Ann Oper Res (2007) 151:269–288
DOI 10.1007/s10479-006-0126-4

Estimating parameters in diffusion processes using an approximate maximum likelihood approach

Erik Lindström

Published online: 21 November 2006

Abstract We present an approximate Maximum Likelihood estimator for univariate Itô stochastic differential equations driven by Brownian motion, based on numerical calculation of the likelihood function. The transition probability density of a stochastic differential equation is given by the Kolmogorov forward equation, known as the Fokker-Planck equation. This partial differential equation can only be solved analytically for a limited number of models, which is the reason for applying numerical methods based on higher order finite differences.

The approximate likelihood converges to the true likelihood, both theoretically and in our simulations, implying that the estimator has many nice properties. The estimator is evaluated on simulated data from the Cox-Ingersoll-Ross model and a non-linear extension of the Chan-Karolyi-Longstaff-Sanders model. The estimates are similar to the Maximum Likelihood estimates when these can be calculated and converge to the true Maximum Likelihood estimates as the accuracy of the numerical scheme is increased. The estimator is also compared to two benchmarks; a simulation-based estimator and a Crank-Nicholson scheme applied to the Fokker-Planck equation, and the proposed estimator is still competitive.

Keywords Approximate likelihood function · Durham-Gallant estimator · Crank-Nicholson scheme · Cox-Ingersoll-Ross model · Non-linear CKLS model

Modelling stochastic, non-linear dynamical systems using continuous time models differ from using discrete time models. The vast majority of statistical or econometric models are discrete time models as these are easier to estimate, but continuous time models have attracted more attention over the last years. There are several reasons for this.

The most important reason is that the laws of nature (interpreted in a broad sense) are usually defined in continuous time. This alone makes continuous time models the appropriate class of models for a large class of dynamical systems. Moreover, continuous time models can handle non-equidistantly sampled data consistently, which can be of importance

E. Lindström (✉)
Centre for Mathematical Sciences, Lund University, Box 118, SE-22100 Lund, Sweden
e-mail: erikl@maths.lth.se

for many macroeconomic applications. Finally, there are areas like financial mathematics where continuous time models (more specifically stochastic differential equations (SDEs) or diffusion processes) are the standard working tool to price contingent claims.

However, a theoretical model is only applicable when the parameters of the model are known. Reliable estimates of parameters in the mathematical models for derivative instruments on financial markets are of great economical importance.

Estimation of parameters in stochastic differential equations is a rapidly expanding area of research, see e.g. Nielsen, Madsen, and Young (2000) or Sørensen (2004). There are mainly two branches in the community of parameter estimation in stochastic differential equations; the branch adopting the Maximum Likelihood framework by approximating the likelihood function and the branch developing estimation techniques based on moment matching. It is sometimes argued that the class of estimators based on matching moments is a more general class of estimators than the class of likelihood-based estimators as the score function could (at least theoretically) be used as a moment. The moment-based estimator would then obtain Cramer-Rao efficiency. Some of the moment-based estimators are

- Generalized Method of Moments (GMM), see Hansen (1982), is an extension of the classical method of moments which, has the advantage of being less sensitive to specification errors than a Maximum Likelihood estimator. GMM has some drawbacks when modeling in continuous time as the moment conditions are simpler to implement if they are calculated for a discretized version of the SDE. However, this discretization introduces bias. Recent research has extended GMM to continuous time Markov processes, see e.g. Hansen and Scheinkman (1995), and Duffie and Glynn (2004).
- Indirect Inference, see Gourieroux, Monfort, and Renault (1993), and Efficient Method of Moments (EMM), see Gallant and Tauchen (1996), are special implementations of GMM using an auxiliary model to fit the parameters in the diffusion process, transforming the problem from finding moment conditions for the diffusion to finding an auxiliary model that can capture the features of the diffusion. Both estimators are consistent and it can be shown that EMM is asymptotically efficient as the score generator of the auxiliary model can approximate the log-likelihood arbitrarily well if enough data is available using e.g. semi non-parametric expansion of the score. However, there are no guarantees for efficiency for finite samples of data.
- Martingale Estimating functions, see e.g. Bibby and Sørensen (1995), is a mixture of the moment-based approach and the Maximum Likelihood approach. Estimation functions solve the estimation problem by finding the zeros of functions of data instead of through optimization. These functions should be chosen to mimic the score function, but conditional moments are sometimes used instead, as the score function usually is not available in closed form. Other choices include eigenfunctions of the generator, see Kessler and Sørensen (1999).

Common for all these estimators is that the efficiency depends on the moment generator (or auxiliary model). Another problem is that the methods mentioned above require higher order moments to be calculated in order to be efficient as the optimal weights for the moments used in estimation are chosen as the inverse of the covariance matrix of these moments. This can be a problem if data is heavy tailed, a common feature for financial data.

The loss of efficiency using moment-based estimators can sometimes be substantial, cf. Durham (2004), who finds Maximum Likelihood estimators to have significantly better precision when estimating parameters in non-linear diffusions. A fast and computationally

reliable method based on the Maximum Likelihood framework would hence be desirable. Previous research on approximate Maximum Likelihood estimation include

- Simulation of likelihood, is a very flexible method that has the drawback of being slow but the virtue of being simple to implement, see Pedersen (1995b). The method is also simple to generalize to multivariate systems. Markov Chain Monte Carlo (MCMC) techniques are suggested in Eraker (2001) and a faster MCMC algorithm was presented in Elerian, Chib, and Shephard (2001). The speed of Monte Carlo simulation-based estimators can be vastly improved by using importance sampling and other variance reduction techniques, see Elerian, Chib, and Shephard (2001) and Durham and Gallant (2002).
- Expansion of the transition probability density in Hermite polynomials, see Aït-Sahalia (1999, 2002), yielding a closed form expression for the transition probability density. This approach has a computational advantage both in terms of calculating the likelihood and in terms of optimizing the likelihood. The expansion is also very accurate, see e.g. Jensen and Poulsen (2002).

  The method corrects for deviation from the standard Gaussian density using an expansion in Hermite polynomials. However, the expansion requires the transition probabilities of the model to be fairly close to standard Gaussian to converge, and this is usually achieved by transforming the process to a representation where the diffusion term is independent of the process, and by standardizing the variance. The increments will then be close to standard Gaussian if the sampling interval is small, cf. the Euler-Maruyama discretization Kloeden and Platen (1992).

  It should be noted that there is a minor technical issue when applying the method. The required transformation does not exist (analytically) for models having complicated diffusion terms, e.g. the diffusion term of the preferred model in Aït-Sahalia (1996), $\sigma(x) = \sqrt{\theta_1 + \theta_2 x + \theta_3 x^{\theta_4}}$, does not belong to the class of (analytically) transformable models. Transforming the model using a numerical scheme would introduce numerical errors which are likely to dominate the errors of the method as the number of included correction terms increases. However, additional research has to be conducted to evaluate the consequences of a numerical transformation.
- Solving the Fokker-Planck equation numerically. The Fokker-Planck equation is a partial differential equation (PDE) describing the evolution of the transition probability density. Generalized Itô processes are treated in Lo (1988), while Poulsen (1999) and Christensen, Poulsen, and Sørensen (2001) estimate parameters in general univariate diffusion processes by solving the Fokker-Planck equation using Crank-Nicholson finite difference procedures.

