# LIKELIHOOD INFERENCE FOR DISCRETELY OBSERVED NON-LINEAR DIFFUSIONS

Ola Elerian, Siddhartha Chib and Neil Shephard*

May 8, 2000

This paper is concerned with the Bayesian estimation of non-linear stochastic differential equations when observations are discretely sampled. The estimation framework relies on the introduction of latent auxiliary data to complete the missing diffusion between each pair of measurements. Tuned Markov chain Monte Carlo (MCMC) methods based on the Metropolis-Hastings algorithm, in conjunction with the Euler-Maruyama discretization scheme, are used to sample the posterior distribution of the latent data and the model parameters. Techniques for computing the likelihood function, the marginal likelihood and diagnostic measures (all based on the MCMC output) are developed. Examples using simulated and real data are presented and discussed in detail.

*Keywords*: Bayes estimation, Non-linear diffusion, Euler-Maruyama approximation, Maximum likelihood, Markov chain Monte Carlo, Metropolis Hastings algorithm, Missing data, Simulation, S-tochastic differential equation.

*We would like to thank the co-editor, Alain Monfort and the four referees for their very helpful comments on a previous draft of this paper. We also thank Peter Burridge, Björn Eraker, Michael K Pitt and participants at various conferences and seminars for discussions on the issues raised by the paper. In addition, we gratefully acknowledge the financial support of Ola Elerian by Nuffield College and Neil Shephard through the ESRC projects R000235488, R000238391 and the EU grant 'Estimation via simulation in econometrics'. All computations reported in this paper were carried out using the Ox language of Doornik (1996).

# 1 INTRODUCTION

## 1.1 *Models*

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

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

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 $\theta$, and $dW(t)$ is the increment of a standard Wiener process.[1] SDEs 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, p. 64)) and suppose that one has measurements $y_t = y(\tau_t)$ at times $\{\tau_1, \ldots, \tau_T\}$, where $\Delta_t^\dagger = \tau_{t+1} - \tau_t \geq 0$, for $t \leq T$. The aim is to estimate $\theta$ given the measurements $Y = (y_1, \ldots, y_T)'$.

In the likelihood context, estimation of $\theta$ 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, i.e., the stochastic differential equation,

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

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 SDEs are rarely available. This has led to growing interest in methods for estimating SDEs on the basis of discretely sampled measurements. Important developments include the indirect inference method of Gourieroux, Monfort, and Renault (1993), the efficient method of moments estimator of Gallant and Long (1997) and the non-parametric approaches of Aït-Sahalia (1996a) and Jiang and Knight (1997). Discretely observed diffusions have also been fit by estimating functions, see Kessler and Sørensen (1999), Sørensen (1997), Florens-Zmirnou (1989) and Hansen and Scheinkman (1995) and by the likelihood based method of Pedersen (1995).

[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 (1998)).

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 (MCMC) 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).[2] Related ideas have been developed by Billio, Monfort, and Robert (1998) in their work applying the Geyer (1999) simulated likelihood ratio method to diffusions and other econometric problems.

### 1.2 *Augmentation for SDEs: Motivation*

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

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

under which the transition density is

$$f(y_{t+1}|y_t, \theta) = \phi\left\{y_{t+1}|y_t + a(y_t, t, \theta)\Delta_t^\dagger,\ b^2(y_t, t, \theta)\Delta_t^\dagger\right\}\,, \tag{1.3}$$

where $\phi(\cdot|m, v)$ denotes the Gaussian density with mean $m$ and variance $v$. Although this is the simplest discrete time approximation of the SDE with respect to the strong convergence criterion Kloeden and Platen (1992, Section 10.2) it is normally too coarse to approximate the true transition density adequately.

In order to describe a modified approach consider any two consecutive times $(\tau_t, \tau_{t+1})$, as in Figure 1, and assume for notational simplicity that the time gap $\Delta_t^\dagger = \Delta^\dagger$ is independent of $t$. Let $\{\tau_{t,1}, ..., \tau_{t,M}\}$ denote $M$ auxiliary times between $(\tau_t, \tau_{t+1})$, assumed to be evenly spaced, with time gap

$$\Delta = \tau_{t,k+1} - \tau_{t,k} = \frac{\Delta^\dagger}{M+1}$$

[2] Our work differs from Eraker (1998) in two crucial respects: (i) we develop a tuned MCMC algorithm for carrying out the calculations, rather than a single move method, and discuss the properties of our method in detail; and (ii) we provide tools for comparing alternative models and conducting model diagnostics.

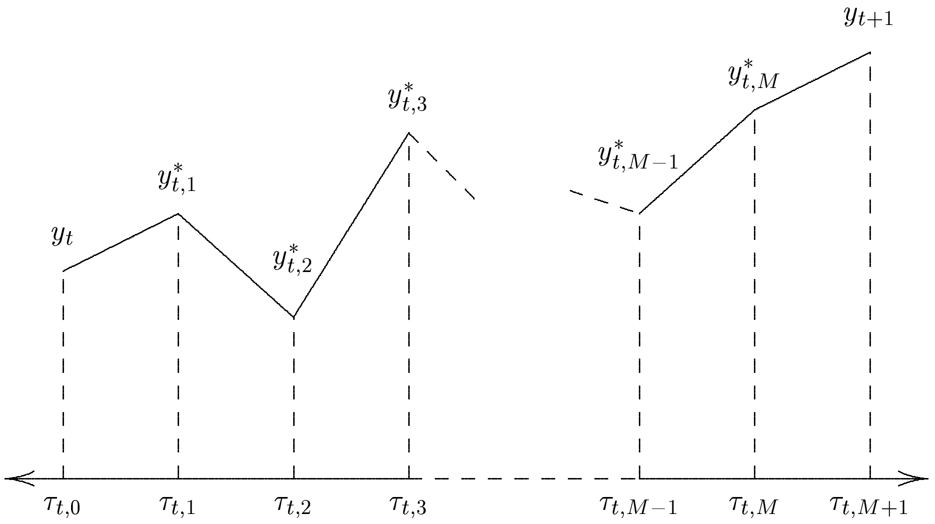

**Figure** 1: *Discretization scheme is shown where the auxiliary points $y^*_t$ are introduced between observed values $y_t$ and $y_{t+1}$.*

for all $t, k$.[3] At each auxiliary time, let $y^*_{t,j} = y^*(\tau_{t,j})$, $j \leq M$, denote the unobserved (or latent) observation and let $y^*_t = (y^*_{t,1}, \dots, y^*_{t,M})$ denote the entire collection of latent observations at the times $\{\tau_{t,1}, ..., \tau_{t,M}\}$.

Then, an improved approximation of the true transition density $g(y_{t+1}|y_t, \theta)$ is given by

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

where

$$f(y^*_{t,j}|y^*_{t,j-1},\theta) = \phi\left\{y^*_{t,j}|y^*_{t,j-1} + a(y^*_{t,j-1},\tau_{t,j-1},\theta)\Delta,\ b^2(y^*_{t,j-1},\tau_{t,j-1},\theta)\Delta\right\}$$

is the transition density using the Euler approximation. It can be shown that $f^M \overset{\text{a.s.}}{\to} g$, as $M \to \infty$ (see Pedersen (1995, Theorem 3) and Kohatsu-Higa and Ogawa (1997)). The global error of the approximation can be measured as $\mathrm{E}\left|f^M(y_{t+1}|y_t,\theta) - g(y_{t+1}|y_t,\theta)\right|$ by specializing a result of Talay and Tubaro (1990) where it is shown that this expectation can be expanded in terms of powers of

[3] The choice of units for $\Delta^\dagger_t$ has a bearing on the scale of $\theta$ and implicitly scales the drift and volatility functions. Because the initial choice of units is merely convenient labelling, chosen for ease of interpretation, the scaling is arbitrary and not reflected in the notation.

$1/M$ (see also Talay (1995) for a discussion), with the leading term being of order $1/M$.[4] A similar type of result would hold for the convergence of the log of the densities which is proved in Bally and Talay (1995) using the Malliavin calculus. The essence of this result is that the discretization error is a function of $M$ rather than solely a function of $\Delta^\dagger$.[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 $y_{t+1}|y_t, \theta$ under the Euler scheme is normal with a mean of $\rho_E^\dagger y_t$ and a variance of $\sigma^2\Delta^\dagger$, where $\rho_E^\dagger = 1 + \mu\Delta^\dagger$. Under the strong solution, the distribution of $y_{t+1}|y_t, \theta$ is also normal with mean $\rho_S^\dagger y_t$ and 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 neighboring points, $y_{t-1}$ and $y_{t+1}$. Then the distribution of the resulting bridge process under the Euler scheme can be expressed as

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

whereas from the strong solution,

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

Similarly, if we consider a block of $M$ latent points $y_t^* = (y_{t,1}^*, \ldots, y_{t,M}^*)$, then $f(y_t^*|y_t, y_{t+1}, \theta)$ is seen to be a Gaussian distribution with expected value

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

where $\rho$ 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 discretizations under the Euler scheme. Figure 2 illustrates the curvature bias with $M = 0, 3$ and 8 latent points denoted as

[4] These results extend to situations where $W$ is a more general Lévy process, and the same expansion can also be established when $g$ is only assumed to be measurable and bounded, see Protter and Talay (1997).

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

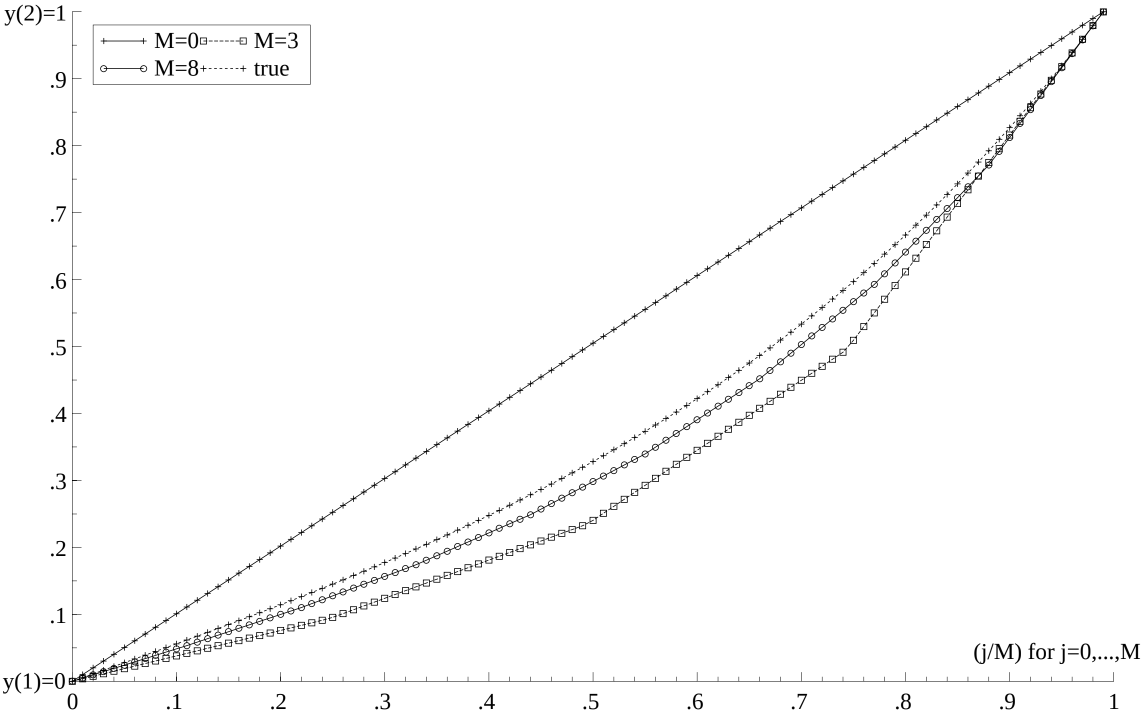

**Figure** 2: *$E(y_1^*|y_1, y_2, \theta)$ is graphed for the OU process for different values of $M$ under the Euler scheme. The number of latent points, $M = 0, 3$ and $8$ are shown for the Euler approximation. The strong solution is denoted by 'true'. The value of $\mu$ is taken to be -1 and $\Delta^\dagger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

$y_1^* = (y_{1,1}^*, \ldots, y_{1,M}^*)$. 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.

