Estimation of Diffusion Parameters for Discretely Observed Diffusion Processes*

Helle Sørensen
Department of Statistics and Operations Research
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø, Denmark
E-mail: hsoeren@stat.ku.dk

January 2001

Abstract

We study estimation of diffusion parameters for one-dimensional, ergodic diffusion processes that are discretely observed. We discuss a method based on a functional relationship between the drift function, the diffusion function and the invariant density and use empirical process theory to show that the estimator is $\sqrt{n}$-consistent and in certain cases weakly convergent. The so-called CKLS model is used as an example and a numerical example is presented.

Keywords: CKLS model; diffusion parameters; ergodic diffusion processes; empirical process theory

*An extended version of this paper was printed as Paper II in Sørensen (2000).

1 Introduction

There is a vast literature on inference for diffusion processes observed at discrete points in time. Important early references are Dacunha-Castelle and Florens-Zmirou (1986) on the effect of discretization and Florens-Zmirou (1989) on simple Gaussian approximations. Later work include Bibby and Sørensen (1995) on martingale estimating functions; Pedersen (1995), Poulsen (1999) and Aït-Sahalia (1998) on advanced approximations to the likelihood; and Elerian, Chib and Shephard (2000) on Bayesian analysis.

One direction of research has been particularly concerned with estimation of the diffusion coefficient, and this paper is yet another contribution to that area. Useful references in parametric settings are Dohnal (1987) on the LAN/LAMN property of the model and lower bounds on the variance of estimators; Genon-Catalot and Jacod (1993), Jacod (1993) and Genon-Catalot and Jacod (1994) on the LAN/LAMN property, contrast estimation and, for the latter two, optimal random sampling times. Estimation in non-parametric models (with the diffusion coefficient either time or state dependent) has been considered by several authors as well: the estimators are based on kernel methods (Florens-Zmirou, 1993; Jiang and Knight, 1997; Soulier, 1998; Jacod, 2000) or wavelet methods (Genon-Catalot, Laredo and Picard, 1992; Hoffmann, 1997; Soulier, 1998; Hoffmann, 1999*a*; Hoffmann, 1999*b*).

The asymptotic results in all the papers mentioned in the previous paragraph concern sampling schemes where the final time-point of observation is fixed, say 1, and the process is observed more and more frequently. As opposed to this, the method from this paper provides consistent estimators for *any fixed sampling frequency and final sampling time increasing to infinity*. This asymptotic scheme is appropriate if, say, daily or weakly observations are available in a sampling period of increasing length.

The set-up is parametric but the estimation method is very much inspired by a non-parametric estimation procedure discussed by Aït-Sahalia (1996). Both methods rely on the following relation: if b is the drift function, σ the diffusion function, and μ the invariant density for a one-dimensional ergodic diffusion process with state space (l, r), then $2b\mu = (\sigma^2\mu)'$, i.e.

$$b(x) = \frac{1}{2}\left(\left(\sigma^2\right)'(x) + \sigma^2(x)\frac{\mu'(x)}{\mu(x)} \right), \quad x \in (l, r) \tag{1}$$

where a prime denotes differentiation with respect to the state variable.

Banon (1978) and Jiang and Knight (1997) use the relation for pointwise estimation of the drift b, plugging in suitable (kernel) estimates of σ^2 and its derivative. Aït-Sahalia (1996) uses the relation for estimation of σ^2,

rather than b. He assumes that $\sigma^2(x)\mu(x) \to 0$ as $x \to l$ so that

$$\sigma^2(x)\mu(x) = 2\int_l^x b(u)\mu(u)\,du, \quad x \in (l,r). \tag{2}$$

For each x, $\sigma^2(x)$ is then estimated by dividing a kernel estimate of twice the integral in (2) by a kernel estimate of $\mu(x)$. The out-coming non-parametric estimator of σ^2 is asymptotically well-behaved, but it is bound to be quite variable in areas with only few observations.

If a non-parametric analysis indicates a certain form of σ^2, then it is natural to specify the diffusion term parametrically and estimate the parameters. For a particular specification $x \to \sigma(x,\theta)$ of the diffusion term it is straight-forward to verify for which parameter values $\sigma^2(x,\theta)\mu(x,\theta)$ actually tends to zero as $x \to l$ such that (2) holds.

The aim of the present paper is to use the relation (2) — and a similar relation involving the integral $\int_x^r b(u)\mu(u,\theta)\,du$ — for parametric estimation. Loosely speaking, the idea is the following. Let $f = \sigma^2\mu$. As we shall see, it is easy for each x to define a consistent estimator $\hat{f}(x)$ of $f(x,\theta)$. We also have an analytical expression for $f(x,\theta)$, and we estimate θ such that the "theoretical" function $f(\cdot,\theta)$ is close to the estimated version $\hat{f}$ in the sense that the uniform distance $\sup_{x\in(l,r)}\left|f(x,\theta) - \hat{f}(x)\right|$ is minimal. The corresponding estimator is consistent under relatively weak regularity conditions (Theorem 4.2) and weakly convergent under somewhat stronger conditions (Theorem 4.7). The asymptotic results are proved by means of empirical process theory. The so-called CKLS model (Chan, Karolyi, Longstaff and Sanders, 1992) is used as an example, and the method seems to work well in a numerical study.

The paper is organized as follows. The model and the basic assumptions are presented in Section 2. We discuss the estimation approach in Section 3 and prove asymptotic properties in Section 4. The CKLS model is discussed in Section 5. Finally, conclusions are drawn in Section 6.

2 Model and notation

In this section we define the diffusion model and introduce notation used throughout the paper.

We consider a one-dimensional, time-homogeneous stochastic differential equation

$$dX_t = b(X_t)\,dt + \sigma(X_t,\theta)\,dW_t \tag{3}$$

where θ is an unknown p-dimensional parameter from the parameter space $\Theta \subseteq \mathbb{R}^p$, W is a one-dimensional Brownian motion and $b : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \times \Theta \to (0,\infty)$ are known continuous functions. Note that the drift function b does not depend on the parameter. We make the following assumptions.

