Bernoulli 1(3), 1995, 257–279

Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes

ASGER ROER PEDERSEN*

Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, DK-8000 Aarhus C, Denmark

Most often the likelihood function based on discrete observations of a diffusion process is unknown, and estimators alternative to the well-behaved maximum likelihood estimator must be found. Traditionally, such estimators are defined with origin in the theory for continuous observation of the diffusion process, and are as a consequence strongly biased unless the discrete observation time-points are close. In contrast to these estimators, an estimator based on an approximation to the (unknown) likelihood function was proposed in Pedersen (1994). We prove consistency and asymptotic normality of this estimator with no assumptions on the distance between the discrete observation time-points.

Keywords: approximate inference; approximate likelihood; approximate transition density; discrete observations; Euler–Maruyama; Ornstein–Uhlenbeck; stochastic differential equation

1. Introduction

Consider the problem of estimating the unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^p$ in the stochastic differential equation

$$dX_t = b(t, X_t; \theta)\,dt + \sigma(t, X_t; \theta)\,dW_t, X_0 = x_0, \qquad t \geq 0, \tag{1}$$

where W is an r-dimensional Wiener process, b: $[0,\infty) \times \mathbb{R}^d \mapsto \mathbb{R}^d$ and σ: $[0,\infty) \times \mathbb{R}^d \mapsto M^{d\times r}$ (the set of $d \times r$ matrices), from discrete observations of X at time-points $0 = t_0 < t_1 < \ldots < t_n$.

If the transition densities $p(s, x, t, y; \theta)$ of X are known, an obvious choice of estimator for θ is the maximum likelihood estimator $\hat{\theta}_n$, which maximizes the log-likelihood function for θ

$$\ell_n(\theta) = \sum_{i=1}^{n} \log\{p(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i}; \theta)\},$$

since it is known in many cases to be consistent and asymptotically normally distributed as n tends to infinity (see Billingsley 1961; Dacunha-Castelle and Florens-Zmirou 1986). Unfortunately the transition densities of X are usually unknown.

In such cases most alternative estimators are defined by making at some stage an approximation

* Present address: Department of Biometry and Informatics, Research Centre Foulum, P.O. Box 23, DK-8830 Tjele, Denmark.

to the estimation theory for continuous observation of X (see Hutton and Nelson 1986; Florens-Zmirou 1989; Genon-Catalot 1990). These estimators are, as a consequence, strongly biased unless the discrete observation time-points are close, that is unless $\max_{1 \leq i \leq n} |t_i - t_{i-1}|$ is small. Moreover, they inherit some difficulties regarding the parameter dependence of the diffusion coefficient σ, from the fact that maximum likelihood estimation of unknown parameters in σ is impossible when it is based on continuous observation of X. A brief review and discussion of such estimators is given in Pedersen (1994).

In contrast to these estimators, an approximation $\ell_{n,N}(\theta)$ to the log-likelihood function $\ell_n(\theta)$ was proposed in Pedersen (1994). The approximate log-likelihood function $\ell_{n,N}(\theta)$ depends on an integer N that is controlled by the statistician. For $N = 1$ it is a generalization of the discretization of the log-likelihood function for θ based on continuous observation of X, and as N tends to infinity it converges for each θ in probability to $\ell_n(\theta)$ (see Section 2). Details of the actual calculation of $\ell_{n,N}(\theta)$ for large values of N can be found in Pedersen (1994), where the performance of the approximate maximum likelihood estimator $\hat{\theta}_{n,N}$ obtained by maximizing $\ell_{n,N}(\theta)$ is studied by simulation.

Here we study the estimator $\hat{\theta}_{n,N}$ from a purely theoretical point of view. Since the derivation of $\hat{\theta}_{n,N}$ is motivated by the good properties of the maximum likelihood estimator $\hat{\theta}_n$, we accordingly restrict attention to cases where $\hat{\theta}_n$ is consistent and asymptotically normally distributed. Under this and various additional conditions, we prove the consistency and asymptotic normality of $\hat{\theta}_{n,N}$ as n and N tend to infinity. In some cases we can prove that $\hat{\theta}_{n,N}$ converges to $\hat{\theta}_n$ in probability as N tends to infinity, and the consistency and asymptotic normality of $\hat{\theta}_{n,N}$ is then an immediate consequence of the consistency and asymptotic normality of $\hat{\theta}_n$. In other cases we use general results concerning consistency and asymptotic normality. We show when these apply for $\hat{\theta}_{n,N}$ under the assumption that they apply for $\hat{\theta}_n$. Thus we prove that $\hat{\theta}_{n,N}$ is as good as $\hat{\theta}_n$, with the significant difference that $\hat{\theta}_{n,N}$ can quite generally be calculated in practice whereas $\hat{\theta}_n$ can only rarely be calculated in practice.

In Section 2 we review the definition and basic properties of $\ell_{n,N}(\theta)$ and give an example. Section 3 contains our general results on the consistency and asymptotic normality of $\hat{\theta}_{n,N}$. These results are presented in a very general setting, and cover for instance cases where the log-likelihood function for some reason must be approximated. It is then shown when the corresponding approximate maximum likelihood estimator in a certain sense inherits presumed asymptotic properties of the maximum likelihood estimator. The results of Section 3 are applied in Section 4 to a class of one-dimensional diffusion processes, and we discuss the possibilities for generalizing the methods.

2. The approximate log-likelihood function

In this section we give the definition and basic properties of the approximate log-likelihood function $\ell_{n,N}(\theta)$, and prove the consistency and asymptotic normality of $\hat{\theta}_{n,N}$ in an example.

Even though the transition densities of X are usually unknown they do in fact exist quite generally (see Friedman 1975; Stroock and Varadhan 1979), and so it makes sense to approximate them when they are unknown. In fact an approximation of the transition densities of X constitutes the basis of the definition of the approximate log-likelihood function. Having defined for each $N \in \mathbb{N}$ the approximate transition densities $p_N(s, x, t, y; \theta)$ with respect to λ^d (the d-dimensional Lebesgue measure), we define for each $N \in \mathbb{N}$ the approximate log-likelihood function

$$\ell_{n,N}(\theta) = \sum_{i=1}^{n} \log\{p_N(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i}; \theta)\}.$$

The approximate transition densities $p_N(s, x, t, y; \theta)$ can be derived under the following very natural assumptions. Of course the stochastic differential equation (1) must have a (weak) solution for all $x_0 \in \mathbb{R}^d$ and $\theta \in \Theta$, and for statistical inference to be meaningful the solutions must be unique in law. This is equivalent to requiring for all $\theta \in \Theta$ that the martingale problem for b and $a = \sigma\sigma^{\mathrm{T}}$ is well posed (see Rogers and Williams 1987). Conditions that ensure this can be found in Rogers and Williams (1987) and Stroock and Varadhan (1979). Sufficient conditions are the local Lipschitz and growth conditions for each $\theta \in \Theta$. Finally, we assume that $a(t, x; \theta)$ is positive definite for all $t \geq 0$, $x \in \mathbb{R}^d$ and $\theta \in \Theta$, and denote by $a(t, x; \theta)^{1/2}$ the positive definite square root of $a(t, x; \theta)$.

Under these assumptions any solution to (1) is also a solution to the stochastic differential equation

$$\mathrm{d}X_t = b(t, X_t; \theta)\,\mathrm{d}t + a(t, X_t; \theta)^{1/2}\,\mathrm{d}\tilde{W}_t, X_0 = x_0, \qquad t \geq 0, \tag{2}$$

where

$$\tilde{W}_t = \int_0^t a(s, X_s; \theta)^{-1/2}\,\mathrm{d}\left(X_s - x_0 - \int_0^s b(u, X_u; \theta)\,\mathrm{d}u\right), \qquad t \geq 0,$$

is a d-dimensional Wiener process. Furthermore, the solutions to (1) and (2) induce for each $\theta \in \Theta$ a unique probability measure P_θ on the space $C([0,\infty), \mathbb{R}^d)$ of continuous trajectories from $[0, \infty)$ into $\mathbb{R}^d$ endowed with its Borel σ-field $\mathscr{B}$. The family $\{P_\theta: \theta \in \Theta\}$ of probability measures on $\mathscr{B}$ is assumed to be uniquely parametrized, that is to say, $P_{\theta_1} \equiv P_{\theta_2}$ implies that $\theta_1 = \theta_2$. Due to the positive definiteness of $a(t, x; \theta)$ a solution to (2) can be realized on the probability space $(C([0,\infty), \mathbb{R}^d), \mathscr{B}, P_\theta)$, since

$$W_t^\theta = \int_0^t a(s, X_s; \theta)^{-1/2}\,\mathrm{d}\left(X_s - x_0 - \int_0^s b(u, X_u; \theta)\,\mathrm{d}u\right), \qquad t \geq 0,$$

is a d-dimensional Wiener process under P_θ and

$$X_t = x_0 + \int_0^t b(s, X_s; \theta)\,\mathrm{d}s + \int_0^t a(s, X_s; \theta)^{1/2}\,\mathrm{d}W_s^\theta, \qquad t \geq 0.$$