In addition, we compare, in Figure 3, the true OU log-likelihood for $\mu$ given by $\sum_{t=1}^{T-1} \log g(y_{t+1}|y_t, \theta)$ and the approximate conditional likelihood $\sum_{t=1}^{T-1} \log f^M(y_{t+1}|y_t, \theta)$ where

$$g(y_{t+1}|y_t, \theta) = \phi\left[y_{t+1} \,\middle|\, \exp(\mu\Delta^\dagger)y_t,\ \frac{\sigma^2}{2\mu}\left\{\exp(2\mu\Delta^\dagger) - 1\right\}\right]$$

and

$$f^M(y_{t+1}|y_t, \theta) = \phi\left[y_{t+1} \,\middle|\, \left(1 + \frac{\mu\Delta^\dagger}{M+1}\right)^{M+1} y_t,\ \sigma^2\left(\frac{\Delta^\dagger}{M+1}\right)\left\{\frac{1 - \left(1 + \frac{\mu\Delta^\dagger}{M+1}\right)^{2(M+1)}}{1 - \left(1 + \frac{\mu\Delta^\dagger}{M+1}\right)^2}\right\}\right],$$

for data generated using $\Delta^\dagger = 4$, $T = 500$, $\mu = -0.5$ and $\sigma^2 = 0.01$ (assumed known). The transition density $f^M(y_{t+1}|y_t,\theta)$ in this case is obtained analytically by integrating out $y^*_{t,M}$ from $f^M(y_{t+1}, y^*_{t,M}|y_t,\theta)$; see (1.4). The approximate likelihoods are computed for various values of $M$. Although the difference between the approximate and the true likelihood is about ten on the log scale, even for $M = 500$, it is interesting to note that the quantiles of the posterior of $\mu$ conditioned on $\sigma^2 = 0.01$ stabilize for $M$ as small as ten, as shown in Figure 4.[6] The quantiles in Figure 4 are computed for values of $M$, ranging from 0 to 1000, but are graphed using a scale of $\log_{10}(M+1)$ on the $x$-axis. It can be seen that the introduction of the auxiliary points provides a better approximation of the likelihood function and, consequently, of the posterior distribution. Further, we are now in a position to control the accuracy of that approximation by our choice of $M$.[7] The benefit of using auxiliary variables will also be demonstrated when we analyze the MCMC output from the estimation procedure (outlined in Section 2) applied to various models. Table I gives the drift and volatility functions for four important processes, which will be considered in the paper.

**Table** I: *Drift $a(y,\theta)$ and volatility $b(y,\theta)^2$ specifications for the OU, the Quadratic drift process, the CIR process, and the Hull-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 $t = 1,\dots,T$. The parameters $\alpha, \beta, \gamma$ and $\sigma$ are all positive and constant, whereas $\mu$ is negative and constant.*

| Process | OU | Quadratic Drift | CIR ($x = \log y$) | Hull-White ($x = \log y$) |
|---|---|---|---|---|
| $a(y,\theta)$ | $\mu y$ | $\mu y^2$ | $\frac{\alpha}{\exp(x)} - \beta - \frac{\sigma^2}{2\exp(x)}$ | $\frac{\alpha}{\exp(x)} - \beta - \frac{\sigma^2}{2}\exp\{2(\gamma-1)x\}$ |
| $b(y,\theta)^2$ | $\sigma^2$ | $\sigma^2$ | $\frac{\sigma^2}{\exp(x)}$ | $\sigma^2\exp\{2(\gamma-1)x\}$ |

The rest of the paper is organized as follows. In Section 2 we present a Markov chain Monte Carlo simulation technique to sample the posterior distribution of the auxiliary variables and the parameters. A method for sampling the latent data in blocks is proposed and evaluated in relation to alternative simulation schemes. In Section 3 we discuss how posterior inferences can be con-

[6] The approximate posteriors are computed over a grid of values for $\mu \in [-0.8, 0])$ and the prior is set to $N(-2, 2)$. The grid points are $[-0.8, -0.6]$, with step size 0.005; $[-0.599, -0.4]$ with step size 0.001; $[-0.395, -0.001]$ with step size 0.005.

[7] Additional results, comparing the difference between the approximate and true log-likelihoods and the average mean square error for the parameters of the OU process, as simulation size varies, are reported in Elerian (1999).

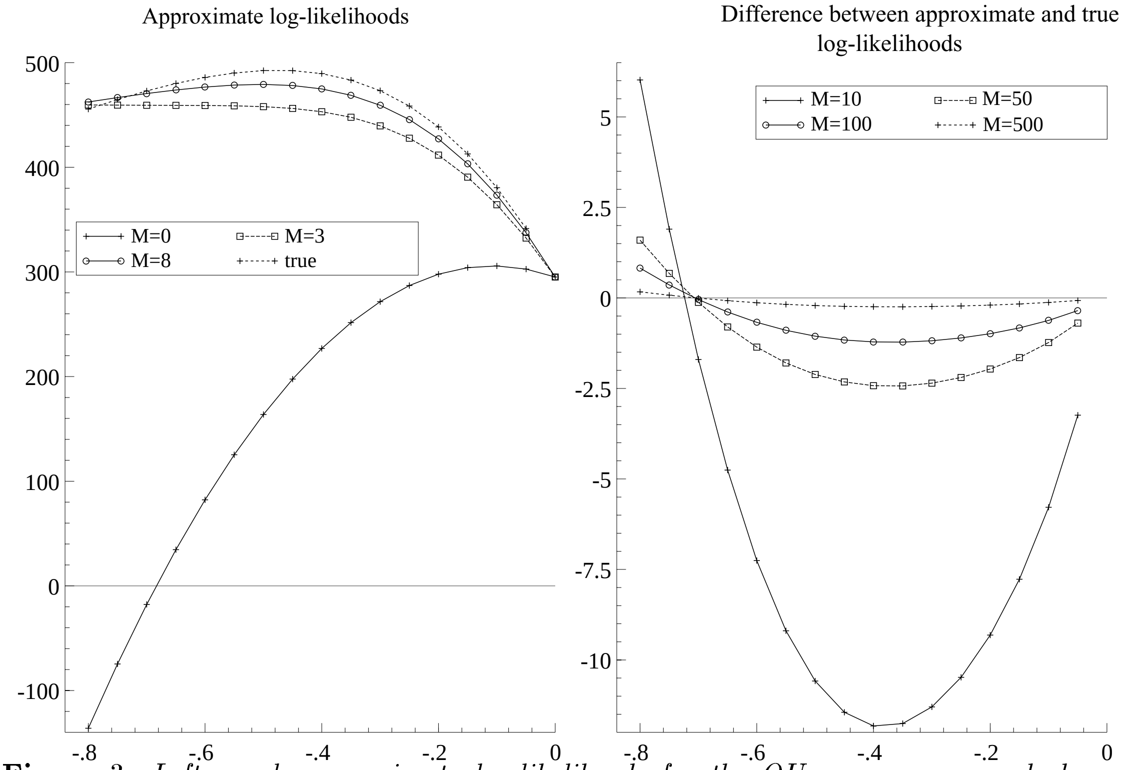

**Figure** 3: *Left panel: approximate log-likelihoods for the OU process are graphed against $\mu$ for different values of $M$. The number of latent points are $M = 0, 3$ and 8. The strong solution is denoted by 'true'. Right panel: The difference between the approximate log-likelihoods and the true log-likelihood given by the strong solution. The number of auxiliary points are $M = 10, 50, 100$ and 500. Each approximation is evaluated for different values of $\mu \in [-0.8, 0]$. For the DGP, $\mu = -0.5$, $\sigma^2 = 0.01$ and $\Delta^\dagger = 4$.*

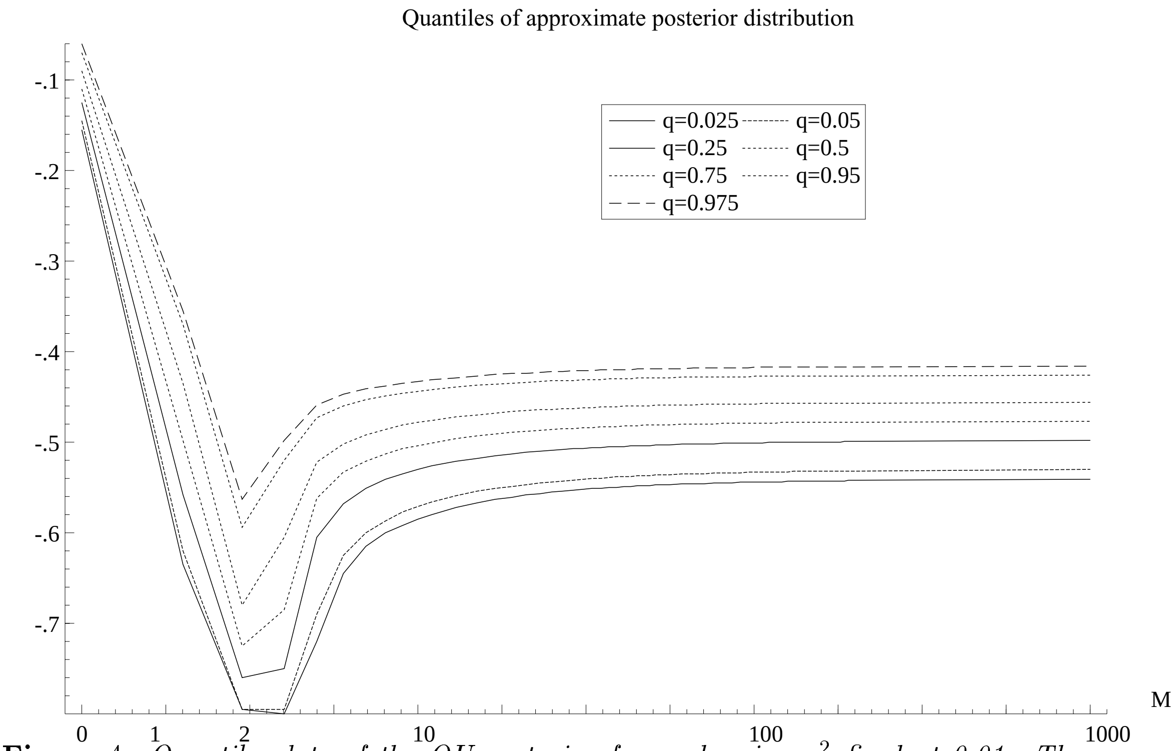

**Figure** 4: *Quantile plots of the OU posterior for $\mu$, keeping $\sigma^2$ fixed at 0.01. The approximate posterior is computed for different values of $M$, ranging from 0 to 1000, but is graphed using a scale of* $\log_{10}(M+1)$ *on the $x$-axis. Again the DGP has $\mu = -0.5$, $\sigma^2 = 0.01$ and $\Delta^\dagger = 4$.*

ducted based on the output of the Markov chain simulation procedure. 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. Some concluding remarks are made in Section 6.

# 2 MCMC-BASED ESTIMATION OF NONLINEAR DIFFUSION-S

## 2.1 *Framework*

To describe our inferential framework, consider the approximate joint density of the observed data $Y = (y_2, ..., y_T)$

$$f^M(Y|y_1, \theta) = \prod_{t=1}^{T-1} f^M(y_{t+1}|y_t, \theta),$$

where $f^M(y_{t+1}|y_t, \theta)$ is the transition density in (1.4). In general, the density $f^M(y_{t+1}|y_t, \theta)$ cannot be computed exactly but, following the discussion of the previous section, an effective way of dealing with this difficulty is to consider the joint posterior distribution of the parameters and the augmented data $Y^* = (y_1^*, y_2^*, ..., y_{T-1}^*)$. Let $\pi(\theta)$ denote the prior density of the parameters and let

$$\pi^M(\theta, Y^*|Y) \propto f(Y^*, Y|y_1, \theta)\pi(\theta) \ ,$$

where

$$f(Y^*, Y|y_1, \theta) = \prod_{t=1}^{T-1} \left\{ \prod_{j=0}^{M} f(y_{t,j+1}^* | y_{t,j}^*, \theta) \right\} ,$$

is the complete data density with $y_{t,0}^* = y_t$ and $y_{t,M+1}^* = y_{t+1}$, denote the posterior density of the parameters and the latent data. This augmented posterior density does not require the computation of the likelihood function $f^M(Y|y_1, \theta)$. To analyze the posterior density we can utilize Markov chain Monte Carlo (MCMC) simulation methods. These methods allow one to sample the augmented posterior density by simulating a Markov chain $\{\theta^j, Y^{*j}\}$ constructed to have $\pi^M(\theta, Y^*|Y)$ as the limiting invariant distribution, see, for example, Gilks, Richardson, and Spiegelhalter (1996) and Chib (2000) for reviews of the literature. The trajectory of the Markov chain, after an initial transient or burn-in stage, provides a sequence of (correlated) draws from $\pi^M(\theta, Y^*|Y)$. Furthermore, the draws $\{\theta^j\}$ are automatically from the marginal distribution

$$\pi^M(\theta|Y) = \int \pi^M(\theta, Y^*|Y) dY^*,$$

and can be used to conduct posterior inferences about $\theta$. For example, the sample mean and the sample standard deviation of the sampled draws are estimates of the corresponding posterior mean and posterior standard deviations; simulation consistency of these estimates is established by ergodic laws of large numbers for Markov chains on continuous state spaces. This leads to full likelihood-based inference for the model even though the likelihood is not evaluated.

