DISCUSSION PAPERS
Institute of Economics
University of Copenhagen

02-08

Paramedic Inference for Diffusion Processes Observed at Discrete Points in Time

Helle Sørensen

Studiestræde 6, DK-1455 Copenhagen K., Denmark
Tel. +45 35 32 30 82 - Fax +45 35 32 30 00
http://www.econ.ku.dk

Parametric inference for diffusion processes observed at discrete points in time: a survey

Helle Sørensen

Department of Economics, University of Copenhagen, Studiestræde 6, DK-1455 Copenhagen K, Denmark. E-mail: helle@econ.ku.dk

Summary

This paper is a survey of existing estimation techniques for stationary and ergodic diffusion processes observed at discrete points in time. The reader is introduced to the following techniques: (i) estimating functions with special emphasis on martingale estimating functions and so-called simple estimating functions; (ii) analytical and numerical approximations of the likelihood which can in principle be made arbitrarily accurate; (iii) Bayesian analysis and MCMC methods; and (iv) indirect inference and EMM which both introduce auxiliary (but wrong) models and correct for the implied bias by simulation.

Key words: Bayesian analysis; diffusion processes; discrete-time observations; Efficient method of moments (EMM); Estimating functions; Indirect inference; Likelihood approximations.

This version: August 8, 2002

1 Introduction

Statistical inference for diffusion processes has been an active research area during the last two or three decades. The work has developed from estimation of linear systems from continuous-time observations (see Breton Le (1974) and the references therein) to estimation of non-linear systems, parametric or non-parametric, from discrete-time observations. This paper is about *parametric inference for (truly) discrete-time observations* exclusively; the models may be linear or non-linear.

Now, why is this an interesting and challenging topic at all? Well, diffusion models have a large range of applications. They have been used for a long time to model phenomena evolving randomly and continuously in time, *e.g.* in physics and biology. During the last thirty years or so the models have also been applied intensively in mathematical finance for describing stock prices, exchange rates, interest rates, *etc.* (although it is well-known that such quantities do not really change continuously in time). At the same time data are essentially always recorded at discrete points in time only (*e.g.* weekly, daily or each minute), no matter the application.

In other words, only discrete-time observations of the continuous-time system is available, and this is exactly what makes the problem challenging. For a few models, estimation is straightforward because the corresponding stochastic differential equation can be solved explicitly. This is the case for (i) the geometric Brownian motion: $dX_t = \alpha X_t\,dt + \sigma X_t\,dW_t$; ($ii$) the Ornstein-Uhlenbeck process: $dX_t = \alpha(\beta - X_t)\,dt + \sigma\,dW_t$; and ($iii$) the square-root/Cox-Ingersoll-Ross process $dX_t = \alpha(\beta - X_t)\,dt + \sigma\sqrt{X_t}\,dW_t$, which have log-normal, normal and non-central chi-square transition probabilities respectively. However, "nature" (or "the market") most often generates data not adequately described by such simple models. For example, empirical studies clearly reveal that increments of logarithmic stock prices are not independent and Gaussian as implied by the geometric Brownian motion classically used for stock price modeling. Accordingly, more complex models allowing for non-linear drift and diffusion functions are needed in order to obtain reasonable agreement with data. (Of course there are other possible extensions of the simple models, *e.g.* stochastic volatility models with unobserved coordinates, but in this survey we stick to pure diffusion models.) This complicates the statistical analysis considerably because the discrete-time transitions, implicitly defined by the model, are no longer known analytically. Specifically, *the likelihood function is usually not tractable*. In other words, one often has to use models for which likelihood analysis is not possible.

Consequently, there is a need for alternative methods. Research in that direction commenced in the mid eighties, with the paper by Dacunha-Castelle and Florens-Zmirou (1986) on the loss of information due to discretization as an important reference, and accelerated in the nineties. Important references from the mid of the decade are Bibby and Sørensen (1995) on martingale estimating functions, Gourieroux *et al.* (1993) on indirect inference, and Pedersen (1995b) on an approximate maximum likelihood method, among others. Later work includes Bayesian analysis (Elerian *et al.*, 2001; Roberts and Stramer, 2001) and further approximate likelihood methods (Aït-Sahalia, 2002b; Poulsen, 1999).

In the following the reader is introduced to the following techniques: (i) estimating functions with special emphasis on martingale estimating functions and so-called simple estimating functions; (ii) three approximations of the likelihood which can in principle be made arbitrarily accurate; (iii) Bayesian analysis and MCMC methods; and finally (iv) indirect inference and EMM which both introduce auxiliary (but wrong) models and correct for the implied bias by simulation. Focus is on fundamental ideas and the reader is referred to the literature for more rigorous treatments. In particular, we consider one-dimensional

diffusions only in order to keep notation simple, although most methods apply in the multi-dimensional case as well (at least in principle; see also comments along the way and in Section 8). Neither do we account for technical assumptions, regularity conditions *etc.*

As was already pointed out we will be concerned with parametric models and discrete-time observations. Specifically we will assume that the process is observed at equidistant time-points $i\Delta$, $i = 1, \ldots, n$. The asymptotic results that we quote, hold for Δ fixed and n tending to infinity. This asymptotic scheme is appropriate if, say, daily or weekly observations are available in a sampling period of increasing length.

Another branch of research has been concerned with estimation in the situation where data are collected up to some fixed point in time but more and more frequently (with the above notation: $\Delta \to 0$ and $T = n\Delta$ is fixed). In the limit, this of course amounts to observation in continuous time. We do not discuss this set-up in this paper, but a few comments are appropriate: Dohnal (1987) and Genon-Catalot and Jacod (1994), among others, have studied parameter estimation via contrasts, local asymptotic (mixed) normality properties, and optimal random sampling times. Several authors have studied non-parametric estimation for the sampling scheme as well: the estimators are based on kernel methods (Florens-Zmirou, 1993; Jacod, 2000) or wavelet methods (Genon-Catalot *et al.*, 1992; Hoffmann, 1999; Honoré, 1997); see also the references in those papers. Not much non-parametric work has been done in the set-up considered in this paper, see however the paper by Aït-Sahalia (1996).

The paper is organized as follows. The model is defined in Section 2, and Section 3 contains preliminary comments on the estimation problem. In Section 4 we discuss estimating functions; in Section 5 approximations of the likelihood. Section 6 is about Bayesian analysis, and Section 7 is about indirect inference and EMM. Finally, in Section 8, we summarize and comment briefly on possible extensions.

2 Model, assumptions and notation

In this section we present the model and the basic assumptions, and introduce notation that will be used throughout the paper. We consider a one-dimensional, time-homogeneous stochastic differential equation

$$dX_t = b(X_t, \theta)\, dt + \sigma(X_t, \theta)\, dW_t \tag{1}$$

defined on a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, Pr)$. Here, W is a one-dimensional Brownian motion and θ is an unknown p-dimensional parameter from the parameter space $\Theta \subseteq \mathbb{R}^p$. The true parameter value is denoted θ_0. The functions $b : \mathbb{R} \times \Theta \to \mathbb{R}$ and $\sigma : \mathbb{R} \times \Theta \to (0, \infty)$ are known and assumed to be suitably smooth.

