Likelihood inference for discretely observed non-linear diffusions

OLA ELERIAN

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

ola.elerian@nuf.ox.ac.uk

SIDDHARTHA CHIB

John M. Olin School of Business, Washington University, St. Louis, MO 63130, USA

chib@simon.wustl.edu

NEIL SHEPHARD

Nuffield College, Oxford, OX1 1NF, UK

neil.shephard@nuf.ox.ac.uk

August 4, 1998

Abstract

This paper is concerned with the Bayesian estimation of non-linear stochastic differential equations when only discrete observations are available. The estimation is carried out using a tuned MCMC method, in particular a blocked Metropolis-Hastings algorithm, by introducing auxiliary points and using the Euler-Maruyama discretisation scheme. Techniques for computing the likelihood function, the marginal likelihood and diagnostic measures (all based on the MCMC output) are presented. Examples using simulated and real data are presented and discussed in detail.

Keywords: Bayes estimation, nonlinear diffusion, Euler-Maruyama approximation, Maximum Likelihood, Markov chain Monte Carlo, Metropolis Hastings algorithm, missing data, Simulation, Stochastic Differential Equation.

1 Introduction

1.1 Models

Consider an Itô stochastic process that satisfies a stochastic differential equation (SDE) of the form

$$dy(t) = a\{y(t), t; \theta\}dt + b\{y(t), t; \theta\}dW(t), \tag{1.1}$$

where $a\{y(t), t; \theta\}$ and $b\{y(t), t; \theta\}$ are the non-anticipative drift and volatility functions, respectively, depending on $y(t)$, time t, and an unknown parameter vector θ, and $dW(t)$ is the increment of a standard Wiener process.[1] SDE's are used extensively in economics: see, for example, the overviews in Dixit (1993) and Merton (1990). Assume that the conditions under which the SDE can be solved for a diffusion $y(t)$ are satisfied (see Øksendal 1995, page 64), and suppose that one has measurements $y_t = y(\tau_t)$ at times $\tau_1, \ldots, \tau_T$, where $\Delta_t^\dagger = \tau_{t+1} - \tau_t$, for $t = 1, \ldots, T$. The aim is to estimate the parameters θ of the process, given the discrete measurements $Y = (y_1, \ldots, y_T)'$.

In the likelihood context, estimation of the parameters is based on the likelihood function $\log L(y_2, \ldots, y_T | y_1; \theta) = \sum_{t=1}^{T-1} \log g(y_{t+1}|y_t; \theta)$, where $g(y_{t+1}|y_t; \theta)$ are the Markovian transition densities. If a strong solution of the underlying SDE process is available, ie., the stochastic differential equation,

$$y(t) = y(0) + \int_0^t a\{y(s), s; \theta\}ds + \int_0^t b\{y(s), s; \theta\}dW(s)$$

can be solved analytically in Itô form, for $t \in (0, T]$, then $g(y_{t+1}|y_t; \theta)$ is available in closed form and likelihood inference is straightforward. The trouble, however, is that analytic solutions of SDE's are rarely available. This has led to growing interest in methods for estimating SDE's on the basis of discrete measurements. Important developments include the indirect inference method (Smith 1993, Broze et al. 1998 and Gourieroux et al. 1993), the efficient method of moments estimator (Gallant and Tauchen 1996 and Gallant and Long 1997), the non-parametric approaches of Aït-Sahalia (1996a, 1996b) and Jiang and Knight (1997) and the likelihood based method of Pedersen (1995).[2]

[1]The same methods developed here can be applied to situations where W(t) is a homogeneous Lévy process, that is, a process with independent increments which is continuous in probability, (see Barndorff-Nielsen, Jensen, and Sørensen 1996).

[2]Discretely observed diffusions have also been fit using estimating functions (see Bibby and Sørensen 1995,

In this paper we propose a new method for dealing with the estimation problem of stochastic differential equations that is likelihood based, can handle non-stationarity and is not dependent on finding an appropriate auxiliary model. Our idea is simply to treat the values of the diffusion between any two discrete measurements as missing data and then to apply tuned Markov chain Monte Carlo methods to learn about the missing data and the parameters. We note that Kim, Shephard, and Chib (1998) have suggested this style of approach in the special case of a stochastic volatility model. This idea has independently also been discussed by Eraker (1998).[3]

1.2 An augmentation based proposal for estimating SDE's

To begin with consider the Euler-Maruyama (or Euler) approximation of the SDE

$$y_{t+1} = y_t + a(y_t, t; \theta)\Delta_t^\dagger + b(y_t, t; \theta)(W_{t+1} - W_t), \tag{1.2}$$

which is the simplest discrete time approximation with respect to the strong convergence criterion [Kloeden and Platen (1992, Section 10.2)]. Under this approximation, the approximate transition distribution is

$$y_{t+1}|y_t, \theta \overset{\text{approx.}}{\sim} N\{y_t + a(y_t, t; \theta)\Delta_t^\dagger, b^2(y_t, t; \theta)\Delta_t^\dagger\}, \tag{1.3}$$

which we can write in density form as $f(y_{t+1}|y_t;\theta)$. Now between any two points $(t, t+1)$, as in Figure 1.1, consider imputing M auxiliary points $y_t^* = (y_{t,1}^*, \ldots, y_{t,M}^*)$ at times $\tau_{t,1} < \ldots < \tau_{t,M}$, where $y_{t,j}^* = y^*(\tau_{t,j})$ for $j = 1, \ldots, M$. For simplicity we assume that the points are evenly spaced, so that $\tau_{t,k+1} - \tau_{t,k} = \frac{\Delta_t^\dagger}{M+1} = \Delta$ for all t, k.[4] Then, the proposal is to approximate the

Kessler and Sørensen 1995, Basawa et al. 1997, Dacunha-Castelle and Florens-Zmirou 1986, Florens-Zmirou 1989, Genon-Catalot and Jacod 1993 and Kessler 1997).

[3]Our work differs from Eraker (1998) in several crucial respects: (i) we develop a tuned MCMC algorithm for carrying out the calculations, rather than a single move method; (ii) we show how to compute the model marginal likelihood and diagnostic checks on the fitted models; and (iii) we include a data-based method for selecting the degree of augmentation to be used. Thus, our techniques provide a complete, efficient and computationally tractable inference engine for stochastic differential equations and various important extensions of it.

[4]$\Delta_t^\dagger$ is introduced for convenience and ease of interpretation. The relationship between θ and $\Delta_t^\dagger$ is determined by the amount of nonlinearity in the drift and volatility functions and the time interval chosen, so that the choice of $\Delta_t^\dagger$ is in fact arbitrary.

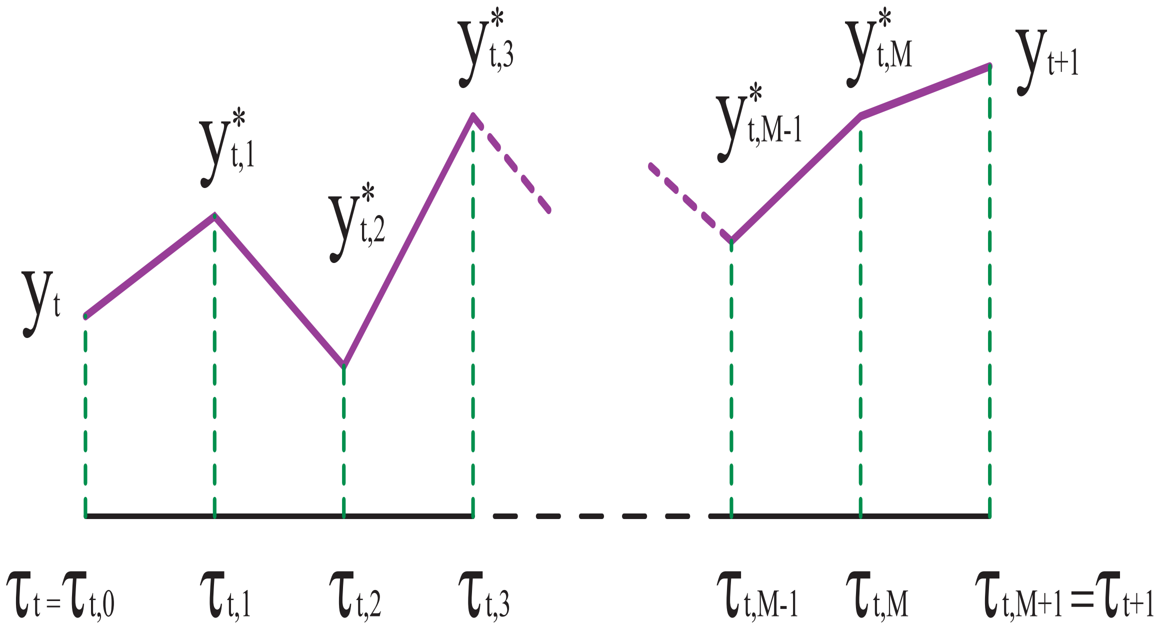

Figure 1.1: *Discretization scheme is shown where the auxiliary points y_t^* are introduced between observed values y_t and y_{t+1}.*

true transition density $g(y_{t+1}|y_t, \theta)$ by

$$\begin{aligned} f^M(y_{t+1}|y_t, \theta) &= \int f(y_{t+1}|y^*_{t,M}, \theta) \left\{ \prod_{j=2}^{M} f(y^*_{t,j}|y^*_{t,j-1}, \theta) \right\} f(y^*_{t,1}|y_t, \theta) dy^*_{t,M}, \ldots, dy^*_{t,1} \\ &= \int f(y_{t+1}|y^*_t, \theta) f(y^*_t|y_t, \theta) dy^*_t \,. \end{aligned} \tag{1.4}$$

As $M \to \infty$, this transition density converges to the true transition density g (Pedersen (1995)). One can, therefore, develop an estimation scheme that utilizes these auxiliary variables to minimize the error of the Euler discretization.[5]

To illustrate the advantages of introducing auxiliary variables, consider the Ornstein-Uhlenbeck (OU) process. In terms of (1.1), $a(y_t, t; \theta) = \mu y_t$, $b^2(y_t, t; \theta) = \sigma^2$ and $\theta = (\mu, \sigma^2)$. The conditional distribution of $y_{t+1}|y_t$ under the Euler scheme is normal with a mean of $\rho_E^\dagger y_t$ and a variance of $\sigma^2 \Delta^\dagger$, where $\rho_E^\dagger = 1 + \mu\Delta^\dagger$. For the strong solution, also normal, the mean of $y_{t+1}|y_t$ is $\rho_S^\dagger y_t$ with variance $\frac{\sigma^2}{2\mu}(\rho_S^{2\dagger} - 1)$, where $\rho_S^\dagger = \exp(\mu\Delta^\dagger)$. Suppose we condition each y_t on its neighbouring points, y_{t-1} and y_{t+1}. Then the distribution of the resulting bridge process

[5]The approach of Gallant and Tauchen (1996) also uses (1.4) as the data generating process. They, however, employ a method of moments approach based on a semi-parametric auxiliary model.

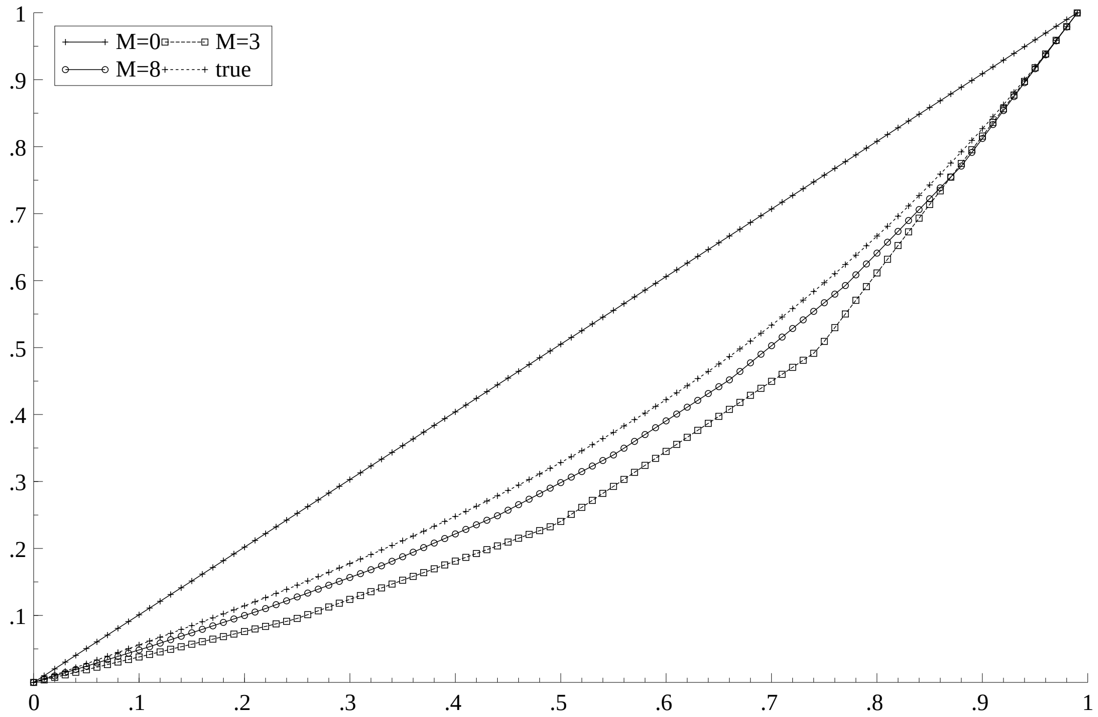

Figure 1.2: *$E(y_t^*|y_1, y_2)$ is graphed for different valies of M under the Euler sheme. The number of latent points, $M = 0, 3$ and 8 are shown for the Euler approximation. The strong solution is denoted by 'true.' The value of μ is taken to be -1 and $\Delta^\dagger = 2$ in all computations, with fixed points $y_1 = 0$ and $y_2 = y_{1,M+1} = 1$ and auxiliary variables $y_1^* = (y_{1,1}^*, \ldots, y_{1,M}^*)$.*

under the Euler scheme can be expressed as

$$y_t|y_{t-1}, y_{t+1}, \theta \sim N\left\{\frac{\rho_E^\dagger}{1+\rho_E^{2\dagger}}(y_{t-1}+y_{t+1}), \frac{\sigma^2\Delta^\dagger}{1+\rho_E^{2\dagger}}\right\}. \tag{1.5}$$

Using the strong solution of the process,

$$y_t|y_{t-1}, y_{t+1}, \theta \sim N\left\{\frac{\rho_S^\dagger}{1+\rho_S^{2\dagger}}(y_{t-1}+y_{t+1}), \frac{\sigma^2\Delta^\dagger}{2\mu}\frac{(\rho_S^{2\dagger}-1)}{1+\rho_S^{2\dagger}}\right\}. \tag{1.6}$$

Similarly, if we consider a block of M latent points $y_t^* = (y_{t,1}^*, \ldots, y_{t,M}^*)$, then we can express $f(y_t^*|y_{t-1}, y_{t+1})$ as a Gaussian distribution with expected value

$$\mathrm{E}(y_t^*|y_{t-1}, y_{t+1}, \theta) = \frac{1}{1+\rho^2+\ldots+\rho^{2M}}\begin{pmatrix} \rho\sum_{i=0}^{M-1}\rho^{2i}y_{t-1} + \rho^M y_{t+1} \\ \rho^2\sum_{i=0}^{M-2}\rho^{2i}y_{t-1} + \rho^{M-1}\sum_{i=0}^{1}\rho^{2i}y_{t+1} \\ \vdots \\ \rho^M y_{t-1} + \rho\sum_{i=0}^{M-1}\rho^{2i}y_{t+1} \end{pmatrix} \tag{1.7}$$

where ρ is $\rho_S = \exp(\mu\Delta)$ under the strong solution and $\rho_E = 1+\mu\Delta$ under the Euler approximation. Given the analytically tractable form of the conditional density for the OU process, we can therefore draw the expected path of the process between two fixed points using the strong solution of the process and illustrate the curvature bias in the paths for different discretisations under the Euler scheme. Figure 1.2 illustrates the curvature bias with $M = 0, 3$ and 8 latent points denoted as $y_1^* = (y_{1,1}^*, \ldots, y_{1,M}^*)$. Note that the case $M = 0$ produces the linear interpolation between the two fixed points, $Y = (y_1, y_2)'$ (where $y_1 = 0$ and $y_2 = y_{1,M+1} = 1$), and completely misses the curvature of the strong solution. Expected paths for non-zero values of M produce a downwards bias, converging to the strong solution of the process, depicted by the second curve from the left. It may be seen that even a small M improves the approximation considerably relative to the $M = 0$ case. Next, we compute the approximate log-likelihoods of $\mu \in [-0.8, 0]$ for the OU process under the Euler scheme using $M = 0, 3$ and 8.[6] These are shown in Figure 1.3 where the strong solution denoted by 'true'. The data generating process (DGP) is defined using $\Delta^\dagger = 4$, $T = 500$, $\mu = -0.5$ and $\sigma^2 = 0.01$. It can be seen that the introduction of the auxiliary points provides a better approximation to the likelihood. Further, we are now in a position to control that approximation by our choice of M.

