1+2\phi^s)\right\} - 1\right]. \end{aligned}$$

This implies

$$\mathsf{Corr}(y_t, y_{t+s}) = \frac{\left[2\exp\left\{\sigma_\alpha^2(1+2\phi^s)\right\} - 1\right]}{\left[2\exp\left(\sigma_\alpha^2\right) - 1\right]}$$

which has a similar structure to an ARMA(1,1) model.

This model can be analysed using the same Gibbs sampling approach used for the SV model, but now the proposal density becomes

$$\mathsf{N}\left[\mu_t + \sigma_t^2\left\{\frac{y_t}{\beta}\exp\left(-\mu_t\right) - 1\right\}, \sigma_t^2\right],$$

with acceptance probability $\exp\left[-\frac{y_t}{\beta}\left\{\exp(-\alpha_t) - \exp(-\mu_t)\left(1 - (\alpha_t - \mu_t)\right)\right\}\right]$, which gives an average acceptance rate of approximately $1 - \frac{y_t}{2\beta}\exp(-\mu_t)\sigma_t^2$. The updating of the parameters is unchanged from the SV model case except for $\beta|y,\alpha \sim \mathsf{G}\left\{n+1, 1/\sum y_t\exp(-\alpha_t)\right\}$. The results of the analysis are given in Figure 4 for this data. Unlike the previous case the sampler works in a satisfactory way, converging quickly. This is due to the much lower level of persistence in the log-mean durations than the corresponding log-volatilities.

2.5 Nonparametric generalized linear models

Our next motivating example will be taken outside the natural field of time dependence, non-parametric regression. Here we will relate some dependent variable y_t to a univariate regressor x_t. Here our generic form will be used as an efficient computational device to allow the application of spline smoothing: see Silverman (1984) and Green and Silverman (1994, pp.47-49) for the connection between spline smoothing and kernel methods. This will be based on the penalised likelihood suggested by Wahba (1978), Kohn and Ansley (1987) and Green and Silverman (1994):

$$\sum_{t=1}^{n}\left[y_{[t]}g(\tau_t) - b\left\{g(\tau_t)\right\} + c(\tau_t)\right] - \frac{\sigma_\eta^2}{2}\int_a^b g''(\tau)d\tau,$$

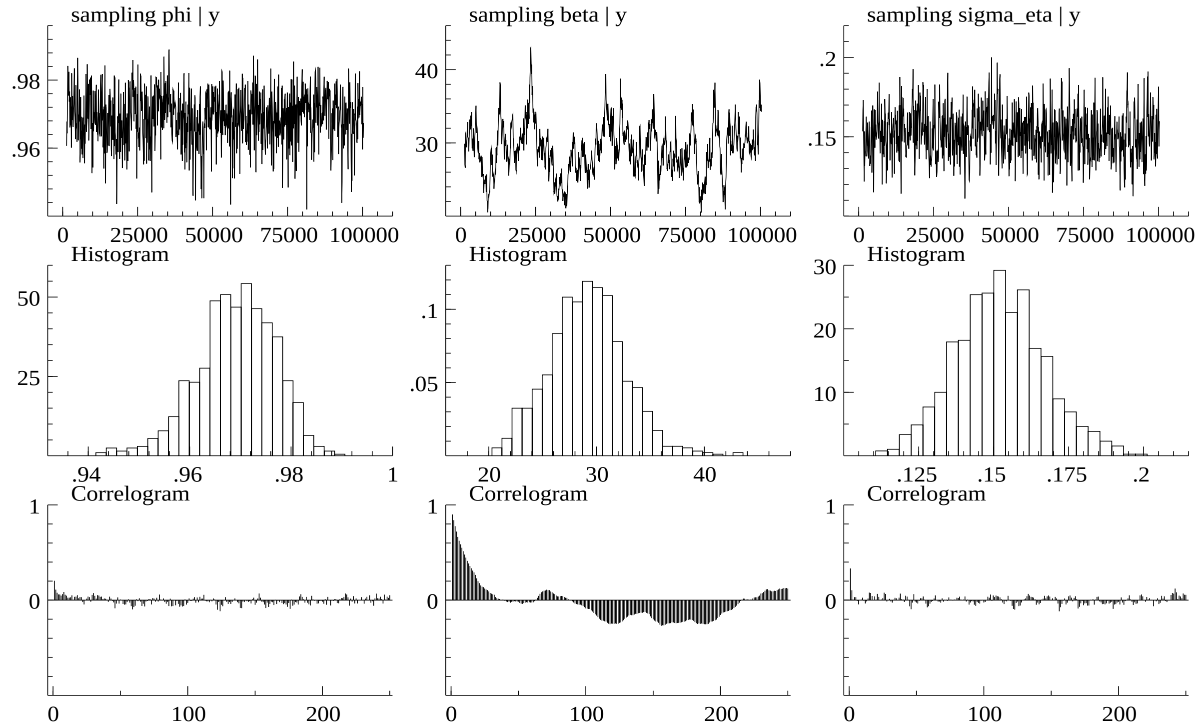

Figure 4: *Single-move Gibbs sampler for the duration time series model. Top graphs: the simulation against iteration number. Middle graphs: histograms of the resulting marginal distributions. Bottom graphs: the corresponding correlograms for the iterations.*

where $\tau_1 = x_{(1)} \leq ... \leq \tau_n = x_{(n)}$ and $y_{[t]}$ corresponds to the $x_{(t)}$ observation. The ability to order the regressors is the crucial idea here. This is a generalized linear model which is based on a canonical link with $\theta_t = g(\tau_t)$.

This penalized likelihood is the joint likelihood of y and g resulting from the Bayes model with the measurement density of the exponential family for $y_t|g(\tau_t)$ and a prior density for the signal $g(\tau)$ is generated by the stochastic differential equation

$$\frac{d^2 g(\tau)}{d\tau^2} = \sigma_\eta \frac{dW(\tau)}{d\tau}, \tag{5}$$

where $W(\tau)$ is a Wiener process, with $W(0) = 0, \mathsf{Var}\{W(1)\} = 1$. Here σ_η is a roughness parameter. One useful way of thinking about this prior distribution is to regard it as a Gaussian prior with variance σ_η^2 on the second derivative of the g function. Hence with σ_η being small, the prior puts most weight on small moves in the curvature of the function, in the limit causing g to be a straight line. On the other-hand a large value of σ_η would generate a very jagged line.

Our approach will give a maximum likelihood or Bayesian estimation of this smoothing parameter and explore the whole of the posterior distribution of $g(\tau_1), ..., g(\tau_n)|y_1, ..., y_n$. Typically we will report the mean and variances of these conditional random variables, but it would also be possible to record any of their quantiles.

The model (5) gives the discrete time Gaussian transition equation for $g(\tau_1), ..., g(\tau_n)$ with $\delta_t = \tau_t - \tau_{t-1}$ and where

$$z_t = (1 \quad 0), \quad T_t = \begin{pmatrix} 1 & \delta_t \\ 0 & 1 \end{pmatrix}, \quad H_t = \sqrt{\delta_t} \begin{pmatrix} \delta_t/\sqrt{3} & 0 \\ \sqrt{3}/2 & \sqrt{1/4} \end{pmatrix}.$$

The addition of explanatory variables to $g(\tau_t)$, gives $g(\tau_t) + x_t'\beta$, a semi-parametric regression model.

2.5.1 Estimating actuarial death rates

An important area of applied actuarial science is the estimation of death rates as a function of age. Here we analyse the London (1985) data on the death rates amongst a population of retired white American females. This data set has previously been used by Green and Silverman (1994, Table 5.1 & pp. 101-104) to illustrate their approach to penalized likelihood.

The data set consists of a series of recordings on the population at risk in a particular year, denoted n_t, and corresponding recorded deaths, y_t, where the index t denotes age and varies between 55 and 104. If the population at risk was regarded as being exogenous, then a sensible basic model for this data would seem to be binomial with a logistic link and a flexible model for the regressor, which in this case is simply a function of age. The crude death rates, y_t/n_t, are displayed in Figure 5. It is important to realise that some of these death rates are more precise than others as, for example, $n_{72} = 20116$ while $n_{103} = 2$.

Here we fit a logistic non-parametric regression by using a Wahba type spline prior. We write $p_t = \exp(\mu_t)/\{1 + \exp(\mu_t)\}$, then $\log f(y_t|n_t, \alpha_t)$ equals $y_t\alpha_t - n_t \log\{1 + \exp(\alpha_t)\}$. Hence the suggested draws for $\alpha_t|y_t, \alpha_{t-1}, \alpha_{t+1}$ would be $\mathsf{N}\left[\mu_t + \sigma_t^2 \{y_t - n_t p_t\}, \sigma_t^2\right]$ with the usual rejection calculation computed using Theorem 1.

	Mean	Monte Carlo S.E.	Posterior variance
$\sigma_\eta \vert y$	0.0493	0.000342	0.000415

Table 2: *Computed death rates. Summaries of the simulations, given in Figure 5 from the 90,000 replications of the single-move Gibbs sampler. The Table focuses on σ_η. The Standard Error of the simulation is computed using 1,000 lags.*

In our fitting procedure we use a non-informative prior on σ_η^2, the only parameter in the model. In our applied work we employed 100,000 iterations of the Gibbs sampler, using 10,000 iterations as a burn-in period. The resulting draws for $\sigma_\eta|y$ are given in Figure 5, together with the resulting correlogram, while the corresponding estimate of the parameter is given in Table 2. The degree of autocorrelation is low. Likewise, the graphs show the correlograms for the draws from the states given the observations. That correlation is small as well.