Assumption 2.1 Assume that

1. the state space, denoted by I, is open and the same for all $\theta \in \Theta$;
2. for any $\theta \in \Theta$ there is a distribution $\mu_\theta(dx) = \mu(x,\theta)\,dx$ on I such that X is strictly stationary and ergodic if $X_0 \sim \mu_\theta$;
3. the drift function b is in $L^1(\mu_\theta)$ for all $\theta \in \Theta$. □

Since X is continuous, the state space I is an interval and we write $I = (l, r)$ where $-\infty \le l < r \le +\infty$. Simple integral conditions ensure stationarity, see Karlin and Taylor (1981, Section 15.6) or Karatzas and Shreve (1991, Section 5.5), for example: Define the scale density $s(\cdot,\theta)$ by $\log s(x,\theta) = -2\int_{x_0}^x b(u)/\sigma^2(u,\theta)\,du$ where $x_0 \in I$ is fixed but arbitrary. If $1/K_0(\theta) = \int_l^r (s(x,\theta)\sigma^2(x,\theta))^{-1}\,dx < +\infty$ and

$$\int_l^{x_0} s(x,\theta)\,dx = \int_{x_0}^r s(x,\theta)\,dx = +\infty \tag{4}$$

then Assumption 2.1.2 holds with

$$\mu(x,\theta) = K_0(\theta)\big(s(x,\theta)\sigma^2(x)\big)^{-1}, \quad (x,\theta) \in I \times \Theta. \tag{5}$$

In the following we let P_θ denote the distribution of X when $X_0 \sim \mu_\theta$ and E_θ the expectation with respect to P_θ. Under P_θ all $X_t \sim \mu_\theta$.

The objective of the paper is estimation of the parameter θ from observations $X_\Delta, \ldots, X_{n\Delta}$ at discrete, equidistant time-points. The estimation method described below is based on the function $f = \sigma^2\mu : I \times \Theta \to (0,\infty)$ which by (5) is given by

$$f_\theta(x) = f(x,\theta) = \frac{K_0(\theta)}{s(x,\theta)} = K_0(\theta)\exp\left(2\int_{x_0}^x \frac{b(u)}{\sigma^2(u,\theta)}\,du\right).$$

For θ fixed we will often write f_θ for the function $f(\cdot,\theta) : I \to (0,\infty)$. Differentiation of f with respect to x yields

$$\frac{\partial f}{\partial x} = 2f\frac{b}{\sigma^2} = 2\sigma^2\mu\frac{b}{\sigma^2} = 2b\mu, \tag{6}$$

and $f(x_0,\theta) = K_0(\theta)$ so $f(x,\theta) = K_0(\theta) + 2\int_{x_0}^x b(u)\mu(u,\theta)\,du$ for $x \in I$ and $\theta \in \Theta$. In particular, for θ fixed f_θ is bounded by $K_0(\theta) + 2\,\mathrm{E}_\theta\,|b(X_0)|$; the limits $f_\theta(l) = f(l,\theta) = \lim_{x\searrow l} f(x,\theta)$ and $f_\theta(r) = f(r,\theta) = \lim_{x\nearrow r} f(x,\theta)$ are well-defined and finite; and

$$f(x,\theta) = f(l,\theta) + 2\int_l^x b(u)\mu(u,\theta)\,du, \quad x \in I \tag{7}$$

$$f(x,\theta) = f(r,\theta) - 2\int_x^r b(u)\mu(u,\theta)\,du, \quad x \in I. \tag{8}$$

The limits $f(l,\theta)$ and $f(r,\theta)$ are non-negative for all $\theta \in \Theta$. For the estimation method below to work *at least one of the limits must be zero for all* $\theta \in \Theta$. Then f_θ is bounded by $2\,\mathrm{E}_\theta\,|b(X_0)|$. Note that (7) is identical to (2) if $f(l,\theta) = 0$.

Some comments: (*i*) If the integral conditions (4) hold then $f(l,\theta) = 0$ ($f(r,\theta) = 0$) holds automatically if $l > -\infty$ ($r < +\infty$). In particular $f(l,\theta) = 0$ for models with state space $(0,\infty)$. (*ii*) If $I = (-\infty,\infty)$ and $b \equiv 0$ so X is on natural scale, then f_θ is constant and the above integral assumption is *not* satisfied. (*iii*) If $b(x) = \alpha + \beta x$ where $\beta < 0$ *and* if the locale martingale part of X is a genuine martingale then $x \to x + \alpha/\beta$ is an eigenfunction for the conditional expectation operator and $f(l,\theta) = f(r,\theta) = 0$ holds automatically (Hansen, Scheinkman and Touzi, 1998, page 10). In particular $f(l,\theta) = f(r,\theta) = 0$ holds for the Ornstein-Uhlenbeck process and the Cox-Ingersoll-Ross model. (*iv*) Generally, one must check that $\int_l^{x_0} b(x)/\sigma^2(x,\theta)dx = +\infty$ for all $\theta \in \Theta$ and/or $\int_{x_0}^r b(x)/\sigma^2(x,\theta)dx = -\infty$ for all $\theta \in \Theta$. These integral conditions are easily checked as they only involve the drift and the diffusion functions.

3 Estimation

In this section we discuss how to define pointwise consistent estimators of $f_\theta = f(\cdot,\theta)$ and how to use them for estimation of θ. Asymptotic results for the estimators are proved in Section 4.

3.1 Basic ideas

If $f(l,\theta) = 0$ we see from (7) that

$$f(x,\theta) = 2\int_l^x b(u)\mu(u,\theta)\,du = 2\,\mathrm{E}_\theta\Big(b(X_0)1_{\{X_0\le x\}}\Big), \quad x \in I,\ \theta \in \Theta.$$

From the right hand side and Assumption 2.1 it follows that

$$\hat{f}_{1,n}(x) = \frac{2}{n}\sum_{i=1}^n \Big(b(X_{i\Delta})1_{\{X_{i\Delta}\le x\}}\Big) \tag{9}$$

is an unbiased and (strongly) consistent estimator of $f(x,\theta)$ with respect to P_θ for all $x \in I$: $\mathrm{E}_\theta\,\hat{f}_{1,n}(x) = f(x,\theta)$ and $\hat{f}_{1,n}(x) \to f(x,\theta)$ almost surely as $n \to \infty$. Also note that $\hat{f}_{1,n}(x) = 0 = f(l,\theta)$ for $x < \min\{X_{i\Delta} : i = 1,\dots,n\}$, hence we write $\hat{f}_{1,n}(l) = 0$.