More generally, we have for each $\theta \in \Theta$ a unique family $\{P_{\theta,s,x}: s \geq 0, x \in \mathbb{R}^d\}$ of probability measures on $(C([0,\infty), \mathbb{R}^d), \mathscr{B})$ induced by the solutions to (1) and (2) for $t \geq s$ with initial conditions $X_s = x$ (see Friedman 1975; Stroock and Varadhan 1979). For each $s \geq 0$, $x \in \mathbb{R}^d$ and $\theta \in \Theta$ we have that

$$P_{\theta,s,x}(X_u = x, 0 \leq u \leq s) = 1.$$

Moreover, we have under $P_{\theta,s,x}$ that

$$X_t = x + \int_s^t b(u, X_u; \theta)\,\mathrm{d}u + \int_s^t a(u, X_u; \theta)^{1/2}\,\mathrm{d}W_u^{\theta,s}, \qquad t \geq s,$$

where

$$W_t^{\theta,s} = \int_s^t a(u, X_u; \theta)^{-1/2}\,\mathrm{d}\left(X_u - x - \int_s^u b(v, X_v; \theta)\,\mathrm{d}v\right), \qquad t \geq s,$$

is a d-dimensional Wiener process after time s. The importance of the probability measures $P_{\theta,s,x}$ is that they determine the transition function $P(s,x,t,A;\theta)$ of X under P_θ. For $0 \le s < t$, $x \in \mathbb{R}^d$, $\theta \in \Theta$ and $A \in \mathscr{B}(\mathbb{R}^d)$ we have that

$$P(s,x,t,A;\theta) = P_{\theta,s,x}(X_t \in A) = P_\theta(X_t \in A \mid X_s = x) = P_\theta \cdot X_t(A \mid X_s = x),$$

where $P_\theta(\cdot \mid X_s = x)$ denotes the conditional probability under P_θ given $X_s = x$, and $P_\theta \cdot X_t(\cdot \mid X_s = x)$ denotes the conditional distribution under P_θ of X_t given $X_s = x$.

The definition of the approximate transition densities $p_N(s,x,t,y;\theta)$, for fixed $0 \le s < t$, $x \in \mathbb{R}^d$, $\theta \in \Theta$ and $N \in \mathbb{N}$, is motivated by the following Euler–Maruyama approximation of X_t under $P_{\theta,s,x}$ (see Kloeden and Platen 1992). Define, for $k = 0, 1, 2, \ldots, N$,

$$\tau_k = s + k\frac{t-s}{N}$$

$$Y_{s,N} = x$$

$$Y_{\tau_k,N} = Y_{\tau_{k-1},N} + \frac{t-s}{N} b(\tau_{k-1}, Y_{\tau_{k-1},N};\theta) + a(\tau_{k-1}, Y_{\tau_{k-1},N};\theta)^{1/2}(W^{\theta,s}_{\tau_k} - W^{\theta,s}_{\tau_{k-1}}).$$

Under the local Lipschitz and growth conditions we have that

$$Y_{\tau_N,N} = Y_{t,N} \to X_t$$

in $L^1(P_{\theta,s,x})$ as $N \to \infty$ (see Kloden and Platen 1992), and we define $y \mapsto p_N(s,x,t,y;\theta)$ to be the density (with respect to λ^d) of the distribution $P_{\theta,s,x} \cdot Y_{t,N}$ of $Y_{t,N}$ under $P_{\theta,s,x}$. For $N = 1$ we can choose the continuous version of the density

$$p_1(s,x,t,y;\theta) = \{2\pi(t-s)\}^{-d/2}|a(s,x;\theta)|^{-1/2}$$
$$\times \exp\left(-\frac{1}{2(t-s)}\{y - x - (t-s)b(s,x;\theta)\}^{\mathrm{T}} a(s,x;\theta)^{-1}\{y - x - (t-s)b(s,x;\theta)\}\right),$$

where $|a(s,x;\theta)|$ denotes the determinant of $a(s,x;\theta)$, and for $N \ge 2$ we have for any version of $p_1(s,x,t,\cdot;\theta)$ the expression

$$p_N(s,x,t,y;\theta) = \mathrm{E}_{P_{\theta,s,x}}(p_1(\tau_{N-1}, Y_{\tau_{N-1},N}, t, y;\theta)).$$

This expression for $p_N(s,x,t,y;\theta)$ enables us to calculate $\ell_{n,N}(\theta)$ and $\hat{\theta}_{n,N}$ in practice by means of simulations of $Y_{\tau_{N-1},N}$ under $P_{\theta,s,x}$, as described in Pedersen (1994).

From the very definition of $p_N(s,x,t,\cdot;\theta)$ and $p(s,x,t,\cdot;\theta)$ as densities with respect to λ^d it is clear that proving the pointwise convergence of $p_N(s,x,t,y;\theta)$ to $p(s,x,t,y;\theta)$ as $N \to \infty$ is a non-trivial task, since it involves choosing definitive versions (for example, continuous) of $p_N(s,x,t,\cdot;\theta)$ and $p(s,x,t,\cdot;\theta)$. This is possible in examples where closed expressions for concrete versions of $p_N(s,x,t,\cdot;\theta)$ and $p(s,x,t,\cdot;\theta)$ are available (see the example below), but in general it is a delicate matter. It has, however, been proved in Pedersen (1994) that (any version of) $p_N(s,x,t,\cdot;\theta)$ converges in $L^1(\lambda^d)$ to (any version of) $p(s,x,t,\cdot;\theta)$ as $N \to \infty$. If $a(t,x;\theta) \equiv a(\theta)$ is independent of t and x this holds under weak assumptions on b (continuous, local Lipschitz and growth conditions), while both the proof and the assumptions become more involved when $a(t,x;\theta)$ is allowed to depend on t and/or x. The $L^1(\lambda^d)$-convergence of $p_N(s,x,t,\cdot;\theta)$ to $p(s,x,t,\cdot;\theta)$ as $N \to \infty$ has the following important consequence, proved in Pedersen (1994).

Corollary 1 If $p_N(s,x,t,\cdot;\theta) \to p(s,x,t,\cdot;\theta)$ in $L^1(\lambda^d)$ as $N \to \infty$ for all $0 \le s < t$, $x \in \mathbb{R}^d$ and $\theta \in \Theta$, then $\ell_{n,N}(\theta) \to \ell_n(\theta)$ in probability under P_{θ_0} as $N \to \infty$ for all $\theta \in \Theta$ and $n \in \mathbb{N}$, where θ_0 denotes the true parameter value.

Example

For the Ornstein–Uhlenbeck process

$$\mathrm{d}X_t = \alpha X_t\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, X_0 = x_0, \qquad t \ge 0$$

with $\theta = (\alpha, \sigma^2)^{\mathrm{T}} \in (-\infty, 0) \times (0, \infty)$ and time-equidistant observations ($t_i = i\Delta$ for $i = 0, 1, 2, \dots$ and some fixed $\Delta > 0$), we have that

$$\psi_n = \sum_{i=1}^{n} X_{i\Delta} X_{(i-1)\Delta} \Big/ \sum_{i=1}^{n} X_{(i-1)\Delta}^2 \to \mathrm{e}^{\Delta\alpha_0}$$

P_{θ_0}-almost surely as $n \to \infty$. Consequently, we have for sufficiently large values of n the following expressions for the maximum likelihood estimators:

$$\hat{\alpha}_n = \frac{1}{\Delta}\log(\psi_n)$$

$$\hat{\sigma}_n^2 = \frac{-2\hat{\alpha}_n}{n(1 - \mathrm{e}^{2\Delta\hat{\alpha}_n})} \sum_{i=1}^{n} (X_{i\Delta} - X_{(i-1)\Delta}\,\mathrm{e}^{\Delta\hat{\alpha}_n})^2.$$

From the equivalence of $\{X_{i\Delta}\}_{i=0}^{n}$ with the AR(1) process, these estimators are easily seen to be consistent and asymptotically normally distributed as $n \to \infty$. In fact