Different estimators for diffusion processes were compared in Christensen, Poulsen, and Sørensen (2001) where the optimal quadratic martingale estimation function and a Fokker-Planck-based approximate Maximum Likelihood estimator were compared to exact and discretized GMM, Quasi Maximum Likelihood and indirect inference estimators. Their conclusion is that the approximate Maximum Likelihood estimator is slightly more accurate than the optimal quadratic martingale estimation function and is more accurate than the other estimators. It was also shown that the computational demand for the approximate Maximum Likelihood estimator is less than for e.g. the indirect inference estimator, making the approximate Maximum Likelihood estimator the estimator of choice.

Different approximate Maximum Likelihood estimators were evaluated in Jensen and Poulsen (2002), using the Vasiček model, the Cox-Ingersoll-Ross model and Geometric Brownian motion as benchmark models. They found that the Hermite polynomial expansion is the fastest method of the approximate Maximum Likelihood estimators, followed by the

Fokker-Planck-based estimator, while the Monte Carlo-based estimator proposed in Pedersen (1995b) lags behind.

But is this conclusion uniformly valid? The introduction of fast simulation-based approximate Maximum Likelihood estimators, see e.g. Durham and Gallant (2002) has presented another option. Furthermore, little is known for processes outside the class of transformable processes. The Fokker-Planck-based estimators and the simulation-based estimators can be applied to a larger class of processes than the class of analytically transformable processes. These estimators allow us to test general non-linear models, which is the reason why we develop an estimator based on solving the Fokker-Planck equation numerically using higher order numerical schemes and comparing the estimator to the Crank-Nicholson scheme in Poulsen (1999) and the simulation-based estimator in Durham and Gallant (2002). The estimators are applied without using any transformations of the models to compare the performance of the estimators for non-Gaussian processes.

1 Models

Consider a univariate stochastic process $\{X_t\}$ defined on a subset of $\mathbb{R}$ and described by a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^d$. Let the process $\{X_t\}$ be a solution to the Itô Stochastic Differential Equation

$$\begin{aligned} dX_t &= \mu(t, X_t; \theta)\, dt + \sigma(t, X_t; \theta)\, dW_t, \\ Y_k &= X_{t_k}, \end{aligned} \tag{1}$$

where the drift term, $\mu : [0, T] \times \mathbb{R} \times \Theta \mapsto \mathbb{R}$ and the diffusion term $\sigma : [0, T] \times \mathbb{R} \times \Theta \mapsto \mathbb{R}$ are sufficiently smooth, measurable functions and W_t is a standard Brownian motion.

We have studied two models in this paper; the Cox-Ingersoll-Ross (CIR) model, see Cox, Ingersoll, and Ross (1985) and a non-linear extension of the Chan-Karolyi-Longstaff-Sanders (CKLS) model, see Aït-Sahalia (1996), and Durham (2004).

1.1 Cox-Ingersoll-Ross

The Cox-Ingersoll-Ross model Cox, Ingersoll, and Ross (1985) was suggested as a model of the short interest rate, although the mathematical model was originally introduced by Feller (1951). Different parameterizations have been presented in the literature, but we have the used the following

$$\begin{aligned} dX_t &= \alpha\,(\beta - X_t)\; dt + \sigma\sqrt{X_t}\, dW_t, \\ X_0 &= x_0 > 0. \end{aligned} \tag{2}$$

By limiting the parameter space Θ to $\{\alpha, \beta, \sigma\} = \{(0, \infty) \times (0, \infty) \times (0, \infty)\}$, the state space is given by $\{X_t, t\} = \{[0, \infty) \times [0, T]\}$. The origin $(x = 0)$ is inaccessible if $2\alpha\beta \geq \sigma^2$, otherwise it is reflecting, see Feller (1951). Similarly, the Maximum Likelihood regularity conditions are valid if $2\alpha\beta \geq \sigma^2$, see Overbeck and Rydén (1997). The success of this model is due to the fact that (given the requirements on the parameters)

- The process is always non-negative.
- The mean converges towards the steady-state mean, β.

– Closed form expressions can be derived for a large class of financial contracts due to the affine form of the drift term and the squared diffusion term.

Furthermore, the parameters in the model can be estimated using Maximum Likelihood estimation since the transition probabilities are explicitly known. It can be shown that the transition probability density is given by

$$\begin{aligned} p(t, x_t; s, x_s; \theta) &= \frac{\partial}{\partial x_t} P_\theta(X_t \leq x_t \mid X_s = x_s, \theta) \\ &= c \exp(-cx_t - c\delta x_s) \left(\frac{x_t}{x_s \delta}\right)^{q/2} I\big(q, 2c\sqrt{x_s x_t \delta}\big), \end{aligned} \tag{3}$$

where

$$\delta = e^{-\alpha(t-s)}, \quad c = \frac{2\alpha}{\sigma^2(1-\delta)}, \quad q = \frac{2\alpha\beta}{\sigma^2} - 1, \tag{4}$$

and $I(q, z)$ is a modified Bessel function of the first kind of order q. We will use the Cox-Ingersoll-Ross model to measure the accuracy of the approximation of the transition probability density, as the model is non-linear, yet has a closed form expression for the transition probability density.

1.2 Non-linear CKLS

A generalization of the CIR model is the CKLS model, see Chan et al. (1992). The difference between these models is that the CKLS-model allows the diffusion term to depend on the state in a general way.

The CKLS model was further extended in Aït-Sahalia (1996), although later studies, see e.g. Durham (2004) did not find statistical support for the full model, specified as

$$dX_t = \left(a_0 + a_1 X_t + a_2 X_t^2 + \frac{a_3}{X_t}\right) dt + \sqrt{\theta_1 + \theta_2 X_t + \theta_3 X_t^{\theta_4}}, \tag{5}$$

where $\theta_1 + \theta_2 x + \theta_3 x^{\theta_4} > 0$ for all values of $x > 0$. This model cannot be estimated using the Hermite expansion method, cf. Aït-Sahalia (2002), but the Fokker-Planck or simulation-based estimators should be able to estimate the parameters.

2 Estimators

We present general approximate Maximum Likelihood Estimators (AMLE) for non-linear diffusion processes. Different numerical schemes for solving the Fokker-Planck equation are compared to the simulation-based estimator presented in Durham and Gallant (2002) and the Fokker-Planck-based estimator presented in Poulsen (1999).

The estimators are evaluated on simulated data (using exact simulation) from the Cox-Ingersoll-Ross model using several different parameter vectors and on simulated data from the non-linear CKLS model. The benchmarks are similar to Durham and Gallant (2002) and Jensen and Poulsen (2002) but has been modified to give an estimate of the criteria derived in Pedersen (1995a).

2.1 Durham & Gallant estimator

Numerical approximation of the likelihood for diffusions processes using Monte Carlo techniques was introduced in Pedersen (1995b). An approximation was obtained using iterated Euler approximations, see e.g. Kloeden and Platen (1992)

$$p^E(t, x_t; s, x_s) \in N(x_s + \mu(s, x_s)(t - s), \sigma^2(s, x_s)(t - s)) \tag{6}$$

to calculate

$$p^{(M)}(t, x_t; s, x_s) = \int \prod_{m=1}^{M} p^E(\tau_m, u_m; \tau_{m-1}, u_{m-1})\, d\lambda(u_1, \ldots, u_{M-1}), \tag{7}$$

where $u_0 = x_s$, $u_M = x_t$, and λ is the Lebesgue measure. A Monte Carlo approximation is obtained by simulating $u_{m,k}$ from the Euler scheme $u_{m,k} = u_{m-1,k} + \mu(\tau_{m-1}, u_{m-1,k})(\tau_m - \tau_{m-1}) + \sigma(\tau_{m-1}, u_{m-1,k})(dW_{m,k} - dW_{m-1,k})$. The approximation of $p^{(M)}(t, x_t; s, x_s)$ is then written as

$$p^{(M,K)}(t, x_t; s, x_s) \approx \frac{1}{K} \sum_{k=1}^{K} p^E(t, x_t; \tau_{M-1}, u_{M-1,k}). \tag{8}$$

Although this method has great theoretical appeal, see Pedersen (1995a), the computational efficiency is not competitive, see Jensen and Poulsen (2002), and Durham and Gallant (2002).