Similarly, if $f(r,\theta) = 0$ then

$$\hat{f}_{2,n}(x) = -\frac{2}{n}\sum_{i=1}^n \Big(b(X_{i\Delta})1_{\{X_{i\Delta}> x\}}\Big) \tag{10}$$

is unbiased and (strongly) consistent for $f(x,\theta)$ under P_θ for all $x \in I$. We write $\hat{f}_{2,n}(r) = 0$ since $\hat{f}_{2,n}(x) = 0$ for $x \geq \max\{X_{i\Delta} : i = 1, \ldots, n\}$.

The estimated functions $\hat{f}_{1,n}$ and $\hat{f}_{2,n}$ are piecewise constant with jumps at each data point $X_{k\Delta}$; the jump size is $2b(X_{k\Delta})/n$. In particular $\hat{f}_{1,n}$ and $\hat{f}_{2,n}$ are increasing (decreasing) at $X_{k\Delta}$ if f_θ is increasing (decreasing) at $X_{k\Delta}$, cf. (6). Note that $\hat{f}_{1,n}(x) - \hat{f}_{2,n}(x) = \frac{2}{n}\sum_{i=1}^n b(X_{i\Delta})$ so the deviation between $\hat{f}_{1,n}(x)$ and $\hat{f}_{2,n}(x)$ is the same for all $x \in I$.

As indicated, the idea is to estimate θ by the value that makes the function f_θ close to its estimator, $\hat{f}_{1,n}$ or $\hat{f}_{2,n}$. To be specific we define the uniform distances

$$U_{i,n}(\theta) = \sup_{x\in I}\left|\hat{f}_{i,n}(x) - f_\theta(x)\right|, \quad i = 1, 2$$

and estimate θ by the value $\hat{\theta}_{1,n}$ that minimizes $U_{1,n}$ if $f(l,\theta) = 0$ and by the value $\hat{\theta}_{2,n}$ that minimizes $U_{2,n}$ if $f(r,\theta) = 0$. Note that $U_{i,n}(\theta)$ is finite since $U_{i,n}(\theta) \leq \frac{2}{n}\sum_{j=1}^n |b(X_{j\Delta})| + 2\mathrm{E}_\theta|b(X_0)|$. One could of course use other measures of distance between $\hat{f}_{i,n}$ and f_θ, such as (an approximation to) the L^2-distance. This will, however, not be discussed any further in this paper.

Now, what if both $f(l,\theta)$ and $f(r,\theta)$ are zero? Then (9) and (10) are both unbiased, consistent estimators of $f(x,\theta)$. Note that $\mathrm{E}_\theta\, b(X_0) = 0$ so $\hat{f}_{1,n}$ and $\hat{f}_{2,n}$ are close for n large; for a moderate size of n, like 500, it might however make a difference whether we use $\hat{f}_{1,n}$ or $\hat{f}_{2,n}$. Also note that either $\hat{f}_{1,n}$ or $\hat{f}_{2,n}$ becomes negative (close to r or l) whereas f_θ is positive on (l,r).

Instead of using either $\hat{f}_{1,n}$ or $\hat{f}_{2,n}$ we use a convex combination of the two. For $\lambda(x) = \big(\lambda_1(x), \lambda_2(x)\big)$ with $\lambda_1(x) + \lambda_2(x) = 1$, define $\hat{f}_{\lambda,n}$ by

$$\begin{aligned}\hat{f}_{\lambda,n}(x) &= \lambda_1(x)\hat{f}_{1,n}(x) + \lambda_2(x)\hat{f}_{2,n}(x)\\ &= \hat{f}_{1,n}(x) - \frac{2}{n}\lambda_2(x)\sum_{i=1}^n b(X_{i\Delta}).\end{aligned}$$

With this notation $\hat{f}_{\lambda,n} = \hat{f}_{1,n}$ for $\lambda \equiv (1,0)$ and $\hat{f}_{\lambda,n} = \hat{f}_{2,n}$ for $\lambda \equiv (0,1)$.

If $\lambda(x)$ is deterministic, then $\hat{f}_{\lambda,n}(x)$ is unbiased for $f(x,\theta)$ and it makes sense to choose $\lambda(x)$ such that the variance of $\hat{f}_{\lambda,n}(x)$ is minimal. In general it is not possible to calculate the variance of $\hat{f}_{\lambda,n}(x)$ since it involves the joint distribution of $X_{i\Delta}$ and $X_{j\Delta}$ for $i \neq j$ which we typically do not know. It is easy, however, to minimize an approximation to the variance: Straight-forward calculations show that *if* the observations $X_\Delta, \ldots, X_{n\Delta}$ were *independent* and identically μ_θ-distributed, then the smallest possible

value of $\mathrm{Var}_\theta\, \hat{f}_{\lambda,n}$ would be obtained for

$$\lambda_{\theta,1}(x) = \frac{V_{\theta,2}(x) + f^2(x,\theta)}{V_{\theta,1}(x) + V_{\theta,2}(x) + 2f^2(x,\theta)} = \frac{\mathrm{E}_\theta\, b^2(X_0) 1_{\{X_0 > x\}}}{\mathrm{E}_\theta\, b^2(X_0)}$$

$$\lambda_{\theta,2}(x) = 1 - \lambda_{\theta,1}(x) = \frac{\mathrm{E}_\theta\, b^2(X_0) 1_{\{X_0 \leq x\}}}{\mathrm{E}_\theta\, b^2(X_0)}.$$

Of course, the observations are *not* independent so these weights are only approximately optimal. Moreover, we do not know the expectations above, but we can use their empirical counterparts and consider

$$\hat{\lambda}_{1,n}(x) = \frac{\sum_{i=1}^n b^2(X_{i\Delta}) 1_{\{X_i > x\}}}{\sum_{i=1}^n b^2(X_{i\Delta})} \quad \text{and} \quad \hat{\lambda}_{2,n}(x) = \frac{\sum_{i=1}^n b^2(X_{i\Delta}) 1_{\{X_i \leq x\}}}{\sum_{i=1}^n b^2(X_{i\Delta})}.$$

In the following we write $\hat{f}_n(x) = \hat{f}_{\hat{\lambda}_n,n}(x)$ for the corresponding estimator.