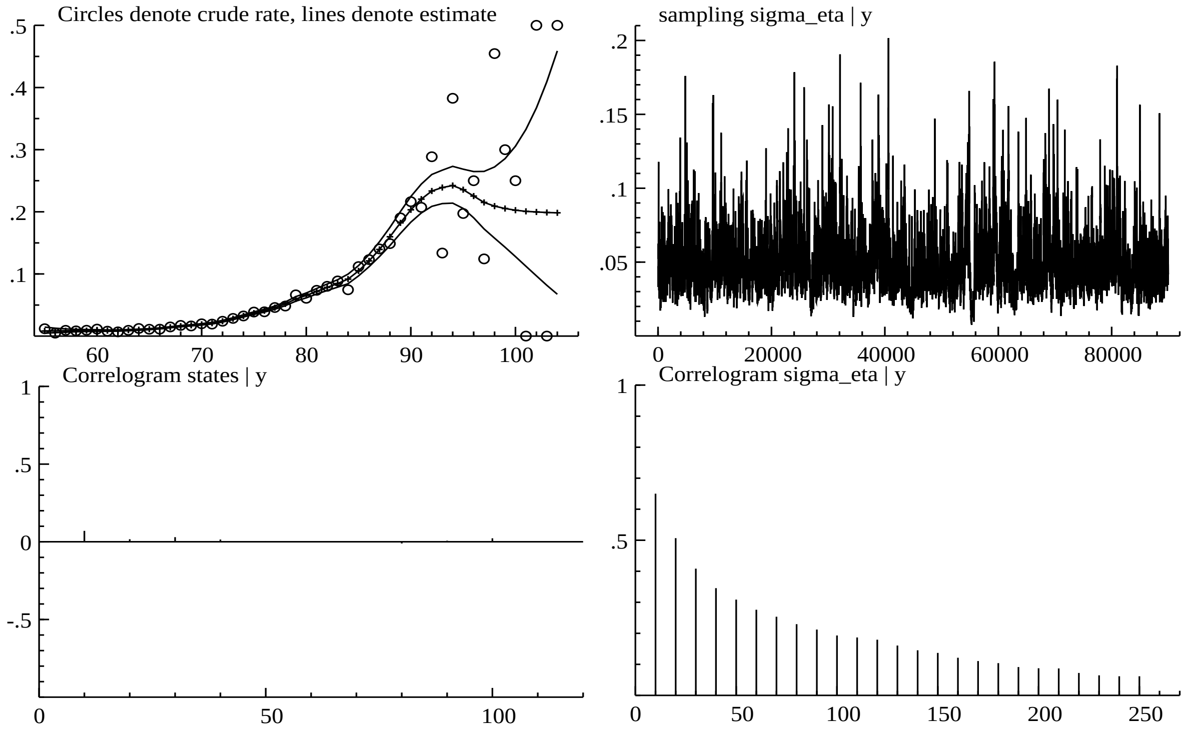

Figure 5: *Raw data on mortality rates, shown by circles, for retired American white females. The circles record y_t/n_t, a very crude death rate. As n_t varies wildly over t, being around 25,000 for each period from 65–72 and only 2 for those of age 102, the precision of the crude rates is sometimes low. The model-based fit records the estimate (line plus crosses) plus and minus two standard errors (line).*

The Gibbs algorithm for this case has worked completely satisfactorily, with low degrees of correlation in the simulations. This is not particularly surprising. First, this is a short data set. Second, the signal-to-noise ratio is quite large here, for σ_η is surprising large for a spline type model. This is due to the quite substantial non-linearities in the curve around the age of 92.

2.5.2 Thickened monthly death rates

To see how things might change, and hence assess the fragility of the single-move computational approach, we will keep the same model but manipulate the data in order to lengthen the data set and reduce the signal-to-noise ratio. This can be carried out by thickening the data using simulations based on the fit of the model previously reported. We will convert the data into pseudo-monthly seasonally-adjusted data for ages 55-98 by assigning each member of the population to a particular month within their true year and then drawing the number of deaths using the fitted estimated death rate. Hence the simulated data will depend on the real data only through their effect on the fitted curve.

The results from the fitting exercise are given in Figure 6. This graph is different from the ones previously displayed. The Figure shows the estimated death rates for given numbers

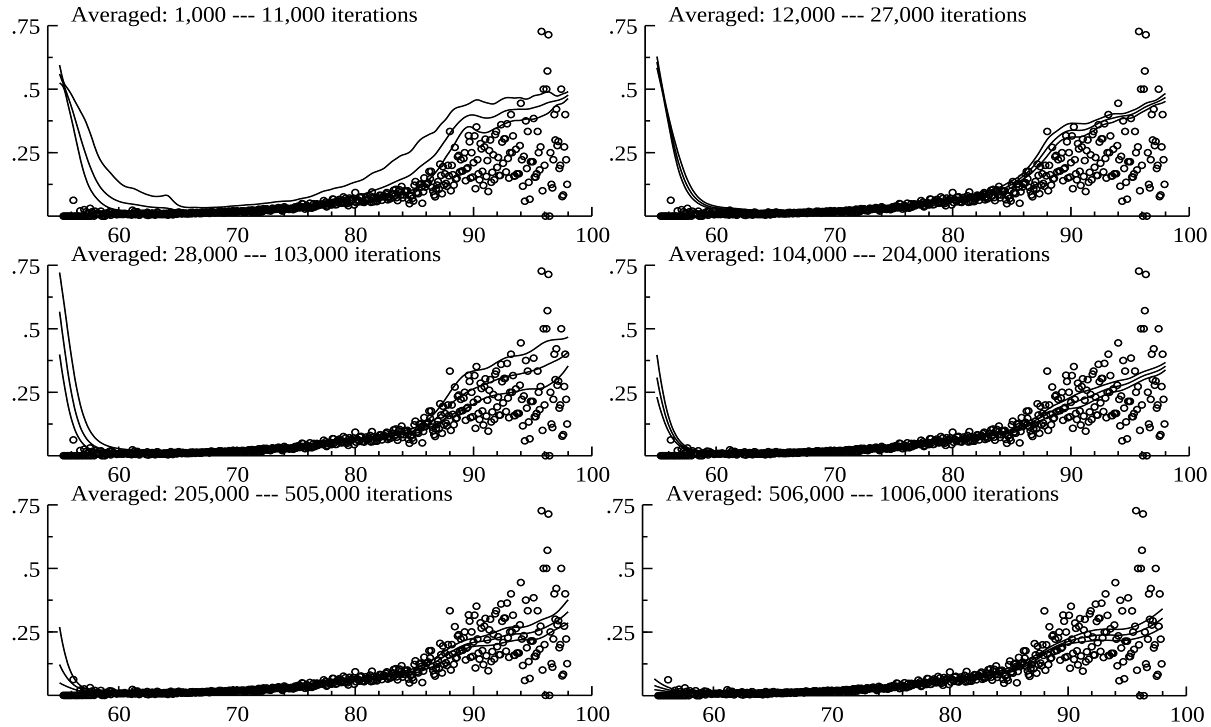

Figure 6: *Thickened data on mortality rates, shown by circles, for retired American white females. Only ages 65–98 are considered here. The circles record y_t/n_t, a very crude death rate. The model-based fit records the estimate (line plus crosses) plus and minus two standard errors (line). The different graphs record these fits for different numbers of iterations from a single-move Gibbs sampler.*

of iterations. The first graph has an initial burn-in of 1,000 iterations, starting off using the $\sigma_\eta^2 = 0.00243/12$ — the value implied for monthly data by the previous data set. The graph displays the results from 10,000 iterations. The next graph averages the next 15,000 iterations, while the following graph reports after another 75,000, 100,000, 300,000, and 500,000 iterations respectively.

The implications of this graph are important. The methods require vast numbers of iterations for them to produce empirically useful results.

2.6 Index models

An influential model in econometrics is the index model, which relates $\mathsf{E}(y_t|x_t'\beta) = g(x_t'\beta)$. In the recent literature there is a desire to allow g to be flexibly determined from the data, rather than by selecting a particular link function: see Stoker (1986) and Stoker (1992). This can be carried out as before by using the likelihood

$$\sum_{t=1}^{n}\Big[y_{[t]}g(\tau_t) - b\left\{g(\tau_t)\right\} + c(\tau_t)\Big] - \frac{\sigma_\eta^2}{2}\int_a^b g''(\tau)d\tau, \text{ where } \tau_t = (x'\beta)_{(t)}.$$

Notice that if the roughness parameter is zero then this model defaults to using the canonical link. This setup is analysed by Green and Silverman (1994, pp. 121-125). Again in our treatment we will give full likelihood analysis of this model.

2.7 Periodogram smoothing

In a recent paper Carter and Kohn (1995) have used MCMC methods to smooth the periodogram. The approach is based on the fact that for moderate samples the periodogram evaluated at the Fourier frequencies are independent with $I(\omega_j) \sim f(\omega_j)\chi_2^2/2$, $\omega_j \neq 0, \pi$ and $I(\omega_j) \sim f(\omega_j)\chi_1^2$, $\omega_j = 0, \pi$ where

$$I(\omega) = \frac{1}{n}\left|\sum_{t=1}^{n} y_t \exp\left(-it\omega\right)\right|, \quad \text{and} \quad \omega_j = \frac{2\pi j}{n}, \quad j = 0, ..., \lfloor n/2 \rfloor,$$

and f is the spectral density. Their approach is to linearise by using a log transformation of the periodogram, following Wahba (1980), and then use a mixture of normals representation of the resulting $\log \chi_2^2/2$ and $\log \chi_1^2$ distributions. This allows the use of the simulation smoothing device of de Jong and Shephard (1995) to estimate the spectral density.

Here we avoid using the log transformation and approximation of the χ^2 distributions. The prior for the density will be $\log f(\omega) = g(\omega)$, where g is given by equation (5) (see Kohn and Ansley (1987)). Again a discrete time state space representation of this process is possible. We can add the constraint that $d \log f(\omega = \pi)/d\omega = 0$.

