Board of the Foundation of the Scandinavian Journal of Statistics

A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations
Author(s): Asger Roer Pedersen
Source: *Scandinavian Journal of Statistics*, Vol. 22, No. 1 (Mar., 1995), pp. 55-71
Published by: Blackwell Publishing on behalf of Board of the Foundation of the Scandinavian Journal of Statistics
Stable URL: http://www.jstor.org/stable/4616340
Accessed: 09/09/2011 07:55

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Blackwell Publishing and *Board of the Foundation of the Scandinavian Journal of Statistics* are collaborating with JSTOR to digitize, preserve and extend access to *Scandinavian Journal of Statistics.*

http://www.jstor.org

© Board of the Foundation of the Scandinavian Journal of Statistics 1995. Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF and 238 Main Street, Cambridge, MA 02142, USA. Vol. 22: 55–71, 1995

A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations

ASGER ROER PEDERSEN

University of Aarhus and National Environmental Research Institute, Silkeborg

ABSTRACT. We consider maximum likelihood estimation for stochastic differential equations based on discrete observations when the likelihood function is unknown. A sequence of approximations to the likelihood function is derived, and convergence results for the sequence are proven. Estimation by means of the approximate likelihood functions is easy and very generally applicable. The performance of the suggested estimators is studied in two examples, and they are compared with other estimators.

Key words: approximate likelihood, approximate transition density, diffusion process, discrete observation, Euler–Maruyama, Ornstein–Uhlenbeck process, stochastic differential equation

1. Introduction

Consider the stochastic differential equation

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

where W is an r-dimensional Wiener process, $\theta \in \Theta \subseteq \mathbf{R}^p$ is an unknown parameter, $b(\cdot, \cdot; \theta): [0, \infty) \times \mathbf{R}^d \mapsto \mathbf{R}^d$ and $\sigma(\cdot, \cdot; \theta): [0, \infty) \times \mathbf{R}^d \mapsto M^{d \times r}$ (the set of real $d \times r$ matrices). To the estimation of θ, from discrete observations of X at time-points $0 = t_0 < t_1 < \cdots < t_n$, the literature has mainly concentrated on two approaches.

If the transition densities $p(s, x, t, y; \theta)$ of X are known, we can use the log-likelihood function

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

for θ. The corresponding maximum likelihood estimator $\hat{\theta}_n$ is known to have the usual good properties (see Billingsley, 1961; Dacunha-Castelle & Florens-Zmirou, 1986). In the case of time-equidistant observations ($t_i = i\Delta$, $i = 0, 1, \ldots, n$ for some fixed $\Delta > 0$) Dacunha-Castelle & Florens-Zmirou (1986) prove consistency and asymptotic normality of $\hat{\theta}_n$ as $n \to \infty$, irrespective of the value of Δ. It is only natural that the number of observations must be large for any estimator to be close to the true value θ_0, and from a practical point of view it is then an important property of $\hat{\theta}_n$ that it can be expected to be close to the true value θ_0 for any value of $\max_{1 \leqslant i \leqslant n} |t_i - t_{i-1}|$ provided n is large enough. Unfortunately the transition densities of X are usually unknown.

When the transition densities of X are unknown the usual alternative to using $l_n(\theta)$ is to approximate the log-likelihood function for θ based on continuous observation of X. For this log-likelihood function to be defined, $\sigma(t, x; \theta) = \sigma(t, x)$ must be known. Under certain conditions (see Liptser & Shiryayev, 1977) the log-likelihood function for θ based on continuous observation of X in the time-interval $[0, t_n]$ is given by

$$\begin{aligned} l^c_{t_n}(\theta) = &\int_0^{t_n} b(s, X_s; \theta)^*(\sigma(s, X_s)\sigma(s, X_s)^*)^{-1}\, dX_s \\ &- \frac{1}{2}\int_0^{t_n} b(s, X_s; \theta)^*(\sigma(s, X_s)\sigma(s, X_s)^*)^{-1} b(s, X_s; \theta)\, ds, \end{aligned}$$

and approximating the integrals in the usual way we obtain the approximate log-likelihood function for θ

$$\tilde{l}_n(\theta) = \sum_{i=1}^{n} b(t_{i-1}, X_{t_{i-1}}; \theta)^*(\sigma(t_{i-1}, X_{t_{i-1}})\sigma(t_{i-1}, X_{t_{i-1}})^*)^{-1}(X_{t_i} - X_{t_{i-1}})$$

$$- \frac{1}{2}\sum_{i=1}^{n} b(t_{i-1}, X_{t_{i-1}}; \theta)^*(\sigma(t_{i-1}, X_{t_{i-1}})\sigma(t_{i-1}, X_{t_{i-1}})^*)^{-1}$$

$$\times\, b(t_{i-1}, X_{t_{i-1}}; \theta)(t_i - t_{i-1})$$

based on the discrete observations of X. If θ admits a partition $\theta = (\theta_1, \theta_2)$ such that $b(\cdot, \cdot; \theta)$ only depends on θ_1 and $\sigma(\cdot, \cdot; \theta)$ is known up to the scalar factor θ_2 ($\sigma(\cdot, \cdot; \theta) = \theta_2\tilde{\sigma}(\cdot, \cdot)$), then $\tilde{l}_n$ can still be used to estimate θ_1 since it essentially depends on θ_1 only, and θ_2^2 can be estimated by a quadratic variation-like formula (see Florens-Zmirou, 1989). In the multidimensional case where $\theta_2 \in \Theta_2 \subseteq M^{q\times r}$, $\tilde{\sigma}: [0, \infty) \times \mathbf{R}^d \mapsto M^{d\times q}$ and $\sigma(\cdot, \cdot; \theta) = \tilde{\sigma}(\cdot, \cdot)\theta_2$, the matrix θ_2 can be estimated similarly in advance. By inserting the estimate in the expression for $\sigma(\cdot, \cdot; \theta)$ this can be assumed "known", and the estimate of $\tilde{l}_n$ thus obtained can be used to estimate θ_1 (see Yoshida, 1990). These restrictions on $\sigma(\cdot, \cdot; \theta)$ regarding parameter dependence can be relaxed by applying the ideas in Hutton & Nelson (1986) to the case of discrete observations. Under some regularity conditions the score function corresponding to $l^c_{t_n}(\theta)$ is given by

$$\dot{l}^c_{t_n}(\theta) = \int_0^{t_n} \dot{b}(s, X_s; \theta)^*(\sigma(s, X_s)\sigma(s, X_s)^*)^{-1}\, dX_s$$

$$- \int_0^{t_n} \dot{b}(s, X_s; \theta)^*(\sigma(s, X_s)\sigma(s, X_s)^*)^{-1} b(s, X_s; \theta)\, ds,$$

where $\dot{l}^c_{t_n}(\theta)$ and $\dot{b}(\cdot, \cdot; \theta)$ denote the derivatives of $l^c_{t_n}(\theta)$ and $b(\cdot, \cdot; \theta)$ with respect to θ, and the maximum likelihood estimator of θ based on continuous observation of X in the time-interval $[0, t_n]$ is the solution to $0 = \dot{l}^c_{t_n}(\theta)$. When $\sigma(\cdot, \cdot; \theta)$ depends on θ, Hutton & Nelson (1986) show that $\dot{l}^c_{t_n}(\theta)$, with $\sigma(\cdot, \cdot)$ replaced by $\sigma(\cdot, \cdot; \theta)$, can under certain regularity conditions still be used to estimate θ. Approximating the integrals in $\dot{l}^c_{t_n}(\theta)$ as before we thus have a way to estimate θ from discrete observations of X when $\sigma(\cdot, \cdot; \theta)$ depends on θ more generally. We shall, however, not pursue this idea, since a common unpleasant feature of the estimation methods for discrete observations with origin in the theory for continuous observation is that the estimators are strongly biased unless $\max_{1\leqslant i\leqslant n} |t_i - t_{i-1}|$ is "small". In the case of time-equidistant observations with Δ fixed, Florens-Zmirou (1989) actually shows that the estimator $\tilde{\theta}_n$ of θ obtained by maximizing $\tilde{l}_n(\theta)$ is inconsistent.