For x close to l we have $\hat{\lambda}_1(x)$ close to 1 and hence $\hat{f}_n(x)$ close to $\hat{f}_{1,n}(x)$. Similarly $\hat{f}_n(x)$ is close to $\hat{f}_2(x)$ when x is close to r. In particular, $\hat{f}_n(x) = 0$ for x outside the range of the observations. Note that $\hat{f}_n(x)$ is consistent for $f(x,\theta)$ — even if b is not in $L^2(\mu_\theta)$, because $\hat{\lambda}_{1,n}$ and $\hat{\lambda}_{2,n}$ are bounded (by 1). However, $\hat{f}_n(x)$ can be biased although $\hat{f}_{1,n}(x)$ and $\hat{f}_{2,n}(x)$ are both unbiased.

For estimation of θ the idea is of course to minimize the uniform distance

$$U_n(\theta) = \sup_{x \in I} \left| \hat{f}_n(x) - f_\theta(x) \right|. \tag{11}$$

between $\hat{f}_n$ and f_θ. We let $\hat{\theta}_n$ denote the corresponding estimator.

3.2 Important comments

Below follow some important remarks on the three estimators of f_θ and the corresponding U-distances.

First an illustration of the difference between the three estimators of f_θ. Figure 1 shows graphs of $\hat{f}_{1,n}$, $\hat{f}_{2,n}$ and $\hat{f}_n$ for 100 hypothetical data points The data are simulated from the model $dX_t = (0.04 - 0.6X_t)\, dt + 0.2X_t^\gamma\, dW_t$ with true parameter value $\gamma_0 = 0.75$ and $\Delta = 1$. The model is discussed in detail in Section 5. For this particular simulation $\sum_{i=1}^n b(X_{i\Delta}) > 0$ so the graph of $\hat{f}_{1,n}$ lies over the graph of $\hat{f}_{2,n}$. The graph of $\hat{f}_n$ is in between; close to $\hat{f}_{1,n}$ for small data values and close to $\hat{f}_{2,n}$ for large data values. The figure also shows the graph of f corresponding to the true parameter value.

[Figure 1]

Second, note that neither $\hat{f}_{1,n}$, $\hat{f}_{2,n}$ or $\hat{f}_n$ would change if the order of the observations was changed. In other words, the observations are treated as if they were independent. This is of course unfortunate since they come from a diffusion model with built-in dependence. For "large" values of Δ

the dependence between observations is minor and we would thus expect the method to perform better for "large" values of Δ than for "small" values of Δ. Still, it turns out that the proposed estimators are consistent as $n \to \infty$ for *any* fixed value of $\Delta > 0$ (Section 4.1).

Third, a practical remark. Despite the definition of $U_n(\theta)$ as a supremum over the all of I, we can calculate $U_n(\theta)$ from the values of f_θ and $\hat{f}_n$ at data points and points where b is zero. To be specific, let $\tilde{X}_1 \leq \cdots \leq \tilde{X}_n$ be the observations ordered according to size and $\tilde{X}_0 = l$. Then, because f_θ is continuous and has a derivative with same sign as b, and because $\hat{f}_n$ is piecewise constant, $U_n(\theta) = \max(N_0, N_1, N_2)$ where

$$
\begin{aligned}
N_1 &= \max_{k=1,\ldots,n} \left| \hat{f}_n(\tilde{X}_k) - f_\theta(\tilde{X}_k) \right| \\
N_2 &= \max_{k=1,\ldots,n} \left| \hat{f}_n(\tilde{X}_{k-1}) - f_\theta(\tilde{X}_k) \right| \\
N_0 &= \sup_{x_0 : b(x_0)=0} \left| \hat{f}_n\big(\tilde{X}(x_0)\big) - f_\theta(x_0) \right|.
\end{aligned}
$$

In the latter expression $\tilde{X}(x_0) = \max_{k=0,\ldots,n}\{\tilde{X}_k : \tilde{X}_k \leq x_0\}$ is the largest observation smaller than x_0 (or l if all observations are larger than x_0). For the most commonly used models b is only zero at very few points. Of course similar formulas apply to $U_{1,n}(\theta)$ ($U_{2,n}(\theta)$) as long as $f(l,\theta) = 0$ ($f(r,\theta) = 0$) for all $\theta \in \Theta$; simply substitute $\hat{f}_n$ by $\hat{f}_{1,n}$ ($\hat{f}_{2,n}$) and remember also to compare $\hat{f}_{1,n}$ ($\hat{f}_{2,n}$) with f_θ at the endpoint r (l).

4 Asymptotic results

In this section we prove asymptotic results for the estimators $\hat{\theta}_{1,n}$, $\hat{\theta}_{2,n}$ and $\hat{\theta}_n$ obtained by minimizing the uniform distances $U_{1,n}$, $U_{2,n}$ and U_n respectively. It is implicitly assumed that the estimators exist (for n large enough).

4.1 Consistency

We first prove consistency. Let θ_0 be the true parameter value and let $U(\theta) = \sup_{x \in I} \left| f_\theta(x) - f_{\theta_0}(x) \right|$ denote the uniform distance between f_θ and f_{θ_0}. Then $U(\theta) = 0$ if and only if $\theta = \theta_0$ because f_θ and $f_{\theta'}$ are identical if and only if $\sigma(\cdot,\theta)$ and $\sigma(\cdot,\theta')$ are identical and because we do not allow parametrizations where $\sigma(\cdot,\theta) = \sigma(\cdot,\theta')$ for $\theta \neq \theta'$. We shall assume that θ_0 is well-separated as a minimum of U in following sense.