2.8 General remarks

In this section we have shown that reasonably simple and general single-move Gibbs samplers or pseudo-dominating Metropolis algorithms can be developed for this wide class of parameter-driven dependence models. Unfortunately, although these methods are sometimes quite useful for empirical work, we have shown that they are fragile. It is very difficult to empirically assess the convergence of these methods due to the enormous dimensions of the problems. Sometimes apparently well-behaved methods hide poor performance. For other problems, small changes in the model or the data set change the methods from being useful to being close to useless.

In our view we need more reliable methods if likelihood methods are going to be widely used for parameter-driven dependence problems. In the next section we develop one such approach.

3 Designing blocks

Liu, Wong, and Kong (1994) suggest that a key feature to improve the speed of convergence for samples is to use blocks. This has motivated a considerable time series statistics literature on this topic in models built out of Gaussianity, but with some non-normality mixed in: see the work of Carter and Kohn (1994), Shephard (1994) and de Jong and Shephard (1995).

However, parameter-driven models are not of this form: they are fundamentally non-Gaussian. Hence there is a substantial need for a new strategy.

Before we discuss our suggestion, we need to be careful in our notation. The states α usually have many identities linking the state variables, for

$$\begin{array}{lllll} h(\theta_t) & = & d_t & = & z_t\alpha_t + x_t\beta \\ & & \alpha_{t+1} & = & W_t\beta + T_t\alpha_t + H_t u_t, \quad u_t \sim \mathsf{NID}(0, I), \\ & & \alpha_1 & \sim & \mathsf{N}(a_{1|0}, p_{1|0}). \end{array}$$

consequently when we simulate from $\alpha \mid y$ we simulate from a possibly highly degenerate density. To avoid this complication we can focus on the fundamental disturbances in the model u_t, ensuring that $u \mid y$ is non-degenerate. Throughout the rest of this section we will ignore the role of the regressors W_t and x_t.

Sampling from $u \mid y$ may be too ambitious as this is highly multivariate and so if n is very large it is likely that we will run into very large rejection frequencies, counteracting the effectiveness of the blocking. Hence we will employ a potentially intermediate strategy. Our sampling method will be based around sampling blocks of disturbances, say $u_{t,k} = (u_{t-1}, ..., u_{t+k-1})'$ given beginning and end conditions, α_{t-1} and α_{t+k+1}, and the observations. We call these end conditions "stochastic knots", the nomenclature being selected by analogy with their role in splines.

At the beginning and end of the data set, there will be no need to use two sided knots. In these cases typically only single knots will be required, simulating, for example, from $u_{0,k}|\alpha_{t+k+1},\, y_1, ..., y_{t+k}$.

In practice k will be a tuning parameter, allowing the lengths of blocks to be selected. Typically it will be chosen to be stochastic, varying the stochastic knots at each iteration of the samplers. If k is too large the sampler will be slow because of rejections, too small it will be correlated because of the structure of the model.

3.1 Proposal density

As far as we can see the only viable way of carrying out sampling from $u_{t,k}|\alpha_{t-1},\alpha_{t+k+1}, y_t, ..., y_{t+k}$ is to place it within a Metropolis algorithm, employing multivariate Gaussian suggestions. We now show how to do this is an effective way. The basis of our work will be the use of a Taylor type expansion of

$$\log f = \log f(u_{t-1}, ..., u_{t+k-1} \mid \alpha_{t-1}, \alpha_{t+k+1}, y_t, ..., y_{t+k})$$

around some preliminary estimate of $u_{t,k}$ made using the conditioning arguments $\alpha_{t-1}, \alpha_{t+k+1}$ and $y_t, ..., y_{t+k}$. These estimates, and the corresponding $\alpha_t, ..., \alpha_{t+k}$ and $\theta_t, ..., \theta_{t+k}$, will be denoted by hats. How they are formed will be discussed in a moment. We will need some notation. We will write $l(\theta_t)$ to denote $\log f(y_t|\theta_t)$ and the derivative of this log-density with respect to d_t as l' and l'' respectively. The expansion is then

$$\begin{array}{rcl}\log f &=& -\frac{1}{2}u_{t,k}'u_{t,k}+\sum_{s=t}^{t+k}l(\theta_s), \quad \alpha_{s+1}=T_s\alpha_s+H_su_s.\\ &\simeq& -\frac{1}{2}u_{t,k}'u_{t+k}+\sum_{s=t}^{t+k}l(\widehat{\theta}_s)+z_s(\alpha_s-\widehat{\alpha}_s)l'(\widehat{\theta}_s)+\frac{1}{2}\left\{z_s(\alpha_s-\widehat{\alpha}_s)\right\}^2D_s(\widehat{\theta}_s)\\ &=& \log g.\end{array}$$

We will require the assumption that the as yet unspecified $D_s(\theta)$ is everywhere strictly negative as a function of θ. Typically we will take $D_s(\widehat{\theta}_s)=l''(\widehat{\theta}_s)$ so that the approximation is a second order Taylor expansion. This will be convenient, for in the vast majority of cases l'' will be everywhere strictly negative. However, we let the method have the possibility that we will not take D_s as the second derivative so that we cover unusual cases as well. Of course for those cases, we will have to provide sensible rules for the selection of D_s.

A crucially attractive feature of this expansion is that the error in the approximation involve only the difference between $l(\theta_s)$ and $l(\widehat{\theta}_s)+z_s(\alpha_s-\widehat{\alpha}_s)l'(\widehat{\theta}_s)+\frac{1}{2}\left\{z_s(\alpha_s-\widehat{\alpha}_s)\right\}^2D_s(\widehat{\theta}_s)$, not the transition equation $u_{t,k}'u_{t,k}$. The implication of this is that the algorithm should not become significantly less effective as the dimension of u_t increases. This can be contrasted with other approaches such as the numerical integration routines used in Kitagawa (1987), whose effectiveness usually deteriorate as the dimension of u_t increases..

Some interesting relations can be found for the crucial l' function. First, in general

$$l'(\widehat{\theta}_s)=\left.\frac{\partial\theta_t}{\partial d_t}\right|_{\theta_t=\widehat{\theta}_t}\times\left.\frac{\partial l(\theta_t)}{\partial\theta_t}\right|_{\theta_t=\widehat{\theta}_s},$$

which for the exponential family becomes $l'(\widehat{\theta}_s)=\left.\frac{\partial\theta_t}{\partial d_t}\right|_{\theta_t=\widehat{\theta}_t}\times\left\{y_t-\dot{b}\,(\widehat{\theta}_t)\right\}$. Similar manipulations are possible for $l''(\widehat{\theta}_t)$. The most interesting case of this is the canonical link, where $l''(\widehat{\theta}_t)=-\ddot{b}\,(\widehat{\theta}_t)$.

This type of expansion also appear in the work of Durbin and Koopman (1992) where a sequential expansion based on filtering and smoothing algorithms was used to provide an approximate likelihood analysis of a wide class of non-Gaussian models. The way it is used will be documented in a moment.

3.1.1 Gaussian approximate model

The density of g is highly multivariate Gaussian. It is clearly not a dominating density for $\log f$, but there is some hope that it will be a good approximation. If we write $v_s^{-1}=-D_s(\widehat{\theta}_s)$, then the joint density can be calculated by using the model:

$$\begin{array}{rcl}\widehat{y}_s &=& z_t\alpha_s+\varepsilon_s, \quad \varepsilon_s\sim\mathsf{N}\left(0,v_s\right)\\ \alpha_{s+1} &=& T_s\alpha_s+H_su_s, \quad u_s\sim\mathsf{NID}(0,I),\\ \widehat{y}_s &=& z_t\widehat{\alpha}_s+v_sl'(\widehat{\theta}_s), \quad s=t,...,t+k.\end{array} \qquad (6)$$

Consequently, it is possible to simulate from $u_{t,k}\mid\alpha_{t-1},\alpha_{t+k+1}$ using the de Jong and Shephard (1995) simulation smoother on the constructed univariate $\widehat{y}_t,...,\widehat{y}_{t+k}$.

As g does not bound f it is not possible to use this simulation smoother inside an accept/reject algorithm within the Gibbs sampler. Rather we will use the simulation smoother to provide suggestions for the pseudo-dominating Metropolis algorithm suggested by Tierney (1994) and discussed in section 2.

A similar expansion to (6), but without the knots, is used in the work of Durbin and Koopman (1992). These authors use an expectation smoother (see the Appendix for details) to provide an estimate of $u|y$ using this Taylor expanded approximation. By iterating, using the latest smoothed values as the basis of the approximation at the start of each iteration, this algorithm converges to the mode of the density of $u|y$ in cases where $\partial^2 \log l/\partial\alpha_t\partial\alpha_t'$ is negative semi-definite; the same condition is needed for generalized linear regression models to have a unique maximum (see McCullagh and Nelder (1989, p. 117), Wedderburn (1976) and Haberman (1977)). This condition is typically stated as a requirement that the link function be log-concave. In the case being addressed here, the iterations are an efficient way of carrying out a Newton-Raphson hill climbing algorithm using analytic first and second derivatives on a concave objective function. Thus the approximations suggested by Durbin and Koopman (1992) can be interpreted as Laplace approximations to a very high dimensional density function. Our algorithm is different. We simply converge to a simulation from $u|y$.

It is informative to look at some special cases of the setup given in (6), to see the effect of non-Gaussianity of the approximating model.