The rest of the paper is organized as follows. Section 2 presents the simulation of the auxiliary variables using a tuned Metropolis-Hastings (MH) step. A method for sampling the latent values in blocks is discussed and its performance is evaluated. The section also contains a discussion of an alternative single-move sampler which is shown to perform poorly. The MCMC algorithm is completed by a description of the methods required to sample the parameters of the model, conditioned on the augmented data. Section 3 discusses posterior inferences based on the output of the Markov chain. Methods for computing the likelihood function, the marginal likelihood and diagnostic measures are presented. The techniques are illustrated first with simulated data in Section 4 and then in Section 5 with a real data example. The paper concludes with Section 6.

2 MCMC inference for nonlinear diffusions

The estimation method proposed in this paper relies on Markov chain Monte Carlo methods (Tanner and Wong (1987), Geweke (1996), Gelfand and Smith (1990), Chib and Greenberg

[6] Approximate log-likelihoods are computed as the sum of $\log f^M(y_{t+1}|y_t)$, where the one-step ahead transition density is found by analytically integrating out $y_{t,M}^*$ from $f^M(y_{t+1}, y_{t,M}^*|y_t)$ for different values of M; see (1.4).

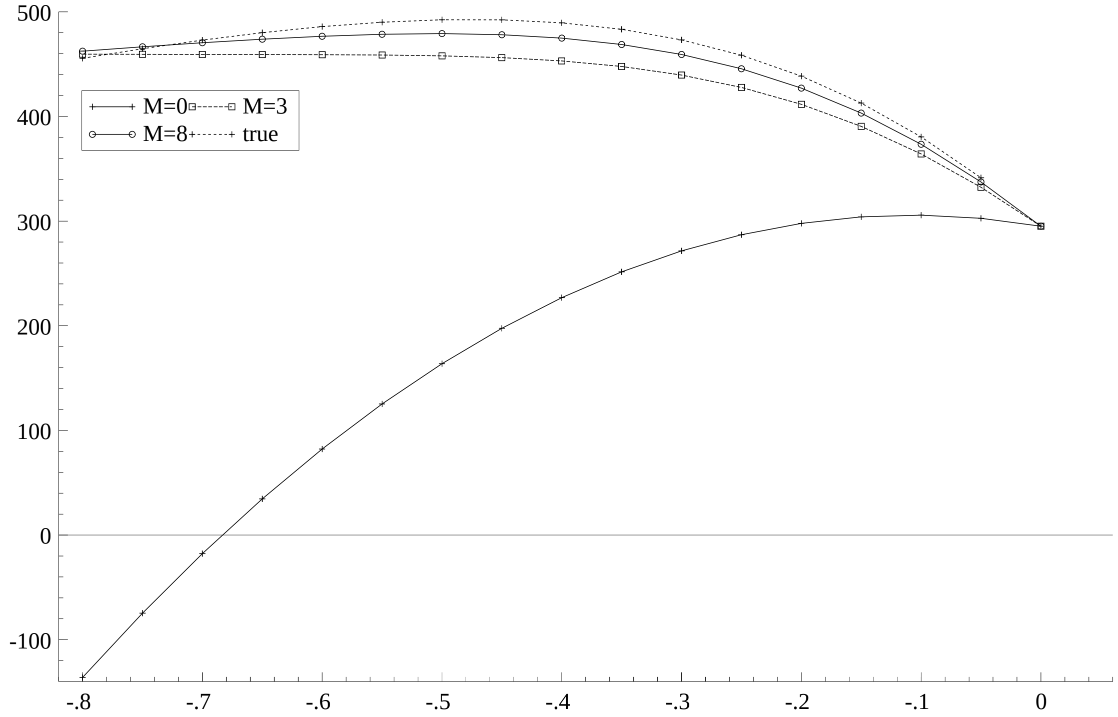

Figure 1.3: *Approximate log-likelihoods are graphed against μ for different values of M under the Euler sheme using 500 observed values. The number of latent points are $M = 0, 3$ and 8. The strong solution is denoted by 'true'. These are evaluated for different values of $\mu \in [-0.8, 0]$. For the DGP, $\mu = -0.5$, $\sigma^2 = 0.01$ and $\Delta^\dagger = 4$.*

(1995) and Chib and Greenberg (1996)). In brief, the aim is to sample the Bayesian posterior density $\pi(\theta, Y^*|Y)$, where $Y = (y_1, \ldots, y_T)'$, by sampling $y_t^*|y_t, y_{t+1}, \theta$, $t = 1, ..., n$, then sampling $\theta|Y, Y^*$ and repeating this process until the chain has converged to its stationary distribution. The key step in this process is the sampling of the latent data $\{y_t^*\}$. Given the dimensionality of the distribution $y_t^*|y_t, y_{t+1}, \theta$ this sampling must be conducted with care in order to produce a Markov chain in which the serial correlations vanish rapidly.

2.1 Simulation of the auxiliary variables from $f(y_t^*|y_t, y_{t+1}, \theta)$

Consider the question of sampling Y^* from $f(Y^*|Y;\theta) = \prod_{t=1}^{T-1} f(y_t^*|y_t, y_{t+1}, \theta)$, where we have used the fact that the y_t^* are conditionally independent, given y_t and y_{t+1}. It therefore suffices to consider the simulation of $y_t^* = (y_{t,1}^*, \ldots, y_{t,M}^*)$ from $f(y_t^*|y_t, y_{t+1}, \theta)$. The target density of interest is

$$f(y_t^*|y_t, y_{t+1}) \propto \prod_{k=0}^{M} f(y_{t,k+1}^*|y_{t,k}^*), \tag{2.8}$$

where $y_t = y^*_{t,0}$ and $y_{t+1} = y^*_{t,M+1}$ and the dependence on θ is suppressed. Each conditional density in this expression can be derived from the discrete-time Euler approximation in equation (1.2). In general, $y^*_{t,k}$ appears nonlinearly in the functions $a(y^*_{t,k})$ and $b(y^*_{t,k})$ and therefore it is not possible to sample $y^*_{t,k}$ directly from the target density. The target density can, however, be sampled by the MH algorithm (see, for example, Chib and Greenberg 1995) with proposal density $q(y^*_t|y_t, y_{t+1})$.

Suppose we update y^*_t in block sizes of m, $(1 \leq m \leq M)$. Starting with observation $y^*_{t,k}$, where $(1 \leq k \leq M)$, the typical block is $y^*_{t(k,m)} = (y^*_{t,k}, y^*_{t,k+1}, \ldots, y^*_{t,k+m-1})$ with neighbours $y^*_{t,k-1}$ and $y^*_{t,k+m}$. Denote this block as $y^*_{t(k,m)} = (w_1, \ldots, w_m) = w$, with $w_0 = y^*_{t,k-1}$ and $w_{m+1} = y^*_{t,k+m}$. The approximate log density for w given its neighbours is then

$$\begin{aligned} l(w) &= \log f(w|w_0, w_{m+1}) \\ &= c + \sum_{j=0}^{m} \log f(w_{j+1}|w_j) = c + \sum_{j=0}^{m} l_j \end{aligned} \tag{2.9}$$

where l_j is given by the Euler approximation $l_j = \log f(w_{j+1}|w_j) = -(1/2)\log\{b^2(w_j\Delta)\} - \{2b^2(w_j)\Delta\}^{-1}[w_{j+1} - \{w_j + a(w_j)\Delta\}]^2$, for $j = 0, \ldots, m$. Details can be found in Appendix A.1. The proposal density to approximate $f(w|w_0, y_{m+1})$ is chosen so that it is easy to sample. To deal with the non-Gaussianity in the likelihood, we propose a second order Taylor expansion of $l(w)$, performed around $\hat{w}$, where the $(j+1)$st element of $\hat{w}$ is given by $\hat{w}_{j+1} = w_0 + \{(j+1)/(m+1)\}(w_{m+1} - w_0)$. A simple form is chosen for the expansion point $\hat{w}$. Details of this expansion for the Euler discretization are given in Appendix A.1. The expansion gives rise to a multivariate Gaussian proposal density

$$q(w) = q(w|w_0, w_{m+1}) \sim N(\tilde{\mu}, \tilde{\Sigma}), \tag{2.10}$$

where the mean $\tilde{\mu}$ and variance matrix $\tilde{\Sigma}$ are defined by equations (A. 1)–(A. 9) of Appendix A.1. Note that the proposal density $q(w|w_0, w_{m+1})$ depends on the neighbouring blocks w_0, w_{m+1} and the parameter θ (whose dependence is not emphasized in the notation). The neighbours w_0 and w_{m+1} are always the most recently sampled values.

To fix ideas about the formulation of the proposal density, consider the SDE in (1.1), and let $a_j = a(y^*_{t,k+j})$ and $g_j = b_j^{-2} = b(y^*_{t,k+j})^{-2}$ specify the diffusion. Four important SDE's are given in Table 1: the Ornstein-Uhlenbeck (OU), Quadratic drift, Cox, Ingersoll and Ross (CIR) (see Cox et al. 1985), and the Hull-White (see Hull and White 1990) processes. For the last two of these processes, we work with the transformation $x_t = \log y_t$ to enforce the restriction that $y_t > 0$. The CIR process, for example, is given by

$$dy(t) = \{\alpha - \beta y(t)\}dt + \sigma\sqrt{y(t)}dW(t)\,,$$

which by Itô's lemma leads to

$$dx(t) = \left\{\frac{\alpha}{\exp x(t)} - \beta - \frac{\sigma^2}{2\exp x(t)}\right\}dt + \frac{\sigma}{\sqrt{\exp x(t)}}dW(t).$$

Model	OU	Quadratic Drift	CIR ($x = \log y$)	Hull-White ($x = \log y$)
a	μy	μy^2	$\frac{\alpha}{\exp(x)} - \beta - \frac{\sigma^2}{2\exp(x)}$	$\frac{\alpha}{\exp(x)} - \beta - \frac{\sigma^2}{2}\exp\{2(\gamma-1)x\}$
b^2	σ^2	σ^2	$\frac{\sigma^2}{\exp(x)}$	$\sigma^2\exp\{2(\gamma-1)x\}$

Table 1: *Drift and volatility specifications for the Ornstein-Uhlenbeck (OU), Quadratic drift process, the Cox, Ingersoll and Ross (CIR) process, and the Hull and White model. For the CIR and Hull-White models, we work with the transformation $x_t = \log y_t$ due to the restriction that $y_t > 0$ for $t = 1, \ldots, T$. The parameters α, β, γ and σ are all positive and constant, whereas μ is negative and constant.*

Table 2 outlines the coefficients needed in the Taylor expansion of the log-likelihood for these four processes using the notation of Appendix A.1. The MH step is now concluded as follows. A proposal value z is drawn from $q(w)$ and is accepted with probability

$$\alpha\{w^{(g)}, z\} = \min\left[1, \frac{f\{z|w_0^{(g+1)}, w_{m+1}^{(g)}\}q\{w^{(g)}\}}{f\{w^{(g)}|w_0^{(g+1)}, w_{m+1}^{(g)}\}q(z)}\right],$$

where $w^{(g)}$ is the current value of w at iteration g of the MCMC sampling. If the proposal value is rejected, the next item in the chain is taken to be $w^{(g)}$.

Remark. The MH algorithm proposed here is referred to as an independence MH sampler (see for example Chib and Greenberg 1995) because the proposal density does not depend on the current value of w. However, because the mean of the proposal depends on the immediate neighbours of w and because the underlying process for the observations is continuous, the mean of the proposal is typically close to w. This has the important consequence that the algorithm proposed here will normally have high acceptance probability and is unlikely to get stuck in the tails of the target distribution.

Improving the sampling

We now focus on the choice of block size m to improve the performance of the sampler given above. Suppose we let $m = M$ and $k = 1$, ie., we update all the auxiliary variables at once. In this case, the sampler may be inefficient since the MH step applied to a high dimensional target distribution often fails to produce a move. One way to generate improved mixing is by randomizing the block size and idea that appears in Shephard and Pitt (1997) in a completely different context. The block to be updated y_t^* is partitioned so that every point is updated within the sweep. We propose using varying random block sizes, selected using a Poisson distribution where λ denotes the average block size. The selection of block sizes to sample as the MCMC method runs through one sweep is carried out using the following algorithm:

1. Initialise λ and M,
2. choose $m_1 \sim 1+$ Poisson$(\lambda - 1)$: if $m_1 \geq M$ set $m_1 = M$ and stop, otherwise set $i = 2$ and
3. choose $m_i \sim 1+$ Poisson$(\lambda - 1)$, $\begin{array}{l} \text{if } \sum_{j=1}^{i} m_j \geq M \text{ set } m_i = M - \sum_{j=1}^{i-1} m_j \text{ and stop,} \\ \text{if } \sum_{j=1}^{i} m_j < M \text{ set } i = i+1 \text{ and repeat.} \end{array}$

Figure 2.4 shows the case when we have an interval of $M = 25$ points between the observed values and we vary the mean λ of the random block size between 2 and 10. The iterations were run over $N = 10,000$ sweeps with $y_t = 1$, $y_{t+1} = 2$ and parameter values $\mu = -0.005$, $\sigma^2 = 0.03$. Though the acceptance rates decline slightly, there is a marked improvement in both the efficiency factors and convergence.[7] Inefficiency factors decrease from 30 to 6 at the middle states indicating that we are only six times as inefficient as if we were to draw i.i.d. samples. This is confirmed by the correlogram, which shows heavy persistence at lags of 60 for the block size of 2 case against a correlogram that dies out by 30 lags for the block size of 10 case.