Assumption 4.1 For all $\delta > 0$ it holds that $C(\delta) > 0$ where

$$C(\delta) = \inf\{U(\theta) : ||\theta - \theta_0|| > \delta\}. \qquad \square$$

The assumption is for example satisfied (i) if $\theta \to f_\theta(x)$ is increasing or decreasing for all $x \in I$ which will often be the case (this makes sense for one-dimensional parameters only); or (ii) if U is continuous and Θ is compact; or (iii) if U is continuous and Θ is open with U bounded away from zero at the boundary.

Theorem 4.2 *Assume that Assumptions 2.1 and 4.1 hold and furthermore that b changes sign only finitely many times on I. If $f(l,\theta) = 0$ ($f(r,\theta) = 0$) for all $\theta \in \Theta$ then $\hat{\theta}_{1,n}$ ($\hat{\theta}_{2,n}$) is consistent for θ, and if $f(l,\theta) = f(r,\theta) = 0$ for all $\theta \in \Theta$ then $\hat{\theta}_n$ is consistent for θ as well.*

Proof It is sufficient to show that the uniform distances converge in P_{θ_0}-probability (or almost surely with respect to P_{θ_0}) to $U(\theta)$, uniformly in θ (van der Vaart and Wellner, 1996, Corollary 3.2.2).

First assume that $f(l,\theta) = 0$ for all $\theta \in \Theta$. By the triangle inequality for the uniform metric, it holds that $|U_{1,n}(\theta) - U(\theta)| \leq U_{1,n}(\theta_0)$ for all $\theta \in \Theta$ so it suffices to show

$$U_{1,n}(\theta_0) = \sup_{x \in I} \left| \hat{f}_{1,n}(x) - f_{\theta_0}(x) \right| \to 0 \tag{12}$$

P_{θ_0}-almost surely. Pointwise convergence follows from the ergodic theorem and Assumption 2.1.2. Recall that $\partial f_{\theta_0}/\partial x$ has same sign as b and that $\hat{f}_{1,n}$ is piecewise constant with jump size $b(X_{k\Delta})$ at $X_{k\Delta}$. Uniform convergence on each of the finitely many subintervals where f_{θ_0} and $\hat{f}_{1,n}$ are either non-increasing or non-decreasing now follows exactly as in the proof of the classical Glivenko-Cantelli theorem (Loève, 1963, page 20). Similarly if $f(r,\theta) = 0$ for all $\theta \in \Theta$. See Sørensen (2000, Section II.4) for more details.

Finally assume that $f(l,\theta) = f(r,\theta) = 0$ for all $\theta \in \Theta$. Recall that $\hat{f}_n(x) = \hat{f}_{1,n}(x) - \frac{2}{n}\hat{\lambda}_{2,n}(x) \sum_{i=1}^n b(X_{i\Delta})$ and $0 \leq \hat{\lambda}_{2,n}(x) \leq 1$. By the triangle inequality for the uniform metric,

$$\begin{aligned} \left|U_n(\theta) - U(\theta)\right| &\leq \sup_{x \in I} \left|\hat{f}_n(x) - f_{\theta_0}(x)\right| \\ &\leq \sup_{x \in I} \left|\hat{f}_{1,n}(x) - f_{\theta_0}(x)\right| + 2\left|\frac{1}{n}\sum_{i=1}^n b(X_{i\Delta})\right| \end{aligned}$$

which converges uniformly in θ to zero P_{θ_0}-almost surely since $\mathrm{E}_{\theta_0} b(X_0) = 0$. This proves consistency of $\hat{\theta}_n$. □

4.2 Rate of convergence of $\hat{\theta}_{1,n}$ and $\hat{\theta}_{2,n}$

In this section we show that $\sqrt{n}(\hat{\theta}_{1,n} - \theta_0)$ and $\sqrt{n}(\hat{\theta}_{2,n} - \theta_0)$ are stochastically bounded (Theorem 4.5). The similar result for $\hat{\theta}_n$ is proved in Section 4.3.

For simplicity we only list the assumptions for a one-dimensional parameter but the convergence results hold for multi-dimensional parameters under similar conditions. One of the conditions concerns the temporal dependence of X, expressed in terms of the β-mixing coefficients β_k, $k \geq 1$. As usual for stationary Markov processes β_k is defined by

$$\beta_k = \int \sup_A \left| p_{k\Delta,\theta_0}(x, A) - \mu_{\theta_0}(A) \right| d\mu_{\theta_0}(x)$$

where $p_{k\Delta,\theta_0}$ is the transition probability from time 0 to time $k\Delta$ and the supremum is taken over all Borel subsets of I.

Assumption 4.3 The true parameter value θ_0 is an inner point of $\Theta \subseteq \mathbb{R}$ and for any $x \in I$ the function $\theta \to f_\theta(x) = f(x, \theta)$ is continuously differentiable in a neighbourhood Θ_0 of θ_0 with first derivative $\dot{f}_\theta = \partial f_\theta / \partial\theta$ satisfying

1. $\dot{f}_{\theta_0}$ is continuous;
2. $\dot{f}_{\theta_0}$ is bounded, i.e. $\sup_{x\in I} |\dot{f}_{\theta_0}(x)| < \infty$;
3. $\sup_{x\in I} \left| \dot{f}_\theta(x) - \dot{f}_{\theta_0}(x) \right| \to 0$ as $\theta \to \theta_0$.

Furthermore,

4. $b \in L^p(\mu_{\theta_0})$ for some $p > 2$;
5. there exist constants $c_1 > 0$ and $0 < c_2 < 1$ such that the β-mixing coefficients for X satisfy $\beta_k \leq c_1 c_2^k$ for all $k \geq 1$. □

Note that conditions 4.3.2 and 4.3.3 imply continuity of U at θ_0.