For the Ornstein–Uhlenbeck process

$$dX_t = \phi X_t\, dt + \sigma\, dW_t, \quad X_0 = x_0, \quad t \geqslant 0,$$

with $(\phi, \sigma) \in (-\infty, 0) \times (0, \infty)$, $t_i = i\Delta$ for $i = 0, 1, 2, \ldots$ and some fixed $\Delta > 0$, the maximum likelihood estimators

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

$$\hat{\sigma}^2_n = \frac{-2\hat{\phi}_n}{n(1-\exp(2\Delta\hat{\phi}_n))}\sum_{i=1}^{n}(X_{i\Delta} - X_{(i-1)\Delta}\exp(\Delta\hat{\phi}_n))^2$$

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

are consistent, whereas the $\tilde{l}_n(\phi)$-estimator $\tilde{\phi}_n$ of ϕ and the quadratic variation-like estimator $\tilde{\sigma}_n^2$ of σ^2 are inconsistent. In fact

$$\tilde{\phi}_n = \frac{1}{\Delta}\left(\frac{\sum_{i=1}^{n} X_{(i-1)\Delta}X_{i\Delta}}{\sum_{i=1}^{n} X_{(i-1)\Delta}^2} - 1\right) \rightarrow \frac{\exp(\Delta\phi_0) - 1}{\Delta} > \phi_0 \tag{2}$$

$$\tilde{\sigma}_n^2 = \frac{1}{n\Delta}\sum_{i=1}^{n} (X_{i\Delta} - X_{(i-1)\Delta})^2 \rightarrow \sigma_0^2 \frac{1 - \exp(\Delta\phi_0)}{-\Delta\phi_0} < \sigma_0^2 \tag{3}$$

in probability as $n \rightarrow \infty$, so if Δ is large, the estimators $\tilde{\phi}_n$ and $\tilde{\sigma}_n^2$ actually concentrate near 0 as $n \rightarrow \infty$.

The dependence of $\tilde{l}_n(\theta)$ on $\max_{1 \leqslant i \leqslant n} |t_i - t_{i-1}|$ and the difficulties regarding the parameter dependence of $\sigma(\cdot, \cdot; \theta)$ are both consequences of $\tilde{l}_n(\theta)$ being an approximation to $l_{t_n}^c(\theta)$. In this paper we derive a sequence $\{l_n^{(N)}(\theta)\}_{N=1}^{\infty}$ of approximations to $l_n(\theta)$, that gives a connection between $\tilde{l}_n(\theta)$ and $l_n(\theta)$. In fact $l_n^{(1)}(\theta)$ can be seen as a generalization of $\tilde{l}_n(\theta)$ with no restrictions on $\sigma(\cdot, \cdot; \theta)$ regarding parameter dependence, while the approximations $l_n^{(N)}(\theta)$ for $N \geqslant 2$ are improvements of $l_n^{(1)}(\theta)$ that approaches $l_n(\theta)$ as $N \rightarrow \infty$. The idea is to approximate the (unknown) transition densities $p(s, x, t, y; \theta)$ of X by a sequence of transition densities $p^{(N)}(s, x, t, y; \theta)$ of approximating Markov processes that converge to $p(s, x, t, y; \theta)$ as $N \rightarrow \infty$, and then to define the approximate log-likelihood functions

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

The sequence $\{l_n^{(N)}(\theta)\}_{N=1}^{\infty}$ is derived under very natural assumptions. In section 2 we list these assumptions and derive the approximating transition densities. Results on their convergence to $p(s, x, t, y; \theta)$ as $N \rightarrow \infty$ and on the convergence of $l_n^{(N)}(\theta)$ to $l_n(\theta)$ as $N \rightarrow \infty$ are given in section 3. Section 4 contains some general considerations about the actual calculation of $l_n^{(N)}(\theta)$ for large values of N, and the proposed estimation method is applied to two examples.

Estimation by means of $l_n^{(N)}(\theta)$ for some fixed N requires no more than the functions $b(t, x; \theta)$ and $\sigma(t, x; \theta)$ themselves, and is easy to implement. The integer N should be chosen as large as possible, but in all considered examples, $N \simeq 25$ has proven to be sufficient. In any case, $l_n^{(1)}(\theta)$ generalizes $\tilde{l}_n(\theta)$ and $l_n^{(N)}(\theta)$ for $N \geqslant 2$ are improvements of $l_n^{(1)}(\theta)$.

2. Derivation of the approximate log-likelihood functions

In this section we derive the approximating densities $p^{(N)}(s, x, t, y; \theta)$, $N = 1, 2, \ldots$. First we list the assumptions.

Of course the stochastic differential equation (1) must have a (weak) solution for all x_0 and θ, and for statistical inference to be meaningful the solutions must be unique in law. This is equivalent to requiring for every θ that the martingale problem for b and $a = \sigma\sigma^*$ is well-posed (see Rogers & Williams, 1987). Conditions that ensure this can be found in Rogers & Williams (1987) and Stroock & Varadhan (1979). Sufficient conditions are the following (A1)–(A2), which must hold for all $\theta \in \Theta$. Let $\|\cdot\|$ denote the Euclidian norm.

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

(A1) *For all* $0 < R < \infty$ *there exists* $0 < K_R < \infty$ *such that*

$$\|\sigma(t, x; \theta) - \sigma(t, y; \theta)\| \leqslant K_R \|x - y\|$$

$$\|b(t, x; \theta) - b(t, y; \theta)\| \leqslant K_R \|x - y\|$$

for all $0 \leqslant t \leqslant R$ *and* $x, y \in \mathbf{R}^d$ *with* $\|x\| \leqslant R$, $\|y\| \leqslant R$.

(A2) *For all* $0 < T < \infty$ *there exists* $0 < C_T < \infty$ *such that*

$$\|\sigma(t, x; \theta)\| + \|b(t, x; \theta)\| \leqslant C_T(1 + \|x\|)$$

for all $0 \leqslant t \leqslant T$ *and* $x \in \mathbf{R}^d$.

Furthermore we shall assume for all $\theta \in \Theta$ that

(A3) $a(t, x; \theta) = \sigma(t, x; \theta)\sigma(t, x; \theta)^*$ *is positive definite for all* $t \geqslant 0$ *and* $x \in \mathbf{R}^d$.

Under the assumptions (A1)–(A3) we have for each $\theta \in \Theta$ a unique family $\{P_{\theta, s, x}: s \geqslant 0, x \in \mathbf{R}^d\}$ of probability measures on $(\Omega, \mathscr{F}) = (C([0, \infty), \mathbf{R}^d), \mathscr{B})$, the space of continuous trajectories from $[0, \infty)$ into $\mathbf{R}^d$ endowed with its Borel σ-field, induced by the solutions to (1) for $t \geqslant s$ with initial conditions $X_s = x$ (see Friedman, 1975; Stroock & Varadhan, 1979). Let $X = (X_t)$ be the coordinate process on Ω and $(\mathscr{F}_t)$ the natural filtration. Furthermore, denote by P_θ the probability measure $P_{\theta, 0, x_0}$. Under $P_{\theta, s, x}$

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

is a d-dimensional Wiener process after time s and

$$X_t = x + \int_s^t b(u, X_u; \theta)\, du + \int_s^t a(u, X_u; \theta)^{1/2}\, dW_u^{\theta, s}, \quad t \geqslant s.$$

The important thing about 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 $0 \leqslant s < t$, $x \in \mathbf{R}^d$ and $A \in \mathscr{B}(\mathbf{R}^d)$

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

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

$$\tau_k = s + k\frac{t - s}{N} \tag{4}$$

$$Y_s^{(N)} = x$$

$$Y_{\tau_k}^{(N)} = Y_{\tau_{k-1}}^{(N)} + \frac{t - s}{N} b(\tau_{k-1}, Y_{\tau_{k-1}}^{(N)}; \theta) + a(\tau_{k-1}, Y_{\tau_{k-1}}^{(N)}; \theta)^{1/2}(W_{\tau_k}^{\theta, s} - W_{\tau_{k-1}}^{\theta, s}). \tag{5}$$

Then (under (A1)–(A3))

$$Y_{\tau_N}^{(N)} = Y_t^{(N)} \to X_t$$

in $L^1(P_{\theta, s, x})$ as $N \to \infty$ (see Kloeden & Platen, 1992), so if the distribution $P_{\theta, s, x} \cdot Y_t^{(N)}$ of $Y_t^{(N)}$ under $P_{\theta, s, x}$ has a density with respect to λ^d (the d-dimensional Lebesgue measure) we may take this as an approximation to the density $y \mapsto p(s, x, t, y; \theta)$.