The degree of data augmentation, $M$, which depends on the space between the observed points, the non-linearity in the drift and volatility functions and the variance between the observed values, influences two aspects of the analysis. First, an increase in $M$ improves the approximation in (1.4), implying that inferences based on $\pi^M(\theta|Y)$ will become less biased as $M$ increases. Second, an increase in $M$ directly increases the dimension of the state space on which the simulation is conducted. Under these circumstances, the sampling process must be more carefully designed to ensure that the simulation output does not display excessive serial dependence. Although one can increase the length of the simulation sample size, a more desirable strategy is to construct samplers that produce good mixing even when $M$ is large.

## 2.2 *Overview of the MCMC method*

Markov chain Monte Carlo sampling from $\theta, Y^*|Y$ is achieved by sampling in turn the full conditional distributions $Y^*|Y,\theta$ and $\theta|Y, Y^*$. One iteration of the Markov chain is completed by revising both $Y^*$ and $\theta$ from these two distributions. A simple calculation (based on the Markov property of the diffusion) shows that the first of these full conditional distributions can be expressed as

$$f(Y^*|y_1, Y, \theta) = \prod_{t=1}^{T-1} f(y_t^*|y_t, y_{t+1}, \theta)\,,$$

due to the fact that the latent data $y_t^*$ is conditionally independent of the remaining latent data, given $(y_t, y_{t+1}, \theta)$. Thus, our simulation procedure in general terms may be described as follows:

`General sampling scheme`

1. Initialize $Y^*, \theta$.
2. Update $y_t^*$ from $y_t^*|y_t, y_{t+1}, \theta$, for $t = 1, 2, ..., T-1$.
3. Update $\theta$ from $\theta|Y^*, Y$.
4. Record the value of $\theta$ and then goto 2.

The most important stage of this procedure is the sampling of the distributions $y_t^*|y_t, y_{t+1}, \theta$ as these are likely to be high dimensional distributions of unknown form and have to be repeated $(T-1)$ times for each sweep of the algorithm.

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

Consider the question of sampling $y_t^* \in \Re^M$ from the target density

$$\begin{aligned} f(y_t^*|y_t, y_{t+1}, \theta) &\propto \prod_{j=0}^{M} f(y_{t,j+1}^*|y_{t,j}^*, \theta) \\ &\propto \prod_{j=0}^{M} \phi\left\{y_{t,j+1}^*|y_{t,j}^* + a(y_{t,j}^*, \tau_{t,j}, \theta)\Delta,\ b^2(y_{t,j}^*, \tau_{t,j}, \theta)\Delta\right\} \end{aligned}$$

where, in general, $y^*_{t,j}$ appears non-linearly in both the drift and diffusion functions. A computationally effective approach for sampling $y^*_t$ from this density can be developed by working in sequence with contiguous subsets of $y^*_t$. Let $y^*_{t(k,m)}$ denote a block of length $m$ $(1 \leq m \leq M-k+1)$ that starts at $y^*_{t,k}$ and ends at $y^*_{t,k+m-1}$ :

$$y^*_{t(k,m)} = \left(y^*_{t,k}, y^*_{t,k+1}, \cdots, y^*_{t,k+m-1}\right)$$

with density conditioned on $(y^*_{t,k-1}, y^*_{t,k+m}, \theta)$ given by

$$f(y^*_{t(k,m)}|y^*_{t,k-1}, y^*_{t,k+m}, \theta) \quad \propto \quad \prod_{j=k-1}^{k+m} \phi\left\{y^*_{t,j+1}|y^*_{t,j} + a(y^*_{t,j}, \tau_{t,j}, \theta)\Delta,\ b^2(y^*_{t,j}, \tau_{t,j}, \theta)\Delta\right\}, \quad (2.8)$$
$$\text{for } k = 1, m-1, 2m-1, ...$$

The idea now is to sample each of the $m$ dimensional vectors $y^*_{t(k,m)}$ in sequence by the Metropolis-Hastings algorithm.

The Metropolis-Hastings (M-H) algorithm (see for example, Chib and Greenberg (1995)) is a general MCMC method for producing sample variates from a given multivariate density such as the ones in (2.8). The method is defined by a user-specified candidate generating density that is used to supply a proposal value and a probability of move that is used to determine if the proposal value should be taken as the next item of the chain. The probability of move is based on the ratio of the target density (evaluated at the proposal value in the numerator and the current value in the denominator) times the ratio of the proposal density (at the current value in the numerator and the proposal value in the denominator). Specifically, let $q(y^*_{t(k,m)}|y^*_{t,k-1}, y^*_{t,k+m}, \theta)$ denote the proposal density conditioned on $(y^*_{t,k-1}, y^*_{t,k+m}, \theta)$ and suppose that the current value of $y^*_{t(k,m)}$ at the end of the $g$th iteration of the Markov chain is $y^{*(g)}_{t(k,m)}$.[8] Then, the M-H step for $y^*_{t(k,m)}$ is implemented by first drawing a candidate value $w \sim q(y^*_{t(k,m)}|y^*_{t,k-1}, y^*_{t,k+m}, \theta)$, computing the probability

$$\alpha\left(y^*_{t(k,m)}, w|y^*_{t,k-1}, y^*_{t,k+m}, \theta\right) = \min\left\{1, \frac{f(w|y^*_{t,k-1}, y^*_{t,k+m}, \theta)q\left(y^{*(g)}_{t(k,m)}|y^*_{t,k-1}, y^*_{t,k+m}, \theta\right)}{f\left(y^{*(g)}_{t(k,m)}|y^*_{t,k-1}, y^*_{t,k+m}, \theta\right)q(w|y^*_{t,k-1}, y^*_{t,k+m}, \theta)}\right\},$$

and then setting $y^{*(g+1)}_{t(k,m)} = w$ with probability $\alpha$ and setting $y^{*(g+1)}_{t(k,m)} = y^{*(g)}_{t(k,m)}$ with probability $(1-\alpha)$. Note that since the probability of moving is based only on ratios of densities one does not need the normalizing constant of the target density.

[8] $y^*_{t,k-1}$ is the updated point obtained at ¿the $g$-th iteration, while $y^*_{t,k+m}$ is set at the value obtained at the $(g-1)$-th iterate.

In the implementation of this method it is vital that one uses a proposal density that allows the chain to efficiently traverse the support of the invariant distribution without staying in one place for many iterations. A simple and general method of specifying such a proposal density is to approximate the target density at the mode by a multivariate-normal or multivariate-t distribution with location given by the mode of $\ln f(\cdot|y^*_{t,k-1}, y^*_{t,k+m}, \theta)$, obtained by a few Newton-Raphson steps, and dispersion given by the negative of the inverse Hessian evaluated at the mode. An early example of this tactic is Chib and Greenberg (1994) while subsequent examples include Chib, Greenberg, and Winkelmann (1998) and Shephard and Pitt (1997).

To develop the proposal density let

$$y^*_{t(k,m)} = (y^*_{t,k}, y^*_{t,k+1}, \ldots, y^*_{t,k+m-1}) = (w_1, \ldots, w_m) = w$$

denote the block of latent values with neigbours $w_0 = y^*_{t,k-1}$ and $w_{m+1} = y^*_{t,k+m}$. Also write

$$\begin{aligned}
a_j &= a(y^*_{t,k+j}) = a(w_{j+1}) \\
g_j &= \{b^2(y^*_{t,k+j})\}^{-1} = \{b^2(w_{j+1})\}^{-1} \\
d_j &= y^*_{t,k+j+1} - (y^*_{t,k+j} + a_j\Delta) = w_{j+2} - (w_{j+1} + a_j\Delta) \\
c_j &= 1 + a'_j\Delta,
\end{aligned}$$

for $j = -1, 0, \ldots, m-1$ and 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. Finally, let

$$\begin{aligned}
u &= \frac{\partial \log f(w|w_0, w_{m+1}, \theta)}{\partial w} = \{u_j\} \quad \text{for } j = 1, \ldots m \\
V &= -\frac{\partial^2 \log f(w|w_0, w_{m+1}, \theta)}{\partial w \partial w'} = \{V_{ij}\} \quad \text{for } i, j = 1, \ldots, m.
\end{aligned}$$

denote the gradient and negative hessian matrix, respectively, of the log target density. Because the only terms in $\log f(w|w_0, w_{m+1}, \theta)$ involving $w_{j+1}$ are

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

we have that

$$u_j = \frac{\partial l_j}{\partial w_{j+1}} + \frac{\partial l_{j+1}}{\partial w_{j+1}} = -\frac{1}{\Delta}\left(g_{j-2}d_{j-2} - g_{j-1}d_{j-1}c_{j-1} + \frac{1}{2}g'_{j-1}d^2_{j-1} - \frac{\Delta g'_{j-1}}{2g_{j-1}}\right),$$

$$V_{jj} = -\frac{1}{\Delta}\left[g_{j-2} + g_{j-1}c_{j-1}^2 - g_{j-1}d_{j-1}a''_{j-1}\Delta - 2g'_{j-1}d_{j-1}c_{j-1} + \frac{1}{2}g''_{j-1}d_{j-1}^2 - \frac{\{g_{j-1}g''_{j-1} - (g'_{j-1})^2\}\Delta}{2g_{j-1}^2}\right],$$

and

$$V_{ij} = \begin{cases} V_{ji} = -\frac{1}{\Delta}(g'_{j-1}d_{j-1} - g_{j-1}c_{j-1}) & \text{for } i = j+1 \\ 0 & \text{for } i > j+1 \end{cases}$$

These equations are used to find the modal value of $w$ and the inverse of $V$ at the mode is used as the dispersion matrix of the proposal density.

We now discuss the choice of $m$ in the above method. First, the choice $m = 1$ which represents single element updating of $y_t^*$ is not recommended. Elerian, Chib and Shephard (1998) show that this leads to poor mixing due to high correlation amongst $(y_{t,k}^*, y_{t,k-1}^*, y_{t,k+1}^*)$. Second, the choice $m = M$ is not practical because it is difficult to sample a high dimensional $y_t^*$ in one block. Thus, values of $m$ that are different from these two choices are preferable. Furthermore, it is not necessary or desirable to fix $m$ at the outset because that means that the blocks always join in the same place, which can foster dependencies in the MCMC sweeps. In order to scramble this type of dependence Shephard and Pitt (1997) found that the chain converged faster if $m$ were selected randomly at each updating stage. This leads to what may be called a random block size M-H algorithm. A simple way of carrying this out is to draw $m-1$ from a Poisson distribution with mean $\lambda$, which leads to an average block size of $\lambda + 1$. Alternatively, other distributions instead of the Poisson can be used to select the block size, as discussed by Wong (1999) in the context of stochastic volatility models.

To summarize, we advocate the following algorithm which is indexed by the tuning parameter $\lambda \geq 1$:

`General sampling scheme with random block sizes`

1. Initialize $Y^*, \theta$.

2. Set $k = 0$.

3. Draw $m \sim Po(\lambda) + 1$; set $k = k + m$. If $k > M$, set $k = M$.

4. Update $y_{t(k,m)}^* | y_{t,k-1}^*, y_{t,k+m}^*, \theta$.

(a) If $k < M$ goto 3.