The approach taken in Durham and Gallant (2002) is to introduce an importance sampler, i.e. to sample more frequently from the relevant parts of the distribution. Formally, let $q(u_1, \ldots, u_{M-1})$ be a probability density on $\mathbb{R}^{M-1}$ and let $\{\mathbf{u}_k = (\mathbf{u}_{1,k}, \ldots, \mathbf{u}_{M-1,k}), k = 1, \ldots, K\}$ be independent samples from $q(u_1, \ldots, u_{M-1})$. The transition probability density can be calculated as

$$p^{(M,K)}(t, x_t; s, x_s) \approx \frac{1}{K} \sum_{k=1}^{K} \frac{\prod_{m=1}^{M} p^E(\tau_m, u_{m,k}; \tau_{m-1}, u_{m-1,k})}{q(u_{k,1}, \ldots, u_{k,M-1})}. \tag{9}$$

It is argued in Durham and Gallant (2002) that this approximation converges to the transition probability density as $M, K \to \infty$. However, the flexibility of the importance sampler allows for good approximations using moderate values of M and K.

The performance of the approximation depends heavily on a clever choice of the importance sampler, $q(u_1, \ldots, u_{M-1})$. We have used the modified bridge sampler in Durham and Gallant (2002) and the Milstein/Elerian subdensity, and the performance was improved further by using normalized variates. This approximation was found to be the best approximation for models with state dependent diffusion terms, see Durham and Gallant (2002).

2.2 Fokker-Planck-based estimators

Another method of calculating the transition probability density $p(t, x_t; s, x_s)$ is to solve the Fokker-Planck equation. We have implemented several different numerical algorithms approximating the time and space derivatives at different levels of accuracy. The partial differential equation was discretized using $O(h^2)$ and $O(h^4)$ discretizations of the space

derivatives and Padé(1,1) and Padé(2,2) approximations of the matrix exponential. Finally, different approximations of the initial condition were also evaluated.

It was shown in Lindström (2003) that Richardson extrapolation applied to the $O(h^2)$ solution gives similar results to the $O(h^4)$ discretization but the algorithm is somewhat slower. It is also less robust, and has thus been excluded.

2.2.1 Fokker-Planck equation

It is known from e.g. Karatzas and Shreve (1999), and Øksendal (2000) that, given some regularity conditions, the time evolution of the transition probability density $p(t, x_t; s, x_s)$ of the solution of a stochastic differential equation $dX_t = \mu(t, X_t; \theta)dt + \sigma(t, X_t; \theta)dW_t$ is given by the Kolmogorov forward equation, or the Fokker-Planck equation,

$$\frac{\partial p}{\partial t} = \mathscr{A}^\star p, \tag{10}$$

$$p(s, x; s, x_s) = \delta(x - x_s), \tag{11}$$

where the operator $\mathscr{A}^\star$ applied on $p(t, x_t; s, x_s)$ is given by

$$\mathscr{A}^\star p(t, x_t; s, x_s) = -\frac{\partial}{\partial x_t}(\mu(t, x_t)p(t, x_t; s, x_s)) + \frac{1}{2}\frac{\partial^2}{\partial x_t^2}(\sigma^2(t, x_t)p(t, x_t; s, x_s)). \tag{12}$$

It is possible to derive the transition probability densities, $p(t, x_t; s, x_s)$ using numerical methods. We have solved the Fokker-Planck equation using the method of lines, see e.g. Smith (1986). This means that a grid in the space-time domain approximates the domain for which the continuous Fokker-Planck equation is defined. The space derivatives in the partial differential equation are approximated by central finite differences of $O(h^2)$ and $O(h^4)$ accuracy, where h is the distance between the lines in the grid, transforming the problem into a system of ordinary differential equations. The resulting system of ordinary differential equations is of the form

$$\frac{d\mathbf{p}}{dt} = \mathsf{A}(t)\mathbf{p} + \mathbf{b}(t), \tag{13}$$

where $\mathsf{A}(t)$ is the finite difference approximation of $\mathscr{A}^\star$ and $\mathbf{b}(t)$ corrects for boundary conditions. $\mathsf{A}(t)$ and $\mathbf{b}(t)$ would be time dependent in the general case but are time invariant for the models used in this paper.

The problem of solving a partial differential equation is transformed into finding the solution to the system of ordinary differential equations as fast and as accurately as possible. This system of linear ordinary differential equations can be solved analytically. Denote the initial condition $\mathbf{p}(0) = \mathbf{p}_0$. The vector valued solution, assuming that A and $\mathbf{b}$ are time invariant, is given by

$$\mathbf{p}(t) = -\mathsf{A}^{-1}\mathbf{b} + \exp(t\mathsf{A})(\mathbf{p_0} + \mathsf{A}^{-1}\mathbf{b}). \tag{14}$$

The system is stiff, i.e. the largest eigenvalue is much larger than the smallest. Stiff systems of ordinary differential equations should preferably be solved using implicit numerical methods. We have tried two different Padé approximations to calculate the matrix exponential, the Padé(1,1) approximation and the Padé(2,2) approximation. The Padé(1,1) approximation is equivalent to the Crank-Nicholson method and is able to handle stiff problems, see Smith (1986). The approximation is given as $\exp(\xi) \approx (1 - \xi/2)^{-1}(1 + \xi/2)$. The Padé(2,2) is

similar but more accurate than the Padé(1,1) approximation, see Smith (1986), and is given as $\exp(\xi) \approx (1 - \xi/2 + \xi^2/12)^{-1}(1 + \xi/2 + \xi^2/12)$. The accuracy of the approximation is improved by splitting the distance between the observations into $Nstep$ equal intervals, applying the approximation to each of the intervals.

Defining the grid over $\mathbb{R}^+$ or $\mathbb{R}$ would lead to an infinite dimensional system of differential equations. This is not necessary as the transition probability density is concentrated to a much smaller subset of $\mathbb{R}$, $\mathbb{S}$. Thus, the equation is only solved for the subset $\mathbb{S}$. The subset is defined by discretizing the model using an Euler approximation, and estimating the parameters using GMM, see e.g. Chan et al. (1992). These estimated parameters are used to define $\mathbb{S}$ as the expected value $\pm 6\sigma$ for the approximated model. The subset is modified if it covers values outside the domain of the process or if the next observation is not covered. The distance between the lines is chosen by considering the conditional distribution implied by the discretized, estimated model. The natural choice of boundary conditions is to fix the solution of the discretized Fokker-Planck equation to be zero at the boundaries. Finally, the grid is constructed by dividing $\mathbb{S}$ in $Hstep$ intervals for each approximate standard deviation, i.e. $h = \sigma(s, x_s; \hat{\theta}_{GMM})\sqrt{\Delta}/Hstep$, where Δ is the sampling distance.

2.2.2 *Improvement using path integration*

It is known from e.g. Smith (1986), and Poulsen (1999) that Crank-Nicholson methods tend to behave badly for non-smooth initial data, especially when the drift is large (and non-linear).

The reason is that the initial probability density is given by a Dirac distribution $p(s, x; s, x_s) = \delta(x - x_s)$. This initial distribution cannot be used for numerical solutions as the finite difference approach is based on smooth and bounded functions. The numerical approximation of a Dirac distribution is all but smooth, resulting in oscillations in the numerical solution.