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

Theorem 1

For fixed $0 \leqslant s < t$, $x \in \mathbf{R}^d$, $\theta \in \Theta$ *and* $N \in \mathbf{N}$ *the distribution of* $Y_t^{(N)}$ *under* $P_{0,s,x}$ *has a density* $p^{(N)}(s, x, t, \cdot\,; \theta)$ *with respect to* λ^d. *For* $N = 1$ *we can choose the continuous version*

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

where $|a(s, x; \theta)|$ *denotes the determinant of* $a(s, x; \theta)$, *and for* $N \geqslant 2$ *we have for any version of* $p^{(1)}(s, x, t, \cdot\,; \theta)$ *the expression*

$$p^{(N)}(s, x, t, y; \theta) = \int_{\mathbf{R}^{d(N-1)}} \prod_{k=1}^{N} p^{(1)}(\tau_{k-1}, \xi_{k-1}, \tau_k, \xi_k; \theta)\, d\xi_1 \ldots d\xi_{N-1} \tag{6}$$

$$= E_{P_{0,s,x}}(p^{(1)}(\tau_{N-1}, Y_{\tau N-1}^{(N)}, t, y; \theta)), \tag{7}$$

with $\xi_0 = x$ *and* $\xi_N = y$.

Proof. Follows from the Markov property of the Markov chain $\{Y_{\tau_k}^{(N)}\}_{k=0}^{N}$ under $P_{0,s,x}$ and the Chapman–Kolmogorov equations. □

Notice that $l_n^{(N)}(\theta)$ is defined with no restrictions on the parameter dependence of $\sigma(\cdot, \cdot\,; \theta)$. In fact $l_n^{(1)}(\theta)$ can in this sense be seen as a generalization of $\tilde{l}_n(\theta)$. Indeed,

$$\begin{aligned} l_n^{(1)}(\theta) = & -\frac{nd}{2}\log(2\pi) - \frac{d}{2}\sum_{i=1}^{n}\log(t_i - t_{i-1}) - \frac{1}{2}\sum_{i=1}^{n}\log(|a(t_{i-1}, X_{t_{i-1}}; \theta)|) \\ & -\frac{1}{2}\sum_{i=1}^{n}(X_{t_i} - X_{t_{i-1}})^* a(t_{i-1}, X_{t_{i-1}}; \theta)^{-1}(X_{t_i} - X_{t_{i-1}})(t_i - t_{i-1})^{-1} \\ & + \sum_{i=1}^{n} b(t_{i-1}, X_{t_{i-1}}; \theta)^* a(t_{i-1}, X_{t_{i-1}}; \theta)^{-1}(X_{t_i} - X_{t_{i-1}}) \\ & -\frac{1}{2}\sum_{i=1}^{n} b(t_{i-1}, X_{t_{i-1}}; \theta)^* a(t_{i-1}, X_{t_{i-1}}; \theta)^{-1} b(t_{i-1}, X_{t_{i-1}}; \theta)(t_i - t_{i-1}), \end{aligned}$$

so when $\sigma(\cdot, \cdot\,; \theta) = \sigma(\cdot, \cdot)$ is independent of θ we see that

$$l_n^{(1)}(\theta) = K_n + \tilde{l}_n(\theta),$$

where K_n is some random variable that does not depend on θ. For the Ornstein–Uhlenbeck process with time-equidistant observations, the $l_n^{(1)}(\theta)$-estimators are $\hat{\phi}_n^{(1)} = \tilde{\phi}_n$ given by (2) and

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

which converges in probability to

$$\sigma_0^2 \frac{3(1 - \exp(\Delta\phi_0))^2}{-2\Delta\phi_0} \tag{8}$$

as $n \to \infty$. The estimator $\hat{\sigma}_n^{2(1)}$ can be thought of as an improvement of $\tilde{\sigma}_n^2$ given by (3) that corrects for the influence of the drift function $b(x; \theta) = \theta x$, and has also been suggested by Yoshida (1990).

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

The approximate log-likelihood functions $l_n^{(N)}(\theta)$ have been derived under the assumption that every time-interval $[t_{i-1}, t_i]$ is divided into N intervals, but if the observation time-points are non-equidistant one might want to choose a larger N for wider time-intervals $[t_{i-1}, t_i]$. More generally we may choose an $N_i \in N$ for each time-interval $[t_{i-1}, t_i]$, thus obtaining the approximate log-likelihood functions

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

where $\underline{N} = (N_1, \ldots, N_n)$.

3. Convergence of the approximate log-likelihood functions

The purpose of the approximate log-likelihood functions in practice is to serve as a tool for calculating approximations to the maximum likelihood estimator $\hat{\theta}_n$ of θ, when $l_n(\theta)$ is unknown. For given observations of X at time-points $0 = t_0 < t_1 < \cdots < t_n$ we may choose any of the approximate log-likelihood functions $l_n^{(N)}(\theta)$ as a substitute for $l_n(\theta)$. In this section we show that $l_n^{(N)}(\theta)$ will indeed be close to $l_n(\theta)$ for large values of N, that is we prove two results on the convergence of the approximating densities $p^{(N)}(s, x, t, y; \theta)$ to $p(s, x, t, y; \theta)$ as $N \to \infty$, and as a consequence we get that $l_n^{(N)}(\theta)$ converges to $l_n(\theta)$ in probability under P_{θ_0} as $N \to \infty$ for all $\theta \in \Theta$ and $n \in \mathbf{N}$.

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 of the densities $p^{(N)}(s, x, t, \cdot\,; \theta)$ 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 example 1 below) but in general it is a delicate matter. We can, however, prove the $L^1(\lambda^d)$-convergence of $p^{(N)}(s, x, t, \cdot\,; \theta)$ to $p(s, x, t, \cdot\,; \theta)$ as $N \to \infty$, thus avoiding the problem of choosing definitive versions.

Theorem 2

In addition to (A1) and (A2) assume for all $\theta \in \Theta$ that:

(i) $(t, x) \mapsto b(t, x; \theta)$ is continuous.

(ii) $a(t, x; \theta) \equiv a(\theta)$ is positive definite.

Then $p(s, x, t, y; \theta)$ exists, and for all $0 \leqslant s < t$, $x \in \mathbf{R}^d$ and $\theta \in \Theta$

$$p^{(N)}(s, x, t, \cdot\,; \theta) \to p(s, x, t, \cdot\,; \theta)$$

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

Proof. Under the assumptions of the theorem, (A1)–(A3) are satisfied for the pairs (b, a) and $(0, a)$, and (see section 2) we have the corresponding families of probability measures $\{P_{\theta, s, x}\}$, $\{Q_{\theta, s, x}\}$ on $(\Omega, \mathscr{F})$. Now let $0 \leqslant s < t$, $x \in \mathbf{R}^d$ and $\theta \in \Theta$ be fixed and define for each $N \in \mathbf{N}$

$$X^{(N)} = (X_{\tau 1}, \ldots, X_{\tau N} = X_t),$$

where the τ_ks are defined by (4). For $k = 1, 2, \ldots, N$

$$X_{\tau_k} = X_{\tau_{k-1}} + a(\theta)^{1/2}(\overline{W}^{\theta, s}_{\tau_k} - \overline{W}^{\theta, s}_{\tau_{k-1}}),$$

where (under $Q_{\theta, s, x}$)

$$\overline{W}^{\theta, s}_t = a(\theta)^{-1/2}(X_t - x), \quad t \geqslant s$$

is a d-dimensional Wiener process after time s. Denote by $\varphi_d(y; \mu, \Sigma)$ the density (with respect to λ^d) of the d-dimensional normal distribution $N_d(\mu, \Sigma)$ with mean value μ and

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

positive definite covariance matrix Σ. Using the Markov property of $\{X_{\tau_k}^{(N)}\}_{k=1}^{N}$ and $\{Y_{\tau_k}^{(N)}\}_{k=1}^{N}$ under $Q_{0,s,x}$ and $P_{0,s,x}$ respectively, we find that:

$$\frac{dQ_{0,s,x}\cdot X^{(N)}}{d\lambda^{dN}}(x_1,\ldots,x_N)=\prod_{k=1}^{N}\varphi_d\left(x_k;x_{k-1},\frac{t-s}{N}a(\theta)\right)>0$$

$$\frac{dP_{0,s,x}\cdot Y^{(N)}}{d\lambda^{dN}}(x_1,\ldots,x_N)=\sum_{k=1}^{N}\varphi_d\left(x_k;x_{k-1}+\frac{t-s}{N}b(\tau_{k-1},x_{k-1};\theta),\frac{t-s}{N}a(\theta)\right)>0,$$

where $x_0=x$, and $Y^{(N)}$ is defined by (5). Consequently, $P_{0,s,x}\cdot Y^{(N)}\sim Q_{0,s,x}\cdot X^{(N)}$, and by direct calculations we see that

$$\frac{dP_{0,s,x}\cdot Y^{(N)}}{dQ_{0,s,x}\cdot X^{(N)}}(x_1,\ldots,x_N)=\exp\left(\sum_{k=1}^{N}b(\tau_{k-1},x_{k-1};\theta)^*a(\theta)^{-1}(x_k-x_{k-1})\right.$$
$$\left.-\frac{1}{2}\sum_{k=1}^{N}b(\tau_{k-1},x_{k-1};\theta)^*a(\theta)^{-1}b(\tau_{k-1},x_{k-1};\theta)\frac{t-s}{N}\right).$$

