A note on the existence of a closed form conditional transition density for the Milstein scheme

OLA ELERIAN

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

ola.elerian@nuf.ox.ac.uk

October 29, 1998

Abstract

This paper is concerned with the estimation of stochastic differential equations when only discrete observations are available. It primarily focuses on deriving a closed form solution for the one-step ahead conditional transition density using the Milstein scheme. This higher order Taylor approximation enables us to obtain an order of improvement in accuracy in estimating the parameters in a non-linear diffusion, as compared to use of the Euler-Maruyama discretization scheme. Examples using simulated data are presented. The method can easily be extended to the situation where auxiliary points are introduced between the observed values. The Milstein scheme can be used to obtain the approximate transition density as in a Pedersen (1995) type of simulated likelihood method or within an MCMC method as proposed in Elerian, Chib, and Shephard (1998).

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

1 Introduction

Stochastic differential equations (SDE's) are used extensively in economics to model many types of data; see, for example, Dixit (1993) , Dixit and Pindyck (1994) and Merton (1990). The SDE, (which is assumed to satisfy certain regularity conditions, formalized in Section 2.1) takes the form

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

where $a(\cdot)$ is the drift function, $b^2(\cdot)$ is the volatility function, $W(t)$ is the Wiener process and $y(t)$ are the observations indexed by time. The parameter θ is the quantity that we are interested in estimating. Assuming that the assumptions under which the diffusion $y(t)$ exists are satisfied (see Øksendal, 1995, pg. 64), we will assume 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 $t = 1, \ldots, T$.

Using discrete observations when the underlying model is assumed to be continuous can lead to considerable bias in the estimation of the parameters. In some cases, exact maximum likelihood estimation is possible as the stochastic differential equation has a strong solution and hence the true transition density, $g(y_{t+1}|y_t)$ is of known closed form. However, in general this method is unfeasible. One can therefore approximate the SDE by using a discretization which implies the approximate transition density $f(y_{t+1}|y_t)$.

The aim is to estimate the parameters θ of the process, given the discrete measurements $Y = (y_1, \ldots, y_T)'$. There has been a 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), Elerian, Chib, and Shephard (1998) and Eraker (1998). Aït-Sahalia (1998) tries to circumvent the discrete time approximation to obtain an expression for the likelihood. Discretely observed diffusions have also been fitted using estimating functions (see Bibby and Sørensen 1995, Kessler and Sørensen 1995, Basawa et al. 1997, Dacunha-Castelle and Florens-Zmirou 1986, Florens-Zmirou 1989, Genon-Catalot and Jacod 1993, Kessler 1997 and Hansen and Scheinkman 1995).

Methods based on approximating the diffusion with a stochastic Taylor scheme use the Euler-Maruyama approximation in obtaining an expression for the one-step ahead transition density, due to its analytical tractibilty. The discretization implied by the Euler scheme is,

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

Taking the increments of W_t to be normal with mean zero and variance $\Delta^\dagger$ implies a normal transition density for $f(y_{t+1}|y_t; \theta)$. This paper makes use of the approximation proposed by Milstein (1978), which takes the form

$$y_{t+1} = y_t + a(y_t, t; \theta)\Delta^\dagger + b(y_t, t; \theta)(W_{t+1} - W_t) + \frac{1}{2}b(y_t, t; \theta)\frac{db(y_t, t; \theta)}{dy_t}\{(W_{t+1} - W_t)^2 - \Delta^\dagger\},$$

the Euler plus an additional term. This implies a non-central chi-squared distribution for the approximate transition density. The one step ahead conditional transition density will be shown to have an analytical form. Use of the Milstein in this paper provides an order of improvement on earlier results obtained when maximizing the likelihood using the Euler scheme.

These can naturally be extended to the case when auxiliary points are placed between the observations. This involves introducing, say $M_t = M$ auxiliary points between each pair of observations so that we have

$$\begin{aligned} y^*_{t,j+1} &= y^*_{t,j} + a(y^*_{t,j}, t; \theta)\Delta^\dagger + b(y^*_{t,j}, t; \theta)(W_{t,j+1} - W_{t,j}) \\ &\quad + \frac{1}{2}b(y^*_{t,j}, t; \theta)\frac{db(y^*_{t,j}, t; \theta)}{dy^*_{t,j}}\{(W_{t,j+1} - W_{t,j})^2 - \Delta^\dagger\}, \end{aligned}$$

where $y^*_{t,j}$ represent the auxiliary points introduced between the observations $y_{t,0} = y_t$ and $y_{t,M+1} = y_{t+1}$. Introduction of the auxiliary points can reduce bias considerably. One can then define the one step ahead transition density as

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

where we use the notation $y^*_t = \{y^*_{t,1}, \ldots, y^*_{t,M}\}$. The Milstein densities are thus integrated out with repect to the auxiliary points. We can therefore perform approximate likelihood inference based upon $f^M(y_{t+1}|y_t; \theta)$, either by using maximum likelihood principles or Bayesian methods. In the former case, the likelihood will be estimated by simulation; the logarithm of the result is then maximized with respect to θ. This type of method is usually called a *maximum simulated likelihood* technique. The Bayesian method will be based on Markov chain Monte Carlo (MCMC) methods, simulating from augmentation and the parameters as in Elerian,

Chib, and Shephard (1998). Though both schemes currently reduce the bias in parameter estimates, as compared to methods that do not introduce auxiliary points, they rely on the Euler approximation. Use of the Milstein scheme in this context, therefore provides a better approximation to the true transition density and hence the parameter estimates.