However, during a short time interval δ, the transition probability density is close to Gaussian and is therefore approximated as such, cf. Kloeden and Platen (1992). The initial density used for solving the PDE is therefore not a Dirac function but Gaussian, according to

$$\begin{aligned} X_{t+\delta} &= X_t + \mu(t, X_t)\delta + \sigma(t, X_t)\delta W, \\ &\in N(X_t + \mu(t, X_t)\delta, \sigma^2(t, X_t)\delta), \end{aligned} \tag{15}$$

where δ is small both compared to the distance between the observations and the dynamics of the model. This initial condition was used in Poulsen (1999), and can be interpreted as a first step in a path integration solution of the transition probability density, see e.g. Skaug (2000).

It is possible to further improve the approximation of the initial condition by considering the Milstein approximation of the stochastic differential equation. Let $\sigma_x(t, x_t) = \partial\sigma(t, X_t)/\partial x$. The Milstein approximation is then given by

$$X_{t+\delta} = X_t + \mu(t, X_t)\delta + \sigma(t, X_t)\delta W + \frac{1}{2}\sigma(t, X_t)\sigma_x(t, X_t)(\delta W^2 - \delta). \tag{16}$$

A closed form expression for the transition probability density of the Milstein approximation was derived in Elerian (1998)

$$f(x_{t+\delta} \mid x_t) = \frac{\exp(-\lambda/2)}{|A|\sqrt{2\pi}} z_{t+\delta}^{-1/2} \exp\left(-\frac{z_{t+\delta}}{2}\right) \cosh\left(\sqrt{\lambda z_{t+\delta}}\right), \tag{17}$$

where

$$z_{t+\delta} = \frac{x_{t+\delta} - B}{A},$$
$$\lambda = \frac{1}{\delta\,(\sigma_x(t, x_t))^2},$$
$$A = \frac{\sigma(t, x_t)\sigma_x(t, x_t)\delta}{2},$$
$$B = -\frac{\sigma(t, x_t)}{2\sigma_x(t, x_t)} + x_t + \mu(t, x_t)\delta - A.$$

The density is defined for $z_{t+\delta} \in \mathbb{R}^+$ and $\sigma_x(t, x_t) \neq 0$ ($\sigma_x(t, x_t) = 0$ would reduce the Milstein scheme to the Euler scheme). The Milstein initial condition is theoretically superior to the Euler initial condition as the rate of convergence is higher, see e.g. Kloeden and Platen (1992).

2.2.3 Convergence

It can be shown that the approximate likelihood converges uniformly to the true likelihood and the speed of convergence can be derived.

Lemma 1. *Assume that a $O(h^{2k})$, $k \geq 1$ central discretization is used to discretize to state derivatives in the Fokker-Planck equation. The approximate solution can then be written as*

$$p(h, x) = p_{True}(x) + c_{2k}(x)h^{2k} + O(h^{2k+2}). \tag{18}$$

The discretization transforms the Fokker-Planck equation into a linear system of ordinary differential equations, which has a closed form solution. The only source of systematic errors is thus the finite difference discretization. Using a Padé(1,1) or Padé(2,2) approximation of the matrix exponential introduces errors in the time integration as well, but the principal error term of the matrix exponential approximations can be chosen to balance (smaller or equal to) the error of the state space error, and is thus included in the error term.

Theorem 1. *Assume that the transition probability density is discretized using a $O(h^{2k})$ discretization and that the true transition probability density is bounded. The approximate solution then converges uniformly to the true likelihood as $h \to 0$ and the error of the likelihood is of order $O(h^{2k})$.*

Proof: We can write the conditional transition probability density (suppressing arguments) as

$$p(h, x) = p_{True}(x) + O(h^{2k}). \tag{19}$$

Using this, we can write the approximated likelihood as

$$L_{AML}(\theta; x, h) = p(h, x_1) \prod_{i=2}^{n} p(h, x_i \mid x_{i-1})$$
$$= p(h, x_1) \prod_{i=2}^{n} (p_{True}(x_i \mid x_{i-1}) + O(h^{2k})). \tag{20}$$

Rewriting (where $p_{True,-i}(\mathbf{x}) = p_{True}^{-1}(x_i \mid x_{i-1}) \prod_{j=2}^{n} p_{True}(x_j \mid x_{j-1})$)

$$= (p_{True}(x_1) + O(h^{2k})) \left[\prod_{i=2}^{n} p_{True}(x_i \mid x_{i-1}) + O(h^{2k}) \left(\sum_{i=2}^{n} p_{True,-i}(\mathbf{x}) + O(h^{2k}) \right) \right] \tag{21}$$

which equals

$$L_{AML}(\theta; x, h) = p_{True}(x_1) \prod_{i=2}^{n} p_{True}(x_i \mid x_{i-1}) + O(h^{2k}) g(x)$$
$$= L(\theta; x) + O(h^{2k}) g(x), \tag{22}$$

where $g(x) = \prod_{i=2}^{n} p_{True}(x_i \mid x_{i-1}) + p_{True}(x_1)(\sum_{i=2}^{n} p_{True,-i}(\mathbf{x}) + O(h^{2k})$.

However, the function $g(x)$ is bounded as the transition probability density is bounded and finite multiplications of bounded terms also are bounded. Hence, $L_{AML}(\theta; x)$ converges uniformly to $L(\theta; x)$ as $h \to 0$. □

A theoretical study of approximate Maximum Likelihood estimators for diffusion processes can be found in Pedersen (1995a). That paper derives conditions for the approximate likelihood function to converge to the true likelihood function as well as alternative conditions to ensure consistency and asymptotic normality of approximate Maximum Likelihood estimators. It is shown that convergence of the estimates is obtained if the approximate likelihood function and the true likelihood function are continuous for all $\theta \in \Theta$ and if the approximate likelihood function converges uniformly in probability to the true likelihood function. Given these conditions, the estimate generated by the approximate Maximum Likelihood estimator will converge in probability to the estimate generated by the true Maximum Likelihood estimator.

The estimator obtained using our methodology converges uniformly to the true Maximum Likelihood estimator. Furthermore, we have obtained a description of how fast the approximate likelihood converges.

2.2.4 Poulsen estimator

The Fokker-Planck-based estimator introduced in Poulsen (1999) is a Crank-Nicholson scheme. Our construction of the grid resembles the construction in Poulsen (1999), although Poulsen (1999) uses an Euler discretization to obtain a smooth starting condition. Richardson extrapolation was applied in Poulsen (1999) but the results presented in Jensen and Poulsen (2002) suggest that the scheme using Richardson extrapolation is less robust and not more accurate than the ordinary Crank-Nicholson scheme. The Poulsen estimator is used

in Christensen, Poulsen, and Sørensen (2001) where numerical experiments showed that it outperformed all non-Maximum Likelihood estimators.

We will implement the Poulsen estimator by using the Euler discretization of the starting condition and the $O(h^2)$-Padé(1,1) discretization of the Fokker-Planck equation, but not applying Richardson extrapolation.