(b) Else update $\theta$ from $\theta|Y, Y^*$ and then goto 2;

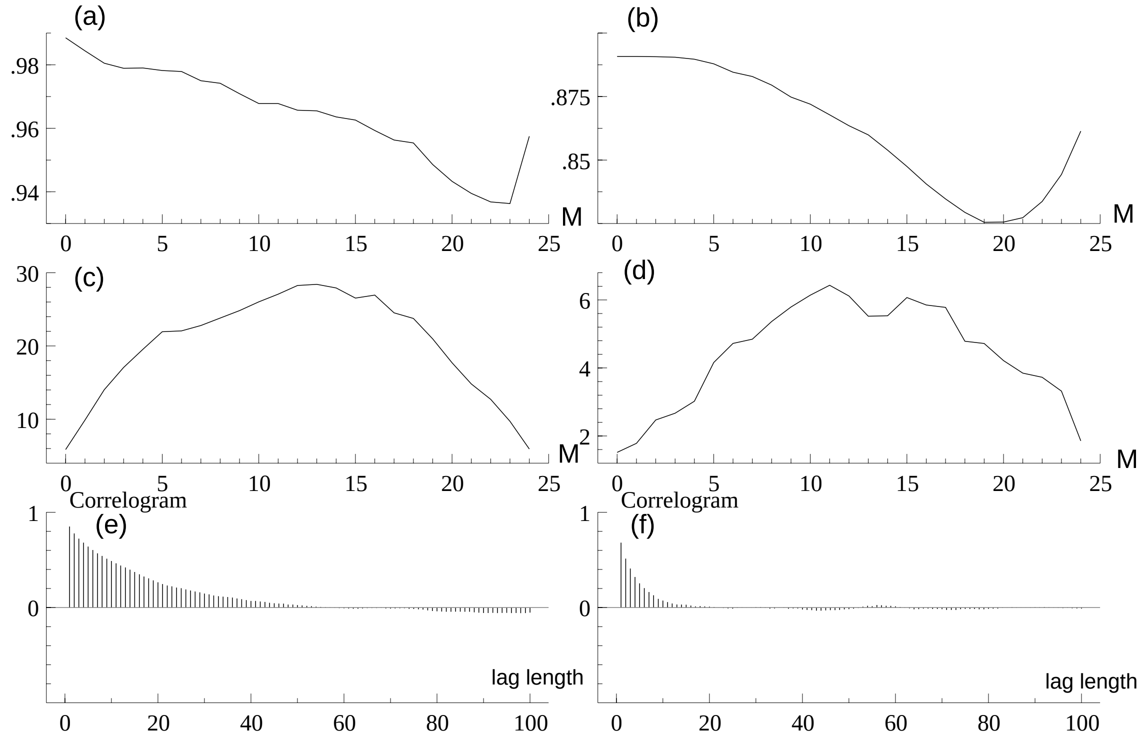

**Figure** 5: *Acceptance rates (a,b), inefficiency factors (c,d, bandwidth $B_N = 100$) and correlogram (e,f) for average block sizes of $\lambda = 2$ (a, c, e) and $\lambda = 10$ (b, d, f), $M = 25$ and $N = 10,000$, with $y_t = 1$, $y_{t+1} = 2$ and parameter values $\mu = -0.005$, $\sigma^2 = 0.03$, for the Quadratic drift process.*

As an example of the value of random-block size sampling, we consider the Quadratic drift model when we have an interval of $M = 25$ points between the observed values and we let the mean $\lambda$ of the random block size be either two or ten. The iterations are run for $N = 10,000$ sweeps with $y_t = 1$ and $y_{t+1} = 2$ fixed; the underlying true parameter values in the data generating process are set to $\mu = -0.005$ and $\sigma^2 = 0.03$. The results, presented in Figure 5, show that the acceptance rates in the M-H step decline slightly but that there is a marked improvement in inefficiency factors.[9]

[9] The inefficiency factor, or the autocorrelation time, of each posterior estimate (computed as a sample average over the simulated values) is defined as the spectrum at zero, $1 + 2\sum_{j=1}^{\infty} \rho_j$, where $\{\rho_j\}$ denotes the autocorrelation function of the simulated values. The inefficiency factor is estimated as $1 + \frac{2N}{N-1}\sum_{j=1}^{B_N} K\left(\frac{j}{B_N}\right)\hat{\rho}_j$, where $\hat{\rho}_j$ is an estimate of the autocorrelation at lag $j$ and $K(\cdot)$ is the Parzen kernel, based on the first $B_N$ autocorrelations. It is equal to the square of the numerical standard error divided by the variance of the posterior estimate under (hypothetical) i.i.d. sampling from the posterior. Geweke (1989) defines the alternative measure of relative numerical efficiency which is the inverse of the inefficiency factor.

The inefficiency factors decrease from thirty to six for the middle states indicating that the sampler is only about six times as inefficient as a hypothetical sampler that produces i.i.d. draws. This is confirmed by the correlogram, which shows heavy persistence until lags of 60 when $\lambda = 2$ against a correlogram that dies out by 30 lags when $\lambda = 10$.

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

In order to complete one cycle of our MCMC sampler we have to sample the full conditional density of $\theta$:

$$\begin{aligned} \pi(\theta|Y, Y^*) &\propto \pi(\theta) f(Y, Y^*|y_1, \theta) \\ &= \pi(\theta) \prod_{t=1}^{T-1} \left\{ \prod_{j=0}^{M} f(y^*_{t,j+1}|y^*_{t,j}, \theta) \right\}, \end{aligned}$$

where $f(y^*_{t,j}|y^*_{t,j-1}, \theta)$ is the normal density given above. Typically $\pi(\theta|Y, Y^*)$ is only available up to an unknown norming constant. In addition, the density is conditioned on both the observed states $Y$ and the simulated auxiliary states $Y^*$ which means that the prior distribution of $\theta$ is being revised using $(T-1)(M+1)$ observations. For some models, the conditional posterior (for a suitable choice of prior) can belong to a known family of distributions. As an example, suppose that

$$y^*_{t,j}|y^*_{t,j-1}, \theta \sim N\left\{y^*_{t,j-1} + \theta y^*_{t,j-1}\Delta,\ b^2\left(y^*_{t,j-1}\right)\Delta\right\},$$

where the unknown parameter $\theta$ indexes the linear drift and the non-linear diffusion is fully known. Then, under a normal prior on $\theta$, $N(\mu_\theta, \sigma^2_\theta)$, the conditional posterior is $\theta|Y, Y^* \sim N(\mu_p, \sigma^2_p)$ with

$$\begin{aligned} \sigma_p^{-2} &= \sigma_\theta^{-2} + \Delta \sum_{t=1}^{T-1} \sum_{j=1}^{M+1} \frac{(y^*_{t,j-1})^2}{b^2(y^*_{t,j-1})}, \quad \text{and} \\ \mu_p &= \sigma^2_p \left\{ \sum_{t=1}^{T-1} \sum_{j=1}^{M+1} \frac{(y^*_{t,j} - y^*_{t,j-1}) y^*_{t,j-1}}{b^2(y^*_{t,j-1})} + \frac{\mu_\theta}{\sigma^2_\theta} \right\}. \end{aligned}$$

The non-linearity in the volatility raises no significant issues. For most problems, however, the distribution $\pi(\theta|Y, Y^*)$ is intractable and must be sampled by (say) the M-H algorithm. The sampling of $\theta$ is model-specific and in some instances it may be possible to update the entire $\theta$ vector at once using a proposal density $q(\theta|Y, Y^*)$ that is matched to the conditional distribution at the modal value. If single block updating results in high rejections then it may be necessary to block $\theta$ into subsets and then to employ the M-H algorithm in sequence to each block. Specific examples of these strategies are provided below.

# 3 POSTERIOR INFERENCES

The techniques outlined so far do not provide an explicit form for the likelihood $g(Y|y_1, \theta)$, or our approximation to it

$$f^M(Y|y_1, \theta) = \prod_{t=1}^{T-1} f^M(y_{t+1}|y_t, \theta).$$

Although we manage to estimate the parameters of the model without computing $f^M(Y|y_1, \theta)$, a one-off estimate of the likelihood function is required for comparing alternative stochastic differential equations using Bayes factors (which are ratios of model marginal likelihoods).

## 3.1 *Likelihood evaluation*

Pedersen (1995) suggested estimating $f^M(y_{t+1}|y_t, \theta)$ by averaging the density of the Euler approximation

$$\widehat{f}^M(y_{t+1}|y_t, \theta) = \frac{1}{R}\sum_{j=1}^{R} f\left(y_{t+1}|y_{t,M_t}^{*(j)}, \theta\right),$$

where $y_{t,M_t}^{*(j)}$ is drawn by iterating the discretized version, defined in (1.2), starting at $y_{t,0}^* = y_t$. Results by Talay and Tubaro (1990) and Kohatsu-Higa and Ogawa (1997) concerning the error induced by using the Euler scheme imply that the absolute value of the expected error, with respect to $g(y_{t+1}|y_t, \theta)$, involves terms of order smaller than or equal to the product of $1/M$ and $1/\sqrt{R}$.

The Pedersen (1995) method can be embedded within a class of importance sampling estimators of $f^M(y_{t+1}|y_t, \theta)$. For further discussion of importance sampling see Kloek and van Dijk (1978), Ripley (1987, pp. 122-3) or Geweke (1989). If we let $q(y_t^*|y_t, y_{t+1}, \theta)$ denote some importance sampling density whose support is the same as that of $f(y_{t+1}, y_t^*|y_t, \theta)$ then

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

The proposal made by Pedersen (1995) is to set

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

which is easy to sample from and has the significant advantage that

$$\frac{f(y_{t+1}, y_t^*|y_t, \theta)}{q(y_t^*|y_t, y_{t+1}, \theta)} = f\left(y_{t+1}|y_{t,M_t}^{*(j)}, \theta\right) .$$

This method is simple, but it does not exploit $y_{t+1}$ in the design of the importance sampler.

A more efficient importance function can be developed according to the approach of Chib, Greenberg, and Winkelmann (1998) and Shephard and Pitt (1997), which relies on a tailoring procedure that is analogous to that discussed above in connection with the M-H proposal density. Let $\hat{\mu}_t$ be the mode of $\log f(y_{t+1}, y_t^*|y_t, \theta)$ as a function of $y_t^*$ and let $\hat{\Sigma}_t$ be the negative of the inverse of the Hessian of $\log f(y_{t+1}, y_t^*|y_t, \theta)$ evaluated at $\hat{\mu}$. Given $\hat{\mu}_t$ and $\hat{\Sigma}_t$, let the importance density be

$$q(y_t^*|y_t, y_{t+1}, \theta) = f_T(y_t^*|\hat{\mu}_t, \hat{\Sigma}_t, \nu) ,$$

a multivariate-t density with mean $\mu_t$, dispersion $\Sigma_t$ and $\nu$ degrees of freedom. This choice leads to the importance sampling density estimator of the form

$$\begin{aligned} \widetilde{f}^M(y_{t+1}|y_t, \theta) &= \frac{1}{R}\sum_{j=1}^{R} \frac{f\left(y_{t+1}, y_t^{*(j)}|y_t, \theta\right)}{f_T\left(y_t^{*(j)}|\hat{\mu}_t, \hat{\Sigma}_t, \nu\right)}, && (3.9) \\ y_t^{*(j)} &\sim f_T(y_t^*|\hat{\mu}_t, \hat{\Sigma}_t, \nu). && (3.10) \end{aligned}$$

The variance of $\tilde{f}^M$, when it exists, can be estimated by the method provided in Geweke (1989). Note that all $M$ states are integrated out by importance sampling in one sweep. The empirical performance of this method is discussed extensively by Elerian (1999).

### 3.2 *Marginal likelihood*

We now consider the question of comparing alternative, potentially non-nested, diffusion models that have been fit to a given data set. A formal Bayesian approach for making this comparison is through the marginal likelihood of each model, where the marginal likelihood is defined as the integral of the likelihood function with respect to the prior density of the parameters. Ratios of marginal likelihoods are called Bayes factors and these provide the evidence in the data in favor of the numerator model, relative to the model in the denominator. Marginal likelihoods can also be used to compute the posterior probability of each model in the collection. Besides providing information on the relative worth of the various models, these posterior probabilities can be used to find the model averaged Bayesian predictive densities, see Raftery, Madigan, and Volinsky (1994).

Chib (1995) has developed a general method for computing the marginal likelihood based on the output produced from the MCMC simulation. Let $\theta$ denote the parameters of a given diffusion model $\mathcal{M}$, with likelihood function $f^M(Y|y_1, \theta, \mathcal{M})$ and prior density $\pi(\theta|\mathcal{M})$, where the likelihood function is computed using the efficient method just outlined. Then, the Chib method exploits the fact that the marginal likelihood of model $\mathcal{M}$ can be written as

$$m(Y|\mathcal{M}) = \frac{f^M(Y|y_1, \theta, \mathcal{M})\pi(\theta|\mathcal{M})}{\pi(\theta|Y, \mathcal{M})}.$$

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

$$\log m(Y|\mathcal{M}) = \log f^M(Y|y_1, \theta^*, \mathcal{M}) + \log \pi(\theta^*|\mathcal{M}) - \log \hat{\pi}(\theta^*|Y, \mathcal{M}), \tag{3.11}$$

where the first term is the value of the log likelihood function at $\theta^*$, found using (3.9) and the second term is the log of the prior density which is available directly. The third term is estimated from the MCMC output by either kernel smoothing (if the dimension of $\theta$ 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 an important feature of this approach is that it requires only a single evaluation of the likelihood function.

### 3.3 *Diagnostic checks on a fitted model*

Another issue that we address is the fit of a given model by judging how well the predictions of the model accord with the data. We carry this out via one-step-ahead prediction distribution functions.