Example 1: canonical link Some simplifications result if a canonical link is assumed, that is $\theta_t = d_t$ and $log f = y_t\theta_t - b(\theta_t) + c(y_t)$. Then

$$v_s^{-1} = \ddot{b}\left(\widehat{\theta}_s\right) \text{ and } \widehat{y}_s = d_s + v_s\left\{y_s - \dot{b}\left(\widehat{\theta}_s\right)\right\}.$$

A special case of this is the poisson model, where $b(\theta_s) = \exp(\theta_s)$ and so

$$v_s^{-1} = \exp(d_s),\ \widehat{y}_s = d_s + \exp(-d_s)\left\{y_s - \exp(d_s)\right\}.$$

Another important example is the binomial, where $b(\theta_s) = n\log\left\{1+\exp(\theta_t)\right\}$. For this model

$$v_s^{-1} = np_t(1-p_t),\ \widehat{y}_s = d_s + (y_s - np_s)/\left\{np_t(1-p_t)\right\}, \text{where } p_t = \exp(d_t)/\left\{1+\exp(d_t)\right\}.$$

Example 2: SV model This model has $\log f(y_t|\alpha_t) = -\alpha_t/2 - y_t^2\exp(-\alpha_t)/2\beta^2$. Thus

$$v_s^{-1} = \frac{y_s^2}{2\beta^2}\exp(-d_s),\ \widehat{y}_s = d_s + \frac{v_s}{2}\left\{y_s^2\exp(-d_s)/\beta^2 - 1\right\}.$$

This case is particularly interesting as v_s depends on y_s, which cannot happen in the canonical link case. At first sight this raises some problems for as $y_s \to 0$ so $v_s^{-1} \to 0$ and $\widehat{y}_s \to -\infty$. This would suggest that the suggestions from the simulation smoother might always be rejected.

However, this observation ignores the role of the prior distribution, which will in effect treat such observations as missing. Of course there is a numerical overflow problem here, but that can be dealt with in a number of ways without resulting in any approximation.

Example 3: heavy tailed SV model This argument can be extended to allow ϵ_t in the SV model, (4), to follow a scaled t-distribution, $\epsilon_t = t_t/\sqrt{(\nu-2)/\nu}$, where $t_t \sim t_\nu$. Then

$$\begin{aligned}
\log f(y_t|\alpha_t) &= -\frac{\alpha_t}{2} - (\nu+1)/2 \log\left[1 + \frac{y_t^2 \exp(-\alpha_t)}{\beta^2(\nu-2)}\right], \quad \text{so} \\
l'(\alpha_t) &= \frac{1}{2}\left[\frac{\frac{(\nu+1)}{2\beta^2(\nu-2)} y_t^2 \exp(-\alpha_t)}{1 + \frac{y_t^2 \exp(-\alpha_t)}{\beta^2(\nu-2)}} - 1\right], \quad \text{and} \\
l''(\alpha_t) &= -\frac{(\nu+1)}{4}\left[\frac{y_t^2 \exp(-\alpha_t)/\beta^2(\nu-2)}{(1 + \frac{y_t^2 \exp(-\alpha_t)}{\beta^2(\nu-2)})^2}\right].
\end{aligned}$$

The resulting v_s^{-1} and $\widehat{y_s}$ are easy to compute.

This approach has some advantages over the generic outlier approaches for Gaussian models suggested in Shephard (1994) and Carter and Kohn (1994), which explicitly use the mixture representation of a t-distribution.

Example 4: factor SV model An economically interesting factor SV model can be constructed, from the corresponding ARCH work of Diebold and Nerlove (1989) and King, Sentana, and Wadhwani (1994). In the simplest univariate case it takes on the form

$$y_t = \epsilon_t \beta \exp(\alpha_t/2) + \omega_t$$

where the model is the same as (4) but with added measurement error $\omega_t \sim \mathsf{NID}(0, \sigma_\eta^2)$. With this model it is not possible to set α_t so as to drive the variance of y_t to zero, as it is bounded from below by σ_ω^2. Interestingly

$$\log f(y_t|\alpha_t) = -\frac{1}{2}\log(\beta^2 \exp(\alpha_t) + \sigma_\omega^2) - \frac{y_t^2}{2(\beta^2 \exp(\alpha_t) + \sigma_\omega^2)},$$

is not necessarily concave in α_t. Notice that

$$l'(\alpha_t) = \frac{\beta^2 \exp(\alpha_t)}{2(\beta^2 \exp(\alpha_t) + \sigma_\omega^2)} \times \left[\frac{y_t^2}{(\beta^2 \exp(\alpha_t) + \sigma_\omega^2)} - 1\right],$$

and so

$$l''(\alpha_t) = -\frac{1}{2}\frac{\beta^4 \exp(2\alpha_t)}{(\beta^2 \exp(\alpha_t) + \sigma_\omega^2)^2} \times \left[\frac{2y_t^2}{(\beta^2 \exp(\alpha_t) + \sigma_\omega^2)} - 1\right],$$

which can be positive if y_t is small. There are a number of approaches which can be suggested to overcome this problem. We could ignore the contribution of the $-1/2\log(\beta^2 \exp(\alpha_t) + \sigma_\omega^2)$ term to the second derivative. This would give us

$$D_t = -\frac{y_t^2 \beta^2 \exp(2\alpha_t)}{(\beta^2 \exp(\alpha_t) + \sigma_\omega^2)^3}.$$

This would be a good approximation if σ^2_ω is small or if y_t^2 is big. In any case the approximation does not affect the validity of the approach, only the rejection probability.

3.1.2 Finding $\widehat{\theta}_t, ..., \widehat{\theta}_{t+s}$

It is important to select sensible values for the sequence $\widehat{\theta}_t, ..., \widehat{\theta}_{t+k}$. The most straightforward choice would be to take them as the mode of $f(\theta_t, ..., \theta_{t+k}|\alpha_{t-1}, \alpha_{t+k+1}, y_t, ..., y_{t+k})$, the points at which the quadratic expansion is carried out. Although we have just noted that Durbin and Koopman (1992) show that iterating the smoothing algorithm on (6) achieves this mode, it is not a very efficient method if we have to iterate this procedure until full convergence (although Durbin and Koopman (1992) report rapid convergence, while we find results very close to the mode after 3 iterations). Instead we suggest using the pseudo-dominating MCMC method and only a fixed number of iterations of the smoothing algorithm to get a reasonably good sequence $\widehat{\theta}_t, ..., \widehat{\theta}_{t+k}$ instead of an "optimal" one. Typically we use 2 to 5 iterations.

3.2 Stochastic knots

The stochastic knots play a crucial role in our work. They ensure that as the sample size increases our algorithm does not fail due to excessive numbers of rejections. In our discussion so far the knots are regarded as being fixed, however in practice there are advantages in selecting them randomly, allowing the points of conditioning to change over the iterations. Our proposal is to work with a collection of stochastic knots, at times $\underset{\sim}{k} = (k_1, ..., k_K)'$ and corresponding values $\underset{\sim}{\alpha} = (\alpha'_{k_1}, ..., \alpha'_{k_K})'$, which appropriately cover the time span of the sample.

The selection of the knots will be carried out randomly and independently of the outcome of the MCMC process. In our work we have used the scheme

$$k_i = n \times (i + U_i)/(K + 2), \quad \text{where } U_i \sim \mathsf{U}(0, 1), \quad i = 1, ..., K. \tag{7}$$

Thus the selection of knots is now indexed by a single parameter K which we control. Notice that in the way we have set this up, it is not possible to recover the single-move algorithm as a special case by the selection of K.

3.3 Illustration on SV model

To illustrate the effect of blocking we will work with the SV model. We have three approaches to this problem: computing analytic convergence rates on simulated data, analysing the output from the MCMC algorithms on simulated and real data.

3.3.1 Theoretical evaluation of proposal

There is little analytic work on the effect of block sampling on the rate of convergence of Metropolis algorithms. However, it is possible to use the results of Roberts and Sahu (1995) on Gibbs samplers for multivariate Gaussian densities to produce some approximate convergence rates for our non-Gaussian problems. The expansion given above indicates a Gaussian approximation to $\alpha|y;\psi$. We look at various blocking strategies on the convergence rate for a Gibbs sampler on this approximate density. Of course these results will not be entirely satisfactory, but we hope the results may be informative.

We have to use a particular dataset to compute these rates. We will use the SV model example and two sets of simulation experiments, designed to reflect daily and weekly financial return data. The daily data have $\sigma_\eta^2 = 0.01$, $\phi = 0.99$ and $\beta = 1.0$ as the true values. This is quite close to the exchange rate data, introduced in the first section of this paper. The weekly data have less persistence, and sets $\sigma_\eta^2 = 0.1$, $\phi = 0.9$ and $\beta = 1.0$. In both cases we use $n = 500$. We compute the covariance of the Gaussian approximation to $\alpha|y$ using the de Jong and MacKinnon (1988) covariance algorithm.

	DUGS Single-move	DUGS $K = 5$	Random block $K = 5$
Daily parameters	0.9952	0.8768	0.3035
Weekly parameters	0.9530	0.6193	0.1270

Table 3: *Analytic convergence rates on the Gaussian approximation to $\alpha|y;\psi$ in the SV case. DUGS denotes deterministic updating Gibbs sampler. K denotes the number of knots. Random block is the approach advocated in this paper.*

The results given in Table 3 show the convergence rate for the single-move Gibbs sampler. As we would expect, the convergence rate for the daily data is much slower than for the weekly data, having half lives of 144 and 14.4 iterations respectively. The deterministic block-sampler is slightly better than these algorithms, when we use $K = 5$ knots with $k_i = i/(K+2)$. Its half lives are 5.3 and 1.45 iterations respectively. However, the random block sampler, with knots selected according to (7), has a half life of 0.58 and 0.34 iterations. This gives efficiency gains of approximately 146 and 19 fold respectively over the single move algorithm. Of course this does not reflect the retardation due to Metropolis rejection.

3.3.2 Output of MCMC algorithms on simulated data

The analytic results are subject to two substantial criticisms: they are based on the approximate model and they ignore the effect of sampling the parameters. Here we repeat the above experiments but now with MCMC output. Table 4 reports some results from a simulation using $n = 1,000$. The table splits into two sections. The first is concerned with estimating the states given the parameters, a pure signal extraction problem. It is clearly difficult to summarise the results for all 1,000 time periods and so we focus on the middle state, α_{500}, in all our calculations. Extensive simulations suggest that our results reported here are representative of these general results.

