

NCER Working Paper Series

Seeing the Wood for the Trees: A Critical Evaluation of Methods to Estimate the Parameters of Stochastic Differential Equations

A. S. Hurn, J. Jeisman and K.A. Lindsay

Working Paper #2

July 2006

Abstract

Maximum-likelihood estimates of the parameters of stochastic differential equations are consistent and asymptotically efficient, but unfortunately difficult to obtain if a closed form expression for the transitional probability density function of the process is not available. As a result, a large number of competing estimation procedures have been proposed. This paper provides a critical evaluation of the various estimation techniques. Special attention is given to the ease of implementation and comparative performance of the procedures when estimating the parameters of the Cox-Ingersoll-Ross and Ornstein-Uhlenbeck equations respectively.

Seeing the Wood for the Trees: A Critical Evaluation of Methods to Estimate the Parameters of Stochastic Differential Equations

A. S. Hurn and J. Jeisman

School of Economics and Finance, Queensland University of Technology

K.A. Lindsay

Department of Mathematics, University of Glasgow

Abstract

Maximum-likelihood estimates of the parameters of stochastic differential equations are consistent and asymptotically efficient, but unfortunately difficult to obtain if a closed form expression for the transitional probability density function of the process is not available. As a result, a large number of competing estimation procedures have been proposed. This paper provides a critical evaluation of the various estimation techniques. Special attention is given to the ease of implementation and comparative performance of the procedures when estimating the parameters of the Cox-Ingersoll-Ross and Ornstein-Uhlenbeck equations respectively.

Keywords

stochastic differential equations, parameter estimation, maximum likelihood, simulation, moments.

JEL Classification C22, C52

Corresponding author

Joseph Jeisman

School of Economics and Finance

Queensland University of Technology

Brisbane, 4001, Australia

email j.jeisman@qut.edu.au

Acknowledgments

We are grateful to Yacine Aït-Sahalia, Adrian Pagan, Vance Martin and Ralf Becker for useful comments on earlier drafts of this manuscript. All remaining errors are the responsibility of the authors.

1 Introduction

The notion of a system of stochastic differential equations (SDEs), defined as a deterministic system of differential equations perturbed by random disturbances that are not necessarily small, has been used profitably in a variety of disciplines including *inter alia* engineering, environmetrics, physics, population dynamics and medicine. SDEs are also central to much of modern finance theory and have been widely used to model the behaviour of key variables such as the instantaneous short-term interest rate, asset prices, asset returns and their volatility (see Sundaresan, 2000). Consequently, the estimation of the parameters of SDEs from discretely-sampled data has received substantial attention in the financial econometrics literature, particularly in the last ten years. It is now reasonable to conjecture that this area is maturing rapidly and is therefore fertile ground for a comprehensive review. Previous surveys of the area have been carried out by Shoji and Ozaki (1997), Jensen and Poulsen (2002), and Durham and Gallant (2002) but all of these papers are limited in scope, in the sense that attention in each case is focussed on a relatively narrow selection of existing estimation methods. The aim of this paper is to provide a comprehensive overview and critical evaluation of many of the existing methods for estimating the parameters of SDEs. The emphasis is on the practical implementation of the various techniques and so effort has been expended in attempting to summarise the details of the algorithms in an accessible way¹. In each case, comments are made about the general applicability of the method, its ease of use and when it is appropriate to apply it. Monte Carlo experiments are performed in order to compare the methods with respect to the accuracy of the parameter estimates and speed.

While this paper attempts to be comprehensive in terms of the variety of estimation methods currently available, it cannot claim to be exhaustive. The estimation methods surveyed here deal exclusively with estimating the parameter vector of the general one-dimensional, time-homogeneous SDE from a single sample of observations at discrete times. As a consequence there are three areas which are not covered in this paper, namely, multi-dimensional models, including latent factor models², the class of estimators that treat the parameters of the drift and diffusion functions separately (see, for example, Yoshida, 1992; Aït-Sahalia, 1996a; Bandi and Phillips, 2003; Bandi and Phillips, 2005; and Phillips and Yu, 2005), and estimators that require panel data for their implementation (see, for example, Hurn and Lindsay, 1997; and McDonald and Sandal, 1999).

A formal statement of the parameter estimation problem to be addressed in this paper is as follows. Given the one-dimensional time-homogeneous SDE

$$dX = \mu(X; \boldsymbol{\theta}) dt + g(X; \boldsymbol{\theta}) dW \tag{1}$$

the task is to estimate the parameters $\boldsymbol{\theta}$ from a sample of $(N + 1)$ observations X_0, \dots, X_N of the

¹To aid in understanding the details of the implementation, the C-code used to implement each estimation method will be made available on request.

²In the context of financial econometrics, this means that estimation methods for models of stochastic volatility are ignored. An excellent survey is to be found in Shephard (2005) and the readings therein.

process at known times t_0, \dots, t_N where $\mu(x; \boldsymbol{\theta})$ and $g^2(x; \boldsymbol{\theta})$ are assumed to be prescribed functions of state.

The maximum-likelihood (ML) estimate of $\boldsymbol{\theta}$ is generated by minimising the negative log-likelihood function of the observed sample, namely

$$-\log \mathcal{L}(\boldsymbol{\theta}) = -\log f_0(X_0 | \boldsymbol{\theta}) - \sum_{k=0}^{N-1} \log f(X_{k+1} | X_k; \boldsymbol{\theta}), \quad (2)$$

with respect to the parameters $\boldsymbol{\theta}$. In this expression, $f_0(X_0 | \boldsymbol{\theta})$ is the density of the initial state and $f(X_{k+1} | X_k; \boldsymbol{\theta}) \equiv f((X_{k+1}, t_{k+1}) | (X_k, t_k); \boldsymbol{\theta})$ is the value of the transitional probability density function (PDF) at (X_{k+1}, t_{k+1}) for a process starting at (X_k, t_k) and evolving to (X_{k+1}, t_{k+1}) in accordance with equation (1). Note that the Markovian property of equation (1) ensures that the transitional PDF satisfies the Fokker-Planck equation

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial x} \left(\frac{1}{2} \frac{\partial (g^2(x; \boldsymbol{\theta}) f)}{\partial x} - \mu(x; \boldsymbol{\theta}) f \right) \quad x \in \mathcal{S}, \quad (3)$$

with suitable initial and boundary conditions³. Unfortunately, exact maximum likelihood (EML) estimation is only feasible in the rare cases in which a closed-form solution to this initial boundary value problem is available. In these cases, EML estimation of the parameters is straightforward since the negative log-likelihood can be computed exactly for any combination of the parameters $\boldsymbol{\theta}$ and its value minimised.

Since EML is usually infeasible, a large number of competing estimation methods have been developed. Figure 1 provides a classification of these estimation procedures into two broad categories. Likelihood-based methods provide one obvious choice as a major category, given the prominent position of ML estimation in econometrics. The other major category includes those methods which may be loosely labelled as “sample DNA matching” procedures. The latter encompass methods which differ greatly in their mode of implementation, but which all have in common the fact that they attempt to match some feature or characteristic of the data to a theoretical counterpart of the model by choice of parameters. Of course, this classification represents a subjective view and others are possible. For example, methods could be grouped on whether or not they are simulation based. In any event, there will always be difficult cases in which the classification is not straightforward. A pertinent case in point is estimation based on the characteristic function. This approach spans both the categories proposed here. In Figure 1 it has been placed under sample DNA matching reflecting the view that, in practice, it is invariably implemented in a moment-matching context.

While for most part this paper makes no claims to originality, given that it is mainly a critical evaluation of existing work, there are a number of interesting issues which emerge that have not yet been recognised in the published literature. In particular, contributions are made in the sections

³A full derivation of the Fokker-Planck equation, based on the flow of probability and the conservation of probability mass, is provided in Appendix 1. This derivation provides a more physical interpretation of the equation than the purely analytical approach favoured by standard texts (see, for example, Karlin and Taylor, 1981).

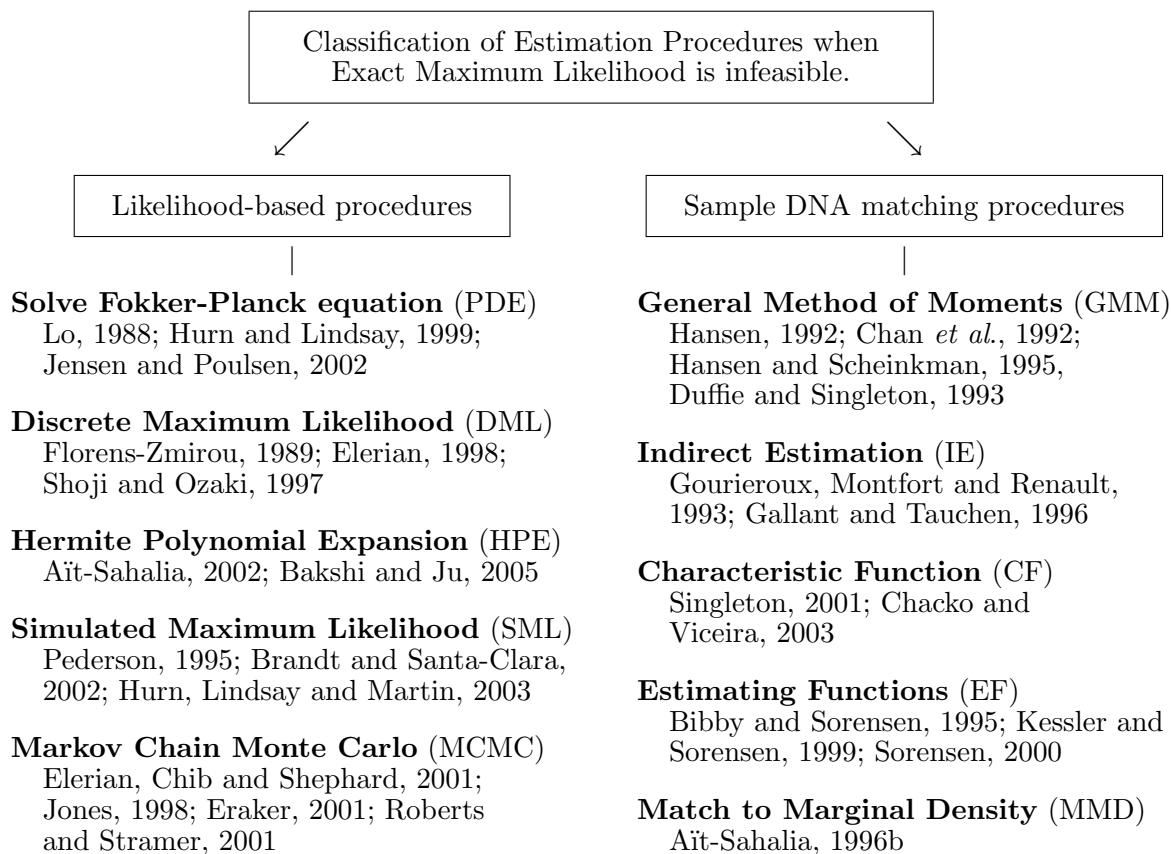


Figure 1: Schematic classification of estimation techniques.

dealing with the numerical solution of the Fokker-Planck equation and the treatment of discrete maximum likelihood respectively. In the former, it is recognised that a more efficient implementation of the numerical scheme can be achieved using the transitional cumulative distribution function (CDF) rather than the transitional PDF itself. The procedure based on the transitional CDF provides a starting condition that can be represented numerically unlike the traditional approach based on a delta function initial condition which has no numerical representation. In terms of discrete maximum likelihood, a potential difficulty in using a likelihood function based on the Milstein approximation introduced by Elerian (1998) and implemented in Durham and Gallant (2002) is highlighted. In order to implement the likelihood based on the Milstein scheme, it may be necessary to alias observations to ensure that they fall within the domain of the likelihood function.

The remainder of this paper comprises five sections. Section 2 introduces the benchmark models considered in this paper. The estimation procedures are all demonstrated using the square-root or CIR process (Cox, Ingersoll and Ross, 1985) and the Ornstein-Uhlenbeck (OU) process associated with Vasicek (1977). Not only are these models relevant in the context of financial econometrics,

but they also have closed-form expressions for their transitional PDFs allowing estimation by EML. The parameter values so obtained provide a useful benchmark with respect to which other estimation methods may be assessed. Section 3 discusses the likelihood-based estimation methods. Section 4 outlines estimation methods that try to match properties of the sample to a theoretical counterpart. Section 5 introduces the Monte Carlo experiments that involve applying the various estimation techniques to the benchmark models. Section 6 concludes.

2 Benchmark models

Given the asymptotic efficiency and consistency properties enjoyed by ML estimation (Dacunha-Castelle and Florens-Zmirou, 1986), it provides a natural benchmark against which the various estimation procedures surveyed in this paper may be compared. Given the desire to use EML as a benchmark method, it is necessary to provide benchmark models which have known closed-form expressions for their transitional PDFs. Two such models that are relevant in the context of financial econometrics are now discussed.

Cox, Ingersoll and Ross model The square-root process proposed by Cox, Ingersoll and Ross (1985) as a model of the instantaneous short term interest rate, commonly referred to as the CIR model, evolves according to the SDE

$$dX = \alpha(\beta - X) dt + \sigma\sqrt{X} dW \quad (4)$$

where α (speed of adjustment), β (the mean interest rate) and σ (volatility control) are positive parameters to be estimated. Thus the CIR process exhibits mean reversion of X to the state $X = \beta$. Most importantly, however, the properties $g(0; \boldsymbol{\theta}) = 0$ and $\mu(0; \boldsymbol{\theta}) > 0$ ensure that $\mathcal{S} = \mathbb{R}^+$.

The transitional PDF of the CIR model is derived in Appendix 2. It is the non-central chi-squared distribution

$$f(x | X_k; \boldsymbol{\theta}) = c \left(\frac{v}{u} \right)^{\frac{q}{2}} e^{-(\sqrt{u}-\sqrt{v})^2} e^{-2\sqrt{uv}} I_q(2\sqrt{uv}) \quad (5)$$

where c , u , v and ν are defined respectively by

$$c = \frac{2\alpha}{\sigma^2(1 - e^{-\alpha(t_{k+1}-t_k)})}, \quad u = cX_k e^{-\alpha(t_{k+1}-t_k)}, \quad v = cx, \quad \nu = \frac{2\alpha\beta}{\sigma^2} - 1. \quad (6)$$

This transitional PDF may now be used in combination with expression (2) to estimate the values of the parameters of the CIR model (4) by EML.

Ornstein-Uhlenbeck model The OU process proposed by Vasicek (1977) evolves according to the SDE

$$dX = \alpha(\beta - X) dt + \sigma dW \quad (7)$$

where α (speed of adjustment), β (the mean interest rate) and σ (volatility control) are again the parameters to be estimated. The OU process also exhibits mean reversion of X to the state $X = \beta$, but unlike the CIR process, the domain of the state variable is unrestricted, that is, $\mathcal{S} = \mathbb{R}$.

The derivation of the transitional PDF of the OU process is similar to that of the CIR process and is also presented in Appendix 2. The transitional PDF is the Normal distribution

$$f(x | X_k; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi V}} \exp \left[-\frac{(x - \bar{x})^2}{2V} \right] \quad (8)$$

where

$$V = \frac{\sigma^2(1 - e^{-2\alpha(t_{k+1}-t_k)})}{2\alpha}, \quad \bar{x} = \beta + (X_k - \beta) e^{-\alpha(t_{k+1}-t_k)}.$$

As for the CIR model, this transitional PDF may now be used in combination with expression (2) to estimate the values of the parameters of the OU model (7) by EML.

Numerical integration In the simulation experiments conducted in Section 5 it is necessary to generate data consistent with these benchmark models⁴. This is efficiently achieved using the Milstein (1978) scheme

$$x_{j+1} = x_j + \mu(x_j) \Delta + g(x_j) \varepsilon_j + \frac{g(x_j)}{2} \frac{\partial g(x_j)}{\partial x} (\varepsilon_j^2 - \Delta) \quad (9)$$

where x_j and x_{j+1} are consecutive realisations of X separated by an interval of duration Δ and $\varepsilon_j \sim N(0, \Delta)$.

In general the Milstein scheme is superior to the Euler-Maruyama scheme in that it exhibits strong convergence of order one by contrast with the order one half convergence of the Euler-Maruyama scheme. However, both algorithms are equivalent when g is a constant function. In the context of the CIR and OU processes, the particularised forms of Milstein's scheme are respectively

$$\begin{aligned} \text{(CIR)} \quad x_{j+1} &= x_j + \alpha(\beta - x_j) \Delta + \sigma \sqrt{x_j} \varepsilon_j + \frac{\sigma^2}{4} (\varepsilon_j^2 - \Delta), \\ \text{(OU)} \quad x_{j+1} &= x_j + \alpha(\beta - x_j) \Delta + \sigma \varepsilon_j. \end{aligned} \quad (10)$$

Having outlined the benchmark models, attention is now focussed on the details of the competing estimation methods illustrated in Figure 1.

3 Likelihood-based procedures

The first group of estimators, highlighted on the left hand column of Figure 1, seek to retain the ML framework by approximating the transitional PDF $f(x | X_k; \boldsymbol{\theta})$ by a numerical method, a discrete approximation or a simulation procedure.

⁴The OU process has an exact solution that may be used to generate data. For consistency, however, a numerical scheme is used in this paper

3.1 Numerical solution of the Fokker-Planck equation

ML estimation relies crucially on the ability to compute the value of the transitional PDF, which is known to satisfy the Fokker-Planck equation

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial x} \left(\frac{1}{2} \frac{\partial(g^2(x; \boldsymbol{\theta})f)}{\partial x} - \mu(x; \boldsymbol{\theta})f \right) \quad (11)$$

for suitable initial condition and boundary conditions. Let the state space of the problem be $\mathcal{S} = [a, b]$ and suppose that the process starts at X_k at time t_k . The appropriate initial condition in this case is

$$f(x | X_k; \boldsymbol{\theta}) = \delta(x - X_k) \quad (12)$$

where δ denotes the Dirac delta function, and the boundary conditions required to conserve unit density within this interval are the zero-flux conditions

$$\lim_{x \rightarrow a^+} \left(\mu f - \frac{1}{2} \frac{\partial(g^2 f)}{\partial x} \right) = 0, \quad \lim_{x \rightarrow b^-} \left(\mu f - \frac{1}{2} \frac{\partial(g^2 f)}{\partial x} \right) = 0. \quad (13)$$

When a closed-form solution to this initial boundary value problem, and hence a closed-form expression for the transitional PDF, is not available, an alternative is to solve the problem numerically. In the context of financial econometrics, the possibility of ML estimation based on the numerical solution of the Fokker-Planck equation was first recognised by Lo (1988) and has since been implemented by Hurn and Lindsay (1999) using spectral approximations and Jensen and Poulsen (2002) using the method of finite differences. The finite-difference procedure is described and implemented here as it is the more familiar technique.

The finite-difference procedure is based on a discretisation of state space into n uniform sub-intervals of length $\Delta x = (b - a)/n$ and a discretisation of the time interval $[t_k, t_{k+1}]$ into m uniform sub-intervals of duration $\Delta t = (t_{k+1} - t_k)/m$. Let the nodes of the finite-difference scheme be denoted by $x_p = a + p \Delta x$ where p is an integer satisfying $0 \leq p \leq n$, let $(t_k =) t_{k,0}, t_{k,1} \dots, t_{k,m} (= t_{k+1})$ denote the subdivision of $[t_k, t_{k+1}]$ into intervals of duration Δt where $t_{k,q} = t_k + q \Delta t$, and let $f_p^{(q)} = f(x_p, t_{k,q})$ be the value of the transitional PDF at x_p at time $t_{k,q}$. Integration of equation (11) over $[t_{k,q}, t_{k,q+1}]$ gives

$$f(x, t_{k,q+1}) - f(x, t_{k,q}) = \frac{1}{2} \frac{\partial^2}{\partial x^2} \left(g^2(x) \int_{t_{k,q}}^{t_{k,q+1}} f(x, t) dt \right) - \frac{\partial}{\partial x} \left(\mu(x) \int_{t_{k,q}}^{t_{k,q+1}} f(x, t) dt \right). \quad (14)$$

Let the auxiliary variables

$$\phi_p = \int_{t_{k,q}}^{t_{k,q+1}} f(x_p, t) dt$$

be defined, then in the usual notation, equation (14) has finite difference approximation

$$f_p^{(q+1)} - f_p^{(q)} = \frac{g_{p+1}^2 \phi_{p+1} - 2g_p^2 \phi_p + g_{p-1}^2 \phi_{p-1}}{2(\Delta x)^2} - \frac{\mu_{p+1} \phi_{p+1} - \mu_{p-1} \phi_{p-1}}{2\Delta x}.$$

The terms in this equation are now regrouped to give

$$f_p^{(q+1)} - f_p^{(q)} = \left(\frac{g_{p-1}^2 + (\Delta x)\mu_{p-1}}{2(\Delta x)^2} \right) \phi_{p-1} - \frac{g_p^2}{(\Delta x)^2} \phi_p + \left(\frac{g_{p+1}^2 - (\Delta x)\mu_{p+1}}{2(\Delta x)^2} \right) \phi_{p+1}$$

and the trapezoidal quadrature used to approximate ϕ_p by the formula

$$\phi_p = \frac{\Delta t}{2} (f_p^{(q+1)} + f_p^{(q)}) + O(\Delta t)^3.$$

The final finite-difference representation of equation (11) now simplifies to give

$$\begin{aligned} -r (g_{p-1}^2 + (\Delta x)\mu_{p-1}) f_{p-1}^{(q+1)} + 2(2 + r g_p^2) f_p^{(q+1)} - r (g_{p+1}^2 - (\Delta x)\mu_{p+1}) f_{p+1}^{(q+1)} \\ = r (g_{p-1}^2 + (\Delta x)\mu_{p-1}) f_{p-1}^{(q)} + 2(2 - r g_p^2) f_p^{(q)} + r (g_{p+1}^2 - (\Delta x)\mu_{p+1}) f_{p+1}^{(q)} \end{aligned} \quad (15)$$

where $r = \Delta t/(\Delta x)^2$ is the Courant number. The procedure used to construct equation (15) is essentially the Crank-Nicolson algorithm, which is well known to exhibit robust numerical properties, for example, it is stable and numerically consistent. Expression (15) forms the core of the finite-difference representation of equation (11). It suggests that the distribution of transitional density at any time is computed by solving a tri-diagonal system of equations given an initial distribution of transitional density and suitable boundary conditions.

As has already been remarked, the initial condition is a delta function and is therefore not representable within the framework of the finite-difference method. Jensen and Poulsen (2002) suggest that this difficulty can be circumvented by starting the finite-difference algorithm with a specification of the distribution of transitional density at $(t_k + \Delta t)$ based on the assumption that the transitional density at this time may be approximated by the normal distribution with mean value $X_k + \mu(X_k; \boldsymbol{\theta})\Delta t$ and variance $g^2(X_k; \boldsymbol{\theta})\Delta t$. The main drawback of this approximation is that once Δt is chosen and the initial state is known, the diffusion occurring over the time interval Δt from the true initial condition determines the size of the interval of state space over which the transitional PDF is significantly different from zero. The resolution Δx of state space must now be chosen to be sufficiently small so as to guarantee that a reasonable number of nodes (say a dozen) lie within this interval of non-zero transitional PDF. Moreover, once a suitable value of Δx is chosen, this discretisation interval must be applied to the entire state space. In practice, this requirement means that $\Delta x = O(\sqrt{\Delta t})$.

A final crucial aspect of the finite-difference procedure is the incorporation of the boundary conditions into the first and last equations in the system. Recall that the solution is sought in the finite interval (x_0, x_n) . For many SDEs of type (1), the sample space is the semi-infinite interval $(0, \infty)$ so that the drift and diffusion specifications will often satisfy $g(x_0) = 0$ and $\mu(x_0) > 0$. Under these conditions the boundary condition at $x = x_0$ is equivalent to the condition $f(x_0, t) = 0$, that is, no density can accumulate at the boundary $x = x_0$. However, no equivalent simplification exists at the boundary $x = x_n$ which must be chosen to be suitably large, but finite⁵. The derivation of the boundary condition at $x = x_n$ is now described.

⁵In the applications here, the state x_n is chosen to be the maximum of the sample plus the range of the sample.