Consider the conditional predictive distribution function

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

We now show that this is uniform $UID(0, 1)$ under the correctness of the model. Clearly $u_{t+1} =$

$G(y_{t+1}|y_t,\theta)$ lies in the interval $(0,1)$. Next,

$$\begin{aligned} \Pr(u_{t+1} &\leq a|y_t,\theta) = \Pr\left\{G(y_{t+1}|y_t,\theta) \leq a|y_t,\theta\right\} \\ &= \Pr\left\{y_{t+1} \leq G^{-1}(a)|y_t,\theta\right\} \\ &= a, \end{aligned}$$

which shows that $u_{t+1}$ is uniform. Finally, since $u_{t+1}$ does not depend on $y_t$ it is also an independent sequence. This result, in a different context, goes back at least to Rosenblatt (1952).

We approximate $G(y_{t+1}|y_t,\theta)$ by

$$F^M\left(y_{t+1}|y_t,\theta\right) = \int \Pr(Y_{t+1} \leq y_{t+1}|y_{t,M}^*,\theta) f^M(y_t^*|y_t,\theta)dy_t^*,$$

where $\Pr(Y_{t+1} \leq y_{t+1}|y_{t,M}^*,\theta)$ is easily computed due to the fact that $Y_{t+1}|y_{t,M}^*,\theta$ is approximately Gaussian. This integral can now be estimated by Monte Carlo by drawing $y_t^{*(j)} \sim f^M(y_t^*|y_t,\theta)$ a large number of times and forming the average:

$$\hat{u}_{t+1} = \widehat{F}^M\left(y_{t+1}|y_t,\theta\right) = \frac{1}{R}\sum_{j=1}^{R} \Pr\left(Y_{t+1} \leq y_{t+1}|y_{t,M}^{*(j)},\theta\right).$$

The direct Monte Carlo estimate can be improved by importance sampling as was suggested for the estimation of the likelihood ordinate above. Again the results of Talay and Tubaro (1990) imply that the error in estimating $G(y_{t+1}|y_t,\theta)$ is again of an order smaller than or equal to the product of $1/M$ and $1/\sqrt{R}$. Adequacy of the model is judged by the serial correlation in the $\{\hat{u}_{t+1}\}$ and by the nature of the distributional shape of both the $\{\hat{u}_{t+1}\}$ and its reflected version $\left\{2\left|\hat{u}_{t+1} - \frac{1}{2}\right|\right\}$, the latter providing information on dispersion. These statistics are also used in, for example, Pedersen (1994) and Smith (1985). The idea of focusing on $2\left|\hat{u}_t - 0.5\right|$ appears in Kim, Shephard, and Chib (1998).[10] As a by-product, we compute the standardized forecast errors (see Pedersen (1994)) as

$$\frac{y_{t+1} - \widehat{\mathrm{E}}(y_{t+1}|y_t,\theta)}{\sqrt{\widehat{\mathrm{var}}(y_{t+1}|y_t,\theta)}},$$

where $\widehat{\mathrm{E}}(y_{t+1}|y_t,\theta) = \frac{1}{R}\sum_{j=1}^{R} Y_{t+1}^{(j)}$ and $\widehat{\mathrm{var}}(y_{t+1}|y_t,\theta) = \frac{1}{R-1}\sum_{j=1}^{R}\{Y_{t+1}^{(j)} - \widehat{\mathrm{E}}(y_{t+1}|y_t,\theta)\}^2$. Properties of these statistics are harder to evaluate than for the $\hat{u}_{t+1}$, but provide a graphical aid to help explain the inadequacies of the fitted model. Application of these methods is presented in Section 5.

[10] In practice we replace $\theta$ by some estimator of the parameter — usually the posterior mean. This leads to another layer of approximation. Alternatively, one could sequentially integrate out the effect of $\theta$ and compute $G(y_{t+1}|y_t)$. Gerlach, Carter, and Kohn (1999) provide methods for doing this in certain time series models but in our context such computations are quite burdensome.

# 4 SIMULATED DATA EXAMPLE

## 4.1 *CIR model*

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

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

Similar detailed calculations for the OU, Quadratic drift and Hull-White models are given in the Appendix of Elerian, Chib, and Shephard (1998). 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 strong solution, (which has a non-central chi-squared distribution), using $\Delta^\dagger = 1$ for daily series and $\Delta^\dagger = 5$ for weekly data. Here $\Delta^\dagger$ can be thought of as the time interval in the observed data and determines the bias in the discretized time gap.

In order to simulate from $\pi^M(\theta, Y^*|Y, y_1)$, we implement the M-H sampler for $\pi^M(Y^*|Y, y_1, \theta)$ as discussed in Section 2. We can exploit the special structure of the CIR model to efficiently update our samples from $f(\theta|Y^*, Y)$ by sampling from $f(\alpha, \beta|Y^*, Y, \sigma^2)$ and then from $f(\sigma^2|Y^*, Y, \alpha, \beta)$. The implementation is discussed in the next two subsections.

## 4.2 *Full conditional density of the drift parameters*

Let $\psi = (\alpha, \beta)'$ and suppose that $\psi$ is given a bivariate normal prior distribution with mean $\gamma_0$ and variance $\Gamma_0^{-1}$. Then a simple calculation shows that the full conditional density given the observations and augmented data, $z = (z_1, \dots, z_n) = (y_1, y^*_{1,1}, \dots, y^*_{1,M}, y_2, \dots, y^*_{T-1,M}, y_T)$ 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

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

### 4.3 *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 $\log f(\sigma^2|z,\alpha,\beta)$, which we can denote by $g(\sigma^2|z,\alpha,\beta)$, (where $f \leq Cg$ for $0 < C < \infty$), and hence obtain the posterior using an Accept/Reject technique (Ripley 1987, pp. 60-1). We generate a candidate value, $\hat{\sigma}^2$, from $g(\sigma^2|z,\alpha,\beta)$ and a uniform random number $U$ on the interval 0 to 1. If $U \leq f(\sigma^2|z,\alpha,\beta)/\{Cg(\sigma^2|z,\alpha,\beta)\}$, the value $\sigma^2 = \hat{\sigma}^2$ is returned. If the value is rejected, another candidate value is drawn and the algorithm is repeated.

Let $\sigma^2$ apriori follow the Inverse Gamma $(\frac{p}{2}, \frac{1}{2\,S_0\,p})$ distribution, 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 $d_{t-1} = \log z_t - \log z_{t-1} - \frac{\alpha\Delta}{z_{t-1}} + \beta\Delta$. Let $\lambda = \log\sigma^2$, then

$$\begin{aligned} \log f(\lambda|z,\alpha,\,\beta) &= \log f(\sigma^2|z,\alpha,\,\beta) + \log\left|\frac{\partial\sigma^2}{\partial\lambda}\right| \\ &= c' - (\xi_1 - 1)\lambda - \xi_2\exp(-\lambda) + p(\lambda)\,, \end{aligned} \tag{4.12}$$

where $\xi_2 > 0$, and $p(\lambda) = S_1\exp(\lambda)$ is concave in $\lambda$, that is $p''(\lambda) = p'(\lambda) = p(\lambda)$ is negative for all $\lambda$, and $p'(\lambda)$ and $p''(\lambda)$ are the first and second derivatives of $p(\lambda)$ with respect to $\lambda$. Then we can sample from $\lambda|z,\alpha,\,\beta$ by making suggestions

$$\log g(\lambda|z,\alpha,\beta) = \exp(\lambda) \sim \text{ Inverse Gamma}\{\xi_1 - p'(\hat{\lambda}) - 1, \xi_2\}, \tag{4.13}$$

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

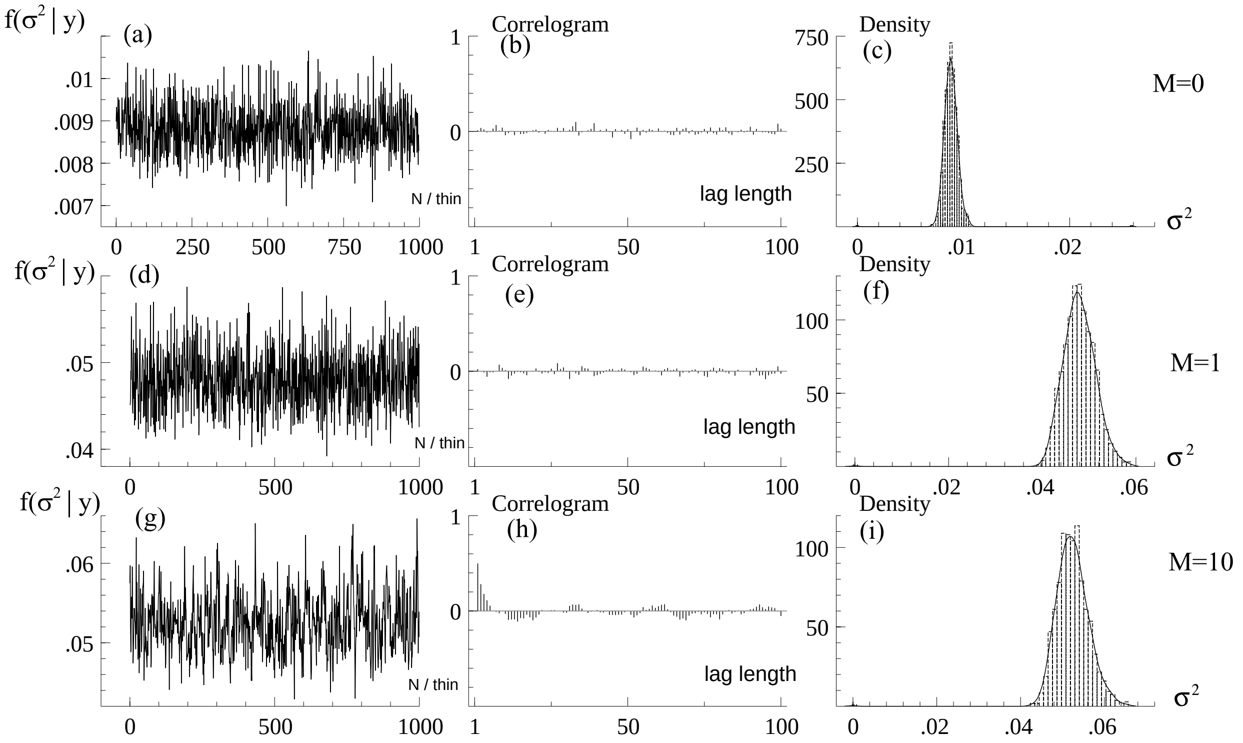

**Figure** 6: *Paths (a, d, g), correlograms (b, e, h) and histograms (c, f, i) for draws from* $\sigma^2|y$, *(true value 0.05). Top graph uses discretization* $M = 0$, *middle graphs have* $M = 1$ *and bottom graphs have* $M = 10$, $\lambda = 3$. $N = 10,000$ *with thin*$= 10$ *and* $\Delta^\dagger = 1$ *in all simulations.*

### 4.4 *MCMC output analysis*

We now apply our 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 6 we show the sample path, autocorrelation function and histogram based on the MCMC output of $\sigma^2$. The output for $\alpha$ and $\beta$ displays similar features. In the top graphs $M = 0$, in the middle graphs $M = 1$ and in the bottom graphs $M = 10$. Figure 7 corresponds to the case $\Delta^\dagger = 5$, where the bias is more pronounced. 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 correspond to the case $\Delta^\dagger = 5$), are reported in Table II. The MCMC algorithm is now run with $M = 0, 1, 10$ ($\lambda = 3$ ), 20 ($\lambda = 5$) and 30 ($\lambda = 9$) for $N = 10,000$ iterations. The inefficiency factors are small for $M < 20$. When $M = 20$, the MCMC sampler is approximately fourteen times less efficient than a hypothetical algorithm that produces independent draws. The inefficiency factor increases to twenty when