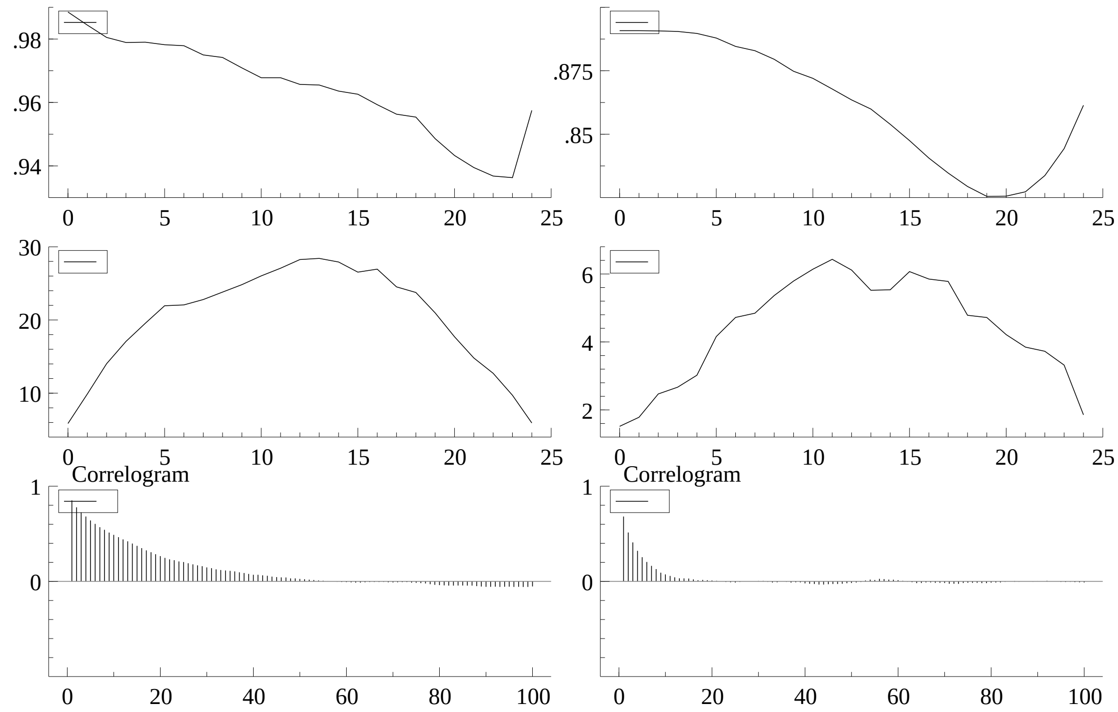

Figure 2.4: *Acceptance rates, inefficiency factors and correlogram for average block sizes of* $\lambda = 2$ *(on the left) and* $\lambda = 10$ *(on the right),* $M = 25$ *and* $N = 10,000$*, with* $y_t = 1$*,* $y_{t+1} = 2$ *and parameter values* $\mu = -0.005$*,* $\sigma^2 = 0.03$*, for the quadratic drift process.*

2.2 Choice of M

An important issue in the implementation of our method is the dimension M of the auxiliary data. In practice it would be convenient to construct a diagnostic check to gauge the adequacy of value selected. We suggest that one compare the fit of the model with a more refined model based on $M_0 = 2M$. We write

$$f(y_{t+1}, y_t^* | y_t, \theta, M_0) = f(y_{t+1} | y_{t,M}^+, \theta) \left\{ \prod_{j=2}^{M} f(y_{t,j}^+ | y_{t,j}^-, \theta) f(y_{t,j}^- | y_{t,j-1}^+, \theta) \right\} f(y_{t,1}^+ | y_{t,1}^-, \theta) f(y_{t,1}^- | y_t, \theta),$$

where $y_t^* = \left(y_t^{-\prime}, y_t^{+\prime}\right)'$. Figure 2.5 shows the discretization scheme using the more refined model. Further, assume that y_t^+ corresponds to the times of the missing data when we use M auxiliary points. Then, the decision on whether M should be increased can be based on the

[7]The inefficiency factor is defined as the ratio of the square of the numerical standard error (computed using a Parzen window) and the variance of the estimate from a hypothetical sampler that produces iid draws (see Geweke 1992).

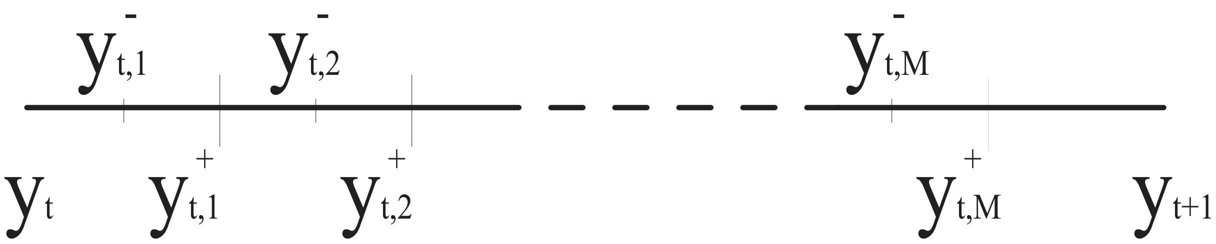

Figure 2.5: *Discretization scheme using the more refined model, where* $y_t^* = \left(y_t^{-\prime}, y_t^{+\prime}\right)'$ *are introduced between observed values* y_t *and* y_{t+1}.

ratio of likelihoods

$$\frac{f_{M_0}(y|\theta)}{f_M(y|\theta)} = \frac{\Pi_{t=1}^T \int f(y_{t+1}, y_t^+, y_t^- | y_t, \theta, M_0) dy_t^+ dy_t^-}{\Pi_{t=1}^T \int f(y_{t+1}, y_t^+ | y_t, \theta, M) dy_t^+}$$

or on the basis of the ratio of marginal likelihoods (the latter being the integral of the likelihood over θ with respect to the prior density of the parameters). The calculation of the marginal likelihood for any given choice of M is illustrated below.

2.3 An alternative method for sampling y_t^*

In this section we consider the MH algorithm which updates $y_t^* = (y_{t,1}^*, \ldots, y_{t,M}^*)'$ one scalar component at a time. Independently of our work Eraker (1997) has used this method to sample the auxiliary variables. We now show that this method generates considerably higher serial correlation in the MCMC output than our preferred method.

As in the previous section consider the target function

$$l(y_{t,k}^*) = \log f(y_{t,k}^* | y_{t,k-1}^*, y_{t,k+1}^*).$$

Equations for the mean and the variance of the proposal distribution are now given by

$$\tilde{\mu} = \tilde{\sigma}^2 u + \hat{y}_{t,k}^* \tag{2.11}$$

and

$$\tilde{\sigma}^2 = \left\{ \frac{1}{\Delta} \left(g_{-1} + g_0 c_0^2 - g_0 d_0 a_0'' \Delta - 2 g_0' d_0 c_0 + \frac{1}{2} g_0'' d_0^2 - \frac{\{g_0 g_0'' - (g_0')^2\}\Delta}{2 g_0^2} \right) \right\}^{-1} \tag{2.12}$$

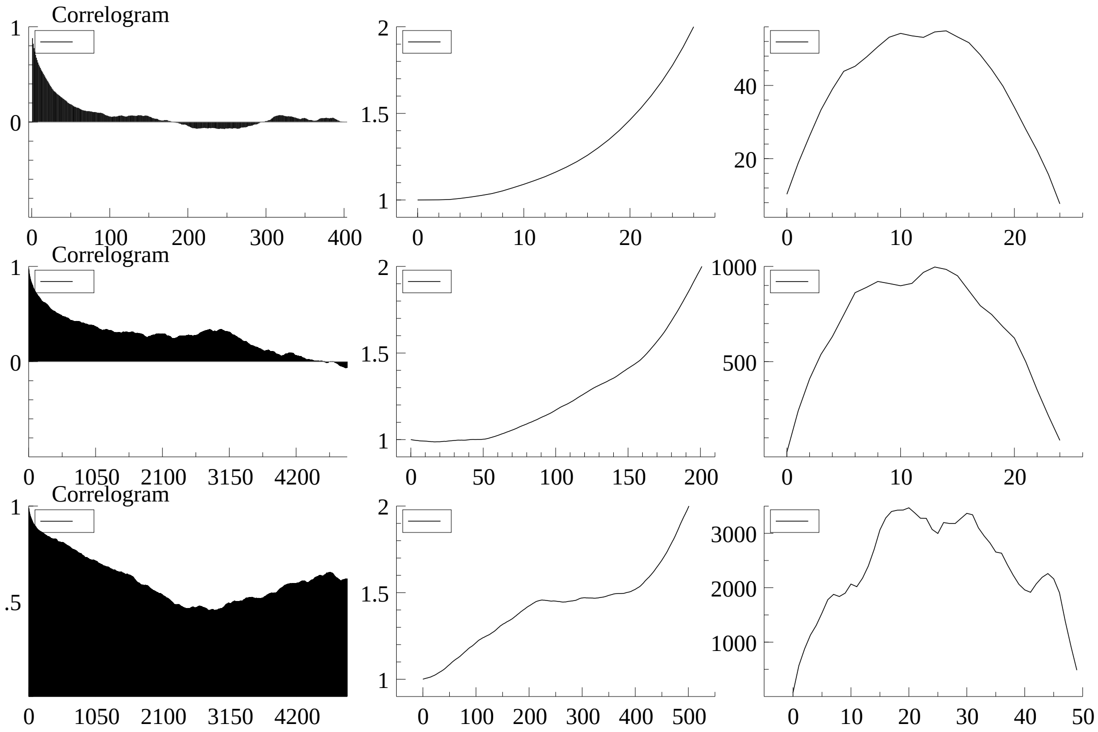

Figure 2.6: *Quadratic drift process: correlograms (left), mean of the simulations (middle) and inefficiency factors (right) for the state $M/2$ for $M = 25$ (top), $N = 10,000$, $M = 200$, (middle) and $M = 500$ (bottom), with $N = 20,000$.*

where

$$u = -\frac{1}{\Delta}\left(g_{-1}d_{-1} - g_0 d_0 c_0 + \frac{1}{2}g_0' d_0^2 - \frac{g_0'\Delta}{2g_0}\right) \tag{2.13}$$

and all quantities are evaluated at $\hat{y}^*_{t,k} = (y^*_{t,k-1} + y^*_{t,k+1})/2$.

The mean of the 20,000 sweeps from the MH algorithm are plotted in Figure 2.6 (middle graphs) for the Quadratic drift process, using $\mu = -0.005$ and $\sigma^2 = 0.03$. One can observe that as the number of points between the observations increases the plots become increasingly less smooth and more iterations are needed for the mean of the simulations to converge. This illustrates the inefficiency in both the Euler scheme and the single-state sampling method. The effect from the initial values is still visible when M is large even after $N = 10,000$ MCMC cycles. From the autocorrelation plots on the left hand side of 2.6, and the value of the inefficiency factors (shown on the right hand side), we conclude that to get reliable estimates for cases larger than $M = 200$ the algorithm would have to be iterated a much larger number of times. For $M = 25$ the single-move MH algorithm is up to 50 times less efficient, while for $M = 200$ it is more than 250 times less efficient than a hypothetical sampler which generates independent samples from the posterior distribution.

Thus, the single-move algorithm suffers from slow convergence, particularly if M is large. This can be explained by the fact that the states $y^*_{t,k}$ are close to each other and hence strongly correlated. Highly correlated components are known to worsen the performance of one-at-a time MCMC samplers, as discussed for example in Liu et al. (1994), Shephard (1994) and Carter and Kohn (1994).

Acceptance rates for sampling the auxiliary variables depend on the choice of M (decreasing slightly as M increases) and the process being considered. In our examples, average acceptance rates varied between 0.99 (OU process) to 0.6 (Hull-White process).

2.4 Sampling of θ from $\pi(\theta|Y, Y^*)$

The next step of the MCMC sampling procedure requires that the parameters θ be sampled from the density $\pi(\theta|Y, Y^*)$. Denote by $z = (y_1, y^*_{1,1}, \dots, y^*_{1,M}, y_2, y^*_{2,1}, \dots, y^*_{T-1,M}, y_T)' = (z_1, \dots z_n)'$ where $n = (T-1)(M+1)+1$ is the length of z. Here both observed states Y and simulated auxiliary states Y^* are treated as the observed data.

Consider the following representation of the conditional likelihood,

$$f(z|y_1;\theta) = f(z_2, \dots z_n|y_1;\theta) = \prod_{t=2}^{n} f(z_t|z_{t-1}, y_1;\theta). \tag{2.14}$$

In general, $f(y_1;\theta)$ is intractable and so we will typically work with the conditional likelihood (2.14) to analyze the posterior distribution of θ, as the effect of omitting the first term is negligible, as when y_t is stationary. Hence, we work with $\prod_{t=2}^{n} f(z_t|z_{t-1};\theta)$, where each $f(z_t|z_{t-1};\theta)$ is a Gaussian density, with mean $z_{t-1} + a(z_{t-1};\theta)\Delta$ and variance $b^2(z_{t-1};\theta)$. Note that these processes are Markov which implies that the prediction decomposition in equation (2.14) is particularly simple.

Given this specification we can estimate the parameters within a Bayesian framework. We know that the Euler discretization of the process yields

$$z_t|z_{t-1}, \theta \sim N\{z_{t-1} + a(z_{t-1})\Delta, b^2(z_{t-1})\Delta\}.$$

where $a(\cdot)$ and $b(\cdot)$ depend on θ. The posterior can be expressed as

$$\pi(\theta|z) \propto \pi(\theta) f(z|y_1;\theta) \tag{2.15}$$

where $\pi(\theta)$ is a prior density for θ. Thus, without loss of generality, we can sample from this posterior inside an MH algorithm by using generic methods such as the Adaptive Rejection Metropolis Sampling (ARMS) procedure (see Gilks, Best, and Tan 1995) sampling one parameter at a time (since this method can only be applied univariately). Of course, in special cases more tailor made algorithms may be better to use. For example, in some models, the posterior (for a suitable choice of prior) will be available in closed form. In other cases one can employ the MH algorithm to sample the conditional posterior. Examples of the models considered earlier are given below.

3 Posterior inferences

3.1 Likelihood evaluation

The approach developed above by-passes the calculation of the likelihood function but nonetheless it is of interest to have a method for evaluating the likelihood function. One way of computing the likelihood function is due to Pedersen (1995). By definition, the approximate likelihood can be expressed in terms of the transition densities

$$f^M(y_{t+1}|y_t,\theta) = \int f(y_{t+1}|y_t^*, y_t, \theta) f(y_t^*|y_t,\theta) dy_t^*$$

Pedersen suggests that this transition density be estimated by importance sampling as

$$\frac{1}{J}\sum_{j=1}^{J} \frac{f(y_{t+1}, y_t^{*(j)}|y_t,\theta)}{f(y_t^{*(j)}|y_t,\theta)} = \frac{1}{J}\sum_{j=1}^{J} f(y_{t+1}|y_{t,M_t}^{*(j)},\theta)$$

where $y_t^{*(j)}$ is drawn from the importance function $f(y_t^*|y_t,\theta)$. This estimate, although easily available, can be poorly behaved if M is large since the importance function conditions only on the first observation.

An alternative method for computing the likelihood contribution, one that minimizes the Monte Carlo variance of the estimate, is based on the importance sampling function $f(y_t^*|y_t, y_{t+1},\theta)$ since then

$$\frac{f(y_{t+1}, y_t^*|y_t,\theta)}{f(y_t^*|y_t, y_{t+1},\theta)} = f^M(y_{t+1}|y_t,\theta)$$

and $f^M(y_{t+1}|y_t,\theta)$ can be computed using one simulated value of y_t^*.

3.2 Marginal likelihood

We now very briefly consider the question of comparing alternative (potentially non-nested) diffusion models that have been fit to a given data set. One useful technique for making this comparison is through the calculation of the model marginal likelihood which is defined as the integral of the likelihood function with respect to the prior density. This quantity can be computed from the MCMC output by a method developed in Chib (1995). Ratios of marginal likelihood give rise to the Bayes factor which represents the evidence in the data in favor of the numerator model, relative to the model in the denominator.

Let θ denote the parameters of a given diffusion model, with likelihood function $f(y|\theta)$ and prior density $\pi(\theta)$, where the likelihood function is computed using one of the two methods discussed above. The dependence of the parameters θ, the likelihood and the prior on the model is not emphasized in the notation for convenience. Then, the marginal likelihood (with all normalizing constants included) can be written as

$$m(y) = \frac{f(y|\theta)\pi(\theta)}{\pi(\theta|y)},$$

an expression that follows from Bayes theorem. The key point is that this expression is an identity in θ and can therefore be evaluated at any appropriately selected point θ^* (say). If θ^* denotes a high density point and $\hat{\pi}(\theta^*|y)$ the estimate of the posterior ordinate at θ^*, then the marginal likelihood on the log scale is estimated as

$$\ln m(y) = \ln f(y|\theta^*) + \ln \pi(\theta^*) - \hat{\pi}(\theta^*|y) \tag{3.16}$$

where the first two terms are typically available directly and the third is estimated from the MCMC output by either kernel smoothing (if the dimension of θ is small) or by a marginal/conditional decomposition of the posterior ordinate followed by reduced MCMC runs to generate the draws necessary to estimate each of the marginal/conditional ordinates (see Chib 1995 for further details). It should be noted that this approach for estimating the marginal likelihood only requires a single evaluation of the likelihood function and the posterior ordinate and is thus computationally inexpensive.