The outline of the paper is as follows; Section 2 describes the Milstein approximation giving a closed form expression for the density. Order of convergence is discussed and compared to the Euler approximation. Section 3.1 presents the likelihood based on the Milstein density and discusses the bias and inconsistency in the log-likelihood inherent in its use without the introduction of auxiliary points. Section 3.2 analyses use of the Milstein density using the approach of Pedersen (1995). This is followed by an outline of the Bayesian approach to diffusions in Section 3.3. An example using simulated data comparing the Euler and Milstein schemes is presented in Section 3.4. The paper concludes with Section 4.

2 The Milstein density

Definitions and the conditions under which a SDE exists are first presented. Two schemes which can be used to approximate a generally unknown true transition density are the Euler and the Milstein approximations. Convergence rates of these two schemes are discussed and transition densities implied by the approximations are shown. The Euler approximation is known to have a closed form and this next Section will show that the Milstein scheme also has an analytic expression for the transition densities.

2.1 Estimation of stochastic differential equations from discrete observations

Suppose $(\Omega, \mathcal{F}, P)$ is a probability space and $\{\mathcal{F}_t,\, t \geq 0\}$ is a non-decreasing family of σ-algebras contained in $\mathcal{F}$. An Itô process can be defined as a stochastic process that satisfies a stochastic differential equation (SDE) of the form

$$dy(t) = a\{y(t), t; \theta\}dt + b\{y(t), t; \theta\}dW(t), \tag{2.1}$$

where the drift function $a(\cdot,\cdot;\theta) : [0,\infty] \times \mathbb{R} \times \mathbb{R}^p \mapsto \mathbb{R}$ and the volatility function $b(\cdot,\cdot;\theta) : [0,\infty] \times \mathbb{R} \times \mathbb{R}^p \mapsto \mathbb{R}$, depend on the observations $y(t)$, time t, and an unknown parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$. They are assumed to be twice continuously differentiable with respect to the states $y(t)$. The process $W(t)$ is a homogeneous Lévy process, that is, a process with independent increments which is continuous in probability, (see Barndorff-Nielsen et al. 1996). A special case of this is where $W(t)$ is a Wiener process, which we consider throughout this paper. The univariate case is presented although it can easily be generalized. So $W = \{W(t), \mathcal{F}, P\}$ is the standard Wiener process, a continuous square integrable martingale. We also define $y(t)$ with respect to $\mathcal{F}(t)$ to be a continuous random process defined on $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{R}$, thus admitting the SDE in equation (2.1). The Itô process is an Itô diffusion when they coincide in law and hence are probabilistically equivalent. We assume that the conditions under which the SDE exists and is unique are satisfied, (see Øksendal 1995, page 64).

It is assumed that the Itô process under consideration in general satisfies the stochastic differential equation,

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

in Itô form, for $t \in [0, T]$.

Suppose we are presented with a set of observations from some financial time series, $y_t = \{y(\tau_t)\}$ for $t = 1, \ldots, T$. As these are only observed at discrete intervals $\Delta_t^\dagger = \tau_{t+1} - \tau_t$, it is in general not possible to determine the analytical form of the density for the observations $g(y_{t+1}|y_t; \theta)$, unless the SDE has a strong solution. $W(t)$ is the standard Wiener process with increment variance $\Delta^\dagger$. Earlier work has focused on estimating the discretized version of the SDE using a time-discrete approximation. We will focus on the implied approximate transition densities, $f(y_{t+1}|y_t; \theta)$. In the analysis presented below, dependence on θ will be suppressed and it will be assumed that $\Delta_t^\dagger = \Delta^\dagger$ for all t.

Suppose we have the form for the SDE as specified by (2.1). The simplest strong approximation is the Euler-Maruyama approximation, also know as the Euler scheme. Suppose for notational convenience that we define $a = a(y_t, t; \theta)$, $b = b(y_t, t; \theta)$ and $b' = \partial b(y_t, t; \theta)/\partial y$. Using the Euler approximation on equation (2.1), (see Kloeden and Platen 1992, Section 10.2), we obtain

$$y_{t+1} = y_t + a\Delta^\dagger + b(W_{t+1} - W_t). \tag{2.2}$$

The Milstein approximation, see Milstein (1978) and Kloeden and Platen (1992, Section 10.3), gives us

$$y_{t+1} = y_t + a\Delta^\dagger + b(W_{t+1} - W_t) + \frac{1}{2}bb'\{(W_{t+1} - W_t)^2 - \Delta^\dagger\}. \tag{2.3}$$

where $(W_{t+1} - W_t)$ has mean zero and variance $\Delta^\dagger$.