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to N_2(0, i(\theta_0, \Delta)^{-1})$$

in distribution under P_{θ_0} as $n \to \infty$, where

$$i(\theta_0, \Delta)^{-1} = \begin{pmatrix} \dfrac{1 - \mathrm{e}^{2\alpha_0\Delta}}{\Delta^2 \mathrm{e}^{2\alpha_0\Delta}} & \dfrac{2\sigma_0^2}{\Delta} + \dfrac{\sigma_0^2}{\alpha_0\Delta^2}\,\dfrac{1 - \mathrm{e}^{2\alpha_0\Delta}}{\mathrm{e}^{2\alpha_0\Delta}} \\ \dfrac{2\sigma_0^2}{\Delta} + \dfrac{\sigma_0^2}{\alpha_0\Delta^2}\,\dfrac{1 - \mathrm{e}^{2\alpha_0\Delta}}{\mathrm{e}^{2\alpha_0\Delta}} & \dfrac{\sigma_0^4}{\alpha_0^2\Delta^2}\,\dfrac{1 - \mathrm{e}^{2\alpha_0\Delta}}{\mathrm{e}^{2\alpha_0\Delta}} + \dfrac{4\sigma_0^4}{\alpha_0\Delta} + \dfrac{2\sigma_0^4(1 + \mathrm{e}^{2\alpha_0\Delta})}{1 - \mathrm{e}^{2\alpha_0\Delta}} \end{pmatrix}$$

is the inverse of the Fisher information matrix. In this example it is possible to choose continuous versions of $p_N(t,x,\cdot;\theta)$ and $p(t,x,\cdot;\theta)$, and for these we have pointwise convergence as $N \to \infty$. Since

$$\beta_N = \left(1 + \frac{\Delta\alpha}{N}\right)^N \to \mathrm{e}^{\Delta\alpha} = \beta$$

$$\tau_N^2 = \sigma^2 \frac{\Delta}{N}\,\frac{1 - \beta_N^2}{1 - \beta_N^{2/N}} \to \sigma^2 \frac{1 - \mathrm{e}^{2\Delta\alpha}}{-2\alpha} = \tau^2$$

as $N \to \infty$ for all $\alpha < 0$ and $\sigma^2 > 0$, we see that

$$p_N(t,x,y;\theta) = \frac{1}{\sqrt{2\pi\tau_N^2}} \exp\left(-\frac{(y-\beta_N x)^2}{2\tau_N^2}\right)$$
$$\to \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(y-\beta x)^2}{2\tau^2}\right)$$
$$= p(t,x,y;\theta)$$

pointwise as $N \to \infty$. Using this expression for $p_N(t,x,y;\theta)$ we find, for sufficiently large values of n, that

$$\hat{\alpha}_{n,N} = \frac{N}{\Delta}(\psi_n^{1/N} - 1)$$
$$\hat{\sigma}^2_{n,N} = \frac{N}{\Delta}\hat{\tau}_n^2 \frac{1-\psi_n^{2/N}}{1-\psi_n^2},$$

where

$$\hat{\tau}_n^2 = \frac{1}{n}\sum_{i=1}^n X^2_{(i-1)\Delta} - \frac{\left(\frac{1}{n}\sum_{i=1}^n X_{i\Delta}X_{(i-1)\Delta}\right)^2}{\frac{1}{n}\sum_{i=1}^n X^2_{(i-1)\Delta}}.$$

Since $-N(1-x^{1/N}) \to \log(x)$ as $N \to \infty$ for all $x > 0$, we have, for sufficiently large values of n, that

$$\hat{\theta}_{n,N} \to \hat{\theta}_n$$

P_{θ_0}-almost surely as $N \to \infty$. Finally, it follows from Lemma A in the Appendix that there exists a subsequence $N(n) \to \infty$ such that

$$\hat{\theta}_{n,N(n)} \to \theta_0$$

in probability under P_{θ_0} as $n \to \infty$, and such that

$$\sqrt{n}(\hat{\theta}_{n,N(n)} - \theta_0) \to N_2(0, i(\theta_0,\Delta)^{-1})$$

in distribution under P_{θ_0} as $n \to \infty$. Furthermore, if $N'(n) \to \infty$ is a faster subsequence ($N'(n) \geq N(n)$ for all $n \in \mathbb{N}$), then the same results hold for this subsequence.

In the rest of the paper we prove consistency and asymptotic normality of $\hat{\theta}_{n,N}$ in the above sense, that is, we prove the existence of a subsequence $N(n) \to \infty$ such that $\hat{\theta}_{n,N(n)}$ is consistent and asymptotically normally distributed as $n \to \infty$. It is important to stress that $N(n) \to \infty$ is not the only subsequence for which $\hat{\theta}_{n,N}$ is consistent and asymptotically normally distributed, since the same results hold for any faster subsequence. In practice, this means that we do not have to worry about choosing the right value of N for a given (large) number n of observations. The message is simply to choose N as large as practically possible (with respect to computer time, computer power, etc.; see Pedersen 1994). Simulations, however, show that moderate values of N (for example, $N = 25$) are sufficient in most cases (see Pedersen 1994).

3. Consistency and asymptotic normality of the approximate maximum likelihood estimator

For all $\theta \in \Theta$, we have that $\ell_{n,N}(\theta) \to \ell_n(\theta)$ in probability under P_{θ_0} as $N \to \infty$ (cf. Corollary 1 and the preceding discussion). If in fact $\ell_{n,N}(\theta)$ converges uniformly in θ to $\ell_n(\theta)$ in probability under P_{θ_0} as $N \to \infty$, then we can prove that $\hat{\theta}_{n,N} \to \hat{\theta}_n$ in probability under P_{θ_0} as $N \to \infty$ (see Theorem 1 below). The consistency and asymptotic normality of $\hat{\theta}_{n,N}$ as n and N tend to infinity is then an immediate consequence of the consistency and asymptotic normality of $\hat{\theta}_n$ as $n \to \infty$ (see Corollary 2 below). In cases where we cannot prove that $\hat{\theta}_{n,N} \to \hat{\theta}_n$ in probability under P_{θ_0} as $N \to \infty$, we may still prove the consistency and asymptotic normality of $\hat{\theta}_{n,N}$ by means of general results on consistency and asymptotic normality (see Theorems 2 and 3 below). In both cases our results are of a general nature, and we present them in a general setting. First, we introduce the general setting.

Consider on some measurable space $(\Omega, \mathcal{F})$ a sequence of random experiments indexed by $n \in \mathbb{N}$, and a uniquely parametrized family of probability measures P_ψ indexed by $\psi \in \Psi \subseteq \mathbb{R}^q$, giving the possible laws of the experiments. On the basis of the nth experiment and the fundamental outcome $\omega \in \Omega$ we estimate ψ by means of some function

$$h_n: \Omega \times \Psi \mapsto \mathbb{R}$$

that measures how likely the different values of ψ are. The higher the value of $h_n(\omega, \psi)$, the more faith we have in ψ. Consequently, if $h_n(\omega, \psi)$ has a unique maximum point $\hat{\psi}_n(\omega) \in \Psi$, then we estimate ψ by $\hat{\psi}_n(\omega)$. A classical situation is when we have an increasing sequence $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}_n \subseteq \ldots$ of sub-σ-algebras of $\mathcal{F}$, representing an increasing amount of information about the random experiments, such that the restricted probability measures $\{P_\psi^n = P_\psi|_{\mathcal{F}_n} : \psi \in \Psi\}$ are equivalent for each fixed $n \in \mathbb{N}$. Then we may take h_n to be the log-likelihood function

$$h_n(\omega, \psi) = \log\left(\frac{\mathrm{d}P_\psi^n}{\mathrm{d}P^n}(\omega)\right), \tag{3}$$

where P is some fixed member of $\{P_\psi : \psi \in \Psi\}$. If, for instance, the outcomes of the random experiments can be represented by a sequence $X_1, X_2, \ldots$ of random vectors on $(\Omega, \mathcal{F})$ and $\mathcal{F}_n$ is generated by $X_1, \ldots, X_n$, then the log-likelihood function (3) is given by the expression

$$h_n(\omega, \psi) = \log\left(\frac{\mathrm{d}P_\psi \cdot (X_1, \ldots, X_n)}{\mathrm{d}P \cdot (X_1, \ldots, X_n)}(X_1(\omega), \ldots, X_n(\omega))\right).$$

In general, however, we merely assume for all $n \in \mathbb{N}$ and $\psi \in \Psi$ that h_n is a measurable function of ω, and denote by $h_n(\psi)$ the random variable $\omega \mapsto h_n(\omega, \psi)$. The application we have in mind is of course

(cf. Section 2)

$$(\Omega, \mathscr{F}) = (C([0,\infty), \mathbb{R}^d), \mathscr{B}), \Psi = \Theta, q = p, P_\psi = P_\theta$$

and

$$h_n(\theta) = \log\left(\frac{\mathrm{d}P_\theta \cdot (X_{t_1}, \ldots, X_{t_n})}{\mathrm{d}\lambda^{dn}}(X_{t_1}, \ldots, X_{t_n})\right)$$
$$= \sum_{i=1}^{n} \log\left(p(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i}; \theta)\right) = \ell_n(\theta).$$

In the general setting we further have for each $n \in \mathbb{N}$ a sequence of functions

$$h_{n,N}\colon \Omega \times \Psi \mapsto \mathbb{R}, \qquad N = 1, 2, \ldots,$$

each with the same interpretation as h_n. We think of $h_{n,N}$ as an approximation to h_n that in some sense improves as $N \to \infty$. Again we merely assume in general that $h_{n,N}$ is a measurable function of ω for all n, $N \in \mathbb{N}$ and $\psi \in \Psi$, and denote by $h_{n,N}(\psi)$ the random variable $\omega \mapsto h_{n,N}(\omega, \psi)$. The intention is, of course, to take

$$h_{n,N}(\theta) = \sum_{i=1}^{n} \log\left(p_N(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i}; \theta)\right)$$

in the application above.