3.3 Diagnostic checks on fitted model

Conventional diagnostic checks on the fitted model usually expand the model under consideration and then perform a test to see if the expansion yields a statistically significant improvement in fit. Although this can be carried out in our framework via the use of Bayes factors or a likelihood ratio statistic, it is also helpful to see if the predictions of the model accord with the data. We will carry this out via one-step-ahead prediction distribution functions. Throughout we will assume that the SDE is univariate, although extensions to the multivariate case are possible.

Our focus will be on

$$G(y_{t+1}|y_t) = \Pr(Y_{t+1} \leq y_{t+1}|y_t),$$

which will be $UID(0,1)$ if the model is correct. We approximate this by $F^M\left(y_{t+1}|y_t\right)$, where

$$\begin{aligned} F^M\left(y_{t+1}|y_t\right) &= \int I(Y_{t+1} < y_{t+1}) f^M(Y_{t+1}|y_t) dY_{t+1} \\ &= \int I(Y_{t+1} < y_{t+1}) f^M(Y_{t+1}, y_t^*|y_t) dY_{t+1} dy_t^* \end{aligned}$$

and then estimate this quantity by drawing $Y_{t+1} \sim f^M(Y_{t+1}|y_t)$ a large number of times and counting the proportion of times the simulated value is below the observed data:

$$u_{t+1} = \widehat{F}^M\left(y_{t+1}|y_t\right) = \frac{1}{R}\sum_{j=1}^{R} I(Y_{t+1}^j < y_{t+1}).$$

Adequacy of the model is judged by the serial correlation in the $\{u_{t+1}\}$ and by the nature of the distributional shape of both the $\{u_{t+1}\}$ and its reflected version $\left\{2\left|u_{t+1} - \frac{1}{2}\right|\right\}$, the latter providing information on dispersion. These statistics are also used in Rosenblatt (1952), Pedersen (1994b), Dawid (1982), Smith (1985) and Shephard (1994). The idea of focusing on $2\left|u_t - 0.5\right|$ appears in Kim, Shephard, and Chib (1998). As a by-product, we compute the standardized forecast errors (see Pedersen 1994a), as

$$\frac{y_{t+1} - \widehat{\mathrm{E}}(y_{t+1}|\mathcal{F}_t)}{\sqrt{\widehat{\mathrm{var}}(y_{t+1}|\mathcal{F}_t)}},$$

where $\widehat{\mathrm{E}}(y_{t+1}|\mathcal{F}_t) = \frac{1}{R}\sum_{j=1}^{R} Y_{t+1}^j$ and $\widehat{\mathrm{var}}(y_{t+1}|\mathcal{F}_t) = \frac{1}{R-1}\sum_{j=1}^{R}\{Y_{t+1}^j - \widehat{\mathrm{E}}(y_{t+1}|\mathcal{F}_t)\}^2$. Properties of these statistics are harder to evaluate then for the u_{t+1}, but provide a graphical aid to help explain the inadequacies of the fitted model.

4 Simulated data example

In this section we apply the methods developed above to the CIR process

$$dy(t) = \{\alpha - \beta y(t)\}dt + \sigma\sqrt{y(t)}dW(t)\,,$$

and defer the details of similar calculations for the OU, Quadratic drift and Hull-White models to Appendix A.2–A.4. With the transformation $x(t) = \log y(t)$ and Itô's lemma, the model under the Euler approximation is given by

$$x_{t+1}|x_t, \alpha, \beta, \sigma^2 \sim N\left\{\left(\frac{\alpha}{\exp x_t} - \beta - \frac{\sigma^2}{2\exp x_t}\right)\Delta + x_t, \frac{\sigma^2}{\exp x_t}\Delta\right\}.$$

The data for our illustration is simulated from the above model with $\alpha = 0.5$, $\beta = 0.2$ and $\sigma^2 = 0.05$. We specify a design to represent typical weekly and daily financial data sets. Using an initial value of $y_0 = 1$, two sets of $T = 500$ observations are obtained using the Euler approximation with $\Delta = \Delta^\dagger/R$ where $\Delta^\dagger = 1$ for daily series and $\Delta^\dagger = 5$ for weekly data and $R = 100,000$. Here $\Delta^\dagger$ can be thought of as the time interval in the observed data and determines the bias in the discretized time gap.

4.1 Full conditional density of the drift parameters

Let $\psi = (\alpha', \beta')'$ and suppose that ψ is given a bivariate normal prior distribution with mean γ_0 and variance Γ_0^{-1}. Then a simple calculation shows that the full conditional density given the augmented data z is

$$\psi|z, \sigma^2 \sim N(V^{-1}u, V^{-1})$$

where

$$u = \left(\frac{X'v}{\sigma^2} + \Gamma_0\gamma_0\right), \; V = \left(\frac{X'X\Delta}{\sigma^2} + \Gamma_0\right)$$

and

$$v = \begin{bmatrix} \sqrt{z_1}\left(\log z_2 + \frac{\sigma^2\Delta}{2z_1} - \log z_1\right) \\ \vdots \\ \sqrt{z_{n-1}}\left(\log z_n + \frac{\sigma^2\Delta}{2z_{n-1}} - \log z_{n-1}\right) \end{bmatrix}, \; X = \begin{pmatrix} z_1^{-\frac{1}{2}} & -z_1^{\frac{1}{2}} \\ \vdots & \vdots \\ z_{n-1}^{-\frac{1}{2}} & -z_{n-1}^{\frac{1}{2}} \end{pmatrix}.$$

4.2 Full conditional density of the volatility

The posterior of the volatility coefficient is no longer conjugate when the log of the process is considered. We propose taking a first-order Taylor expansion of the non-conjugate part to upperbound $\log f(\sigma^2|z,\alpha,\beta)$ and hence obtain the posterior using an Accept/Reject algorithm.

Suppose we take a prior of Inverse Gamma $(\frac{p}{2}, \frac{1}{2\,S_0\,p})$ for σ^2, where $S_0 = c\,\Delta$, p and c are constants, (typically 10 and 0.001 respectively); the log-full conditional density is

$$\log f(\sigma^2|z;\alpha,\beta) = c' - \xi_1(\log\sigma^2) - \xi_2\sigma^{-2} - S_1\sigma^2$$

where $\xi_1 = \left(\frac{n+p}{2}\right)$, $\xi_2 = \frac{1}{2\Delta}(\sum_{t=2}^n z_{t-1}d_{t-1}^2 + S_0\,p\,\Delta)$, $S_1 = \frac{\Delta}{8}\sum_{t=2}^n z_{t-1}^{-1}$ and $d_{t-1} = \log z_t - \log z_{t-1} - \frac{\alpha\Delta}{z_{t-1}} + \beta\Delta$. Suppose we let $\lambda = \log\sigma^2$, then

$$\begin{aligned}\log f(\lambda|z;\alpha,\,\beta) &= \log f(\sigma^2|z;\alpha,\,\beta) + \log\left|\frac{\partial\sigma^2}{\partial\lambda}\right| \\ &= c' - (\xi_1 - 1)\lambda - \xi_2\exp(-\lambda) + p(\lambda) \qquad (4.17)\end{aligned}$$

$\xi_2 > 0$, and $p(\lambda)$ is concave in λ, that is $p''(\lambda)$ is negative for all λ. Then we can sample from $\lambda|z,\alpha,\,\beta$ by making suggestions

$$exp(\lambda) \sim \text{ Inverse Gamma}\{\xi_1 - p'(\hat\lambda) - 1, \xi_2\} \qquad (4.18)$$

where $\xi_1 - p'(\hat\lambda) > 1$, $\xi_2 > 0$ and $\hat\lambda$ is an arbitrary fixed value of λ. These proposed values are accepted with probability $\exp\{p(\lambda) - p(\hat\lambda) - p'(\hat\lambda)(\lambda - \hat\lambda)\}$.

4.3 MCMC output analysis

We now apply our efficient MCMC sampling procedure to the first simulated data set which corresponds to the case $\Delta^\dagger = 1$. The MCMC algorithm is run for $N = 10,000$ cycles with $\Delta^\dagger = 1$. In Figure 4.7 we show the sample path, autocorrelation function and histogram based on the MCMC output of σ^2. The output for α and β displays similar features. In the top graphs $M = 0$, in the middle graphs $M = 1$, and in the bottom graphs $M = 10$. These graphs demonstrate the clear advantage of using auxiliary variables in the estimation procedure. Bias decreases quickly and the autocorrelation in the sampler is low.

Summaries of the MCMC output for the second data set (which corresponds to the case $\Delta^\dagger = 5$) are reported in Table 3. The MCMC algorithm is now run with $M = 0, 1, 10$ ($\lambda = 3$

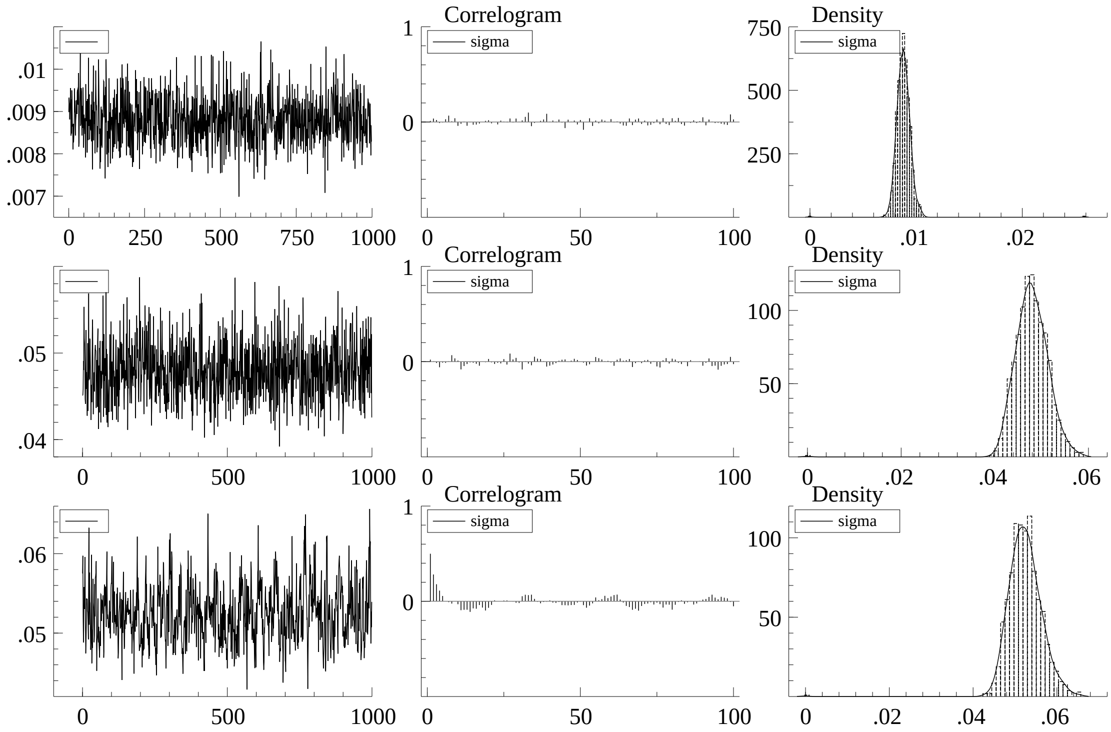

Figure 4.7: *Paths, correlograms and histograms for draws from $\sigma^2|y$ (sigma), (true value 0.05). Top graph uses discretisation $M = 0$, middle graphs have $M = 1$ and bottom graphs have $M = 10$, $\lambda = 3$. $N = 10,000$ and $\Delta^\dagger = 1$ in all simulations.*

), 20 ($\lambda = 5$) and 30 ($\lambda = 9$) for $N = 10,000$ iterations. The simulation inefficiency factors are very small for $M < 20$. When $M = 20$, the MCMC sampler is approximately 14 times less efficient than a hypothetical algorithm that produces independent draws. The inefficiency factor increases to 20 when $M = 30$. The autocorrelation plots in Figure 4.7 reinforce this point, dying down more quickly for smaller values of M. There is strong correlation between α and β (0.99) regardless of M, while correlation between σ^2 and α, β initially increases with M, fluctuating around 0.75 and 0.8 for higher values of M.

4.3.1 Monte Carlo experiment

Table 4 shows the results from a small simulation study for the parameters of the CIR process. Ten data sets of length $T = 500$, $R = 100,000$ and $\Delta^\dagger = 5$ were generated, and the MH algorithm was run with $M = 0, 1, 10, 20$ and 30 for $N = 10,000$ iterations. In general, $N = 10,000$ since the simulation inefficiency hovers around 20. On the basis of the Monte Carlo results reported in table 4, we see that as M increases there is a marked decrease in bias, though the difference in results for $M = 10, 20$ and $M = 30$ is negligible. Hence, the initial increase in M gives more precise estimates of the parameters, but there are diminishing returns after certain values, (for example $M > 20$). There is a corresponding increase in the standard deviations as M increases as the estimation is carried out over a larger number of latent states, though these seem to level out for higher values of M. The MSE for σ^2 is smallest for $M = 20$. Computer time to run 100 iterations of the sampler for the case $M = 3$, $T = 500$ is 258.60 seconds on a Pentium/133, using the matrix language Ox by Doornik (1996).

5 Real data example

We now consider the analysis of a diffusion model of the type outlined in Aït-Sahalia (1996b) and apply this (and related models) to the 7-day Eurodollar deposit spot rate (measured as the mid-point of the bid-ask rates).[8] The model presented in 5.19 is specified by the diffusion functions

$$\begin{aligned} a(y;\theta) &= \alpha_0 + \alpha_1 y + \alpha_2 y^2 + \alpha_3 y^{-1} \\ b^2(y;\theta) &= \beta_0 + \beta_1 y + \beta_2 y^{\beta_3} \end{aligned} \tag{5.19}$$

where $\theta = (\alpha_0, \alpha_1, \alpha_3, \beta_0, \beta_1, \beta_2, \beta_3)'$ and

$\beta_0 \geq 0$ (and $\beta_2 > 0$ if $\beta_0 = 0$ and $0 < \beta_3 < 1$, or $\beta_1 > 0$ if $\beta_0 = 0$ and $\beta_3 > 1$)

$\beta_2 > 0$ if either $\beta_3 > 1$ or $\beta_1 = 0$, and $\beta_1 > 0$ if either $0 < \beta_3 < 1$ or $\beta_2 = 0$

$\alpha_2 \leq 0$ and $\alpha_1 < 0$ if $\alpha_2 = 0$

$\alpha_3 > 0$ and $2\alpha_3 \geq \beta_0 \geq 0$, or $\alpha_3 = 0, \alpha_0 > 0$, $\beta_0 = 0$, $\beta_3 > 1$ and $2\alpha_0 \geq \beta_1 > 0$.

The first two sets of constraints are necessary for b^2 to be positive, the third ensures that the drift is mean reverting for large values of y while the last ensures that the values are positive. Note that instead of letting β_3 be restricted to the region $(0, 1) \cup (1, \infty)$, we could let $\beta_3 \in (1, \infty)$ given the characteristics of the data at hand (the sample data contains $T = 5505$ observations with a mean of 0.083621 and variance of 0.0012893. The observations range from 0.029150 to 0.24333). The other model specifications are outlined in Table 5. One reason we consider these alternative model specifications is because the above model is practically unidentified for these data.

[8] The data can be downloaded from http://gsbwww.uchicago.edu/fac/yacine.aitsahalia.