Now it is well-known that under the assumptions of the theorem

$$Q_{0,s,x}|_{\mathscr{F}_t}\sim P_{0,s,x}|_{\mathscr{F}_t}.$$

Let

$$S=\left.\frac{dP_{0,s,x}}{dQ_{0,s,x}}\right|_{\mathscr{F}_t}=\exp\left(\int_s^t b(u,X_u;\theta)^*a(\theta)^{-1}\,dX_u-\frac{1}{2}\int_s^t b(u,X_u;\theta)^*a(\theta)^{-1}b(u,X_u;\theta)\,du\right),$$

and define for $N=1,2,\ldots$

$$S_N=\frac{dP_{0,s,x}\cdot Y^{(N)}}{dQ_{0,s,x}\cdot X^{(N)}}(X^{(N)}).$$

From (i) and (ii) it follows that $S_N\to S$ in probability (under $Q_{0,s,x}$) as $N\to\infty$ (see Jacod & Shiryaev, 1987, ch. 4; Revuz & Yor, 1991, ch. IV), and so we have a sequence $\{S_N\}_{N=1}^{\infty}$ of non-negative random variables on $(\Omega,\mathscr{F},Q_{0,s,x})$ that converges in probability to the random variable S as $N\to\infty$. Furthermore,

$$E_{Q_{0,s,x}}(S_N)=E_{Q_{0,s,x}}(S)=1$$

for all N, and so it follows from Dunford–Pettis' theorem that $S_N\to S$ in $L^1(Q_{0,s,x})$ as $N\to\infty$. Finally, since

$$X_t\sim N_d(x,(t-s)a(\theta))$$

under $Q_{0,s,x}$ it follows that

$$E_{Q_{0,s,x}}(S_N\mid X_t=\cdot)\varphi_d(\cdot\,;x,(t-s)a(\theta))\to E_{Q_{0,s,x}}(S\mid X_t=\cdot)\varphi_d(\cdot\,;x,(t-s)a(\theta))$$

in $L^1(\lambda^d)$ as $N\to\infty$ (see Hoffmann-Jørgensen, 1993). But

$$p^{(N)}(s,x,t,y;\theta)=E_{Q_{0,s,x}}(S_N\mid X_t=\cdot)\varphi_d(\cdot\,;x,(t-s)a(\theta))$$

and

$$p(s,x,t,y;\theta)=E_{Q_{0,s,x}}(S\mid X_t=\cdot)\varphi_d(\cdot\,;x,(t-s)a(\theta)).\qquad\square$$

Because $a(t,x;\theta)$ in theorem 2 was assumed to be independent of t and x it was possible to give a direct proof of the $L^1(\lambda^d)$-convergence of $p^{(N)}(s,x,t,\cdot\,;\theta)$ to $p(s,x,t,\cdot\,;\theta)$ as $N\to\infty$. When $a(t,x;\theta)$ is allowed to depend on t and x more advanced techniques are

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

needed. Let's consider the time-homogeneous case ($b(t, x; \theta) = b(x, \theta)$, $\sigma(t, x; \theta) = \sigma(x; \theta)$). From theorem 1 we have that

$$p^{(N)}(t, x, y; \theta) = E_{P_{\theta, x}}\left(p^{(1)}\left(\frac{t}{N}, Y^{(N)}_{(N-1)t/N}, y; \theta\right)\right),$$

where $p^{(1)}(t/N, u, \cdot\,; \theta)$ is the density of $N_d(u + (t/N)b(u; \theta), (t/N)a(u; \theta))$ with respect to λ^d. Loosely speaking this means that

$$p^{(1)}\left(\frac{t}{N}, u, y; \theta\right) \to \delta_y(u)$$

$$Y^{(N)}_{(N-1)t/N} \to X_t$$

as $N \to \infty$, where δ_y is the Dirac delta function, and so

$$p^{(N)}(t, x, y; \theta) \to E_{P_{\theta, x}}(\delta_y(X_t)) \tag{9}$$

as $N \to \infty$, where the expectation in (9) has to be understood properly. Ikeda & Watanabe (1989) actually prove existence theorems for $p(t, x, y; \theta)$ by showing that it is the generalized expectation of the generalized Wiener functional $\delta_y(X_t)$. Generalized expectations and generalized Wiener functionals are fundamental concepts in Malliavin calculus, and so it is not surprising that the proof of the next result is based on Malliavin calculus.

Theorem 3

Assume for all $\theta \in \Theta$ that $b(\cdot\,; \theta)\colon \mathbf{R}^d \mapsto \mathbf{R}^d$ and $\sigma(\cdot\,; \theta)\colon \mathbf{R}^d \mapsto M^{d \times r}$ are bounded with bounded derivatives of any order. Furthermore, assume that $a(\cdot\,; \theta) = \sigma(\cdot\,; \theta)\sigma(\cdot\,; \theta)^*$ is strongly positive definite, that is there exists an $\varepsilon(\theta) > 0$ such that $a(x; \theta) - \varepsilon(\theta)I_d$ is non-negative definite for all $x \in \mathbf{R}^d$. Then $p(t, x, y; \theta)$ exists, and for fixed $t \geqslant 0$, $x \in \mathbf{R}^d$ and $\theta \in \Theta$

$$p^{(N)}(t, x, \cdot\,; \theta) \to p(t, x, \cdot\,; \theta)$$

in $L^1(\lambda^d)$ as $N \to \infty$. Furthermore,

$$p(t, x, y; \theta) = \liminf_N p^{(N)}(t, x, y; \theta)$$

for λ^d-almost all $y \in \mathbf{R}^d$, and so if $p^{(N)}(t, x, \cdot\,; \theta)$ converges pointwise, then it converges to $p(t, x, \cdot\,; \theta)$.

Proof. The proof is based on Malliavin calculus and presumes a good acquaintance with the book by Bell (1987). The proof is given in the appendix. □

The assumptions in theorem 3 are very restrictive but there are several possibilities for improvements. It should be possible to replace the boundedness conditions on $b(\cdot\,; \theta)$ and $\sigma(\cdot\,; \theta)$ by conditions such as (A1) and (A2). Only a few derivatives of $b(\cdot\,; \theta)$ and $\sigma(\cdot\,; \theta)$ are really needed, as will be clear from the proof. Furthermore, it should be straightforward to extend the results to the time-inhomogeneous case. Finally, the assumption of strongly positive definiteness of $a(\cdot\,; \theta)$ can be considerably weakened, however not by conditions that are easy to check. What is really needed is that the Malliavin covariance matrix (for fixed t, x and θ) is invertible with a sufficiently integrable inverse. This is implied by the very weak (but abstract) Hörmander condition, which in turn is implied by the strongly positive definiteness of $a(\cdot\,; \theta)$ (see Bell, 1987).

We now give our main justification for using $l_n^{(N)}(\theta)$ as a substitute for $l_n(\theta)$ for large values of N, namely

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

Theorem 4
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 $0 \leqslant s < t$, $x \in \mathbf{R}^d$ and $\theta \in \Theta$, then $l_n^{(N)}(\theta) \to l_n(\theta)$ in probability under P_{θ_0} as $N \to \infty$ for all $\theta \in \Theta$ and $n \in \mathbf{N}$.

Proof. For fixed $0 \leqslant s < t$ and $\theta \in \Theta$ we prove that

$$p^{(N)}(s, X_s, t, X_t; \theta) \to p(s, X_s, t, X_t; \theta)$$

in probability under P_{θ_0} as $N \to \infty$. Clearly it is enough to prove that

$$p^{(N)}(s, X_s, t, X_t; \theta) \to p(s, X_s, t, X_t; \theta)$$

in probability under $P_{\theta_0, s, x}$ as $N \to \infty$ for (P_{θ_0}-almost) all $x \in \mathbf{R}^d$. Now choose any subsequence $N_k \to \infty$ and an $x \in \mathbf{R}^d$. Then there exists a further subsequence $N_{k_i}^x \to \infty$ and a Borel set $\mathcal{N}_x \subseteq \mathbf{R}^d$ such that $\lambda^d(\mathcal{N}_x) = 0$ and

$$p^{(N_{k_i}^x)}(s, x, t, y; \theta) \to p(s, x, t, y; \theta)$$

as $i \to \infty$ for all $y \in \mathbf{R}^d \backslash \mathcal{N}_x$. But since $P_{\theta_0, s, x} \cdot X_t$ is absolutely continuous with respect to λ^d we see that

$$p^{(N_{k_i}^x)}(s, X_s, t, X_t; \theta) \to p(s, X_s, t, X_t; \theta)$$

$P_{\theta_0, s, x}$-almost surely as $i \to \infty$. Consequently,

$$p^{(N)}(s, X_s, t, X_t; \theta) \to p(s, X_s, t, X_t; \theta)$$

in probability under $P_{\theta_0, s, x}$ as $N \to \infty$. □

Another sequence of approximations to $p(s, x, t, y; \theta)$ can be based on the following theorem.