The approximate maximum likelihood estimator $\hat{\theta}_{n,N}$ is meant in practice to serve as a substitute for $\hat{\theta}_n$ when this cannot be calculated. For the Ornstein–Uhlenbeck process (cf. Section 2) we already know that $\hat{\theta}_{n,N}$ is indeed a good substitute for $\hat{\theta}_n$ for large values of N, in that we have proved that $\hat{\theta}_{n,N} \to \hat{\theta}_n$ in probability under P_{θ_0} as $N \to \infty$. In Section 4 we prove this for a class of one-dimensional diffusion processes by applying the following theorem.

Theorem 1 Let $\Psi \subseteq \mathbb{R}^q$ be a compact subset and assume that the following two conditions are satisfied P_{ψ_0}-almost surely for some fixed $n \in \mathbb{N}$:

(i) $\psi \mapsto h_n(\psi)$ is continuous and has a unique maximum point $\hat{\psi}_n \in \Psi$.
(ii) $\psi \mapsto h_{n,N}(\psi)$ is continuous, at least when N is larger than some fixed $N_0 \in \mathbb{N}$.

Finally, assume that

(iii) $\sup_{\psi \in \Psi} |h_{n,N}(\psi) - h_n(\psi)| \to 0$ in probability under P_{ψ_0} as $N \to \infty$.

Then there exist (P_{ψ_0}-almost surely) sequences $\{\hat{\psi}_{n,N}\}_{N=1}^{\infty} \subseteq \Psi$ of maximum points for the functions $\{h_{n,N}(\psi)\}_{N=1}^{\infty}$, and for any such sequence we have that

$$\hat{\psi}_{n,N} \to \hat{\psi}_n$$

in probability under P_{ψ_0} as $N \to \infty$.

Proof
Any subsequence $N_k \to \infty$ has a further subsequence $N_{k_i} \to \infty$ for which the convergence in (iii)

holds P_{ψ_0}-almost surely, and for this subsequence

$$\hat{\psi}_{n,N_{k_i}} \to \hat{\psi}_n$$

P_{ψ_0}-almost surely as $i \to \infty$. □

Corollary 2 Let the assumptions of Theorem 1 hold for all n larger than some fixed $n_0 \in \mathbb{N}$, and assume that the estimator $\hat{\psi}_n$ is consistent and asymptotically normally distributed, that is:

(i) $\hat{\psi}_n \to \psi_0$ in probability under P_{ψ_0} as $n \to \infty$;
(ii) there exists a sequence $\{A_n(\psi_0)\}_{n=1}^{\infty}$ of non-random and non-singular $q \times q$ matrices such that

$$A_n(\psi_0)(\hat{\psi}_n - \psi_0) \to N_q(0, V(\psi_0))$$

in distribution under P_{ψ_0} as $n \to \infty$, where $V(\psi_0)$ is some non-random positive definite $q \times q$ matrix.

Then there exists a subsequence $N(n) \to \infty$ such that

$$\hat{\psi}_{n,N(n)} \to \psi_0$$

in probability under P_{ψ_0} as $n \to \infty$, and such that

$$A_n(\psi_0)(\hat{\psi}_{n,N(n)} - \psi_0) \to N_q(0, V(\psi_0))$$

in distribution under P_{ψ_0} as $n \to \infty$. Furthermore, if $N'(n) \to \infty$ is a faster subsequence, then the same results hold for this subsequence.

Proof
Apply Theorem 1 and Lemma A in the appendix. □

The general idea underlying the proofs of consistency and asymptotic normality of $\hat{\theta}_{n,N}$ is to prove that $\hat{\theta}_{n,N}$ in some sense inherits these properties from $\hat{\theta}_n$. In cases where this does not take place in the strong sense of Theorem 1 and Corollary 2, we may still prove the consistency and asymptotic normality of $\hat{\theta}_{n,N}$ by means of some general results on consistency and asymptotic normality, in that we show when these apply for $\hat{\theta}_{n,N}$ under the assumption that they apply for $\hat{\theta}_n$. For this purpose we use the general results in Jensen (1986) (see also Sweeting 1980), but we could just as well have used similar results such as those in Billingsley (1961), Dacunha-Castelle and Duflo (1983), Dacunha-Castelle and Florens-Zmirou (1986) or Barndorff-Nielsen and Sørensen (1994). One notable difference between the results in Billingsley (1961), Sweeting (1980), Jensen (1986) and Barndorff-Nielsen and Sørensen (1994), and the results in Dacunha-Castelle and Duflo (1983) and Dacunha-Castelle and Florens-Zmirou (1986) is, however, that the concepts of consistency and asymptotic normality are local properties in the former whereas they are global properties in the latter. The main difference between the two concepts is whether the properties hold for all (global) sequences of maximum estimators or for at least one (local) sequence of maximum estimators. It is, however, immaterial for our purposes which concept we consider. The main idea is to show that even if $\hat{\theta}_{n,N}$ does not converge to $\hat{\theta}_n$ as N tends to infinity for any n, it may still enjoy the same asymptotic properties as $\hat{\theta}_n$ as n and N tend to infinity, and this can be done assuming either of the two concepts of consistency and asymptotic normality mentioned above for $\hat{\theta}_n$. We proceed by formulating the assumptions and results in Jensen (1986) in the general setting above. Throughout we assume that $\Psi \subseteq \mathbb{R}^q$ is an open subset.

The three assumptions in Jensen (1986) are given below.

Assumption 1
The functions $\{h_n\}_{n=1}^{\infty}$ are measurable in ω for all $\psi \in \Psi$ and twice continuously differentiable in ψ for P_{ψ_0}-almost all $\omega \in \Omega$.

For any twice continuously differentiable function $f\colon \Psi \mapsto \mathbb{R}$ we denote by $\dot{f}$ and $\ddot{f}$ respectively the vector of first derivatives and the matrix of second derivatives of f with respect to ψ. Furthermore, for any list $\Gamma = [\psi_1, \ldots, \psi_q]$ of vectors $\psi_i \in \Psi$, $i = 1, \ldots, q$, we denote by $\ddot{f}(\Gamma)$ the $q \times q$ matrix of which the ith row equals the ith row of $\ddot{f}(\psi_i)$. Moreover, we denote by $\dot{h}_n(\psi)$, $\ddot{h}_n(\psi)$ and $\ddot{h}_n(\Gamma)$ the random variable/vector/matrix $\omega \mapsto \dot{h}_n(\omega, \psi)$, $\omega \mapsto \ddot{h}_n(\omega, \psi)$ and $\omega \mapsto \ddot{h}_n(\omega, \Gamma)$. On the space of $q \times q$ matrices we use for convenience the norm

$$\|A\|_{\mathrm{M}} = \max_{1 \le i, j \le q} |A_{ij}|,$$

which is equivalent to the usual Euclidean norm. We are now able to formulate Assumptions 2 and 3.

Assumption 2
There exists a sequence $\{A_n(\psi_0)\}_{n=1}^{\infty}$ of non-random and non-singular $q \times q$ matrices such that:

(i) $A_n(\psi_0)^{-1} \to 0$ as $n \to \infty$.
(ii) $(A_n(\psi_0)^{-1})^{\mathrm{T}} \ddot{h}_n(\psi_0) A_n(\psi_0)^{-1} \to -F(\psi_0)$ in probability under P_{ψ_0} as $n \to \infty$, where $F(\psi_0)$ is some non-random positive definite $q \times q$ matrix.
(iii) For all $\delta > 0$:

$$\sup_{\Gamma \in \Psi_{\delta n}} \|(A_n(\psi_0)^{-1})^{\mathrm{T}} (\ddot{h}_n(\Gamma) - \ddot{h}_n(\psi_0)) A_n(\psi_0)^{-1}\|_{\mathrm{M}} \to 0$$

in probability under P_{ψ_0} as $n \to \infty$, where $\Psi_{\delta n}$ is the set of lists $\Gamma = [\psi_1, \ldots, \psi_q]$ for which $\|A_n(\psi_0)(\psi_i - \psi_0)\| \le \delta$ for all $i = 1, \ldots, q$.

Assumption 3
$(A_n(\psi_0)^{-1})^{\mathrm{T}} \dot{h}_n(\psi_0) \to N_q(0, G(\psi_0))$ in distribution under P_{ψ_0} as $n \to \infty$, where $G(\psi_0)$ is some non-random positive definite $q \times q$ matrix.