2.2 Orders of Convergence

We shall use a stochastic Taylor expansion to derive a time discrete approximation with respect to the strong convergence criterion, subsequently defined. A time discrete approximation x^δ with maximum step size *converges strongly with order* $\gamma > 0$ (in absolute error) at time T if there exists a positive constant C, which does not depend on γ, and a $\delta_0 > 0$ such that

$$\epsilon(\delta) = E(|y_T - x^\gamma(T)|) \leq C\delta^\gamma$$

for each $\delta \in (0, \delta_0)$, where the expectation is taken with respect to the random variable y_T, (for details, see Kloeden and Platen 1992, Section 9.6). The assumptions are presented in Appendix A.1. The Euler scheme converges strongly with order $\gamma = 0.5$.

The addition of one more term to the Euler scheme yields the Milstein scheme which is an order 1.0 strong Taylor scheme. Hence the strong convergence order is increased from $\gamma = 0.5$ to $\gamma = 1$. Suppose we impose the assumptions (A. 2), (A. 3), (A. 9), (A. 12) and (A. 14). Then for the Milstein approximation x^δ, the estimate

$$\mathrm{E}\{|y_T - x^\delta|\} \leq K_5\delta$$

where the constant K_5 does not depend on δ. A proof can again be found in Kloeden and Platen (1992, Section 10.3 and 10.6).

2.3 Approximate transition densities

Both the Euler and the Milstein schemes can trivially be used to simulate diffusions. Furthermore, both imply an approximate transition density $f(y_{t+1}|y_t)$ which can be used for estimation.

For the Euler approximation, we can express conditional distribution of y_{t+1} given y_t as

$$y_{t+1}|y_t \sim N\{y_t + a(y_t, t; \theta)\Delta^\dagger, b^2(y_t, t; \theta)\Delta^\dagger\}. \tag{2.4}$$

Unfortunately, for estimation of the unknown parameters, methods based on discrete observations when the underlying process is motivated for continuous observations, it is well-known that estimators are strongly biased unless the interval between the observations, $\Delta^\dagger$ tends to zero. Use of a higher order scheme reduces this inherent discretization bias. The Milstein approximation is an alternative that can be used and as far as we know, the density has not appeared in the literature before.

The transition density implied by the Milstein is given by the next theorem.

Theorem 2.1 *Given the SDE specified by (2.1), if the usual conditions under which the SDE exists and is unique are satisfied and further, that assumptions (A. 2), (A. 3) and (A. 9)–(A. 14) hold then for the approximate one-step ahead transition density implied by (2.3), can be expressed as*

$$f(y_{t+1}|y_t) = \frac{\exp(-\lambda/2)}{|A|\sqrt{2\pi}} z_{t+1}^{-1/2} \exp(-\frac{z_{t+1}}{2}) \cosh(\sqrt{\lambda z_{t+1}}) \tag{2.5}$$

where

$$\begin{aligned} z_{t+1} &= \frac{y_{t+1} - B}{A} \\ \lambda &= \frac{1}{\Delta^\dagger (b')^2} \\ A &= \frac{bb'\Delta^\dagger}{2} \\ B &= -\frac{b}{2b'} + y_t + a\Delta^\dagger - \frac{bb'\Delta^\dagger}{2}. \end{aligned}$$

The density is defined for $z_{t+1} \in \mathbb{R}^+$ and $b' \neq 0$.

Proof 2.1 *Denote y_{t+1} by w, given a fixed value of y_t. From (2.3) it follows that*

$$\begin{aligned} w &= y_t + a\Delta^\dagger + b\sqrt{\Delta^\dagger}U + \frac{1}{2}bb'\Delta^\dagger(U^2 - 1) \\ &= A(U + \delta)^2 + B \\ &= Az + B \end{aligned}$$

where $U \sim N(0,1)$, $\delta = 1/(b'\sqrt{\Delta^\dagger})$ *and* $z = (U+\delta)^2$. *Thus* z *follows a noncentral chi-squared distribution with* $\nu = 1$ *degrees of freedom and non-centrality parameter* δ^2. *(For a more extensive coverage see Johnson, Kotz, and Balakrishnan, 1995, chapter 29). Letting* $\lambda = \delta^2$, *the density of* z *is given by*

$$f_z(z) = \frac{1}{2}\exp\left\{-\frac{(\lambda+z)}{2}\right\}\left(\frac{z}{\lambda}\right)^{-1/4} I_{-\frac{1}{2}}(\sqrt{\lambda z})$$

(see Fisher 1928) where

$$I_{-\frac{1}{2}}(x) = \sqrt{\frac{2}{x}}\sum_{j=0}^{\infty}\frac{(x/2)^{2j}}{j!\,\Gamma(j+\frac{1}{2})} = \sqrt{\frac{2}{\pi x}}\cosh(x)$$

is a modified Bessel function of the first kind of order $-\frac{1}{2}$, *(see Abramowitz and Stegun, 1970 Ch. 10; the expression in terms of* cosh *is given in equation 10.2.14). A justification of why this follows naturally is provided in Appendix A.2. So the density of* w *is given by*

$$\begin{aligned} f(w) &= \frac{1}{|A|}f_z(\frac{w-B}{A}) \\ &= \frac{1}{2|A|}\exp\left\{-\frac{(A\lambda+w-B)}{2A}\right\}\left(\frac{w-B}{\lambda A}\right)^{-\frac{1}{4}}\sqrt{\frac{2}{\pi}}\left\{\frac{\lambda(w-B)}{A}\right\}^{-\frac{1}{4}}\cosh\left\{\sqrt{\frac{\lambda(w-B)}{A}}\right\} \\ &= \frac{1}{|A|}\frac{1}{\sqrt{2\pi}}\left(\frac{w-B}{A}\right)^{-\frac{1}{2}}\exp\left\{-\frac{(A\lambda+w-B)}{2A}\right\}\cosh\left\{\sqrt{\frac{\lambda(w-B)}{A}}\right\} \end{aligned}$$

This concludes the proof.