The second section of the Table looks at the estimation of the states at the same time as estimating the parameters of the model. Hence for that simulation the problem is a four-fold one: estimate the states and three parameters.

The simulations are analysed using a Parzen type window, see Andrews (1991). The table reports the ratio of the resulting variance of the single-move sampler to the multi-move sampler. Numbers bigger than one reflect gains from using a multi-move sampler. One interpretation of the Table is that if the ratio is x then the single-move sampler has to be iterated x times more than the multi-move sampler to achieve the same degree of precision in the estimates of interest. So if a sample is 10 times more efficient, then it produces the same degree of accuracy from 1,000 iterations as 10,000 iterations from the inferior simulator.

Weekly parameters	K=0	K=1	K=3	K=5	K=10	K=20	K=50	K=100	K=200
States|parameters	1.7	4.1	7.8	17	45	14	12	4.3	3.0
States	20	22	32	28	39	12	12	21	1.98
σ_η	1.3	1.3	1.1	1.7	2.2	1.7	1.5	1.7	1.5
ϕ	1.5	1.4	1.1	1.6	2.6	1.7	1.5	2.0	1.5
β	16	14	32	14	23	6	9	10	1

Daily parameters	K=0	K=1	K=3	K=5	K=10	K=20	K=50	K=100	K=200
States|parameters	66	98	98	85	103	69	25	8.5	2.5
States	91	40	30	60	47	80	14	27	18
σ_η	2.7	3.1	2.9	2.8	3.4	3.8	2.6	3.6	1.8
ϕ	16	13	18	18	16	18	6.8	11	5.7
β	93	51	51	76	65	106	23	27	26

Table 4: *Relative efficiency of block sampler to single-move Gibbs sampler. K denotes the number of stochastic knots used. The figures are the ratio of the computed variances, and so reflect efficiency gains. The variances are computed using 10,000 lags and 100,000 iterations in all cases except for the single-move sampler on daily parameters cases. For that problem 100,000 lags and 1,000,000 iterations were used. In all cases the burn-in period is the same as the number of lags.*

Table 4 indicates a number of results. When the number of knots is chosen to be extremely small, say 0, there is a possibility for low persistence cases that the sampler will actually perform worse than a single-move algorithm. This is because of large rejection rates. However, this result is very specialised. In the vast majority of cases the gains are quite significant and most importantly, they are largest in cases where the single-move algorithms do badly. This suggests that the use of block samplers does increase the reliability of MCMC techniques for parameter-driven models.

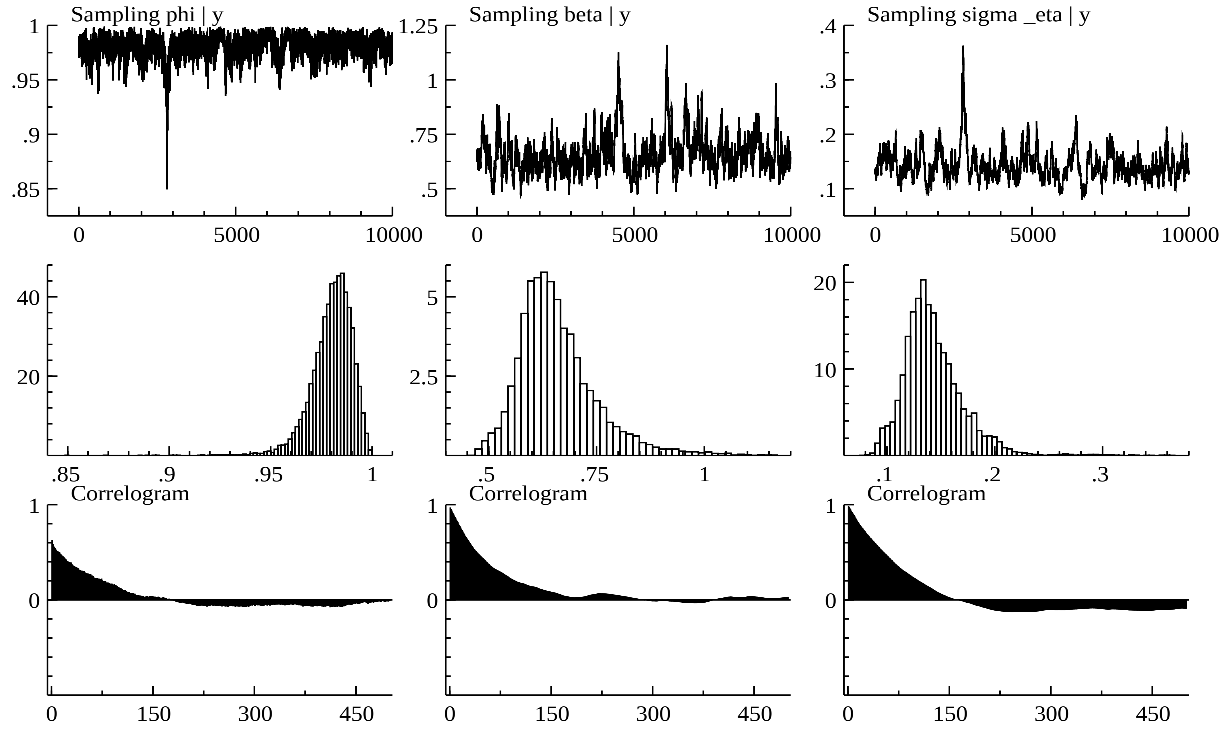

Figure 7: *Daily returns for the Pound against the US Dollar. Top graphs: the simulation aginst iteraions number for block sampler using 10 knots. Middle graphs: histograms of the resulting marginal distributions. Bottom graphs: the corresponding correlogram for the iterations.*

3.3.3 Output of MCMC algorithms on real data

To illustrate the effectiveness of this method we will return to the application of the SV model considered in Section 2. We maintain exactly the same model used earlier, but now use 10 stochastic knots in our block sampler. The basic results are displayed in Figure 7. This was generated by using 200 iterations using the initial parameter, 300 iterations updating the parameters and states and finally we record 10,000 iterations from the equilibrium path of the sampler.

The Figure shares the features of Figure 2, with the same distribution for the parameters. However, the correlations amongst the simulations are now quite manageable. This seems a workable tool.

It is often useful to report the precision of the results from the simulation estimation. Here we use the Parzen type window, using 1,000 lags. The results given in Table 5 are consistent with those given in Table 1, for the 1,000,000 iterations of the single-move algorithm. However, the precision achieved with the 1,000,000 iterations is broadly the same as that achieved by 10,000 iterations from the multi-move sampler.

	Mean	Monte Carlo S.E.	Covariance & *Correlation* of Posterior		
$\phi\|y$	0.9802	0.000734	0.000105	*-0.689*	*0.294*
$\sigma_\eta\|y$	0.1431	0.00254	-0.000198	0.000787	*-0.178*
$\beta\|y$	0.6589	0.0100	0.000273	-0.000452	0.00823
Computer time	322				

Table 5: *Daily returns for Pound against Dollar. Summaries of Figure 7, 10,000 replications of the multi-move sampler, using 10 stochastic knots. The Standard Error of the simulation is computed using 1,000 lags. The figures in italics are correlations. Computer time is in seconds on a P5/133.*

3.4 Automatic choice of block size

One of the powerful aspects of the approach we are advocating is that the block size controls the efficiency of the methods. It is important to note, however, that the methods work whatever the block size.

It would be convenient if we had a method which would automatically find the "optimal" value for the block size K. But, it seems likely that this size will be different for different aspects of the problem and in any case our results suggest extensive robustness of the efficiency gains, with large efficiency gains for blocks between 50 and 500.

3.5 Reparameterisation

There is an alternative to blocking which may reduce correlation and hence increase the speed on convergence. Reparameterisation has been highlighted by a number of statisticians in various problems; leading references include Gelfand, Sahu, and Carlin (1995), Hills and Smith (1992) and Geweke (1995) . There is some possibility of there application for dependence problems. One reparameterisation is towards sampling the disturbances, u_{it}, rather than the states, α_{it}. This has been investigated by Gamerman (1995). Although this has some advantages for some problems, a simple illustration will suffice to show that it is not satisfactory. Suppose we have a time series problem with $y_t = \alpha_t + \epsilon_t$ and $\alpha_{t+1} = \alpha_t + u_t$. Here ϵ_t and u_t are independent of one another with variances of σ_ϵ^2 and $\sigma_\epsilon^2 q$ respectively.

Interestingly as q goes to zero, the single-move Gibbs sampler's rate of convergence improves, becoming instantaneous in the limit. Unfortunately, as q goes to infinity, the sampler breaks down, never converging. This result is precisely the complement of the results for a Gibbs sampler on the states! This suggests combining samplers of disturbances and states in some manner, but such an approach would be very close to the block sampler we have suggested here. Further, the disturbances are quite difficult to Gibbs sample as the computations are in general of $O(n^2)$ per sweep, which is unsatisfactory. Hence, at the moment it looks like direct disturbance samplers have only very limited attractions.

4 Classical estimation

So far this paper has focused on Bayesian estimation of the parameters which index the models. We now turn our attention to classical inference. Our suggestion is to construct the likelihood via the use of importance sampling (Ripley (1987, pp. 122-23)). The idea will be to draw samples from the Gaussian approximation to the smoothing density we developed in section (6). Notice there will be no use of stochastic knots, nor acceptance/rejection sampling. We write this density as $f(u|y;\psi)^{\tilde{}}$, using the disturbances rather than the states, and then use the manipulation:

$$f(y;\psi) = \int \frac{f(y|u;\psi)f(u)}{f(u|y;\psi)^{\tilde{}}} f(u|y;\psi)^{\tilde{}} du.$$

The estimator of this integral will take on the form

$$\widehat{f(y;\psi)}_M = \frac{1}{M}\sum_{j=1}^{M} \frac{f(y|u^j;\psi)f(u^j)}{f(u^j|y;\psi)^{\tilde{}}}, \quad \text{where} \quad u^j \sim f(u|y;\psi)^{\tilde{}}. \tag{8}$$

Crucially, this estimator is smooth in ψ due to the Gaussianity of the smoothing density. If $f(u|y;\psi)^{\tilde{}} = f(u|y;\psi)$ then this estimate will be exact even if $M = 1$.