Theorem 2 Under Assumptions 1–3 there exists, with P_{ψ_0}-probability that tends to 1 as $n \to \infty$, a sequence $\{\hat{\psi}_n\}_{n=1}^{\infty} \subseteq \Psi$ of local maximum points for the functions $\{h_n(\psi)\}_{n=1}^{\infty}$ such that

$$\hat{\psi}_n \to \psi_0$$

in probability under P_{ψ_0} as $n \to \infty$, and such that

$$A_n(\psi_0)(\hat{\psi}_n - \psi_0) \to N_q(0, F(\psi_0)^{-1} G(\psi_0) F(\psi_0)^{-1})$$

in distribution under P_{ψ_0} as $n \to \infty$.

This result is proved in Jensen (1986). Among its many applications it can be seen (see also Billingsley 1961) that if the diffusion process corresponding to the stochastic differential equation (1) is time-homogeneous and ergodic, the observation time-points are equidistant, and the transition densities (exist and) satisfy some weak regularity conditions, then Theorem 2 applies with $F(\theta_0) = G(\theta_0) = I_p$ and $A_n(\theta_0) = \sqrt{n}i(\theta_0)^{1/2}$, where $i(\theta_0)$ denotes the Fisher information matrix. In this case we further have that

$$-\ddot{\ell}_n(\theta_0)^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow N_p(0, I_p)$$

in distribution under P_{θ_0} as $n \rightarrow \infty$.

In the next theorem we show when Theorem 2 applies for the functions $\{h_{n,N}(\psi)\}_{n=1,N=1}^{\infty,\infty}$ under the assumption that it applies for the functions $\{h_n(\psi)\}_{n=1}^{\infty}$. This method for proving consistency and asymptotic normality of $\hat{\theta}_{n,N}$ is applied in Section 4 to a class of one-dimensional diffusion processes. The random vectors/matrices $\dot{h}_{n,N}(\psi)$, $\ddot{h}_{n,N}(\psi)$ and $\ddot{h}_{n,N}(\Gamma)$ are defined below analogously to $\dot{h}_n(\psi)$, $\ddot{h}_n(\psi)$ and $\ddot{h}_n(\Gamma)$.

Theorem 3 In addition to Assumptions 1–3, assume, for all n and N larger than some fixed values $n_0 \in \mathbb{N}$ and $N_0 \in \mathbb{N}$ respectively, that $h_{n,N}$ is measurable in ω for all $\psi \in \Psi$ and twice continuously differentiable in ψ for P_{ψ_0}-almost all $\omega \in \Omega$. Assume, furthermore, that:

(i) $\dot{h}_{n,N}(\psi_0) \rightarrow \dot{h}_n(\psi_0)$ in probability under P_{ψ_0} as $N \rightarrow \infty$;
(ii) there exists an open subset $\Phi \subseteq \mathbb{R}^q$ such that $\psi_0 \in \Phi \subseteq \Psi$ and

$$\sup_{\psi \in \Phi} \|\ddot{h}_{n,N}(\psi) - \ddot{h}_n(\psi)\|_{\mathrm{M}} \rightarrow 0$$

in probability under P_{ψ_0} as $N \rightarrow \infty$.

Then there exists a subsequence $N(n) \rightarrow \infty$ and, with a P_{ψ_0}-probability that tends to 1 as $n \rightarrow \infty$, a sequence $\{\hat{\psi}_{n,N(n)}\}_{n=1}^{\infty}$ of local maximum points for the functions $\{h_{n,N(n)}(\psi)\}_{n=1}^{\infty}$ such that

$$\hat{\psi}_{n,N(n)} \rightarrow \psi_0$$

in probability under P_{ψ_0} as $n \rightarrow \infty$, and such that

$$A_n(\psi_0)(\hat{\psi}_{n,N(n)} - \psi_0) \rightarrow N_q(0, F(\psi_0)^{-1}G(\psi_0)F(\psi_0)^{-1})$$

in distribution under P_{ψ_0} as $n \rightarrow \infty$. Furthermore, if $N'(n) \rightarrow \infty$ is a faster subsequence, then the same results hold for this subsequence.

Proof

According to assumptions (i) and (ii) and Lemma A in the Appendix we can find a subsequence $N_1(n) \rightarrow \infty$ of the desired type such that Assumptions 2(i), 2(ii) and 3 are satisfied for $\{A_n(\psi_0)\}_{n=1}^{\infty}$ and $\{h_{n,N_1(n)}(\psi)\}_{n=1}^{\infty}$. Now it only remains to prove the existence of a subsequence $N_2(n) \rightarrow \infty$ of the desired type such that

$$\sup_{\Gamma \in \Psi_{\delta n}} \|(A_n(\psi_0)^{-1})^{\mathrm{T}}(\ddot{h}_{n,N_2(n)}(\Gamma) - \ddot{h}_{n,N_2(n)}(\psi_0))A_n(\psi_0)^{-1}\|_{\mathrm{M}} \rightarrow 0$$

in probability under P_{ψ_0} as $n \rightarrow \infty$ for all $\delta > 0$, since then we have that Assumptions 1–3 are

Since

$$\sup_{\theta\in[\alpha,\beta]} \mathrm{E}_{Q_x}\{\exp(2\theta S_{t,N} - \theta^2 L_{t,N})| X_t = y\}$$
$$\leq \mathrm{E}_{Q_x}\{\exp(2\alpha S_{t,N})| X_t = y\} + \mathrm{E}_{Q_x}\{\exp(2\beta S_{t,N})| X_t = y\}$$

and

$$\sup_{\theta\in[\alpha,\beta]} \mathrm{E}_{Q_x}\{[S_{t,N} - \theta L_{t,N}]^2 \exp(2\theta S_{t,n} - \theta^2 L_{t,N})| X_t = y\}$$
$$\leq 2\sqrt{\mathrm{E}_{Q_x}(S_{t,N}^4| X_t = y)}\left\{\sqrt{\mathrm{E}_{Q_x}(\exp(4\alpha S_{t,N})| X_t = y)}\right.$$
$$\left. + \sqrt{\mathrm{E}_{Q_x}(\exp(4\beta S_{t,N})| X_t = y)}\right\}$$
$$+ 2\beta^2 t^2 M^4[\mathrm{E}_{Q_x}\{\exp(2\alpha S_{t,N})| X_t = y\} + \mathrm{E}_{Q_x}\{\exp(2\beta S_{t,N})| X_t = y\}]$$

and finally

$$\sup_{\theta\in[\alpha,\beta]} \mathrm{E}_{Q_x}\{[(S_{t,N} - \theta L_{t,N})^2 - L_t]^2 \exp(2\theta S_{t,N} - \theta^2 L_{t,N})| X_t = y\}$$
$$\leq 16\sqrt{\mathrm{E}_{Q_x}(S_{t,N}^8| X_t = y)}\left\{\sqrt{\mathrm{E}_{Q_x}(\exp(4\alpha S_{t,N})| X_t = y)}\right.$$
$$\left. + \sqrt{\mathrm{E}_{Q_x}(\exp(4\beta S_{t,N})| X_t = y)}\right\}$$
$$+ 2t^2 M^4(8\beta^4 t^2 M^4 + 1)[\mathrm{E}_{Q_x}\{\exp(2\alpha S_{t,N})| X_t = y\}$$
$$+ \mathrm{E}_{Q_x}\{\exp(2\beta S_{t,N})| X_t = y\}],$$

it is clearly enough to prove that

$$\mathrm{E}_{Q_x}\{\exp(\gamma S_{t,N})| X_t = y\} < \infty, \qquad \forall \gamma \in \mathbb{R} \tag{11}$$

in order to prove (6), (8) and (10). Similarly, it is enough to prove that

$$\mathrm{E}_{Q_x}\{\exp(\gamma S_t)| X_t = y\} < \infty, \qquad \forall \gamma \in \mathbb{R} \tag{12}$$

in order to prove (5), (7) and (9). But for a given $\gamma \in \mathbb{R}$ we have that

$$\mathrm{E}_{Q_x}\{\exp(\gamma S_{t,N})| X_t = y\}$$
$$\leq \mathrm{E}_{Q_x}\left\{\exp\left(|\gamma| M \sum_{k=1}^{N} | X_{kt/N} - X_{(k-1)t/N}|\right)\middle| X_t = y\right\}$$
$$= \mathrm{E}\left\{\exp\left(|\gamma| M \sum_{k=1}^{N} | Y_k|\right)\middle| \sum_{k=1}^{N} Y_k = y - x\right\},$$

where $Y_1, \ldots, Y_N$ are stochastically independent and distributed as $N(0, t/N)$, from which (11) follows. Furthermore, we have, by means of Itô's formula, that

$$\begin{aligned}
&\mathrm{E}_{Q_x}\{\exp(\gamma S_t) \mid X_t = y\} \\
&\leq \exp\left(\gamma \int_x^y b(u)\,\mathrm{d}u\right) \mathrm{E}_{Q_x}\left\{\exp\left(|\gamma| \int_0^t |b'(X_s)|\,\mathrm{d}s\right) \middle| X_t = y\right\} \\
&\leq \exp\left(\gamma \int_x^y b(u)\,\mathrm{d}u\right) \exp(t|\gamma| M') < \infty,
\end{aligned}$$

and thus (12) is proved. □

Lemma 2 For all $n \in \mathbb{N}$

$$\sup_{\theta \in [\alpha, \beta]} |\ell_{n,N}(\theta) - \ell_n(\theta)| \to 0$$

$$\sup_{\theta \in [\alpha, \beta]} |\dot{\ell}_{n,N}(\theta) - \dot{\ell}_n(\theta)| \to 0$$

$$\sup_{\theta \in [\alpha, \beta]} |\ddot{\ell}_{n,N}(\theta) - \ddot{\ell}_n(\theta)| \to 0$$

in probability under P_{θ_0} as $N \to \infty$. In particular,

$$\dot{\ell}_{n,N}(\theta_0) \to \dot{\ell}_n(\theta_0)$$

$$\ddot{\ell}_{n,N}(\theta_0) \to \ddot{\ell}_n(\theta_0)$$

in probability under P_{θ_0} as $N \to \infty$.