The mean and variance of z are given by $\mathrm{E}(z) = v+\lambda$ and $\mathrm{var}(z) = 2(v+\lambda)$ which give $E(y_{t+1}|y_t) = y_t + a\Delta^\dagger$ and $var(y_{t+1}|y_t) = b^2\Delta^\dagger + (bb'\Delta^\dagger)^2/2$. These results can easily be verified from (2.3). This density is non-symmetric, unlike the Euler case. It is also worth noting that the Milstein collapses to the Euler when b' is close to zero, see Appendix A.3. For numerical reasons, it is often easier to work with the exponential function when evaluating the hyperbolic cosine function and so we use $\cosh(x) = (1/2)\{\exp(x)+\exp(-x)\}$.

3 Estimation

One can approximate the likelihood using some discrete scheme. This is achieved using either the Euler or the Milstein scheme which are outlined below. Essentially, this involves maximizing the likelihood, based on the approximate transition density. We will also motivate the

extensions of applying the Milstein density. These involve introducing missing paths between the observed values, integrating these auxiliary (latent) points out of the joint density and then computing the implied conditional transition density. Hence one could follow a simulated likelihood method, see for example, Pedersen (1995), or a Bayesian Markov Chain Monte Carlo (MCMC) method, conditioning on both neighboring observations, see Elerian, Chib, and Shephard (1998).

3.1 The likelihood based on the Milstein density

When the transition densities are known, the log-likelihood for the data, $l_T(\theta)$ in terms of the conditional distributions of y_{t+1} given y_t is given by

$$l_T(\theta) = \prod_{t=0}^{T-1} \log g(y_{t+1}|y_t;\theta) + \log g(y_1;\theta)$$

where $g(y_{t+1}|y_t;\theta)$ is used to denote the true transition densities. In general, $g(y_1;\theta)$ is intractable and so we will typically omit it and work with the conditional likelihood to estimate θ. The impact of the initial distribution is one term relative to the $T-1$ conditional likelihoods and so it can be argued that though $g(y_1;\theta)$ contains information on θ, the effect of omitting the term is small. This can be easily formalized if y_t is stationary or mixing. Hence we work with

$$\sum_{t=1}^{T-1} \log g(y_{t+1}|y_t;\theta) \tag{3.6}$$

which is maximized. If the strong solution of the underlying process is available, $g(y_{t+1}|y_t;\theta)$ will be known, and so the maximum likelihood estimator will be consistent. In general, no such strong solution exists and so we approximate (3.6) using the conditional transition densities $f(y_{t+1}|y_t;\ theta)$ implied by the Euler or the Milstein schemes.

For discretely observed diffusions, the MLE has, under relatively weak conditions, the usual properties of consistency, asymptotic normality and efficiency, see for example Billingsley (1961) and Dacunha-Castelle and Florens-Zmirou (1986).

Considering the SDE defined by equation (2), where the log-likelihood takes the form

$$\hat{l}_T(\theta) = \sum_{t=1}^{T-1} \log\{f(y_{t+1}|y_t;\theta)\}$$

where the transition densities $f(\cdot)$ of Y are known; T includes all the observed points. In this set-up, a necessary condition for the existence of a consistent estimator of θ is that $T\Delta \to \infty$, that is, $\hat{\theta}_T \to \theta$ as $T \to \infty$.

Dacunha-Castelle and Florens-Zmirou (1986) give quantities for the loss in precision due to discretization when the volatility coefficient is known and unknown. The approximation based on discrete observations without the introduction of auxiliary points is biased and inconsistent, see Florens-Zmirou (1993); implying that the MLE will be inconsistent under both the Euler and Milstein schemes when the true transition density is unknown. An example of this is shown in table 1. The approximate log-likelihood is computed for two simulated data sets, corresponding to the cases $\Delta^\dagger = 2$ and $\Delta^\dagger = 5$ for the CIR process, Cox, Ingersoll, and Ross (1985) which will be defined in Section 3.4.[1]. The results show that the bias and MSE are lower under the Milstein scheme for the drift parameters. For the volatility coefficient, there was negligible difference between the two schemes.

3.2 A Pedersen type approach to diffusions using the Milstein scheme

Pedersen (1995) treats the values on a trajectory connecting two discrete observations as missing and so introduces auxiliary points and discretizes the process using the Euler scheme. Given y_t, he simulates forwards obtaining $y_t^* = (y_{t,1}^*, \ldots, y_{t,M_t}^*)$, for $t = 1, \ldots, T$, where M_t is the number of points introduced between the observations y_t and y_{t+1}. This method can be extended to the case when the Milstein scheme is used to discretize the process.