4.1 Technical issues

The estimate (8) involves simulating from $f(u|y;\psi)^{\tilde{}}$ and evaluating $f(u|y;\psi)^{\tilde{}}$, $f(u^j)$ and $f(y|u^j;\psi)$. The term $f(u^j|y;\psi)^{\tilde{}}$ can be found using the smoothing density result of de Jong and Shephard (1995) given in the Appendix, while $f(u^j)$ and $f(y|u^j;\psi)$ are trivial to evaluate. The simulation is straightforward using the simulation smoother, noting that the Kalman filter will have to be run only once, while the recursions for C_t and N_t have only to be compute once in order to draw M simulations.

4.1.1 Behaviour of estimator

The behaviour of this estimator can be found by the use of Proposition 1.

Proposition 1

Define $X_j = f(u^j|y;\psi)/f(u^j|y;\psi)^{\tilde{}}$ where $u^j \sim f(u^j|y;\psi)^{\tilde{}}$, then X_j is (conditional on y) independent and identically distributed with the moments

$$E(X_j|y) = 1 \quad \text{and} \quad Var(X_j|y;\psi) = \int \left(\frac{f(u|y;\psi)}{f(u|y;\psi)^{\tilde{}}} - 1\right)^2 f(u|y;\psi)^{\tilde{}} du = \sigma^2(y;\psi).$$

We define $f(y;\psi)^j = f(y|u^j;\psi)f(u^j)/f(u^j|y;\psi)^{\tilde{}}$ and so

$$f(y;\psi)^j = f(y;\psi)\frac{f(u^j|y;\psi)}{f(u^j|y;\psi)^{\tilde{}}} = f(y;\psi)X_j.$$

Consequently the estimator $f\widehat{(y;\psi)}_M$ is unbiased with variance of $\sigma^2(y;\psi)/M$.

For classical inference it is the log-likelihood which is often the focus of attention. One estimator of this $\log f\widehat{(y;\psi)}_M$ has the property

$$\log f\widehat{(y;\psi)}_M = \log f(y;\psi) + \log \frac{1}{M}\sum_{j=1}^{M} X_j.$$

The implication of this is that

$$E\left[\log\left\{1+\frac{1}{M}\sum_{j=1}^{M}(X_j-1)\right\}\right] = -\frac{\sigma^2(y;\psi)}{2M} + O(M^{-2}),$$

and so the estimator of $\log f\widehat{(y;\psi)}_M$ will be biased to $O_p(M^{-1})$. As the $\sigma^2(y;\psi)$ depends on the parameters of the model this will cause an $O_p(M^{-1})$ bias in the estimates of the parameters of the model. How big this bias will be will depend on the merit of the importance sampling device. However, importantly, it is likely that $\sigma^2(y;\psi) = O(n)$. This means that as the sample size increases the importance sampling bias will become larger unless M increases at a rate larger than n. This is unfortunate.

Simulation: small sample behaviour A straightforward way of studying the effect of the simulation size on the estimators of the parameters in the model is to conduct a simple Monte Carlo experiment. Here we repeatedly estimate the SV model using a variety of values of n, varying M. The results, given in Figure 8, suggest that if n is small the value of M can indeed be important. However, for moderate values of the samples, which would be expected in finance, the sensitivity of the results to M seems to be small.

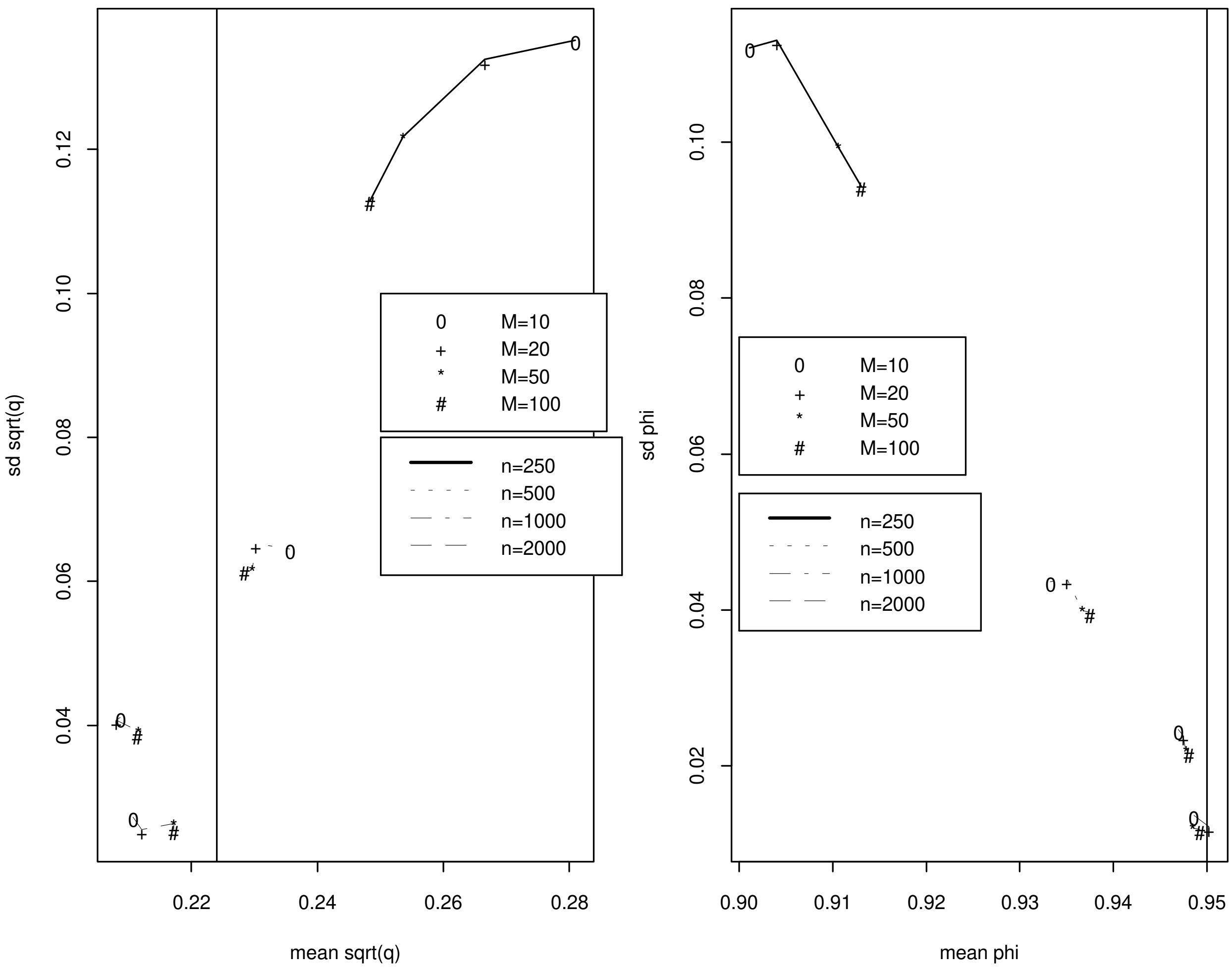

Figure 8: Maximum simulated likelihood estimators of the SV model with true values $\phi = 0.95$ and $q = 0.05$ (with $\beta = 1$ and known). Simulations of size 100, using different values of M and n. The two parameters estimated are ϕ and q. Plot maps the mean of the simulations against the standard deviation.

Removing the $O_p(M^{-1})$ bias Clearly

$$\frac{1}{M-1}\sum_{j=1}^{M}\left(f(y;\psi)^j - \widehat{f(y;\psi)}_M\right)^2,$$

is an unbiased estimator of $\sigma^2(y;\psi)$. Hence

$$\log \widehat{f(y;\psi)}_M + \frac{1}{2M}\frac{1}{M-1}\sum_{j=1}^{M}\left(f(y;\psi)^j - \widehat{f(y;\psi)}_M\right)^2,$$

will give estimates which are unbiased to $O_p(M^{-2})$. Of course this does not mean that the resulting estimates of ψ will be superior, for it could be that the introduction of this added term will sufficiently increase the variance of the estimator of the log-likelihood so as to compensate for the reduction in bias.

4.1.2 Antithetic variables

It is straightforward to exploit antithetic variables (Ripley (1987, pp. 129–132)) for this problem. In particular we will exploit antithetic simulations from $f(u|y;\psi)$ by simply switching the signs on the κ_t in the simulation smoother (9). Notice this is not the same as changing the signs of the u. In our work we have found correlations of between -0.3 and -0.65 between our antithetic draws. These seem to be substantial gains.

5 Conclusions

The "parameter-driven" theme can be used to bring together many different statistical and econometric dependence literatures. This paper uses simulation to provide a likelihood basis for these models. The paper argues that single-move algorithms are likely to be unreliable for real applications due to their slow convergence and high correlation properties. Instead we argue that the development of Taylor expansion based multi-move simulation smoothing algorithms can offer the delivery of reliable methods.

The methods have four basic advantages. (1) they integrate into the analysis of non-Gaussian models the role of the Kalman filter and simulation smoother, so fully exploiting the structure of the model to improve the speed of the methods. (2) the central expansion in the paper has been previously used in various approximate methods suggested in the time series and spline literatures. (3) the expansion only approximates $\log f(y_t|\theta_t)$ and so as the dimension of the state increases the computational efficiency of the method should not diminish significantly. (4) the new methods can be straightforwardly extended to many multivariate cases.