The backward-difference representation of the boundary condition (13) at $x = x_n$ is

$$\frac{1}{2} \left(\frac{3g_n^2 f_n^{(q)} - 4g_{n-1}^2 f_{n-1}^{(q)} + g_{n-2}^2 f_{n-2}^{(q)}}{2\Delta x} \right) - \mu_n f_n^{(q)} + O(\Delta x)^2 = 0. \quad (16)$$

These terms are regrouped and the truncation error ignored to obtain

$$(3g_n^2 - 4(\Delta x)\mu_n) f_n^{(q)} - 4g_{n-1}^2 f_{n-1}^{(q)} + g_{n-2}^2 f_{n-2}^{(q)} = 0. \quad (17)$$

This boundary condition is now used at $(t_k + q\Delta t)$ and $(t_k + (q+1)\Delta t)$ to eliminate $f_n^{(q)}$ and $f_n^{(q+1)}$ respectively from equation (15) evaluated at $p = n - 1$. The final result is

$$P f_{n-2}^{(q+1)} - (Q - R) f_{n-1}^{(q+1)} = -P f_{n-2}^{(q)} + (Q + R) f_{n-1}^{(q)} \quad (18)$$

where

$$P = g_{n-2}^2 (3(\Delta x)\mu_n - 2g_n^2) - (\Delta x)\mu_{n-2} (3g_n^2 - 4(\Delta x)\mu_n),$$

$$Q = g_{n-1}^2 (4(\Delta x)\mu_n - 2g_n^2), \quad R = \frac{4}{r} (3g_n^2 - 4(\Delta x)\mu_n).$$

When it is not possible to assume that $f(x_0, t) = 0$, the lower boundary condition can be derived using a similar procedure. The result is an identical expression to equation (18) but with the subscripts n , $n - 1$ and $n - 2$ replaced by 0, 1 and 2 respectively, and the negative sign between the two terms in P replaced by a positive sign.

Assuming that $f(x_0, t) = 0$, the final specification of the numerical problem starts with equation

$$2(2 + rg_1^2) f_1^{(q+1)} - r(g_2^2 - (\Delta x)\mu_2) f_2^{(q+1)} = 2(2 - rg_1^2) f_1^{(q)} + r(g_2^2 - (\Delta x)\mu_2) f_2^{(q)}, \quad (19)$$

which is the particularisation of the general equation (15) at x_1 taking account of the requirement that $f(x_0, t) = 0$. There now follow $(n - 3)$ equations with general form (15) in which the index p takes values from $p = 2$ to $p = n - 2$, followed finally by equation (18). Taken together, these equations form a tri-diagonal system of linear equations to be solved for the transitional PDF at time $t_{k,q+1}$ given the PDF at $t_{k,q}$. Note that the tri-diagonal system described by equations (15), (18) and (19) is solved for the transitional PDF at the nodes x_1, \dots, x_{n-1} . Here the transitional PDF at x_0 is known *a priori* to be zero, while the transitional PDF at x_n is obtained directly from relation (17).

There is an equivalent statement of this problem in terms of the transitional cumulative distribution function (CDF), $F(x, t)$, which is defined in terms of the transitional PDF, $f(x, t)$, by

$$F(x, t) = \int_a^x f(u, t) du. \quad (20)$$

When expressed in terms of $F(x, t)$, equation (11) takes the form

$$\frac{\partial^2 F}{\partial x \partial t} = \frac{\partial}{\partial x} \left[\frac{1}{2} \frac{\partial}{\partial x} \left(g \frac{\partial F}{\partial x} \right) - \mu \frac{\partial F}{\partial x} \right] \quad (21)$$

which can be integrated with respect to x to give

$$\frac{\partial F}{\partial t} = \left[\frac{1}{2} \frac{\partial}{\partial x} \left(g \frac{\partial F}{\partial x} \right) - \mu \frac{\partial F}{\partial x} \right] + C(t) \quad (22)$$

where $C(t)$ is an arbitrary function of integration. The boundary conditions for this equation require that $F(a, t) = 0$ and $F(b, t) = 1$ which in turn require that $C(t) = 0$. Therefore $F(x, t)$ satisfies the partial differential equation

$$\frac{\partial F}{\partial t} = \frac{1}{2} \frac{\partial}{\partial x} \left(g \frac{\partial F}{\partial x} \right) - \mu \frac{\partial F}{\partial x} \quad (23)$$

with Dirichlet boundary conditions $F(a, t) = 0$ and $F(b, t) = 1$. The initial condition $F(x, t_k)$ for a transition from (X_k, t_k) is constructed from the definition (12) to obtain

$$F(x, t_k) = \begin{cases} 0 & x < X_k, \\ 1/2 & x = X_k, \\ 1 & x > X_k. \end{cases} \quad (24)$$

One important advantage of this approach is that the delta function initial condition required in the computation of transitional PDF is now replaced by a step function initial condition in the computation of the transitional CDF. The latter has a precise numerical representation whereas the delta function (12) must be approximated⁶.

It is clear that the estimation procedure based on the numerical solution of the Fokker-Planck equation is closest in spirit to EML, and it is perhaps surprising, therefore, that it has not enjoyed much support in the literature particularly given that modern numerical methods allow the transitional PDF to be recovered to high accuracy. Although the construction of the tri-diagonal system of equations appears to be algebraically complicated, the component parts of these equations are developed from well-known central-difference formulae. Once constructed, the numerical scheme does not require that any special constitutive assumptions be made about the nature of the drift and diffusion functions, for example, that they are affine functions of state. In other words, the method is completely general in that any SDE can be easily accommodated without further work. In conclusion it should be noted that this approach requires a moderate level of computational effort to implement properly. In the main, this stems from the delta function initial condition and the concomitant implications for the resolution of the solution in state space. The numerical effort in this estimation algorithm is consumed by the repeated solution of a large system of simultaneous equations, albeit a tri-diagonal system of equations.

3.2 Discrete maximum likelihood

The central idea in the “discrete” maximum likelihood (DML) approach to parameter estimation is the approximation of the transitional PDF by a closed-form expression involving the parameters of the SDE. The traditional (and most direct) way to achieve this objective is to use the Euler-Maruyama algorithm with one step of duration $\Delta = (t_{k+1} - t_k)$ to generate the approximate solution

$$X = X_k + \mu(X_k; \boldsymbol{\theta}) \Delta + g(X_k; \boldsymbol{\theta}) \varepsilon_k \quad (25)$$

⁶Preliminary research indicates that formulating the problem in terms of the transitional CDF is a promising line of enquiry.

of equation (1) where $\varepsilon_k \sim N(0, \Delta)$. The transitional PDF of X is therefore approximated by the Normal PDF with mean value $(X_k + \mu(X_k; \boldsymbol{\theta})\Delta)$ and variance $g^2(X_k; \boldsymbol{\theta})\Delta$. Thus the simplest version of DML replaces the true transitional PDF $f(X_{k+1} | X_k; \boldsymbol{\theta})$ (required by EML) in expression (2) by

$$\frac{1}{g(X_k; \boldsymbol{\theta}) \sqrt{2\pi\Delta}} \exp \left[- \frac{(X_{k+1} - X_k - \mu(X_k; \boldsymbol{\theta}) \Delta)^2}{2g^2(X_k; \boldsymbol{\theta}) \Delta} \right]. \quad (26)$$

However, the estimates of the parameters of the SDE obtained using this DML approach are inconsistent for any fixed sampling interval because of the bias introduced by discretisation. Broze, Scaillet and Zakoïan (1998) provide a detailed discussion of this bias. Although the DML estimator converges to the EML estimator as the sampling interval approaches zero (Florens-Zmirou, 1989), this asymptotic result may not be sufficiently accurate at the sampling frequencies for which economic and financial data are usually available.

Note that for some common SDEs, including those governing the CIR and OU processes, this direct implementation of the DML approach leads to a negative log-likelihood function which can be minimised without the need for a numerical optimisation routine. In other words, analytical expressions can be obtained for the parameters of the SDE that minimise the negative log-likelihood function based on simple DML. The relevant details are provided in Appendix 3 for the CIR and OU processes. While at first sight this finessing of the traditional DML algorithm might seem to be nothing more than an exercise in algebra, the traditional DML algorithm is intrinsic to various parameter estimation procedures to be discussed in this paper, *e.g.* Indirect Estimation (IE). The ability to “write down” optimal parameter estimates based on the traditional DML approach without the need for numerical optimisation has a dramatic impact on the accuracy and numerical efficiency of these estimation procedures.

A variety of more accurate discrete approximations to the transitional PDF of the SDE (1) have been suggested, but only two of these will be outlined here. See Durham and Gallant (2002) for an excellent summary of these approaches.

Milstein variant Elerian (1998) developed a more accurate implementation of DML based on the integration of equation (1) using the Milstein algorithm with one step of duration $\Delta = (t_{k+1} - t_k)$. The approximate solution of (1) generated in this way is

$$X = X_k + \mu(X_k; \boldsymbol{\theta}) \Delta + g(X_k; \boldsymbol{\theta}) \varepsilon_k + \frac{g(X_k; \boldsymbol{\theta})}{2} \frac{dg(X_k; \boldsymbol{\theta})}{dx} (\varepsilon_k^2 - \Delta) \quad (27)$$

where $\varepsilon_k \sim N(0, \Delta)$. This equation is now treated as a mapping defining the random variable X in terms of the random variable $Y = \varepsilon_k$. When expressed in terms of the constants A_k , B_k and C_k defined by

$$A_k = \frac{\Delta}{2} g(X_k; \boldsymbol{\theta}) g'(X_k; \boldsymbol{\theta}), \quad B_k = \frac{1}{\Delta [g'(X_k; \boldsymbol{\theta})]^2}, \quad C_k = X_k + \mu(X_k; \boldsymbol{\theta}) \Delta - A_k(1 + B_k),$$

the mapping embodied in the Milstein solution (27) takes the simplified form

$$\left(\frac{Y}{\sqrt{\Delta}} + \text{sign}(A_k)\sqrt{B_k} \right)^2 = \frac{(X - C_k)}{A_k}. \quad (28)$$

The transitional PDF of X may be constructed from the mapping (28) by first noting that each value of X arises from two values of Y , namely

$$Y_1 = -\sqrt{\Delta} \left[\text{sign}(A_k)\sqrt{B_k} + \sqrt{\frac{X - C_k}{A_k}} \right], \quad Y_2 = -\sqrt{\Delta} \left[\text{sign}(A_k)\sqrt{B_k} - \sqrt{\frac{X - C_k}{A_k}} \right]. \quad (29)$$

The usual formula for the PDF, f_X , of X in terms of the PDF, f_Y , of Y gives immediately

$$f_X = f_{Y_1} \left| \frac{dY_1}{dX} \right| + f_{Y_2} \left| \frac{dY_2}{dX} \right| = \frac{1}{2} \sqrt{\frac{\Delta}{A_k(X - C_k)}} \frac{1}{\sqrt{2\pi\Delta}} \left(\exp \left[-\frac{Y_1^2}{2\Delta} \right] + \exp \left[-\frac{Y_2^2}{2\Delta} \right] \right).$$

It is now straightforward algebra to replace Y_1 and Y_2 by formulae (29) to get

$$f_X = \frac{e^{-B_k/2}}{\sqrt{2\pi A_k(X - C_k)}} \exp \left[-\frac{(X - C_k)}{2A_k} \right] \cosh \left(\sqrt{\frac{B_k(X - C_k)}{A_k}} \right). \quad (30)$$

Elerian (1998) simply replaces the true transitional PDF $f(X_{k+1} | X_k; \boldsymbol{\theta})$ (required by EML) in expression (2) with the transitional PDF given by expression (30).

It seems not to have been recognised in the existing literature that there may be a problem with the Milstein variant of DML. The derivation of expression (30) requires that the argument of the hyperbolic cosine be real valued. This in turn requires that $B_k(X_{k+1} - C_k)/A_k \geq 0$. There is, however, no guarantee that this condition will hold because the observation X_{k+1} may not arise from a realisation of ε_k . For example, when the Milstein variant is applied to a transition from X_k of the CIR process, the constants A_k , B_k and C_k become

$$A_k = \frac{\sigma^2 \Delta}{4}, \quad B_k = \frac{4X_k}{\sigma^2 \Delta}, \quad C_k = \Delta \left[\alpha(\theta - X_k) - \frac{\sigma^2}{4} \right].$$

Since A_k and B_k are positive, the scheme experiences difficulties whenever $X_{k+1} < C_k$, that is, whenever two successive datums satisfy

$$X_{k+1} < \Delta \left[\alpha(\theta - X_k) - \frac{\sigma^2}{4} \right].$$

To overcome this problem, the density (30) can be replaced by the numerically robust expression

$$f_X = \frac{e^{-B_k/2}}{\sqrt{2\pi |A_k| |X - C_k|}} \exp \left[-\frac{|X - C_k|}{2|A_k|} \right] \cosh \left(\sqrt{\frac{B_k |X - C_k|}{|A_k|}} \right) \quad (31)$$

which is identical to expression (30) whenever $A_k(X - C_k) \geq 0$. Clearly this decision has repercussions for the quality of the estimation procedure since any observation X_{k+1} for which $(X_{k+1} - C_k)/A_k < 0$ will behave as though $C_k + \text{sign}(A_k)|X_{k+1} - C_k|$ was observed and not X_{k+1} . Moreover, the frequency with which this happens will change in an unpredictable way as the parameters $\boldsymbol{\theta}$ are modified by the optimisation procedure.

Local linearisation Shoji and Ozaki (1998) develop another approximation which they call local linearisation. Their approach requires a constant diffusion function. This objective can be achieved in equation (1) by changing the state variable from X to Y where

$$Y = \int^X \frac{du}{g(u; \boldsymbol{\theta})}. \quad (32)$$

Ito's lemma

$$dY = \left(\mu(X; \boldsymbol{\theta}) \frac{dy}{dx} + \frac{g^2(X, \boldsymbol{\theta})}{2} \frac{d^2y}{dx^2} \right) dt + g(X; \boldsymbol{\theta}) dW$$

may now be used to show that $Y(t)$ satisfies the stochastic differential equation

$$dY = \hat{\mu}(Y; \boldsymbol{\theta}) dt + dW, \quad \hat{\mu}(Y; \boldsymbol{\theta}) = \frac{\mu(X; \boldsymbol{\theta})}{g(X; \boldsymbol{\theta})} - \frac{1}{2} \frac{dg(X; \boldsymbol{\theta})}{dx} \quad (33)$$

where it is assumed that occurrences of X in the definition of $\hat{\mu}(Y; \boldsymbol{\theta})$ are replaced by a function of Y via the mapping defined by equation (32). The Taylor series expansion of the drift function about Y_k is

$$\hat{\mu}(Y; \boldsymbol{\theta}) = \hat{\mu}(Y_k; \boldsymbol{\theta}) + \frac{d\hat{\mu}(Y_k; \boldsymbol{\theta})}{dY} (Y - Y_k) + \frac{(Y - Y_k)^2}{2} \frac{d^2\hat{\mu}(Y_k; \boldsymbol{\theta})}{dY^2} + O(Y - Y_k)^3.$$

If Y is now taken to be the solution of equation (33), then $(Y - Y_k)^2 = \Delta + O(\Delta^{3/2})$ and $(Y - Y_k)^3 = O(\Delta^{3/2})$, and the expansion of $\hat{\mu}(Y; \boldsymbol{\theta})$ about Y_k to order Δ now becomes

$$\hat{\mu}(Y; \boldsymbol{\theta}) = \hat{\mu}(Y_k; \boldsymbol{\theta}) + \hat{\mu}'(Y_k; \boldsymbol{\theta})(Y - Y_k) + \frac{\Delta}{2} \hat{\mu}''(Y_k; \boldsymbol{\theta}). \quad (34)$$

With this approximation of the drift process, the random variable Y now satisfies the SDE

$$dY = \left(\hat{\mu}(Y_k; \boldsymbol{\theta}) + \hat{\mu}'(Y_k; \boldsymbol{\theta})(Y - Y_k) + \frac{\Delta}{2} \hat{\mu}''(Y_k; \boldsymbol{\theta}) \right) dt + dW, \quad (35)$$

which conforms to an OU process with parameter specification

$$\alpha = -\hat{\mu}'(Y_k; \boldsymbol{\theta}), \quad \beta = Y_k - \frac{2\hat{\mu}(Y_k; \boldsymbol{\theta}) + \Delta \hat{\mu}''(Y_k; \boldsymbol{\theta})}{2\hat{\mu}'(Y_k; \boldsymbol{\theta})}, \quad \sigma = 1. \quad (36)$$

In conclusion, the transitional PDF of Y is approximated by the transitional PDF of the OU process with parameters given by equations (36). The transitional PDF of X is now computed from that of the OU process by the usual formula

$$f_X = f_Y \frac{dY}{dX} = \frac{f_Y}{g(X; \boldsymbol{\theta})}$$

where f_Y is given by expression (8) with α , β and σ taking the values in equation (36).

In a comparative study of various DML approximation methods Durham and Gallant (2002) found the local linearisation method to be among the most accurate of this class of estimators. They also suggest that almost all of the methods perform more accurately if the SDE is converted to one with a unit diffusion by the transformation in (32) even if their implementation does not necessarily require it. This is most likely due to the fact that the transitional PDF of the new variable Y is closer to

that of a Gaussian distribution than that of X . Note, however, that it must be possible to compute $\hat{\mu}(y; \boldsymbol{\theta})$ in closed form to take advantage of this procedure. If the derivatives of $\hat{\mu}(y; \boldsymbol{\theta})$ are taken numerically then it is likely that some of the advantage of this procedure will be sacrificed.

In conclusion, the major advantages of traditional DML are its ease of implementation, speed and the fact that it does not impose any restrictions on the nature of the drift and diffusion functions of the underlying SDE. These features have led to the popularity of DML both as an estimation procedure in its own right and also as a component of more complex methods. Of course, the accuracy of the traditional implementation of DML suffers from discretisation bias for the size of sampling interval commonly encountered. The variants of DML try to reduce this bias, but at the expense of some of the simplicity both in terms of coding and in terms of the need for computing gradients of the drift and diffusion functions.

3.3 Hermite polynomial expansion approaches

Aït-Sahalia (2002) develops two estimation procedures in which the unknown transitional PDF is approximated by means of an expansion based on modified Hermite polynomials. The modified Hermite polynomial of degree n , here denoted conveniently by $H_n(z)$ but not to be confused with the conventional Hermite polynomial, is defined by

$$H_n(z) = e^{z^2/2} \frac{d^n}{dz^n} (e^{-z^2/2}) \quad n \geq 0 \quad (37)$$

and satisfies the important orthogonality property

$$\int_{-\infty}^{\infty} \phi(z) H_n(z) H_m(z) dz = \begin{cases} 0 & n \neq m \\ n! & n = m \end{cases} \quad (38)$$

where $\phi(z)$ is the PDF of the standard normal distribution. This property of modified Hermite polynomials can be established directly from the identity

$$\int_{-\infty}^{\infty} \phi(z) H_n(z) H_m(z) dz = \int_{-\infty}^{\infty} \frac{d^n}{dz^n} (\phi(z)) H_m(z) dz = (-1)^n \int_{-\infty}^{\infty} \phi(z) \frac{d^n H_m(z)}{dz^n} dz \quad (39)$$

which is constructed by applying integration by parts n times to the middle integral and differentiating the modified Hermite polynomial on each occasion. Since $H_m(z)$ is a polynomial of degree m , then the right hand integral in equation (39) has value zero when $m < n$. Similarly, symmetry demands that the value of this integral is also zero when $m > n$. Thus result (38) is established when $n \neq m$. The result when $n = m$ follows by first noting that $d^n H_n(z)/dz^n = n!$, and then taking advantage of the fact that $\phi(z)$ is a PDF.

Both of the estimation algorithms proposed by Aït-Sahalia (2002) begin by transforming the variable in the original SDE (1) from X to

$$Y = \int^X \frac{du}{g(u; \boldsymbol{\theta})}. \quad (40)$$

The procedure is identical to that used by Shoji and Ozaki (1998) for the local linearisation algorithm described in Subsection 3.2. Briefly, Ito's lemma is used to show that Y satisfies the SDE

$$dY = \hat{\mu}(Y; \boldsymbol{\theta}) dt + dW \quad (41)$$

with drift specification

$$\hat{\mu}(Y; \boldsymbol{\theta}) = \frac{\mu(X; \boldsymbol{\theta})}{g(X; \boldsymbol{\theta})} - \frac{1}{2} \frac{\partial g(X; \boldsymbol{\theta})}{\partial X} \quad (42)$$

where all occurrences of X on the right hand side of equation (42) are understood to be replaced by a function of Y that is constructed by inverting the monotonically increasing mapping defined by equation (40). This strategy serves two purposes; first it reduces the technical complexity of the problem from two arbitrary functions in the original SDE written for X to a single arbitrary function in the SDE written for Y , and second, the procedure provides a canonical form into which many SDE can be reduced by an appropriate change of random variable.

Let Y_k be the value of Y corresponding to X_k , and define the auxiliary random variable Z by

$$Z = \frac{Y - Y_k}{\sqrt{\Delta}} \quad (43)$$

where $\Delta = (t_{k+1} - t_k)$ is the interval between observations. When Δ is small so that the current distribution of Z is not close to its marginal distribution - the most common situation to occur in practice - the transitional PDF of Z at t_{k+1} , namely the solution $f(Z | Y_k; \boldsymbol{\theta})$ to the Fokker-Planck equation for Z , is intuitively well approximated by the normal distribution with mean value $\hat{\mu}(Y_k; \boldsymbol{\theta})\sqrt{\Delta}$ and unit variance. This suggests that the transitional PDF of Z at t_{k+1} can be well approximated by the Fourier-Hermite series expansion

$$f(z) = \phi(z) \sum_{n=0}^{\infty} \eta_n H_n(z). \quad (44)$$

Finite Hermite expansion Ait-Sahalia's first procedure involves truncating the infinite sum in equation (44) so as to approximate the transitional PDF of Z at t_{k+1} by the finite Hermite expansion

$$f(z | Y_k; \boldsymbol{\theta}) = \phi(z) \sum_{j=0}^J \eta_j(\Delta, Y_k; \boldsymbol{\theta}) H_j(z). \quad (45)$$

The transitional PDF of X can be constructed from that of Z in the usual way to get

$$f(x | X_k; \boldsymbol{\theta}) = \frac{f(z | Y_k; \boldsymbol{\theta})}{\sqrt{\Delta} g(X; \boldsymbol{\theta})}, \quad (46)$$

and so the efficacy of this estimation procedures depends on the ease with which the coefficients $\eta_0(\Delta, Y_k; \boldsymbol{\theta}), \dots, \eta_j(\Delta, Y_k; \boldsymbol{\theta})$ can be calculated. By multiplying equation (45) by $H_m(z)$ and integrating the result over \mathbb{R} , it follows immediately from the orthogonality property (38) that the coefficients of this expansion are

$$\eta_j(\Delta, Y_k; \boldsymbol{\theta}) = \frac{1}{j!} \int_{-\infty}^{\infty} H_j(z) f(z | Y_k; \boldsymbol{\theta}) dz = \frac{1}{j!} \text{E} [H_j(Z) | Y(t_k) = Y_k; \boldsymbol{\theta}]. \quad (47)$$

Since $H_0(z) = 1$ then $\eta_0(\Delta, Y_k; \boldsymbol{\theta}) = 1$, which makes intuitive sense as the first term in the Hermite expansion of the transitional PDF of Z must be the standard normal. The additional terms in the expansion serve to refine this approximation, but unfortunately the coefficients of these additional terms are more difficult to determine.