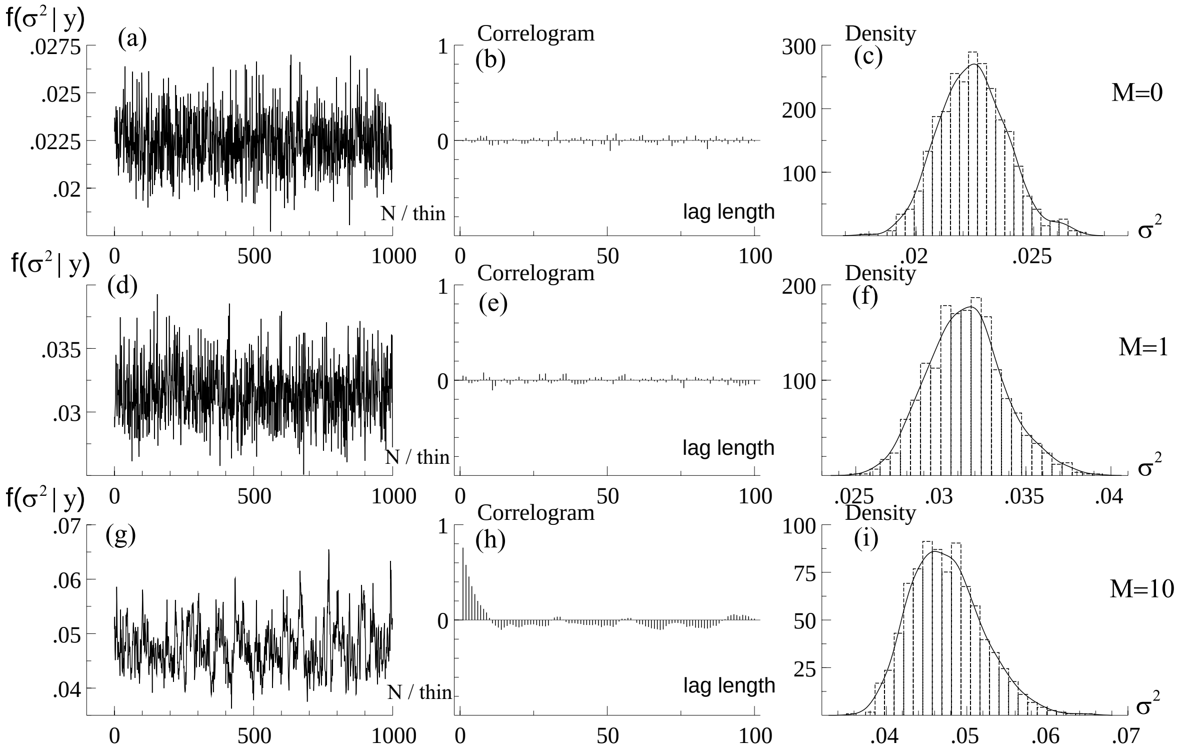

**Figure** 7: *Paths (a, d, g), correlograms (b, e, h) and histograms (c, f, i) for draws from $\sigma^2|y$, (true value 0.05). Top graph uses discretization $M = 0$, middle graphs have $M = 1$ and bottom graphs have $M = 10$, $\lambda = 3$. $N = 10,000$ with thin$= 10$ and $\Delta^\dagger = 5$ in all simulations.*

$M = 30$. The autocorrelation plots in Figure 6 reinforce this point, dying down more quickly for smaller values of $M$. There is strong correlation between $\alpha$ and $\beta$ (0.99) regardless of $M$, while the correlation between $\sigma^2$ and $\alpha, \beta$ initially increases with $M$, fluctuating around 0.75 and 0.8 for higher values of $M$. Table III shows the results from a small simulation study for the parameters of the CIR process. Ten data sets of length $T = 500$ and $\Delta^\dagger = 5$ were generated using the strong solution and the M-H algorithm was run with $M = 0, 1, 10, 20$ and 30 for $N = 10,000$ iterations. In general, $N = 10,000$ iterations were used since the simulation inefficiency for the largest $M$ hovered around 20. On the basis of the Monte Carlo results reported in Table III, 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 $M > 20$ ). The MSE for $\sigma^2$ is smallest for $M = 20$.[11] Additional simulation experiments, reported elsewhere, show

[11] To monitor the convergence of each parameter, we can additionally compute the Gelman-Rubin r statistic: see Gelman and Rubin (1992) and Gelman, Carlin, Stern, and Rubin (1995, pp. 331-332). Suppose we denote one of the parameters of $\theta$ as $\phi$ and run $J$ parallel sequences of the M-H algorithm with overdispersed starting values. Each

that an increase in the DGP drift parameter values, holding the volatility parameter constant, is associated with an small increase in the bias, MSE and inefficiency factors for all three parameters. Interestingly, increasing the volatility parameter, while keeping the drift parameters fixed, has the opposite effect. The computer time to run 100 iterations of the sampler, with $T = 500$ observations, is 66.08 seconds for $M = 3$ ($\lambda = 1$), 125.82 seconds for $M = 10$ ($\lambda = 3$) and 245.58 seconds for $M = 30$ ($\lambda = 5$), using the matrix language Ox developed 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). The data, which consists of daily observations over the period 1st of June, 1973 to 25th of February, 1995, has also been considered by Aït-Sahalia (1996b).[12] The model is specified by the diffusion functions

$$\begin{aligned} a(y,\theta) &= \alpha_0 + \alpha_1 y + \alpha_2 y^2 + \alpha_3 y^{-1} \\ b^2(y,\theta) &= \beta_0 + \beta_1 y + \beta_2 y^{\beta_3}, \end{aligned} \tag{5.14}$$

sequence is run for $N$ iterations, to obtain draws $\phi_{ij}$, $(i = 1,\ldots,N; j = 1,\ldots,J)$. The *between*, $B$ and *inbetween*, $W$, sequence variances can be computed as

$$\begin{aligned} B &= \frac{N}{J-1}\sum_{j=1}^{J}(\bar{\phi}_{.j} - \bar{\phi}_{..})^2, \quad \text{where } \bar{\phi}_{.j} = \frac{1}{N}\sum_{i=1}^{N}\phi_{ij}, \ \ \bar{\phi}_{..} = \frac{1}{J}\sum_{j=1}^{J}\bar{\phi}_{.j}, \\ W &= \frac{1}{J}\sum_{j=1}^{J}s_j^2, \quad \text{where } s_j^2 = \frac{1}{N}\sum_{i=1}^{N}(\phi_{ij} - \bar{\phi}_{.j})^2. \end{aligned}$$

The estimate $\text{var}(\phi|Y)$ is then given by a weighted average of $B$ and $W$ as

$$\text{var}^+(\phi|Y) = \frac{N-1}{N}W + \frac{1}{N}B,$$

which is in fact an overestimate of the variance of the marginal posterior density. The convergence can be assessed through $\sqrt{\hat{R}}$, where

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{\text{var}}^+(\phi|Y)}{W}} \overset{N\to\infty}{\to} 1.$$

For the parameters $(\alpha,\beta,\sigma^2)$ of the CIR model in one of the experiments, the corresponding $\sqrt{\hat{R}}$ statistic is given by $(1.0001, 1.0001, 1.0013)$. The M-H algorithm was run using $M = 10$ ($\lambda = 3$), $N = 5000$ and a burn-in of 500 simulations.

[12] The data can be downloaded from http://www.princeton.edu/ yacine/research/.

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

$$\begin{aligned}
&\beta_0 \geq 0 \text{ (and } \beta_2 > 0 \text{ if } \beta_0 = 0 \text{ and } 0 < \beta_3 < 1, \text{ or } \beta_1 > 0 \text{ if } \beta_0 = 0 \text{ and } \beta_3 > 1) \\
&\beta_2 > 0 \text{ if either } \beta_3 > 1 \text{ or } \beta_1 = 0, \text{ and } \beta_1 > 0 \text{ if either } 0 < \beta_3 < 1 \text{ or } \beta_2 = 0 \\
&\alpha_2 \leq 0 \text{ and } \alpha_1 < 0 \text{ if } \alpha_2 = 0 \\
&\alpha_3 > 0 \text{ and } 2\alpha_3 \geq \beta_0 \geq 0, \text{ or } \alpha_3 = 0, \alpha_0 > 0, \ \beta_0 = 0, \ \beta_3 > 1 \text{ and } 2\alpha_0 \geq \beta_1 > 0.
\end{aligned}$$

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 fourth ensures that the values are positive. Note that instead of letting $\beta_3$ 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 IV. These alternative specificatons are considered because the above model is practically unidentified for these data.

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 Table V, 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.[13] The results are based on $M = 10$ ( $\lambda = 3$) using $N = 2000$ MCMC iterations for the Vasicek (shown) and CIR models and $N = 5000$ for the Affine CIR model. For each of these models, the output paths, correlograms and histograms of $\theta = (\alpha_0, \alpha_1, \beta_1)'$ are given in Figures 8, 9 and 10. We selected $B_N = 100$ (400) in the Parzen computations of the Monte Carlo standard errors for $\alpha$ ($\beta$). We set standard diffuse normal priors on the alpha coefficients and diffuse inverse Gamma priors on $\beta$. In the Affine CIR and general parametric model, however, $\beta_1$ is given a diffuse normal prior.

It will be seen that for each model the correlogram of $\alpha$ dissipates quickly, though that of $\beta_1$ in the Affine CIR model shows persistence up to lag 100. In all models, the posterior mean of $\alpha_0$ is positive and that of $\alpha_1$ is negative. The results on $\beta_0$ in the Vasicek model and those of $\beta_1$ in the CIR model are consistent with those in the Affine CIR model. The Figures show that $\beta_0$ is close to zero and that the variance is mainly determined by $\beta_1 y$. The $\alpha$ and $\beta$ parameters are close to

[13] The $\alpha$% HPD region represents the shortest interval that contains $\alpha$% of the points of highest posterior density, for example, see Besag, Green, Higdon, and Mengersen (1995).

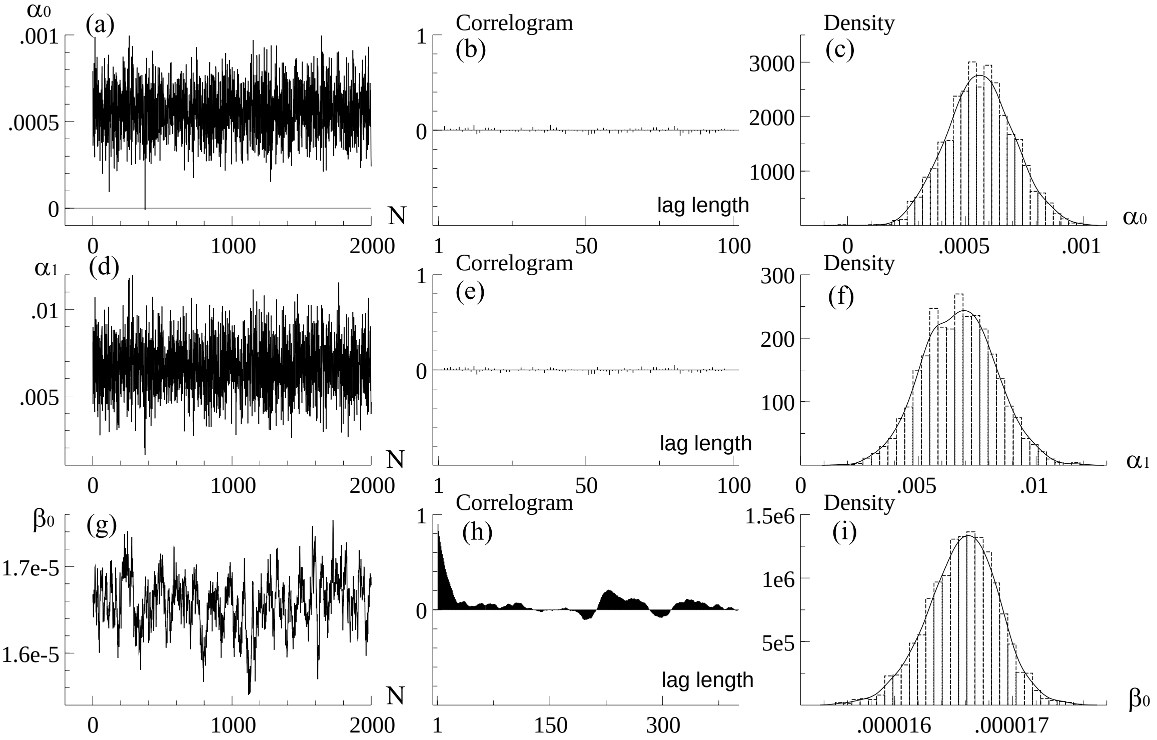

**Figure** 8: *Paths (a, d, g), correlograms (b, e, h) and histograms (c, f, i) for $\alpha_0$, $\alpha_1$ and $\beta_0$ for the Vasicek process fitting the Eurodollar short-rate. The algorithm was run with $M = 10$ and $\lambda = 3$.*