2.3 Criteria of estimator performance

The most important condition in Pedersen (1995a) is the uniform convergence of the approximate likelihood to the true likelihood for all values of $\theta \in \Theta$. This condition is impossible to test numerically, but it is possible to test if the approximate likelihood converges for a fixed θ. A conservative approximation of the distance between the approximate and true likelihood function is given by

$$\left| \sum_{i=1}^{n} \log \hat{p}(t_i, x_{t_i}; t_{i-1}, x_{t_{i-1}}) - \log p(t_i, x_{t_i}; t_{i-1}, x_{t_{i-1}}) \right| \tag{23}$$

$$\leq \sum_{i=1}^{n} |\log \hat{p}(t_i, x_{t_i}; t_{i-1}, x_{t_{i-1}}) - \log p(t_i, x_{t_i}; t_{i-1}, x_{t_{i-1}})|. \tag{24}$$

By weighting the distance by $p(t_i, x_{t_i}; t_{i-1}, x_{t_{i-1}})$ and scaling by the number of observations, we derive the mean absolute error (MAE) of the log-likelihood function

$$\begin{aligned} MAE &= \frac{1}{N} \sum_{i=1}^{N} |\log \hat{p}(x_{i+1} \mid x_i) - \log p(x_{i+1} \mid x_i)| \\ &\approx \int |\log \hat{p}(x_{i+1} \mid x_i) - \log p(x_{i+1} \mid x_i)| \, p(x_i, x_{i-1}) dx_i dx_{i-1}. \end{aligned} \tag{25}$$

This criterion is similar to the criterion in Durham and Gallant (2002) where the convergence is measured as the root mean square error (RMSE) of the log-likelihood function

$$RMSE = \left(\frac{1}{N} \sum_{i=1}^{N} (\log \hat{p}(x_{i+1} \mid x_i) - \log p(x_{i+1} \mid x_i))^2 \right)^{1/2}. \tag{26}$$

However, the RMSE is less robust and is slightly more restrictive than the MAE. Also, one of the benchmarks in Jensen and Poulsen (2002) can be obtained using a Taylor expansion of the MAE (assuming that $\log \hat{p}(x_{i+1} \mid x_i) \approx \log p(x_{i+1} \mid x_i)$) as

$$\begin{aligned} MAE &= \frac{1}{N} \sum_{i=1}^{N} |\log \hat{p}(x_{i+1} \mid x_i) - \log p(x_{i+1} \mid x_i)| \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\hat{p}(x_{i+1} \mid x_i) - p(x_{i+1} \mid x_i)}{p(x_{i+1} \mid x_i)} \right|. \end{aligned} \tag{27}$$

The MAE was evaluated on a data series of $N = 100\,000$ observations, and gives an upper bound of the errors in the log-likelihood function. However, the MAE of the log-likelihood

can only be calculated if the true likelihood is known, and does not tell us whether the estimates are sufficiently close to the true parameter values or not. Following Durham and Gallant (2002), we also introduce the root mean squared error of the parameter estimates

$$RMSE_{True-MLE} = \left\{ \frac{1}{J} \sum_{j=1}^{J} \left(\theta_0^{(j)} - \hat{\theta}_{MLE}^{(j)} \right)^2 \right\}^{1/2}, \tag{28}$$

$$RMSE_{MLE-AMLE} = \left\{ \frac{1}{J} \sum_{j=1}^{J} \left(\hat{\theta}_{MLE}^{(j)} - \hat{\theta}_{AMLE}^{(j)} \right)^2 \right\}^{1/2}. \tag{29}$$

The parameters are estimated on $J = 500$ data sets of length $T = 1000$ observations for three different parameter vectors.

This criterion cannot be used to compare the estimators applied to the non-linear CKLS model, as the Maximum Likelihood estimate is unknown. Two other measures are used

$$RMSE_{True-AMLE} = \left\{ \frac{1}{J} \sum_{j=1}^{J} \left(\theta_0^{(j)} - \hat{\theta}_{AMLE}^{(j)} \right)^2 \right\}^{1/2}, \tag{30}$$

$$RMSE_{AMLE_1-AMLE_2} = \left\{ \frac{1}{J} \sum_{j=1}^{J} \left(\hat{\theta}_{AMLE_1}^{(j)} - \hat{\theta}_{AMLE_2}^{(j)} \right)^2 \right\}^{1/2}. \tag{31}$$

We can compare the *RMSE* of the different approximate Maximum Likelihood estimators, but the estimators are expected to generate similar estimates (and also similar to the Maximum Likelihood estimate). This makes it difficult to draw any firm conclusions from $RMSE_{True-AMLE}$.

We therefore test if they generate similar estimates. This would indicate that they are able to work as a proxy for the true Maximum Likelihood estimator, especially if the $RMSE_{AMLE_1-AMLE_2}$ is significantly smaller than the $RMSE_{True-AMLE}$. The $RMSE_{AMLE_1-AMLE_2}$ should be small as

$$\|\hat{\theta}_{AMLE_1} - \hat{\theta}_{AMLE_2}\| = \left\| \left(\hat{\theta}_{AMLE_1} - \hat{\theta}_{MLE} \right) - \left(\hat{\theta}_{AMLE_2} - \hat{\theta}_{MLE} \right) \right\| \tag{32}$$

which can be bounded by the triangle inequality as

$$\leq \|\hat{\theta}_{AMLE_1} - \hat{\theta}_{MLE}\| + \|\hat{\theta}_{AMLE_2} - \hat{\theta}_{MLE}\|. \tag{33}$$

However, these terms $\to 0$ as the estimates converges to the true Maximum Likelihood estimates in probability.

We have used the BHHH algorithm for maximizing the log-likelihood function, see Berndt et al. (1974). This is a Quasi-Newton optimization algorithm using the sample covariance matrix of the score functions to approximate the Hessian of the log-likelihood function. The advantage of this algorithm is that the covariance matrix is always positive definite, thus implying that the likelihood increases for each iteration.

3 Monte Carlo studies

The estimators were evaluated on simulated data, using the criteria defined in the previous section. All simulations were run using Matlab 6.5 on a 1.8 GHz personal computer. Matlab is an interpreting language, which can be significantly slower than a compiled language, especially if the code cannot be vectorized properly. The results should therefore be interpreted with caution, especially when comparing the results to e.g. Durham and Gallant (2002). Still the complexity of the algorithms does not depend on the type of language.

3.1 Cox-Ingersoll-Ross

The Cox-Ingersoll-Ross model is used mainly because of the closed form expression of the transition probability density, making it a suitable benchmark model. We have used a time series of $N = 100\,000$ observations to evaluate the first benchmark. The parameters used were specified as $\alpha = 0.5$, $\beta = 0.06$ and $\sigma = 0.15$ and $\Delta = 1/12$. These parameters are close to calibrated parameters for the monthly U.S. treasury bill rate, see e.g. Durham and Gallant (2002) and Aït-Sahalia (2002). The series was simulated using exact simulation starting from the stationary distribution.

The MAE of the log-likelihood for the simulated data series is presented in Fig. 1. The Durham-Gallant estimator is the simulation-based estimator presented in Durham and Gallant (2002), the Poulsen estimator is a Crank-Nicholson-based estimator using an Euler approximation as initial condition while the Order2-Padé(1,1) estimator is a Crank-Nicholson scheme using a Milstein approximation as initial condition. Finally, the Order4-Padé(2,2) is a $O(h^4)$ finite difference scheme using a Milstein approximation as initial condition. It can be seen that the higher order finite difference approximation is significantly better at