On the assumption that $\eta_j(\Delta, Y_k; \boldsymbol{\theta})$ is K times continuously differentiable with respect to Δ in a neighbourhood of the origin, then $\eta_j(\Delta, Y_k; \boldsymbol{\theta})$ has MacLaurin expansion

$$\eta_j(\Delta, Y_k; \boldsymbol{\theta}) = \sum_{n=0}^K \frac{\Delta^n}{n!} \frac{d\eta_j^n(0)}{d\Delta^n} + O(\Delta^{K+1}). \quad (48)$$

The infinitesimal operator⁷ of Z (which is identical to the infinitesimal operator of Y) is defined by

$$\mathcal{A}_{\boldsymbol{\theta}}(\psi) = \widehat{\mu}(y; \boldsymbol{\theta}) \frac{d\psi}{dy} + \frac{1}{2} \frac{d^2\psi}{dy^2} \quad (49)$$

and has the property that it expresses the time derivative of an expected value as an expected value taken in state space. This property is now used to replace derivatives with respect to Δ in equation (48) with expectations taken in state space to obtain

$$\eta_j(\Delta, Y_k; \boldsymbol{\theta}) = \frac{1}{j!} \sum_{n=0}^K \lim_{\Delta \rightarrow 0^+} E[\mathcal{A}_{\boldsymbol{\theta}}^n[H_j(z)]] \frac{\Delta^n}{n!} + O(\Delta^{K+1}). \quad (50)$$

Note that as $\Delta \rightarrow 0^+$, the transitional PDF approaches a delta function, and therefore the integral defining the expectation in equation (50) may be replaced by $\mathcal{A}_{\boldsymbol{\theta}}^n[H_j(z)]$ evaluated at $z = Z_k$, or equivalently at $y = Y_k$, to get

$$\eta_j(\Delta, Y_k; \boldsymbol{\theta}) = \frac{1}{j!} \sum_{n=0}^K \lim_{y \rightarrow Y_k} \mathcal{A}_{\boldsymbol{\theta}}^n[H_j(z)] \frac{\Delta^n}{n!} + O(\Delta^{K+1}). \quad (51)$$

This strategy is used to develop explicit expressions for the coefficients $\eta_1(\Delta, Y_k; \boldsymbol{\theta}), \dots, \eta_J(\Delta, Y_k; \boldsymbol{\theta})$ which in turn allows the Fourier-Hermite expansion of the transitional PDF of X to be constructed from identity (46). In practice, these coefficients are complicated expressions involving Δ , $\widehat{\mu}(Y_k; \boldsymbol{\theta})$ and derivatives of $\widehat{\mu}(y; \boldsymbol{\theta})$ evaluated at $y = Y_k$. The technical details are unpleasant, and so only a few steps in the calculations are given. For example, the components

$$\begin{aligned} \mathcal{A}_{\boldsymbol{\theta}}[H_1(z)] &= -\frac{\widehat{\mu}}{\sqrt{\Delta}}, & \mathcal{A}_{\boldsymbol{\theta}}^2[H_1(z)] &= -\frac{2\widehat{\mu}\widehat{\mu}' + \widehat{\mu}''}{2\sqrt{\Delta}}, \\ \mathcal{A}_{\boldsymbol{\theta}}^3[H_1(z)] &= -\frac{4\widehat{\mu}^2\widehat{\mu}'' + 4\widehat{\mu}(\widehat{\mu}')^2 + 4\widehat{\mu}\widehat{\mu}''' + 6\widehat{\mu}'\widehat{\mu}'' + \widehat{\mu}''''}{4\sqrt{\Delta}} \end{aligned}$$

are sufficient to compute $\eta_1(\Delta, Y_k; \boldsymbol{\theta})$ to $o(\Delta^3)$. Ait-Sahalia (2002) uses this strategy to develop

⁷A general introduction to the infinitesimal operator is given in Appendix 4.

explicit expressions for $\eta_1(\Delta, Y_k; \boldsymbol{\theta}), \dots, \eta_6(\Delta, Y_k; \boldsymbol{\theta})$ to $o(\Delta^3)$. The coefficients are

$$\begin{aligned}
\eta_0 &= 1, \\
\eta_1 &= -\widehat{\mu}\Delta^{1/2} - [2\widehat{\mu}\widehat{\mu}' + \widehat{\mu}''] \Delta^{3/2}/4 \\
&\quad - [4\widehat{\mu}(\widehat{\mu}')^2 + 4\widehat{\mu}^2\widehat{\mu}'' + 6\widehat{\mu}'\widehat{\mu}''' + 4\widehat{\mu}\widehat{\mu}'''' + \widehat{\mu}''''] \Delta^{5/2}/24 \\
\eta_2 &= [\widehat{\mu}^2 + \widehat{\mu}'] \Delta/2 + [6\widehat{\mu}^2\widehat{\mu}' + 4(\widehat{\mu}')^2 + 7\widehat{\mu}\widehat{\mu}'' + 2\widehat{\mu}'''] \Delta^2/12 \\
&\quad + [28\widehat{\mu}^2(\widehat{\mu}')^2 + 28\widehat{\mu}^2\widehat{\mu}''' + 16(\widehat{\mu}')^3 + 16\widehat{\mu}^3\widehat{\mu}'' + 88\widehat{\mu}\widehat{\mu}'\widehat{\mu}'' \\
&\quad + 21(\widehat{\mu}'')^2 + 32\widehat{\mu}'\widehat{\mu}'''' + 16\widehat{\mu}\widehat{\mu}'''' + 3\widehat{\mu}'''''] \Delta^3/96 \\
\eta_3 &= -[\widehat{\mu}^3 + 3\widehat{\mu}\widehat{\mu}' + \widehat{\mu}''] \Delta^{3/2}/6 - [12\widehat{\mu}^3\widehat{\mu}' + 28\widehat{\mu}(\widehat{\mu}')^2 \\
&\quad + 22\widehat{\mu}^2\widehat{\mu}'' + 24\widehat{\mu}'\widehat{\mu}'' + 14\widehat{\mu}\widehat{\mu}'''' + 3\widehat{\mu}''''] \Delta^{5/2}/48 \\
\eta_4 &= [\widehat{\mu}^4 + 6\widehat{\mu}^2\widehat{\mu}' + 3(\widehat{\mu}')^2 + 4\widehat{\mu}\widehat{\mu}'' + \widehat{\mu}'''] \Delta^2/24 \\
&\quad + [20\widehat{\mu}^4\widehat{\mu}' + 50\widehat{\mu}^3\widehat{\mu}'' + 100\widehat{\mu}^2(\widehat{\mu}')^2 + 50\widehat{\mu}^2\widehat{\mu}'''' + 23\widehat{\mu}\widehat{\mu}'''' \\
&\quad + 180\widehat{\mu}\widehat{\mu}'\widehat{\mu}'' + 40(\widehat{\mu}')^3 + 34(\widehat{\mu}'')^2 + 52\widehat{\mu}'\widehat{\mu}'''' + 4\widehat{\mu}'''''] \Delta^3/240 \\
\eta_5 &= -[\widehat{\mu}^5 + 10\widehat{\mu}^3\widehat{\mu}' + 15\widehat{\mu}(\widehat{\mu}')^2 + 10\widehat{\mu}^2\widehat{\mu}'' + 10\widehat{\mu}'\widehat{\mu}'' + 5\widehat{\mu}\widehat{\mu}'''' + \widehat{\mu}^4] \Delta^{5/2}/120 \\
\eta_6 &= [\widehat{\mu}^6 + 15\widehat{\mu}^4\widehat{\mu}' + 15(\widehat{\mu}')^3 + 20\widehat{\mu}^3\widehat{\mu}'' + 15\widehat{\mu}'\widehat{\mu}'''' + 45\widehat{\mu}^2(\widehat{\mu}')^2 \\
&\quad + 10(\widehat{\mu}'')^2 + 15\widehat{\mu}^2\widehat{\mu}'''' + 60\widehat{\mu}\widehat{\mu}'\widehat{\mu}'' + 6\widehat{\mu}\widehat{\mu}'''' + \widehat{\mu}'''''] \Delta^3/720
\end{aligned} \tag{52}$$

It should be noted that in Ait-Sahalia's procedure the availability of a closed form expression for $\widehat{\mu}(y; \boldsymbol{\theta})$ appears to be a crucial prerequisite for the computation of the coefficients. However, if this closed form expression is difficult (or impossible) to derive in practice, it is still possible to construct the coefficients in equation (52) by noting that the values of $\widehat{\mu}(y; \boldsymbol{\theta})$ and its derivatives at Y_k can be obtained directly from the gradient of the mapping function and the properties of the primitive drift and diffusion functions⁸ without an explicit expression for X in terms of Y . For example, $\widehat{\mu}(Y_k; \boldsymbol{\theta})$ can be obtained directly from equation (42) and its first derivative can be computed as

$$\frac{d\widehat{\mu}}{dy} = \frac{\partial\widehat{\mu}}{\partial x} \frac{dx}{dy} = g \frac{\partial\widehat{\mu}}{\partial x} = \frac{\partial\mu}{\partial x} - \frac{\mu}{g} \frac{\partial g}{\partial x} - \frac{g}{2} \frac{\partial^2 g}{\partial x^2} \tag{53}$$

where $\mu = \mu(x; \boldsymbol{\theta})$ and $g = g(x; \boldsymbol{\theta})$ are the primitive expressions for drift and diffusion. Higher derivatives of $\widehat{\mu}(y; \boldsymbol{\theta})$ can be obtained by a similar procedure.

CIR process: Let X satisfy the CIR process with drift specification $\mu(x; \boldsymbol{\theta}) = \alpha(\beta - x)$ and diffusion specification $g(x; \boldsymbol{\theta}) = \sigma\sqrt{x}$, then the new random variable defined in equation (40) is $Y = 2\sqrt{X}/\sigma$ and the underlying SDE satisfied by Y becomes

$$dY = \widehat{\mu}(Y; \boldsymbol{\theta}) dt + dW, \quad \widehat{\mu}(y; \boldsymbol{\theta}) = \left(\frac{2\alpha\beta}{\sigma^2} - \frac{1}{2} \right) \frac{1}{y} - \frac{\alpha}{2} y. \tag{54}$$

Although $\widehat{\mu}(y; \boldsymbol{\theta})$ is given by a simple closed-form expression in y which can be differentiated as required, nevertheless the coefficients in expression (52) remain cumbersome.

⁸This fact has also been noted by Ait-Sahalia in a companion working paper and by Bakshi and Ju (2005).

OU process: The OU process with drift specification $\mu(x; \boldsymbol{\theta}) = \alpha(\beta - x)$ and diffusion specification $g(x; \boldsymbol{\theta}) = \sigma$ is the more tractable of the two models used for illustrative purposes in this survey. In this case $Y = X/\sigma$ and the underlying SDE satisfied by Y becomes

$$dY = \widehat{\mu}(Y; \boldsymbol{\theta}) dt + dW, \quad \widehat{\mu}(y; \boldsymbol{\theta}) = \frac{\alpha\beta}{\sigma} - \alpha y. \quad (55)$$

The second derivative and all higher derivatives of $\widehat{\mu}(y; \boldsymbol{\theta})$ are zero which in turn simplifies considerably the coefficients of the finite Fourier-Hermite expansion of the transitional PDF of Z .

The accuracy of the Hermite polynomial procedure is controlled by the choice of the order of the modified Hermite expansion and the order of the Maclaurin series used for the expansion in Δ . As equations (52) make clear, one obvious disadvantage of this approach lies in the inherent complexity of the expressions for the Fourier-Hermite coefficients. From equations (52) it is clear that $\eta_j(\Delta, Y_k; \boldsymbol{\theta}) = O(\Delta^{j/2})$ and therefore the extension of this procedure to a higher degree of temporal accuracy would inevitably entail significantly more calculation, both in terms of the number of coefficients and the degree of accuracy required from each coefficient. Ait-Sahalia (2002) suggests that a Maclaurin series of accuracy $o(\Delta^3)$ and a Hermite expansion of order one or two is sufficient to achieve the required degree of precision in most applications. This situation can be expected to occur whenever the transitional PDF remains well approximated by a Gaussian PDF after diffusion has taken place over an interval of duration Δ . One final drawback of this approach is that the Hermite polynomials have infinite domain and consequently there is nothing in this method to force the transitional PDF of a process with semi-infinite domain (such as the CIR process) to be zero at the origin. Therefore, this procedure may leak density at the origin when estimating the transitional PDF of processes with semi-infinite domain.

Infinite Hermite expansion Ait-Sahalia's second procedure involves rewriting equation (44) as

$$f(y, t) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y - Y_k)^2}{2t}\right) \exp\left(\int_{Y_k}^y \widehat{\mu}(u) du\right) \psi(y, t). \quad (56)$$

The right hand side of equation (56) is the product of the standard normal PDF for the variable Z ($\phi(z)$ in equation (44)) expressed in terms of Y , and two further terms which will assume the role of the infinite Hermite sum in equation (44). The objective of this second procedure is to express the function $\psi(y, t)$ as a convergent power series in Δ in which the coefficients of the series capture the contribution made by the entire family of Hermite polynomials at each order in Δ . By contrast, the coefficients η_0, η_1, \dots in the infinite Hermite sum in equation (44) capture the total contribution from a single Hermite polynomial to the sum. The transitional PDF of X can be constructed from that of Y in the usual way to get

$$f(x | X_k; \boldsymbol{\theta}) = \frac{f(y | Y_k; \boldsymbol{\theta})}{g(X; \boldsymbol{\theta})}. \quad (57)$$

The analysis begins by rewriting the Fokker-Planck equation for the unit diffusion (41) as

$$\frac{1}{f} \frac{\partial f}{\partial t} = \frac{1}{2f} \frac{\partial^2 f}{\partial y^2} - \frac{\widehat{\mu}}{f} \frac{\partial f}{\partial y} - \frac{\partial \widehat{\mu}}{\partial y}. \quad (58)$$

It is a straight forward calculation to show that the solution of equation (58) may be represented by expression (56) provided $\psi(y, t)$ satisfies

$$\frac{\partial \psi}{\partial t} = \frac{1}{2} \frac{\partial^2 \psi}{\partial y^2} - \frac{y - Y_k}{t} \frac{\partial \psi}{\partial y} + \lambda \psi, \quad \lambda = -\frac{1}{2} \left(\widehat{\mu}^2 + \frac{\partial \widehat{\mu}}{\partial y} \right). \quad (59)$$

A solution to equation (59) in the form of a power series expansion in time is now sought in the form

$$\psi = \sum_{n=0}^{\infty} \frac{c_n(y) t^n}{n!} \quad (60)$$

in which the functions $c_0(y), c_1(y), \dots$ are determined by matching the coefficients of powers of t . This representation of ψ , when substituted into equation (59), leads to the identity

$$\sum_{n=1}^{\infty} \frac{c_n(y) t^{n-1}}{(n-1)!} = \frac{1}{2} \sum_{n=0}^{\infty} \frac{d^2 c_n(y) t^n}{dy^2 n!} - \frac{y - Y_k}{t} \sum_{n=0}^{\infty} \frac{dc_n(y) t^n}{dy n!} + \sum_{n=0}^{\infty} \frac{\lambda(y) c_n(y) t^n}{n!} \quad (61)$$

Some of the summations in equation (61) are re-indexed and the terms rearranged to give

$$\sum_{n=0}^{\infty} \frac{t^n}{n!} \left[c_{n+1}(y) + \frac{(y - Y_k)}{n+1} \frac{dc_{n+1}(y)}{dy} - \left(\frac{1}{2} \frac{d^2 c_n(y)}{dy^2} + \lambda(y) c_n(y) \right) \right] = -\frac{y - Y_k}{t} \frac{dc_0(y)}{dy} \quad (62)$$

from which it follows immediately that

$$\begin{aligned} \frac{dc_0(y)}{dy} &= 0, \\ c_{n+1}(y) + \frac{(y - Y_k)}{(n+1)} \frac{dc_{n+1}(y)}{dy} &= \frac{1}{2} \frac{d^2 c_n(y)}{dy^2} + \lambda(y) c_n(y). \end{aligned} \quad (63)$$

The first condition asserts that $c_0(y)$ is a constant function. In order to maintain the correct short time asymptotic expression for the transitional PDF of the process, this constant function must be taken to be $c_0(y) = 1$. Furthermore, the requirement that $\psi(y, t)$ be finite at $y = Y_k$ for all $t > 0$ forces the solution of the second condition to be

$$c_{n+1}(y) = \frac{n+1}{(y - Y_k)^{n+1}} \int_{Y_k}^y (u - Y_k)^n \left(\frac{1}{2} \frac{d^2 c_n(u)}{du^2} + \lambda(u) c_n(u) \right) du \quad n \geq 0. \quad (64)$$

In conclusion, equation (64) in combination with $c_0(y) = 1$ enables the coefficients $c_0(y), c_1(y), \dots$ to be determined recursively. With the aid of numerical packages such as *Mathematica* or *Maple* the integral in equation (64) can be computed analytically and the process can proceed without error. However, if the solution to equation (64) is to be found by numerical means then this approach may be problematic since numerical error occurring in the computation of $c_n(y)$ generates further numerical error in the computation of $c_{n+1}(y)$, that is, the system of equations may quickly become unstable. This appears to occur in the case of the CIR process with the onset of the instability being further hastened by the presence of the singularity in the modified drift specification $\widehat{\mu}$.

3.4 Simulated maximum likelihood

The method of “simulated” maximum likelihood (SML) is developed independently⁹ in Pedersen (1995) and Brandt and Santa-Clara (2002) with an alternative algorithm given in Hurn, Lindsay and Martin (2003). In each of the methods to be described, SML aims to obtain an estimate of $f(x | X_k; \theta)$ by simulation. A common feature in all the SML approaches is the use of a numerical scheme to advance the solution of the SDE from t_k to t_{k+1} in m small steps of duration $\Delta t = (t_{k+1} - t_k)/m$. Let $(X_k =) x_{k,0}^*, x_{k,1}^*, \dots, x_{k,m}^*$ denote the sequence of states accessed by the solution in this integration procedure at the respective times $(t_k =) t_{k,0}, t_{k,1}, \dots, t_{k,m}(= t_{k+1})$ where $t_{k,q} = t_k + q\Delta t$ and q takes all integer values between $q = 0$ and $q = m$. Each sequence of states from (X_k, t_k) to a final state (X, t_{k+1}) is called a “path”. Figure 2 illustrates a typical path starting at (X_k, t_k) and passing through the unobserved states $(x_{k,1}^*, t_{k,1}), (x_{k,2}^*, t_{k,2}), \dots, (x_{k,m-1}^*, t_{k,m-1})$. The final state is either enforced to be (X_{k+1}, t_{k+1}) or is left unrestricted, denoted (X, t_{k+1}) . As the relationship between state and time is clear from Figure 2, for ease of notation, time will be suppressed in the subsequent discussion of SML.

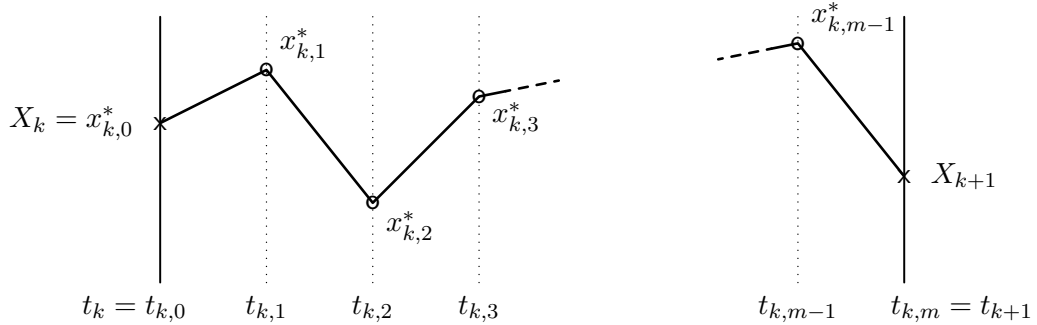


Figure 2: A simulation path connecting the observed states X_k and X_{k+1} (denoted by X) via $(m-1)$ unobserved states $x_{k,1}^*, x_{k,2}^*, \dots, x_{k,m-1}^*$ at the respective times $t_{k,1}, t_{k,2}, \dots, t_{k,m-1}$ (denoted by O).

Kernel estimate of transitional PDF The kernel estimation of the transitional PDF developed by Hurn *et al.* (2003) is the most direct implementation of SML. For ease of explanation it is expedient to start with this approach. A numerical scheme (for example Euler-Mayumara or Milstein) is used to simulate M paths of the SDE starting from X_k with the final state unrestricted. Each simulated value of the terminal state represents an independent draw from the transitional PDF. These simulated values can therefore be used to construct a nonparametric kernel density estimate of the value of the transitional density at X_{k+1} , namely,

$$f(X_{k+1} | X_k; \theta) = \frac{1}{Mh} \sum_{i=1}^M K\left(\frac{X_{k+1} - X_{k+1}^{(i)}}{h}\right) \quad (65)$$

⁹Although the paper by Pedersen appears to predate that by Brandt and Santa-Clara, the latter is essentially a revised version of the ideas developed in Santa-Clara (1995).

where K is a kernel function, $X_{k+1}^{(i)}$ is the i -th simulated value of X at time t_{k+1} and h is the kernel bandwidth¹⁰. Underlying this method is the idea that the solution of the Fokker-Planck equation at t_{k+1} is the limiting density that would be achieved by an infinite number of simulations of the SDE starting from X_k .

Despite its intuitive appeal, this direct approach suffers from all the problems associated with estimating a density function from a finite data set. Most important of these are the problems of bandwidth selection and leakage of density into state space not accessible to the process. Although substantial progress on these problems has been made in the context of kernel methods for density estimation, all the proposed improvements add to what is already a computationally expensive procedure.

Monte Carlo simulation of transitional PDF Pedersen's (1995) method is similar to the kernel procedure just described but differs crucially from it in the respect that the transitional PDF is not estimated by a kernel procedure. Each simulated path of the SDE now terminates at the penultimate step, namely at $t_{k,m-1}$. The last transition required to advance the process to X_{k+1} , is over a small interval Δt and is therefore well approximated¹¹ by the Gaussian PDF with mean value and variance determined by the penultimate state $x_{k,m-1}^*$. The value of the transitional PDF to be used in the computation of likelihood, therefore, is the average value of the M estimates of the likelihood of the transition to X_{k+1} from the penultimate state of the simulated paths. A desirable consequence of this procedure is that a nonparametric estimate of the transitional PDF is no longer required.

Importance sampling approach The Pedersen (1995) approach can be nested within a family of importance sampling estimators of transitional density (Elerian, Chib and Shephard, 2001). For the purpose of explanation, it is convenient to simplify notation further and let $\mathbf{x}^* = (x_{k,1}, \dots, x_{k,m-1})$ denote the vector of unobserved states as the stochastic process X evolves from the state X_k to the state X_{k+1} . The value of the transitional PDF at X_{k+1} satisfies the identity

$$f(X_{k+1} | X_k; \boldsymbol{\theta}) = \int f(X_{k+1}, \mathbf{x}^* | X_k; \boldsymbol{\theta}) d\mathbf{x}^* \quad (66)$$

where $f(X_{k+1}, \mathbf{x}^* | X_k)$ is the joint probability density of the final state X_{k+1} and all possible unobserved paths for the stochastic process evolving from X_k in accordance with the SDE (1). The identity (66) suggests that $f(X_{k+1} | X_k; \boldsymbol{\theta})$ may be estimated by a Monte Carlo integration in which paths \mathbf{x}^* are chosen and the value of $f(X_{k+1} | X_k; \boldsymbol{\theta})$ computed and then averaged over the paths to get an estimate of the transitional PDF.

The central idea of importance sampling is to weight the selection of paths in favour of those that are thought to make the most significant contribution to the value of the integral on the right hand side of equation (66) whilst simultaneously correcting for the distortion introduced by the new probability

¹⁰Hurn *et al.* (2003) use a Gaussian kernel with a bandwidth given by the normal reference rule (see Scott, 1992).

¹¹Recall that this approximation was also used by Jensen and Poulsen (2002) to initialise the transitional PDF in the finite-difference algorithm.

measure. Let $\psi(\mathbf{x}^* | X_{k+1}, X_k; \boldsymbol{\theta})$ denote the PDF resulting from a user-supplied selection criterion for possible paths terminating at X_{k+1} . The identity (66) may now be rewritten in equivalent algebraic form

$$f(X_{k+1} | X_k; \boldsymbol{\theta}) = \int \frac{f(X_{k+1}, \mathbf{x}^* | X_k; \boldsymbol{\theta})}{\psi(\mathbf{x}^* | X_{k+1}, X_k; \boldsymbol{\theta})} \psi(\mathbf{x}^* | X_{k+1}, X_k; \boldsymbol{\theta}) d\mathbf{x}^*. \quad (67)$$

Equation (67) now asserts that the transitional PDF $f(X_{k+1} | X_k; \boldsymbol{\theta})$ may be regarded as the expected value of the ratio $f(X_{k+1}, \mathbf{x}^* | X_k; \boldsymbol{\theta}) / \psi(\mathbf{x}^* | X_{k+1}, X_k; \boldsymbol{\theta})$ when calculated over all paths generated using the importance sampling procedure. Once again the integral (67) is estimated by Monte Carlo integration. Specifically, if M paths are drawn using the importance sampler, then the transitional PDF is taken to be the mean value of the ratio $f(X_{k+1}, \mathbf{x}^* | X_k; \boldsymbol{\theta}) / \psi(\mathbf{x}^* | X_{k+1}, X_k; \boldsymbol{\theta})$.

Intuitively, the numerator of this ratio is calculated for the path in accordance with the process underlying the SDE whereas the denominator is calculated for the same path but now for the process underlying the importance sampler. In practice, the numerator is computed as the product of likelihoods of the transitions from $x_{k,q}^*$ to $x_{k,q+1}^*$ for $q = 0 \dots m - 1$, that is

$$f(X_{k+1}, \mathbf{x}^* | X_k; \boldsymbol{\theta}) = \prod_{q=0}^{m-1} f(x_{k,q+1}^* | x_{k,q}^*)$$

where it is understood that $X_k = x_{k,0}^*$ and $X_{k+1} = x_{k,m}^*$. In this computation $f(x_{k,q+1}^* | x_{k,q}^*)$ is approximated by the traditional DML likelihood (see Subsection 3.2).

The denominator of the ratio is computed as the likelihood of the path as measured by its generating process. Of course, the form of the likelihood depends on the specification of the importance sampler. For example, the algorithm used by Pedersen (1995) may be regarded as the special case in which the importance sampler is the same scheme as that used to approximate the SDE. In this case, the likelihood of the path under both the SDE and the generating process are identical, so the ratio in equation (67) consequently collapses to $f(X_{k+1} | x_{k,m-1}^*; \boldsymbol{\theta})$.