The state space is denoted $I = (l, r)$ for $-\infty \leq l < r \leq +\infty$ (implicitly assuming that it is open and the same for all θ). We shall assume that for any $\theta \in \Theta$ and any $\mathcal{F}_0$-measurable initial condition U with state space I, equation (1) has a unique strong solution X with $X_0 = U$. Assume furthermore that there exists an *invariant distribution* $\mu_\theta = \mu(x, \theta)dx$ such that the solution to (1) with $X_0 \sim \mu_\theta$ is strictly stationary and ergodic. It is well-known that sufficient conditions for this can be expressed in terms of the so-called scale function and speed measure (see the textbooks by Karatzas and Shreve (1991, Section 5.5) or Karlin and Taylor (1981, Section 15.6), for example), and that $\mu(x, \theta)$ is given by

$$\mu(x, \theta) = \big(M(\theta)\sigma^2(x, \theta)s(x, \theta)\big)^{-1} \tag{2}$$

where $\log s(x,\theta) = -2\int_{x_0}^{x} b(y,\theta)/\sigma^2(y,\theta)\,dy$ for some $x_0 \in I$ and $M(\theta)$ is a normalizing constant.

For all $\theta \in \Theta$ the distribution of X with $X_0 \sim \mu_\theta$ is denoted P_θ. Under P_θ all $X_t \sim \mu_\theta$. Further, let for $t \geq 0$ and $x \in I$, $p_\theta(t,x,\cdot)$ denote the conditional density (transition density) of X_t given $X_0 = x$. Since X is time-homogeneous $p_\theta(t,x,\cdot)$ is actually the density of X_{s+t} conditional on $X_s = x$ for all $s \geq 0$. The transition probabilities are most often analytically intractable whereas the invariant density is easy to find (at least up the normalizing constant).

We are going to need some matrix notation: Vectors in $\mathbb{R}^p$ are considered as $p \times 1$ matrices and A^T is the transpose of A. For a function $f = (f_1,\ldots,f_q)^T : \mathbb{R}\times\Theta \to \mathbb{R}^q$ we let $\dot{f}(x,\theta) = \partial_\theta f(x,\theta)$ denote the $q \times p$ matrix of partial derivatives with respect to θ, *i.e.* $\dot{f}_{jk} = \partial f_j/\partial\theta_k$, assuming that the derivatives exist.

Finally, introduce the differential operator $\mathcal{A}_\theta$ given by

$$\mathcal{A}_\theta f(x,\theta) = b(x,\theta)f'(x,\theta) + \tfrac{1}{2}\sigma^2(x,\theta)f''(x,\theta) \tag{3}$$

for twice continuously differentiable functions $f : \mathbb{R}\times\Theta \to \mathbb{R}$. Here f' and f'' are the first and second derivatives with respect to x. When restricted to a suitable subspace, $\mathcal{A}_\theta$ is the *infinitesimal generator* of X (see Rogers and Williams (1987), for example).

3 Preliminary comments on estimation

The objective of this paper is estimation of the parameter θ. First note that if X was observed *continuously* from time zero to time T then parameters from the diffusion coefficient could — in principle, at least — be determined (rather than estimated) from the quadratic variation process of X since

$$\sum_{i=1}^{2^n} \left(X_{t\wedge k/2^n} - X_{t\wedge(k-1)/2^n}\right)^2 \to \int_0^t \sigma^2(X_s,\theta)\,ds$$

in P_{θ_0}-probability for any $t \geq 0$. Thereafter, parameters from the drift could be estimated by maximum likelihood: if the diffusion function is completely known, that is $\sigma(x,\theta) = \sigma(x)$, then the likelihood function for the continuous observation $X_{0\leq t\leq T}$ is given by

$$L_T^c(\theta) = \exp\left(\int_0^T \frac{b(X_s,\theta)}{\sigma^2(X_s)}\,dX_s - \frac{1}{2}\int_0^T \frac{b^2(X_s,\theta)}{\sigma^2(X_s)}\,ds\right). \tag{4}$$

An informal argument for this formula is given below; for a proper proof see Lipster and Shiryayev (1977, Chapter 7).

From now on we shall consider the situation where X is observed at discrete timepoints only. For convenience we consider equidistant points in time: $\Delta, 2\Delta, \ldots, n\Delta$ for some $\Delta > 0$. Conditional on the initial value X_0, the likelihood function is given as the product

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta(\Delta, X_{(i-1)\Delta}, X_{i\Delta})$$

because X is Markov. Ideally, θ should be estimated by the value maximizing $L_n(\theta)$, but since the transition probabilities are not analytically known, neither is the likelihood function.

There are a couple of obvious, very simple alternatives which unfortunately are not satisfactory. First, one could ignore the dependence structure and simply approximate the conditional densities by the marginal density. Then all information due to the time evolution of X is lost, and it is usually not possible to estimate the full parameter vector; see Section 4.2 for further details.

As a second alternative, one could use the *Euler scheme* (or a higher-order scheme) given by the approximation

$$X_{i\Delta} \approx X_{(i-1)\Delta} + b(X_{(i-1)\Delta}, \theta)\Delta + \sigma(X_{(i-1)\Delta}, \theta)\sqrt{\Delta}\epsilon_i \tag{5}$$

where ϵ_i, $i = 1, \ldots, n$ are independent, identically $N(0,1)$-distributed. This approximation is good for small values of Δ but may be bad for larger values. The approximation is two-fold: the moments are not the true conditional moments, and the true conditional distribution need not be Gaussian. The moment approximation introduces bias implying that the corresponding estimator is inconsistent as $n \to \infty$ for any fixed Δ (Florens-Zmirou, 1989). The Gaussian approximation introduces no bias *per se*, but usually implies inefficiency: if the conditional mean and variance are replaced by the true ones, but the Gaussian approximation is maintained, then the corresponding approximation to the score function is a non-optimal martingale estimating function, see Section 4.1.

Note that the Euler approximation provides an informal explanation of formula (4): if σ does not depend on θ, then the discrete-time likelihood function is, except for a constant, via (5) approximated by

$$\exp\left(\sum_{i=1}^{n} \frac{b(X_{(i-1)\Delta}, \theta)}{\sigma^2(X_{(i-1)\Delta})}\big(X_{i\Delta} - X_{(i-1)\Delta}\big) - \frac{1}{2}\Delta \sum_{i=1}^{n} \frac{b^2(X_{(i-1)\Delta}, \theta)}{\sigma^2(X_{(i-1)\Delta})}\right) \tag{6}$$

which is the Riemann-Itô approximation of (4).

4 Estimating functions

Estimating functions provide estimators in very general settings where an unknown p-dimensional parameter θ is to be estimated from data X^{obs} of size n. Basically, an estimating function F_n is simply a $\mathbb{R}^p$-valued function which takes the data as well as the unknown parameter as arguments. An estimator is obtained by solving $F_n(X^{\mathrm{obs}}, \theta) = 0$ for the unknown parameter θ.

The prime example of an estimating function is the score function, yielding the maximum likelihood estimator. When the score function is not available an alternative estimating function should of course be chosen with care. In order for the corresponding estimator to behave (asymptotically) "nicely" it is crucial that the estimating function is unbiased and is able to distinguish the true parameter value from other values of θ:

$$\mathrm{E}_{\theta_0} F_n(X^{\mathrm{obs}}, \theta) = 0 \quad \text{if and only if} \quad \theta = \theta_0. \tag{7}$$

The general theory for estimating functions is reviewed in the textbook by Heyde (1997) (including various applications) and by Sørensen (1999a) (mostly asymptotic theory).

Now, let us turn to the case of discretely observed diffusions again. The score function

$$S_n(\theta) = \partial_\theta \log L_n(\theta) = \sum_{i=1}^{n} \partial_\theta \log p_\theta(\Delta, X_{(i-1)\Delta}, X_{i\Delta})$$

is a sum of n terms where the i'th term depends on data through $(X_{(i-1)\Delta}, X_{i\Delta})$ only. As we are trying to mimic the behaviour of the score function, it is natural to look for estimating functions with the same structure. Hence, we shall consider estimating functions of the form

$$F_n(\theta) = \sum_{i=1}^{n} f(X_{(i-1)\Delta}, X_{i\Delta}, \theta) \tag{8}$$

where we have omitted the dependence of data on F_n from the notation. Condition (7) simplifies to: $\mathrm{E}_{\theta_0} f(X_0, X_\Delta, \theta) = 0$ if and only if $\theta = \theta_0$.