Our first set of results are for the Vasicek, CIR and Affine CIR models, where we have fit the model after taking logs and applying the Itô transformation. The results are given in Tables 6, 7 and 8 where we report the posterior means, Monte Carlo standard errors, inefficiency factors, and the covariances and correlations of the parameters. The *Highest Probability Density* (HPD) regions are also reported (for example, see Besag, Green, Higdon, and Mengersen 1995). These results are based on $M = 10$ ($\lambda = 3$) using $N = 2000$ MCMC iterations for the Vasicek, CIR and $N = 5000$ for the Affine CIR models. For each of these models, the output paths, correlograms and histograms of $\theta = (\alpha_0, \alpha_1, \beta_1)'$ are given in Figures 5.8 , 5.9, and 5.10. The bandwidth $B_N = 100$ (400) is used in the Parzen computations of the Monte Carlo standard errors for α (β). We set standard diffuse normal priors on the alpha coefficients and diffuse inverse Gamma priors on β. In the Affine CIR and general parametric model, however, β_1 is given a diffuse normal prior centered at zero.

It will be seen that for each model the corrleogram of α dissipate quickly, though that of β_1 in the Affine CIR model shows persistence up to lag 100. In all models the posterior mean of α_0 is positive and that of α_1 is negative. The results on β_0 in the Vasicek model and those of β_1 in the CIR model are consistent with those in the Affine CIR model. The Figures show that β_0 is close to zero and that the variance is mainly determined by $\beta_1 y$. The α and β parameters are close to being uncorrelated. One point to note is that the inefficiency factors are low for α. For β , we are up to 25 times less efficient.

In Figure 5.11 we plot the standardized forecast errors, defined in section 3.3, for all models. A plot of the autocorrelation functions is provided in Figure 5.12 with the corresponding QQ plots. The residuals and their reflected versions are computed using the one-step ahead prediction distribution.

We now turn our attention to the general parametric model given by equation 5.19. We first show that β_2 and β_3 are difficult to identify separately from our data. In Figure 5.13 we show a plot of $\beta_2 y^{\beta_3}$ against a range of y values determined from the data. The function is plotted for five pairs of values of $(\beta_2, \beta_3) : (4.511 \times 10^{-4}, 1.5)$, $(6.9338 \times 10^{-4}, 1.7)$, $(7.021 \times 10^{-4}, 1.8)$, $(1.312 \times 10^{-3}, 2)$, and $(1.5177 \times 10^{-3}, 2.1)$. These were obtained from the MCMC algorithm fixing a value of β_3 and evaluating the posterior mean of β_2. In the bottom graph we plot the kernel density estimate of the y data. We see that for the range of y values with mass (essentially values of y below 0.16), the five different pairs of (β_2, β_3) values produce virtually

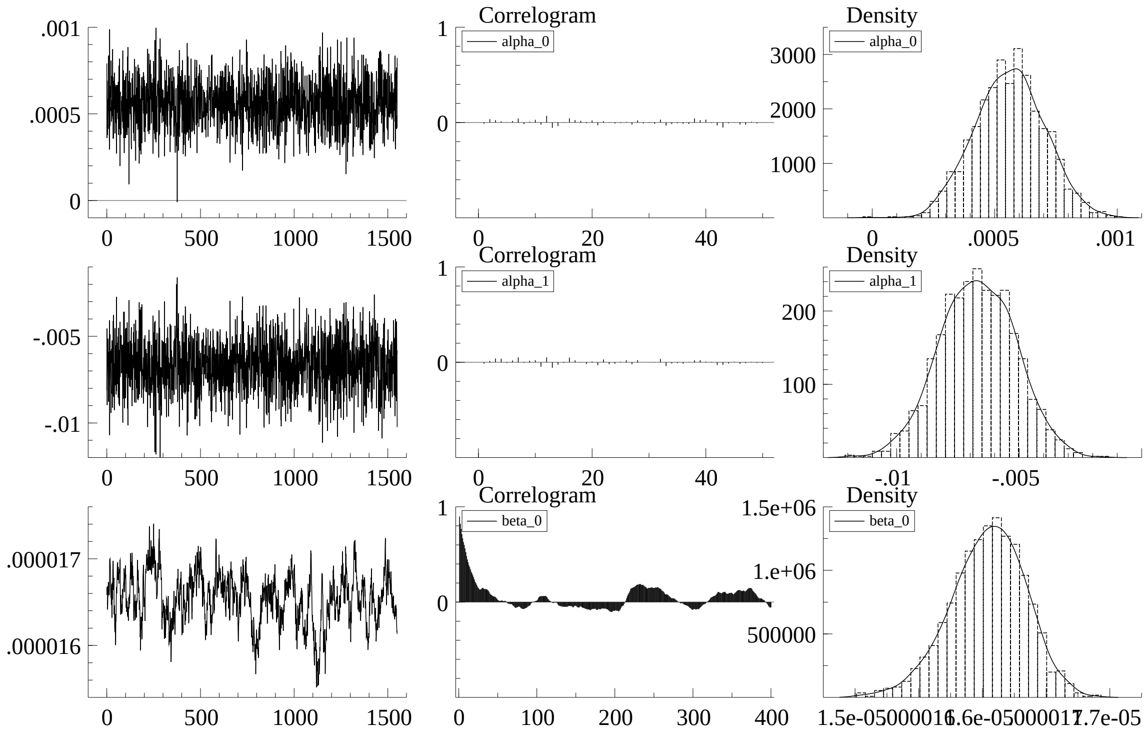

Figure 5.8: *Paths, correlograms and histograms for α_0, α_1 and β_0 are shown using the Vasicek process applied to the Eurodollar short-rate. The MH algorithm was run with $M = 10$ ($\lambda = 3$), for $N = 10,000$ iterations. The bandwidth used in the Parzen computations, $B_N = 100$ for α_0 and α_1, and $B_N = 400$ for β_0.*

the same value of $\beta_2 y^{\beta_3}$. On the basis of this analysis (see also Tauchen 1997) we estimate an alternative model in which the volatility specification is cubic, preserving the quadratic shape exhibited above

$$\beta_0 + \beta_1 y + \beta_2 y^2 + \beta_3 y^3,$$

as suggested by Figure 5.13. Results are shown in table 9. Similar diagnostics were obtained on setting $\beta_3 = 0$. Corresponding posterior means, correlograms and histograms are shown in Figures 5.14 and 5.15. Standardized forecast errors in Figure 5.11 show that the outliers are not being picked up, and the autocorrelation functions in Figure 5.12 confirm that the models have not provided a good fit to the data.

To complete the analysis we also compute the marginal likelihoods of these various models. The log marginal likelihood computations are shown in table 10. We infer that there is almost equal support in the data for the CIR and the affine CIR models but that the general parametric model receives overwhelming support in relation to the other three models. Nonetheless, volatility clustering in the residuals suggests that each of these models should be elaborated

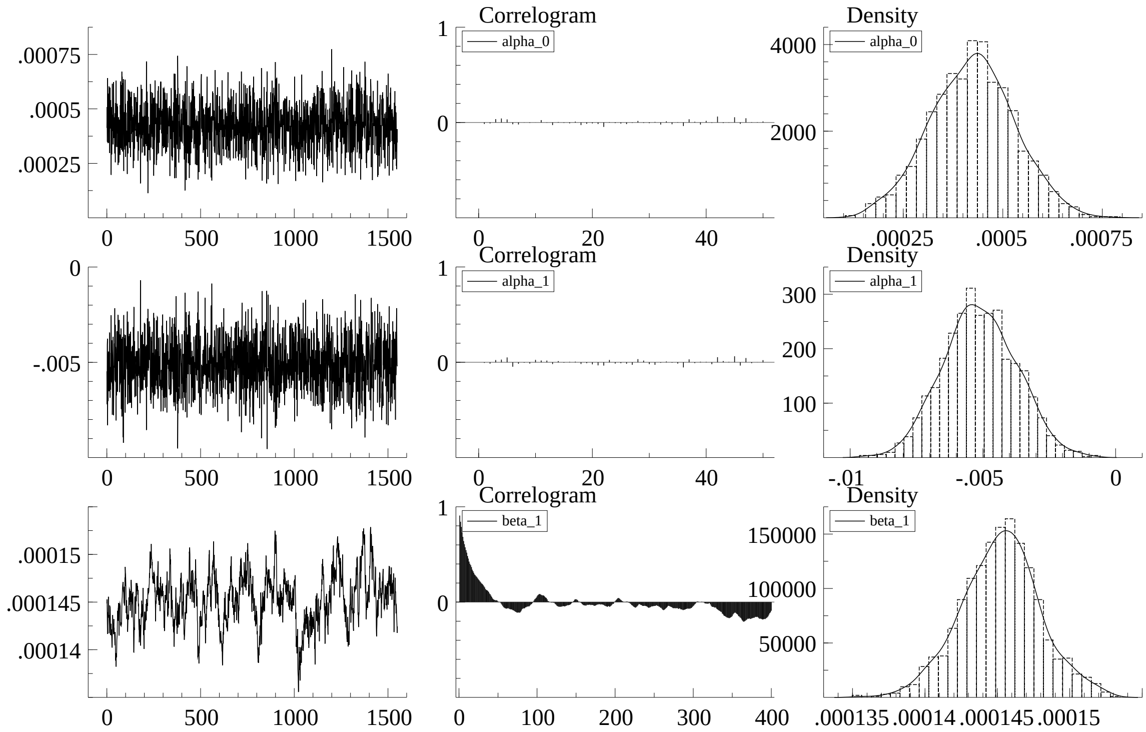

Figure 5.9: *Paths, correlograms and histograms for* α_0, α_1, *and* β_1 *are shown using the CIR process applied to the Eurodollar short-rate data. The MH algorithm was run with* $M = 10$ *(*$\lambda = 3$*), for* $N = 1500$ *iterations. The bandwidth used in the computations,* $B_N = 100$ *for* α_0 *and* α_1 *and* $B_N = 400$ *for* β_1.

to include a heavy tailed stochastic volatility component, perhaps along the lines of (Andersen and Lund 1997).

6 Concluding remarks and further topics

In this paper we have provided a likelihood approach to the analysis of discretely observed diffusions. Our approach is based around the introduction of auxiliary observations which are then integrated out of the likelihood function. We have proposed efficient ways of performing Bayesian calculations on these problems and provided a broad toolkit for conducting inferences in these models that includes methods for finding the model marginal likelihood (to compare alternative stochastic differential equations) and methods for computing model fit measures, both based on the MCMC output.

This paper differs from much of the recent econometric literature on the estimation of diffusions. In comparison with the EMM/indirect inference literature, no auxiliary model to

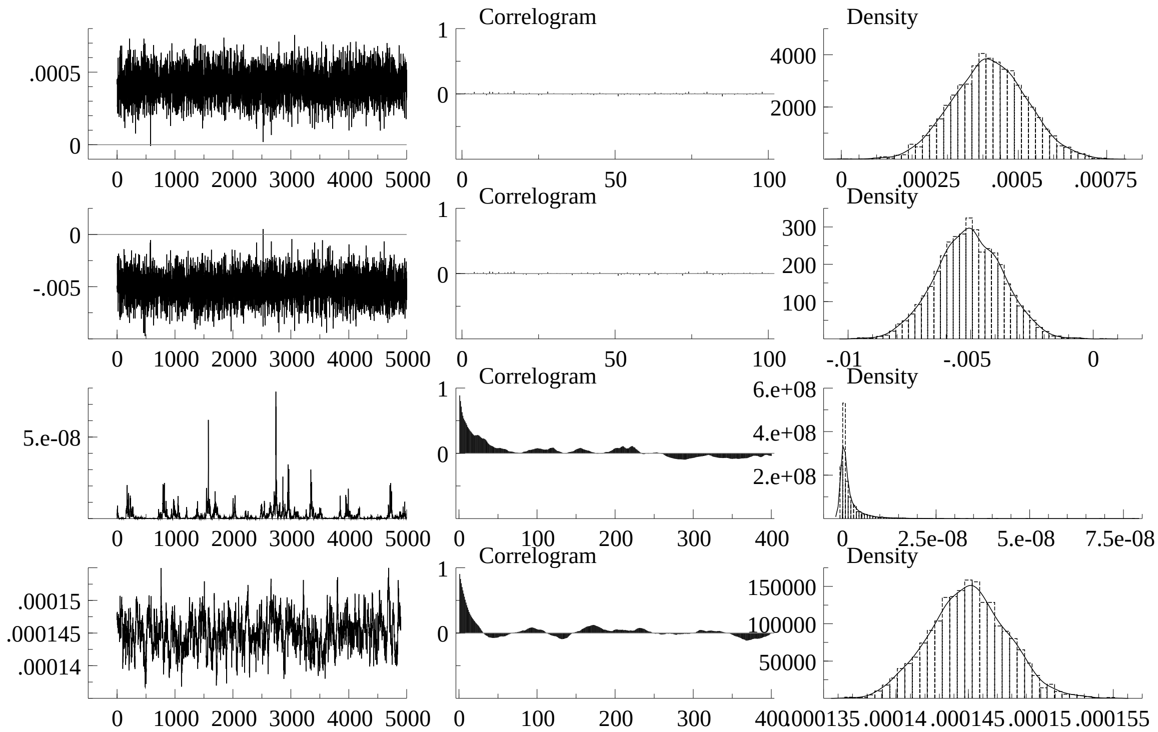

Figure 5.10: *Paths, correlograms and histograms for α_0, α_1, β_0 and β_1 are shown using the Affine CIR process applied to the Eurodollar short-rate data. The MH algorithm was run with $M = 10$ ($\lambda = 3$), for $N = 5000$ iterations. The bandwidth used in the Parzen computations, $B_N = 100$ for α_0 and α_1 and $B_N = 400$ for β_0 and β_1.*

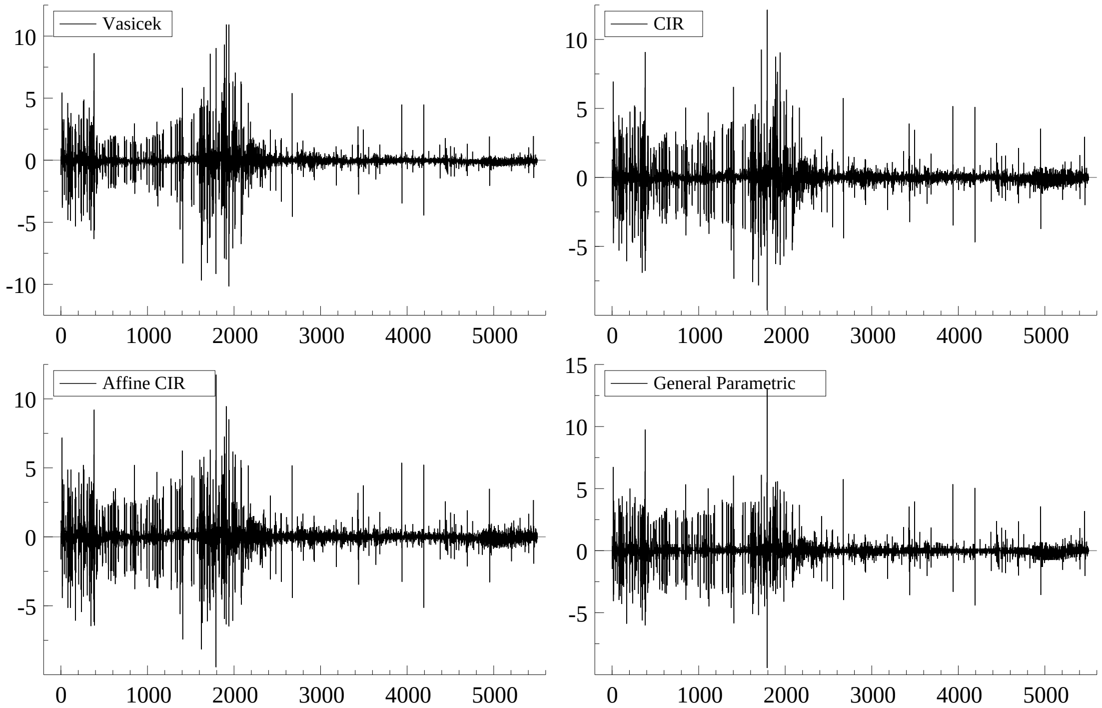

Figure 5.11: *Standardized forecast errors for the Vasicek, CIR, Affine CIR and general parametric models.*