In general, a more judicious choice for ψ can improve the efficacy of the estimation procedure. Specifically, it is beneficial to generate paths which make use of the known terminal state X_{k+1} of the path. The modified Brownian bridge proposed by Durham and Gallant (2002) is one such procedure. They suggest that the unobserved datums \mathbf{x}^* can be generated in sequence starting with $x_{k,1}^*$ and ending with $x_{k,m-1}^*$. They propose that the unobserved datum $x_{k,q}^*$ should be generated from $x_{k,q-1}^*$ and X_{k+1} by making a draw from a normal distribution with mean and variance given respectively by

$$\frac{(m-q)x_{k,q-1}^* + X_{k+1}}{(m-q+1)}, \quad g^2(x_{k,q-1}^*) \frac{(m-q)\Delta t}{(m-q+1)}. \quad (68)$$

The motivation for the choice of the parameters in equation (68) starts with the identity

$$p(x_{k,q}^* | x_{k,q-1}^*, X_{k+1}) = \frac{p(x_{k,q}^* | x_{k,q-1}^*) p(X_{k+1} | x_{k,q}^*)}{p(X_{k+1} | x_{k,q-1}^*)} \quad (69)$$

in which $p(x_{k,q}^* | x_{k,q-1}^*, X_{k+1})$ is the conditional distribution of $x_{k,q}^*$ given the current unobserved state $x_{k,q-1}^*$ and the next observed state, namely X_{k+1} . Each PDF on the right hand side of equation

(69) is approximated by a Normal PDF. The selections

$$\begin{aligned} p(x_{k,q}^* | x_{k,q-1}^*) &\sim N(x_{k,q-1}^* + \mu(x_{k,q-1}^*)\Delta t, g^2(x_{k,q-1}^*)\Delta t), \\ p(X_{k+1} | x_{k,q-1}^*) &\sim N(x_{k,q-1}^* + \mu(x_{k,q-1}^*)(m-q+1)\Delta t, g^2(x_{k,q-1}^*)(m-q+1)\Delta t) \end{aligned} \quad (70)$$

take their mean and variance from the usual Euler-Maruyama approximation of the solution to the SDE over the appropriate time intervals. The same logic would dictate that

$$p(X_{k+1} | x_{k,q}^*) \sim N(x_{k,q}^* + \mu(x_{k,q}^*)(m-q)\Delta t, g^2(x_{k,q}^*)(m-q)\Delta t).$$

However the values of $\mu(x_{k,q}^*)$ and $g^2(x_{k,q}^*)$ are unknown, and so to resolve this impasse, Durham and Gallant (2002) approximate these unknown values by $\mu(x_{k,q-1}^*)$ and $g^2(x_{k,q-1}^*)$ respectively. Consequently, the construction of $p(x_{k,q}^* | x_{k,q-1}^*, X_{k+1})$ is based on the assumption that

$$p(X_{k+1} | x_{k,q}^*) \sim N(x_{k,q}^* + \mu(x_{k,q-1}^*)(m-q)\Delta t, g^2(x_{k,q-1}^*)(m-q)\Delta t). \quad (71)$$

The Normal PDFs underlying each of the distributions in equations (70) and (71) are introduced into the conditional probability density defined in equation (69) and after some straightforward algebra, it becomes clear that

$$p(x_{k,q}^* | x_{k,q-1}^*, X_{k+1}) \sim N\left(\frac{(m-q)x_{k,q-1}^* + X_{k+1}}{(m-q+1)}, g^2(x_{k,q-1}^*)\frac{(m-q)\Delta t}{(m-q+1)}\right). \quad (72)$$

The expression for $E[x_{k,q}^*]$ proposed in equation (68) corresponds exactly to that which would be constructed by linear interpolation, through time, of the current state $x_{k,q-1}^*$ and the final state X_{k+1} and conforms with the intuitive idea that the generation procedure must progressively focus the unobserved path towards the next observed state. Although less obvious, the expression for the variance proposed in equation (68) likewise corresponds exactly to that which would be constructed by linear interpolation, through time, of the variance $g^2(x_{k,q-1}^*)$ at the current state $x_{k,q-1}^*$ and the variance (zero) at the final state X_{k+1} .

Elerian *et al.* (2001) propose drawing the unobserved states from a multivariate normal distribution with mean values chosen to maximise the likelihood of a transition from X_k to X_{k+1} under the Euler discretisation and covariance matrix given by the inverse of the negative Hessian evaluated at the mean. The calculation of the mean of this distribution generally requires the use of a numerical optimisation routine. This technique will be described in more detail in the Markov Chain Monte Carlo approach in Subsection 3.5.

Each method discussed in this subsection requires the simulation of the SDE, but differs in how the information in the paths is used to construct the estimate of the transitional PDF and thus the simulated likelihood function. Although the direct method proposed by Hurn *et al.* (2003) has intuitive appeal and is easy to implement, the need to use kernel estimates of the transitional PDF compromises its accuracy. Durham and Gallant (2002) describe a comparative study of the performance of the Pedersen (1995), the modified Brownian bridge process and the procedure suggested

by Elerian *et al.* (2001). They conclude that the importance sampler approach using the modified Brownian bridge outperforms the others. One final comment concerns the major drawback of SML. If this method is to provide comparable accuracy to the other methods surveyed here, it is bound to be computationally very expensive.

3.5 Markov Chain Monte Carlo

Bayesian analysis of discretely observed SDEs has been studied independently by Elerian, Chib and Shephard (2001), Eraker (2001), Roberts and Stramer (2001) and Jones (1998). The Bayesian approach to estimating the parameters of SDEs involves iteration of the following two steps.

1. **Data augmentation** Each transition in the data, say from t_k to t_{k+1} , is divided into m uniform intervals of size Δt and the observed data is augmented by introducing $(m - 1)$ unobserved datums at the intermediate times $t_{k,1}, \dots, t_{k,m-1}$. In total, $N(m - 1)$ unobserved datums are constructed giving an augmented sample containing $mN + 1$ data points.
2. **Drawing parameters** The likelihood of the augmented sample under traditional DML is treated as the joint distribution of the parameters θ of the SDE from which a new set of parameters are drawn.

The iterations of these two steps are assumed to converge in the sense that, after a burn-in period, the likelihood function generated by the augmented sample will be insensitive to any particular augmented data set and therefore each draw of parameters is now a draw from their true marginal distribution. The estimate of each parameter is then obtained as the sample mean of these repeated draws from the marginal distribution. Each step is now described in more detail following Elerian *et al.* (2001).

Data augmentation Suppose that a parameter vector θ and an augmented sample are available, then the first task is to update the values for the unobserved data using the Metropolis-Hastings algorithm. Elerian *et al.* (2001) recommend updating the unobserved data between two observations in random sized blocks, where the number of unobserved datums in the block, say B , is drawn from a Poisson distribution with appropriately chosen mean. It is convenient to maintain the notation used in Figure 2 in which the typical path starts at X_k , passes through the unobserved states \mathbf{x}^* before being forced to the final state X_{k+1} . For ease of notation let $X_k = x_{k,0}^*$ and $X_{k+1} = x_{k,m}^*$.

Let $x_{k,i}^*, \dots, x_{k,i+B-1}^*$ be a typical block with lefthand neighbour $x_{k,i-1}^*$ and righthand neighbour $x_{k,i+B}^*$. Note that blocks must contain at least one unobserved datum to be updated and cannot straddle the observed data points. Elerian *et al.* (2001) suggest drawing possible new values for the unobserved block from a multivariate normal distribution¹². They provide analytical expressions for the components of a Newton-Raphson iterative procedure that can be used to find the mean and

¹²Chib and Shephard (2001) suggest that it may be possible to use the modified Brownian bridge proposed by Durham

covariance matrix of this multivariate normal. The mean maximises the log-likelihood of going from $x_{k,i-1}^*$ to $x_{k,i+B}^*$ under the Euler discretisation, and the covariance matrix is given by the inverse of the negative Hessian evaluated at the mean. Note that, in practice, there is no guarantee that the negative Hessian matrix will be positive definite at the computed maximum. To fix this problem, a generalised Choleski decomposition is used in which the negative Hessian is factored as LDL^T , where L is a lower triangle matrix and D is a diagonal matrix. By replacing each nonzero entry of D by its absolute value, the negative Hessian is forced to be positive definite.

Once the proposed replacement block $x_{k,i}^{*p}, \dots, x_{k,i+B-1}^{*p}$ is generated, the next step is to decide whether or not to replace the current block with this new block. The Metropolis-Hastings algorithm states that the current block should be replaced with probability

$$P = \min \left[\frac{f(x_{k,i}^{*p}, \dots, x_{k,i+B-1}^{*p} | x_{t,i-1}^*, x_{t,i+B}^*; \boldsymbol{\theta}) \psi(x_{k,i}^*, \dots, x_{k,i+B-1}^* | x_{t,i-1}^*, x_{t,i+B}^*; \boldsymbol{\theta})}{f(x_{k,i}^*, \dots, x_{k,i+B-1}^* | x_{t,i-1}^*, x_{t,i+B}^*; \boldsymbol{\theta}) \psi(x_{k,i}^{*p}, \dots, x_{k,i+B-1}^{*p} | x_{t,i-1}^*, x_{t,i+B}^*; \boldsymbol{\theta})}, 1 \right] \quad (73)$$

where

$$f(x_{k,i}^*, \dots, x_{k,i+B-1}^* | x_{t,i-1}^*, x_{t,i+B}^*; \boldsymbol{\theta}) \propto \prod_{q=i-1}^{i+B-1} N(x_{t,q}^* + \mu(x_{t,q}^*)\Delta t, g^2(x_{t,q}^*)\Delta t)$$

provides an estimate of the likelihood of the process arising from the Euler discretisation, and $\psi(x_{k,i}^{*p}, \dots, x_{k,i+B-1}^{*p} | x_{t,i-1}^*, x_{t,i+B}^*; \boldsymbol{\theta})$ provides the same estimate for the proposal generating density, in this case the multivariate normal as described previously.

Drawing parameters The next stage in the MCMC procedure involves drawing a new set of parameters, conditioned on the augmented sample. In the absence of an informed prior, the likelihood of the augmented sample

$$f(X^*, X | \boldsymbol{\theta}) = \prod_{k=0}^{N-1} \left(\prod_{q=0}^{m-1} f(x_{k,q+1}^* | x_{k,q}^*; \boldsymbol{\theta}) \right) \quad (74)$$

is treated as a PDF from which the next set of parameters $\boldsymbol{\theta}$ are to be drawn. For simple models, including SDEs commonly encountered in the finance literature, drawing new parameters can be straightforward. For more complex models, however, it may be difficult to draw from the marginal distribution of the parameters. In these cases it is necessary to use an accept-reject strategy to draw parameters from the marginal distribution (see Chib and Greenberg, 1995).

One important class of SDE, which occurs frequently in finance, divides the parameters $\boldsymbol{\theta}$ into two disjoint sets, namely, parameters $\boldsymbol{\theta}_\mu$ which control the drift function and a single volatility control parameter, say σ which scales the diffusion. In such models, the volatility control parameter conditioned on $\boldsymbol{\theta}_\mu$, is inverse-Gamma distributed (Zellner, 1971). Moreover, if the drift specification is itself a linear function of the parameters $\boldsymbol{\theta}_\mu$, then the parameters of the drift function are multivariate student-t distributed. The CIR and OU processes only fall into this category if the definition of and Gallant (2002), introduced in Section 3.4, as an alternative method for data augmentation. In our experience, this approach proved less satisfactory.

the drift function is modified to $\mu(x) = \kappa - \alpha x$, where $\kappa = \beta\alpha$. A procedure for drawing from the marginal distributions of the parameters for these processes is outlined in Appendix 5.

Initialisation The MCMC approach to estimating the parameters of SDEs is initialised by selecting starting values for the parameter vector θ and constructing unobserved states by linear interpolation between the observed states. After the burn-in period, the effect of the initial conditions has decayed to the point beyond which the likelihood is insensitive to any particular realisation of the augmented sample.

To conclude, by its very nature the MCMC method is computationally intensive. Not only do the augmented samples have to be generated by simulation, many samples are required to construct the distribution of the estimates of the model parameters. Furthermore, the generation of unobserved states can be problematic and requires intervention by the user. A further drawback is that the parameter drawing procedure is problem specific, that is, the marginal distributions for the parameters are known in close form only for a few common SDEs. Despite these drawbacks, MCMC shows good accuracy and extends naturally to multivariate models, including those with latent factors.

4 Sample DNA matching

The second broad class of estimators, highlighted in the right-hand column of Figure 1, estimate parameters by aligning user-defined features of the model with those of the data. The most obvious features of the data to match are the moments although there are a variety of features that are proposed by estimators in this class. For this reason, the rather eclectic title of “sample DNA matching” has been chosen for this group of estimators.

4.1 General method of moments

The general method of moments (GMM) developed by Hansen (1982) has been applied to the estimation of the parameters of SDEs. The crux of this method is the specification of a number of moment conditions¹³ $\psi_1(X; \theta), \dots, \psi_K(X; \theta)$, so that at the true parameter values, θ_{true} , each moment condition satisfies

$$\mathbb{E}[\psi_j(X_k; \theta_{\text{true}})] = 0 \quad j = 1, \dots, K. \quad (75)$$

Let $\Psi(X; \theta)$ be the $K \times N$ matrix of sample values of the functions evaluated at θ

$$\Psi(X; \theta) = \begin{bmatrix} \psi_1(X_1; \theta) & \dots & \psi_1(X_N; \theta) \\ \vdots & \ddots & \vdots \\ \psi_K(X_1; \theta) & \dots & \psi_K(X_N; \theta) \end{bmatrix} \quad (76)$$

¹³In order to identify the parameters, the number of moment conditions must be at least as large as the number of parameters to be estimated.

The GMM estimate of the parameter vector $\hat{\theta}$ is obtained by minimising the objective function

$$J(\theta) = \text{E} [\Psi(X; \theta)]^T \Omega \text{E} [\Psi(X; \theta)] \quad (77)$$

where Ω is a $(K \times K)$ positive definite weighting matrix yet to be determined. Note that $\text{E}[\Psi(X; \theta)]$ is a $(K \times 1)$ vector containing the expected value of the rows in $\Psi(X; \theta)$, that is, the vector contains the sample analogues of the moment conditions (75) for the current estimate of the parameter vector.

Hansen (1982) shows that the optimal weighting matrix has the form

$$\Omega = \left[\frac{1}{N} \Psi(X; \theta_{\text{true}}) \Psi(X; \theta_{\text{true}})^T \right]^{-1} \quad (78)$$

which is guaranteed to be positive definite but is an infeasible choice in practice as θ_{true} is unknown. Hansen, Heaton and Yaron (1996) outline a variety of algorithms for constructing the weighting matrix. An iterative procedure is adopted here. An initial estimate of the parameters is obtained by minimising the objective function (77) with the identity matrix as the weighting matrix. These consistent estimates of the parameters are then used to construct a new weighting matrix and the objective function is minimised with this parameter-dependent weighting matrix until convergence.

As already noted, the central element of GMM is the specification of the moment conditions. In the estimation of the parameters of SDEs, moment conditions have been obtained by a number of different routes. Chan, Karolyi, Longstaff, and Sanders (1992) derive a set of approximate moment conditions from a first order Euler-Maruyama discretisation of the SDE. As noted in Subsection 3.2, this discretisation implies that $X_{k+1} \sim \text{N}(X_k + \mu(X_k; \theta)\Delta, g^2(X_k; \theta)\Delta)$. Accordingly, the two moment conditions

$$\begin{aligned} \psi_1(X; \theta) : & \quad \text{E} [X_{k+1} - X_k - \Delta\mu(X_k; \theta)] = 0, \\ \psi_2(X; \theta) : & \quad \text{E} [(X_{k+1} - X_k - \Delta\mu(X_k; \theta))^2 - \Delta g^2(X_k; \theta)] = 0 \end{aligned} \quad (79)$$

follow directly from the Euler-Maruyama discretisation of the SDE. These primitive moment conditions may be used to generate any number of further conditions using the generating relationships

$$\psi_{2j+1}(X; \theta) = \psi_1(X; \theta)X^j \quad \psi_{2j+2}(X; \theta) = \psi_2(X; \theta)X^j \quad j = 1, 2, \dots$$

For example, the next pair of moment conditions are

$$\begin{aligned} \psi_3(X; \theta) : & \quad \text{E} [(X_{k+1} - X_k - \Delta\mu(X_k; \theta))X_k] = 0, \\ \psi_4(X; \theta) : & \quad \text{E} [((X_{k+1} - X_k - \Delta\mu(X_k; \theta))^2 - \Delta g^2(X_k; \theta))X_k] = 0. \end{aligned} \quad (80)$$

Another way to generate moment conditions for estimating the parameters of SDEs is suggested by Hansen and Scheinkman (1995) who advocate the use of the infinitesimal generator to characterise continuous-time Markov processes. This approach is not discussed here, mainly because it is difficult to provide moment conditions to estimate the parameters of the diffusion function by this route. Moment conditions have also been generated using simulated moments, which are calculated as

expected values of moments across a large number of simulations of the stochastic process. This approach, referred to as the simulated method of moments (SMM), was developed by Duffie and Singleton (1993) and is similar in spirit to indirect estimation, which is discussed in detail in the following subsection.

Like DML, the major advantages of GMM are its ease of implementation, speed and the fact that it does not impose any restrictions on the nature of the drift and diffusion functions of the SDE. Similarly, its accuracy is hindered by discretisation bias at least for the method outlined here where discrete moments conditions are used.

4.2 Indirect estimation

The class of indirect estimators developed in Gourieroux, Monfort and Renault (1993) and Gallant and Tauchen (1996) are offshoots of SMM. In this approach, parameters of the SDE are not estimated directly from the (usually intractable) likelihood denoted in equation (2), but indirectly through another model. The efficacy of the indirect estimation technique therefore depends critically on finding an auxiliary model which is easy to estimate by ML, but which also provides a suitable approximation to the true likelihood function.

As usual, let $\boldsymbol{\theta}$ denote the parameter vector of the SDE to be fitted to the sample data. Now define an auxiliary model with independent parameters $\boldsymbol{\xi}$ which can be estimated easily by ML. The objective of the indirect estimation algorithm is to extract information about $\boldsymbol{\theta}$ indirectly through the ML estimates of $\boldsymbol{\xi}$. The crux of the procedure is to recognise that the auxiliary model is misspecified, and, therefore, the ML estimates of $\boldsymbol{\xi}$, say $\boldsymbol{\xi}^*$, are linked to the true parameters of the SDE by the so-called binding function

$$\boldsymbol{\xi}^* = \phi(\boldsymbol{\theta}_{\text{true}}),$$

where the dimension of $\boldsymbol{\xi}$ must be at least that of $\boldsymbol{\theta}$ for identification purposes. If the dimensions of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are identical then the binding function is invertible and $\boldsymbol{\theta}$ can be estimated indirectly by

$$\boldsymbol{\theta}_{\text{true}} = \phi^{-1}(\boldsymbol{\xi}).$$

In practice, the binding function may be recovered by simulation. Observations are simulated from the SDE for given $\boldsymbol{\theta}$, say $\tilde{\boldsymbol{\theta}}$, and then used to find the corresponding estimate for $\boldsymbol{\xi}$, say $\tilde{\boldsymbol{\xi}}$. In effect, a realisation of the binding function has been constructed.

In general terms the indirect estimation procedure proceeds as follows. Auxiliary parameters are obtained from the auxiliary model by minimising its negative log-likelihood to obtain

$$\boldsymbol{\xi}^* = \underset{\boldsymbol{\xi}}{\text{Argmin}} \left(- \sum_{k=0}^{N-1} \log f^{(\text{aux})}(X_{k+1} | X_k; \boldsymbol{\xi}) \right) \quad (81)$$

where $f^{(\text{aux})}(X_{k+1} | X_k; \boldsymbol{\xi})$ is the transitional PDF of the auxiliary model. As already noted, it is essential that these ML estimates are easy to compute. For a given $\boldsymbol{\theta}$, the SDE is simulated to obtain

M samples $X_t^{(1)}(\boldsymbol{\theta}), \dots, X_t^{(M)}(\boldsymbol{\theta})$. There are now two ways to use the information provided by these simulations in order to extract the form of the binding function.

The procedure proposed by Gouriéroux *et al.* (1993) (GMR) proceeds as follows. For each simulated data set, the auxiliary model is re-estimated to provide M estimates $\tilde{\boldsymbol{\xi}}^{(1)}(\boldsymbol{\theta}), \tilde{\boldsymbol{\xi}}^{(2)}(\boldsymbol{\theta}), \dots, \tilde{\boldsymbol{\xi}}^{(M)}(\boldsymbol{\theta})$ of the parameters of the auxiliary model. The Gouriéroux *et al.* (1993) indirect estimator $\hat{\boldsymbol{\theta}}_{\text{GMR}}$ is defined by

$$\hat{\boldsymbol{\theta}}_{\text{GMR}} = \underset{\boldsymbol{\theta}}{\text{Argmin}} \left(\boldsymbol{\xi}^* - \frac{1}{M} \sum_{i=1}^M \tilde{\boldsymbol{\xi}}^{(i)}(\boldsymbol{\theta}) \right)^{\text{T}} \Omega \left(\boldsymbol{\xi}^* - \frac{1}{M} \sum_{i=1}^M \tilde{\boldsymbol{\xi}}^{(i)}(\boldsymbol{\theta}) \right) \quad (82)$$

where Ω is a positive-definite weighting matrix.

Gallant and Tauchen (1996) (GT) suggest another indirect estimator, commonly known as the efficient method of moments (EMM). Let \mathbf{G} be the matrix of scores of the auxiliary model where the k -th column of \mathbf{G} is the vector

$$\frac{\partial \log f^{(\text{aux})}(X_{k+1}|X_k; \boldsymbol{\xi}^*)}{\partial \boldsymbol{\xi}}.$$

With this notation in place, the EMM estimator, $\hat{\boldsymbol{\theta}}_{\text{GT}}$, is defined by

$$\hat{\boldsymbol{\theta}}_{\text{GT}} = \underset{\boldsymbol{\theta}}{\text{Argmin}} \left(\frac{1}{M} \sum_{i=1}^M \text{E}[\tilde{\mathbf{G}}^{(i)}(\boldsymbol{\theta})] \right)^{\text{T}} \Omega \left(\frac{1}{M} \sum_{i=1}^M \text{E}[\tilde{\mathbf{G}}^{(i)}(\boldsymbol{\theta})] \right), \quad (83)$$

where $\text{E}[\tilde{\mathbf{G}}^{(i)}(\boldsymbol{\theta})]$ is a vector with the same dimension as $\boldsymbol{\xi}$ containing the expected value of the rows of $\tilde{\mathbf{G}}^{(i)}$. Essentially $\tilde{\mathbf{G}}^{(i)}$ is the analogue of \mathbf{G} constructed from the i -th simulated sample instead of the observed data.

In both these estimators, the positive-definite weighting matrix Ω may be set equal to the inverse of the outer product estimate of the Hessian matrix of the auxiliary log-likelihood function. Specifically, the weighting matrix has the following form

$$\Omega = \left[\frac{1}{N} \mathbf{G} \mathbf{G}^{\text{T}} \right]^{-1}.$$

Note that for the models used in this paper, analytical expressions for the Hessian are available.

For the CIR and OU processes, the natural choice of auxiliary model is the Euler-Maruyama discretisations of the SDE, namely

$$\begin{aligned} \text{(CIR)} \quad X_{k+1} - X_k &= \alpha^*(\theta^* - X_k) + \sigma^* \sqrt{X_k} \varepsilon_k, \\ \text{(OU)} \quad X_{k+1} - X_k &= \alpha^*(\theta^* - X_k) + \sigma^* \varepsilon_k, \end{aligned} \quad (84)$$

with $\varepsilon_k \sim \text{IIN}(0, 1)$. The analytical solutions to the ML problem for the parameters of these models is given in Appendix 3.

The GMR approach needs M estimations of the auxiliary model (one for each simulated sample) for each value of $\boldsymbol{\theta}$ in the minimisation procedure required by equation (82). Of course, this is not

an arduous requirement if estimation of the auxiliary model is easy, that is, closed form solutions are available for estimating ξ . In general, however, the EMM variant of indirect estimation is to be preferred because, being a simulation method, the core procedure is already computationally quite expensive even without the added burden of repeated optimisation. The fundamental performance issue in either approach depends critically on the auxiliary model. If a natural auxiliary model is available then indirect estimation works well even in relatively small samples. When no natural auxiliary model is available, for example, as occurs in the CIR process with the levels effect parameter not constrained to be $1/2$, the performance deteriorates significantly (see Hurn *et al.*, 2003).