In the following we shall concentrate on two special types of estimating functions, namely *martingale estimating functions* ($F_n(\theta)$ being a P_θ-martingale) and *simple estimating functions* (each term in F_n depending on one observation only). For more comprehensive overviews, see Sørensen (1997) and Jacobsen (2001) and in particular Bibby *et al.* (2002).

4.1 Martingale estimating functions

There are (at least) two good reasons for looking at estimating functions that are martingales: (i) the score function which we are basically trying to imitate is a martingale; and (ii) we have all the machinery from martingale theory (*e.g.* limit theorems) at our disposal. Moreover, martingale estimating functions are important as any asymptotically well-behaved estimating function is asymptotically equivalent to a martingale estimating function (Jacobsen, 2001).

Definition, asymptotic results and optimality

Consider the conditional moment condition

$$\mathrm{E}_\theta\big(\tilde{h}(X_0, X_\Delta, \theta)|X_0 = x\big) = \int_I \tilde{h}(x, y, \theta) p_\theta(\Delta, x, y)\, dy = 0, \quad x \in I, \theta \in \Theta \tag{9}$$

for a function $\tilde{h} : I^2 \times \Theta \to \mathbb{R}$. If all coordinates of f from (8) satisfy this condition, and $(\mathcal{G}_i)$ is the discrete-time filtration generated by the observations, then $F_n(\theta)$ is a P_θ-martingale with respect to $(\mathcal{G}_i)$ since

$$\mathrm{E}_\theta\big(F_n(\theta)|\mathcal{G}_{n-1}\big) = F_{n-1}(\theta) + \mathrm{E}_\theta\big(f(X_{(n-1)\Delta}, X_{n\Delta}, \theta)|X_{(n-1)\Delta}\big) = F_{n-1}(\theta).$$

Suppose that $h_1, \ldots, h_N : I^2 \times \Theta \to \mathbb{R}$ all satisfy (9) and let $\alpha_1, \ldots, \alpha_N : I \times \Theta \to \mathbb{R}^p$ be *arbitrary weight functions*. Then each coordinate of f defined by

$$f(x, y, \theta) = \sum_{j=1}^{N} \alpha_j(x, \theta) h_j(x, y, \theta) = \alpha(x, \theta) h(x, y, \theta)$$

satisfies (9) as well. Here we have used the notation α for the $\mathbb{R}^{p \times N}$-valued function with (k, j)'th element equal to the k'th element of α_j and h for the function $(h_1, \ldots, h_N)^T$ with values in $\mathbb{R}^{N \times 1}$. Note that the score function is obtained as a special case: for $N = p$, $h(x, y, \theta) = (\partial_\theta \log p_\theta(\Delta, x, y))^T$ and $\alpha(x, \theta)$ equal to the $p \times p$ unit matrix.

Classical limit theory for stationary martingales (Billingsley, 1961) can be applied for asymptotic results of F_n with f as above. Under differentiability and integrability

conditions $\dot{F}_n(\theta)/n \to A(\theta)$ in P_{θ_0}-probability for all θ and $F_n(\theta_0)/\sqrt{n} \to N(0, V_0)$ in distribution wrt. P_{θ_0}. Here,

$$A(\theta) = \mathrm{E}_{\theta_0} \dot{f}(X_0, X_\Delta, \theta) = \mathrm{E}_{\theta_0} \alpha(X_0, \theta)\dot{h}(X_0, X_\Delta, \theta)$$
$$V_0 = \mathrm{E}_{\theta_0} f(X_0, X_\Delta, \theta_0) f(X_0, X_\Delta, \theta_0)^T = \mathrm{E}_{\theta_0} \alpha(X_0, \theta_0)\tau_h(X_0, \theta_0)\alpha^T(X_0, \theta_0),$$

where $\tau_h(x, \theta) = \mathrm{Var}_\theta(h(X_0, X_\Delta, \theta)|X_0 = x)$. Sørensen (1999a) proved the following asymptotic result: If the convergence $\dot{F}_n(\theta)/n \to A(\theta)$ is suitably uniform in θ and $A_0 = A(\theta_0)$ is non-singular then a solution $\tilde{\theta}_n$ to $F_n(\theta) = 0$ exists with a probability tending to 1, $\tilde{\theta}_n \to \theta_0$ in probability, and $\sqrt{n}(\tilde{\theta}_n - \theta_0) \to N(0, A_0^{-1} V_0 A_0^{-1^T})$ in distribution wrt. P_{θ_0} . Sørensen (2000, Section 2.3.1) discusses the non-singularity condition of A_0 thoroughly and explains it in terms of reparametrizations.

For $h_1 \ldots, h_N$ given it is easy to find optimal weights $\alpha^\star$ in the sense that the corresponding estimator has the smallest asymptotic variance, where $V \leq V'$ as usual means that $V' - V$ is positive semi-definite (Sørensen, 1997):

$$\alpha^\star(x, \theta) = \Big(\tau_h(x, \theta)^{-1} \mathrm{E}_\theta\big(\dot{h}(X_0, X_\Delta, \theta)|X_0 = x\big)\Big)^T.$$

Calculation of $\alpha^\star$ may, however, give rise to serious numerical problems; in practice an approximation to $\alpha^\star$ is therefore often used instead (Bibby and Sørensen, 1995). This does not infuence the consistency of the estimator, only the efficiency.

How to construct martingale estimating functions in practice

The question on how to choose $h_1, \ldots, h_N$ (and N) is far more subtle (when the score function is not known), and the optimal $h_1, \ldots, h_N$ within some class (typically) change with Δ. Jacobsen (2000, 2001) investigates optimality as $\Delta \to 0$, and it is clear that the score for the invariant measure is optimal as $\Delta \to \infty$. Not much work has been done for fixed values of Δ in between.

To our best knowledge all martingale estimating functions used in the literature so far are based on functions on the form

$$h_j(x, y, \theta) = g_j(y, \theta) - \mathrm{E}_\theta(g_j(X_\Delta, \theta)|X_0 = x)$$

for some (simple) functions $g_j : I \times \Theta \to \mathbb{R}$ in $L^1(\mu_\theta)$, $j = 1, \ldots, N$. Obviously, h_j satisfies (9).

Most often polynomials in y have been used, namely $g_j(y, \theta) = y^{k_j}$ for some (small) integers k_j (Bibby and Sørensen, 1995, 1996, 1997). Then θ only appears in the conditional expectation. In some models low-order conditional moments are known analytically although the transition probabilities are not. But even if this is not the case, the conditional moments are relatively easy to calculate by simulation. Kessler and Paredes (2002) investigates the influence of simulations on the asymptotic properties of the estimator.

Alternatively, eigenfunctions have been used: Let $g_j(\cdot, \theta) : I \to \mathbb{R}$, $j = 1, \ldots, N$ be eigenfunctions for $\mathcal{A}_\theta$ with eigenvalues $\lambda_j(\theta)$. Under mild conditions (Kessler and Sørensen, 1999), $\mathrm{E}_\theta(g_j(X_\Delta, \theta)|X_0 = x) = \exp(-\lambda_j(\theta)\Delta)g_j(x, \theta)$ so

$$h_j(x, y, \theta) = g_j(y, \theta) - e^{-\lambda_j(\theta)\Delta} g_j(x, \theta)$$

is of the above type. The estimating functions based on eigenfunctions have two advantages: they are invariant to twice continuously differentiable transformations of data and the optimal weights are easy to simulate (Sørensen, 1997). However, the applicability is rather limited as the eigenfunctions are known only for a few models; see Kessler and Sørensen (1999) for some non-trivial examples, though.