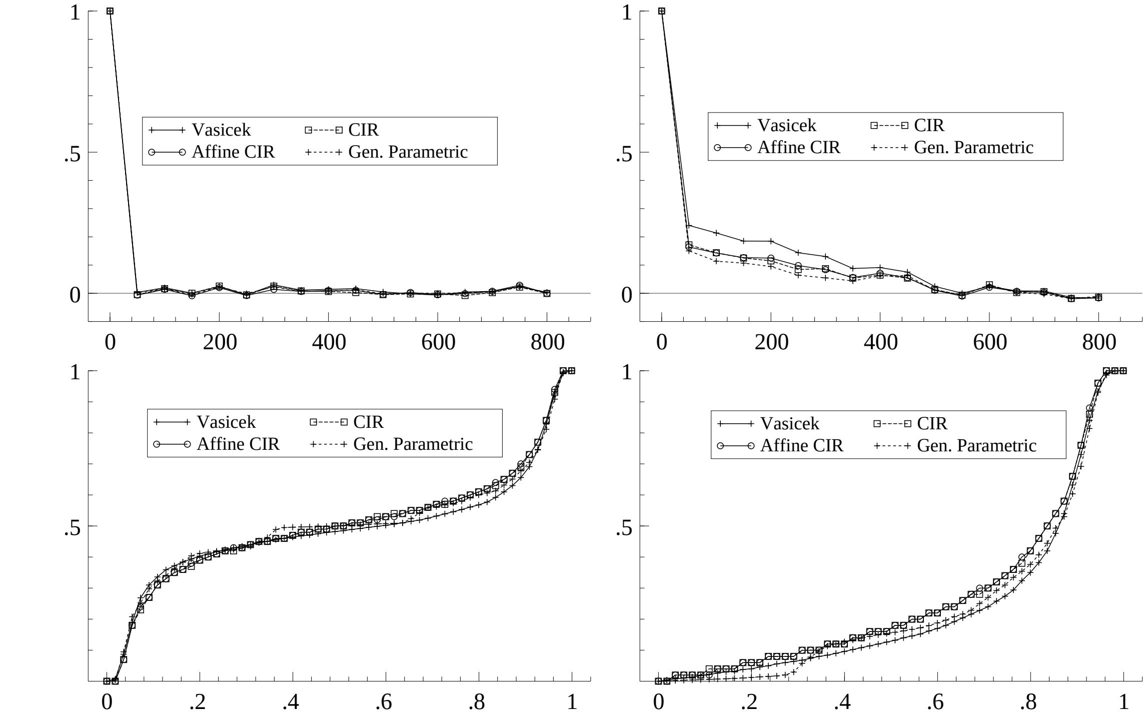

Figure 5.12: *Top graphs: Autocorrelation functions of residuals (left panel) and reflected residuals (right panel) for the Vasicek, CIR, Affine CIR and general parametric process. Bottom graphs: The corresponding QQ plots against observations.*

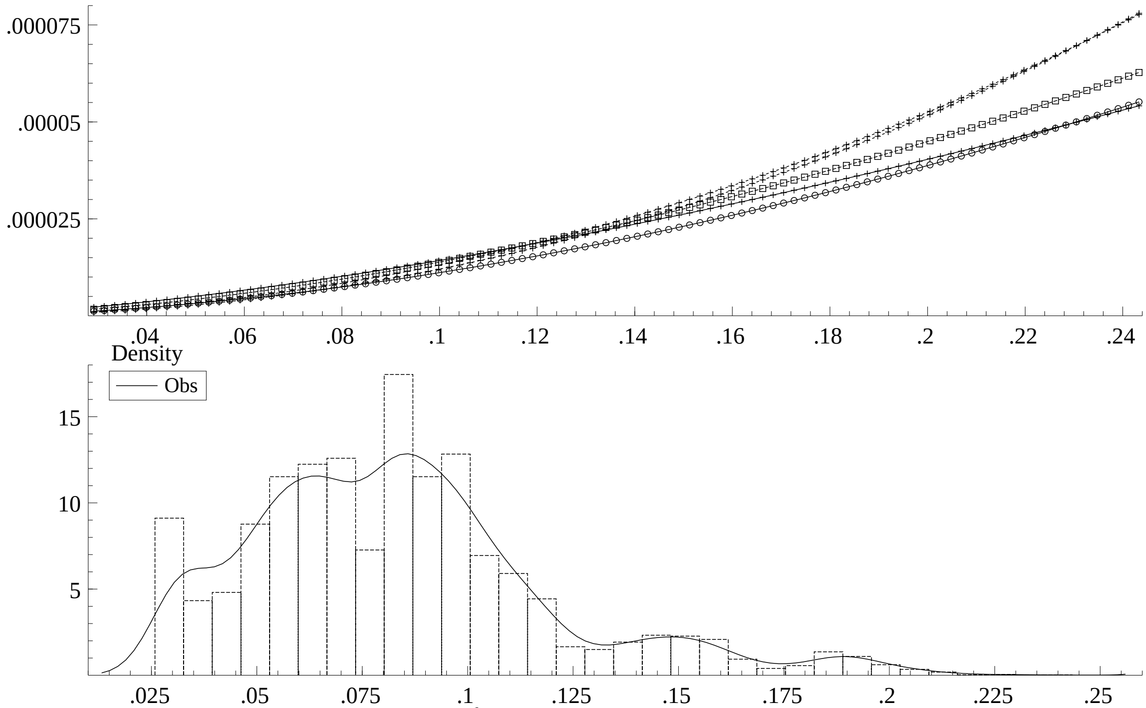

Figure 5.13: *Top graph plots $\beta_2 y^{\beta_3}$ against a range of values for y between the minimum and maximum values observed from the data, for five pairs of values of (β_2, β_3). Bottom graph plots the kernel estimate of the density using the interest data.*

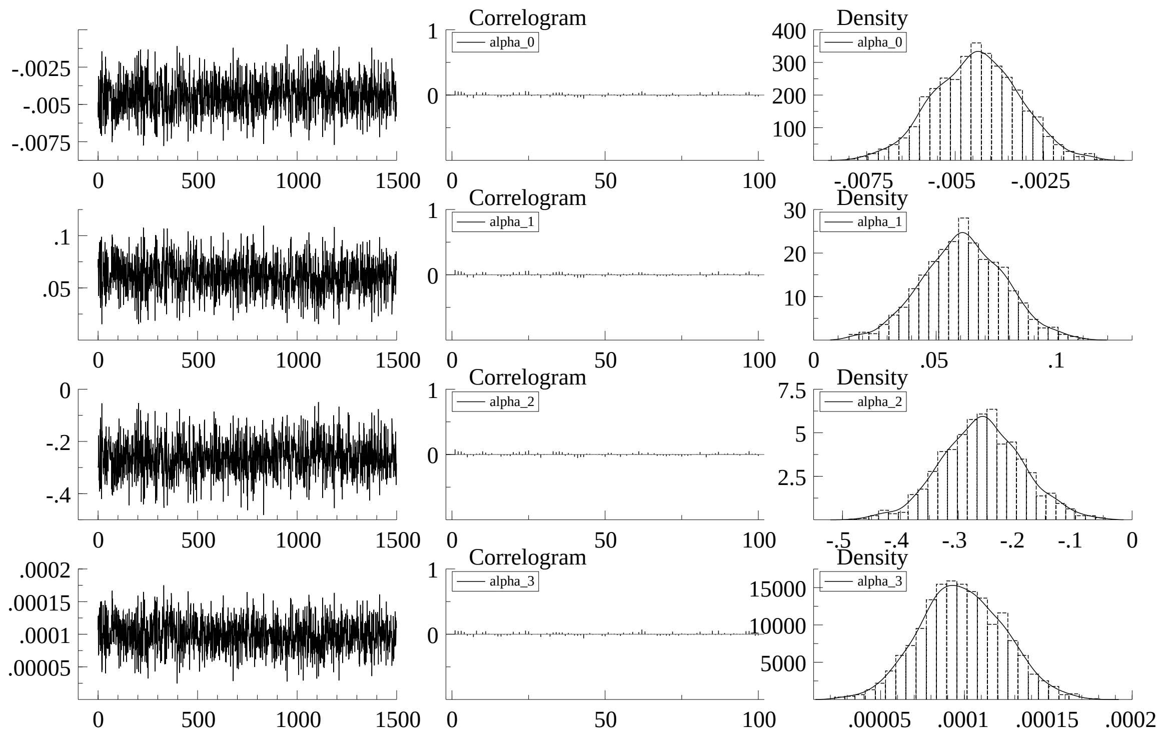

Figure 5.14: *Paths, correlograms and histograms for α_0, α_1, α_2 and α_3 are shown using the Eurodollar short-rate data applied to the general parametric process. The MH algorithm was run with $M = 10$ ($\lambda = 3$), for $N = 1500$ iterations. The bandwidth used in the computations,* $B_N = 100$.

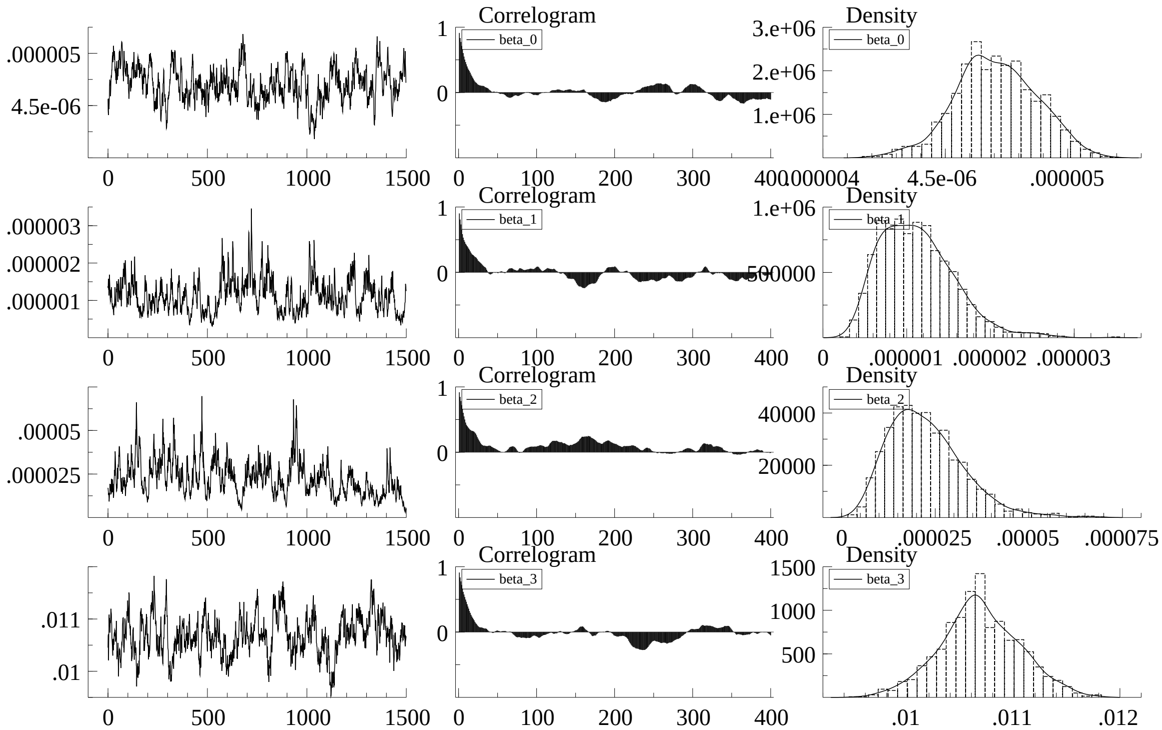

Figure 5.15: *Paths, correlograms and histograms for* β_0*,* β_1*,* β_2 *and* β_3 *are shown using the Eurodollar short-rate data applied to the general parametric process. The MH algorithm was run with* $M = 10$ *(*$\lambda = 3$*), for* $N = 1500$ *iterations. The bandwidth used in the computations,* $B_N = 400$.

sample the latent data is required. This simplifies the estimation and makes possible the corresponding analysis of multivariate stochastic differential equation models where finding good auxiliary models is known to be difficult. Although we have used a prior in our analysis — for large samples this information is dominated by the likelihood function and so our conclusions are insensitive to this aspect of the analysis. Indeed in such situations the Bayesian analysis becomes very close to that of maximum likelihood inheriting the efficiency of that method. However, unlike maximum likelihood, the Bayesian method developed here is computationally quite tractable.

One important characteristic of our approach is that it can be extended to deal with partially observed diffusions (ie., diffusions containing an unobserved state variable such as a stochastic volatility component), multivariate observations and non-stationary data. Further work on these problems has been initiated. Finally, the approach can be easily modified to include the more sophisticated Milstein approximation as the basis of the discretization scheme (Elerian (1998)). Comparable results based on this modification will be reported elsewhere.

7 Acknowledgments

Ola Elerian would like to thank Nuffield College for their financial support. Neil Shephard would like to thank the ESRC for their financial support through the project 'Estimation via simulation in Econometrics' and the EU through their grant 'Econometric inference using Simulation Techniques'. All the computations reported in this paper were carried out using the Ox language of Doornik (1996).

A Appendix

A.1 Obtaining the proposal for block sampling

Consider a block of size m $(1 \leq m \leq M)$ starting with observation $y^*_{t,k}$, $(1 \leq k \leq M)$. So we have a block $y^*_{t(k,m)} = (y^*_{t,k}, y^*_{t,k+1}, \ldots, y^*_{t,k+m-1}) = (w_1, \ldots, w_m) = w$ with neighbours $y^*_{t,k-1} = w_0$ and $y^*_{t,k+m} = w_{m+1}$. To deal with the non-Gaussian element of the likelihood, we propose to perform a second order Taylor expansion of $l(w)$ around a fixed value of w, denoted by $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_m)$. This gives

$$l(w) \approx c + x'u - \frac{1}{2}x'Vx = c' - \frac{1}{2}(x - V^{-1}u)'V(x - V^{-1}u)$$

where $x = (w - \hat{w})$ and

$$\begin{aligned} u &= \frac{\partial l(\hat{w})}{\partial w} = \{u_j\} \quad \text{for } j = 1, \ldots m \\ V &= -\frac{\partial^2 l(\hat{w})}{\partial w \partial w'} = \{V_{ij}\} \quad \text{for } i, j = 1, \ldots, m. \end{aligned}$$

This shows that $l(w)$ is approximately the log of a multivariate Gaussian density with mean $\tilde{\mu} = \hat{w} + V^{-1}u$ and variance matrix $\tilde{\Sigma} = V^{-1}$. To obtain expressions for these parameters, define

$$a_j = a(y^*_{t,k+j}) = a(w_{j+1}) \quad \text{(A. 1)}$$

$$g_j = \{b^2(y^*_{t,k+j})\}^{-1} = \{b^2(w_{j+1})\}^{-1} \quad \text{(A. 2)}$$

$$d_j = y^*_{t,k+j+1} - (y^*_{t,k+j} + a_j\Delta) = w_{j+2} - (w_{j+1} + a_j\Delta) \quad \text{(A. 3)}$$

$$c_j = 1 + a'_j\Delta \quad \text{(A. 4)}$$

for $j = -1, 0, \dots, m-1$. Let a'_j and a''_j denote the first and second derivatives of $a(w_{j+1})$ with respect to w_{j+1}, and define g'_j and g''_j similarly. Consider a given element, w_{j+1} ($j = 0, \dots, m-1$) of the block. The only terms in $l(w)$ involving w_{j+1} are

$$\begin{aligned} l_j &= \frac{1}{2}\log\left(\frac{g_{j-1}}{\Delta}\right) - \frac{g_{j-1}}{2\Delta}[w_{j+1} - (w_j + a_{j-1}\Delta)]^2 \\ l_{j+1} &= \frac{1}{2}\log\left(\frac{g_j}{\Delta}\right) - \frac{g_j}{2\Delta}[w_{j+2} - (w_{j+1} + a_j\Delta)]^2 \end{aligned}$$

which implies that

$$\frac{\partial l(w)}{\partial w_{j+1}} = \frac{\partial l_j}{\partial w_{j+1}} + \frac{\partial l_{j+1}}{\partial w_{j+1}} = -\frac{1}{\Delta}\left(g_{j-1}d_{j-1} - g_j d_j c_j + \frac{1}{2}g'_j d_j^2 - \frac{\Delta g'_j}{2g_j}\right).$$