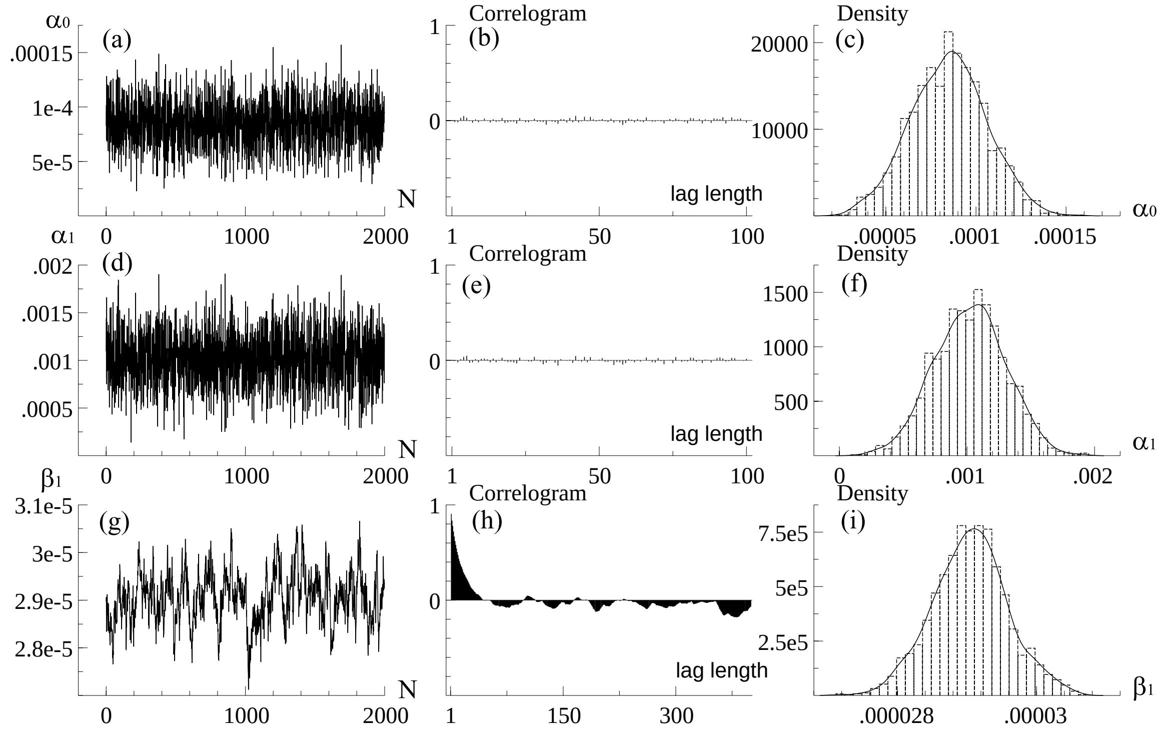

**Figure** 9: *Paths (a, d, g), correlograms (b, e, h) and histograms (c, f, i) for $\alpha_0$, $\alpha_1$ and $\beta_1$ for the CIR process fitting the Eurodollar short-rate. The algorithm was run with $M = 10$ and $\lambda = 3$.*

being uncorrelated. One point to note is that the inefficiency factors are low for $\alpha$. For $\beta$, we are up to twenty five times less efficient.

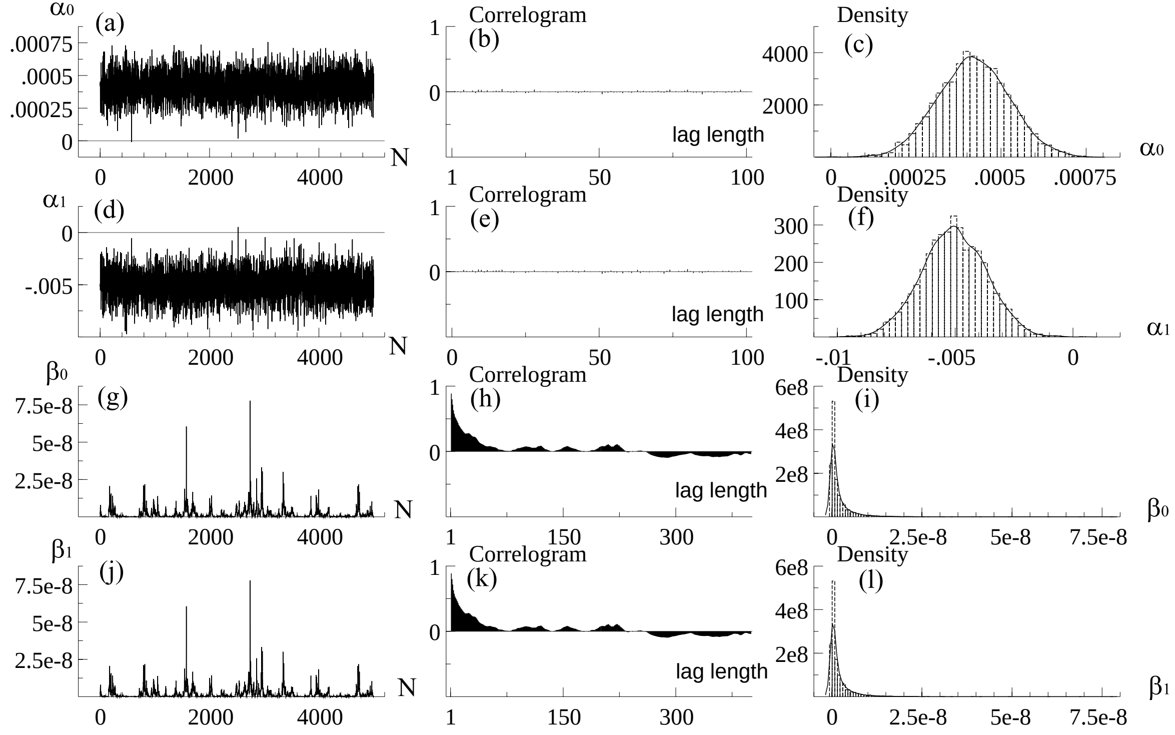

**Figure** 10: *Paths (a, d, g, j), correlograms (b, e, h, k) and histograms (c, f, i, l) for $\alpha_0$ $\alpha_1$, $\beta_0$ and $\beta_1$ for the Affine CIR process fitting the Eurodollar short-rate. The algorithm was run with $M = 10$ and $\lambda = 3$.*

In Figure 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 12 with the corresponding QQ plots. The residuals and their reflected versions are computed using the one-step-ahead prediction distribution. The standardized forecast errors show that the extremes in the data are not being picked up for a large part of the data. This is further confirmed by the correlograms which show remaining structure not being accounted for. The lack of fit is also highlighted by the amount of activity present, which occurs in concentrated clumps of observations; an indication that these Markov models do not account for the volatility clustering present in the series.

We now turn our attention to the general parametric model given by equation 5.14. We first show that $\beta_2$ and $\beta_3$ are difficult to identify separately from our data. In Figure 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 $(\beta_2, \beta_3)$ : $(4.511 \times 10^{-4}, 1.5)$, $(6.9338 \times 10^{-4}, 1.7)$, $(7.021 \times 10^{-4}, 1.8)$, $(1.312 \times 10^{-3}, 2)$

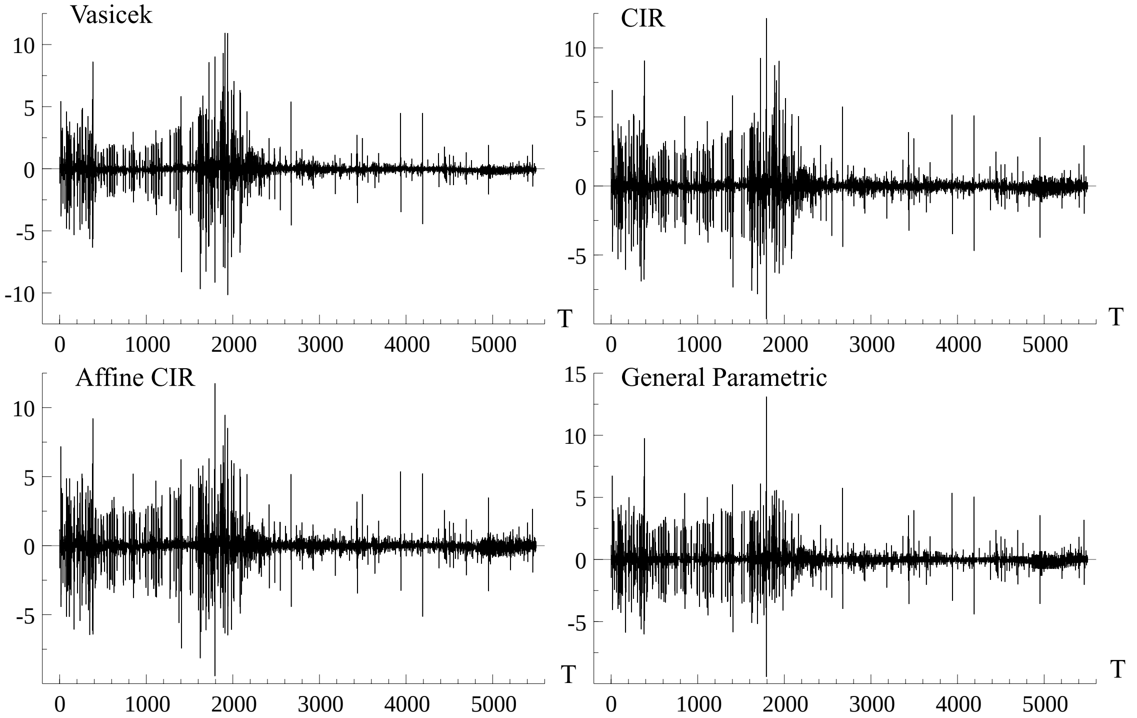

**Figure** 11: *Standardized one-step-ahead forecast errors for the Vasicek, CIR, Affine CIR and general parametric models. We graph* $y_{t+1} - \hat{E}(y_{t+1}|y_t, \theta)/\sqrt{var(y_{t+1}|y_t, \theta)}$*, taking* $\theta = E(\theta|y)$.

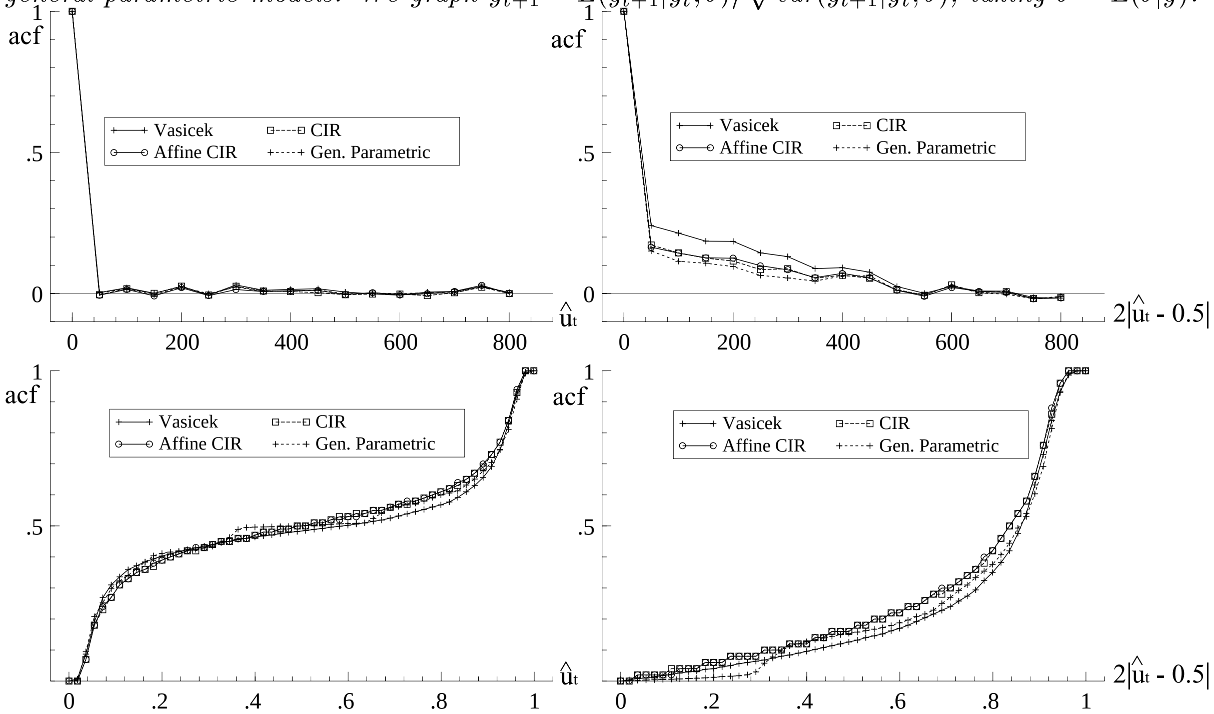

**Figure** 12: *Top graphs: Autocorrelation functions of residuals,* $\{\hat{u}_t\}$ *(left panel) and reflected residuals,* $\{2|\hat{u}_t - 0.5|\}$ *(right panel) for the Vasicek, CIR, Affine CIR and general parametric process. Bottom graphs: The corresponding QQ plots against observations.*

and $(1.5177 \times 10^{-3}, 2.1)$. These were obtained from the MCMC algorithm fixing a value of $\beta_3$ and evaluating the posterior mean of $\beta_2$. 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 $(\beta_2, \beta_3)$ values produce virtually 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 13. Results are shown in Table VI. Similar diagnostics were obtained on setting $\beta_3 = 0$. Standardized forecast errors in Figure 11 show that the extremes in the data are not being picked up, and the autocorrelation functions in Figure 12 confirm that the models have not provided a good fit to the data.