4.2 Simple estimating functions

An estimating function is called *simple* if it has the form $F_n(\theta) = \sum_{i=1}^n f(X_{i\Delta}, \theta)$ where $f : I \times \Theta \to \mathbb{R}^p$ takes only one state variable as argument (Kessler, 2000). The crucial condition (7) simplifies to: $\mathrm{E}_{\theta_0} f(X_0, \theta) = 0$ if and only if $\theta = \theta_0$. This condition involves the marginal distribution only which has two important consequences: First, since the invariant distribution is known explicitly, it is easy to find functionals f analytically with $\mathrm{E}_{\theta_0} f(X_0, \theta_0) = 0$. Second, simple estimating functions completely ignore the dependence structure of X and can only be used for estimation of (parameters in) the marginal distribution. This is of course a very serious objection.

Kessler (2000) shows asymptotic results for the corresponding estimators and is also concerned with optimality. This work was continued by Jacobsen (2001) who characterizes the optimal simple estimating function, see also Conley *et al.* (1997, Appendix C.4). In practice, however, it is usually not possible to use this characterization and f is chosen somewhat ad hoc.

An obvious possibility is the score corresponding to the invariant distribution, $f = \partial_\theta \log \mu$. Another is moment generated functions $f_j(x, \theta) = x^{k_j} - \mathrm{E}_\theta X_0^{k_j}$, $j = 1, \ldots, p$. Also, functions could be generated by the infinitesimal generator $\mathcal{A}_\theta$ defined by (3): let $h_j : I \times \Theta \to \mathbb{R}$, $j = 1, \ldots, p$, be such that the martingale part of $h_j(X, \theta)$ is a true martingale wrt. P_θ. Then $f = (\mathcal{A}_\theta h_1, \ldots, \mathcal{A}_\theta h_p)^T$ gives rise to an unbiased, simple estimating function. This particular type of estimating function was first introduced by Hansen and Scheinkman (1995) and later discussed by Kessler (2000).

Kessler (2000) suggests to use low-order polynomials for $h_1, \ldots, h_p$, regardless of the model. Sørensen (2001) studies the *model-dependent* choice $(h_1, \ldots, h_p) = \partial_\theta \log \mu$ and recognizes that the corresponding estimating function based on $f_j = \mathcal{A}_\theta(\partial_{\theta_j} \log \mu)$, $j = 1, \ldots, p$, may be interpreted as an approximation to minus twice the continuous-time score function when σ does not depend on θ. Intuitively, one would thus expect it to work well for small values of Δ, and it is indeed small Δ-optimal in the sense of Jacobsen (2001); still if σ does not depend on θ. Note the crucial differences from the usual Riemann-Itô approximation of the continuous-time score, that is, the logarithmic derivative wrt. θ of (6): the above approximation is unbiased while the Riemann-Itô approximation is not.

Finally, note the following relation between the simple estimating function $F_n(\theta) = \sum_{i=1}^n f(X_{i\Delta}, \theta)$ and a class of martingale estimating functions: Define

$$h_f(x, y, \theta) = U_\theta f(y, \theta) - \big(U_\theta f(x, \theta) - f(x, \theta)\big)$$

where U_θ is the potential operator, $U_\theta f(x, \theta) = \sum_{k=0}^\infty \mathrm{E}_\theta(f(X_{k\Delta}, \theta) | X_0 = x)$. Then h_f satisfies condition (9), and $F_n(\theta)$ is asymptotically equivalent to the martingale estimating function $\sum_{i=1}^n h_f(X_{(i-1)\Delta}, X_{i\Delta}, \theta)$ (Jacobsen, 2001). Recall that the martingale estimating function may be improved by introducing weights α (unless of course the optimal weight $\alpha^\star(\cdot, \theta)$ is constant). In this sense martingale estimating functions are always better (or at least as good) as simple estimating functions. In practice it is not very helpful, though, as the potential operator in general is not known! Also, the improvement may be very small as was demonstrated in an example in Sørensen (2000, page 15)

4.3 Comments

Obviously, there are lots of unbiased estimating functions that are neither martingales nor simple. For example,

$$f(x, y, \theta) = h_2(y, \theta)\mathcal{A}_\theta h_1(x, \theta) - h_1(x, \theta)\mathcal{A}_\theta h_2(y, \theta)$$

generates a class of estimating functions which are transition dependent and yet explicit (Hansen and Scheinkman, 1995; Jacobsen, 2001).

Estimating functions of different kinds may of course be combined. For example, one could firstly estimate parameters from the invariant distribution by solving a simple estimating equation and secondly estimate parameters from the conditional distribution one step ahead. See Bibby and Sørensen (2001) for a successful application. Also, note that estimating functions work for multivariate diffusions as well; however, simple estimating functions are less useful than in the univariate setting as they are most often not explicit.

Finally, estimating functions may be used as building blocks for the *generalized method of moments* (GMM), the much favored estimation method in the econometric literature (Hansen, 1982). Estimation via GMM is essentially performed by choosing an estimating function F_n of dimension $p' > p$ and minimizing the quadratic form $F_n(\theta)^T \Omega F_n(\theta)$ for some weight matrix Ω.

5 Approximate maximum likelihood estimation

Estimating functions can be thought of as relatively simple imitations of the score function. We now turn to three quite ambitious approximate maximum likelihood methods. They all supply approximations, analytical or numerical, of $p_\theta(\Delta, x, \cdot)$ for fixed x and θ. Hence, they supply approximations of $p_\theta(\Delta, X_{(i-1)\Delta}, X_{i\Delta})$, $i = 1, \ldots, n$, and therefore of $L_n(\theta)$. The approximate likelihood is finally maximized over $\theta \in \Theta$.

5.1 An analytical approximation

A naive, explicit approximation of the conditional distribution of X_Δ given $X_0 = x$ is provided by the Euler approximation (5). The Gaussian approximation may be poor even if the conditional moments are replaced by accurate approximations (or perhaps even the true moments). A sequence of *explicit, non-Gaussian approximations* of $p_\theta(\Delta, x, \cdot)$ is suggested by Aït-Sahalia (2002b); see also Aït-Sahalia (1999) and Aït-Sahalia (2002a) for an application and an extension to multivariate diffusions. For fixed x and θ the idea is to (i) transform X to a process Z which, conditional on $X_0 = x$, has $Z_0 = 0$ and Z_Δ "close" to standard normal; (ii) define a truncated Hermite series expansion of the density of Z_Δ around the standard normal density; and (iii) invert the Hermite approximation in order to obtain an approximation of $p_\theta(\Delta, x, \cdot)$.

For step (i) define $Z = g_{x,\theta}(X)$ where

$$g_{x,\theta}(y) = \frac{1}{\sqrt{\Delta}} \int_x^y \frac{1}{\sigma(u, \theta)} \, du.$$

Then Z solves $dZ_t = b_Z(Z_t, \theta)\, dt + 1/\sqrt{\Delta}\, dW_t$ with drift function given by Itô's formula and $Z_0 = 0$ (given $X_0 = x$). Note that $g'_{x,\theta}(y) = (\Delta \sigma^2(y, \theta))^{-1/2} > 0$ for all $y \in I$ so that $g_{x,\theta}$ is injective. The data are not actually transformed (this would also be impossible since the transformation depends on θ); the transformation is just a device for the density approximation as explained below.