Theorem 5
In addition to (A1)–(A3) assume for all $\theta \in \Theta$ that:

(i) $(t, x) \mapsto b(t, x; \theta)$ and $(t, x) \mapsto \sigma(t, x; \theta)$ are continuous.

(ii) There exists a mapping $G(\cdot, \cdot\,; \theta)\colon [0, \infty) \times \mathbf{R}^d \mapsto \mathbf{R}^d$ with continuous partial derivatives

$$\frac{\partial G_k}{\partial t}, \frac{\partial G_k}{\partial x_i}, \frac{\partial^2 G_k}{\partial x_i\, \partial x_j}, \quad \forall i, j, k$$

such that

$$a(\cdot, \cdot\,; \theta)^{-1/2} = \frac{\partial G}{\partial x}(\cdot, \cdot\,; \theta) \quad \text{and} \quad x \mapsto G(t, x; \theta)$$

is a bijection for all $t \geqslant 0$. Then $p(s, x, t, y; \theta)$ exists and is given by the expression

$$\begin{aligned} p(s, x, t, y; \theta) = {} & (2\pi(t-s))^{-d/2} |a(t, y; \theta)|^{-1/2} \\ & \times \exp\left(-\frac{1}{2(t-s)} \|G(t, y; \theta) - G(s, x; \theta)\|^2\right) \\ & \times E\left\{\exp\left[\int_s^t B(u, V_u^{G(s, x; \theta), G(t, y; \theta)}; \theta)^* \, dV_u^{G(s, x; \theta), G(t, y; \theta)} \right.\right. \\ & \left.\left. - \frac{1}{2}\int_s^t \|B(u, V_u^{G(s, x; \theta), G(t, y; \theta)}; \theta)\|^2 \, du\right]\right\}, \end{aligned}$$

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

where $(G(u, z; \theta) = v)$

$$B(u, v; \theta) = a(u, z; \theta)^{-1/2} b(u, z; \theta) + \frac{\partial G}{\partial t}(u, z; \theta)$$
$$+ \left\{ \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2 G_k}{\partial x_i \, \partial x_j}(u, z; \theta) a_{ij}(u, z; \theta) \right\}_{k=1}^{d}$$

and $(V_u^{\alpha, \beta})_{s \leqslant u \leqslant t}$ *is a d-dimensional Brownian bridge with* $V_s^{\alpha, \beta} = \alpha$ *and* $V_t^{\alpha, \beta} = \beta$.

Proof. This theorem is a straightforward generalization of a one-dimensional result in Dacunha-Castelle & Florens-Zmirou (1986) so the proof will be sketchy. First assume that $a(t, x; \theta) \equiv I_d$ and calculate $p(s, x, t, y; \theta)$ by using the fact that the conditional distribution under $Q_{0, s, x}$ of $(X_u)_{s \leqslant u \leqslant t}$ given $X_t = y$ is the distribution of $(V_u^{x, y})_{s \leqslant u \leqslant t}$. Then reduce the general case to this case by applying Itô's formula to the transformation $U_t^0 = G(t, X_t; \theta)$. The assumptions on G now guarantee that the usual theorem on transformation of densities in $\mathbf{R}^d$ can be applied to the transformation $X_t = G^{-1}(t, U_t^0; \theta)$, from which the expression for $p(s, x, t, y; \theta)$ follows, at least for $y \in D$, where $D \subseteq \mathbf{R}^d$ is an open subset for which $P_{0, s, x}(X_t \in D) = 1$. □

The expression for $p(s, x, t, y; \theta)$ in theorem 5 can be used to derive the following sequence of approximations to $l_n(\theta)$. As with $l_n^{(N)}(\theta)$ the procedure is to approximate $p(s, x, t, y; \theta)$ by some functions $q^{(N)}(s, x, t, y; \theta)$ for $N \in \mathbf{N}$, and then to define

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

for $N = 1, 2, \ldots$. Since $(t, x) \mapsto B(t, x; \theta)$ is continuous for all θ the integrals in $p(s, x, t, y; \theta)$ can be approximated by Riemann sums. For fixed $0 \leqslant s < t$, $x, y \in \mathbf{R}^d$, $\theta \in \Theta$ and $N \in \mathbf{N}$

$$\int_s^t B(u, V_u^{G(s, x; \theta), G(t, y; \theta)}; \theta)^* \, dV_u^{G(s, x; \theta), G(t, y; \theta)}$$
$$\simeq \sum_{k=1}^{N} B(\tau_{k-1}, V_{\tau_{k-1}}^{G(s, x; \theta), G(t, y; \theta)}; \theta)^* (V_{\tau_k}^{G(s, x; \theta), G(t, y; \theta)} - V_{\tau_{k-1}}^{G(s, x; \theta), G(t, y; \theta)}),$$

and similarly for the Lebesgue integral. Even though the functions $q^{(N)}(s, x, t, y; \theta)$ thus obtained may not be densities the functions $h_n^{(N)}(\theta)$ will of course be close to $l_n(\theta)$ for N large. However, we shall not study these approximations any further since any attempt to calculate $h_n^{(N)}(\theta)$ (see section 4) will require explicit expressions for $G(\cdot, \cdot; \theta)$ and $G^{-1}(t, \cdot; \theta)$, which are only available in some special cases. Notice however that closed expressions for $G(\cdot, \cdot; \theta)$ and $G^{-1}(t, \cdot; \theta)$ are given in the setting of theorem 2.

4. Estimation by means of the approximate log-likelihood functions: some considerations and two examples

Employing $l_n^{(N)}(\theta)$ in some maximization procedure means that we have to calculate it in a finite number of points $\theta \in \Theta$. For $N = 1$ we have a closed expression for $l_n^{(N)}(\theta)$ but for $N \geqslant 2$ this is in general not the case. In this section we propose a general method for calculating $l_n^{(N)}(\theta)$ for $N \geqslant 2$, and apply the approximate log-likelihood functions in two numerical examples.

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

To calculate $l_n^{(N)}(\theta)$ in general we must be able to calculate $p^{(N)}(s, x, t, y; \theta)$ for all $0 \leqslant s < t$, $x, y \in \mathbf{R}^d$ and $\theta \in \Theta$. For large values of N the Lebesgue-integral form (6) of $p^{(N)}(s, x, t, y; \theta)$ is of no use, but expression (7)

$$p^{(N)}(s, x, t, y; \theta) = E_{P_{0,s,x}}(p^{(1)}(\tau_{N-1}, Y^{(N)}_{\tau_{N-1}}, t, y; \theta))$$

enables us to calculate $p^{(N)}(s, x, t, y; \theta)$ by appealing to the Strong Law of Large Numbers. Let $\{U_k^m\}_{k=1, m=1}^{N-1, M}$ be an i.i.d. sample from the r-dimensional standard normal distribution. Then $\{Y^m\}_{m=1}^M = \{Y_{N-1}^m\}_{m=1}^M$ given by

$$Y_0^m = x, m = 1, \ldots, M$$

$$Y_k^m = Y_{k-1}^m + \frac{t-s}{N} b(\tau_{k-1}, Y_{k-1}^m; \theta) + \sqrt{\frac{t-s}{N}} \sigma(\tau_{k-1}, Y_{k-1}^m; \theta) U_k^m$$

for $k = 1, \ldots, N-1$ and $m = 1, \ldots, M$ has the same distribution as an i.i.d. sample of $Y^{(N)}_{\tau_{N-1}}$ under $P_{0,s,x}$. Choosing the integer M sufficiently large we can calculate $p^{(N)}(s, x, t, y; \theta)$ with any accuracy by

$$\frac{1}{M} \sum_{m=1}^{M} p^{(1)}(\tau_{N-1}, Y^m, t, y; \theta).$$

The sample $\{U_k^m\}_{k=1, m=1}^{N-1, M}$ may of course be used to calculate $p^{(N)}(s, x, t, y; \theta)$ for all values of $0 \leqslant s < t$, $x, y \in \mathbf{R}^d$ and $\theta \in \Theta$. In applications this means that the U_k^ms can be simulated once and for all (see Kloeden & Platen, 1992), stored in the computer, and then be used to calculate $l_n^{(N)}(\theta)$ at any point $\theta \in \Theta$. Initial values of θ for a numerical maximization of $l_n^{(N)}(\theta)$ can be found by maximizing $l_n^{(1)}(\theta)$.