There is much work to be carried out in this area. We have not provided any model checking devices for our fitted models. In principle this is straightforward if based on one-step-ahead prediction densities. Some progress has recently been made in finding methods for computing these densities; see Gordon, Salmond, and Smith (1993) and Muller (1991).

6 Acknowledgements

We would like to thank the ESRC for their financial support for this research under the grant "Estimation via simulation in econometrics." Various parts of this paper have been presented under a number of titles at: American Statistical Association annual meeting, Florida August 1995; ESRC Econometric study group conference, Bristol July 1995; Longitudinal Data conference, Lancaster August 1995; World Congress of the Econometric Society, Tokyo August 1995, LSE; UC Berkeley; Chicago; NBER/NSF Time Series conference, Harvard November 1995 and $(\mathrm{EC})^2$ Conference at Aarhus, December 1995. We thank Robert Kohn, Siem Jan

Koopman, Sid Chib, Aurora Manrique and David Hendry for various suggestions. Finally, all the computations and graphs reported in this program were generated using `C++` code dynamically linked to the `Ox` Matrix Programming Language (Doornik (1995)). We thank Jurgen Doornik for extremely helpful computational advice.

7 Appendix

Proof of theorem 1.

The basis of this will be that, ignoring constants

$$\begin{aligned}\log f(\alpha_t|\alpha_{t-1},\alpha_{t+1},y_t) &= l(\alpha_t)-\tfrac{1}{2}(\alpha_t-\mu_t)'\Sigma_t^{-1}(\alpha_t-\mu_t),\\ &\leq l(m_t)+l'(m_t)'(\alpha_t-m_t)-\tfrac{1}{2}(\alpha_t-\mu_t)'\Sigma_t^{-1}(\alpha_t-\mu_t),\\ &= \log g(\alpha_t,m_t)\end{aligned}$$

by Taylor's mean value theorem $l''(\alpha_t)$ is negative semi-definite for all α_t. Bounding means we can use the standard accept/reject results to prove the validity of the sampler. This gives the proof of the first result.

Recall we have written the densities f and g only up to constants of proportionality. We will express

$$\log g(\alpha_t,m_t)=\log G_t(m_t)-\frac{1}{2}(\alpha_t-a_t)'\Sigma_t^{-1}(\alpha_t-a_t)-\frac{1}{2}\mu_t'\Sigma_t^{-1}\mu_t,\quad \text{where}\quad a_t=\mu_t+\Sigma_t l'(m_t),$$

and

$$\log G_t(m_t)=l(m_t)+l'(m_t)'m_t+\frac{1}{2}m_t'\Sigma_t^{-1}m_t.$$

It is straightforward to show that the probability of accepting is $F_t/G_t(m_t)$, where

$$F_t=\int f(\alpha_t)d\alpha_t \quad \text{and} \quad G_t(m_t)=\int g(\alpha_t,m_t)d\alpha_t.$$

Hence to maximise the acceptance probability we can select m_t to minimise $\log G_t(m_t)$. However,

$$\begin{aligned}\frac{\partial \log G_t(m_t)}{\partial m_t} &= l'(m_t)-l''(m_t)m_t-l'(m_t)+a_t'l''(m_t)\\ &= \{\mu_t+\Sigma_t l'(m_t)-m_t\}'\,l''(m_t).\end{aligned}$$

Equating to zero, this has a solution $\widehat{m_t}$ which satisfies the equation $\widehat{m_t}=\mu_t+\Sigma_t l'(\widehat{m_t})$. However, the solution to this equation is also the mode of f. This proves the second part of the theorem.

7.1 Acceptance frequencies

It is straightforward to make good approximations to the acceptance rates for the accept/reject algorithms used in the single-move Gibbs samplers.

7.1.1 SV model

The acceptance rate, taking the expectation over the suggestions

$$\alpha_t|y_t, \alpha_{t-1}, \alpha_{t+1} \sim \mathsf{N}\left[\mu_t + \frac{\sigma_t^2}{2}\left\{\frac{y_t^2}{\beta^2}\exp\left(-\mu_t\right) - 1\right\}, \sigma_t^2\right],$$

is

$$\begin{aligned} acc_t &= E\exp\left[-\tfrac{y_t^2}{2\beta^2}\left\{\exp(-\alpha_t) - \exp(-\mu_t)\left(1-(\alpha_t-\mu_t)\right)\right\}\right] \\ &\simeq 1 - E\tfrac{y_t^2}{2\beta^2}\left\{\exp(-\alpha_t) - \exp(-\mu_t)\left(1-(\alpha_t-\mu_t)\right)\right\} \\ &= 1 - \tfrac{y_t^2}{2\beta^2}\exp(-\mu_t)\left\{\exp(-r_t+\sigma_t^2/2) - (1-r_t)\right\}, \text{where } r_t = \tfrac{\sigma_t^2}{2}\left\{\tfrac{y_t^2}{\beta^2}\exp\left(-\mu_t\right) - 1\right\} \\ &\simeq 1 - \tfrac{y_t}{4\beta^2}\exp(-\mu_t)\sigma_t^2. \end{aligned}$$

A typical situation is where $\sigma_t^2 = 0.01$. Notice that it is unlikely that $y_t\exp(-\mu_t)/\beta^2$ will be very large as μ_t is a smoothed draw from the log-mean of y_t given the data and so reflects the variation in y_t. An extreme case $y_t\exp(-\mu_t)/\beta^2 = 100$ implies $acc_t \simeq 0.75$. More reasonably when $y_t\exp(-\mu_t)/\beta^2 = 2$, $acc_t \simeq 0.995$, while when the value is one, $acc_t \simeq 0.9975$. In our experience an acceptance rate of over 0.99 seems usual for real financial datasets.

7.1.2 Exponential duration model

The acceptance rate, taking the expectation over the suggestions

$$\mathsf{N}\left[\mu_t + \sigma_t^2\left\{\frac{y_t}{\beta}\exp\left(-\mu_t\right) - 1\right\}, \sigma_t^2\right],$$

is

$$\begin{aligned} acc_t &= E\exp\left[-\tfrac{y_t}{\beta}\left\{\exp(-\alpha_t) - \exp(-\mu_t)\left(1-(\alpha_t-\mu_t)\right)\right\}\right] \\ &\simeq 1 - E\tfrac{y_t}{\beta}\left\{\exp(-\alpha_t) - \exp(-\mu_t)\left(1-(\alpha_t-\mu_t)\right)\right\} \\ &= 1 - \tfrac{y_t}{\beta}\exp(-\mu_t)\left\{\exp(-r_t+\sigma_t^2/2) - (1-r_t)\right\}, \text{where } r_t = \sigma_t^2\left\{\tfrac{y_t}{\beta}\exp\left(-\mu_t\right) - 1\right\} \\ &\simeq 1 - \tfrac{y_t}{2\beta}\exp(-\mu_t)\sigma_t^2. \end{aligned}$$

This has similar acceptance rates to the SV model given above.

7.2 State space models

This Appendix focuses on the computational issues of filtering and smoothing. It provides no direct insight into the models.

7.2.1 Gaussian state space form

The state space form

$$\begin{aligned} y_t &= x_t\beta + Z_t\alpha_t + G_tu_t, \quad u_t \sim \mathsf{NID}(0, I), \\ \alpha_{t+1} &= W_t\beta + T_t\alpha_t + H_tu_t, \\ \alpha_1|Y_0 &\sim N(a_{1|0}, P_{1|0}), \end{aligned}$$

has a prominent role in modern time series (see Harvey (1989) and Hannan and Deistler (1988)). It provides, by appropriate selection of Z_t, G_t, T_t, and H_t, a unified representation of all linear Gaussian time series models. For simplicity we assume that $G_t' H_t = 0$ and we write the non-zero rows of H_t as M_t, $G_t G_t' = \Sigma_{\epsilon t}$ and $H_t H_t' = \Sigma_{\eta t}$. An example of this is an AR(2) with measurement error. Using an obvious notation:

$$Z_t = \begin{pmatrix} 1 & 0 \end{pmatrix}, G_t = \begin{pmatrix} \sigma_\varepsilon & 0 \end{pmatrix}, T_t = \begin{pmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{pmatrix}, H_t = \begin{pmatrix} 0 & \sigma_\eta \\ 0 & 0 \end{pmatrix}, M_t = \begin{pmatrix} 0 & \sigma_\eta \end{pmatrix}.$$

7.2.2 Kalman filter

The Kalman filter (de Jong (1991)) plays a crucial computational role in the analysis of models in state space form. In particular, if we write $a_{t|t-1} = E\alpha_t|Y_{t-1},\beta$ and $P_{t|t-1} = MSE(\alpha_t|Y_{t-1},\beta)$, then the Kalman filter computes these quantities recursively for $t = 1, \ldots, n$,

$$\begin{array}{lllllllll} a_{t+1|t} & = & W_t\beta + T_t a_{t|t-1} + K_t v_t, & P_{t+1|t} & = & T_t P_{t|t-1} L_t' + \Sigma_{\eta t}, & v_t & = & y_t - Z_t a_{t|t-1} - x_t\beta, \\ F_t & = & Z_t P_{t|t-1} Z_t' + \Sigma_{\varepsilon t}, & K_t & = & T_t P_{t|t-1} Z_t' F_t^{-1}, & L_t & = & T_t - K_t Z_t. \end{array}$$

Bi-product of the filter are the innovations v_t, which are the one-step ahead forecast errors, and their corresponding mean square errors, F_t. Together they deliver the likelihood

$$\log f(y_1, ..., y_n) = const - \frac{1}{2}\sum \log|F_t| - \frac{1}{2}\sum v_t' F_t^{-1} v_t$$

for the Gaussian state space form.