For step (ii) note that $N(0, 1)$ is a natural approximation of the conditional distribution of Z_Δ given $Z_0 = 0$, as increments of Z over time intervals of length Δ has approximately unit variance. Let $p_\theta^Z(\Delta, 0, \cdot)$ denote the true conditional density of Z_Δ given $Z_0 = 0$ and let $p_\theta^{Z,J}(\Delta, 0, \cdot)$ be the *Hermite series expansion truncated after J terms* of $p_\theta^Z(\Delta, 0, \cdot)$ around the standard normal density. That is, the first term is simply the $N(0, 1)$-density; the remaining terms are corrections given in terms of the Hermite polynomials.

For step (iii) note that the true densities $p_\theta(\Delta, x, \cdot)$ and $p_\theta^Z(\Delta, 0, \cdot)$ are related by

$$p_\theta(\Delta, x, y) = \frac{1}{\sqrt{\Delta}\sigma(x,\theta)} p_\theta^Z\big(\Delta, 0, g_{x,\theta}(y)\big), \quad y \in I$$

and apply this formula to invert the approximation $p_\theta^{Z,J}(\Delta, 0, \cdot)$ of $p_\theta^Z(\Delta, 0, \cdot)$ into an approximation $p_\theta^J(\Delta, x, \cdot)$ of $p_\theta(\Delta, x, \cdot)$ in the natural way:

$$p_\theta^J(\Delta, x, y) = \frac{1}{\sqrt{\Delta}\sigma(x,\theta)} p_\theta^{Z,J}\big(\Delta, 0, g_{x,\theta}(y)\big), \quad y \in I.$$

Then $p_\theta^J(\Delta, x, y)$ converges to $p_\theta(\Delta, x, y)$ as $J \to \infty$, suitably uniformly in y and θ. Furthermore, if $J = J(n)$ tends to infinity fast enough as $n \to \infty$ then the estimator maximizing $\prod_{i=1}^n p_\theta^{J(n)}(\Delta, X_{(i-1)\Delta}, X_{i\Delta})$ is asymptotically equivalent to the maximum likelihood estimator (Aït-Sahalia, 2002b, Theorems 1 and 2).

Note that the coefficients of the Hermite series expansion cannot be computed explicitly but could be replaced by analytical approximations in terms of the infinitesimal generator. Hence, the technique provides *explicit*, though *very complex*, approximations to $p_\theta(\Delta, x, \cdot)$. Aït-Sahalia (2002b) performs very persuasive numerical experiments indicating that the approximate maximum likelihood estimates are very close to the true maximum likelihood estimates, even when only a few terms are included in the Hermite series expansion.

5.2 Numerical solutions of the Kolmogorov forward equation

A classical result from stochastic calculus states that the transition densities under certain regularity conditions are characterized as solutions to the *Kolmogorov forward equations*. Lo (1988) uses a similar result and finds explicit expressions for the likelihood function for a log-normal diffusion with jumps and a Brownian motion with zero as an absorbing state. Poulsen (1999) seems to be the first one to construct numerical solutions for non-trivial diffusion models.

For x and θ fixed the forward equation for $p_\theta(\cdot, x, \cdot)$ is a partial differential equation: for $(t, y) \in (0, \infty) \times I$,

$$\frac{\partial}{\partial t} p_\theta(t, x, y) = -\frac{\partial}{\partial y}\big(b(y,\theta) p_\theta(t, x, y)\big) + \frac{1}{2}\frac{\partial^2}{\partial (y)^2}\big(\sigma^2(y,\theta) p_\theta(t, x, y)\big),$$

with initial condition $p_\theta(0, x, y) = \delta(x - y)$ where δ is the Dirac delta function. In order to calculate the likelihood $L_n(\theta)$ one has to solve n of the above forward equations, one for each $x = X_{(i-1)\Delta}$, $i = 1, \ldots, n$. Note that the forward equation for $X_{(i-1)\Delta}$ determines $p_\theta(t, X_{(i-1)\Delta}, y)$ for *all* values of (t, y), but that we only need it at a single point, namely at $(\Delta, X_{i\Delta})$.

Poulsen (1999) uses the so-called Crank-Nicholson finite difference method for each of the n forward equations. For fixed θ he obtains a second order approximation of $\log L_n(\theta)$ in the sense that the numerical approximation $\log L_n^h(\theta)$ satisfies

$$\log L_n^h(\theta) = \log L_n(\theta) + h^2 f_n^\theta(X_0, X_\Delta, \ldots, X_{n\Delta}) + o(h^2) g_n^\theta(X_0, X_\Delta, \ldots, X_{n\Delta})$$

for suitable functions f_n^θ and g_n^θ. The parameter h determines how fine-grained a (t, y)-grid used in the numerical procedure is (and thus the accuracy of approximation). If $h = h(n)$ tends to zero faster than $n^{-1/4}$ as $n \to \infty$ then the estimator maximizing

$\log L_n^h(\theta)$ is asymptotically equivalent to the maximum likelihood estimator (Poulsen, 1999, Theorem 3).

Poulsen (1999) fits the Chan-Karolyi-Longstaff-Sanders (CKLS) model (Chan *et al.*, 1992) model to a dataset of 655 observations and is able to do so in quite reasonable time. Although n partial differential equations must be solved the method seems to be much faster than the simulation based method below. On the other hand the Crank-Nicholson method is less accurate than, and comparable in computing-time to, the method of Section 5.1 (Jensen and Poulsen, 2002).

5.3 Approximation via simulation

Pedersen (1995b) defines a sequence of approximations to $p_\theta(\Delta, x, \cdot)$ via a missing data approach. The basic idea is to (i) split the time interval from 0 to Δ into pieces short enough that the Euler approximation holds reasonably well; (ii) consider the joint Euler likelihood for the augmented data consisting of the observation X_Δ and the values of X at the endpoints of the subintervals; (iii) integrate the unobserved variable out of the joint Euler density; and (iv) calculate the resulting expectation by simulation. The method has been applied successfully to the CKLS model (Honoré, 1997).

To be precise, let x and θ be fixed, consider an integer $N \geq 0$, and split the interval $[0, \Delta]$ into $N+1$ subintervals of length $\Delta_N = \Delta/(N+1)$. Use the notation $X_{0,k}$ for the (unobserved) value of X at time $k\Delta/(N+1)$, $k = 1, \ldots, N$. Then (with $x_{0,0} = x$ and $x_{0,N+1} = y$),

$$\begin{aligned} p_\theta(\Delta, x, y) &= \int_I p_\theta\big(N\Delta_N, x, x_{0,N}\big) p_\theta\big(\Delta_N, x_{0,N}, y\big)\, dx_{0,N} \\ &= \mathrm{E}_\theta\Big(p_\theta\big(\Delta_N, X_{0,N}, y\big)\big|X_0 = x\Big), \quad y \in I \end{aligned} \tag{10}$$

where we have used the Chapman-Kolmogorov equations.

Now, for Δ_N small (N large), $p_\theta(\Delta_N, x_{0,N}, \cdot)$ is well approximated by the Gaussian density with mean $x_{0,N} + b(x_{0,N}, \theta)\Delta_N$ and variance $\sigma^2(x_{0,N}, \theta)\Delta_N$. Denote this density by $\tilde{p}_\theta^N(\Delta_N, x_{0,N}, \cdot)$. Following (10),

$$p_\theta^N(\Delta, x, y) = \mathrm{E}_\theta\Big(\tilde{p}_\theta^N\big(\Delta_N, X_{0,N}, y\big)\big|X_0 = x\Big)$$

is a natural approximation of $p_\theta(\Delta, x, y)$, $y \in I$. Note that $N = 0$ corresponds to the simple Euler approximation.