Whenever the mappings $G(\cdot, \cdot; \theta)$ and $G^{-1}(t, \cdot; \theta)$ in theorem 5 are known, the functions $h_n^{(N)}(\theta)$ can be calculated in a similar way, since independent replicates of a d-dimensional Brownian bridge (see Karatzas & Shreve, 1988) at discrete time-points are easily simulated.

Example 1. For the Ornstein-Uhlenbeck process we have continuous versions

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

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

of $p^{(N)}(t, x, \cdot; \theta)$ and $p(t, x, \cdot; \theta)$ respectively, where

$$\beta_N = \left(1 + \frac{\Delta\phi}{N}\right)^N \to \beta = \exp(\Delta\phi)$$

$$\tau_N^2 = \sigma^2 \frac{\Delta}{N} \frac{1 - \beta_N^2}{1 - \beta_N^{2/N}} \to \tau^2 = \sigma^2 \frac{1 - \exp(2\Delta\phi)}{-2\phi}$$

as $N \to \infty$ for all $\phi < 0$ and $\sigma^2 > 0$. Thus, the continuous version of $p^{(N)}(t, x, \cdot; \theta)$ converges pointwise to the continuous version of $p(t, x, \cdot; \theta)$ as $N \to \infty$. Since

$$\psi_n = \sum_{i=1}^{n} X_{i\Delta} X_{(i-1)\Delta} \Big/ \sum_{i=1}^{n} X_{(i-1)\Delta}^2 \to \exp(\Delta\phi_0)$$

P_{θ_0}-almost surely as $n \to \infty$ we have for sufficiently large values of n the following expressions

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

for the maximum likelihood estimators and the $l_n^{(N)}(\theta)$-estimators of θ:

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

$$\hat{\sigma}_n^2 = \frac{-2\hat{\phi}_n}{n(1-\exp(2\Delta\hat{\phi}_n))}\sum_{i=1}^{n}(X_{i\Delta} - X_{(i-1)\Delta}\exp(\Delta\hat{\phi}_n))^2$$

$$\hat{\phi}_n^{(N)} = \frac{N}{\Delta}(\psi_n^{1/N} - 1)$$

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

where

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

From these expressions and the fact that $-N(1-x^{1/N}) \to \log(x)$ as $N \to \infty$ for all $x > 0$, we deduce for sufficiently large values of n that

$$\hat{\theta}_n^{(N)} = \begin{pmatrix} \hat{\phi}_n^{(N)} \\ \hat{\sigma}_n^{2(N)} \end{pmatrix} \to \begin{pmatrix} \hat{\phi}_n \\ \hat{\sigma}_n^2 \end{pmatrix} = \hat{\theta}_n$$

P_{θ_0}-almost surely as $N \to \infty$. Now it is well-known that $\hat{\theta}_n$ is 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

$$i(\theta_0, \Delta)^{-1} = \begin{Bmatrix} \dfrac{1-\exp(2\phi_0\Delta)}{\Delta^2\exp(2\phi_0\Delta)} & \dfrac{2\sigma_0^2}{\Delta} + \dfrac{\sigma_0^2}{\phi_0\Delta^2}\dfrac{1-\exp(2\phi_0\Delta)}{\exp(2\phi_0\Delta)} \\ \dfrac{2\sigma_0^2}{\Delta} + \dfrac{\sigma_0^2}{\phi_0\Delta^2}\dfrac{1-\exp(2\phi_0\Delta)}{\exp(2\phi_0\Delta)} & \dfrac{\sigma_0^4}{\phi_0^2\Delta^2}\dfrac{1-\exp(2\phi_0\Delta)}{\exp(2\phi_0\Delta)} + \dfrac{4\sigma_0^4}{\phi_0\Delta} + \dfrac{2\sigma_0^4(1+\exp(2\phi_0\Delta))}{1-\exp(2\phi_0\Delta)} \end{Bmatrix}$$

is the inverse of the Fisher-information matrix. Since both convergence in probability and convergence in distribution can be viewed as convergences in metric spaces we may conclude 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 ($N'(n) \geqslant N(n)$) for all $n \in \mathbf{N}$), then the same results hold for this subsequence.

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

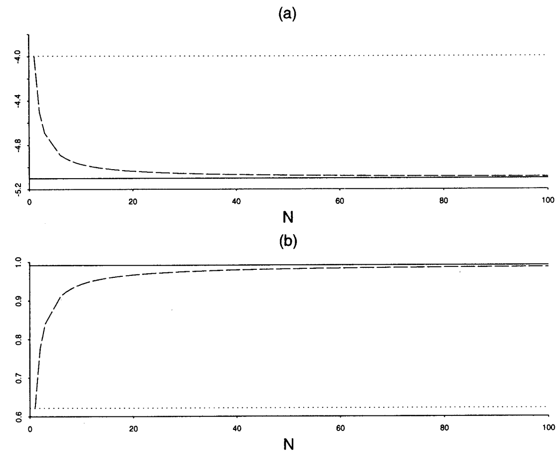

Fig. 1. (Example 1) (a): The two levels are $\hat{\phi}_n^{(1)}$ (the dotted line) and $\hat{\phi}_n$ (the solid line), and the dashed line is the sequence $\{\hat{\phi}_n^{(N)}\}_{N=1}^{100}$. (b): As in (a), but for the estimates of σ^2. The true values are $(\phi_0, \sigma_0^2) = (-5, 1)$.

For a typical simulation of $\{X_{i\Delta}\}_{i=0}^n$ with $(\phi_0, \sigma_0^2) = (-5, 1)$, $n = 1000$, $\Delta = 0.1$ and $x_0 = 0$ the estimates $(\hat{\phi}_n, \hat{\sigma}_n^2)$ and $(\hat{\phi}_n^{(N)}, \hat{\sigma}_n^{2(N)})$ for $N = 1, 2, \ldots, 500$ have been calculated. The results for $N \leqslant 100$ are shown in Fig. 1 since the estimates $(\hat{\phi}_n^{(N)}, \hat{\sigma}_n^{2(N)})$ are virtually equal to $(\hat{\phi}_n, \hat{\sigma}_n^2)$ for $N > 100$. Notice that the asymptotic values (as $n \to \infty$) of $\hat{\phi}_n^{(1)}(= -3.9955)$ and $\hat{\sigma}_n^{2(1)}(=0.6217)$ are respectively -3.9347 (equation (2)) and 0.4645 (equation (8)). It is remarkable how poor the estimators $(\hat{\phi}_n^{(1)}, \hat{\sigma}_n^{2(1)})$ are and how fast the estimators $(\hat{\phi}_n^{(N)}, \hat{\sigma}_n^{2(N)})$ approach $(\hat{\phi}_n, \hat{\sigma}_n^2)$.

Example 2. For the one-dimensional stochastic differential equation

$$dX_t = -\theta X_t\,dt + \theta\sqrt{1 + \frac{X_t^2}{1+X_t^2}}\;dW_t, \quad X_0 = 0, t \geq 0,$$

the log-likelihood function $l_n(\theta)$ is unknown, and the estimation methods in Florens-Zmirou (1989) and Yoshida (1990) do not apply since $b(\cdot\,; \theta)$ and $\sigma(\cdot\,; \theta)$ depend on the same

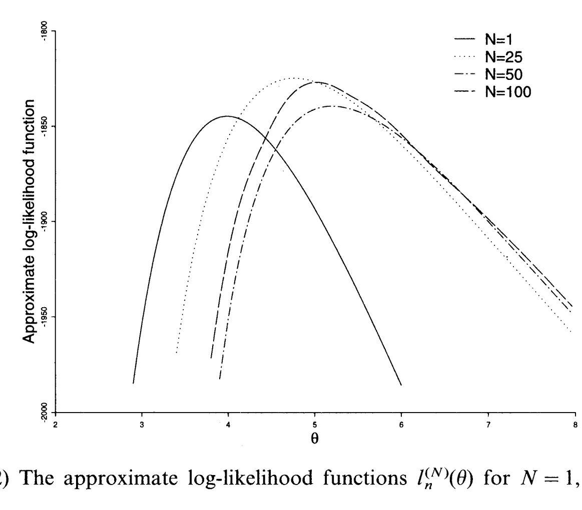

Fig. 2. (Example 2) The approximate log-likelihood functions $l_n^{(N)}(\theta)$ for $N = 1, 25, 50, 100$. The true value is $\theta_0 = 5$.