Proof

The first three convergences can be proved by the same procedure. Here we prove that

$$\sup_{\theta \in [\alpha, \beta]} |\ddot{\ell}_{n,N}(\theta) - \ddot{\ell}_n(\theta)| \to 0$$

in probability under P_{θ_0} as $N \to \infty$ for all $n \in \mathbb{N}$. Let $n \in \mathbb{N}$ be fixed. Clearly it is enough to prove that

$$\sup_{\theta \in [\alpha, \beta]} |\xi_N(s, t, \theta) - \xi(s, t, \theta)| \to 0$$

in probability under P_{θ_0} as $N \to \infty$ for all $0 \leq s < t$, where

$$\xi_N(s, t, \theta) = \frac{p_N(t-s, X_s, X_t; \theta)\,\ddot{p}_N(t-s, X_s, X_t; \theta) - \dot{p}_N(t-s, X_s, X_t; \theta)^2}{p_N(t-s, X_s, X_t; \theta)^2}$$

$$\xi(s, t, \theta) = \frac{p(t-s, X_s, X_t; \theta)\,\ddot{p}(t-s, X_s, X_t; \theta) - \dot{p}(t-s, X_s, X_t; \theta)^2}{p(t-s, X_s, X_t; \theta)^2}.$$

But this follows from Lemmas B and C in the Appendix, with

$$U = (0, \infty) \times \mathbb{R} \times \mathbb{R}, \qquad f(x, y, z) = (xz - y^2)/x^2, \qquad T = [\alpha, \beta]$$

and

$$X_N(\theta) = (p_N(t-s, X_s, X_t; \theta), \dot{p}_N(t-s, X_s, X_t; \theta), \ddot{p}_N(t-s, X_s, X_t; \theta))^{\mathsf{T}}$$

$$X(\theta) = (p(t-s, X_s, X_t; \theta), \dot{p}(t-s, X_s, X_t; \theta), \ddot{p}(t-s, X_s, X_t; \theta))^{\mathsf{T}},$$

as soon as we have established that

$$\sup_{\theta \in [\alpha, \beta]} |p_N(t-s, X_s, X_t; \theta) - p(t-s, X_s, X_t; \theta)| \to 0$$

$$\sup_{\theta \in [\alpha, \beta]} |\dot{p}_N(t-s, X_s, X_t; \theta) - \dot{p}(t-s, X_s, X_t; \theta)| \to 0$$

$$\sup_{\theta \in [\alpha, \beta]} |\ddot{p}_N(t-s, X_s, X_t; \theta) - \ddot{p}(t-s, X_s, X_t; \theta)| \to 0$$

in probability under P_{θ_0} as $N \to \infty$ for all $0 \leq s < t$. These three convergences are proved in exactly the same manner so we prove only the last. Applying the arguments used in the proof of Corollary 1, it is enough to prove that

$$\sup_{\theta \in [\alpha, \beta]} |\ddot{p}_N(t, x, \cdot\,; \theta) - \ddot{p}(t, x, \cdot\,; \theta)| \to 0 \tag{13}$$

in $L^1(\lambda)$ as $N \to \infty$ for all $t \geq 0$ and $x \in \mathbb{R}$. Now let $t \geq 0$ and $x \in \mathbb{R}$ be fixed, and define for all $\theta \in [\alpha, \beta]$ and $N \in \mathbb{N}$

$$U_{t,N}(\theta) = [(S_{t,N} - \theta L_{t,N})^2 - L_{t,N}] \exp\left(\theta S_{t,N} - \frac{\theta^2}{2} L_{t,N}\right)$$

$$U_t(\theta) = [(S_t - \theta L_t)^2 - L_t] \exp\left(\theta S_t - \frac{\theta^2}{2} L_t\right).$$

Then it follows from Dunford-Pettis' theorem that if

$$\sup_{\theta \in [\alpha, \beta]} |U_{t,N}(\theta) - U_t(\theta)| \to 0 \tag{14}$$

in probability under Q_x as $N \to \infty$ and

$$\sup_N \mathrm{E}_{Q_x}\Big(\sup_{\theta \in [\alpha, \beta]} |U_{t,N}(\theta) - U_t(\theta)|^2\Big) < \infty, \tag{15}$$

then

$$\mathrm{E}_{Q_x}\Big(\sup_{\theta \in [\alpha, \beta]} |U_{t,N}(\theta) - U_t(\theta)| \,\Big|\, X_t = \cdot\Big)\,\varphi(\cdot\,; x, t) \to 0$$

in $L^1(\lambda)$ as $N \to \infty$. But since

$$\sup_{\theta \in [\alpha,\beta]} |\ddot{p}_N(t,x;y;\theta) - \ddot{p}(t,x,y;\theta)| \leq \mathrm{E}_{Q_x}\Big(\sup_{\theta \in [\alpha,\beta]} |U_{t,N}(\theta) - U_t(\theta)| \,\Big|\, X_t = y \Big) \varphi(y;x,t)$$

we see that (13) holds in that case. Thus it only remains to prove (14) and (15).

To prove (14), we apply Lemmas B and C in the Appendix with

$$U = \mathbb{R}^2, \qquad f(x,y) = x\,\mathrm{e}^y, \qquad T = [\alpha,\beta]$$

and

$$X_N(\theta) = \left((S_{t,N} - \theta L_{t,N})^2 - L_{t,N}, \theta S_{t,N} - \frac{\theta^2}{2} L_{t,N} \right)^{\mathrm{T}}$$

$$X(\theta) = \left((S_t - \theta L_t)^2 - L_t, \theta S_t - \frac{\theta^2}{2} L_t \right)^{\mathrm{T}}.$$

Notice that condition (i) in Lemma B in the Appendix is fulfilled in this case since

$$\sup_{\theta \in [\alpha,\beta]} |X_N^{(1)}(\theta) - X^{(1)}(\theta)| \leq |L_{t,N} - L_t| + |S_{t,N}^2 - S_t^2| + \beta^2 |L_{t,N}^2 - L_t^2| + 2|\beta|\,|L_{t,N} S_{t,N} - L_t S_t| \to 0$$

in probability under Q_x as $N \to \infty$ (see Jacod and Shiryayev 1987; Revuz and Yor 1991; Pedersen 1994) and

$$\sup_{\theta \in [\alpha,\beta]} |X_N^{(2)}(\theta) - X^{(2)}(\theta)| \leq |\beta|\,|S_{t,N} - S_t| + \tfrac{1}{2}\beta^2 |L_{t,N} - L_t| \to 0$$

in probability under Q_x as $N \to \infty$.

To prove (15), we treat each term on the right-hand side of the inequality

$$\sup_N \mathrm{E}_{Q_x}\Big(\sup_{\theta \in [\alpha,\beta]} |U_{t,N}(\theta) - U_t(\theta)|^2 \Big) \leq 2 \sup_N \mathrm{E}_{Q_x}\Big(\sup_{\theta \in [\alpha,\beta]} U_{t,N}(\theta)^2 \Big) + 2\mathrm{E}_{Q_x}\Big(\sup_{\theta \in [\alpha,\beta]} U_t(\theta)^2 \Big)$$

separately. By direct calculation we see that

$$\begin{aligned} &\mathrm{E}_{Q_x}\Big(\sup_{\theta \in [\alpha,\beta]} U_{t,N}(\theta)^2 \Big) \\ &\leq 16 \sqrt{\mathrm{E}_{Q_x}(S_{t,N}^8)} \left\{ \sqrt{\mathrm{E}_{Q_x}(\exp(4\alpha S_{t,N}))} + \sqrt{\mathrm{E}_{Q_x}(\exp(4\beta S_{t,N}))} \right\} \\ &\quad + 2t^2 M^4 (8\beta^4 t^2 M^4 + 1)[\mathrm{E}_{Q_x}\{\exp(2\alpha S_{t,N})\} + \mathrm{E}_{Q_x}\{\exp(2\beta S_{t,N})\}]. \end{aligned}$$