4.3 Characteristic function approaches

Singleton (2001), Jiang and Knight (2002), and Chacko and Viciara (2003) have revisited the idea of utilising the Fokker-Planck equation in a procedure to estimate the parameters of SDEs. However, rather than solving for the transitional PDF they solve for the characteristic function of X defined by

$$\tilde{f}(p, t) = \int_{\mathcal{S}} e^{-px} f(x, t) dx \quad (85)$$

where $f(x, t) \equiv f((x, t) | (X_0, t_0); \theta)$ is the transitional PDF of X and p is a parameter which is restricted to those regions of the complex plane for which the convergence of the integral defining $\tilde{f}(p, t)$ is assured. If $\mathcal{S} = \mathbb{R}$, as occurs in the case of the OU process, then $p = i\omega$ where ω is real-valued, $i^2 = -1$ and the characteristic function is the familiar Fourier transform of the transitional PDF. On the other hand, if $\mathcal{S} = \mathbb{R}^+$ (or any semi-infinite interval) as happens in the case of the CIR process, then p is a complex number that is typically restricted to the half-plane $\text{Re}(p) > b$, where b is defined by the requirement that $|f(x, t)| < Me^{bx}$ for all $x > 0$, and the characteristic function is now the Laplace transform of the transitional PDF. Once the characteristic function has been obtained, it is possible to use the inverse transform to obtain the unknown transitional PDF.

When the characteristic function, $\tilde{f}(\omega, t)$, is the Fourier transform (as is the case for the OU process), the transitional PDF is recovered from

$$f(x, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{f}(\omega, t) e^{i\omega x} d\omega \quad (86)$$

by numerical integration. The procedure is not straightforward since the kernel of the integrand is oscillatory to mention but one of the difficulties. The usual way to estimate $f(x, t)$ is by means of Filon's method. On the other hand, when the characteristic function is the Laplace transform (as is the case for the CIR process), the transitional PDF can be obtained by means of Mellin's formula

$$f(x, t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \tilde{f}(p, t) e^{px} dp, \quad c > b. \quad (87)$$

When it becomes necessary to evaluate $f(x, t)$ by Mellin's formula, the procedure will usually involve the method of Residue Calculus often in the presence of a slit plane to accommodate the fact that the integrand is not single valued.

For the affine¹⁴ class of SDEs, it is possible to obtain the characteristic function in closed form as an exponential of an affine function (see Duffie and Kan, 1996). Although the affine class contains a number of popular SDEs in finance (including the CIR and OU processes) it is a rather limited subset of the SDEs of interest. For non-affine SDEs the characteristic function must be approximated. A more promising offshoot of this literature, therefore, has been the development of a procedure to estimate the parameters of SDEs in a method of moments framework that does not require the use of the inverse transform. It is for this reason that the characteristic function method appears in the sample DNA matching section and not the section dealing with likelihood-based methods. There are two ways to generate moments.

Spectral moments The central idea here is that the expected value of the characteristic function at time t_{k+1} is known. If the characteristic function is a Laplace transform then

$$\mathbb{E}[\tilde{f}(p, t_{k+1})] = \int_0^\infty e^{-px} f(x, t_{k+1}) dx = \int_0^\infty e^{-px} \delta(x - X_{k+1}) dx = e^{-pX_{k+1}} \quad (88)$$

whereas if the characteristic function is a Fourier transform then

$$\mathbb{E}[\tilde{f}(\omega, t_{k+1})] = \int_{-\infty}^\infty e^{-i\omega x} f(x, t_{k+1}) dx = \int_{-\infty}^\infty e^{-i\omega x} \delta(x - X_{k+1}) dx = e^{-i\omega X_{k+1}}. \quad (89)$$

Moment conditions can therefore be obtained by matching the expected value of the characteristic function with $e^{-pX_{k+1}}$ as in equation (88) at various (arbitrary) choices of p for the Laplace transform, and by matching the real and imaginary parts of the expected value of the characteristic function with the real and imaginary parts of $e^{-i\omega X_{k+1}}$ as in equation (89) at various (arbitrary) choices of ω for the Fourier transform. Further moment conditions can be generated by taking products of these primitive moment conditions with powers of X_k .

Temporal moments The j -th moment of the Laplace and Fourier characteristic functions can be expressed respectively as

$$\begin{aligned} \text{(Laplace)} \quad & (-1)^j \frac{d^j \tilde{f}(p, t_{k+1})}{dp^j} \Big|_{p=0} = \int_0^\infty x^j f(x, t_{k+1}) dx \\ \text{(Fourier)} \quad & i^j \frac{d^j \tilde{f}(\omega, t_{k+1})}{d\omega^j} \Big|_{\omega=0} = \int_{-\infty}^\infty x^j f(x, t_{k+1}) dx. \end{aligned} \quad (90)$$

For example, taking $j = 1$ and $j = 2$ for the Laplace transform gives the moment conditions

$$\mathbb{E} \left[(-1) \frac{d\tilde{f}(p, t_{k+1})}{dp} \Big|_{p=0} - X_{k+1} \right] = 0 \quad \mathbb{E} \left[(-1)^2 \frac{d^2 \tilde{f}(p, t_{k+1})}{dp^2} \Big|_{p=0} - X_{k+1}^2 \right] = 0 \quad (91)$$

respectively, provided that these derivatives can be evaluated analytically.

The characteristic function procedure implemented in the Monte Carlo experiments of Section 5 is based on spectral moment conditions. The derivation of the transitional PDFs of the CIR and OU processes in Appendix 2 relies crucially on their characteristic functions, given in equations (134) and (139) respectively. These results are now used to establish the following moment conditions.

¹⁴Affine SDEs have drift and diffusion functions that are linear in the state variable X .

CIR process The primitive moment condition for the CIR process is

$$\mathbb{E} \left[c^{\nu+1} e^{-u} (p+c)^{-\nu-1} \exp \left[\frac{cu}{p+c} \right] - \exp[-pX_{k+1}] \right] = 0 \quad (92)$$

where c , v , u and ν are as defined in the discussion following equation (134). Each choice of p provides only one moment condition, as this condition is real valued. Further moment conditions can be obtained by taking products of the primitive moment condition with powers of X_k .

OU process The primitive moment condition used in the estimation of the parameters of the OU SDE is given by

$$\mathbb{E} \left[\exp \left[-i\omega(X_k + (\beta - X_k)(1 - e^{-\alpha\Delta}) - \frac{\omega^2\sigma^2}{4\alpha}(1 - e^{-2\alpha\Delta})) \right] - \exp[-i\omega X_{k+1}] \right] = 0 \quad (93)$$

where $\Delta = t_{k+1} - t_k$. In theory an infinite number of moment conditions are available by (arbitrary) choice of values for ω . Note also, that each choice of ω provides two moment conditions as the real and complex parts of the two components of expression (93) are matched separately. Further moment conditions can also be generated by multiplying the primitive moment condition with powers of X_k .

Once the moment conditions have been chosen, either by choice of p or ω , or by choice of j (if temporal moments are used), parameter estimation then proceeds in an identical way to GMM. Note that when using the spectral moment approach, some care is needed in the choice for the values of p or ω at which the moment conditions are generated. Injudicious choices can lead to matrices that are poorly conditioned and therefore difficult to invert when it becomes necessary to construct the weighting matrix in the GMM procedure (Singleton, 2001).

Although this method of estimating the parameters of SDEs is analytically appealing, it does suffer from the major drawback that its applicability is largely limited to the class of SDEs with affine drift and diffusion functions. This is because it is possible to generate exact closed-form expressions for the characteristic function for this class of model. If the drift and diffusion functions are not intrinsically linear functions of state, then they must be approximated by linear functions.

4.4 Estimating function approaches

Like the characteristic function approach to estimating the parameters of SDEs, the estimating function method could easily be classified as a likelihood-based procedure. Ultimately the decision to classify it as a DNA matching method is driven by the recognition that, at its heart, the estimating function approach comprises a set of conditions that may be regarded as moments and implemented within the GMM framework.

Kessler and Sørensen (1999) recognise that moment conditions may be obtained from the eigenfunctions of the infinitesimal generator of the process. It is demonstrated in Appendix 4 that

$$\frac{d\mathbb{E}[\phi]}{dt} = \mathbb{E}[\mathcal{A}_\theta\phi] \quad (94)$$

where $\phi(x)$ is a suitably differentiable function of state and \mathcal{A}_θ is the usual infinitesimal operator. Suppose now that ϕ is an eigenfunction of the infinitesimal operator \mathcal{A}_θ with eigenvalue λ , then

$$\mathcal{A}_\theta \phi = \lambda \phi. \quad (95)$$

In this instance equation (94) becomes

$$\frac{d\mathbb{E}[\phi]}{dt} = \lambda \mathbb{E}[\phi] \quad (96)$$

with solution

$$\mathbb{E}[\phi](t_{k+1}) = e^{\lambda(t_{k+1}-t_k)} \phi(X_k). \quad (97)$$

Given J eigenfunctions ϕ_1, \dots, ϕ_J with associated distinct eigenvalues $\lambda_1, \dots, \lambda_J$, moment conditions may be generated from the primitive moment generating expression

$$\mathbb{E}[\phi_j(X) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)] = [\phi_j(X_{k+1}) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)] = 0. \quad (98)$$

Before demonstrating how these moment conditions are used in an estimating function approach, it is necessary to outline the basic ideas behind the general estimating function method.

An estimating function is defined to be a function of the model parameters with the property that the optimal parameters are returned when this function takes the value zero. The most familiar estimating function is the score function, the zeros of which are the ML estimates of the model parameters. For estimating the parameters of SDEs, however, the likelihood function is generally intractable and the score function is not available.

Bibby and Sørensen (1995) consider the score function arising from DML, but correct for the discretisation bias by subtracting a compensator that depends on the conditional moments of the process. The resulting estimating function is

$$G_{\text{BS}} = \sum_{k=0}^{N-1} \frac{[X_{k+1} - \mathbb{E}[X_{k+1} | X_k; \theta]]}{V(X_k; \theta)} \frac{\partial \mathbb{E}[X_{k+1} | X_k; \theta]}{\partial \theta} \quad (99)$$

where, as usual, θ is the parameter vector, $\mathbb{E}[X_{k+1} | X_k; \theta]$ is the conditional mean of the process and $V(X_k; \theta) = \mathbb{E}[(X_{k+1} - \mathbb{E}[X_{k+1} | X_k; \theta])^2 | X_k; \theta]$ is its conditional variance. Bibby and Sørensen (1995) show that the resulting parameter estimates are asymptotically consistent (but not unbiased) and are normally distributed. Of course, one drawback of this approach is that analytical expressions for the conditional moments are not usually known. Bibby and Sørensen (1995) suggest that these moments can be well approximated by simulation, but this is a time consuming process, particularly when calculating partial derivatives with respect to the parameters θ .

Kessler and Sørensen (1999) have further developed the procedure of Bibby and Sørensen by recognising that equation (95) can be reformulated as a Sturm-Liouville eigenvalue problem. Sturm-Liouville theory asserts that continuous functions $\psi(x)$ lying within the function space defined by the Sturm

operator can be approximated to arbitrary accuracy by linear combination of the eigenfunctions of \mathcal{A}_θ , that is,

$$\psi(x) \approx \sum_{j=1}^J a_j \phi_j(x)$$

The estimating function method then naturally leads to the conclusion that the eigenfunctions of the infinitesimal generator may be used as a basis for approximating the scores. For a particular transition, let S_i be the i -th component of the vector of scores, that is

$$S_i = \frac{\partial \log f(x, t_{k+1} | X_k; \boldsymbol{\theta})}{\partial \theta_i}.$$

The approximation is achieved on a transition-by-transition basis, with the optimal coefficients for a particular transition being obtained by minimising

$$\Phi_k^{(i)} = \int \left(S_i - \sum_{j=1}^J a_j^{(i)} [\phi_j(x) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)] \right)^2 f(x, t_{k+1} | X_k; \boldsymbol{\theta}) dx \quad (100)$$

with respect to the coefficients $a_j^{(i)}$. The function $\Phi_k^{(i)}$ may be interpreted as the expected value of the squared difference between the score for parameter θ_i and the estimating function. The optimal values of $a_j^{(i)}$ are obtained by setting to zero the partial derivative of $\Phi_k^{(i)}$ with respect to $a_j^{(i)}$ for each integer $j = 1, 2, \dots, J$. This procedure leads to the first-order conditions

$$\begin{aligned} & \int \sum_{p=1}^J a_p^{(i)} [\phi_p(x) - e^{\lambda_p(t_{k+1}-t_k)} \phi_p(X_k)] [\phi_j(x) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)] f(x, t_{k+1} | X_k; \boldsymbol{\theta}) dx \\ &= \int \frac{\partial f(x, t_{k+1} | X_k; \boldsymbol{\theta})}{\partial \theta_i} [\phi_j(x) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)] dx \end{aligned} \quad (101)$$

which has matrix representation $A^{(i)} \mathbf{a}^{(i)} = \mathbf{b}^{(i)}$ where

$$\begin{aligned} A_{pj}^{(i)} &= \int [\phi_p(x) - e^{\lambda_p(t_{k+1}-t_k)} \phi_p(X_k)] [\phi_j(x) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)] f(x, t_{k+1} | X_k; \boldsymbol{\theta}) dx, \\ b_j^{(i)} &= \int \frac{\partial f(x, t_{k+1} | X_k; \boldsymbol{\theta})}{\partial \theta_i} [\phi_j(x) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)] dx. \end{aligned} \quad (102)$$

By observing that

$$\begin{aligned} b_j &+ \int \frac{\partial}{\partial \theta_i} [\phi_j(x) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)] f(x, t_{k+1} | X_k; \boldsymbol{\theta}) dx \\ &= \frac{\partial}{\partial \theta_i} \int [\phi_j(x) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)] f(x, t_{k+1} | X_k; \boldsymbol{\theta}) dx \\ &= \frac{\partial}{\partial \theta_i} \left(\mathbf{E}[\phi_j](t_{k+1}) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k) \right) = 0, \end{aligned}$$

it follows that b_j can be further simplified to

$$b_j = - \int \frac{\partial}{\partial \theta_i} [\phi_j(x) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)] f(x, t_{k+1} | X_k; \boldsymbol{\theta}) dx. \quad (103)$$

Solution of this linear system for each transition yields the optimal coefficient values $\widehat{a}_j^{(i)}(X_k)$. Once these optimal coefficients have been obtained, Kessler and Sørensen (1999) construct the approximation to the score as

$$G_{KS}^{(i)} = \sum_{k=0}^{N-1} \sum_{j=1}^J \widehat{a}_j^{(i)}(X_k) [\phi_j(X_{k+1}) - e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k)]. \quad (104)$$

The estimates of the model parameters $\boldsymbol{\theta}$ are now found by solving simultaneously the equations $G_{KS}^{(i)} = 0$ for each value of i . This is done using a Newton-Raphson scheme with a Jacobian matrix calculated by numerical differentiation. In order for this scheme to converge, it is imperative that the initial values for the parameters are reasonably accurate. In practice this would mean that the parameters of the SDE would first be estimated by (for example) traditional DML in order to obtain a good starting estimate for the optimal value of $\boldsymbol{\theta}$.

With regard to the practical implementation of the estimating function method, significant simplifications are possible when the eigenfunctions $\phi_j(x)$ ($j \geq 1$) are polynomials of degree j , that is,

$$\phi_j(x) = \sum_{p=0}^j c_p^{(j)}(\boldsymbol{\theta}) x^p \quad (105)$$

where $c_p^{(j)}(\boldsymbol{\theta})$ are known coefficients. Some common SDEs fall into this category including those describing the CIR and OU processes. When the eigenfunctions are polynomials, the computation of expressions (102) and (103) is equivalent to computing the moments $\int x^p f(x, t_{k+1} | X_k; \boldsymbol{\theta}) dx$ for $1 \leq p \leq 2J$. Kessler and Sørensen (1999) show that these moments can be computed from expression (97) by constructing and then solving the system of linear equations

$$e^{\lambda_j(t_{k+1}-t_k)} \phi_j(X_k) = \sum_{p=0}^j c_p^{(j)} \int x^p f(x, t_{k+1} | X_k; \boldsymbol{\theta}) dx \quad j = 1, \dots, 2J. \quad (106)$$

Alternatively, if a closed form expression for the characteristic function exists, the moments can be obtained by the procedure described in equations (90). If the eigenfunctions are not polynomials the evaluation of the integrals in (102) and (103) usually requires the process to be simulated, which is again a time consuming exercise.

CIR and OU processes Appendix 6 demonstrates that the eigenfunctions for the CIR process are generalised Laguerre polynomials with argument $2\alpha x/\sigma^2$ and for the OU process are Hermite polynomials with argument $\sqrt{\alpha}(x-\beta)/\sigma$. The coefficients $c_p^{(j)}$ of the generalised Laguerre polynomials may be determined by setting $c_0^{(j)} = 1$ and then using the iterative scheme

$$c_{p+1}^{(j)} = \frac{2\alpha(j-p)}{\sigma^2(p+1)(p+1+\nu)} c_p^{(j)}, \quad \nu = \frac{2\alpha\beta}{\sigma^2} - 1, \quad (107)$$

to determine the remaining coefficients. It can be shown that the first and second eigenfunctions for the CIR and OU processes are respectively

$$\begin{aligned} \text{(CIR)} \quad \phi_1(x) &= \frac{2\sqrt{\alpha}(\beta - x)}{\sigma}, & \phi_2(x) &= \frac{4\alpha(\beta - x)^2}{\sigma^2} - 2, \\ \text{(OU)} \quad \phi_1(x) &= 1 + \frac{2\alpha x}{\beta\sigma^2}, & \phi_2(x) &= 1 + \frac{\alpha x}{\beta\sigma^2} + \frac{4\alpha^3 x^2}{\sigma^4(\sigma^2 + 2\alpha\beta)}. \end{aligned}$$

Two other types of estimating functions have been introduced in the literature. Sørensen (2000) considers prediction based estimating functions which involve approximating conditional moments with expressions based on unconditional moments. By contrast, Kessler (2000) suggests that the moment conditions derived in Hansen and Schienkman (1995) (see Section 4.1) can be used to construct the simple estimating function

$$G_K(\boldsymbol{\theta}) = \sum_{k=0}^{N-1} \mathcal{A}_{\boldsymbol{\theta}}\phi(x_k) \tag{108}$$

for a suitable choice of test functions ϕ . Parameter estimates are obtained by solving for the values of $\boldsymbol{\theta}$ that set each equation in the system of equations (108) to zero. However, this approach is not particularly useful in practice. For example, in the cases of the CIR and OU processes, irrespective of the choices of ϕ it will not be possible to identify both the α and σ parameters simultaneously.

The EF approach is a complex procedure and to be viable it requires that closed-form expressions be available for the eigenfunctions. In practice this means that the associated Sturm-Liouville eigenvalue problem must have a known closed-form expression for the solution. For this solution to be a polynomial, as in the CIR and OU processes, the drift and diffusion functions of the SDE will most likely need to be polynomials and, in particular, affine functions of state. If the eigenfunctions are not polynomials the complexity of this method is likely to be severe and will involve simulation. It is clear, therefore, that this method is particularly problem specific.

4.5 Matching marginal density

Aït-Sahalia (1996b) develops an approach to estimating $\boldsymbol{\theta}$ based on matching the marginal density of the SDE with a kernel density estimate of the marginal density constructed from the data. Under this approach the optimal parameters are given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{Argmin}} \frac{1}{N} \sum_{k=0}^{N-1} (\pi(X_k; \boldsymbol{\theta}) - \hat{\pi}_0(X_k))^2 \tag{109}$$

where $\pi(x; \boldsymbol{\theta})$ is the marginal density of the SDE and $\hat{\pi}_0(x)$ is the kernel density estimated from the data.

The major disadvantage of this approach is that time-series data tends to be highly correlated whereas the kernel procedure presupposes that the data are independent and identically distributed observations of the process. For example, Pritsker (1998) demonstrates that the informational content of the

sample for which Aït-Sahalia (1996b) proposed this approach is small, and therefore a kernel estimate of the marginal density is likely to be inaccurate. This estimation approach is not pursued further in the light of this recognition. Similarly, another approach suggested by Aït-Sahalia (1996a), based on the transitional PDF, also requires a kernel estimate of the marginal density and is not discussed.

5 Simulation experiments

In this section the estimators discussed previously are used to estimate the parameters of the CIR and OU SDEs. In comparing the different algorithms particular attention will be given to the following criteria.

1. **Implementation** This refers to the ease with the estimation procedure can be used in practice. This includes whether or not the procedure requires specific forms for the drift and diffusion functions to be implemented easily.
2. **Accuracy** The accuracy of the method is judged by comparing the parameter estimates with the known values used in the simulations and with reference to EML, the benchmark method.
3. **Speed** There are large variations in the time required to estimate parameters by the various methods. The speed will be judged in terms of the the average time taken to obtain estimates in repeated experiments.

Of course, it is not possible to provide a totally objective comparison. Two specific cases spring to mind. In the first instance, it is difficult to choose models that are neutral to all the procedures. For example, the two SDEs chosen as the benchmark models involve linear specifications of the drift and diffusion, which in turn favour methods that can take advantage of this property (the MCMC, characteristic function and estimating function methods) relative to more generic methods (such as the finite-difference procedure). The second case in point is the natural conflict between accuracy and speed that characterises many of these methods, a trade-off that is influenced by a range of subjective decisions made by the researcher. For example, the choice of the level of discretisation of the state space and time in the finite-difference approach, the number of moments in a GMM approach, the number of simulation paths in a simulation method, the number of eigenfunctions used in an estimating function method to name but a few. As a guiding principle, these subjective choices over the operating regimen of the estimating procedure were made with the overarching objective of achieving *reasonable accuracy* in a sensible timeframe.

5.1 Experimental design

The estimation algorithms are compared by generating 2000 samples of 500 observations from the CIR model

$$dX = \alpha(\beta - X) dt + \sigma\sqrt{X} dW$$

with true parameters $\alpha = 0.20$, $\beta = 0.08$ and $\sigma = 0.10$, and from the OU process

$$dX = \alpha(\beta - X) dt + \sigma dW$$

with $\alpha = 0.20$, $\beta = 0.08$ and $\sigma = 0.03$. The synthetic samples are generated using the Milstein scheme with each time interval between observations $\Delta = 1/12$ (representing monthly data) broken into 1000 steps to ensure that the observed data are accurate realisations of the process. The estimation procedures used in the comparison and the choices made in their implementation are now summarised¹⁵.

1. Likelihood-based procedures

- (a) **EML** The benchmark method, EML, involves no subjective decisions.
- (b) **PDE** The method of finite differences requires the specification of the units of discretisation of state space ($\Delta x = 0.001$) and time ($\Delta t = 1/120$). This is a somewhat coarse discretisation of time but it is sufficient to obtain reasonable accuracy for the processes being considered. The discretisation of state space needs to be fine enough that the initial density for each transition is adequately resolved. The initial density used here is the Normal approximation suggested by Jensen and Poulsen (2002) which becomes more diffuse as Δt increases. Therefore, the relatively coarse discretisation of time permits a similarly coarse discretisation of state space.
- (c) **DML** Both traditional DML, based on the Euler discretisation, and the local linearisation approach (DML-SO) of Shoji and Ozaki (1998) are employed. In order to use the local linearisation method for the CIR process, the SDE is first converted to a unit diffusion by the transformation in equation (32).
- (d) **SML** Three SML procedures are used
 - **SML-KD** The kernel density approach of Hurn *et al.* (2003) is implemented with time discretisation $\Delta t = 1/120$. The kernel density approach generally requires a large number of simulations to deliver satisfactory accuracy and therefore $M = 200$ simulated paths are used.
 - **SML-IS-EUL** The importance sampling approach of Elerian *et al.* (2001) is implemented with sub-transition densities obtained from the Euler scheme, time discretisation $\Delta t = 1/120$ and $M = 100$ simulation paths. The simulated time paths between observations are generated from the modified Brownian bridge developed by Durham and Gallant (2002).
 - **SML-IS-SO** The importance sampling approach of Elerian *et al.* (2001) is implemented with identical specification to SML-IS-EUL but with sub-transition densities obtained by local linearisation.

¹⁵The notation used in each case should be interpreted in the context of the subsection in which the details of the procedure is discussed.

- (e) **HPE** Two HPE procedures are used
- **FHPE** In the finite HPE approach proposed by Aït-Sahalia (2002), seven polynomial terms are included in the Hermite expansion so that all the expressions for the coefficients of the expansion derived by Aït-Sahalia are used.
 - **IHPE** In the infinite HPE approach proposed by Aït-Sahalia (2002), contributing terms up to order three in Δ are included in the power series expansion.
- (f) **MCMC** The MCMC approach involves setting the unit of temporal discretisation ($\Delta t = 1/60$) and deciding on the number of iterations of the procedure to be done. Here 2500 parameter draws are made and the first 500 are discarded when calculating the final parameter estimates. Simulated realisations of the process are generated from the Elerian *et al.* (2001) proposal density.