By definition, the approximate likelihood can be expressed in terms of the transition densities

$$f^M(y_{t+1}|y_t,\theta) = \int f(y_{t+1}|y_t^*, y_t, \theta) f(y_t^*|y_t,\theta) dy_t^* \; = \mathrm{E}\{f(y_{t+1}|y_{t,M_t};\theta)|y_t\}, \tag{3.7}$$

where the expectations is taken with respect to the density

$$\begin{aligned} f(y_{t,M_t}|y_t) &= \int f(y_t^*|y_t)\, dy_{t,1}^*, \ldots, dy_{t,M_t-1}^* \\ &= \int f(y_{t,M_t}^*|y_{t,M_t-1}^*) f(y_{t,M_t-1}^*|y_{t,M_t-2}^* \ldots f(y_{t,1}^*|y_t)\, dy_{t,1}^*, \ldots, dy_{t,M_t-1}^*, \end{aligned} \tag{3.8}$$

[1]Data was obtained using $\Delta = \Delta^\dagger/R$, where R is typically taken as 10,000. $T = 500$ data points are simulated using the Milstein scheme and recorded every R iterations. The Monte Carlo results of a small simulation study are shown in table 1.

Monte Carlo Results (CIR process), $\Delta^\dagger = 2$					
	true value	mean	bias	s.d.	MSE
Euler					
α	0.5	0.25976	-0.24024	0.01358	0.05790
β	0.2	0.10449	-0.09551	0.00592	0.00916
σ^2	0.05	0.03700	-0.01300	0.00219	0.00017
Milstein					
α	0.5	0.34381	-0.15619	0.033150	0.025495
β	0.2	0.13845	-0.06155	0.01319	0.00396
σ^2	0.05	0.03600	-0.01400	0.00207	0.00020

Monte Carlo Results (CIR process), $\Delta^\dagger = 5$					
	true value	mean	bias	s.d.	MSE
Euler					
α	0.5	0.17871	-0.32129	0.00835	0.10329
β	0.2	0.07271	-0.12729	0.00340	0.01621
σ^2	0.05	0.02499	-0.02501	0.00173	0.00063
Milstein					
α	0.5	0.26631	-0.23369	0.01324	0.05478
β	0.2	0.10834	-0.09166	0.00535	0.00843
σ^2	0.05	0.02292	-0.02708	0.00126	0.00074

Table 1: *The results from a small simulation study for the parameters of the CIR process are shown using the Euler and Milstein scheme. One hundred sets of data with* $T = 500$, $R = 10,000$, $\Delta^\dagger = 2$ *and* $\Delta^\dagger = 5$ *were generated and used to evaluate the approximate log-likelihoods using a numerical optimization scheme. The true parameter value, mean, bias, standard deviation and mean square errors are shown.*

where dependence on θ has been suppressed. Monte Carlo integration is used to estimate equation (3.7) and equation (3.8) is used to generate the values of y^*_{t,M_t}, where we set $M_t = M$ for all t. As each $y^*_{t,j}$ is drawn from $f(y^*_{t,j}|y^*_{t,j-1};\theta)$, the resulting $y^*_{t,M}$ will be a variate drawn from $f(y^*_{t,M}|y_t;\theta)$. A sequence of N by M standard normal random variates, $U_i^{(k)}$ are generated for $k = 1,\ldots,N$ and $i = 1,\ldots,M$. Each point $y^{(k)*}_{t,i+1}$ is simulated forwards where $y^{(k)*}_{t,i+1}|y^{(k_*)}_{t,i}$ is normal given the Euler approximation

$$y^{*(k)}_{t,i+1} = y^{*(k)}_{t,i} + a\{y^{*(k)}_{t,i};\theta\}\Delta + b\{y^{*(k)}_{t,i};\theta\}\sqrt{\Delta}U^{(k)}_{i+1} \tag{3.9}$$

where $\Delta = \frac{\Delta^\dagger}{M+1}$, for $i = 0,\ldots,M-1$. An estimate of $f^M(y_{t+1}|y_t;\theta)$ is then given by

$$\hat{f}^M(y_{t+1}|y_t;\theta) \;=\; \frac{1}{N}\sum_{k=1}^{N} f(y_{t+1}|y^{*(k)}_{t,M};\theta) \tag{3.10}$$

where the density is given by equation (3.9) and the estimated log-likelihood is given by

$$\hat{l}_T^{(M)}(\theta) = \sum_{t=1}^{T-1} \log \hat{f}^M(y_{t+1}|y_t;\theta). \tag{3.11}$$

Note that $\hat{f}^M$ is unbiased for f^M, but the log is biased.

Estimates of the θ are obtained by maximizing the approximate log-likelihood with respect to θ. The same procedure can be carried out using the Milstein.