Now let $\gamma \in \mathbb{R}$. Since

$$\exp\left(\gamma S_{t,N} - \frac{\gamma^2}{2} L_{t,N}\right) \to \exp\left(\gamma S_t - \frac{\gamma^2}{2} L_t\right)$$

in probability under Q_x as $N \to \infty$, we see as in the proof of Theorem 2 in Pedersen (1994) that the convergence is in fact in $L^1(Q_x)$. Consequently, there exists a constant $0 < C_1 < \infty$ such that

$$\mathrm{E}_{Q_x}\left\{\exp\left(\gamma S_{t,N} - \frac{\gamma^2}{2} L_{t,N}\right)\right\} < C_1$$

for all $N \to \mathbb{N}$, and so

$$\begin{aligned}\mathrm{E}_{Q_x}\{\exp(\gamma S_{t,N})\} &\leq \mathrm{E}_{Q_x}\left\{\exp\left(\gamma S_{t,N} - \frac{\gamma^2}{2} L_{t,N}\right)\right\}\exp\left(\frac{\gamma^2}{2} t M^2\right) \\ &< C_1 \exp\left(\frac{\gamma^2}{2} t M^2\right)\end{aligned}$$

for all $N \in \mathbb{N}$. Define

$$b_{s,N} = b(X_{(k-1)t/N}), \qquad (k-1)t/N \leq s < kt/N,$$

for all $0 \leq s \leq t$ and $N \in \mathbb{N}$. Then

$$\begin{aligned}\mathrm{E}_{Q_x}(S_{t,N}^{2m}) &= \mathrm{E}_{Q_x}\left\{\left(\int_0^t b_{s,N}\,\mathrm{d}W_s\right)^{2m}\right\} \\ &\leq C_m \mathrm{E}_{Q_x}(L_{t,N}^m) \\ &\leq C_m t^m M^{2m} \end{aligned} \tag{16}$$

according to the Burkholder–Gundy–Davis inequality, where $0 < C_m < \infty$ in some constant that only depends on $m \in \mathbb{N}$. Consequently,

$$\mathrm{E}_{Q_x}\Big(\sup_{\theta \in [\alpha,\beta]} U_{t,N}(\theta)^2\Big) < C < \infty$$

for all $N \in \mathbb{N}$, where $0 < C < \infty$ is some constant that does not depend on $N \in \mathbb{N}$. Notice that inequality (16) also holds for S_t. Using this and the inequality

$$\begin{aligned}\mathrm{E}_{Q_x}\{\exp(\gamma S_t)\} &\leq \mathrm{E}_{Q_x}\left\{\exp\left(\gamma S_t - \frac{\theta^2}{2} L_t\right)\right\}\exp\left(\frac{\gamma^2}{2} t M^2\right) \\ &= \exp\left(\frac{\gamma^2}{2} t M^2\right), \qquad \forall \gamma \in \mathbb{R},\end{aligned}$$

we see that

$$\mathrm{E}_{Q_x}\Big(\sup_{\theta\in[\alpha,\beta]} U_t(\theta)^2\Big) < \infty,$$

and (15) is proved. □

Thus if $\hat\theta_n$ is unique (P_{θ_0}-almost surely) for some $n \in \mathbb{N}$, then we have from Lemmas 1 and 2 and Theorem 1 that $\hat\theta_{n,N}$ exists for all $N \in \mathbb{N}$ and that

$$\hat\theta_{n,N} \to \hat\theta_n$$

in probability under P_{θ_0} as $N \to \infty$. Furthermore, if $\hat\theta_n$ is unique (P_{θ_0}-almost surely) for all n larger than some $n_0 \in \mathbb{N}$, then it follows from Corollary 2 that consistency and asymptotic normality of $\hat\theta_n$ implies the same for $\hat\theta_{n,N}$. The maximum likelihood estimator $\hat\theta_n$ is, for instance, consistent and asymptotically normally distributed if the log-likelihood function satisfies Assumptions 1–3 in Section 2 (cf. the discussion after Theorem 2), and according to Lemmas 1 and 2 and Theorem 3 we then have that the same holds for $\hat\theta_{n,N}$. In conclusion, both approaches in Section 3 for proving consistency and asymptotic normality of $\hat\theta_{n,N}$ apply to the diffusion process solving (4).

The rather restrictive assumptions in this section on the drift coefficient b were made entirely in order to simplify the exposition and can be relaxed considerably by refining the arguments. More importantly, there are good possibilities for extending the use of the results in Section 3 to the time-inhomogeneous case and to multidimensional diffusion processes and parameters, since the arguments used in the present section essentially rely on the general Lemmas B and C in the Appendix and on the closely related expressions for $p_N(s,x,t,y;\theta)$ and $p(s,x,t,y;\theta)$. Such expressions also exist in the following special case of the stochastic differential equation (1):

$$\mathrm{d}X_t = b(t,X_t;\theta)\,\mathrm{d}t + \sigma(\theta)\,\mathrm{d}W_t, \qquad X_0 = x_0, \qquad t \geq 0.$$

In this case (see Pedersen 1994)

$$\begin{aligned} p(s,x,t,y;\theta) &= \mathrm{E}_{Q_{\theta,s,x}}\left(\frac{\mathrm{d}P_{\theta,s,x}}{\mathrm{d}Q_{\theta,s,x}}\bigg|_{\mathcal{F}_t}\bigg| X_t = y\right)\frac{\mathrm{d}Q_{\theta,s,x}\cdot X_t}{\mathrm{d}\lambda^d}(y) \\ &= \mathrm{E}_{Q_{\theta,s,x}}\{\exp(S_{s,t}(\theta) - \tfrac{1}{2}L_{s,t}(\theta))|\, X_t = y\}\,\varphi_d(y;x,(t-s)a(\theta)) \\ p_N(s,x,t,y;\theta) &= \mathrm{E}_{Q_{\theta,s,x}}\{\exp(S_{s,t,N}(\theta) - \tfrac{1}{2}L_{s,t,N}(\theta))|\, X_t = y\}\,\varphi_d(y;x,(t-s)a(\theta)), \end{aligned}$$

where $Q_{\theta,s,x}$ is the probability measure $P_{\theta,s,x}$ (cf. Section 2) corresponding to the case $b \equiv 0$. $\varphi_d(\cdot\,;\mu,\Sigma)$ denotes the density with respect to λ^d of the d-dimensional normal distribution $N_d(\mu,\Sigma)$, and

$$\begin{aligned} S_{s,t}(\theta) &= \int_s^t b(u,X_u;\theta)^{\mathrm{T}} a(\theta)^{-1}\,\mathrm{d}X_u \\ S_{s,t,N}(\theta) &= \sum_{k=1}^{N} b(\tau_{k-1},X_{\tau_{k-1}};\theta)^{\mathrm{T}} a(\theta)^{-1}(X_{\tau_k} - X_{\tau_{k-1}}) \\ L_{s,t}(\theta) &= \int_s^t b(u,X_u;\theta)^{\mathrm{T}} a(\theta)^{-1} b(u,X_u;\theta)\,\mathrm{d}u \end{aligned}$$

$$L_{s,t,N}(\theta) = \sum_{k=1}^{N} b(\tau_{k-1}, X_{\tau_{k-1}};\theta)^{\mathrm{T}} a(\theta)^{-1} b(\tau_{k-1}, X_{\tau_{k-1}};\theta) \frac{t-s}{N}$$

$$\tau_k = s + k\frac{t-s}{N}, \qquad k = 0, 1, \ldots, N.$$

If, in addition, σ is independent of θ and b depends linearly on θ, then it should be straightforward to extend the methods of this section, but also more generally it should be possible to apply the results of Section 3 using the expressions above. As long as $Q_{\theta,s,x} \cdot X_t$ is absolutely continuous with respect to λ^d we still have this kind of expression for $p(s,x,t,y;\theta)$, but the expression for $p_N(s,x,t,y;\theta)$ depends very much on the fact that σ is independent of both t and x. Thus we find, as in the proofs of the $L^1(\lambda^d)$ convergence of $p_N(s,x,t,\cdot;\theta)$ to $p(s,x,t,\cdot;\theta)$ as $N \to \infty$ (see Pedersen 1994), that the dependency of σ on t and/or x determines the complexity of the analysis, at least by our methods.

Appendix

Lemma A Let $\{X_{n,k}\}_{n=1,k=1}^{\infty,\infty}$, $\{X_n\}_{n=1}^{\infty}$ and X be $\mathbb{R}^d$-valued random vectors. Suppose that $X_{n,k} \to X_n$ in probability (distribution) as $k \to \infty$ for all $n \in \mathbb{N}$ and that $X_n \to X$ in probability (distribution) as $n \to \infty$. Then there exists a subsequence $k(n) \to \infty$ such that $X_{n,k(n)} \to X$ in probability (distribution) as $n \to \infty$, and if $k'(n) \to \infty$ is a faster subsequence ($k'(n) \geq k(n)$ for all $n \in \mathbb{N}$), then the same holds for this subsequence.