The approximate likelihood functions $L_n^N(\theta) = \prod_{i=1}^n p_\theta^N(\Delta, X_{(i-1)\Delta}, X_{i\Delta})$ converge in probability to $L_n(\theta)$ as $N \to \infty$ (Pedersen, 1995b, Theorems 3 and 4). Furthermore, there exists a sequence $N(n)$ such that the estimator maximizing $L_n^{N(n)}(\theta)$ is asymptotically equivalent (as $n \to \infty$) to the maximum likelihood estimator (Pedersen, 1995a, Theorem 3).

In practice one would calculate $p_\theta^N(\Delta, x, y)$ as the average of a large number of values $\{\tilde{p}_\theta^N(\Delta_N, X_{0,N}^r, y)\}_r$ where $X_{0,N}^r$ is the last element of a simulated discrete-time path $X_0, X_{0,1}^r, \ldots, X_{0,N}^r$ started at x. Note that the paths are simulated conditional on $X_0 = x$ only which implies that the simulated values $X_{0,N}^r$ at time $N\Delta_N$ may be far from the observed value at time Δ. This is not very appealing as the continuity of X makes a large jump over a small time interval unlikely to occur in practice. Also, it has the unfortunate numerical implication that a very large number of simulations is needed in order to obtain convergence of the average. Elerian *et al.* (2001, Section 3.1) suggest an importance sampling technique conditioning on the observation at time Δ as well (see Section 6).

6 Bayesian analysis

Bayesian analysis of discretely observed diffusions has been discussed by Eraker (2001), Elerian *et al.* (2001) and Roberts and Stramer (2001). The unknown model parameter is treated as a missing data point, and Markov Chain Monte Carlo (MCMC) methods are used for simulation of the posterior distribution of the parameter with density

$$f(\theta|X_0, X_\Delta, \ldots, X_{n\Delta}) \propto f(X_0, X_\Delta, \ldots, X_{n\Delta}|\theta) f(\theta). \tag{11}$$

The Bayesian estimator of θ is simply the mean (say) of this posterior. Note that we use f generically for densities. In particular, $f(\theta)$ denotes the prior density of the parameter and $f(X_0, \ldots, X_{n\Delta}|\theta)$ denotes the likelihood function evaluated at θ.

The Bayesian approach deals with the intractability of $f(X_0, \ldots, X_{n\Delta}|\theta)$ in a way very similar to that of Pedersen (1995b), namely by introducing auxiliary data and using the Euler approximation over small time intervals. However, the auxiliary data are generated and used quite differently in the two approaches.

As in Section 5.3 each interval $[(i-1)\Delta, i\Delta]$ is split into $N+1$ subintervals of length $\Delta_N = \Delta/(N+1)$. We use the notation $X_{i\Delta,k}$ for the value of X at time $i\Delta + k\Delta/(N+1)$, $i = 0, \ldots, n-1$ and $k = 0, \ldots, N+1$. The value is observed for $k = 0$ and $k = N+1$, and $X_{i\Delta,N+1} = X_{(i+1)\Delta,0}$. Further, let $\tilde{X}_{i\Delta}$ be the collection of latent variables $X_{i\Delta,1}, \ldots, X_{i\Delta,N}$ between $i\Delta$ and $(i+1)\Delta$, let $\tilde{X} = (\tilde{X}_0, \ldots, \tilde{X}_{(n-1)\Delta})$ be the nN-vector of all auxiliary variables, and let X^{obs} be short for the vector of observations $X_0, X_\Delta, \ldots, X_{n\Delta}$.

For N large enough the Euler approximation is quite good so the density of $(X^{\mathrm{obs}}, \tilde{X})$, conditional on θ (and X_0) is roughly

$$f^N(X^{\mathrm{obs}}, \tilde{X}|\theta) = \prod_{i=0}^{n-1} \prod_{k=1}^{N+1} \phi\Big(X_{i\Delta,k}, X_{i\Delta,k-1} + b(X_{i\Delta,k-1}, \theta)\Delta_N, \sigma^2(X_{i\Delta,k-1}, \theta)\Delta_N\Big) \tag{12}$$

where $\phi(\cdot, m, v)$ is the density of $N(m, v)$. The idea is now to generate a Markov chain $\{\tilde{X}^j, \theta^j\}_j$ with invariant (and limiting) density equal to the approximate posterior density

$$f^N(\tilde{X}, \theta|X^{\mathrm{obs}}) = \frac{f^N(X^{\mathrm{obs}}, \tilde{X}|\theta) f(\theta)}{f(X^{\mathrm{obs}})} \propto f^N(X^{\mathrm{obs}}, \tilde{X}|\theta) f(\theta). \tag{13}$$

Then $\{\theta^j\}_j$ has invariant density equal to the marginal of $f^N(\tilde{X}, \theta|X^{\mathrm{obs}})$. This is interpreted as an approximation of the posterior (11) of θ and the Bayes estimator of θ is simply the average of the simulated values $\{\theta^j\}_j$ after some burn-in time.

In order to start off the Markov chain, θ^0 is drawn according to the prior density $f(\theta)$, and $\tilde{X}^0$ is defined by linear interpolation, say, between the observed values of X. The j'th iteration in the Markov chain is conducted in two steps: first, $\tilde{X}^j = (\tilde{X}^j_0, \ldots, \tilde{X}^j_{(n-1)\Delta})$ is updated from $f(\tilde{X}|X^{\mathrm{obs}}, \theta^{j-1})$, and second, θ^j is updated from $f(\theta|X^{\mathrm{obs}}, \tilde{X}^j)$.

For the first step, note that the Markov property of X implies that the conditional distribution of $\tilde{X}_{i\Delta}$ given $(X^{\mathrm{obs}}, \theta)$ depends on $(X_{i\Delta}, X_{(i+1)\Delta}, \theta)$ only, so the vectors $\tilde{X}^j_{i\Delta}$, $i = 0, \ldots, n-1$ may be drawn one at a time. We focus on how to draw $\tilde{X}_0 = (X_{0,1}, \ldots, X_{0,N})$ conditional on $(X_0, X_\Delta, \theta^{j-1})$; the target density being proportional to

$$\prod_{k=1}^{N+1} \phi\Big(X_{0,k}, X_{0,k-1} + b(X_{0,k-1}, \theta^{j-1})\Delta_N, \sigma^2(X_{0,k-1}, \theta^{j-1})\Delta_N\Big),$$

cf. (12). Note the crucial difference from the simulation approach in Section 5.3 where $\tilde{X}_{i\Delta}$ was simulated conditional on $X_{i\Delta}$ only: here $\tilde{X}_{i\Delta}$ is simulated conditional on both $X_{i\Delta}$ and $X_{(i+1)\Delta}$. It is (usually) not possible to find the normalizing constant so direct sampling from the density is not feasible. However, *the Metropolis-Hastings algorithm* may be applied; for example with suitable Gaussian proposals. Eraker (2001) suggests to sample only one element of $\tilde{X}_0$ at a time whereas Elerian *et al.* (2001) suggests to sample block-wise, with random block size. Roberts and Stramer (2001) take a slightly different approach as they sample *transformations* of the missing data in order to improve the rate of convergence of the Markov chain (of course, all the usual problems with convergence of the chain should be investigated). Moreover, they sample all missing data points between two consecutive observations at once, using Brownian bridge arguments.

For the second step it is sometimes possible to find the posterior of θ explicitly from (13) in which case θ is updated by direct sampling from the density. Otherwise the Metropolis-Hastings algorithm is imposed again.

The method is relatively easily extended to the multi-dimensional case. Also, it applies to models that are only partially observed (*e.g.* stochastic volatility models) in which case the values of the unobserved coordinates are simulated like $\tilde{X}$ above (Eraker, 2001). Eraker (2001) analyses US interest rate data and simulated data, using the CKLS model $dX_t = \alpha(\beta - X_t)\,dt + \sigma X_t^\gamma\,dW_t$ as well as a stochastic volatility model. Elerian *et al.* (2001) and Roberts and Stramer (2001) apply the method on simulated data as well as interest rate data using the CIR model (the CKLS model with $\gamma = 1/2$) and various other models.