We first prove uniform weak convergence of

$$M_{i,n}(h) = \sup_{x\in I} \left| n^{1/2} \big(\hat{f}_{i,n}(x) - f_{\theta_0 + h/\sqrt{n}}(x) \big) \right|, \quad h \in H$$

for any compact subset H of $\mathbb{R}$ (Proposition 4.4). We write $M_{i,n}(h) = \sup_{x\in I} \left| M'_{i,n}(x) - M''_n(h, x) \right|$ where

$$M'_{i,n}(x) = n^{1/2} \big(\hat{f}_{i,n}(x) - f_{\theta_0}(x) \big)$$
$$M''_n(h, x) = n^{1/2} \big(f_{\theta_0 + h/\sqrt{n}}(x) - f_{\theta_0}(x) \big)$$

and note that the processes M''_n and $M_{i,n}$ are well-defined for n large enough.

Recall that $|\hat{f}_{i,n}(x)| \leq \frac{2}{n} \sum_{j=1}^n |b(X_{j\Delta})|$ for all $x \in I$ and that $|f_\theta(x)| \leq 2\, \mathrm{E}_\theta\, |b(X_0)|$ for all $(x, \theta) \in I \times \Theta$. Hence, M''_n takes values in $l^\infty(H \times I)$ (since H is compact), $M'_{i,n}$ in $l^\infty(I)$, and thus $M_{i,n}$ in $l^\infty(H)$. Here we have used the notation $l^\infty(T)$ for the set of uniformly bounded, real functions on T; $l^\infty(T) = \{g : \sup_{t\in T} |g(t)| < \infty\}$.

The process M_n'' is non-stochastic and clearly $M_n''(h,x) \to \dot{f}_{\theta_0}(x)h$ pointwise as $n \to \infty$. Under Assumption 4.3.3 the convergence is suitably uniform. Moreover, it is well-known that the finite-dimensional marginals of M_n' converge in distribution to Gaussian limits (with quite complicated variance structure, however) see Florens-Zmirou (1989). As will be clear from below, conditions 4.3.4 and 4.3.5 ensure that the convergence is uniform implying uniform weak convergence of $M_{i,n}(h)$:

Proposition 4.4 *Let H be an arbitrary compact subset of $\mathbb{R}$. Under Assumptions 2.1 and 4.3, $\{M_{1,n}(h)\}_{h\in H}$ converges weakly if $f(l,\theta) = 0$ for all $\theta \in \Theta$ and $\{M_{2,n}(h)\}_{h\in H}$ converges weakly if $f(r,\theta) = 0$ for all $\theta \in \Theta$.*

Proof Assume first that $f(l,\theta) = 0$ for all $\theta \in \Theta$. We will use Theorem 2.1 from Arcones and Yu (1994) to show that $\{M_{1,n}'(x)\}_{x\in I}$ converges weakly to a Gaussian process. By Assumption 4.3.5 the required mixing condition is satisfied: $k^{p/(p-2)}(\log k)^{2(p-1)/(p-2)}\beta_k \to 0$ as $k \to \infty$ with p from Assumption 4.3.4.

Define for $x \in I$ the function $F_x : I \to \mathbb{R}$ by $F_x(y) = 2b(y)1_{\{y\le x\}}$ and let $\mathcal{F} = \{F_x\}_{x\in I}$. Then, $\mathrm{E}_\theta\, F_x(X_0) = f_\theta(x)$ and by definition of $\hat{f}_{1,n}$,

$$M_{1,n}'(x) = n^{-1/2}\sum_{i=1}^{n}\big(F_x(X_{i\Delta}) - \mathrm{E}_{\theta_0}\, F_x(X_0)\big).$$

The function $F_x(y)$ is jointly measurable in (x,y) and the envelope function $\sup_{x\in I}|F_x| = 2|b|$ of $\mathcal{F}$ has finite p'th moment by Assumption 4.3.4. Furthermore, $\mathcal{F}$ is a so-called Vapnik-Červonenkis (VC) subgraph class of functions. This follows from lemmas in van der Vaart and Wellner (1996): the indicator functions $H_x(y) = 1_{\{y\le x\}} = 1_{\{(-\infty,0]\}}(y-x)$ form a VC subgraph class of functions (Lemma 2.6.16) and $F_x = bH_x$; now use Lemma 2.6.18.

We conclude (Arcones and Yu, 1994) that $M_{1,n}'$ converges weakly in $l^\infty(I)$ to a tight, Gaussian process with P_{θ_0}-almost all paths uniformly bounded and uniformly continuous (with respect to the metric d on I given by $d(x,y)^2 = \int(F_x - F_y)^2 d\mu_{\theta_0}$).

Uniform convergence of M_n'' follows from Assumption 4.3.3, and the limit process M'' given by $M''(h,x) = \dot{f}_{\theta_0}(x)h$ is in $l^\infty(H\times I)$ by Assumption 4.3.2 (since H is compact). It now follows by Slutsky's Theorem that $M_{1,n}' - M_n''$ converges weakly in $l^\infty(H\times I)$ and finally by the continuous mapping theorem that $M_{1,n}$ converges in $l^\infty(H)$.

Similarly for $M_{2,n}$ if $f(r,\theta) = 0$ for all $\theta \in \Theta$. □

Theorem 4.5 *Assume that Assumptions 2.1, 4.1 and 4.3 hold and that $\dot{f}_{\theta_0}(x_0) \neq 0$ for an $x_0 \in I$. Then $\sqrt{n}(\hat{\theta}_{1,n} - \theta_0)$ is $O_p(1)$ if $f(l,\theta) = 0$ for all $\theta \in \Theta$ and $\sqrt{n}(\hat{\theta}_{2,n} - \theta_0)$ is $O_p(1)$ if $f(r,\theta) = 0$ for all $\theta \in \Theta$.*

Proof Recall that $\hat{\theta}_{i,n}$ minimizes $U_{i,n}(\theta) = \sup_{x\in I} |\hat{f}_{i,n}(x) - f_\theta(x)|$ and that $U_{i,n}(\theta) \to U(\theta) = \sup_{x\in I} |f_{\theta_0}(x) - f_\theta(x)|$ P_{θ_0}-almost surely as $n \to \infty$. We first show that $\sqrt{n}U(\hat{\theta}_{i,n})$ is $O_p(1)$: By the triangle inequality

$$\sqrt{n}U(\hat{\theta}_{i,n}) \le \sqrt{n}U_{i,n}(\hat{\theta}_{i,n}) + \sqrt{n}U_{i,n}(\theta_0) \le 2\sqrt{n}U_{i,n}(\theta_0)$$

and $\sqrt{n}U_{i,n}(\theta_0) = M_{i,n}(0)$ converges weakly and is hence $O_p(1)$.