Proof
Let $x_{n,k}$, x_n and x be elements of some metric space (M,d) for all $n,k \in \mathbb{N}$. Assume that $d(x_{n,k}, x_n) \to 0$ as $k \to \infty$ for all $n \in \mathbb{N}$ and that $d(x_n, x) \to 0$ as $n \to \infty$. Define $k(n) \in \mathbb{N}$ for all $n \in \mathbb{N}$ such that $d(x_{n,k}, x_n) \leq 2^{-n}$ for all integers $k \geq k(n)$. Then

$$\begin{aligned} d(x_{n,k(n)}, x) &\leq d(x_{n,k(n)}, x_n) + d(x_n, x) \\ &\leq 2^{-n} + d(x_n, x), \end{aligned}$$

and the same inequality holds for all integers $k \geq k(n)$. Since both convergences in the lemma can be viewed as convergences in metric spaces, the lemma is proved. □

Lemma B Let $X_n(\theta)$, $n = 1, 2, \ldots$, and let $X(\theta)$ be $\mathbb{R}^d$-valued random vectors on some probability space $(\Omega, \mathcal{F}, P)$ for all $\theta \in T \subseteq \mathbb{R}^p$, and let f be a real-valued function on $\mathbb{R}^d$ that is continuous on some open subset $U \subseteq \mathbb{R}^d$. Assume that:

(i) $\sup_{\theta \in T} |X_n^{(i)}(\theta) - X^{(i)}(\theta)| \to 0$ in probability as $n \to \infty$ for $i = 1, \ldots, d$, where $X_n^{(i)}(\theta)$ and $X^{(i)}(\theta)$ denote the ith coordinate of $X_n(\theta)$ and $X(\theta)$, respectively;
(ii) for all $\epsilon > 0$ there exists a compact subset $K \subseteq \mathbb{R}^d$ such that $K \subseteq U$ and

$$P(X(\theta) \in K, \forall \theta \in T) > 1 - \epsilon.$$

Then

$$\sup_{\theta \in T} |f(X_n(\theta)) - f(X(\theta))| \to 0$$

in probability as $n \to \infty$.

Proof

For given $\delta > 0$ and $\epsilon > 0$ we must show that there exists an $N \in \mathbb{N}$ such that

$$P\Big(\sup_{\theta \in T} |f(X_n(\theta)) - f(X(\theta))| \le \delta\Big) > 1 - \epsilon$$

for all $n \ge N$. According to assumption (ii) there exists a compact subset $K \subseteq \mathbb{R}^d$ such that $K \subsetneq U$ and

$$P(X(\theta) \in K, \forall \theta \in T) > 1 - \epsilon/2.$$

Now we can find a $\delta_1 > 0$ such that

$$B[K, \delta_1] = \{x \in \mathbb{R}^d \mid \inf_{y \in K} \|x - y\| \le \delta_1\}$$

is a subset of U. Notice that $B[K, \delta_1]$ is a compact subset of $\mathbb{R}^d$ and that

$$u \in K, \qquad \|u - v\| \le \delta_1 \Rightarrow v \in B[K, \delta_1].$$

Since f is absolutely continuous on $B[K, \delta_1]$ we can find a $\delta_2 \in (0, \delta_1)$ such that $|f(u) - f(v)| < \delta$ for all $u, v \in B[K, \delta_1]$ with $\|u - v\| < \delta_2$. Assumption (i) implies the existence of an $N \in \mathbb{N}$ such that

$$P\Big(\sup_{\theta \in T} \|X_n(\theta) - X(\theta)\| < \delta_2\Big) > 1 - \epsilon/2$$

for all $n \ge N$. Let

$$\begin{aligned}
A &= \{X(\theta) \in K, \forall \theta \in T\} \\
B_n &= \{X_n(\theta) \in B[K, \delta_1], \forall \theta \in T\} \\
C_n &= \{\sup_{\theta \in T} \|X_n(\theta) - X(\theta)\| < \delta_2\} \\
D_n &= \{\|X_n(\theta) - X(\theta)\| < \delta_2, \forall \theta \in T\}.
\end{aligned}$$

Then $P(A \cap C_n) > 1 - \epsilon$ for all $n \ge N$. Now

$$\begin{aligned}
A \cap C_n &\subseteq A \cap D_n \\
&= A \cap B_n \cap D_n \\
&\subseteq \{|f(X_n(\theta)) - f(X(\theta))| < \delta, \forall \theta \in T\} \\
&\subseteq \{\sup_{\theta \in T} |f(X_n(\theta)) - f(X(\theta))| \le \delta\}
\end{aligned}$$

and consequently we have for all $n \ge N$ that

$$P\Big(\sup_{\theta \in T} |f(X_n(\theta)) - f(X(\theta))| \le \delta\Big) \ge P(A \cap C_n) > 1 - \epsilon. \qquad \square$$

Lemma C Let $X(\theta)$ be an $\mathbb{R}^d$-valued random vector on some probability space $(\Omega, \mathscr{F}, P)$ for all $\theta \in T \subseteq \mathbb{R}^p$, and let $U \subseteq \mathbb{R}^d$ be an open subset such that

$$P(X(\theta) \in U, \forall\theta \in T) = 1.$$

Assume the following:

(i) There exists an increasing sequence $\{K_n\}_{n=1}^{\infty}$ of compact subsets of $\mathbb{R}^d$ such that $K_n \subseteq U$ for all $n \in \mathbb{N}$, and such that $K_n \uparrow U$ as $n \to \infty$. Furthermore, the interior of K_1 is non-empty.
(ii) $T \subseteq \mathbb{R}^p$ is a compact subset.
(iii) The mapping $\theta \mapsto X(\theta)$ is continuous with probability 1.

Then condition (ii) of Lemma B is satisfied, that is, for all $\epsilon > 0$ there exists a compact subset $K \subseteq \mathbb{R}^d$ such that $K \subseteq U$ and

$$P(X(\theta) \in K, \forall\theta \in T) > 1 - \epsilon.$$

Proof
Let

$$C(T, U) = \{x: T \mapsto U \mid x \text{ is continuous}\}$$

be equipped with the usual metric of uniform convergence on T. Then

$$C_n(T, U) = \{x \in C(T, U) \mid x(\theta) \in K_n, \forall\theta \in T\}$$

is a Borel set in $C(T, U)$ for all $n \in \mathbb{N}$ since it is determined by a countable number of coordinate mappings on $C(T, U)$. Since $x(T) \subseteq U$ is a compact subset for all $x \in C(T, U)$ we have that

$$C_n(T, U) \uparrow C(T, U)$$

as $n \to \infty$, and consequently,

$$P(X(\theta) \in K_n, \forall\theta \in T) = P \cdot X(C_n(T, U)) \to 1$$

as $n \to \infty$, where $P \cdot X$ denotes the distribution on $C(T, U)$ of the random function

$$\omega \mapsto (\theta \mapsto X(\theta, \omega)) \in C(T, U). \qquad \square$$

References

Barndorff-Nielsen, O.E. and Sørensen, M. (1994) A review of some aspects of asymptotic likelihood theory for stochastic processes. *Internat. Statist. Rev.*, **62**, 133–165.

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

Dacunha-Castelle, D. and Duflo, M. (1983) *Probabilités et Statistiques*. Paris: Masson.

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

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

Friedman, A. (1975) *Stochastic Differential Equations and Applications, Volume 1*. New York: Academic Press.

Genon-Catalot, V. (1990) Maximum contrast estimation for diffusion processes from discrete observations. *Statistics*, **21**, 99–116.
Hutton, J.E. and Nelson, P.I. (1986) Quasi-likelihood estimation for semi-martingales. *Stochastic Process. Appl.*, **22**, 245–257.
Jacod, J. and Shiryayev, A.N. (1987) *Limit Theorems for Stochastic Processes*. Berlin: Springer-Verlag.
Jensen, J.L. (1986) *Nogle Asymptotiske Resultater*. Aarhus University (in Danish).
Kloeden, P.E. and Platen, E. (1992) *Numerical Solution of Stochastic Differential Equations*. New York: Springer-Verlag.
Pedersen, A.R. (1995) A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.*, **22**, 55–71.
Revuz, D. and Yor, M. (1991) *Continuous Martingales and Brownian Motion*. Berlin: Springer-Verlag.
Rogers, L.C.G. and Williams, D. (1987) *Diffusions, Markov Processes, and Martingales, Volume 2: Itô Calculus*. Chichester: Wiley.
Stroock, D.W. and Varadhan, S.R.S. (1979) *Multidimensional Diffusion Processes*. Berlin: Springer-Verlag.
Sweeting, T.J. (1980) Uniform asymptotic normality of the maximum likelihood estimator. *Ann. Statist.*, **8**, 1375–1381.

Received December 1993 and revised August 1994