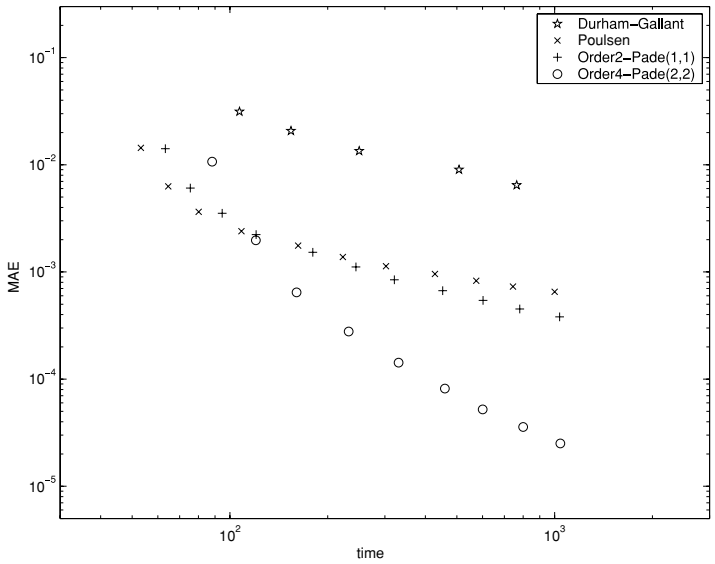

Fig. 1 MAE of the log-likelihood versus time for simulated data from the Cox-Ingersoll-Ross model. The rate of convergence of the Fokker-Planck-based approximations is faster than the convergence rate for the simulation-based approximation

Table 1 Mean absolute error of the log-likelihood, calculated using parameters similar to parameters estimated using the US 3-month Treasury bill

Approximation errors of the log-likelihood								
Estimator	M	K	Hstep	Nstep	Order	Padé	MAE	time
Euler							0.0780	4.7
Milstein							0.0207	11.7
Durham & Gallant	4	4					0.031356	107.0172
Durham & Gallant	8	8					0.020737	154.1856
Durham & Gallant	16	16					0.013459	250.1704
Durham & Gallant	32	32					0.009002	507.8327
Durham & Gallant	32	64					0.0064544	764.3522
Poulsen			2	2	2	(1,1)	0.014391	53.1654
Poulsen			3	3	2	(1,1)	0.006296	64.5757
Poulsen			4	4	2	(1,1)	0.0036436	80.1255
Poulsen			5	5	2	(1,1)	0.0024	108.5382
Poulsen			6	6	2	(1,1)	0.0017632	162.2198
Poulsen			7	7	2	(1,1)	0.00138	222.8214
Poulsen			8	8	2	(1,1)	0.0011307	302.3388
Poulsen			9	9	2	(1,1)	0.00095702	427.9107
Poulsen			10	10	2	(1,1)	0.00082857	573.8564
Poulsen			11	11	2	(1,1)	0.00073033	743.9218
Poulsen			12	12	2	(1,1)	0.00065297	1000.4605
Milstein-Order2-Padé(1,1)			2	2	2	(1,1)	0.014131	63.2776
Milstein-Order2-Padé(1,1)			3	3	2	(1,1)	0.0060665	75.54
Milstein-Order2-Padé(1,1)			4	4	2	(1,1)	0.0035273	94.6593
Milstein-Order2-Padé(1,1)			5	5	2	(1,1)	0.0022319	120.3769
Milstein-Order2-Padé(1,1)			6	6	2	(1,1)	0.001527	180.085
Milstein-Order2-Padé(1,1)			7	7	2	(1,1)	0.0011131	244.4255
Milstein-Order2-Padé(1,1)			8	8	2	(1,1)	0.00084503	320.9568
Milstein-Order2-Padé(1,1)			9	9	2	(1,1)	0.00066839	452.2627
Milstein-Order2-Padé(1,1)			10	10	2	(1,1)	0.00054272	602.0248
Milstein-Order2-Padé(1,1)			11	11	2	(1,1)	0.00045083	780.6822
Milstein-Order2-Padé(1,1)			12	12	2	(1,1)	0.00038096	1036.4098
Milstein-Order4-Padé(2,2)			2	2	4	(2,2)	0.010692	88.1356
Milstein-Order4-Padé(2,2)			3	3	4	(2,2)	0.0019766	120.1704
Milstein-Order4-Padé(2,2)			4	4	4	(2,2)	0.00064452	160.478
Milstein-Order4-Padé(2,2)			5	5	4	(2,2)	0.00027783	232.1686
Milstein-Order4-Padé(2,2)			6	6	4	(2,2)	0.00014225	330.6465
Milstein-Order4-Padé(2,2)			7	7	4	(2,2)	8.1583e-05	459.3695
Milstein-Order4-Padé(2,2)			8	8	4	(2,2)	5.2168e-05	600.3513
Milstein-Order4-Padé(2,2)			9	9	4	(2,2)	3.5735e-05	799.7702
Milstein-Order4-Padé(2,2)			10	10	4	(2,2)	2.5068e-05	1041.5521

approximating the log-likelihood and that the rate of convergence is faster than lower order finite difference approximations, which in turn are better at approximating the log-likelihood than the simulation-based estimators.

The results are also presented in Table 1. Note that the Milstein initial condition is superior to the Euler initial condition as the overall accuracy improves, while not much is gained when the grid is coarse.

Table 2 The three different sets of parameters used to evaluate the algorithms

Parameters for the Cox-Ingersoll-Ross models				
Model	α	β	σ	Δ
A	0.5	0.06	0.15	1/12
B	5.0	0.06	0.15	1/12
C	0.5	0.06	0.22	1/12

A consequence of the higher rate of convergence obtained, using the higher order finite difference schemes is that the "curse of dimensionality" of deterministic numerical techniques can be reduced. This indicates that the higher order finite difference technique might be competitive for (low dimensional) multivariate diffusions as well, although additional research is needed to verify this indication.

The numerical techniques were also evaluated in terms of parameter estimates using simulated data from different Cox-Ingersoll-Ross models. The parameters and sampling distance between the observations for the models are given in Table 2.

Model A is using the same parameters as the previous test. The second model (B) is designed to test the numerical properties of estimators when the drift term is large, as large drift terms are usually troublesome for discrete time approximations. Finally, model (C) is designed to test the estimators when $2\alpha\beta/\sigma^2$ is close to unity. This parameter setting will stress the numerical approximations as the transition probability density is almost singular at the origin. The estimators were evaluated using 500 independent time series, each series consisting of 1000 observations for the three different models.

The performance of the estimators is evaluated by comparing an Euler approximation, a Milstein approximation, the Durham-Gallant simulation-based estimator using $M = 16$ and $K = 16$, the Poulsen estimator and finally a finite difference $O(h^4)$, Padé(2,2) scheme using the Milstein initial condition. The Crank-Nicholson-based estimator used $Nstep = 5$ and $Hstep = 5$, while the higher order finite difference scheme used $Nstep = 4$ and $Hstep = 4$.

The estimates from model A are summarized in Table 3. The Durham-Gallant estimator is a magnitude better than the discretized approximation. The Fokker-Planck-based estimators are performing at least as well as the Durham-Gallant estimator and the higher order finite difference scheme is almost a magnitude better than the Durham-Gallant estimator, as expected from Fig. 1. Of importance is that the error generated by any of the advanced numerical techniques is much smaller than the sample variation.

Similar results are obtained for the second model (B), see Table 4. The Euler and Milstein estimates are biased while the simulation-based and Fokker-Planck-based estimators are far better off, having an error much smaller than the sample variation. The higher order finite difference scheme is performing well once again.