3.3 Bayesian approach to diffusions using Milstein densities

We propose to estimate the parameters within a Bayesian context to conduct finite sample inference and calculate posterior distributions. The algorithm relies on MCMC methods, see Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), Hastings (1970), Geman and Geman (1984) for early work, and subsequently Tanner and Wong (1987), Gelfand and Smith (1990), Gilks, Richardson, and Spiegelhalter (1996) and Tanner, (ch. 6, 1996). An excellent exposition of the above is provided by Chib and Greenberg (1996). Essentially, MCMC methods concentrate on producing variates from a given multivariate distribution, (the posterior), by repeatedly sampling a Markov chain whose invariant distribution is the target density of interest. We sample the conditional densities using an MH algorithm, see, Chib and Greenberg (1995) and proceed to sample the states in blocks. One should note that as these typically form a high dimensional correlated sample from the target density of interest, it is important to monitor the correlograms and the inefficiency factors (the variance of the sample mean of the MCMC sampling scheme divided by the variance of the sample mean from a hypothetical sampler drawing independent and identically distributed random variables from the posterior), to gauge how long one should run the sampler for and how accurate the degree of the posterior estimates. Emphasis has been on constructing a proposal that is easy to sample from and an algorithm that has good mixing properties.

Our approach is to generate M values of $y_t^* = (y_{t,1}, \ldots, y_{t,M})$ between each pair of successive observed values y_t and y_{t+1}. We will then generate the parameters from their posterior distributions, given the data Y and $Y^* = (y_1^{*\prime}, \ldots, y_{T-1}^{*\prime})'$, which for these models will often be relatively easy to sample from. An outline of the method using the Euler scheme is presented in Elerian, Chib, and Shephard (1998).

The aim is to sample Y^* from $f(Y^*|Y;\theta)$, where we will make use of the conditional independence of y_t^* given the observations y_t and y_{t+1}. Suppose we focus on a pair of observations y_t and y_{t+1}; we wish to simulate y_t^* from $f(y_t^*|y_t, y_{t+1}) \propto \Pi_{k=0}^{M} f(y_{t,k+1}^*|y_{t,k}^*)$. Each of these conditional distributions is given by the Milstein approximation defined in equation (2.3). Parameters are then estimated in a Bayesian framework. An application outlining the workings of the algorithm and comparison with Pedersen's method will be reported elsewhere.

3.4 Example

3.4.1 The Cox, Ingersoll and Ross model

The CIR model, also known as the *square-root process*, was developed by Cox et al. (1985) for term structure modelling. It can be expressed as

$$dy_t = (\alpha - \beta y_t)dt + \sigma\sqrt{y_t}dW_t$$

where y_t is, for example, the short-term interest rate. The presence of the square-root in the diffusion coefficient means that the diffusion only takes positive values, though it can reach zero. For our purposes, we are interested in finding the Milstein approximation and taking a second order Taylor expansion of the log- likelihood. Suppose we take the transformation $x_t = \log y_t$, due to the restriction that $y_t > 0$.[2] Then on applying Itô's lemma (see for example Øksendal 1995), we obtain

$$dx_t = \left\{ \frac{\alpha}{\exp(x_t)} - \beta - \frac{\sigma^2}{2\exp(x_t)} \right\} dt + \frac{\sigma}{\sqrt{\exp(x_t)}} dW_t.$$

Considering the Milstein scheme in equation (2.3), we obtain, for $x_{t,k+1}$, the following specification

$$\begin{aligned} x_{t,k+1} &= x_{t,k} + \left\{ \frac{\alpha}{\exp(x_{t,k})} - \beta - \frac{\sigma^2}{2\exp(x_{t,k})} + \frac{\sigma^2}{4\exp(2x_{t,k})} \right\} \Delta + \frac{\sigma}{\exp\frac{x_{t,k}}{2}} \epsilon_{t,k+1} \\ &\quad - \frac{\sigma^2}{4\exp(2x_{t,k})} \epsilon_{t,k+1}^2 \end{aligned}$$

[2] Although the CIR is always positive, when we do Euler or Milstein, this is not necessarily true, so one needs to impose this restriction by transforming the process.

where $\epsilon \sim N(0, \Delta)$. Hence the conditional mean and variance are given by

$$\begin{aligned} \mathrm{E}(x_{t,k+1}|x_{t,k}) &= x_{t,k} + \left\{ \frac{\alpha}{\exp(x_{t,k})} - \beta - \frac{\sigma^2}{2\exp(x_{t,k})} \right\} \Delta \\ \mathrm{var}(x_{t,k+1}|x_{t,k}) &= \frac{\sigma^2 \Delta}{\exp(x_{t,k})} + \frac{\sigma^2 \Delta^2}{8\exp(4x_{t,k})}. \end{aligned}$$

Hence the Milstein density for $f(y_{t+1}|y_t)$ is given by (2.5), with

$$\begin{aligned} a &= \frac{\alpha}{\exp(x_{t,k})} - \beta - \frac{\sigma^2}{2\exp(x_{t,k})} + \frac{\sigma^2}{4\exp(2x_{t,k})} \\ b &= \frac{\sigma}{\exp \frac{x_{t,k}}{2}} \\ b' &= -\frac{\sigma}{2\exp \frac{x_{t,k}}{2}} \end{aligned}$$

4 Conclusion

A closed form solution for the one-step ahead conditional transition density using the Milstein scheme has been shown to exist. This higher order Taylor approximation enables us to obtain an order of improvement in accuracy in estimating the parameters in a non-linear diffusion, as compared to the use of the Euler-Maruyama discretization scheme. The method can easily be extended to estimate the parameters using either the method proposed by Pedersen (1995), or a using MCMC scheme as proposed in Elerian, Chib, and Shephard (1998). Both of these will be considered elsewhere.