2. Sample DNA matching procedures

- (a) **GMM** GMM is implemented using the discrete moments proposed by Chan *et al.* (1992) described in equations (79) and (80).
- (b) **CF** Spectral moment conditions are generated from equations (93) and (92) by choosing two arbitrary frequencies ($p = 1, 5$ or $\omega = 1, 5$). The first moment of the basic conditions is also used, to give a total of four conditions for the CIR process and eight conditions for the OU process. The OU process has twice as many conditions because at any frequency two conditions arise from matching the real and complex components, whereas for the CIR process there are only real components to match.
- (c) **IE** In the class of indirect estimators, the efficient method of moments (EMM) is chosen as the variant to implement. It requires the selection of a discretisation of time ($\Delta t = 1/120$) and the number of simulation paths ($M = 100$). Note that closed-form expressions given in Appendix 3 are used to compute the local scores.
- (d) **EF** The final estimator used is the eigenfunction approximation to the score suggested by Kessler and Sørensen (1999). Two eigenfunctions, polynomials of degree one and two, are used in the EF approach. DML estimates are used to initialise the Newton-Raphson scheme.

5.2 Estimation results

The results of the comparative study are reported in Tables 1 and 2. The bias and root mean square error (RMSE) of the estimates of the parameters of the CIR and OU processes over 2000 samples are presented together with the average computational time per parameter estimation. To avoid problems of ensuring similar levels of numerical efficiency across different methods¹⁶ all the computer

¹⁶Computational comparisons in programs such as *Gauss*, *Matlab* and *Ox* may be distorted by differing levels of vectorisation.

code to implement the various methods were written in C and compiled with the Intel C++ Compiler, version 9.0. The simulation experiments were all run on a Pentium(R)4 2.80GHz desktop computer with 0.5Gb of RAM. Prior to discussing these results reported in Tables 1 and 2 in detail, some general comments on the first criterion for comparison, namely ease of implementation, are given.

Implementation The difficulties in implementation may be classified as conceptual and/or technical. EML, conventional DML and GMM are relatively straightforward to implement. DML with local linearisation, PDE, all the SML methods and IE are moderately difficult to implement. In our experience the most difficult algorithms to implement are MCMC, CF and EF and attention is focused on these methods in the following discussion.

To take MCMC first, the most important drawback is its problem-specific nature, particularly in regard to parameter drawings. When the model is linear, the drawing procedure is tractable, but when the problem is nonlinear, the parameters may need to be drawn from an incomplete density or the procedure may need to use an accept-reject strategy. For example, the drift specification $\alpha(\beta - x)$ cannot be estimated because the likelihood function is incomplete. Instead the drift is rewritten as $\kappa - \alpha x$ where $\kappa = \alpha\beta$ and so β is calculated indirectly. A second difficulty with MCMC relates to the drawing of the unobserved states between observations by means of a high-dimensional optimisation. In any application of this method there is a significant probability that at some stage in drawing the unobserved states, the negative Hessian at the chosen parameter values will not be positive-definite. There is a practical fix to remedy this difficulty but it is one further level of complexity to be overcome in implementing the procedure successfully.

The CF approach suffers from the need to specify a characteristic function which in practice means that the drift and diffusion functions have to be linear functions of state. Both the CIR and OU processes satisfy this condition but other specifications will require the drift and diffusion functions to be approximated by linear specifications. The CF procedure will then need to be implemented on these approximate forms. Furthermore, the choice of frequency on which to base the spectral moments may be critical to the accuracy of the procedure.

The EF approach is a complex procedure and is only viable for the processes discussed here for two reasons. First, the eigenfunctions for these models are well-known polynomials which allow the optimal coefficients of the expansion of the score function to be computed from the moments of the process. Second, because the processes are linear, the characteristic function exists and the moments can be computed directly.

From the perspective of implementation, the HPE methods are curious; they are either very simple or very difficult. The FHPE procedure is straightforward to implement provided that the coefficients supplied by Ait-Sahalia (2002) are sufficient to generate the required accuracy. Note that this result extends to situations in which a simple closed-form expression for the drift of the unit diffusion is not readily available. In this case, the chain rule for differentiation can be used to derive all the components required to compute the coefficients. The real difficulty arises when the level of temporal

accuracy supplied by Aït-Sahalia's specification is inadequate. In these situations it will be necessary to extend the Hermite expansion which will require not only the computation of additional coefficients, but also the re-computation of all existing coefficients to higher accuracy. These calculations can, however, often be assisted by numerical packages such as *Mathematica* or *Maple*. The computation of the coefficients of the power series expansion in the IHPE procedure is likely to be so complex as to be infeasible without the assistance of programs such as these.

Accuracy In terms of the accuracy of the parameter estimation, all the methods perform adequately, in the sense that the parameter estimates recovered are comparable with the accuracy of EML. The speed of adjustment parameter, α , is by far the most difficult parameter to estimate. By comparison, the long-term mean, β , and the volatility control parameter, σ , are easier to estimate and are determined more accurately.

In terms of bias, traditional DML, GMM based on the discretised moments of Chan *et al.* (1992) and the CF approach are the most interesting methods. The average estimates of α delivered by traditional DML are less biased than those provided by other procedures. This is, however, most likely due to a beneficial incidence of the (usually problematic) discretisation bias that afflicts this procedure. The local linearisation based DML approach (DML-SO) returns parameter values that are more in line with the majority of the other procedures. Of course, its good performance when applied to the OU process is unsurprising as it is essentially equivalent to EML in this (isolated) instance. The GMM and CF approaches return the results that are most distinct from the other procedures. In particular, they struggle to resolve the β parameter for the CIR process and provide mixed results for the α parameter of both processes.

In terms of RMSE, the poor performance of the GMM and CF approaches with regards to the β parameter of the CIR process is reflected in the large RMSE associated with these estimates. The kernel density based SML procedure (SML-KD) also returns larger RMSEs than most of the other procedures, particularly in the case of the CIR process. This reflects the difficulty in obtaining accurate kernel estimates of density from relatively few independent drawings. The relatively poorer performance in the CIR case may stem from the fact that density can be leaked into negative state space, a problem which cannot occur in the OU process.

Of course, if EML is to be regarded as a benchmark method, an interesting aspect of the performance of the estimators is their ability to deliver the EML estimates, in each and every repetition of the experiment and not merely on average. This aspect of the performance of the estimators is not apparent from Tables 1 and 2. A measure of this ability is obtained by examining the difference between the various estimates and the EML estimates. Tables 3 and 4 show the mean and standard deviations of these differences taken over the 2000 repetitions. From these results it is clear that the HPE, EF, PDE and SML-IS methods provide similar parameter estimates to EML in each and every sample.

Speed The clear losers in terms of speed are the two variants of the SML procedure (SML-IS-EUL and SML-IS-SO). The reason for this is to be found in the computation of the modified Brownian bridge suggested by Durham and Gallant (2002). This involves repeated computation of exponents which is well-known to be numerically intensive. The PDE, MCMC, SML-KD and EMM methods rate next in terms of speed. For the finite-difference procedure, time is spent solving a tri-diagonal system of equations a large number of times. In the case of MCMC, the time is spent in the simulation of the unobserved datums. Although the simulation paths in the case of the SML-KD approach are easy to generate, large numbers of paths are required to resolve adequately the kernel estimate of density. Although it remains moderately intensive, EMM runs surprisingly quickly for a simulation based procedure. This result is explained by the fact that analytical expressions for the scores of the auxiliary models are available in this experiment. As an aside, it may appear to be a curious result that MCMC is superior to SML in terms of speed, since both methods are similar in the sense that the sample is augmented in each case by generating data. The resolution of this paradox is to be found in the guiding principle of reasonable accuracy. In order to obtain broadly similar accuracy, the time discretisation in the case of SML ($\Delta t = 1/120$) is finer than that required by MCMC ($\Delta t = 1/60$). This discrepancy stems from the fact that the augmented datums are used differently by the two procedures. Moreover, as noted previously, the implementation of MCMC does take advantage of the particular forms of the benchmark models in the drawing of parameters.

To conclude, the best combinations of accuracy and speed are delivered by the EF and HPE approaches. As noted earlier, however, both these approaches are particularly suited to these specific problems and this result should not be taken as a blanket recommendation that they are to be regarded as the automatic methods of choice. Although inferior to EF and HPE in terms of the speed, the PDE and SML-IS methods also deliver the EML parameter estimates but have the additional advantage of being truly generic in the sense that once the methods are coded, they can be applied to all parameter estimation problems, whether linear or non-linear, simply by changing the specification of the drift and diffusion functions in the code.

CIR Process	Mean error in α (RMS error in α)	Mean error in β (RMS error in β)	Mean error in σ (RMS error in σ)	Time (sec)
EML	0.1097 (0.1803)	0.0011 (0.0224)	0.0002 (0.0032)	0.0374
DML	0.1054 (0.1746)	0.0012 (0.0224)	-0.0009 (0.0033)	0.0008
DML-SO	0.1088 (0.1794)	0.0011 (0.0223)	0.0002 (0.0032)	0.0143
PDE	0.1102 (0.1806)	0.0012 (0.0224)	0.0000 (0.0032)	8.1283
SML-IS-EUL	0.1094 (0.1797)	0.0012 (0.0224)	0.0000 (0.0032)	13.0960
SML-IS-S0	0.1098 (0.1803)	0.0012 (0.0224)	0.0002 (0.0032)	20.6624
SML-KD	0.1093 (0.1924)	0.0073 (0.0360)	-0.0008 (0.0042)	4.3654
FHPE	0.1097 (0.1802)	0.0012 (0.0223)	0.0002 (0.0032)	0.0324
IHPE	0.1097 (0.1802)	0.0012 (0.0223)	0.0002 (0.0032)	0.0362
GMM	0.1176 (0.1874)	0.0139 (0.3511)	-0.0013 (0.0036)	0.0124
CF	0.0861 (0.1701)	0.0130 (0.3857)	-0.0003 (0.0038)	0.0459
EMM	0.1095 (0.1809)	0.0013 (0.0227)	0.0001 (0.0005)	2.0555
EF	0.1100 (0.1805)	0.0012 (0.0224)	0.0002 (0.0033)	0.0151
MCMC	0.1109 (0.1802)	0.0013 (0.0224)	0.0003 (0.0032)	8.6636

Table 1: Bias and RMSE of parameter estimates for the CIR process, for 2000 repetitions of the experiment with a sample size of 500. RMSEs are shown in parentheses.

OU Process	Mean error in α (RMS error in α)	Mean error in β (RMS error in β)	Mean error in σ (RMS error in σ)	Time (sec)
EML	0.1101 (0.1780)	-0.0006 (0.0227)	0.0001 (0.0010)	0.0015
DML	0.1055 (0.1716)	-0.0005 (0.0227)	-0.0003 (0.0010)	0.0007
DML-SO	0.1101 (0.1780)	-0.0006 (0.0227)	0.0001 (0.0010)	0.0033
PDE	0.1100 (0.1779)	-0.0005 (0.0227)	0.0000 (0.0010)	8.1121
SML-IS-EUL	0.1096 (0.1773)	-0.0006 (0.0227)	0.0000 (0.0009)	12.1218
SML-IS-S0	0.1099 (0.1778)	-0.0006 (0.0227)	0.0001 (0.0010)	17.4021
SML-KD	0.1103 (0.1897)	-0.0005 (0.0269)	-0.0003 (0.0012)	1.4544
FHPE	0.1100 (0.1779)	-0.0006 (0.0227)	0.0001 (0.0010)	0.0270
IHPE	0.1101 (0.1780)	-0.0006 (0.0227)	0.0001 (0.0010)	0.0255
GMM	0.1051 (0.1723)	-0.0006 (0.0228)	-0.0004 (0.0010)	0.0075
CF	0.1292 (0.1986)	-0.0008 (0.0231)	-0.0004 (0.0011)	0.1096
EMM	0.1089 (0.1774)	-0.0006 (0.0228)	0.0000 (0.0010)	0.9790
EF	0.1102 (0.1780)	-0.0005 (0.0227)	0.0001 (0.0010)	0.0151
MCMC	0.1116 (0.1788)	-0.0005 (0.0227)	0.0001 (0.0001)	6.5898

Table 2: Bias and RMSE of parameter estimates for the OU process, for 2000 repetitions of the experiment with a sample size of 500. RMSEs are shown in parentheses.

CIR Process	α Mean diff. from EML (Std. dev.)	β Mean diff. from EML (Std. dev.)	σ Mean diff. from EML (Std. dev.)
DML	-0.0043 (0.0138)	0.0001 (0.0024)	-0.0011 (0.0007)
DML-SO	-0.0009 (0.0112)	0.0000 (0.0027)	0.0000 (0.0001)
PDE	0.0005 (0.0077)	0.0001 (0.0025)	-0.0002 (0.0001)
SML-IS-EUL	-0.0003 (0.0083)	0.0001 (0.0026)	-0.0002 (0.0001)
SML-IS-SO	0.0001 (0.0059)	0.0001 (0.0025)	0.0000 (0.0001)
SML-KD	-0.0004 (0.0630)	0.0061 (0.0258)	-0.0010 (0.0024)
FHPE	0.0000 (0.0017)	0.0001 (0.0023)	0.0000 (0.0000)
IHPE	0.0000 (0.0015)	0.0001 (0.0023)	0.0000 (0.0000)
GMM	0.0079 (0.0618)	0.0128 (0.3472)	-0.0015 (0.0012)
CF	-0.0236 (0.0375)	0.0119 (0.3813)	-0.0005 (0.0020)
EMM	-0.0003 (0.0273)	0.0002 (0.0036)	-0.0001 (0.0005)
EF	0.0002 (0.0063)	0.0001 (0.0024)	0.0000 (0.0003)
MCMC	0.0012 (0.0090)	0.0001 (0.0025)	0.0001 (0.0006)

Table 3: Mean difference between the various parameter estimates and the EML estimates for the CIR process, for 2000 repetitions of the experiment with a sample size of 500. The standard deviation of these differences is shown in parentheses.

OU Process	α Mean diff. from EML (Std. dev.)	β Mean diff. from EML (Std. dev.)	σ Mean diff. from EML (Std. dev.)
DML	-0.0046 (0.0053)	0.0001 (0.0030)	-0.0004 (0.0002)
DML-SO	0.0000 (0.0000)	0.0000 (0.0001)	0.0000 (0.0000)
PDE	-0.0001 (0.0058)	0.0001 (0.0031)	0.0000 (0.0000)
SML-IS-EUL	-0.0005 (0.0040)	0.0000 (0.0005)	0.0000 (0.0000)
SML-IS-SO	-0.0002 (0.0063)	0.0000 (0.0005)	0.0000 (0.0000)
SML-KD	0.0002 (0.0623)	0.0001 (0.0147)	-0.0003 (0.0007)
FHPE	0.0000 (0.0006)	0.0000 (0.0001)	0.0000 (0.0000)
IHPE	0.0000 (0.0003)	0.0000 (0.0000)	0.0000 (0.0000)
GMM	-0.0050 (0.0150)	0.0000 (0.0017)	-0.0005 (0.0002)
CF	-0.0100 (0.0112)	0.0000 (0.0015)	-0.0004 (0.0003)
EMM	-0.0011 (0.0126)	0.0000 (0.0035)	0.0000 (0.0001)
EF	0.0191 (0.0372)	-0.0002 (0.0045)	-0.0004 (0.0003)
MCMC	0.0015 (0.0043)	0.0000 (0.0030)	0.0001 (0.0001)

Table 4: Mean difference between the various parameter estimates and the EML estimates for the OU process, for 2000 repetitions of the experiment with a sample size of 500. The standard deviation of these differences is shown in parentheses.

6 Conclusion

There are now a large number of methods for estimating the parameters of SDEs. This paper provides a comprehensive evaluation and comparison of most of these methods. Attention is focused on the univariate SDE with no latent factors. In terms of the evidence presented in this paper regarding the ease of implementation, accuracy and speed of the estimation methods considered, three estimators perform particularly well. These are the procedures based on Hermite polynomial expansions, the method of finite differences and simulated maximum likelihood based on an importance sampling algorithm. Of these three estimators, the procedures based on Hermite polynomial expansions are by far the quickest to run, but the other two are more generic procedures that are easier to generalise to provide greater accuracy when applied to non-standard problems. While the performance of an estimating function approach based on eigenfunctions is also impressive in terms of accuracy and speed, the procedure is difficult to implement and is particularly problem specific.

References

- Aït-Sahalia, Y. (1996a). Nonparametric 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. (1999). Transition Densities for Interest Rates and Other Nonlinear Diffusions. *Journal of Finance*, **54**, 499-547.
- Aït-Sahalia, Y. (2002). Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approximation Approach. *Econometrica*, **70**, 223-262.
- Bandi, F.M. and Phillips P.C. (2003). Fully Nonparametric Estimation of Scalar Diffusion Models. *Econometrica*, **71**, 241-283.
- Bandi, F.M. and Phillips P.C. (2005). A Simple Approach to the Parametric Estimation of Potentially Nonstationary Diffusions. *Discussion Paper, Cowles Foundation, Yale University*, **1522**.
- Bakshi, G. and Ju, N. (2005). A Refinement to Aït-Sahalia's (2002) "Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approximation Approach". *Journal of Business*, **Forthcoming**.
- Bibby, B.M. and Sørensen, M. (1995). Martingale Estimation Functions for Discretely Observed Diffusion Processes. *Bernoulli*, **1**, 17-39.
- Brandt, M.W. and Santa-Clara, P. (2002). Simulated Likelihood Estimation of Diffusions with an Application to Exchange Rate Dynamics in Incomplete Markets. *Journal of Financial Economics*, **63**, 161-210.
- Broze, L., Scaillet, O. and Zakoïan, J-M. (1998). Quasi-Indirect Inference for Diffusion Processes. *Econometric Theory*, **14**, 161-186.
- Chacko, G. and Viceira, L.M. (2003). Spectral GMM Estimation of Continuous-Time Processes. *Journal of Econometrics*, **116**, 259-292.
- Chan, K.C., Karolyi, G.A., Longstaff, F.A. and Sanders, A.B. (1992). An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. *Journal of Finance*, **47**, 1209-1227.
- Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, **49**, 327-335.
- Chib, S. and Shephard, N. (2001). Comment on Garland B. Durham and A. Ronald Gallant's "Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes". *Unpublished mimeo*.

- Cox, J.C., Ingersoll, J.E. and Ross, S.A. (1985). A Theory of the Term Structure of Interest Rates. *Econometrica*, **53**, 385-407.
- Dacunha-Castelle, D. and Florens-Zmirou, D. (1986). Estimation of the Coefficients of a Diffusion from Discrete Observations. *Stochastics*, **19**, 263-284.
- Duffie, D. and Singleton, K.J. (1993). Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica*, **61**, 929-952.
- Duffie, D. and Kan, R. (1996). A Yield-Factor Model of Interest Rates. *Mathematical Finance*, **6**, 379-406.
- Durham, G.B. and Gallant, A.R. (2002). Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes. *Journal of Business and Economic Statistics*, **20**, 297-316.
- Elerian, O. (1998). A Note on the Existence of a Closed Form Conditional Transition Density for the Milstein Scheme. *Working Paper, Nuffield College, Oxford University*.
- Elerian, O., Chib, S. and Shephard, N. (2001). Likelihood Inference for Discretely Observed Non-linear Diffusions. *Econometrica*, **69**, 959-993.
- Eraker, B. (2001). MCMC Analysis of Diffusion Models with Application to Finance. *Journal of Business and Economic Statistics*, **19**, 177-191.
- Bateman Manuscript Project, Vol I and II (1954). Erdelyi, A. (ed.) McGraw-Hill Book Co. Inc., New York, London, Toronto.
- Florens-Zmirou, D. (1989). Approximate Discrete-Time Schemes for Statistics of Diffusion Processes. *Statistics*, **20**, 547-557.
- Gallant, A.R. and Tauchen, G. (1996). Which Moments to Match? *Econometric Theory*, **12**, 657-681.
- Gourieroux, C., Monfort, A. and Renault, E. (1993). Indirect Inference. *Journal of Applied Econometrics*, **8**, 85-118.
- Hansen, L.P. (1982). Large Sample Properties of Generalised Method of Moments Estimators. *Econometrica*, **50**, 1029-1054.
- Hansen, L.P. and Scheinkman, J.A. (1995). Back to the Future: Generating Moment Implications for Continuous-Time Markov Processes. *Econometrica*, **63**, 767-804.
- Hansen, L.P., Heaton, J. and Yaron, A. (1996). Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business and Economic Statistics*, **14**, 262-280.
- Hurn, A.S. and Lindsay, K.A. (1997). Estimating the Parameters of Stochastic Differential Equations by Monte Carlo Methods. *Mathematics and Computers in Simulation*, **43**, 495-501.

- Hurn, A.S. and Lindsay, K.A. (1999). Estimating the Parameters of Stochastic Differential Equations. *Mathematics and Computers in Simulation*, **48**, 373-384.
- Hurn, A.S., Lindsay, K.A. and Martin, V.L. (2003). On the Efficacy of Simulated Maximum Likelihood for Estimating the Parameters of Stochastic Differential Equations. *Journal of Time Series Analysis*, **24**, 45-63.
- Jensen, B. and Poulsen, R. (2002). Transition Densities of Diffusion Processes: Numerical Comparison of Approximation Techniques. *Journal of Derivatives*, **9**, 18-32.
- Jiang, G.J. and Knight, J.L. (2002). Estimation of Continuous-Time Processes via the Empirical Characteristic Function. *Journal of Business and Economic Statistics*, **20**, 198-212.
- Jones, C.S. (1998). A Simple Bayesian Method for the Analysis of Diffusion Processes. *Working Paper, Simon School of Business, University of Rochester*.
- Karlin, S. and Taylor, H.M. (1981). *A Second Course in Stochastic Processes*. Academic Press Inc.: New York.
- Kessler, M. and Sørensen, M. (1999). Estimating Equations Based on Eigenfunctions for a Discretely Observed Diffusion Process. *Bernoulli*, **5**, 299-314.
- Kessler, M. (2000). Simple and Explicit Estimating Functions for a Discretely Observed Diffusion Process. *Scandinavian Journal of Statistics*, **27**, 65-82.
- Lo, A.W. (1988). Maximum Likelihood Estimation of Generalized Ito Processes with Discretely Sampled Data. *Econometric Theory*, **4**, 231-247.
- McDonald, A.D. and Sandal, L.K. (1999). Estimating the Parameters of Stochastic Differential Equations Using a Criterion Function Based on the Kolmogorov-Smirnov Statistic. *Journal of Statistics, Computation and Simulation*, **64**, 235-250.
- Milstein, G. (1978). A Method of Second Order Accuracy Integration of Stochastic Differential Equations. *Theory of Probability and Its Applications*, **23**, 396-401.
- Pedersen, A.R. (1995). A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations. *Scandinavian Journal of Statistics*, **22**, 55-71.
- Phillips P.C. and Yu, J. (2005). A Two-stage Realised Volatility Approach to the Estimation for Diffusion Processes from Discrete Observations. *Discussion Paper, Cowles Foundation, Yale University*, **1523**.
- Pritsker, M. (1998). Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models. *The Review of Financial Studies*, **11**, 449-487.

- Roberts, G.O. and Stramer, O. (2001). On Inference for Partially Observed Non-linear Diffusion Models using the Metropolis-Hastings Algorithm. *Biometrika*, **88**, 603-621.
- Santa-Clara, P. (1995). Simulated Likelihood Estimation of Diffusions with an Application to the Short-Term Interest Rate. *Ph.D. Dissertation, INSEAD*.
- Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualisation*. John Wiley: New York.
- Shephard, N. (2005). *Stochastic Volatility: Selected Readings*. Oxford University Press: Oxford.
- Shoji, I. and Ozaki, T. (1997). Comparative Study of Estimation Methods for Continuous Time Stochastic Processes. *Journal of Time Series Analysis*, **18**, 485-506.
- Shoji, I. and Ozaki, T. (1998). Estimation for Nonlinear Stochastic Differential Equations by a Local Linearization Method. *Stochastic Analysis and Applications*, **16**, 733-752.
- Singleton, K.J. (2001). Estimation of Affine Asset Pricing Models Using the Empirical Characteristic Function. *Journal of Econometrics*, **102**, 111-141.
- Sørensen, M. (2000). Prediction Based Estimating Functions. *Econometrics Journal*, **3**, 123-147.
- Sundaresan, S. M. (2000). Continuous-time Methods in Finance: A Review and an Assessment. *Journal of Finance*, **55**, 1569-1622.
- Vasicek, O. (1977). An Equilibrium Characterization of the Term Structure. *Journal of Financial Economics*, **5**, 177-188.
- Yoshida, N. (1992). Estimation for Diffusion Processes from Discrete Observation. *Journal of Multivariate Analysis*, **41**, 220-242.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, John Wiley: New York.