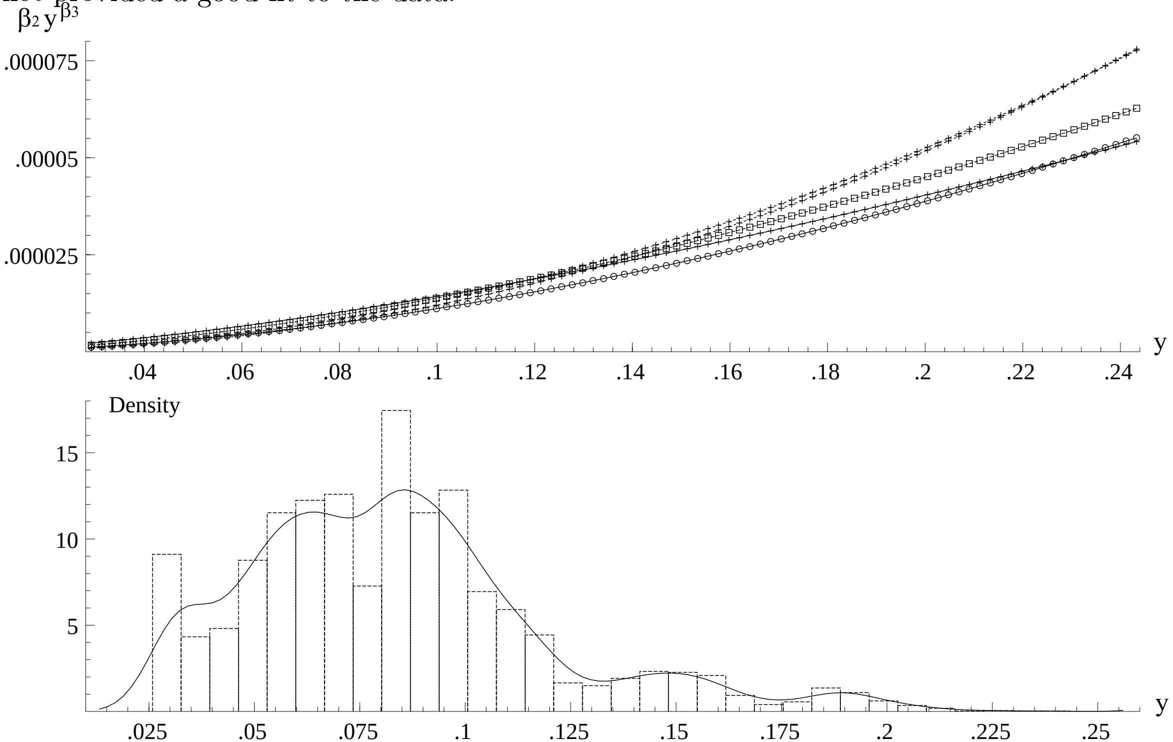

**Figure** 13: *Top graph plots $\beta_2 y^{\beta_3}$ against a range of values for y between the minimum and maximum values observed from the data, for five pairs of values of $(\beta_2, \beta_3)$. Bottom graph plots the kernel estimate of the density using the interest data.*

To complete the analysis we also compute the marginal likelihoods of these various models. The required likelihood ordinate is computed by importance sampling, as discussed earlier, using $M = 10$ latent values between every two data points and the posterior ordinate is found by the method of kernel smoothing. Both these quantities and the prior ordinate are evaluated at the posterior mean of the parameters. The resulting log marginal likelihood estimates are shown in Table VII. 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 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 full Bayesian 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 by tuned Markov chain Monte Carlo simulation methods. We have proposed efficient ways of summarizing the posterior distributions in these problems and provided methods for finding the model marginal likelihood (to compare alternative stochastic differential equations) and 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 sample the latent data is required. This feature is likely to be particularly helpful in the 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, the results are largely determined by information in the likelihood function and not the prior given the sample sizes that are encountered in this area.

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

*Nuffield College, Oxford, OX1 1NF, UK; ola.elerian@nuf.ox.ac.uk,*
*http://www.nuff.ox.ac.uk/Users/Elerian/*

and

*John M. Olin School of Business, Washington University, St. Louis, MO 63130, USA; chib@olin.wustl.edu, http://www.olin.wustl.edu/faculty/chib/*

and

*Nuffield College, Oxford, OX1 1NF, UK; neil.shephard@nuf.ox.ac.uk, http://www.nuff.ox.ac.uk/economics/people/shephard.htm*

## References

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

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

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.

Bally, V., and D. Talay (1995): "The law of the Euler scheme for stochastic differential equations: error analysis with Malliavin calculus," *Mathematics and Computers in Simulation*, 38, 35–41.

Barndorff-Nielsen, O. E., J. L. Jensen, and M. Sørensen (1998): "Some stationary processes in discrete and continuous time," *Advances in Applied Probability*, 30, 989–1007.

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

Billio, M., A. Monfort, and C. P. Robert (1998): "The simulated likelhood ratio method," Doc. Travail Crest, INSEE, Paris.

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

——— (2000): "Markov chain Monte Carlo methods: Computation and Inference," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. Leamer, vol. 5. North-Holland, Amsterdam, Forthcoming.

Chib, S., and E. Greenberg (1994): "Bayes inference in regression models with ARMA $(p, q)$ errors," *Journal of Econometrics*, 64, 183–206.

——— (1995): "Understanding the Metropolis–Hastings algorithm," *The American Statistician*, 49, 327–336.

Chib, S., E. Greenberg, and R. Winkelmann (1998): "Posterior simulation and Bayes factors in panel count data models," *Journal of Econometrics*, 86, 33–54.

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

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

Elerian, O. (1998): "A note on the existence of a closed-form conditional transition density for the Milstein scheme," Economics discussion paper 1998-W18, Nuffield College, Oxford.

——— (1999): "Simulation estimation of continous time series models with applications to finance," D.Phil, Nuffield College, Oxford.

Elerian, O., S. Chib, and N. Shephard (1998): "Likelihood inference for discretely observed non-linear diffusions," Economics discussion paper 146, Nuffield College, Oxford.

Eraker, B. (1998): "Markov Chain Monte Carlo Analysis of Diffusion Models with Application to Finance," HAE thesis, Norwegian School of Economics and Business Administration.

Florens-Zmirnou, 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.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995): *Bayesian Data Analysis.* Chapman and Hall, London.

Gelman, A., and D. B. Rubin (1992): "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7(4), 457–472.

GERLACH, R., C. CARTER, AND R. KOHN (1999): "Diagnostics for time series analysis," *Journal of Time Series Analysis*, 21, 309–330.

GEWEKE, J. (1989): "Bayesian inference in econometric models using Monte Carlo integration," *Econometrica*, 57, 1317–1339.

GEYER, C. J. (1999): "Likelihood inference for spatial point processes," in *Current Trends in Stochastic Geometry and Applications*, ed. by O. E. Barndorff-Nielsen, W. S. Kendall, and M. N. M. Lieshout, pp. 141–172. Chapman & Hall, London.

GILKS, W. R., S. RICHARDSON, AND D. J. SPIEGELHALTER (1996): *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

GOURIEROUX, C., A. MONFORT, AND E. RENAULT (1993): "Indirect inference," *Journal of Applied Econometrics*, 3, S85–S118.

HANSEN, L., AND J. A. SCHEINKMAN (1995): "Back to the future: generating moment implications for continuous-time Markov processes," *Econometrica*, 63, 767–804.

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., AND M. SØRENSEN (1999): "Estimating equations based on eigenfunctions for a discretely observed diffusion process," *Bernoulli*, 5, 299–314.

KIM, S., N. SHEPHARD, AND S. CHIB (1998): "Stochastic volatility: likelihood inference and comparison with ARCH models," *Review of Economic Studies*, 65, 361–393.

KLOEDEN, P. E., AND E. PLATEN (1992): *Numerical Solutions to Stochastic Differential Equations*. Springer, New York.

KLOEK, T., AND H. K. VAN DIJK (1978): "Bayesian estimates of equation system parameters: an application on integration by Monte Carlo," *Econometrica*, 46, 1–20.

KOHATSU-HIGA, A., AND S. OGAWA (1997): "Weak rate of convergence for an Euler scheme of nonlinear SDE's," *Monte Carlo Methods and Applications*, 3, 327–345.

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

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

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

——— (1995): "A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations," *Scandinavian Journal of Statistics*, 22, 55–71.

Protter, P., and D. Talay (1997): "The Euler scheme for Lévy stochastic differential equations," *Annals of Probability*, 25, 393–423.

Raftery, A. E., D. Madigan, and C. T. Volinsky (1994): "Accounting for model uncertainty in survival analysis improves predictive performance," in *Bayesian Statistics*, ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, vol. 5, pp. 323–349. Oxford University Press, New York.

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

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

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

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

Sørensen, M. (1997): "Estimating functions for discretely observed diffusions: a review," in *Selected Proceedings of the Symposium on Estimating Functions*, ed. by I. V. Basawa, V. P. Godambe, and R. L. Taylor, Monograph Series, pp. 305–325. IMS Lecture Notes.

Talay, D. (1995): "Simulation and numerical analysis of stochastic differential systems: a review," in *Probabilistic Methods in Applied Physics*, ed. by P. Krée, and W. Wedig, vol. 451, chap. 3, pp. 54–96. Springer-Verlag, Berlin.

Talay, D., and L. Tubaro (1990): "Expansion of the global error for numerical schemes solving stochastic differential equations," *Stochastic Analysis and Applications*, 8, 94–120.

TAUCHEN, G. E. (1997): "New minimum chi-squared methods in empirical finance," in *Advances in Economics and Econometrics: Theory and Applications, 7th World Congress*, ed. by D. M. Kreps, and K. F. Wallis, vol. 3, chap. 9, pp. 279–317. Cambridge University Press, Cambridge.

WONG, C. K. J. (1999): "New methods for performing efficient inference for linear and log-normal volatility models," M.Phil. thesis, University of Oxford.

**Table** II: *The posterior means, inefficiency factors, (bandwidth* $B_N = 100$*), covariances and correlations 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 M-H algorithm was run with* $M = 0, 1, 10$ *(*$\lambda = 3$*),* $20$ *(*$\lambda = 5$*) and* $30$ *(*$\lambda = 9$*) for* $N = 10,000$ *iterations.*

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

**Table** III: *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 M-H algorithm was run with* $M = 0, 1, 10, 20$ *and* 30 *for* $N = 10{,}000$ *iterations.*

| **Monte Carlo Results (CIR process)** | | | | | |
|---|---|---|---|---|---|
| | Mean | Bias | s.d. | Mean(s.e) | MSE(×100) |
| $\alpha$ (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 |
| $\beta$ (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 |
| $\sigma$ (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** IV: *Models to be considered for the short-rate process: Vasicek, CIR, Affine CIR, Aït-Sahalia (1996b) model, and general parametric model.*

| **Model** | **Drift function** $a(y,\theta)$ | **Volatility function** $b^2(y,\theta)$ |
|---|---|---|
| Vasicek | $\alpha_0 + \alpha_1 y$ | $\beta_0$ |
| 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** V: *The posterior means, HPD, Monte Carlo standard errors, inefficiency factors (Ineff, bandwidth* $B_N = 100$*), covariances and correlations (in italics) are shown for the parameters using the Eurodollar short-rate data applied to the Vasicek process. The M-H algorithm was run with* $M = 10$ *(*$\lambda = 3$*), for* $N = 2000$ *iterations.*

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

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

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

**Table** VII: *The log marginal likelihood computations are shown for* $x = \log y$*.* $f(x|\theta^*)$ *and* $\pi(\theta^*|x)$ *are computed using* $M = 10$*, burn-in=200 and* $N = 2000$*. Priors used are Inverse Gamma* $\left\{\frac{\theta^*}{(15\sigma^*)^2} + 2, \theta^*(\sigma^* - 1)\right\}$ *for* $\beta$ *and Normal* $\{\theta^* + 4\sigma^*,\ (15\sigma^*)^2\}$ *for* $\alpha$*, (*$\beta_1$ *uses normal specification in general parametric model), where* $\theta^*$ *and* $\sigma^*$ *are the respective posterior means and standard deviations, based on a training sample data set. The numerical standard error of the marginal likelihood (log scale) is denoted by (numerical se).*

| **Model** | $\log m(x\|\mathcal{M})$ | **Numerical se** |
|---|---|---|
| Vasicek | 8328.9 | 0.52857 |
| CIR | 9430.5 | 0.25093 |
| Affine CIR | 9436.9 | 0.26587 |
| General parametric model | 9716.8 | 0.60919 |