Model C is more complicated to estimate for the Fokker-Planck-based estimators. The problem is caused by the fact that the process is approaching the origin, which causes some difficulties as the finite difference approximation "needs a few grid lines" on each side of the observation to reduce errors caused by the boundaries. Similarly, the trajectories generated in simulation-based technique might choose points outside of the domain of the true process, although it did not happen frequently in our simulation study. The difficulty encountered for the Fokker-Planck-based estimators was approached by transforming the process using a logarithmic transformation $Y = \log X$ and solving the Fokker-Planck equation for Y instead of X. This transformation was also used in Elerian, Chib, and Shephard (2001)

Table 3 Data A. Parameter estimates of the Cox-Ingersoll-Ross process using 500 series of 1000 observations each

Approximation errors for parameter estimates				
		α	β	σ
True-MLE	Mean	0.0550	−0.0002	0.0003
True-MLE	Std dev.	0.1207	0.0082	0.0035
True-MLE	RMSE	0.1327	0.0082	0.0035
MLE-Euler	Mean	0.013772	−1.2068e-06	0.0024639
MLE-Euler	Std dev.	0.023918	6.6018e-05	0.0010001
MLE-Euler	RMSE	0.027599	6.6029e-05	0.0026591
MLE-Milstein	Mean	0.019924	−7.9294e-05	0.0028864
MLE-Milstein	Std dev.	0.027828	0.0005338	0.0008221
MLE-Milstein	RMSE	0.034225	0.0005397	0.0030012
MLE-Durham&Gallant	Mean	−0.0005841	0.0001058	0.0002172
MLE-Durham&Gallant	Std dev.	0.0042021	3.3036e-05	7.6156e-05
MLE-Durham&Gallant	RMSE	0.0042425	0.0001109	0.0002301
MLE-Poulsen	Mean	−0.0009426	7.181e-08	−0.0001226
MLE-Poulsen	Std dev.	0.0014637	2.7703e-05	0.0001606
MLE-Poulsen	RMSE	0.001741	2.7703e-05	0.0002020
MLE-Milstein-o4p4	Mean	−4.2562e-05	−7.872e-08	2.4287e-05
MLE-Milstein-o4p4	Std dev.	0.00097889	3.1255e-06	1.7733e-05
MLE-Milstein-o4p4	RMSE	0.00097981	3.1265e-06	3.0072e-05

Table 4 Data B. Parameter estimates of the Cox-Ingersoll-Ross process using 500 series of 1000 observations each

Approximation errors for parameter estimates				
		α	β	σ
True-MLE	Mean	0.0326	−0.0001	−0.0003
True-MLE	Std dev.	0.4322	0.0008	0.0040
True-MLE	RMSE	0.4334	0.0008	0.0040
MLE-Euler	Mean	0.92758	−1.0475e-07	0.026055
MLE-Euler	Std dev.	0.15171	8.2478e-07	0.0024026
MLE-Euler	RMSE	0.93991	8.314e-07	0.026165
MLE-Milstein	Mean	0.96832	−3.2509e-06	0.026341
MLE-Milstein	Std dev.	0.1554	1.1482e-05	0.0023772
MLE-Milstein	RMSE	0.98071	1.1933e-05	0.026448
MLE-Durham&Gallant	Mean	0.058365	−1.0361e-07	0.0019015
MLE-Durham&Gallant	Std dev.	0.015325	3.2651e-06	0.0002069
MLE-Durham&Gallant	RMSE	0.060343	3.2668e-06	0.0019127
MLE-Poulsen	Mean	0.014664	6.0541e-08	−0.0008903
MLE-Poulsen	Std dev.	0.014292	1.0316e-05	0.00016686
MLE-Poulsen	RMSE	0.020476	1.0317e-05	0.00090581
MLE-Milstein-o4p4	Mean	0.011159	−6.0457e-09	0.00027036
MLE-Milstein-o4p4	Std dev.	0.0025044	1.3372e-06	2.8816e-05
MLE-Milstein-o4p4	RMSE	0.011436	1.3372e-06	0.00027189

Table 5 Data C. Parameter estimates of the Cox-Ingersoll-Ross process using 500 series of 1000 observations each

Approximation errors for parameter estimates				
		α	β	σ
True-MLE	Mean	0.0462	−0.0005	−0.0000
True-MLE	Std dev.	0.1279	0.0113	0.0053
True-MLE	RMSE	0.1360	0.0113	0.0053
MLE-Euler	Mean	0.015674	2.7577e-06	−0.0021681
MLE-Euler	Std dev.	0.10593	0.0003723	0.004705
MLE-Euler	RMSE	0.10709	0.00037231	0.0051805
MLE-Milstein	Mean	−0.079539	−0.0010445	0.0050127
MLE-Milstein	Std dev.	0.097955	0.0030055	0.0024278
MLE-Milstein	RMSE	0.12618	0.0031819	0.0055697
MLE-Durham&Gallant	Mean	0.0062517	−0.0023201	0.0003192
MLE-Durham&Gallant	Std dev.	0.039425	0.0033177	0.0019566
MLE-Durham&Gallant	RMSE	0.039918	0.0040485	0.0019824
MLE-Poulsen	Mean	0.0091496	0.0001223	0.0007945
MLE-Poulsen	Std dev.	0.018645	0.0001144	0.0008521
MLE-Poulsen	RMSE	0.020769	0.0001675	0.001165
MLE-Milstein-o4p4	Mean	−0.008879	−0.0002351	0.000561
MLE-Milstein-o4p4	Std dev.	0.015284	0.0012609	0.0005237
MLE-Milstein-o4p4	RMSE	0.017676	0.0012826	0.0007675

to transform the state space into $\mathbb{R}$. The calculated density $p_Y(t, y_t; s, y_s)$ was transformed back to $p_X(t, x_t; s, x_s)$, which was used for forming the estimator. The results are presented in Table 5.

The Fokker-Planck-based estimators are once again competitive although the difference between the Poulsen and the higher order finite difference scheme is relatively small, expect for β where the Poulsen estimator outperforms all other estimators.

3.2 Non-linear CKLS

A substantial generalization of the CKLS model was introduced in Aït-Sahalia (1996)

$$dX_t = \left(a_0 + a_1 X_t + a_2 X_t^2 + \frac{a_3}{X_t}\right) dt + \sqrt{\theta_1 + \theta_2 X_t + \theta_3 X_t^{\theta_4}} dW_t. \tag{34}$$

Later studies, see e.g. Durham (2004) has found this model too general, but it is a minor problem when using simulated data. 100 series has been simulated, each series consisting of 3000 observations using the Milstein scheme. The bias in the simulation was reduced by taking 100 intermediate steps between the observations and discarding the first 1000 observations as burn-in. The sampling distance was $\Delta = 1/52$ implying that each series is equivalent to approximately 40 years of weekly data. This set-up is similar to Durham (2004), and we have used the parameters estimated for *GEN4* in Durham (2004) applied to weekly observations of the 3-month US Treasury bill rate to generate the data, although a different

Table 6 RMSE for the parameter estimates for the non-linear CKLS model. The RMSE between the advanced approximations is small compared to $RMSE_{True-AMLE}$ and the RMSE of the Euler estimator

Root Mean Squared Errors				
	α_0	α_1	α_2	α_3
True-$O(h^4)$	0.7907	0.5109	0.1084	15.9805
True-Poulsen	0.7907	0.5109	0.1080	15.9805
True-Durham & Gallant	0.7911	0.5105	0.1104	15.9804
True-Euler	1.3648	66.1792	2.7729	114.7802
$O(h^4)$-Poulsen	0.0013	0.0012	0.0038	0.0001
$O(h^4)$-Durham & Gallant	0.0021	0.0029	0.0049	0.0005
Poulsen-Durham & Gallant	0.0020	0.0026	0.0043	0.0005
$O(h^4)$-Euler	1.3760	66.1538	2.7627	111.7918
	θ_1	θ_2	θ_3	θ_4
True-$O(h^4)$	0.2638	0.0950	0.0031	0.4634
True-Poulsen	0.2637	0.0950	0.0031	0.4643
True-Durham & Gallant	0.2565	0.0883	0.0014	0.4543
True-Euler	0.2593	0.0898	0.0014	0.4686
$O(h^4)$-Poulsen	0.0069	0.0033	0.0001	0.0243
$O(h^4)$-Durham & Gallant	0.0371	0.0194	0.0020	0.0770
Poulsen-Durham & Gallant	0.0361	0.0191	0.0020	0.0739
$O(h^4)$-Euler	0.0420	0.0219	0.0020	0.1000

parameterization was used

$$dX_t = \alpha_0 \left(\alpha_1 - X_t + \alpha_2 X_t^2 + \frac{\alpha_3}{X_t} \right) dt + \sqrt{\theta_1 + \theta_2 X_t + \theta_3 X_t^{\theta_4}} dW_t. \tag{35}$$

We used an Euler approximation, the Durham & Gallant estimator, the Poulsen estimator and the $O(h^4)$, Padé(2,2) finite difference estimator to estimate the parameters using the same options for the numerical estimators as before. The results can be found in Table 6.