Appendix 1: Fokker-Planck equation

Recall the one-dimensional, time-homogeneous SDE

$$dX = \mu(X; \boldsymbol{\theta}) dt + g(X; \boldsymbol{\theta}) dW \quad (110)$$

where dW is the differential of the Wiener process and the instantaneous drift $\mu(x; \boldsymbol{\theta})$ and instantaneous diffusion $g^2(x; \boldsymbol{\theta})$ are prescribed functions of state. The task is to estimate the parameters $\boldsymbol{\theta}$ from a sample of $(N + 1)$ observations X_0, \dots, X_N of the stochastic process at known times t_0, \dots, t_N . ML estimation of the parameters $\boldsymbol{\theta}$ requires the construction of the likelihood function which, in turn, requires that the transitional PDF of the process be known for each transition. This function is the solution of the Fokker-Planck equation and therefore this equation and its solution play a central role in parameter estimation for SDEs.

Let $X(t)$ be the solution of equation (110) satisfying the initial condition $X(t_0) = X_0$, then $X(t)$ is a stochastic process with a state space \mathcal{S} that is determined by the form of the drift and diffusion specifications. For $t > t_0$, the distribution of $X(t)$ is captured by $f(x, t) \equiv f((x, t) | (X_0, t_0); \boldsymbol{\theta})$, the transitional PDF of the process at time t . It will be shown that $f(x, t)$ satisfies the Fokker-Planck equation

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial x} \left(\frac{1}{2} \frac{\partial (g^2(x; \boldsymbol{\theta}) f)}{\partial x} - \mu(x; \boldsymbol{\theta}) f \right) \quad x \in \mathcal{S}, t > t_0, \quad (111)$$

with initial and boundary conditions

$$\begin{aligned} f(x, t_0) &= \delta(x - X_0), \quad x \in \mathcal{S}, \\ q &= \mu(x; \boldsymbol{\theta}) f - \frac{1}{2} \frac{\partial (g^2(x; \boldsymbol{\theta}) f)}{\partial x} = 0, \quad x \in \partial \mathcal{S}, t > t_0, \end{aligned} \quad (112)$$

where $\delta(x)$ is the usual Dirac delta function, $q = q(x, t)$ is the flux of probability density at time t and state x and $\partial \mathcal{S}$ denotes the boundary of the region \mathcal{S} . Briefly, the initial condition in (112) asserts that the process is known to start at X_0 at time t_0 and the boundary conditions assert that no probability can cross the boundaries of \mathcal{S} thereby ensuring that the probability mass within \mathcal{S} is conserved as the process evolves.

In essence, the Fokker-Planck equation is a conservation law expressing the fact that probability mass cannot be created or destroyed, that is, it is a conserved quantity. The mass of probability contained within the interval $[x - \Delta x, x + \Delta x] \subset \mathcal{S}$ is

$$\int_{x-\Delta x}^{x+\Delta x} f(u, t) du$$

where $f(u, t)$ is the transitional PDF of $X(t)$ at time t . Conservation of probability mass requires that the rate of change of the probability mass within $[x - \Delta x, x + \Delta x]$ exactly balances the rate at which probability mass enters this interval across its boundary points, that is,

$$\frac{d}{dt} \int_{x-\Delta x}^{x+\Delta x} f(u, t) du = q(x - \Delta x, t) - q(x + \Delta x, t) \quad (113)$$

where $q(u, t)$ is the flux of probability mass at state u and time t . The conservation law from which the Fokker-Planck equation is derived is obtained from equation (113) by means of the identities

$$\begin{aligned}\frac{\partial f(x, t)}{\partial t} &= \lim_{\Delta x \rightarrow 0^+} \frac{1}{2\Delta x} \int_{x-\Delta x}^{x+\Delta x} \frac{\partial f(u, t)}{\partial t} du, \\ -\frac{\partial q(x, t)}{\partial x} &= \lim_{\Delta x \rightarrow 0^+} \frac{q(x - \Delta x, t) - q(x + \Delta x, t)}{2h}.\end{aligned}\quad (114)$$

Equation (113) is divided by $2\Delta x$, and then identities (114) are used to deduce the conservation law

$$\frac{\partial f(x, t)}{\partial t} = -\frac{\partial q(x, t)}{\partial x}.\quad (115)$$

The Fokker-Planck equation is simply an application of equation (115) in which the SDE (110) is used to construct the relationship between the probability flux $q(x, t)$ and the transitional PDF $f(x, t)$.

Let $\psi(x, t)$ be an arbitrary differentiable function of state. The following analysis shows how to compute the flux of ψ when X evolves in accordance with equation (110). This general result, when applied to the transitional PDF $f(x, t)$, will establish the Fokker-Planck equation (111). The argument relies on the observation that in a small interval $(t, t + \Delta t)$, the local evolution of the process in the interval $(u, u + \Delta u)$ is approximately Gaussian with mean value $u + \mu(u)\Delta t$ and variance $g^2(u)\Delta t$ where it should be noted that the dependence of the drift and diffusion functions on the parameters θ has been suppressed since these parameters play no role in the analysis. Let Φ denote the CDF of the standard normal, and let $u = x$ be a rigid boundary in \mathcal{S} . The fraction of the probability located in $(u, u + \Delta u)$ at time t where $u < x$ that has diffused into the region $u > x$ is asymptotically

$$1 - \Phi\left(\frac{x - u - \mu(u)\Delta t}{g(u)\sqrt{\Delta t}}\right) = \Phi\left(\frac{-x + u + \mu(u)\Delta t}{g(u)\sqrt{\Delta t}}\right),$$

while the fraction of the probability located in $(u, u + \Delta u)$ at time t where $u > x$ that has diffused into the region $u < x$ is asymptotically

$$\Phi\left(\frac{x - u - \mu(u)\Delta t}{g(u)\sqrt{\Delta t}}\right).$$

Therefore the imbalance between diffusion of ψ from the region $u < x$ into the region $u > x$ and from the region $u > x$ into the region $u < x$ is asymptotically

$$\int_{-\infty}^x \psi(u, t) \Phi\left(\frac{-x + u + \mu(u)\Delta t}{g(u)\sqrt{\Delta t}}\right) du - \int_x^{\infty} \psi(u, t) \Phi\left(\frac{x - u - \mu(u)\Delta t}{g(u)\sqrt{\Delta t}}\right) du.$$

The flux of ψ is therefore

$$q_\psi = \lim_{\Delta t \rightarrow 0^+} \frac{1}{\Delta t} \left[\int_{-\infty}^x \psi(u, t) \Phi\left(\frac{-x + u + \mu(u)\Delta t}{g(u)\sqrt{\Delta t}}\right) du - \int_x^{\infty} \psi(u, t) \Phi\left(\frac{x - u - \mu(u)\Delta t}{g(u)\sqrt{\Delta t}}\right) du \right]. \quad (116)$$

This expression for the flux of ψ is simplified by the substitutions $u = x - \lambda\sqrt{\Delta t}$ in the first integral of expression (116) and $u = x + \lambda\sqrt{\Delta t}$ in the second integral of expression (116) to obtain

$$\begin{aligned}q_\psi &= \lim_{\Delta t \rightarrow 0^+} \frac{1}{\sqrt{\Delta t}} \left[\int_0^{\infty} \psi(x - \lambda\sqrt{\Delta t}, t) \phi\left(\frac{-\lambda + \mu(x - \lambda\sqrt{\Delta t})\sqrt{\Delta t}}{g(x - \lambda\sqrt{\Delta t})}\right) d\lambda \right. \\ &\quad \left. - \int_0^{\infty} \psi(x + \lambda\sqrt{\Delta t}, t) \phi\left(\frac{-\lambda - \mu(x + \lambda\sqrt{\Delta t})\sqrt{\Delta t}}{g(x + \lambda\sqrt{\Delta t})}\right) d\lambda \right].\end{aligned}\quad (117)$$

The computation of this limit is best achieved by expanding the integrand as a Maclaurin expansion in $\sqrt{\Delta t}$. The procedure begins by observing that

$$\begin{aligned}\frac{-\lambda + \mu(x - \lambda\sqrt{\Delta t})\sqrt{\Delta t}}{g(x - \lambda\sqrt{\Delta t})} &= -\frac{\lambda}{g(x)} - \left(\frac{\mu(x)}{g(x)} - \frac{\lambda^2 g'(x)}{g^2(x)}\right)\sqrt{\Delta t} + O(\Delta t) \\ \frac{-\lambda - \mu(x + \lambda\sqrt{\Delta t})\sqrt{\Delta t}}{g(x + \lambda\sqrt{\Delta t})} &= -\frac{\lambda}{g(x)} - \left(\frac{\mu(x)}{g(x)} - \frac{\lambda^2 g'(x)}{g^2(x)}\right)\sqrt{\Delta t} + O(\Delta t).\end{aligned}$$

Thereafter the limiting procedure is straightforward and leads to the expression

$$q_\psi = \int_0^\infty \left[-2\frac{\partial\psi(x,t)}{\partial x}\Phi(-\lambda/g) + 2\psi(x,t)\Phi'(-\lambda/g)\left(\frac{\mu(x)}{g(x)} - \frac{\lambda^2 g'(x)}{g^2(x)}\right) \right] d\lambda. \quad (118)$$

This integral is now simplified by means of the change of variable $y = \lambda/g$ to give

$$q_\psi = \int_0^\infty \left[-2g^2(x)\frac{\partial\psi(x,t)}{\partial x}y\Phi(-y) + 2\psi(x,t)\Phi'(-y)(\mu(x) - y^2g(x)g'(x)) \right] dy. \quad (119)$$

It remains to note that $\Phi(y)$, the CDF of the standard normal, satisfies

$$\int_0^\infty \Phi'(-y) dy = \frac{1}{2}, \quad \int_0^\infty y^2 \Phi'(-y) dy = \frac{1}{2}, \quad \int_0^\infty y \Phi(-y) dy = \frac{1}{4}.$$

In conclusion, expression (119) for the flux of ψ finally yields

$$q_\psi(x,t) = \mu(x; \boldsymbol{\theta})\psi(x,t) - \frac{1}{2}\frac{\partial}{\partial x}(g^2(x; \boldsymbol{\theta})\psi(x,t)). \quad (120)$$

Expression (120) gives the flux of an arbitrary function $\psi(x,t)$ when the underlying stochastic process evolves in accordance with equation (110). When ψ is taken to be the transitional PDF of the process itself, then expression (120) now relates the probability flux $q(x,t)$ of the process to the transitional PDF $f(x,t)$ by the formula

$$q(x,t) = \mu(x; \boldsymbol{\theta})f(x,t) - \frac{1}{2}\frac{\partial}{\partial x}(g^2(x; \boldsymbol{\theta})f(x,t)). \quad (121)$$

The Fokker-Planck equation in one dimension follows immediately from the conservation law (115).

Appendix 2: Derivation of the transitional PDF of the OU and CIR processes

The Fokker-Planck equation is linear in the unknown transitional PDF and therefore one way to obtain solutions of this equation is by means of integral transforms. The characteristic function is defined to be the Fourier transform of the transitional PDF if $\mathcal{S} = \mathbb{R}$ and the Laplace transform if $\mathcal{S} = \mathbb{R}^+$. After presenting a general introduction to the characteristic function, the method of characteristics is used to derive closed-form expressions for the transitional PDF of the OU and CIR models respectively.

Characteristic function Suppose that X is the stochastic process satisfying the initial boundary value problem posed by

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial x} \left(\frac{1}{2} \frac{\partial (g^2(x; \boldsymbol{\theta}) f)}{\partial x} - \mu(x; \boldsymbol{\theta}) f \right) \quad x \in \mathcal{S}, t > t_0, \quad (122)$$

with initial and boundary conditions

$$\begin{aligned} f(x, t_0) &= \delta(x - X_0), \quad x \in \mathcal{S}, \\ q &= \mu(x; \boldsymbol{\theta}) f - \frac{1}{2} \frac{\partial (g^2(x; \boldsymbol{\theta}) f)}{\partial x} = 0, \quad x \in \partial \mathcal{S}, t > t_0, \end{aligned} \quad (123)$$

where $\delta(x)$ is the usual Dirac delta function, $q = q(x, t)$ is the flux of probability density at time t and state x and $\partial \mathcal{S}$ denotes the boundary of the region \mathcal{S} .

The characteristic function of X is defined by

$$\tilde{f}(p, t) = \int_{\mathcal{S}} e^{-px} f(x, t) dx \quad (124)$$

where $f(x, t) \equiv f((x, t) | (X_0, t_0); \boldsymbol{\theta})$ is the transitional PDF of X and p is a parameter which is restricted to those regions of the complex plane for which the convergence of the integral defining $\tilde{f}(p, t)$ is assured. If $\mathcal{S} = \mathbb{R}$, as occurs in the case of the OU process, then $p = i\omega$ where ω is real-valued, $i^2 = -1$ and the characteristic function is the familiar Fourier transform of the transitional PDF. On the other hand, if $\mathcal{S} = \mathbb{R}^+$ (or any semi-infinite interval) as happens in the case of the CIR process, then p is a complex number that is typically restricted to the half-plane $\text{Re}(p) > b$ where b is defined by the requirement that $|f(x, t)| < M e^{bx}$ for all $x > 0$ and the characteristic function is now the Laplace transform of the transitional PDF.

Independent of the choice of transform, the Fokker-Planck equation given in (122) is first used to show that

$$\frac{\partial \tilde{f}}{\partial t} = \int_{\mathcal{S}} e^{-px} \frac{\partial f(x, t)}{\partial t} dx = - \int_{\mathcal{S}} e^{-px} \frac{\partial q(x, t)}{\partial x} dx = -p \int_{\mathcal{S}} e^{-px} q(x, t) dx = -p \tilde{q} \quad (125)$$

where $\tilde{q}(p, t)$ is the transform of the probability flux and the boundary conditions $q = 0$ required to conserve transitional probability density in \mathcal{S} likewise causes the contributions from the boundary of \mathcal{S} to vanish in the derivation of equation (125). A further integration by parts gives

$$\tilde{q} = \int_{\mathcal{S}} e^{-px} \mu(x; \boldsymbol{\theta}) f(x, t) dx - \frac{1}{2} \left[e^{-px} g^2(x; \boldsymbol{\theta}) f(x, t) \right]_{\partial \mathcal{S}} - \frac{p}{2} \int_{\mathcal{S}} e^{-px} g^2(x; \boldsymbol{\theta}) f(x, t) dx. \quad (126)$$

Typically the product $g^2(x; \boldsymbol{\theta})f(x, t) \rightarrow 0$ as $x \rightarrow \partial\mathcal{S}$, and with this proviso, the characteristic function¹⁷ of X is the solution of the initial value problem

$$\frac{\partial \tilde{f}}{\partial t} = \frac{p^2}{2} \int_{\mathcal{S}} e^{-px} g^2(x; \boldsymbol{\theta}) f(x, t) dx - p \int_{\mathcal{S}} e^{-px} \mu(x; \boldsymbol{\theta}) f(x, t) dx, \quad \tilde{f}(p, t_0) = e^{-pX_0}. \quad (127)$$

Once the characteristic function has been found, the transitional PDF may, in principle, be obtained through the use of an inverse transform.

Cox, Ingersoll and Ross model The square-root process proposed by Cox, Ingersoll and Ross (1985) as a model of the instantaneous short term interest rate, commonly referred to as the CIR model, evolves according to the SDE

$$dX = \alpha(\beta - X) dt + \sigma\sqrt{X} dw \quad (128)$$

where α (speed of adjustment), β (the mean interest rate) and σ (volatility control) are positive parameters to be estimated. Thus the CIR process exhibits mean reversion of X to the state $X = \beta$. Most importantly, however, the properties $g(0; \boldsymbol{\theta}) = 0$ and $\mu(0; \boldsymbol{\theta}) > 0$ ensure that $\mathcal{S} = \mathbb{R}^+$.

The transitional probability density function of the CIR process will now be constructed using the method of characteristics. Since the state space of the CIR process is \mathbb{R}^+ , the Laplace transform provides a suitable characterisation of the process. The crucial idea is to recognise that whenever $\mu(x; \boldsymbol{\theta})$ and $g^2(x; \boldsymbol{\theta})$ are affine functions of state, their Laplace (or Fourier) transforms can be expressed as linear combination of the characteristic function itself and its partial derivative with respect to the parameter p . The drift and diffusion specifications for the CIR process in equation (128) are respectively $\mu(x; \boldsymbol{\theta}) = \alpha(\beta - x)$ and $g^2(x; \boldsymbol{\theta}) = \sigma^2 x$, and therefore

$$\int_0^\infty e^{-px} g^2(x; \boldsymbol{\theta}) f(x, t) dx = -\sigma^2 \frac{\partial \tilde{f}}{\partial p}, \quad \int_0^\infty e^{-px} \mu(x; \boldsymbol{\theta}) f(x, t) dx = \alpha\beta \tilde{f} + \alpha \frac{\partial \tilde{f}}{\partial p}. \quad (129)$$

It now follows immediately from equation (127) that the characteristic function of the CIR process starting in state X_k at time t_k satisfies the initial value problem

$$\frac{\partial \tilde{f}}{\partial t} = -\left(\frac{p^2\sigma^2}{2} + \alpha p\right) \frac{\partial \tilde{f}}{\partial p} - p\alpha\beta \tilde{f}, \quad \tilde{f}(p, t_k) = e^{-pX_k}.$$

This partial differential equation can be solved immediately using the method of characteristics. Briefly, on the curve defined by the solution of the initial value problem

$$\frac{dp}{dt} = \frac{p^2\sigma^2}{2} + \alpha p, \quad p(t_k) = s, \quad (130)$$

the characteristic function \tilde{f} satisfies the initial value problem

$$\frac{d\tilde{f}}{dt} = \frac{\partial \tilde{f}}{\partial t} + \frac{\partial \tilde{f}}{\partial p} \frac{dp}{dt} = -p\alpha\beta \tilde{f}, \quad \tilde{f}(s, t_k) = e^{-sX_k}. \quad (131)$$

¹⁷It should be noted that the procedure described here in one dimension extends to multi-dimensions. The Fokker-Planck equation becomes the conservation equation $\partial f(\mathbf{x}, t)/\partial t + \text{div } \mathbf{q} = 0$ and integration by parts generalises to Gauss's theorem.

Equation (130) is a Bernoulli equation with particular solution

$$p = \frac{2\alpha s e^{\alpha(t-t_k)}}{2\alpha + s\sigma^2(1 - e^{\alpha(t-t_k)})}. \quad (132)$$

This solution for p is now substituted into equation (131) and the resulting differential equation integrated with respect to t to obtain

$$\tilde{f}(s, t) = e^{-sX_k} \left[1 + \frac{s\sigma^2}{2\alpha} (1 - e^{\alpha(t-t_k)}) \right]^{2\alpha\beta/\sigma^2}. \quad (133)$$

The characteristic function of X at time t_{k+1} is now constructed by eliminating the parameter s between equations (132) and (133). After some straightforward algebra it can be shown that

$$\tilde{f}(p, t_{k+1}) = c^{\nu+1} e^{-u} (p+c)^{-\nu-1} \exp\left[\frac{cu}{p+c}\right] \quad (134)$$

where c , u , v and ν are defined respectively by

$$c = \frac{2\alpha}{\sigma^2(1 - e^{-\alpha(t_{k+1}-t_k)})}, \quad u = cX_k e^{-\alpha(t_{k+1}-t_k)}, \quad v = cx, \quad \nu = \frac{2\alpha\beta}{\sigma^2} - 1. \quad (135)$$

The final stage in this argument is to observe from tables of Laplace transforms¹⁸ that

$$\int_0^\infty \left(\frac{x}{\eta}\right)^{\nu/2} I_\nu(2\sqrt{\eta x}) e^{-px} dx = p^{-\nu-1} e^{\eta/p}, \quad \text{Re}(\nu) > -1$$

where $I_\nu(x)$ is the modified Bessel function of the first kind of order ν . This identity in combination with the shift theorem for Laplace transforms, namely that the function $e^{-cx} f(x, t_{k+1})$ has Laplace transform $\tilde{f}(p+c, t_{k+1})$, indicates that the CIR process which started at X_k at time t_k diffuses to a final state at time t_{k+1} that is non-central chi-squared distributed with transitional PDF

$$f(x | X_k; \boldsymbol{\theta}) = c \left(\frac{v}{u}\right)^{\frac{q}{2}} e^{-(\sqrt{u}-\sqrt{v})^2} e^{-2\sqrt{uv}} I_q(2\sqrt{uv}) \quad (136)$$

where c , u , v and ν are defined in (135). This transitional PDF may now be used in combination with expression (2) to estimate the values of the parameters of the CIR model (128) by EML.

Ornstein-Uhlenbeck model The OU process proposed by Vasicek (1977) evolves according to the SDE

$$dX = \alpha(\beta - X) dt + \sigma dW \quad (137)$$

where α (speed of adjustment), β (the mean interest rate) and σ (volatility control) are again the parameters to be estimated. The OU process also exhibits mean reversion of X to the state $X = \beta$, but unlike the CIR process, the domain of the state variable is unrestricted, that is, $\mathcal{S} = \mathbb{R}$. The analysis of the OU process mirrors that of the CIR process with the exception that the characteristic

¹⁸One useful source of Fourier and Laplace transforms with their inverses is provided in Volumes I and II of the Bateman Manuscript Project (1954). The Laplace transform needed here is result (35) on page 245 of Volume I.

function is now the Fourier transform of the transitional PDF. In this case $p = i\omega$ where $i^2 = -1$. Briefly, the characteristic function of the OU process satisfies the initial value problem

$$\frac{\partial \tilde{f}}{\partial t} = -\left(\frac{\omega^2 \sigma^2}{2} + i\omega \alpha \beta\right) \tilde{f} - \alpha \omega \frac{\partial \tilde{f}}{\partial \omega}, \quad \tilde{f}(\omega, t_k) = e^{-i\omega X_k}. \quad (138)$$

The characteristic procedure described in detail in the treatment of the CIR process is now used to solve equation (138) and eventually leads to the characteristic function

$$\tilde{f}(\omega, t_{k+1}) = \exp \left[-i\omega (X_k + (\beta - X_k)(1 - e^{-\alpha(t_{k+1}-t_k)})) - \frac{\omega^2 \sigma^2}{4\alpha} (1 - e^{-2\alpha(t_{k+1}-t_k)}) \right]. \quad (139)$$

This is the characteristic function of the Normal distribution with mean value $X_k + (\beta - X_k)(1 - e^{-\alpha(t_{k+1}-t_k)})$ and variance $\sigma^2(1 - e^{-2\alpha(t_{k+1}-t_k)})/2\alpha$. The transitional PDF of X at time t_{k+1} for the process starting at X_k at time t_k therefore has closed-form expression

$$f(x | X_k; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi V}} \exp \left[-\frac{(x - \bar{x})^2}{2V} \right] \quad (140)$$

where

$$V = \frac{\sigma^2(1 - e^{-2\alpha(t_{k+1}-t_k)})}{2\alpha}, \quad \bar{x} = \beta + (X_k - \beta) e^{-\alpha(t_{k+1}-t_k)}.$$

This transitional PDF may now be used in combination with expression (2) to estimate the values of the parameters of the OU model (137) by EML.