7.2.3 Simulation smoother

Traditionally the posterior density of α given Y_n is called the smoothing density. In the Gaussian case there are algorithms which compute the mean and covariance terms of this highly multivariate distribution. We call these algorithms expectation smoothers. More recently a number of algorithms, labelled simulation smoothers, have been proposed for drawing random numbers from $\alpha|Y_n,\beta$. This is useful in MCMC problems. The first algorithms were proposed by Fruhwirth-Schnatter (1994) and Carter and Kohn (1994). We exploit the approach of de Jong and Shephard (1995), which is computationally simpler, more efficient and avoids singularities. This requires that F_t^{-1}, v_t and K_t be stored from the run of the Kalman filter.

Setting $r_n = 0$ and $N_n = 0$, we run for $t = n, ..., 1$

$$\begin{array}{lll} C_t & = & M_t M_t' - M_t H_t' N_t H_t M_t', \\ r_{t-1} & = & Z_t' F_t^{-1} v_t + L_t' r_t - L_t' N_t H_t M_t' C_t^{-1} \kappa_t, \\ \kappa_t & \sim & \mathsf{N}(0, C_t), \\ N_{t-1} & = & Z_t' F_t^{-1} Z_t + L_t' N_t L_t + L_t' N_t H_t M_t' C_t^{-1} M_t H_t' N_t L_t, \end{array} \tag{9}$$

storing the simulated $M_t u_t$ as $\widehat{M_t u_t} = M_t H_t' r_t + \kappa_t$. It will then be convenient to add to $\widehat{M_t u_t}$ the corresponding zero rows so that we simulate from the whole $H_t u_t$ vector (recall H_t is M_t plus some rows of zeros). We will write this as $\widehat{\eta}_t$.

The end condition $\widehat{\eta_0}$ is calculated by

$$\begin{aligned} C_0 &= P_{1|0} - P_{1|0} N_0 P_{1|0}, \quad \kappa_0 \sim \mathsf{N}(0, C_0), \\ \widehat{\eta_0} &= P_{1|0} r_0 + \kappa_0. \end{aligned}$$

The α vector to be simulated via the forward recursion, starting with $\alpha_0 = 0$,

$$\alpha_{t+1} = W_t \beta + T_t \alpha_t + \widehat{\eta}_t, \quad t = 0, ..., n-1. \tag{10}$$

7.2.4 Expectation smoothing

Expectation smoothing is useful in estimating the unobserved stochastic process. In particular the disturbance smoother, due to Koopman (1993), computes $a_{t|n} = E\alpha_t | Y_n, \beta$. It requires that F_t^{-1}, v_t and K_t be stored from the Kalman filter. Then setting $r_n = 0$ it computes backwards:

$$r_{t-1} = Z_t' F_t^{-1} v_t + L_t' r_t, \quad t = n, \ldots, 1.$$

Then it records $\widehat{M_t u_t} = M_t H_t' r_t$ and completes, as above, $\widehat{\eta}_t$ allowing a recursion very similar to (10), but now for the conditional expectations:

$$a_{t+1|n} = W_t \beta + T_t a_{t|n} + \widehat{\eta}_t, \quad t = 1, ..., n.$$

References

Amit, Y. and U. Grenander (1991). Comparing sweep strategies for stochastic relaxation. *Journal of Multivariate Analysis 37*, 197–222.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*, 817–858.

Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis.* New York: Oxford University Press.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory.* Chichester: John Wiley.

Carlin, B. P., A. E. Gelfand, and A. F. M. Smith (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics 41*, 389–405.

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*, 541–553.

Carter, C. K. and R. Kohn (1995). Semiparametric Bayesian inference for time series with mixed spectra. Unpublished working paper: Australian Graduate School of Management, University of New South Wales.

Chan, K. S. and J. Ledolter (1995). Monte Carlo EM estimation for time series models involving counts. *J. Am. Statist. Assoc. 89*, 242–252.

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

Chib, S. and E. Greenberg (1995). Hierarchical analysis of SUR models with extensions to correlated serial errors and time varying-parameter models. *J. Econometrics 68*, 339–360.

Cox, D. R. (1955). Some statistical models connected to a series of events (with discussion). *J. R. Statist. Soc. B 17*, 129–64.

Cox, D. R. (1981). Statistical analysis of time series: some recent developments. *Scandanavian J Statist 8*, 93–115.

Cox, J. C., J. E. Ingersoll, and S. A. Ross (1985). A theory of the term structure of interest rates. *Econometrica 53*, 385–407.

de Jong, P. (1991). The diffuse Kalman filter. *Ann. Statist. 19*, 1073–1083.

de Jong, P. and M. J. MacKinnon (1988). Covariances for smoothed estimates in state space models. *Biometrika 75*, 601–602.

de Jong, P. and N. Shephard (1995). The simulation smoother for time series models. *Biometrika 82*, 339–50.

Diebold, F. X. and M. Nerlove (1989). The dynamics of exchange rate volatility: a multivariate latent factor ARCH model. *J. Appl. Econometrics 4*, 1–21.

Doornik, J. A. (1995). *Ox: Object oriented Matrix Programming, 1.0.* Oxford: Nuffield College, Oxford. Unpublished, `http://hicks.nuff.ox.ac.uk/doornik/doornik.htm`.

Durbin, J. and S. J. Koopman (1992). Filtering, smoothing and estimation for time series models when the observations come from exponential family distributions. Unpublished paper: Department of Statistics, LSE.

Engle, R. F. and J. R. Russell (1994). Forecasting transaction rates: the autoregressive conditional duration model. Unpublished paper: Department of Economics, University of California at San Diego.

Fahrmeir, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalised linear models. *J. Am. Statist. Assoc. 87*, 501–509.

Fruhwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Analysis 15*, 183–202.

Gamerman, D. (1992). A dynamic approach to the statistical analysis of point processes. *Biometrika 79*, 39–50.

Gamerman, D. (1995). Monte carlo markov chains for dynamic generalized linear models. Discussion Paper: Instituto de Matematica, Universidade Federal do Rio de Janerio.

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.

Geweke, J. (1995). Bayesian reduced rank regression in econometrics. *J. Econometrics*. Forthcoming.

Geweke, J. (1996). Posterior simulators in econometrics. In K. F. Wallis (Ed.), *Advances in Econometrics: Seventh World Congress*. Cambridge: Cambridge University Press. Forthcoming.

Goldstein, H. (1995). *Multilevel Models*. London: Edward Arnold.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993). A novel approach to non-linear and non-Gaussian Bayesian state estimation. *IEE-Proceedings F 140*, 107–133.

Green, P. and B. W. Silverman (1994). *Nonparameteric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Green, P. J. and X. L. Han (1990). Metropolis methods, gaussian proposals and antithetic variables. In P. Barone, A. Frigessi, and M. Piccioni (Eds.), *Stochastic Models, Statistical Methods and Algorithms in Image Analysis*, Volume 74, pp. 142–164. Berlin: Springer-Verlag.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist. 5*, 815–41.

Hannan, E. J. and M. Deistler (1988). *The Statistical Theory of Linear Systems*. New York: John Wiley.

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

Harvey, A. C., E. Ruiz, and N. Shephard (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.

Hastings, W. K. (1970). Monte-carlo sampling methods using markov chains and their applications. *Biometrika 57*, 97–109.

Hills, S. E. and A. F. M. Smith (1992). Parameterization issues in bayesian inference. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics, 4*, pp. 641–649. Oxford: Oxford University Press.

Hull, J. and A. White (1987). The pricing of options on assets with stochastic volatilities. *J. Finance 42*, 281–300.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994). Bayesian analysis of stochastic volatility models (with discussion). *J. Business and Economic Statist. 12*, 371–417.

King, M., E. Sentana, and S. Wadhwani (1994). Volatility and links between national stock markets. *Econometrica 62*, 901–933.

Kitagawa, G. (1987). Non-Gaussian statee space modelling of non-stationary time series. *J. Am. Statist. Assoc. 82*, 503–514.

Kohn, R. and C. F. Ansley (1987). A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM J Sci. Statistcal Computing 8*, 33–48.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika 80*, 117–126.

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

London, D. (1985). *Graduation: the Revision of Estimates.* Winsted, Connecticut: ACTEX Publications.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models.* London: Chapman and Hall. 2nd Edition.

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

Muller, P. (1991). Numerical integration in general dynamic models. *Contemporary Mathematics 115*, 145–163.

Pawitan, Y. and F. O'Sullivan (1994). Nonparametric spectral density estimation using penalized Whittle likelihood. *J. Am. Statist. Assoc. 89*, 600–610.

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

Ripley, B. D. and M. D. Kirkland (1990). Iterative simulation methods. *Journal of Computational and Applied Mathematics 31*, 165–172.

Roberts, G. O. and S. K. Sahu (1995). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. Unpublished, Statistical Laboratory, Cambridge University. Available at `http://www.statslab.cam.ac.uk/ mcmc/html`.

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

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist. 12*, 898–916.

Smith, A. F. M. and G. Roberts (1993). Bayesian computations via the Gibbs sampler and related Markov Chain Monte Carlo methods. *J. R. Statist. Soc. B 55*, 3–23.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica 54*, 1461–1481.

Stoker, T. M. S. (1992). *Lectures on Semiparametric Econometrics.* Louvain-la-Neuve, Belgium: Core Lecture Series.

Tierney, L. (1994). Markov Chains for exploring posterior distributions (with discussion). *Ann. Statist. 21*, 1701–1762.

Wahba, G. (1978). Improper priors, spline smoothing, and the problems of guarding against model errors in regression. *J. R. Statist. Soc. B 40*, 364–372.

Wahba, G. (1980). Automatic smoothing of the log periodogram. *J. Am. Statist. Assoc. 75*, 122–132.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates of certain generalized linear models. *Biometrika 63*, 27–32.

West, M., P. J. Harrison, and H. S. Migon (1985). Dynamic generalised models and Bayesian forecasting (with discussion). *J. Am. Statist. Assoc. 80*, 73–97.

Wold, H. (1948). On stationary point processes and Markov chains. *Skand. Aktuar. 31*, 229–240.