Recall $C(\delta)$ from Assumption 4.1 and note that $P(\sqrt{n}|\hat{\theta}_{i,n} - \theta_0| > \delta) \le P\big(\sqrt{n}U(\hat{\theta}_{i,n}) \ge \sqrt{n}C(\delta/\sqrt{n})\big)$ for all $\delta > 0$. Hence, if

$$\sqrt{n}C(\delta/\sqrt{n}) > c\delta \tag{13}$$

for all $\delta > 0$, some constant $c > 0$ not depending on δ and n large enough, then $\sqrt{n}\big(\hat{\theta}_{i,n} - \theta_0\big)$ is $O_p(1)$.

To prove (13), choose $c, \eta > 0$ such that $U(\theta) > c|\theta - \theta_0|$ for all θ with $|\theta - \theta_0| \le \eta$. This is possible by differentiability of $\theta \to f_\theta(x_0)$ (use e.g. $c = |\dot{f}_{\theta_0}(x_0)|/2$). For $n > \delta^2/\eta^2$,

$$\begin{aligned} C(\delta/\sqrt{n}) &= \inf\big\{U(\theta) : |\theta - \theta_0| > \delta/\sqrt{n}\big\} \\ &= \min\Big(\inf\big\{U(\theta) : \delta/\sqrt{n} < |\theta - \theta_0| \le \eta\big\}, \inf\big\{U(\theta) : |\theta - \theta_0| > \eta\big\}\Big) \\ &= \min\Big(\inf\big\{U(\theta) : \delta/\sqrt{n} < |\theta - \theta_0| \le \eta\big\}, C(\eta)\Big). \end{aligned}$$

Now, $C(\eta) > 0$ by Assumption 4.1 and $\inf\big\{U(\theta) : \delta/\sqrt{n} < |\theta - \theta_0| < \eta\big\} \to 0$ as $n \to \infty$ since $U(\theta_0) = 0$ and U is continuous at θ_0. Hence, for n large enough

$$C(\delta/\sqrt{n}) = \inf\big\{U(\theta) : \delta/\sqrt{n} < |\theta - \theta_0| < \eta\big\} > c\delta/\sqrt{n}$$

which proves (13) and thus the theorem. □

4.3 Convergence in distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$

We finally show that $\sqrt{n}\big(\hat{\theta}_n - \theta_0\big)$ is $O_p(1)$ and even converges weakly (Theorem 4.7). Let $M_n'(x) = n^{1/2}\big(\hat{f}_n(x) - f_{\theta_0}(x)\big)$ and $M_n(h) = \sup_{x\in I} |M_n'(x) - M_n''(h, x)|$. We first give a uniform convergence result for M_n similar to Proposition 4.4.

Proposition 4.6 *Assume that Assumptions 2.1 and 4.3 hold and $f(l, \theta) = f(r, \theta) = 0$ for all $\theta \in \Theta$. Then $\{M_n(h)\}_{h\in H}$ converges weakly for any compact set $H \subseteq \mathbb{R}$.*

Proof Recall that $\hat{f}_n = \hat{\lambda}_{1,n}\hat{f}_{1,n} + \hat{\lambda}_{2,n}\hat{f}_{2,n}$ where $\hat{\lambda}_{j,n}$ converges pointwise P_{θ_0}-almost surely to $\lambda_j := \lambda_{\theta_0,j}$, $j = 1, 2$ (see Section 3.1 for definitions of the various λ's). The convergence is even uniform: indeed, note that

λ_1 is continuous and decreasing and argue as in the proof of the classical Glivenko-Cantelli theorem (Loève, 1963, page 20).

We first argue that it suffices to consider $\lambda_1\hat{f}_{1,n} + \lambda_2\hat{f}_{2,n}$ instead of $\hat{f}_n$: By adding and subtracting $\lambda_1\hat{f}_{1,n}$ and $\lambda_2\hat{f}_{2,n}$ we get

$$\begin{aligned}\hat{f}_n &= \big(\hat{\lambda}_{1,n} - \lambda_1\big)\hat{f}_{1,n} + \big(\hat{\lambda}_{2,n} - \lambda_2\big)\hat{f}_{2,n} + \lambda_1\hat{f}_{1,n} + \lambda_2\hat{f}_{2,n} \\ &= \big(\hat{\lambda}_{1,n} - \lambda_1\big)\big(\hat{f}_{1,n} - f_{\theta_0}\big) + \big(\hat{\lambda}_{2,n} - \lambda_2\big)\big(\hat{f}_{2,n} - f_{\theta_0}\big) + \lambda_1\hat{f}_{1,n} + \lambda_2\hat{f}_{2,n}\end{aligned}$$

and hence,

$$M'_n = \big(\hat{\lambda}_{1,n} - \lambda_1\big)M'_{1,n} + \big(\hat{\lambda}_{2,n} - \lambda_2\big)M'_{2,n} + M'_{\lambda,n} \tag{14}$$

where $M'_{\lambda,n}(x) = n^{1/2}\big(\lambda_1(x)\hat{f}_{1,n}(x) + \lambda_2(x)\hat{f}_{2,n}(x) - f_{\theta_0}(x)\big)$. In the proof of Proposition 4.4 we showed that $M'_{1,n}$ and $M'_{2,n}$ converge weakly, and it now follows from Slutsky's Theorem that M'_n converges in $l^\infty(I)$ if $M'_{\lambda,n}$ does.