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

Table 1. *(Example 2) The estimates corresponding to the approximate log-likehood functions in Fig. 2*

N	1	25	50	100
$\hat{\theta}_n^{(N)}$	3.98	4.76	5.20	5.04

unknown parameter $\theta > 0$. The martingale problem for b and a is, however, well-posed for all $\theta > 0$ (see th. 34.1 in Rogers & Williams (1987)) and assumption (A3) is satisfied for all $\theta > 0$, so the approximate log-likelihood functions $\{l_n^{(N)}(\theta)\}_{N=1}^{\infty}$ can be used to estimate θ. For a typical simulation of $\{X_{i\Delta}\}_{i=0}^{n}$ (using the Milstein scheme with time-step $\Delta/1000$, see Kloeden & Platen (1992)) with $\theta_0 = 5$, $n = 1000$ and $\Delta = 0.1$ the approximate log-likelihood functions $l_n^{(N)}(\theta)$ have been calculated for $N = 1, 25, 50, 100$. A plot of these is given in Fig. 2 and the corresponding estimates are shown in Table 1.

Notice that the maximum quasi-likelihood estimator $\hat{\theta}_{t_n}^q$ of θ as defined by Hutton & Nelson (1986), in this case equals

$$\hat{\theta}_{t_n}^q = -\frac{\displaystyle\int_0^{t_n} \frac{X_s(1+X_s^2)}{1+2X_s^2}\,dX_s}{\displaystyle\int_0^{t_n} \frac{X_s^2(1+X_s^2)}{1+2X_s^2}\,ds}.$$

Approximating the integrals in $\hat{\theta}_{t_n}^q$ as usual we obtain the approximate maximum quasi-likelihood estimator $\tilde{\theta}_n^q$ of θ given by

$$\tilde{\sigma}_n^q = -\frac{\displaystyle\frac{1}{\Delta}\sum_{i=1}^{n} \frac{X_{(i-1)\Delta}(1+X_{(i-1)\Delta}^2)}{1+2X_{(i-1)\Delta}^2}(X_{i\Delta} - X_{(i-1)\Delta})}{\displaystyle\sum_{i=1}^{n} \frac{X_{(i-1)\Delta}^2(1+X_{(i-1)\Delta}^2)}{1+2X_{(i-1)\Delta}^2}}.$$

The same expression is obtained if $l_{t_n}^c(\theta)$ is approximated by the usual integral approximations and $\tilde{\theta}_n^q$ is defined to be the solution of the corresponding approximation to the equation $0 = l_{t_n}^c(\theta)$. For the given simulation $\tilde{\theta}_n^q = 3.8053$, which is worse than $\hat{\theta}_n^{(1)}$.

5. Concluding remarks

The approximate log-likelihood functions $\{l_n^{(N)}(\theta)\}_{N=1}^{\infty}$ have been derived as approximations to the log-likelihood function based on discrete observations, thus avoiding the inevitable problems encountered when estimation methods for discrete observations are based on the theory for continuous observation. This is the main difference between this and other approaches.

The basic idea behind the approximation $p^{(N)}(s, x, t, y; \theta)$ was to approximate X_t given $X_s = x$ by the Euler–Maruyama approximation. We could instead have used a stochastic Itô–Taylor approximation with a higher convergence rate with respect to time-discretization (see Kloeden & Platen, 1992), to get a faster convergence in N. However, in order to obtain an approximate transition density that can be calculated we need to know the transition densities of the approximating Markov chain, which is only the case for the Euler–Maruyama approximation. From these considerations we may, however, expect a faster convergence of $p^{(N)}(s, x, t, y; \theta)$ when $\sigma(t, x; \theta)$ does not depend on x, since the Euler–Maruyama approximation is then equal to the faster Milstein approximation.

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

In this paper we have focused on the derivation of the approximate log-likelihood functions, which was possible under weak and natural assumptions. No assumptions were made about the functional dependence of $b(t, x; \theta)$ and $\sigma(t, x; \theta)$ on θ, since we have not been concerned with the approximate log-likelihood functions as functions of θ. Consequently, no results were given on the existence and (asymptotic) properties of the estimator $\hat{\theta}_n^{(N)}$ obtained by maximizing $l_n^{(N)}(\theta)$, except for the Ornstein–Uhlenbeck process. Such results can be found in Pedersen (1993), where sufficient conditions for the convergence of $\hat{\theta}_n^{(N)}$ to $\hat{\theta}_n$ in probability under P_{θ_0} as $N \to \infty$ are given. If this convergence holds, then the consistency and asymptotic normality of $\hat{\theta}_n^{(N)}$, as for the Ornstein–Uhlenbeck process, is an immediate consequence of the consistency and asymptotic normality of $\hat{\theta}_n$. If the convergence of $\hat{\theta}_n^{(N)}$ to $\hat{\theta}_n$ cannot be established (even for large values of n) it is shown in Pedersen (1993) how consistency and asymptotic normality of $\hat{\theta}_n^{(N)}$ can still be proved.

Acknowledgement

I would like to thank Svend Erik Graversen for substantial help with the proof of theorem 3.

References

Bell, D. R. (1987). *The Malliavin calculus*. Wiley, New York.

Billingsley, P. (1961). *Statistical inference for Markov processes*. Chicago University Press, Chicago.

Dacunha-Castelle, D. & 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*, Vol. 1. Academic Press, New York.

Hoffmann-Jørgensen, J. (1993). *Probability with a view to statistics*. Van Nostrand, New York.

Hutton, J. E. & Nelson, P. I. (1986). Quasi-likelihood estimation for semimartingales. *Stochast. Process. Applic.* **22**, 245–257.

Ikeda, N. & Watanabe, S. (1989). *Stochastic differential equations and diffusion processes*, 2nd Edn. North-Holland, Amsterdam.

Jacod, J. & Shiryaev, A. N. (1987). *Limit theorems for stochastic processes*. Springer, Berlin.

Karatzas, I. & Shreve, S. E. (1988). *Brownian motion and stochastic calculus*. Springer, New York.

Kloeden, P. E. & Platen, E. (1992). *Numerical solution of stochastic differential equations*. Springer, New York.

Liptser, R. S. & Shiryayev, A. N. (1977). *Statistics of random processes I: general theory*. Springer, New York.

Pedersen, A. R. (1993). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. Research Report No. **273**, Department of Theoretical Statistics, University of Aarhus.

Revuz, D. & Yor, M. (1991). *Continuous martingales and Brownian motion*. Springer, New York.

Rogers, L. C. G. & Williams, D. (1987). *Diffusions, Markov processes, and martingales, Volume 2: Itô calculus*. Wiley, Chichester.

Stroock, D. W. & Varadhan, S. R. S. (1979). *Multidimensional diffusion processes*. Springer, Berlin.

Yoshida, N. (1990). Estimation for diffusion processes from discrete observations. Preprint.

Received June 1993, in final form December 1993.

Asger Roer Pedersen, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus, DK-8000 Aarhus C, Denmark and Department of Freshwater Ecology, National Environmental Research Institute, Vejlsovej 25, DK-8600, Silkeborg, Denmark.