5 Acknowledgments

I would like to thank Neil Shephard, Bent Nielsen and Jouni Kuha for their comments and advice. I would also like to thank Nuffield College for their financial support. All the computations reported in this paper were carried out using the Ox language of Doornik (1996).

A Appendix

A.1 Assumptions under the Euler and Milstein schemes

We will assume the following conditions in Section 2.2 under the Euler scheme, using y_0 and x_0 as starting points:

$$\begin{aligned}
\mathrm{E}(|y_0|^2) &< \infty && \text{(A. 1)}\\
\mathrm{E}(|y_0 - x_0^\delta|^2)^{\frac{1}{2}} &\leq K_1\delta^{\frac{1}{2}} && \text{(A. 2)}\\
|a(t,y) - a(t,x)| + |b(t,y) - b(t,x)| &\leq K_2|y-x| && \text{(A. 3)}\\
|a(t,y)| + |b(t,y)| &\leq K_3(1+|y|) && \text{(A. 4)}\\
|a(s,y) - a(t,y)| + |b(s,y) - b(s,x)| &\leq K_4(|1+|y|))|s-t|^{\frac{1}{2}} && \text{(A. 5)}
\end{aligned}$$

where the constants $K_1, \ldots, K_4$ do not depend on δ, $s,t \in [0,T]$ and $y,x \in \mathbb{R}^d$ then for the Euler approximation x^δ, we obtain

$$\mathrm{E}\{|y_T - x^\delta(T)|\} \leq K_5\delta^{\frac{1}{2}}$$

where the constant K_5 does not depend on δ. Given conditions A. 4 and A. 5, the result follows from applying Doob's inequality and the Gronwall inequality. A proof of this assertion can be found in Kloeden and Platen (Section 10.2, 1992).

Under the Milstein scheme we make the following additional assumptions:

$$\begin{aligned}
|a^*(t,y) - a^*(t,x)| &\leq K_2|y-x| && \text{(A. 6)}\\
|b(t,y) - b(t,x)| &\leq K_2|y-x| && \text{(A. 7)}\\
|Lb(t,y) - Lb(t,x)| &\leq K_2|y-x| && \text{(A. 8)}\\
|a^*(t,y)| + |La(t,y)| &\leq K_3(1+|y|) && \text{(A. 9)}\\
|b(t,y)| + |Lb(t,y)| &\leq K_3(1+|y|) && \text{(A. 10)}\\
|L^2b(t,y)| &\leq K_3(1+|y|) && \text{(A. 11)}\\
|a^*(s,y) - a^*(t,y)| &\leq K_4(1+|y|)|s-t|^{\frac{1}{2}} && \text{(A. 12)}\\
|b(s,y) - b(t,y)| &\leq K_4(1+|y|)|s-t|^{\frac{1}{2}} && \text{(A. 13)}\\
|Lb(s,y) - Lb(t,y)| &\leq K_4(1+|y|)|s-t|^{\frac{1}{2}} && \text{(A. 14)}
\end{aligned}$$

for all $s, t \in [0, T]$, $y, x \in \mathbb{R}^d$ where

$$\begin{aligned} a^* &= a - \frac{1}{2} b b' \\ L &= b \frac{\partial}{\partial y}, \end{aligned}$$

and the constants $K_2, \ldots, K_4$ do not depend on δ.

A.2 Proof of Milstein density

The modified Bessel function follows naturally by first defining the double factorial notation and expressing $(2j)!! = (2j)(2j-2)\ldots(4)(2)$ and $(2j-1)! = (2j-1)(2j-3)\ldots(3)(1)$ which is the j'th moment of a chi-squared random variable with one degree of freedom. Hence $(2j)! = (2j)!!(2j-1)!!$ and $2^j(j!) = (2j)!!$ implying that

$$2^j\,\Gamma(j+\frac{1}{2}) = 2^j(j-\frac{1}{2})(j-\frac{3}{2})\ldots(\frac{1}{2})\Gamma(\frac{1}{2}) = (2j-1)!!\,\Gamma(\frac{1}{2}),$$

so that

$$4^j\,j!\,\Gamma(j+\frac{1}{2}) = 2^j\,j!\,(2j-1)!!\,\Gamma(\frac{1}{2}) = (2j)!\,\Gamma(\frac{1}{2}).$$

Hence, going back to the infinite sum,

$$\sum_{j=0}^{\infty} \left\{ \frac{\left(\frac{x^2}{4}\right)^j}{j!\,\Gamma(j+\frac{1}{2})} \right\} = \frac{1}{\Gamma(\frac{1}{2})} \sum_{j=0}^{\infty} \frac{x^{2j}}{(2j)!} = \frac{1}{\sqrt{\pi}} \cosh(x),$$

using $\mathcal{F}_{-\frac{1}{2}(x)} = \sqrt{\frac{2}{\pi x}} \cos x$; see Gradshteyn and Ryzhik 1965, Section 8.464. Here, $\mathcal{F}$ is the real argument component of $I_{\nu(x)} = \exp\left(-\frac{\pi}{2}\nu\, i\right)\, \mathcal{F}_\nu \left\{\exp\left(\frac{\pi}{2i}\right) x\right\}$. For an additional reference, see Watson 1944.