7 Estimation based on auxiliary models

We now discuss *indirect inference* (Gourieroux *et al.*, 1993) and the so-called *efficient method of moments*, or EMM for short (Gallant and Tauchen, 1996). The methods are essentially applicable whenever simulation from the model is possible and there exists a suitable auxiliary model (hence also for multivariate diffusions). Therefore the methods have gained popularity among econometricians.

The idea is most easily described in a relatively general set-up: let $(Y_1, \ldots, Y_n)$ be data from a (complicated) time series model Q_θ, indexed by the parameter of interest θ. Estimation is performed in two steps: First, the model Q_θ is approximated by a simpler one $\tilde{Q}_\rho$ — *the auxiliary model*, indexed by ρ — and the auxiliary parameter ρ is estimated. Second, the two parameters ρ and θ are linked in order to obtain an estimate of θ. This is done via a GMM procedure, and the first step may simply be viewed as a way of finding moment functionals for the GMM procedure.

Let us be more specific. Assume that $(Y_1, \ldots, Y_n)$ has density $\tilde{q}_n$ wrt. $\tilde{Q}_\rho$ and let $\hat{\rho}_n$ be the maximum likelihood estimator of ρ, that is,

$$\hat{\rho}_n = \operatorname{argmax}_\rho \log \tilde{q}_n(Y_1, \ldots, Y_n, \rho),$$

with first-order condition

$$\tfrac{\partial}{\partial \rho} \log \tilde{q}_n(Y_1, \ldots, Y_n, \hat{\rho}_n) = 0. \tag{14}$$

Loosely speaking, $\hat{\theta}_n$ is now defined such that simulated data drawn from $Q_{\hat{\theta}_n}$ resembles data drawn from $\tilde{Q}_{\hat{\rho}_n}$.

For $\theta \in \Theta$ let $Y_1^\theta, \ldots, Y_R^\theta$ be a long trajectory simulated from Q_θ and let $\hat{\rho}_R(\theta)$ be the maximum likelihood estimator of ρ based on the simulated data. The indirect inference estimator of θ is the value minimizing the quadratic form

$$\left[\hat{\rho}_n - \hat{\rho}_R(\theta)\right]^T \Omega \left[\hat{\rho}_n - \hat{\rho}_R(\theta)\right]$$

where Ω is some positive semidefinite matrix of size $\dim(\rho) \times \dim(\rho)$. In EMM computation of $\hat{\rho}_R(\theta)$ is avoided as

$$\left[\tfrac{\partial}{\partial \rho} \log \tilde{q}_R(Y_1^\theta, \ldots, Y_n^\theta, \hat{\rho}_n)\right] \tilde{\Omega} \left[\tfrac{\partial}{\partial \rho} \log \tilde{q}_R(Y_1^\theta, \ldots, Y_R^\theta, \hat{\rho}_n)\right]^T$$

with $\tilde{\Omega}$ like Ω above is minimized, *cf.* 14.

Both estimators of θ are consistent and asymptotically normal, and they are asymptotically equivalent (if Ω and $\tilde{\Omega}$ are chosen appropriately). If θ and ρ have same dimension, then the two estimators coincide and simply solve $\hat{\rho}_R(\hat{\theta}_n) = \hat{\rho}_n$. However, as the auxiliary model should be both easy to handle statistically and flexible enough to resemble the original model, it is often necessary to use one with higher dimension than the original model.

Of course, the quality of the estimator depends on the auxiliary model. So how should we choose it? For the diffusion models considered in this paper the discrete-time Euler scheme

$$X_{i\Delta} = X_{(i-1)\Delta} + b(X_{(i-1)\Delta}, \rho)\Delta + \sigma(X_{(i-1)\Delta}, \rho)\sqrt{\Delta}\, U_i$$

with $U_1, \ldots, U_n$ independent and identically $N(0,1)$-distributed, is a natural suggestion (Gourieroux *et al.*, 1993). The second step in the estimation procedure corrects for the discrepancy between the true conditional distributions and those suggested by the Euler scheme. In a small simulation study for the Ornstein-Uhlenbeck process (solving $dX_t = \theta X_t\, dt + \sigma dW_t$) the indirect inference estimator was highly inefficient (compared to the maximum likelihood estimator). In the EMM literature it is generally suggested to use auxiliary densities based on expansions of a non-parametric density (Gallant and Long, 1997). Under certain (strong) conditions EMM performed with these auxiliary models is claimed to be as efficient as maximum likelihood.

The suggested auxiliary models are, however, fairly incomprehensible, and also computationally burdensome. We believe that the approximate maximum likelihood methods from Section 5 are as fast and efficient — and far more comprehensible — and that they should therefore be preferred.

8 Conclusion

In this paper we have reviewed various estimation techniques for univariate diffusion processes. We finish by summarizing important points and finally by commenting on possible extensions of the techniques.

8.1 Concluding remarks

Maximum likelihood estimation is typically not possible for diffusion processes that have been observed at discrete time-points only. In this paper we have reviewed a number of alternatives.

From a classical point of view, the most appealing methods are those based on approximations of the true likelihood which in principle can be made arbitrarily accurate.

We reviewed three types: Two of them rely on numerical techniques, one on numerical solutions to partial differential equations and one on simulations. Even with today's efficient computers both methods are quite computationally demanding. The last approximation provides analytical, yet very accurate, approximations to the likelihood function. The expressions are quite complicated, though, even for low-order approximations, and simpler procedures are often valuable.

Estimation via estimating functions is generally faster. So-called simple estimating functions are available in explicit form but provide only estimators for parameters from the marginal distribution. Still, they may be useful for preliminary analysis, for example in combination with martingale estimating functions. The latter are analytically available for a few models but must in general be calculated by simulated. This basically amounts to simulating conditional expectations, which is faster than calculating conditional densities as required by the numerical likelihood approximations mentioned above.

The Bayesian approach is to consider the parameter as random and make simulations from its (posterior) distribution. This is quite hard and requires simulation, conditional on the observations, of the diffusion process at a number of time-points in between those where it was observed. The simulation strategy may prove useful for non-Bayesian analysis as well.

Indirect inference and EMM remove bias due to the discrete-time auxiliary model by simulation methods. The quality of the estimators is bound to depend on the auxiliary model which is chosen somewhat arbitrarily, and we believe that more direct approaches are preferable.

Summarizing, a recommendable approach may be to carry out preliminary analysis, for example find good starting values for later numerical optimizations, by way of estimating functions (which is not too hard) and use an approximate likelihood approach for a more sophisticated study. For the latter, the explicit approximation from Section 5.1 is often to be preferred.

8.2 Outlook

As already mentioned most of the methods in principle apply to multivariate diffusions as well. With a few exceptions this has yet to be demonstrated in practice, though, as there has been very few applications in that direction. Moreover, the computational burden will be even more substantial than for univariate processes. In other words, the properties of the methods still have to be explored in multivariate settings.

There has recently been some focus on so-called stochastic volatility models, that is, two-dimensional diffusions where one of the coordinate processes is completely unobserved. This of course complicates the analysis even further; see Sørensen (2000, Section 3.4) for a survey of estimation techniques. Estimating functions have been developed (Sørensen, 1999b) and the Bayesian approach as well as indirect inference and EMM have been applied (Andersen and Lund, 1997; Eraker, 2001; Gourieroux *et al.*, 1993). It is not yet clear if the approximate likelihood methods from Section 5 can be extended to cover such models, but there are other suggestions on approximate likelihood analysis (Sørensen, 2003). Still, inference for stochastic volatility models, is far from fully explored.