For the second derivatives,

$$\frac{\partial^2 l}{\partial w_{j+1}^2} = -\frac{1}{\Delta}\left[g_{j-1} + g_j c_j^2 - g_j d_j a''_j\Delta - 2g'_j d_j c_j + \frac{1}{2}g''_j d_j^2 - \frac{\{g_j g''_j - (g'_j)^2\}\Delta}{2g_j^2}\right].$$

The only term involving w_{j+1} and w_{j+2} is l_{j+1}, implying that

$$\frac{\partial^2 l(w)}{\partial w_{j+1}\partial w_{j+2}} = \frac{\partial^2 l_{j+1}}{\partial w_{j+1}\partial w_{j+2}} = -\frac{1}{\Delta}(g'_j d_j - g_j c_j)$$

In summary, the proposal density $q(w|w_0, w_{m+1})$ obtained from the second-order Taylor expansion of $l(w)$ is Gaussian with mean

$$\tilde{\mu} = \hat{w} + \tilde{\Sigma}u \tag{A. 5}$$

and variance matrix

$$\tilde{\Sigma} = V^{-1} \tag{A. 6}$$

where u is an $m \times 1$ vector with elements

$$u_{j+1} = -\frac{1}{\Delta}\left(g_{j-1}d_{j-1} - g_j d_j c_j + \frac{1}{2}g'_j d_j^2 - \frac{g'_j\Delta}{2g_j}\right) \tag{A. 7}$$

for $j = 0, \dots, m-1$, and V is an $m \times m$ matrix with elements

$$V_{j+1,j+1} = \frac{1}{\Delta}\left[g_{j-1} + g_j c_j^2 - g_j d_j a''_j\Delta - 2g'_j d_j c_j + \frac{1}{2}g''_j d_j^2 - \frac{\{g_j g''_j - (g'_j)^2\}\Delta}{2g_j^2}\right] \tag{A. 8}$$

$$V_{j+1,j+2} = V_{j+2,j+1} = \frac{1}{\Delta}(-g_j c_j + g'_j d_j) \tag{A. 9}$$

and $V_{j+1,j+s} = 0$ for all j when $s > 2$. All quantities in (A. 7)–(A. 9) are evaluated at $w = \hat{w}$.

A.2 The posterior for the OU process

Assuming a conjugate prior of $\chi_p^{-2} S_0 p$ for σ^2, for the OU process, where typically $p = 10$ and $S_0 = c\Delta$, and c is a constant (typically 0.001) that does not depend on p, we achieve the posterior density

$$\sigma^2|\mu, z \sim \chi_{n+p}^{-2} \Delta^{-1} \left[\sum_{t=2}^{n} \{z_t - (1+\mu\Delta) z_{t-1}\}^2 - z_1^2(2\mu\Delta + \mu^2\Delta^2) + S_0 p \right] \tag{A. 10}$$

where μ is the most recently updated value, over the iterations. Suppose we wish to ensure stationarity on μ. This can be acheived by employing a beta prior on $(1+\phi)/2$, where $\phi = 1+\mu\Delta$ with $\mathrm{E}(\mu) = \frac{1}{\Delta}\left(\frac{2\delta}{\delta+\gamma} - 2\right)$. We typically use values $\delta = 20$ and $\gamma = 1.5$ giving 0.86 and 0.11 for the mean and variance of ϕ. Then the prior of μ is proportional to

$$f\left(\mu|\sigma^2\right) \sim \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} \left(\frac{\mu\Delta + 2}{2}\right)^{\gamma-1} \left(\frac{-\mu\Delta}{2}\right)^{\delta-1} \frac{2}{\Delta}.$$

So

$$\begin{aligned}
\log f(\mu|\sigma^2, z) &\propto \log f(z|\mu, \sigma^2) + \log f(\mu|\sigma^2) \\
&= \text{constant } - \frac{1}{2\sigma^2\Delta} \sum_{t=2}^{n} \{z_t - (1+\mu\Delta) z_{t-1}\}^2 + \frac{z_1^2}{2\sigma^2\Delta}(2\mu\Delta + \mu^2\Delta^2) \\
&\quad - \frac{1}{2}\log(2\mu\Delta + \mu^2\Delta^2) + (\gamma-1)\log(\mu\Delta+2) + (\delta-1)\log(-\mu\Delta) \\
&= \text{constant } + l(\mu) + p(\mu),
\end{aligned}$$

where $p(z_1; \mu, \sigma^2)$ has unconditional distribution $N(0, \frac{\sigma^2}{1-\phi^2})$. The first term is Gaussian but the second term, due to the choice of the prior is non-Gaussian. Suppose we define $s_1^2 = \sum_{t=2}^{n} z_{t-1}^2$ and $s_{0,1} = \sum_{t=2}^{n} z_t z_{t-1}$, then the first term is in fact

$$l(\mu) = \text{constant } - \frac{\Delta s_1^2}{2\sigma^2} \left\{\mu - \left(\frac{s_{0,1}}{s_1^2} - 1\right)\right\}^2$$

which is normal with mean $\{(s_{0,1}/s_1^2) - 1\}$ and variance $\{\sigma^2/(\Delta s_1^2)\}$. The posterior is non-conjugate and so a first order approximation of $p(\mu)$ is taken and an Accept/Reject algorithm is performed to obtain an estimate for μ. Applying a first order Taylor approximation to $p(\mu)$ we obtain

$$\begin{aligned}
p(\mu) &\approx p(\hat{\mu}) + (\mu - \hat{\mu}) p'(\hat{\mu}) \\
&= (\gamma-1)\log(\hat{\mu}\Delta+2) + (\delta-1)\log(-\hat{\mu}\Delta) + (\mu-\hat{\mu})\left\{\frac{(\gamma-1)\Delta}{\hat{\mu}\Delta+2} + \frac{\Delta(\delta-1)}{\hat{\mu}\Delta}\right\}.
\end{aligned}$$

So we can bound the true posterior density from above by the following density

$$\log g(\mu|\sigma^2, z) \quad = \quad c - \frac{\Delta}{2\sigma^2 s_1^2}\left[s_1^2\mu - \left\{s_{0,1} - s_1^2 + \frac{\sigma^2(\gamma-1)}{\hat{\mu}\Delta+2} + \frac{\sigma^2(\delta-1)}{\hat{\mu}\Delta}\right\}\right]^2. \quad \text{(A. 11)}$$

Values of μ are drawn from (A. 11), and we can apply the result on acceptance rates outlined in Shephard and Pitt (1997), section 2, which states that the acceptance probability for making proposals for the parameters is accepted with probability $\exp\{p(\mu) - p(\hat{\mu}) - p'(\hat{\mu})(\mu - \hat{\mu})\}$. For the Accept/Reject algorithm, (see Rubinstein (1981) and Ripley, pg 60, (1987)), if a value is rejected, another candidate value is proposed using (A. 11) until acceptance, (in the stationary region), producing variates from $\log f(\mu|\sigma^2, z)$. As we have resticted values of μ to lie in the region of stationarity, the proposals may often be rejected. To overcome this, one could employ the probability integral transform theorem proposed by Feller (1971) which enables sampling from a truncated distribution, (denoted by $TN(\cdot)$), without rejection. A proposals μ^* of μ is drawn from $TN_{[-(2/\Delta),0]}\{\tilde{\mu}, \text{var}(\tilde{\mu})\}$, where $\tilde{\mu} = \frac{s_{0,1}}{s_1^2} - 1 + \frac{\sigma^2(\gamma-1)}{s_1^2(\hat{\mu}\Delta+2)} + \frac{\sigma^2(\delta-1)}{s_1^2\hat{\mu}\Delta}$ and $\text{var}\tilde{\mu} = \frac{\Delta s_1^2}{\sigma^2}$.

A.3 The posterior for the Quadratic drift process

For the Quadratic drift process, on choosing conjugate priors for the drift and volatility, the posterior for σ^2 is given by

$$\sigma^2|z;\mu \sim \chi^{-2}_{n+p}\Delta^{-1}\left\{\sum_{t=2}^{n}(z_t - z_{t-1} - \mu z_{t-1}^2\Delta)^2 + S_0 p\right\}. \quad \text{(A. 12)}$$

The posterior for $\mu|z$ can be expressed as $f(\mu|z) \propto \exp\left\{-\frac{1}{2\tau_1^2}(\mu - \mu_1)^2\right\}$, where on defining $s_{0,1}^2 = \sum_{t=2}^{n} z_t z_{t-1}^2$ and $s_1^r = \sum_{t=2}^{n} z_{t-1}^r$,

$$\mu_1 = \frac{\frac{s_{0,1}^2}{\sigma^2} - \frac{s_1^3}{\sigma^2} - \frac{\mu_0}{\tau_0^2}}{s_1^4\Delta}\sigma^2 + \frac{1}{\tau_0^2}, \; \tau_1^2 = \left(\frac{s_1^4\Delta}{\sigma^2} + \frac{1}{\tau_0^2}\right)^{-1},$$

where the prior is given by $\mu \sim N(\mu_0, \tau_0^2)$.

A.4 The posterior for the Hull-White model

For the Hull-White process, estimation of the parameters is analogous to the CIR case except that it involves an additional parameter γ. Suppose we put a non-informative prior on γ, a discrete uniform distribution $U(0, 2)$, with I support points. For computational simplicity, we consider only a discrete set of possible values for γ. To obtain the full conditional posterior of γ, we use the following

$$P(\gamma=\gamma_i|z;\alpha,\beta,\sigma^2)=\frac{f(z|\gamma=\gamma_i,\alpha,\beta,\sigma^2)\,P(\gamma=\gamma_i)}{\sum_{i=1}^{I}f(z|\gamma=\gamma_i,\alpha,\beta,\sigma^2)\,P(\gamma=\gamma_i)}. \tag{A. 13}$$

I represents the number of mass points in the prior distribution of γ. The posterior mean is obtained by sampling from the discrete distribution. A similar strategy is adopted in Albert and Chib (1993). To implement the procedure, variates from a multinomial distribution need to be generated. We propose to use the *Direct* method first proposed by Dagpunar, pp. 168, (1988). An additional computational issue to avoid overflow involves subtracing a constant from the log-likelihood of the states given the parameters. This additional term will cancel in equation (A. 13).

References

Aït-Sahalia, Y. (1996a). Non-parametric pricing of interest rate derivative securities. *Econometrica 64*, 527–560.

Aït-Sahalia, Y. (1996b). Testing continuous-time models of the spot interest rate. *Review of Financial Studies 9*, 385–426.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polichotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Andersen, T. G. and J. Lund (1997). Estimating continuous-time stochastic volatility models of the short-term interest rate. *Journal of Econometrics 77*, 343–377.

Barndorff-Nielsen, O. E., J. L. Jensen, and M. Sørensen (1996). Some stationary processes in discrete and continuous time. *Advances in Applied Probability* , forthcoming.

Basawa, I. V., V. P. Godambe, and R. L. Taylor (Eds.) (1997). *Estimating functions for discretely observed diffusions: A review*, Volume 32 of *IMS Lecture Notes - Monograph Series*. Selected Proceedings of the Symposium on Estimating Functions.

Besag, J., P. Green, D. Higdon, and K. Mengersen (1995). Bayesian computation and Stochastic systems. *Statistical Science 10*, 3–66.

Bibby, B. M. and M. Sørensen (1995). Martingale estimating functions for discretely observed diffusion processes. *Bernoulli 1*, 17–39.

Broze, L., O. Scaillet, and J.-M. Zakaïan (1998). Quasi-Indirect Inference for diffusion processes. *Econometric Theory 14*, 161–186.

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

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician 49*, 327–336.

Chib, S. and E. Greenberg (1996). Markov chain Monte Carlo simulation methods in econometrics. *Econometric Theory 12*, 409–431.

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

Dacunha-Castelle, D. and D. Florens-Zmirou (1986). Estimation of the coefficients of a diffusion from discrete observations. *Stochastics 19*, 263–284.

Dagpunar, J. (1988). *Principles of Random Variate Generation.* Oxford: Clarendon Press.

Dawid, A. P. (1982). The well-calibrated Bayesian (with discussion). *Journal of the American Staistical Assiociation 77*, 605–613.

Dixit, A. (1993). *The Art of Smooth Pasting.* Switzerland: Harwood.

Doornik, J. A. (1996). *Ox: Object Oriented Matrix Programming, 1.10.* London: Chapman & Hall.

Elerian, O. (1998). Simulation estimation of continous time models with applications to finance. D.Phil in progress, Nuffield College, Oxford.

Eraker, B. (1997). *Markov Chain Monte Carlo Analysis of Diffusion Models with Application to Finance.* Ph. D. thesis, Norwegian School of Economics and Business Administration. Master's thesis.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications* (Second ed.), Volume 2. New York: Wiley.

Florens-Zmirou, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics 20*, 547–557.

Gallant, A. R. and J. R. Long (1997). Estimating stochastic differential equations efficiently by minimum Chi-squared. *Biometrika 84*, 125–141.

Gallant, A. R. and G. Tauchen (1996). Which moments to match. *Econometric Theory 12*, 657–681.

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

Genon-Catalot, V. and J. Jacod (1993). On the estimation of the diffusion coefficient for multidimensional diffusion processes. *Annales de l'I.H.P., Probabilités-Statistiques 29*, 119–151.

Geweke, J. (1992). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion)*, Volume 4, pp. 169–193. Oxford: Oxford University Press. Eds. Bernado, J M and Berger, J O and Dawid, A P and Smith, A F M.

Geweke, J. (1996). *Posterior simulators in econometrics*, Volume 3 of *Advances in Economics and Econometrics: theory and applications, 7th World Congress*, pp. 128–165. Cambridge.

Gilks, W. R., N. G. Best, and K. K. C. Tan (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *JRSSC 44*, 155–173.

Gourieroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics 3*, S85–S118.

Hull, J. and A. White (1990). Pricing interest-rate derivative securities. *Review of Financial Studies 3*, 573–592.

Jiang, G. J. and J. L. Knight (1997). A Nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest-rate model. *Econometric Theory 13*, 615–645.

Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics 24*, 211–229.

Kessler, M. and M. Sørensen (1995). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli* , . forthcoming.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic Volatility: Likelihood inference and comparison with Arch models. *Review of Economic Studies 65*, 361–394.

Kloeden, P. E. and E. Platen (1992). *Numerical Solutions to Stochastic Differential Equations*. New York: Springer.

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

Merton, R. (1990). *Continuous-Time Finance*. Cambridge: Blackwell.

Øksendal, B. (1995). *Stochastic Differential Equations: An Introduction with Applications* (Fourth ed.). Berlin: Springer-Verlag.

Pedersen, A. R. (1994a). Quasi-likelihood inference for discretely observed diffusion processes. Research Report No. 295, Department of Theoretical Statistics, University of Aarhus.

Pedersen, A. R. (1994b). Uniform residuals for discretely observed diffusions. Unpublished paper: Department of Statistics, University of Aarhus.

Pedersen, A. R. (1995). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion process. *Bernoulli 1*, 257–279.

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

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics 23*, 470–472.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method* (First ed.). New York: John Wiley.