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

Appendix: the proof of theorem 3

Under the assumptions of the theorem the stochastic differential equation (1) is pathwise unique and strong in the sense of Ikeda & Watanabe (1989), and the solution can be realized on the r-dimensional Wiener space $C_0([0, \infty), \mathbf{R}^d) = \{w \in C([0, \infty), \mathbf{R}^d) \mid w(0) = 0\}$ in a very strong way (see Ikeda & Watanabe, 1989). There exists (θ fixed) a mapping $(t, x, w) \mapsto X(t, x, w)$ from $[0, \infty) \times \mathbf{R}^d \times C_0([0, \infty), \mathbf{R}^d)$ into $\mathbf{R}^d$ such that $(t, w) \mapsto X(t, x, w)$ is a pathwise unique solution to (1) with initial condition $X_0 = x$, for all $x \in \mathbf{R}^d$. Obviously it is enough to consider the solution on $C_0([0, T], \mathbf{R}^d)$ for every $T > 0$, and as in Bell (1987) we take $T = 1$. Now let $0 < t < 1$, $x \in \mathbf{R}^d$ and $\theta \in \Theta$ be fixed. From now on we shall for convenience adopt the notation in Bell (1987) since the proof of theorem 3 makes extensive use of the results in Bell (1987). The Wiener measure on $C_0 = C_0([0, 1], \mathbf{R}^d)$ is denoted by γ and $g_t(w) = X(t, x, w)$ for all $w \in C_0$. *Let* $g_t^m(w)$ denote the mth Euler–Maruyama approximation of $g_t(w)$ for all $w \in C_0$ and put

$$\sigma^m(w) = Dg_t^m(w)Dg_t^m(w)^\dagger$$

for all $m \in \mathbf{N}$ and $w \in C_0$, where $Dg_t^m(w)$ is the differential operator for the smooth functional g_t^m and † denotes the adjoint of any linear map between the Cameron–Martin subspace of C_0 and $\mathbf{R}^d$ (see Ikeda & Watanabe, 1989). The strongly positive definiteness of $a(\cdot\,; \theta)$ implies (see Bell, 1987; Ikeda & Watanabe, 1989) that the Malliavin covariance matrix σ corresponding to g_t is invertible (γ-almost surely) and that its inverse is in $L^p(\gamma)$ for all $p \geqslant 1$. Consequently, the assumptions of th. 4.10 in Bell (1987) are fulfilled, thus $F(y) := p(t, x, y; \theta)$ exists and is C^∞. Moreover, we may use the results obtained in the proof of this theorem. For each $m \in \mathbf{N}$ define

$$C_0^m = \{w \in C_0 \mid \sigma^m(w) \text{ is invertible}\}.$$

Denote by $GL(d)$ the set of invertible real $d \times d$ matrices. Then since $GL(d) \subseteq M^{d\times d}$ is an open subset and $\sigma^m \to \sigma$ in γ-measure as $m \to \infty$ (see Bell, 1987) we have that $\gamma(C_0^m) \to 1$ as $m \to \infty$. As in Bell (1987) we define for all $N \in \mathbf{N}$

$$R_N(\alpha) = \begin{cases} \psi_N(\|\alpha^{-1}\|^2), & \alpha \in GL(d) \\ 0, & \alpha \in M^{d\times d} \backslash GL(d), \end{cases}$$

where $\{\psi_N\}_{N=1}^\infty$ is a sequence of smooth functions from $[0, \infty)$ into $[0, 1]$ such that the following two conditions are satisfied for all $N \in \mathbf{N}$:

(i) $\psi_N(u) = 1$ for $0 \leqslant u \leqslant N^2$ and $\psi_N(u) = 0$ for $u \geqslant (N+1)^2$.

(ii) $\sup_{N, u} |D^q \psi_N(u)| < \infty$ for all $q \in \mathbf{N}$.

Define further for all $m, N \in \mathbf{N}$

$$\gamma_N^m(dw) = R_N(\sigma^m(w))\gamma(dw)$$

$$\gamma_\infty^m(dw) = 1_{C_0^m}(w)\gamma(dw).$$

Since $0 \leqslant R_N(\sigma^m(w)) \leqslant 1$ for all $w \in C_0$, $R_N(\sigma^m(w)) = 0$ for all $w \notin C_0^m$ and $R_N(\sigma^m(w)) \to 1$ as $N \to \infty$ for all $w \in C_0^m$ we see that $\gamma_N^m \leqslant \gamma_\infty^m \leqslant \gamma$ and (by the Dominated Convergence theorem) that $\|\gamma_N^m - \gamma_\infty^m\| \to 0$ as $N \to \infty$ (convergence in variation). Define for all $m, N \in \mathbf{N}$

$$\nu = \gamma \cdot g_t$$

$$\nu^m = \gamma \cdot g_t^m$$

$$\nu_N^m = \gamma_N^m \cdot g_t^m.$$

$$\nu_\infty^m = \gamma_\infty^m \cdot g_t^m.$$

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.

Then we know that $v^m \ll \lambda^d$ with density $F^m(y) = p^{(m)}(t, x, y; \theta)$. Furthermore, from what we have just shown: $v_N^m \leqslant v_\infty^m \leqslant v^m$ and $\|v_N^m - v_\infty^m\| \to 0$ as $N \to \infty$, and it is easily seen that $\|v^m - v_\infty^m\| \to 0$ as $m \to \infty$. Now let $\phi: \mathbf{R}^d \mapsto$ be C^∞ with compact support and let $y_1, \ldots, y_b$ be unit vectors in $\mathbf{R}^d$ for some $b \in \mathbf{N}$. Define as in Bell (1987)

$$\phi_b(z) = D^b \phi(z)(y_1, \ldots, y_b)$$

for all $z \in \mathbf{R}^d$. Then

$$\int_{\mathbf{R}^d} D^b \phi(z)(y_1, \ldots, y_b) v_N^m(dz) = \int_{C_0} \phi_b(g_t^m(w)) R_N(\sigma^m(w)) \gamma(dw).$$

From the proof of theorem 4.10 in Bell (1987) follows that

$$\left| \int_{\mathbf{R}^d} D^b \phi(z)(y_1, \ldots, y_b) v_N^m(dz) \right| \leqslant \|\phi\|_\infty \cdot C_b,$$

where $0 < C_b < \infty$ is some constant independent of m and N. Consequently, since $\|v_N^m - v_\infty^m\| \to 0$ as $N \to \infty$ we have that

$$\left| \int_{\mathbf{R}^d} D^b \phi(z)(y_1, \ldots, y_b) v_\infty^m(dz) \right| \leqslant \|\phi\|_\infty \cdot C_b$$

for all $m \in \mathbf{N}$. This means the lem. 1.14 in Bell (1987) can be applied to the family $\{v_\infty^m\}_{m=1}^\infty$ of finite ($\leqslant 1$) Borel-measures on $\mathscr{B}(\mathbf{R}^d)$. In fact the same constants C_b can be used for all $m \in \mathbf{N}$, and we may conclude that $v_\infty^m \ll \lambda^d$ with C^∞-density F_∞^m. Moreover, for all $n > d$ there exists a constant $0 < A(n, d) < \infty$ depending only on n and d such that

$$\|D^{n-d-1} F_\infty^m\|_\infty \leqslant A(n, d) \cdot \max_{1 \leqslant i \leqslant n} C_i$$

for all $m \in \mathbf{N}$, Taking $n = d + 1$ and $n = d + 2$ we see that the family $\{F_\infty^m\}_{m=1}^\infty$ of densities of $\mathbf{R}^d$ is uniformly bounded and equicontinuous. Since $v^m \to v$ weakly as $m \to \infty$ and $\|v_\infty^m - v^m\| \to 0$ as $m \to \infty$ we have that $v_\infty^m \to v$ weakly as $m \to \infty$. This means that for all bounded and continuous functions $\psi: \mathbf{R}^d \mapsto \mathbf{R}$

$$\int_{\mathbf{R}^d} F_\infty^m(y) \psi(y)\, dy \to \int_{\mathbf{R}^d} F(y) \psi(y)\, dy \tag{10}$$

as $m \to \infty$. Since the family $\{F_\infty^m\}_{m=1}^\infty$ of densities on $\mathbf{R}^d$ is uniformly bounded and equicontinuous we know by the Arzela–Ascoli theorem that any subsequence of these has a further subsequence that converges locally uniformly to a limit $h: \mathbf{R}^d \mapsto \mathbf{R}$ as $m \to \infty$, so (10) implies that

$$\int_{\mathbf{R}^d} h(y) \psi(y)\, dy = \int_{\mathbf{R}^d} F(y) \psi(y)\, dy$$

for all continuous functions $\psi: \mathbf{R}^d \mapsto \mathbf{R}$ with compact support. Since h and F are continuous it is easily seen that $h = F$, thus F_∞^m converges locally uniformly to F as $m \to \infty$. Consequently, $F_\infty^m \to F$ in $L^1(\lambda^d)$ as $m \to \infty$ by use of Dunford-Pettis' theorem. Since $F_\infty^m \leqslant F^m$ for all $m \in \mathbf{N}$ and $\|v^m - v_\infty^m\| \to 0$ as $m \to \infty$ we have that

$$\int_{\mathbf{R}^d} |F_\infty^m(y) - F^m(y)|\, dy \to 0$$

as $m \to \infty$, thus $F^m \to F$ in $L^1(\lambda^d)$ as $m \to \infty$. The last claim in theorem 3 is now an easy consequence of Fatou's lemma. □

© Board of the Foundation of the Scandinavian Journal of Statistics 1995.