Also, more research on non-parametric estimation would be interesting. For the sampling scheme considered in this paper (with Δ fixed), we are only aware of the paper by Aït-Sahalia (1996) in this area; see also the comments in the introduction regarding other sampling schemes.

Finally, recall that the fundamental problem is the combination of a continuous-time model and discrete-time sampling. Of course the diffusion structure has been used intensively for some approaches, but the ideas may also prove useful for other types of continuous-time models with discrete-time observations. And certainly the work has widened the spectrum of models for which proper statistical analysis is possible.

References

Yacine Aït-Sahalia (1996). Nonparametric pricing of interest rate derivative securities. *Econometrica*, **64**:527–560.

Yacine Aït-Sahalia (1999). Transition densities for interest rate and other nonlinear diffusions. *J. Finance*, **54**:1361–1395.

Yacine Aït-Sahalia (2002). Closed-form expansions for multivariate diffusions. Working Paper, Department of Economics, Princeton University.

Yacine Aït-Sahalia (2002). Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica*, **70**:223–262.

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

Bo Martin Bibby, Martin Jacobsen, and Michael Sørensen (2002). Estimating functions for discretely sampled diffusion-type models. In Yacine Aït-Sahalia and Lars Peter Hansen, editors, *Handbook of Financial Econometrics*. North-Holland, Amsterdam. Forthcoming.

Bo Martin Bibby and Michael Sørensen (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, **1**:17–39.

Bo Martin Bibby and Michael Sørensen (1996). On estimation for discretely observed diffusions: A review. *Theory of Stochastic processes*, **2**:49–56.

Bo Martin Bibby and Michael Sørensen (1997). A hyperbolic diffusion model for stock prices. *Finance Stoch.*, **1**:25–41.

Bo Martin Bibby and Michael Sørensen (2001). Simplified estimating functions for diffusion models with a high-dimensional parameter. *Scand. J. Statist.*, **28**:99–112.

Patrick Billingsley (1961). The Lindeberg-Levy Theorem for martingales. *Proc. Amer. Math. Soc.*, **12**:788–792.

Alain Breton Le (1974). Parameter estimation in a linear stochastic differential equation. In *Transactions of the Seventh Prague Conference and of the European Meeting of Statisticians*, pages 353–366.

K. C. Chan, G. Andrew Karolyi, Francis A. Longstaff, and Anthony B. Sanders (1992). An empirical comparison of alternative models of the short-term interest rate. *J. Finance*, **47**:1209–1227.

Timothy G. Conley, Lars Peter Hansen, Erzo G. J. Luttmer, and José A. Scheinkman (1997). Short-term interest rates as subordinated diffusions. *Review of Financial Studies*, **10**:525–577.

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

Gejza Dohnal (1987). On estimating the diffusion coefficient. *J. Appl. Probab.*, **24**:105–114.

Ola Elerian, Siddhartha Chib, and Neil Shephard (2001). Likelihood inference for discretely observed non-linear diffusions. *Econometrica*, **69**:959–993.

Björn Eraker (2001). MCMC analysis of diffusion models with application to finance. *J. Bus. and Econom. Statist.*, **19**:177–191.

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

Danielle Florens-Zmirou (1993). On estimating the diffusion coefficient from discrete observations. *J. Appl. Probab.*, **30**:790–804.

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

A. Ronald Gallant and George Tauchen (1996). Which moments to match? *Econometric Theory*, **12**:657–681.

Valentine Genon-Catalot and J. Jacod (1994). Estimation of the diffusion coefficient for diffusion processes: random sampling. *Scand. J. Statist.*, **21**:193–221.

Valentine Genon-Catalot, Catherine Laredo, and D. Picard (1992). Non-parametric estimation of the diffusion coefficient by wavelets methods. *Scand. J. Statist.*, **19**:317–335.

C. Gourieroux, A. Monfort, and E. Renault (1993). Indirect inference. *J. Appl. Econometrics*, **8**:S85–S118.

Lars Peter Hansen (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**:1029–1054.

Lars Peter Hansen and José Alexandre Scheinkman (1995). Back to the future: generating moment implications for continuous-time Markov processes. *Econometrica*, **63**:767–804.

Christopher C. Heyde (1997). *Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation*. Springer-Verlag, New York.

Marc Hoffmann (1999). Adaptive estimation in diffusion processes. *Stochastic Process. Appl.*, **79**:135–163.

Peter Honoré (1997). Maximum likelihood estimation of non-linear continuous-time term-structure models. Working paper 1997-7, Department of Finance, Aarhus School of Business.

Martin Jacobsen (2000). Optimality and small Δ-optimality of martingale estimating functions. Preprint 2000-5, Department of Theoretical Statistics, University of Copenhagen. To appear in *Bernoulli*.

Martin Jacobsen (2001). Discretely observed diffusions: classes of estimating functions and small Δ-optimality. *Scand. J. Statist.*, **28**:123–149.

Jean Jacod (2000). Non-parametric kernel estimation of the coefficient of a diffusion. *Scand. J. Statist.*, **27**:83–96.

Bjarke Jensen and Rolf Poulsen (2002). Transition densities of diffusion processes: numerical comparison of approximation techniques. *J. Derivatives*, **9**:18–32.

Ioannis Karatzas and Steven E. Shreve (1991). *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York, 2nd edition.

Samuel Karlin and Howard M. Taylor (1981). *A Second Course in Stochastic Processes*. Academic Press, New York.

Mathieu Kessler (2000). Simple and explicit estimating functions for a discretely observed diffusion process. *Scand. J. Statist.*, **27**:65–82.

Mathieu Kessler and Silvestre Paredes (2002). Computational aspects related to martingale estimating functions for a discretely observed diffusion. Technical report, Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena. To appear in *Scand. J. Statist.*

Mathieu Kessler and Michael Sørensen (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, **5**:299–314.

R. S. Lipster and A. N. Shiryayev (1977). *Statistics of Random Processes*, volume 1. Springer-Verlag, New York.

Andrew W. Lo (1988). Maximum likelihood estimation of generalized Itô processes with discretely sampled data. *Econometric Theory*, **4**:231–247.

Asger Roer Pedersen (1995). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, **1**(3):257–279.

Asger Roer Pedersen (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.*, **22**: 55–71.

Rolf Poulsen (1999). Approximate maximum likelihood estimation of discretely observed diffusion processes. Working paper 29, Centre for Analytical Finance, Aarhus.

Gareth O. Roberts and Osnat Stramer (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, **88**:603–621.

L. C. G. Rogers and David Williams (1987). *Diffusions, Markov Processes, and Martingales*, volume 2: Itô Calculus. Wiley.

Helle Sørensen (2000). *Inference for Diffusion Processes and Stochastic Volatility Models*. PhD thesis, Department of Statistics and Operations Research, University of Copenhagen.

Helle Sørensen (2001). Discretely observed diffusions: approximation of the continuous-time score function. *Scand. J. Statist.*, **28**:113–121.

Helle Sørensen (2003). Simulated likelihood approximations for stochastic volatility models. *Scand. J. Statist.*, **30**:Forthcoming.

Michael Sørensen (1997). Estimating functions for discretely observed diffusions: A review. In Ishwar V. Basawa, V. P. Godambe, and Robert L. Taylor, editors, *Selected Proceedings of the Symposium on Estimating Functions*, volume 32, pages 305–325. IMS Lecture notes.

Michael Sørensen (1999). On asymptotics of estimating functions. *Braz. J. Probab. Statist.*, **13**:111–136.

Michael Sørensen (1999). Prediction-based estimating functions. *Econometrics J.*, **3**: 123–147.