Appendix 3: Discrete maximum likelihood

CIR process The CIR process evolves according to the SDE

$$dx = \alpha(\beta - x)dt + \sigma\sqrt{x}dW \quad (141)$$

where α , β and σ are parameters to be estimated from $(N + 1)$ observations X_0, \dots, X_N of the process at the deterministic times t_0, \dots, t_N . Traditional DML uses the Euler-Maruyama algorithm to approximate the solution of equation (141) by one integration step of duration $\Delta_k = (t_{k+1} - t_k)$ to obtain

$$X = X_k + \alpha(\beta - X_k)\Delta_k + \sigma\sqrt{X_k}\varepsilon_k \quad (142)$$

where $\varepsilon_k \sim N(0, \Delta_k)$. To maintain generality, the DML procedure is developed for observations that are spaced non-uniformly in time, although in practice it is often the case that $\Delta_k = \Delta$ for all values of k , that is, the observations are spaced uniformly in time. Equation (142) indicates that the transitional density of (X, t_{k+1}) for a CIR process starting at (X_k, t_k) is

$$f((X, t_{k+1}) | (X_k, t_k); \theta) = \frac{1}{\sqrt{2\pi\sigma^2 X_k \Delta_k}} \exp\left[-\frac{(X - X_k - \alpha(\beta - X_k)\Delta_k)^2}{2\sigma^2 X_k \Delta_k}\right],$$

and therefore the likelihood \mathcal{L} of observing the states X_0, \dots, X_N at the times t_0, \dots, t_N is

$$\mathcal{L} = \prod_{k=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2 X_k \Delta_k}} \exp\left[-\frac{(X - X_k - \alpha(\beta - X_k)\Delta_k)^2}{2\sigma^2 X_k \Delta_k}\right]. \quad (143)$$

By choice of the parameters α , β and σ in expression (143), the traditional DML procedure either seeks to maximise the value of \mathcal{L} , or alternatively, seeks to minimise the value of the negative log-likelihood

$$-\log \mathcal{L} = \frac{1}{2} \sum_{k=0}^{N-1} \frac{(X_{k+1} - X_k - \alpha(\beta - X_k)\Delta_k)^2}{\sigma^2 X_k \Delta_k} + \frac{N}{2} \log(2\pi\sigma^2 X_k \Delta_k). \quad (144)$$

The partial derivatives of the negative log-likelihood function with respect to the parameters α , β and σ are respectively

$$\begin{aligned} -\frac{\partial \log \mathcal{L}}{\partial \alpha} &= -\frac{1}{\sigma^2} \sum_{k=0}^{N-1} \frac{(\beta - X_k)(X_{k+1} - X_k - \alpha(\beta - X_k)\Delta_k)}{X_k}, \\ -\frac{\partial \log \mathcal{L}}{\partial \beta} &= -\frac{\alpha}{\sigma^2} \sum_{k=0}^{N-1} \frac{X_{k+1} - X_k - \alpha(\beta - X_k)\Delta_k}{X_k}, \\ -\frac{\partial \log \mathcal{L}}{\partial \sigma} &= -\frac{1}{\sigma^3} \sum_{k=0}^{N-1} \frac{(X_{k+1} - X_k - \alpha(\beta - X_k)\Delta_k)^2}{X_k \Delta_k} + \frac{N}{\sigma}. \end{aligned} \quad (145)$$

The optimal values of the parameters, say $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}$, are determined by the requirement that the gradient of the negative log-likelihood function is zero. The third equation of (145) indicates that

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{k=0}^{N-1} \frac{(X_{k+1} - X_k - \hat{\alpha}(\hat{\beta} - X_k)\Delta_k)^2}{X_k \Delta_k}, \quad (146)$$

and so $\hat{\sigma}$ can be recovered immediately once the optimal values of $\hat{\alpha}$ and $\hat{\beta}$ are determined. On the other hand, the optimal values of $\hat{\alpha}$ and $\hat{\beta}$ satisfy

$$\begin{aligned} \sum_{k=0}^{N-1} \frac{(\hat{\beta} - X_k)(X_{k+1} - X_k - \hat{\alpha}(\hat{\beta} - X_k)\Delta_k)}{X_k} &= 0, \\ \sum_{k=0}^{N-1} \frac{X_{k+1} - X_k - \hat{\alpha}(\hat{\beta} - X_k)\Delta_k}{X_k} &= 0. \end{aligned} \quad (147)$$

The second of equations (147) is used to simplify the first of equations (147) to obtain

$$X_N - X_0 = \sum_{k=0}^{N-1} X_{k+1} - X_k = \hat{\alpha} \sum_{k=0}^{N-1} (\hat{\beta} - X_k)\Delta_k = \hat{\alpha} \left(\hat{\beta} \sum_{k=0}^{N-1} \Delta_k - \sum_{k=0}^{N-1} X_k \Delta_k \right).$$

The second equation in (147) may itself be reorganised to give

$$\sum_{k=0}^{N-1} \frac{X_{k+1} - X_k}{X_k} = \hat{\alpha} \left(\hat{\beta} \sum_{k=0}^{N-1} \frac{\Delta_k}{X_k} - \sum_{k=0}^{N-1} \Delta_k \right).$$

To summarise the situation to date, the optimal values of the parameters α and β satisfy the equations

$$\begin{aligned} \hat{\alpha} \left(\hat{\beta} \sum_{k=0}^{N-1} \Delta_k - \sum_{k=0}^{N-1} X_k \Delta_k \right) &= X_N - X_0, \\ \hat{\alpha} \left(\hat{\beta} \sum_{k=0}^{N-1} \frac{\Delta_k}{X_k} - \sum_{k=0}^{N-1} \Delta_k \right) &= \sum_{k=0}^{N-1} \frac{X_{k+1} - X_k}{X_k}. \end{aligned} \quad (148)$$

Each summation appearing in these equations is computable from the observations and times. Clearly $\hat{\alpha}$ and $\hat{\beta}$ may be calculated directly from equations (148), for example, by dividing the first equation by the second equation to eliminate $\hat{\alpha}$ and generate a linear equation to be solved for $\hat{\beta}$. Once $\hat{\alpha}$ and $\hat{\beta}$ are calculated, $\hat{\sigma}$ is determined from equation (146).

OU process The OU process evolves according to the SDE

$$dx = \alpha(\beta - x)dt + \sigma dW \quad (149)$$

where α , β and σ are parameters to be estimated from $(N+1)$ observations X_0, \dots, X_N of the process at the deterministic times t_0, \dots, t_N . The Euler-Maruyama algorithm is again used to approximate the solution of equation (149) by one integration step of duration $\Delta_k = (t_{k+1} - t_k)$ to get

$$X = X_k + \alpha(\beta - X_k)\Delta_k + \sigma \varepsilon_k \quad (150)$$

where $\varepsilon_k \sim N(0, \Delta_k)$. As with the DML development of the CIR process, the treatment of the OU process again maintains generality by assuming that the observations are spaced non-uniformly in time, although in practice this is often not the case. Equation (150) indicates that the transitional density of (X, t_{k+1}) for an OU process starting at (X_k, t_k) is

$$f((X, t_{k+1}) | (X_k, t_k); \theta) = \frac{1}{\sqrt{2\pi\sigma^2\Delta_k}} \exp \left[-\frac{(X - X_k - \alpha(\beta - X_k)\Delta_k)^2}{2\sigma^2\Delta_k} \right],$$

and therefore the likelihood \mathcal{L} of observing the states X_0, \dots, X_N at the times t_0, \dots, t_N is

$$\mathcal{L} = \prod_{k=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2\Delta_k}} \exp \left[-\frac{(X - X_k - \alpha(\beta - X_k)\Delta_k)^2}{2\sigma^2\Delta_k} \right]. \quad (151)$$

By choice of the parameters α , β and σ in expression (151), the traditional DML procedure either seeks to maximise the value of \mathcal{L} , or alternatively, seeks to minimise the value of the negative log-likelihood

$$-\log \mathcal{L} = \frac{1}{2} \sum_{k=0}^{N-1} \frac{(X_{k+1} - X_k - \alpha(\beta - X_k)\Delta_k)^2}{\sigma^2\Delta_k} + \frac{N}{2} \log(2\pi\sigma^2\Delta_k). \quad (152)$$

The partial derivatives of the negative log-likelihood function with respect to the parameters α , β and σ are respectively

$$\begin{aligned} -\frac{\partial \log \mathcal{L}}{\partial \alpha} &= -\frac{1}{\sigma^2} \sum_{k=0}^{N-1} (\beta - X_k)(X_{k+1} - X_k - \alpha(\beta - X_k)\Delta_k), \\ -\frac{\partial \log \mathcal{L}}{\partial \beta} &= -\frac{\alpha}{\sigma^2} \sum_{k=0}^{N-1} (X_{k+1} - X_k - \alpha(\beta - X_k)\Delta_k), \\ -\frac{\partial \log \mathcal{L}}{\partial \sigma} &= -\frac{1}{\sigma^3} \sum_{k=0}^{N-1} \frac{(X_{k+1} - X_k - \alpha(\beta - X_k)\Delta_k)^2}{\Delta_k} + \frac{N}{\sigma}. \end{aligned} \quad (153)$$

The optimal values of the parameters, say $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}$, are determined by the requirement that the gradient of the negative log-likelihood function is zero. The third equation of (153) indicates that

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{k=0}^{N-1} \frac{(X_{k+1} - X_k - \hat{\alpha}(\hat{\beta} - X_k)\Delta_k)^2}{\Delta_k}, \quad (154)$$

and so $\hat{\sigma}$ can be recovered immediately once the optimal values of $\hat{\alpha}$ and $\hat{\beta}$ are determined. On the other hand, the optimal values of $\hat{\alpha}$ and $\hat{\beta}$ satisfy

$$\begin{aligned} \sum_{k=0}^{N-1} (\hat{\beta} - X_k)(X_{k+1} - X_k - \hat{\alpha}(\hat{\beta} - X_k)\Delta_k) &= 0, \\ \sum_{k=0}^{N-1} (X_{k+1} - X_k - \hat{\alpha}(\hat{\beta} - X_k)\Delta_k) &= 0. \end{aligned} \quad (155)$$

The second of equations (155) may be reorganised immediately to give

$$X_N - X_0 = \sum_{k=0}^{N-1} X_{k+1} - X_k = \hat{\alpha} \sum_{k=0}^{N-1} (\hat{\beta} - X_k)\Delta_k = \hat{\alpha} \left(\hat{\beta} \sum_{k=0}^{N-1} \Delta_k - \sum_{k=0}^{N-1} X_k \Delta_k \right).$$

Alternatively, the second of equations (155) may be used to simplify the first of equations (155) to get

$$\sum_{k=0}^{N-1} (X_{k+1} - X_k)X_k = \hat{\alpha} \left(\hat{\beta} \sum_{k=0}^{N-1} X_k \Delta_k - \sum_{k=0}^{N-1} X_k^2 \Delta_k \right).$$

To summarise the situation to date, the optimal values of the parameters α and β satisfy the equations

$$\begin{aligned} \widehat{\alpha} \left(\widehat{\beta} \sum_{k=0}^{N-1} \Delta_k - \sum_{k=0}^{N-1} X_k \Delta_k \right) &= X_N - X_0, \\ \widehat{\alpha} \left(\widehat{\beta} \sum_{k=0}^{N-1} X_k \Delta_k - \sum_{k=0}^{N-1} X_k^3 \Delta_k \right) &= \sum_{k=0}^{N-1} X_k (X_{k+1} - X_k). \end{aligned} \tag{156}$$

Each summation appearing in these equations is computable from the observations and times. Clearly $\widehat{\alpha}$ and $\widehat{\beta}$ may be calculated directly from equations (156), for example, by dividing the first equation by the second equation to eliminate $\widehat{\alpha}$ and generate a linear equation to be solved for $\widehat{\beta}$. Once $\widehat{\alpha}$ and $\widehat{\beta}$ are calculated, $\widehat{\sigma}$ is determined from equation (154).

Appendix 4: Infinitesimal operator

The infinitesimal operator has the property that it expresses the time derivative of an expected value as an expected value taken in state space. This property is central to many estimation methods, including the Hermite polynomial, method of moments and estimating function approaches.

Let $\psi = \psi(x)$ be a function of state, then the expected value of ψ is

$$\mathbb{E}[\psi](t) = \int_{\mathcal{S}} \psi(x) f(x, t) dx \quad (157)$$

and evolves as a function of time with initial value $\psi(X_0)$. The conservation law representation of the Fokker-Planck equation given in (115) is now used to show that

$$\frac{d\mathbb{E}[\psi]}{dt} = \int_{\mathcal{S}} \psi(x) \frac{\partial f(x, t)}{\partial t} dx = - \int_{\mathcal{S}} \psi(x) \frac{\partial q(x, t)}{\partial x} dx = \int_{\mathcal{S}} q(x, t) \frac{d\psi}{dx} dx \quad (158)$$

where the usual contribution to the value of the integral from boundary terms vanishes in equation (158) since $q = 0$ on $\partial\mathcal{S}$. A further integration by parts applied to equation (158) gives

$$\frac{d\mathbb{E}[\psi]}{dt} = \int_{\mathcal{S}} \left(\mu(x; \boldsymbol{\theta}) \frac{d\psi}{dx} + \frac{g^2(x; \boldsymbol{\theta})}{2} \frac{d^2\psi}{dx^2} \right) f(x, t) dx \quad (159)$$

where the boundary contributions to the value of the integral are again assumed to vanish on the basis that $g^2(x; \boldsymbol{\theta}) f(x, t) \psi'(x) \rightarrow 0$ for all $t > t_0$ as $x \rightarrow \partial\mathcal{S}$. The *infinitesimal operator* of the diffusion X is now defined by

$$\mathcal{A}_{\boldsymbol{\theta}}(\psi) = \mu(x; \boldsymbol{\theta}) \frac{d\psi}{dx} + \frac{g^2(x; \boldsymbol{\theta})}{2} \frac{d^2\psi}{dx^2} \quad (160)$$

and has the property that

$$\frac{d\mathbb{E}[\psi]}{dt} = \mathbb{E}[\mathcal{A}_{\boldsymbol{\theta}}(\psi)]. \quad (161)$$

Furthermore, provided the drift and diffusion specifications of the model SDE in (1) are autonomous functions of state, then result (161) is a special case of the general result

$$\frac{d^n \mathbb{E}[\psi]}{dt^n} = \mathbb{E}[\mathcal{A}_{\boldsymbol{\theta}}^n(\psi)] \quad (162)$$

where n is any positive integer.

Appendix 5: Marginal distributions of the parameters of the CIR and OU processes

Given a sample of $(N + 1)$ observations of the process

$$dX = \mu(X; \boldsymbol{\theta}) dt + g(X; \boldsymbol{\theta}) dW \quad (163)$$

MCMC methods proceed by dividing the transition from (X_k, t_k) to (X_{k+1}, t_{k+1}) into m sub-transitions of duration $\Delta t = (t_{k+1} - t_k)/m$ to give mN sub-transitions in total. The likelihood function for this augmented sample is constructed from the product of the traditional DML likelihoods of these sub-transitions and is given by

$$\begin{aligned} \mathcal{L} &= \prod_{j=0}^{mN-1} \frac{1}{\sqrt{2\pi g^2(x_j^*; \boldsymbol{\theta}) \Delta t}} \exp \left[-\frac{(x_{j+1}^* - x_j^* - \mu(x_j^*; \boldsymbol{\theta}) \Delta t)^2}{2g^2(x_j^*; \boldsymbol{\theta}) \Delta t} \right] \\ &= \exp \left[-\sum_{j=0}^{mN-1} \frac{(x_{j+1}^* - x_j^* - \mu(x_j^*; \boldsymbol{\theta}) \Delta t)^2}{2g^2(x_j^*; \boldsymbol{\theta}) \Delta t} \right] \prod_{j=0}^{mN-1} \frac{1}{\sqrt{2\pi g^2(x_j^*; \boldsymbol{\theta}) \Delta t}}. \end{aligned} \quad (164)$$

Usually little is known about the prior distribution of the parameters, and consequently the likelihood (164) is treated as a joint probability density for the parameters given the current values for the unobserved datums. The CIR and OU processes to be considered in this article have drift specification $\mu(x; \boldsymbol{\theta}) = \kappa - \alpha x$ and diffusion specification $g^2(x; \boldsymbol{\theta}) = \sigma^2 h(x)$ where $h(x) = 1$ for the OU process and $h(x) = x$ for the CIR process. In particular, the drift specification is a linear function of its parameters¹⁹ and is independent of σ . These properties of the drift and diffusion functions facilitate the development of a procedure for drawing from the marginal distributions of the parameters.

Let $\boldsymbol{\theta}_\mu$ denote the model parameters excluding σ and let $\nu = mN$, then for the class of models characterised by the constitutive specifications $\mu = \mu(x; \boldsymbol{\theta}_\mu)$ and $g^2 = \sigma^2 h(x)$, the likelihood of the augmented sample given initially by expression (164) can be further simplified into the form

$$\mathcal{L} = \frac{e^{-\psi/\sigma^2}}{(2\pi\Delta t)^{\nu/2} \sigma^\nu} \left(\prod_{j=0}^{\nu-1} h(x_j^*) \right)^{-1/2} \quad (165)$$

where

$$\psi = \frac{1}{2\Delta t} \sum_{j=0}^{\nu-1} \frac{(x_{j+1}^* - x_j^* - \mu(x_j^*; \boldsymbol{\theta}_\mu) \Delta t)^2}{h(x_j^*)}.$$

The likelihood (165) is now scaled to a probability density function in σ and $\boldsymbol{\theta}_\mu$ to obtain

$$f_\sigma(\sigma, \boldsymbol{\theta}_\mu) = A \frac{e^{-\psi/\sigma^2}}{\sigma^\nu} \left(\prod_{j=0}^{\nu-1} h(x_j^*) \right)^{-1/2} \quad (166)$$

where A is chosen to ensure that $f_\sigma(\sigma, \boldsymbol{\theta}_\mu)$ encloses unit mass in parameter space. Let $Z = \psi/\sigma^2$, then the density of Z and $\boldsymbol{\theta}_\mu$ is constructed from $f_\sigma(\sigma, \boldsymbol{\theta}_\mu)$ in the usual way to get

$$f_Z(z, \boldsymbol{\theta}_\mu) = f_\sigma(\sigma, \boldsymbol{\theta}_\mu) \frac{d\sigma}{dz} = \frac{A}{2} \frac{z^{(\nu-3)/2} e^{-z}}{\psi^{(\nu-1)/2}} \left(\prod_{j=0}^{\nu-1} h(x_j^*) \right)^{-1/2}. \quad (167)$$

¹⁹Of course, in order to make this so, the drift function has been rewritten with α/β replaced by κ .

Since $f_Z(z, \boldsymbol{\theta}_\mu)$ is a product of a function of Z and a function of $\boldsymbol{\theta}_\mu$ then it follows immediately that Z and $\boldsymbol{\theta}_\mu$ are independently distributed random variables. Clearly Z is Gamma-distributed with parameter $(\nu - 1)/2$, and therefore a realisation of σ is obtained directly from $\sigma = \sqrt{\psi(\boldsymbol{\theta}_\mu)/Z}$ by drawing Z from the Gamma distribution once a realisation of $\boldsymbol{\theta}_\mu$ is available.

Again, it is clear from expression (167) for the joint probability density function of Z and $\boldsymbol{\theta}_\mu$ that $\boldsymbol{\theta}_\mu$ has multivariate PDF proportional to

$$f_{\boldsymbol{\theta}_\mu}(\boldsymbol{\theta}_\mu) = \frac{B}{\psi^{(\nu-1)/2}}. \quad (168)$$

where B is a suitable scaling constant. Provided the specification of $\mu(x_j^*; \boldsymbol{\theta}_\mu)$ is a linear function of $\boldsymbol{\theta}^*$, as happens in both the CIR and OU processes where $\mu(x; \boldsymbol{\theta}_\mu) = \kappa - \alpha x$, then ψ is a quadratic form in the parameters $\boldsymbol{\theta}_\mu$ and therefore the parameters themselves are multivariate student-t distributed.

CIR and OU processes The procedure for drawing the parameters $\boldsymbol{\theta}_\mu$ is now illustrated for the drift specification $\mu(x, \boldsymbol{\theta}_\mu) = \kappa - \alpha x$. In this case

$$\psi = \frac{1}{2\Delta t} \sum_{j=0}^{\nu-1} \frac{(x_{j+1}^* - x_j^* - (\kappa - \alpha x_j^*)\Delta t)^2}{h(x_j^*)} = \alpha^2 C_1 - 2\kappa\alpha C_2 + \kappa^2 C_3 + C_4 + 2\alpha C_5 - 2\kappa C_6 \quad (169)$$

where

$$\begin{aligned} C_1 &= \frac{\Delta t}{2} \sum_{j=0}^{\nu-1} \frac{x_j^{*2}}{h(x_j^*)}, & C_2 &= \frac{\Delta t}{2} \sum_{j=0}^{\nu-1} \frac{x_j^*}{h(x_j^*)}, \\ C_3 &= \frac{\Delta t}{2} \sum_{j=0}^{\nu-1} \frac{1}{h(x_j^*)}, & C_4 &= \frac{1}{2\Delta t} \sum_{j=0}^{\nu-1} \frac{(x_{j+1}^* - x_j^*)^2}{h(x_j^*)}, \\ C_5 &= \frac{1}{2} \sum_{j=0}^{\nu-1} \frac{x_j^*(x_{j+1}^* - x_j^*)}{h(x_j^*)}, & C_6 &= \frac{1}{2} \sum_{j=0}^{\nu-1} \frac{(x_{j+1}^* - x_j^*)}{h(x_j^*)}. \end{aligned} \quad (170)$$

The expression for ψ is manipulated into standard form in two stages. First, the solutions $\hat{\alpha}$ and $\hat{\kappa}$ of the equations $\partial\psi/\partial\alpha = \partial\psi/\partial\kappa = 0$ are sought. Thus $\hat{\alpha}$ and $\hat{\kappa}$ satisfy the linear equations

$$\hat{\alpha}C_1 - \hat{\kappa}C_2 = -C_5 \quad -\hat{\alpha}C_2 + \hat{\kappa}C_3 = C_6$$

with solutions

$$\hat{\alpha} = \frac{C_2C_6 - C_3C_5}{C_1C_3 - C_2^2}, \quad \hat{\kappa} = \frac{C_1C_6 - C_2C_5}{C_1C_3 - C_2^2}. \quad (171)$$

The solutions for $\hat{\alpha}$ and $\hat{\kappa}$ are now used to re-express ψ in the more convenient form

$$\psi = C_1 \left[(\alpha - \hat{\alpha}) - \frac{C_2}{C_1} (\kappa - \hat{\kappa}) \right]^2 + \left(C_3 - \frac{C_2^2}{C_1} \right) (\kappa - \hat{\kappa})^2 + \chi \quad (172)$$

where

$$\chi = \frac{C_3C_5^2 + C_1C_6^2 - 2C_2C_5C_6 + C_2^2C_4 - C_1C_3C_4}{C_2^2 - C_1C_3}.$$

Two new variables ζ and η are now defined by

$$\zeta = \sqrt{\frac{C_1}{\chi}} \left[(\alpha - \hat{\alpha}) - \frac{C_2}{C_1} (\kappa - \hat{\kappa}) \right], \quad \eta = \sqrt{\frac{C_1 C_3 - C_2^2}{C_1 \chi}} (\kappa - \hat{\kappa}). \quad (173)$$

With respect to these new variables $\psi = \chi(\zeta^2 + \eta^2 + 1)$ and the PDF from which the parameters α and κ are to be drawn now takes the form

$$f_{\boldsymbol{\theta}_\mu}(\boldsymbol{\theta}_\mu) = \frac{B}{\chi^{(\nu-1)/2}} \frac{1}{(\zeta^2 + \eta^2 + 1)^{(\nu-1)/2}}. \quad (174)$$

The value of α and κ are now straightforward to draw. First the marginal density of η is constructed in the usual way by integrating ζ out of expression (174). The result of this calculation is that $\eta\sqrt{\nu-3}$ is student-t distributed with $(\nu-3)$ degrees of freedom. The random variable $\eta\sqrt{\nu-3}$ is now drawn and the value of η determined. Given η , expression (174) indicates that $\zeta\sqrt{\nu-2}/\sqrt{1+\eta^2}$ is student-t distributed with $(\nu-2)$ degrees of freedom. Thus the value of ζ may be determined by drawing $\zeta\sqrt{\nu-2}/\sqrt{1+\eta^2}$ from a student-t distribution with $(\nu-2)$ degrees of freedom. Thus the value of ψ is now determined from which the parameter σ is obtained by drawing Z from the appropriate Gamma distribution. Of course, knowledge of η and ζ enables the values of κ and α to be backed out from formulae (173).

Appendix 6: Eigenfunctions of the infinitesimal operator

CIR process The eigenfunctions $\phi(x)$ of the infinitesimal operator of the CIR process

$$dX = \alpha(\beta - X)dt + \sigma\sqrt{X} dW \quad (175)$$

are solutions of the ordinary differential equation

$$\frac{\sigma^2 x}{2} \frac{d^2 \phi}{dx^2} + \alpha(\beta - x) \frac{d\phi}{dx} - \lambda \phi = 0. \quad (176)$$

Under the change of variable $z = 2\alpha x/\sigma^2$ equation (176) becomes

$$z \frac{d^2 \phi}{dz^2} + (\nu + 1 - z) \frac{d\phi}{dz} - \frac{\lambda}{\alpha} \phi = 0, \quad \nu = \frac{2\alpha\beta}{\sigma^2} - 1.$$

When $\lambda = -\alpha j$ this equation has the generalised Laguerre polynomial of degree j , namely $L_j^\nu(z)$, as one solution and so the estimating function procedure for the CIR process is based on the choice

$$\phi_j(x) = L_j^\nu(2\alpha x/\sigma^2), \quad \lambda_j = -\alpha j, \quad j > 0. \quad (177)$$

OU process The eigenfunctions $\phi(x)$ of the infinitesimal operator of the OU process

$$dX = \alpha(\beta - X)dt + \sigma dW \quad (178)$$

are solutions of the ordinary differential equation

$$\frac{\sigma^2}{2} \frac{d^2 \phi}{dx^2} + \alpha(\beta - x) \frac{d\phi}{dx} - \lambda \phi = 0. \quad (179)$$

Under the change of variable $z = \sqrt{\alpha}(x - \beta)/\sigma$ equation (179) becomes

$$\frac{d^2 \phi}{dz^2} - 2z \frac{d\phi}{dz} - \frac{2\lambda}{\alpha} \phi = 0$$

which is the generalised Hermite equation. This equation has the Hermite polynomial of degree j , namely $H_j(z)$, as one solution whenever $\lambda = -\alpha j$ and so the estimating function procedure for the OU process is based on the choice

$$\phi_j(x) = H_j(\sqrt{\alpha}(x - \beta)/\sigma), \quad \lambda_j = -\alpha j, \quad j > 0. \quad (180)$$