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

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*, 653–667.

Smith, A. (1993). Estimating non-linear time series models using simulated vector autoregressions. *Journal of Applied Econometrics 8*, 63–84.

Smith, J. Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting 4*, 283–291.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of American Statistical Association 82*, 528–550.

Tauchen, G. (1997). *New minimum Chi-squared methods in empirical finance*, Volume 3 of *Advances in Economics and Econometrics: Theory and Applications, 7th World Congress*, Chapter 9, pp. 279–317. Econometric Society Monographs.

Model	OU	Quadratic Drift
a_j	μy_{k+j}	μy_{k+j}^2
a'_j	μ	$2\mu y_{k+j}$
a''_j	0	0
g_j	σ^{-2}	σ^{-2}
g'_j	0	0
g''_j	0	0
d_j	$y_{k+j+1} - (1+\mu\Delta)y_{k+j}$	$y_{k+j+1} - (1+\mu\Delta y_{k+j})y_{k+j}$
c_j	$(1+\mu\Delta)$	$-(1+2\mu\Delta y_{k+j})$

Model	CIR	Hull White
a_j	$\frac{\alpha}{\exp(x_{k+j})} - \beta - \frac{\sigma^2}{2\exp(x_{k+j})}$	$\frac{\alpha}{\exp(x_{k+j})} - \beta - \frac{\sigma^2}{2}\exp\{2(\gamma-1)x_{k+j}\}$
a'_j	$-\frac{\alpha}{\exp(x_{k+j})} + \frac{\sigma^2}{2\exp(x_{k+j})}$	$-\frac{\alpha}{\exp(x_{k+j})} - \sigma^2(\gamma-1)\exp\{2(\gamma-1)x_{k+j}\}$
a''_j	$\frac{\alpha}{\exp(x_{k+j})} - \frac{\sigma^2}{2\exp(x_{k+j})}$	$\frac{\alpha}{\exp(x_{k+j})} - 2\sigma^2(\gamma-1)^2\exp\{2\gamma-1)x_{k+j}\}$
g_j	$\frac{\exp(x_{k+j})}{\sigma^2}$	$\frac{1}{\sigma^2\exp\{2(\gamma-1)x_{k+j}\}}$
g'_j	$\frac{\exp(x_{k+j})}{\sigma^2}$	$\frac{1}{\sigma^2} - \frac{2(\gamma-1)}{\exp\{2(\gamma-1)x_{k+j}\}}$
g''_j	$\frac{\exp(x_{k+j})}{\sigma^2}$	$\frac{4(\gamma-1)^2}{\sigma^2\exp\{2(\gamma-1)x_{k+j}\}}$
d_j	$x_{k+j+1} - x_{k+j} - a_j\Delta$	$x_{k+j+1} - x_{k+j} - a_j\Delta$
c_j	$1 + a'_j\Delta$	$1 + a'_j\Delta$

Table 2: *Coefficients needed in the Taylor expansion of the log-likelihood for the Ornstein-Uhlenbeck (OU), Quadratic drift process, the Cox, Ingersoll and Ross (CIR) process, and the Hull and White model. For the CIR and Hull-White models, we work with the transformation* $x_t = \log y_t$ *due to the restriction that* $y_t > 0$ *for* $t = 1, \ldots, T$. *The parameters* α, β, γ *and* σ *are all positive and constant, whereas* μ *is negative and constant.*

Summary statistics for the parameters (CIR process)					
	Posterior Mean	Inefficiency	Covariance& *Correlation*		
$M = 0$					
$\alpha\|y$	0.31760	0.77854	0.0054081	*0.97279*	*-0.092305*
$\beta\|y$	0.12770	0.84724	0.00021815	9.2992e-005	*-0.21829*
$\sigma^2\|y$	0.022571	1.1696	-1.2008e-005	-1.1775e-005	3.1291e-005
$M = 1$					
$\alpha\|y$	0.40436	1.1150	0.0013566	*0.98856*	*0.54185*
$\beta\|y$	0.16207	1.1429	0.00054035	0.00022024	*0.53330*
$\sigma^2\|y$	0.031452	0.81728	4.8755e-005	1.9335e-005	5.9681e-006
$M = 3$					
$\alpha\|y$	0.46317	1.2305	0.0026897	*0.99218*	*0.68820*
$\beta\|y$	0.18467	1.2050	0.0010667	0.00042976	*0.67840*
$\sigma^2\|y$	0.039716	1.0100	0.00012994	5.1199e-005	1.3253e-005
$M = 5$					
$\alpha\|y$	0.48435	2.9272	0.0029436	*0.99234*	*0.75960*
$\beta\|y$	0.19274	2.8637	0.0011658	0.00046890	*0.75262*
$\sigma^2\|y$	0.043112	3.4651	0.00017067	6.7491e-005	1.7150e-005
$M = 10$					
$\alpha\|y$	0.51020	3.2109	0.0031933	*0.99134*	*0.75791*
$\beta\|y$	0.20257	3.2473	0.0012596	0.00050556	*0.75184*
$\sigma^2\|y$	0.047446	3.7755	0.00020091	7.9299e-005	2.2005e-005
$M = 20$					
$\alpha\|y$	0.51063	11.035	0.0037585	*0.99268*	*0.80889*
$\beta\|y$	0.20264	11.189	0.0014996	0.00060718	*0.80235*
$\sigma^2\|y$	0.048186	14.682	0.00026019	0.00010373	2.7529e-005
$M = 30$					
$\alpha\|y$	0.50255	14.210	0.0033777	*0.99158*	*0.78960*
$\beta\|y$	0.19932	14.063	0.0013367	0.00053798	*0.78177*
$\sigma^2\|y$	0.047606	19.590	0.000222342	8.8281e-005	2.3704e-005

Table 3: *The posterior means, covariances and correltaions are shown for the parameters of the CIR process. True values are* $\alpha = 0.5$, $\beta = 0.2$ *and* $\sigma^2 = 0.05$. *Data was obtained using* $T = 500$, $R = 100,000$ *and* $\Delta^{\dagger} = 5$. *The MH algorithm was run with* $M = 0, 1, 10$ $(\lambda = 3)$, 20 $(\lambda = 5)$ *and* 30 $(\lambda = 9)$ *for* $N = 10,000$ *iterations. The bandwidth used in the Parzen computations,* $B_N = 100$ *in all computations.*

Monte Carlo Results (CIR process)					
	mean	bias	s.d.	mean(s.e)	MSE($\times$100)
α (0.5)					
$M = 0$	0.310	-0.190	0.0161	0.0227	3.6190
$M = 1$	0.393	-0.107	0.0276	0.0359	1.2302
$M = 3$	0.449	-0.051	0.0375	0.0480	0.4032
$M = 5$	0.470	-0.031	0.0403	0.0515	0.2570
$M = 10$	0.496	-0.004	0.0461	0.0543	0.2136
$M = 20$	0.498	-0.0019	0.0451	0.0580	0.2035
$M = 30$	0.494	-0.0061	0.0445	0.0579	0.2021
β (0.2)					
$M = 0$	0.127	-0.074	0.0069	0.0095	5.4814
$M = 1$	0.159	-0.041	0.0115	0.0146	0.1806
$M = 3$	0.181	-0.019	0.0154	0.0194	0.0605
$M = 5$	0.189	-0.011	0.0165	0.0208	0.0400
$M = 10$	0.199	-0.000	0.0188	0.0219	0.0355
$M = 20$	0.200	-0.000	0.0185	0.0236	0.0341
$M = 30$	0.198	-0.002	0.0182	0.0234	0.0337
σ (0.05)					
$M = 0$	0.023	-0.027	0.0008	0.0056	0.0738
$M = 1$	0.032	-0.018	0.0013	0.0025	0.0325
$M = 3$	0.040	-0.010	0.0022	0.0036	0.0097
$M = 5$	0.044	-0.006	0.0026	0.0042	0.0044
$M = 10$	0.048	-0.002	0.0033	0.0048	0.0013
$M = 20$	0.049	-0.001	0.0033	0.0052	0.0011
$M = 30$	0.049	-0.001	0.0034	0.0051	0.0013

Table 4: *The results from a small simulation study for the parameters of the CIR process are shown. Ten sets of data with $T = 500$, $R = 100,000$ and $\Delta^{\dagger} = 5$ were generated, and the MH algorithm was run with $M = 0, 1, 10, 20$ and 30 for $N = 10,000$ iterations.*

Model	Drift function $a(y;\theta)$	Volatility function $b^2(y;\theta)$
Vasicek	$\alpha_0 + \alpha_1 y$	β_0
CIR	$\alpha_0 + \alpha_1 y$	$\beta_1 y$
Affine CIR	$\alpha_0 + \alpha_1 y$	$\beta_0 + \beta_1 y$
Aït-Sahalia (1996b)	$\alpha_0 + \alpha_1 y + \alpha_2 y^2 + \frac{\alpha_3}{y}$	$\beta_0 + \beta_1 y + \beta_2 y^{\beta_3}$
General parametric model	$\alpha_0 + \alpha_1 y + \alpha_2 y^2 + \frac{\alpha_3}{y}$	$\beta_0 + \beta_1 y + \beta_2 y^2 + \beta_3 y^3$

Table 5: *Models to be considered for the short-rate process: Vasicek, CIR, Affine CIR, Aït-Sahalia (1996b) model, and general parametric model.*

Parameters of Vasicek applied to Eurodollar short-rate data								
	Mean	HPD		MCse	Ineff	Covariance & *Correlation*		
$M = 10$								
$\alpha_0 \vert y$	0.000563	0.000271	0.000821	8.74e-08	1.2	1.98e-08	*-0.917*	*0.065*
$\alpha_1 \vert y$	-0.00673	-0.00972	-0.00377	8.85e-07	1.2	-1.97e-07	2.33e-06	*-0.080*
$\beta_0 \vert y$	1.66e-05	1.60e-05	1.71e-05	3.37e-09	22.4	2.73e-12	-3.67e-11	9.04e-14

Table 6: *The posterior means, HPD, Monte Carlo standard errors, inefficiency factors (Ineff), covariances and correlations (in italics) are shown for the parameters using the Eurodollar short-rate data applied to the Vasicek process. The MH algorithm was run with* $M = 10$ *(*$\lambda = 3$*), for* $N = 2000$ *iterations. The bandwidth used in the Parzen computations,* $B_N = 100$ *in all computations.*

Parameters of CIR applied to Eurodollar short-rate data								
	Mean	HPD		MCse	Ineff	Covariance & *Correlation*		
$M = 10$								
$\alpha_0 \vert y$	0.00042392	0.00022572	0.00063118	4.6733e-008	0.9	1.09e-08	*-0.911*	*0.085*
$\alpha_1 \vert y$	-0.0051101	-0.00246	-0.00779	5.99e-07	0.9	-1.32e-07	1.92e-06	*0.0797*
$\beta_1 \vert y$	0.000145	0.000140	0.000150	3.06e-08	22.8	2.38e-11	2.96e-10	7.19e-12

Table 7: *The posterior means, HPD, Monte Carlo standard errors, inefficiency factors, covariances and correlations (in italics) are shown for the parameters using the CIR process applied to the Eurodollar short-rate data. The MH algorithm was run with* $M = 10$ *(*$\lambda = 3$*), for* $N = 2000$ *iterations. The bandwidth used in the Parzen computations,* $B_N = 100$ *in all computations.*

Parameters of Affine CIR applied to Eurodollar short-rate data									
	Mean	HPD		MCse	Ineff	Covariance & *Correlation*			
$M = 10$									
$\alpha_0 \| y$	0.000420	0.000202	0.000607	2.84e-08	1.4	1.07e-08	*-0.908*	*0.00202*	*0.0766*
$\alpha_1 \| y$	-0.00506	-0.00770	-0.00245	3.39e-07	1.3	-1.27e-07	1.82e-06	*-0.00439*	*-0.0728*
$\beta_0 \| y$	1.92e-09	4.10e-12	7.95e-09	1.92e-11	25.4	7.81e-16	-2.24e-14	1.43e-17	*0.0323*
$\beta_1 \| y$	0.000145	0.000140	0.000150	1.21e-08	20.2	2.37e-11	-2.94e-10	3.65e-16	8.96e-12

Table 8: *The posterior means, HPD, Monte Carlo standard errors, inefficiency factors (Ineff), covariances and correlations (in italics) are shown for the parameters using the Affine CIR process applied to the Eurodollar short-rate data. The MH algorithm was run with* $M = 10$ *(*$\lambda = 3$*), for* $N = 5000$ *iterations. The bandwidth used in the Parzen computations,* $B_N = 100$ *in all computations.*

	Mean	MCse	HPD		Ineff
$M = 10$					
$\alpha_0 \| y$	-0.00441	1.25e-06	-0.00674	-0.00215	1.6
$\alpha_1 \| y$	0.0612	1.83e-05	0.0280	0.0924	1.7
$\alpha_2 \| y$	-0.259	7.72e-05	-0.387	-0.118	1.7
$\alpha_3 \| y$	9.85e-05	2.53e-08	4.97e-05	0.000145	1.5
$\beta_0 \| y$	4.71e-06	2.32e-09	4.40e-06	5.04e-06	21.2
$\beta_1 \| y$	1.13e-06	6.32e-09	3.90e-07	1.95e-07	21.6
$\beta_2 \| y$	2.25e-05	1.61e-07	6.62e-06	4.21e-05	24.1
$\beta_3 \| y$	0.00106	5.21e-06	0.00997	0.0115	21.0

Covariance & *Correlation*							
1.40e-06	*-0.983*	*0.928*	*-0.982*	*0.00393*	*-0.0208*	*0.0190*	*-0.0488*
-1.93e-05	0.000274	*-0.978*	*0.937*	*-0.0173*	*0.0347*	*-0.0230*	*0.0549*
7.61e-05	-0.00112	0.00480	*-0.861*	*0.0238*	*-0.0448*	*0.0284*	*-0.0594*
-2.88e-08	3.84e-07	-1.48e-06	6.13e-10	*0.0190*	*0.00429*	*-0.0170*	*0.0409*
7.68e-13	-4.72e-11	2.72e-10	7.76e-014	2.72e-14	*-0.185*	*-0.0962*	*-0.286*
-1.08e-11	2.52e-10	-1.36e-09	4.67e-14	-1.34e-14	1.93e-13	*-0.0182*	*-0.136*
2.25e-10	-3.80e-09	1.96e-08	-4.21e-12	-1.59e-13	-7.98e-14	9.99e-11	*-0.266*
-2.15e-08	3.38e-07	-1.53e-06	3.77e-10	-1.76e-11	-2.23e-11	-9.90e-10	1.39e-07

Table 9: *The posterior means, standard errors, HPD, inefficiency factors (Ineff), covariances and correlations (in italics) are shown for the parameters using the Eurodollar short-rate data applied to the general parametric process. The MH algorithm was run with* $M = 10$ *(*$\lambda = 3$*), for* $N = 1500$ *iterations. The bandwidth used in the Parzen computations,* $B_N = 100$ *in all computations.*

Model	$\log m(x)$	numerical se
Vasicek	6591.4	0.24090
CIR	8319.9	0.12940
Affine CIR	8321.1	0.22718
General parametric model	9083.7	0.31326

Table 10: *The log marginal likelihood computations are shown for* $x = \log y$*.* $f(x|\theta^*)$ *and* $\pi(\theta^*|x)$ *are computed using* $M = 10$*,* $\lambda = 3$ *burn-in=200 and* $N = 2000$*. Priors used are Inverse Gamma* $\left\{\frac{\theta^*}{(15\sigma^*)^2} + 2, \theta^*(\sigma^* - 1)\right\}$ *for* β*, and Normal* $\{\theta^* + 4\sigma^*,\ (15\sigma^*)^2\}$ *for* α*, (*β_1 *uses normal specification in general parametric model), where* θ^* *and* σ^* *are the respective posterior means and standard deviations, based on a training sample data set. Numerical standard error of marginal likelihood (log scale) is denoted by (numerical se).*