All advanced approximate Maximum Likelihood estimators obtain good estimates, having lower RMSE than the Euler estimates. The approximate Maximum Likelihood estimates are also similar indicating that any of the advanced methods can be used to obtain estimates similar to the true Maximum Likelihood estimates.

4 Conclusions

An Approximate Maximum Likelihood method based on numerical solution of the Fokker-Planck equation using higher order finite differences and Padé approximations have been proposed.

It was shown that the approximate likelihood function converges to the true likelihood function as the discretization approaches the continuous model implying that the estimates converge in probability to the true Maximum Likelihood estimates.

The Fokker-Planck-based estimators have also nice practical properties and are competitive compared to the Poulsen and Durham & Gallant estimators. An interesting extension

is the possible use of the higher order finite difference schemes for multivariate models as it might be able to beat the "curse of dimensionality."

Acknowledgments We thank the anonymous referees and the editor for helpful comments.

References

Aït-Sahalia, Y. (1996). "Testing Continuous-Time Models of the Spot Interest Rate." *The Review of Financial Studies*, 9(2), 385–426.

Aït-Sahalia, Y. (1999). "Transition Densities for Interest Rate and Other Nonlinear Diffusions." *Journal of Finance*, LIV(4), 1361–1395.

Aït-Sahalia, Y. (2002). "Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed–Form Approximation Approach." *Econometrica*, 70, 223–262.

Berndt, E., B. Hall, R. Hall, and J. Hausman. (1974). "Estimation and Inference in Nonlinear Structural Models." *Annals of Economic and Social Measurement*, 3, 653–665.

Bibby, B.M. and M. Sørensen. (1995). "Martingale Estimation Functions for Discretely Observed Diffusion Processes." *Bernoulli*, 1, 17–39.

Chan, K., G. Karolyi, F. Longstaff, and A. Sanders. (1992). "An Empirical Comparison of Alternative Models of the Short Term Interest Rate." *Journal of Finance*, XLVII, 1209–1227.

Christensen, B.J., R. Poulsen, and M. Sørensen. (2001). "Optimal Inference in Diffusion Models of the Short Rate of Interest." Working paper 102, Centre for Analytical Finance, University of Aarhus, Aarhus, Denmark.

Cox, J.C., J.E. Ingersoll, and S.A. Ross. (1985). "A Theory of the Term Structure of Interest Rates." *Econometrica*, 53, 385–407.

Duffie, D. and P. Glynn. (2004). "Estimation of Continuous-Time Markov Processes Sampled at Random Time Intervals." *Econometrica*, 72(6), 1773–1808.

Durham, G.B. (2004). "Likelihood-Based Specification Analysis of Continuous-Time Models of the Short-Term Interest Rate." *Journal of Financial Economics*, 70(3).

Durham, G.B. and A.R. Gallant. (2002). "Numerical Techniques for Maximum Likelihood Estimation of Continuous Time Processes." *Journal of Business and Economic Statistics*, 20(3), 297–316.

Elerian, O. (1998). "A Note on the Existence of a Closed form Conditional Transition Density for the Milstein Scheme." Technical report, Nuffield College, Oxford University.

Elerian, O., S. Chib, and N. Shephard. (2001). "Likelihood Inference for Discretely Observed Non-Linear Diffusions." *Econometrica*, 69, 959–993.

Eraker, B. (2001). "MCMC Analysis of Diffusion Models With Applications to Finance." *Journal of Business and Economic Statistics*, 19(2), 177–191.

Feller, W. (1951). "Two Singular Diffusion Problems." *Annals of Mathematics*, 54, 173–182.

Gallant, A.R. and G.E. Tauchen. (1996). "Which Moments to Match?" *Econometic Theory*, 12(4), 657–681.

Gourieroux, C., A. Monfort, and E. Renault. (1993). "Indirect Inference." *Journal of Applied Econometrics*, 8, 85–118.

Hansen, L.P. (1982). "Large Sample Properties of Generalized Method of Moment Estimators." *Econometrica*, 50(4), 1029–1054.

Hansen, L.P. and J.A. Scheinkman. (1995). "Back to the Future: Generating Moment Implications for Continuous Time Markov Processes." *Econometrica*, 63(4), 767–804.

Jensen, B. and R. Poulsen. (2002). "Transition Densities of Diffusion Processes: Numerical Comparison of Approximation Techniques." *Journal of Derivatives*, 9(4), 18–32.

Karatzas, I. and S.E. Shreve. (1999). *Brownian Motion and Stochastic Calculus*, 2nd edn. Vol 113, Graduate texts in mathematics. New York: Springer-Verlag.

Kessler, M. and M. Sørensen. (1999). "Estimating Equations Based on Eigenfunctions for a Discretely Observed Diffusion Process." *Bernoulli*, 5, 299–314.

Kloeden, P.E. and E. Platen. (1992). *Numerical Solution of Stochastic Differential Equations*. Berlin: Springer-Verlag.

Lindström, E. (2003). *Estimation and Model Validation of Diffusion Processes*. Licentiate thesis LUTFMS-2007-2003, Mathematical Statistics, Centre for Mathematical Sciences, Sweden: Lund University.

Lo, A.W. (1988). "Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data." *Econometric Theory*, 4(2), 231–247.

Nielsen, J.N., H. Madsen, and P. C. Young. (2000). Parameter Estimation in Stochastic Diffential Equations: An overview." *IFAC Annual Reviews in Control*, 24(1), 83–94.

Øksendal, B. (2000). *Stochastic Differential Equations: An Introduction with Applications*, 5th edn. Berlin: Springer-Verlag.

Overbeck, L. and T. Rydén. (1997). "Estimation in the Cox-Ingersoll-Ross model." *Econometric Theory*, 13, 430–461.

Pedersen, A.R. (1995a). "Consistency and Asymptotic Normality of an Approximative Maximum Likelihood Estimator for Discretely Observed Diffusion Processes." *Bernoulli*, 1(3), 257–279.

Pedersen, A.R. (1995b). "A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations." *Scandinavian Journal of Statistics*, 22, 55–71.

Poulsen, R. (1999). "Approximate Maximum Likelihood Estimation of Discretely Observed Diffusion Processes." Working paper 29, Centre for Analytical Finance, University of Aarhus, Aarhus, Denmark.

Skaug, C. (2000). *Random Vibration and the Path Integral Method*. PhD Thesis, Department of Structural Engineering, Norwegian University of Science and Technology, Trondheim, Norway.

Smith, G.D. (1986). *Numerical Solution of Partial Differential Equations: Finite Difference Methods*, 3rd edn. Oxford University Press.

Sørensen, H. (2004). "Parametric Inference for Diffusion Processes Observed at Discrete Points in Time: A Survey." *International Statistical Review*, 72(3), 337–354.