In general, Bessel functions with 'half' parameters ($I_{\frac{1}{2}+\nu}$, $\mathcal{F}_{\frac{1}{2}+\nu}$ for $\nu \in \mathbb{Z}$) can be written in terms of trigonometric functions. Other Bessel functions cannot be expressed in such explicit forms.

A.3 Milstein reduces to Euler for small b'

Suppose we denote $x = y_{t+1} - y_t - a\Delta$. Then for the Milstein scheme:

$$\begin{aligned} y_{t+1} - B &= x + \frac{b}{2b'} + \frac{bb'\Delta}{2} \\ z_{t+1} &= \frac{y_{t+1} - B}{A} = \frac{x + \frac{b}{2b'} + \frac{bb'\Delta}{2}}{\frac{bb'\Delta}{2}} = 1 + \frac{2x}{bb'\Delta} + \frac{1}{(b')^2\Delta}. \end{aligned}$$

For small b':

$$\begin{aligned} \sqrt{z} &= \sqrt{1 + \frac{2x}{bb'\Delta} + \frac{1}{(b')^2\Delta}} = \sqrt{\frac{1}{(b')^2\Delta}\left\{1 + \frac{2xb'}{b} + (b')^2\Delta\right\}} \\ &\approx \frac{1}{b'\sqrt{\Delta}}\left[1 + \frac{1}{2}\left\{\frac{2xb'}{b} + (b')^2\Delta\right\} + \frac{1}{8}\left\{\frac{2xb'}{b} + (b')^2\Delta\right\}^2 \ldots\right] \\ &\approx \frac{1}{b'\sqrt{\Delta}}\left\{1 + \frac{xb'}{b} + \frac{(b')^2\Delta}{2} - \frac{x^2(b')^2}{2b^2} + \ldots\right\} \end{aligned}$$

so that

$$\begin{aligned} \sqrt{z\lambda} &\approx \frac{1}{(b')^2\Delta}\left\{1 + \frac{xb'}{b} + \frac{(b')^2\Delta}{2} - \frac{x^2(b')^2}{2b^2} + \ldots\right\} \\ \cosh\sqrt{z\lambda} &\approx \frac{1}{2}\exp(\sqrt{z\lambda}). \end{aligned}$$

This implies that

$$\begin{aligned} 2\exp\left(-\frac{\lambda + z}{2}\right)\cosh\sqrt{z\lambda} &\approx \exp\left(-\frac{1}{2} - \frac{x}{bb'\Delta} - \frac{1}{(b')^2\Delta}\right)\exp\left\{\frac{1}{(b')^2\Delta} + \frac{x}{bb'\Delta} + \frac{1}{2} + \frac{x^2}{2b^2}\right\} \\ &= \exp\left(\frac{-x^2}{2b^2}\right). \end{aligned}$$

So, assuming $b > 0$,

$$2|A|\sqrt{z} \to b\sqrt{\Delta}$$

proving that for small b', the Milstein density reduces to that of the Euler.

References

Abramowitz, M. and I. Stegun (1970). *Handbook of Mathematical Functions* (Ninth ed.). New York: Dover Publications Inc.

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.

Aït-Sahalia, Y. (1998). Maximum likelihood estimation of discretely sampled diffusions: A closed-form approach. Pre-print.

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.

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

Billingsley, P. (1961). *Statistical Inference for Markov Processes*. Chicago: Chicago University Press.

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

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.

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

Dixit, A. and R. S. Pindyck (1994). *Investment Under Uncertainty*. Priceton University Press.

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

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. Norwegian School of Economics and Business Administration, HAE thesis.

Fisher, R. A. (Ed.) (1928). *The General Sampling Distribution of the Multiple Correlation Coefficient*, Volume 121 A, London. Royal Society of London. 654-673.

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

Florens-Zmirou, D. (1993). On estimating the diffusion coefficient from discrete observations. *Journal of Applied Probability 30*, 790–804.

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.

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

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.

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

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

Gradshteyn, I. S. and I. M. Ryzhik (1965). *Table of Integrals, Series and Products*. San Diego: Academic Press.

Hansen, L. and J. A. Scheinkman (1995). Back to the future: Generating moment implications for continuous-time markov processes. *Econometrica 63*, 767–804.

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

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.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous Univariate Distributions* (Second ed.), Volume 2. New York: John Wiley.

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.

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

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

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

Milstein, G. N. (1978). A method of second-order accuracy integration of stochastic differential equations. *Theory of Probability and its Applications 23*, 396–401.

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

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

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

Tanner, M. A. (1996). *Tools for Statistical Inference: methods for exploration of posterior distributions and likelihood functions* (Third ed.). New York: Springer-Verlag.

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.

Watson, G. N. (1944). *A Treatise on the Theory of Bessel Functions*. Cambridge: Cambridge University Press.