Now, let $\mathcal{F} = \{F_x\}_{x\in I}$ where $F_x : I \to \mathbb{R}$ is defined by

$$\begin{aligned}F_x(y) &= 2\lambda_1(x)b(y)1_{\{y\le x\}} - 2\lambda_2(x)b(y)1_{\{y>x\}} \\ &= 2b(y)\big(\lambda_1(x) - 1_{\{y>x\}}\big), \quad y \in I.\end{aligned}$$

Then $\mathrm{E}_\theta\, F_x(X_0) = f_\theta(x)$ and $M'_{\lambda,n}(x) = n^{-1/2}\sum_{i=1}^n\big(F_x(X_{i\Delta}) - f_{\theta_0}(x)\big)$. The function $F_x(y)$ is jointly measurable in (x, y) and the envelope function $\sup_{x\in I}|F_x| \le 4|b|$ of $\mathcal{F}$ has finite p'th moment by Assumption 4.3.4.

Let Q be an arbitrary probability measure on I with $b \in L^2(Q)$ and let $||\cdot||_Q$ be the $L^2(Q)$-norm. By continuity and boundedness of λ_1 and $x \to \int_l^x b^2\,dQ$ it easily follows that the $||\cdot||_Q$-covering number $N(\varepsilon, \mathcal{F}, ||\cdot||_Q)$, that is, the minimal number of $||\cdot||_Q$-balls of radius ε needed to cover $\mathcal{F}$, is at most $32\int_l^r b^2\,dQ/\varepsilon^2$ (at least for small ε). For further details, see the proof of Proposition II.8 in Sørensen (2000). Similar arguments show that $N(\varepsilon, \mathcal{F}, ||\cdot||_p) \le C/\varepsilon^p$ where $||\cdot||_p$ is the L^p-norm with respect to μ_{θ_0} (p being the number from Assumption 4.3.4) and $C > 0$ is a constant not depending on ε. This implies $\int_0^\infty(\log N(\varepsilon, \mathcal{F}, ||\cdot||_p))^{1/2}d\varepsilon < \infty$.

It follows (Arcones and Yu, 1994, Lemma 2.1) that $M'_{\lambda,n}$ converges in $l^\infty(I)$[1] and hence from (14) that M'_n converges in $l^\infty(I)$. Finally, weak convergence of M''_n and M_n follows as in the proof of Proposition 4.4. □

Theorem 4.7 *Assume that Assumptions 2.1, 4.1, and 4.3 hold and that $f(l, \theta) = f(r, \theta) = 0$ for all $\theta \in \Theta$. If, in addition, $\dot{f}_{\theta_0}(x_0) \ne 0$ for some $x_0 \in I$ then $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is $O_p(1)$ and if furthermore $\dot{f}_{\theta_0}(x) \ne 0$ for all $x \in I$, then $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges weakly.*

[1] This also follows from $\mathcal{F}$ being a $\mathcal{F}$ a Vapnik-Červonenkis subgraph class of functions; a quite tedious proof may be found in Sørensen (2000, Lemma II.12).

Proof The stochastic boundedness of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ follows exactly as in the proof of Theorem 4.5. For the weak convergence it then suffices to show that P_{θ_0}-almost all paths of the limit M of M_n has a unique minimum (van der Vaart and Wellner, 1996, Theorem 3.2.2).

The limit process M has the form $M(h) = \sup_{x \in I} |M'(x) - \dot{f}_{\theta_0}(x)h|$ where M' is the Gaussian limit of M_n'. All paths $h \to M(h)$ satisfy $M(h) \to \infty$ as $h \to \pm\infty$ since $M(h) \geq \left|M'(x) - \dot{f}_{\theta_0}(x)h\right|$ and $\dot{f}_{\theta_0}(x) \neq 0$ for any fixed $x \in I$ (for this it suffices that $\dot{f}_{\theta_0}(x_0) \neq 0$ for some $x_0 \in I$). All paths are continuous since $|M(h_2) - M(h_1)| \leq |h_2 - h_1| \sup_{x \in I} |\dot{f}_{\theta_0}(x)|$ for all $h_1, h_2 \in \mathbb{R}$ and hence have a minimum. We must show that the minimum is unique for almost all paths.

Now, it holds P_{θ_0}-almost surely that M' is continuous and satisfies $M'(x) \to 0$ as $x \searrow l$ and $x \nearrow r$ (by Portmanteau's theorem). Consider a path $h \to M(h)$ for which this is the case and assume that $h_1 < h_2$ both minimize M. Let $m = M(h_1) = M(h_2)$ be the minimum value. All paths of M are obviously *weakly* convex so $M(\bar{h}) = m$ where $\bar{h} = (h_1 + h_2)/2$ is the mid point between h_1 and h_2.

By definition, $M(\bar{h}) = \sup_{x \in I} \left|M'(x) - \dot{f}_{\theta_0}(x)\bar{h}\right|$. Choose a sequence (x_n) from I such that $\left|M'(x_n) - \dot{f}_{\theta_0}(x_n)\bar{h}\right| \geq m - 1/n$ for each $n \geq 1$. For $j = 1, 2$ and all $n \geq 1$,

$$m = M(h_j) \geq \left|M'(x_n) - \dot{f}_{\theta_0}(x_n)h_j\right|$$

implying that $\left|\dot{f}_{\theta_0}(x_n)\right|(h_2 - h_1)/2 \leq 1/n$ (due to the special form of the graph of $x \to |M'(x) - \dot{f}_{\theta_0}(x)h_j|$). Hence, $\left|\dot{f}_{\theta_0}(x_n)\right| \to 0$ as $n \to \infty$.

Since $\dot{f}_{\theta_0}$ is continuous and $\dot{f}_{\theta_0}(x) \neq 0$ for all $x \in I$ it thus holds for any $l < x_1 < x_2 < r$ that $x_n \notin [x_1, x_2]$ for n large enough and hence $M'(x_n) \to 0$ as $n \to \infty$. It follows that

$$m = M(\bar{h}) = \lim_{n \to \infty} \left|M'(x_n) - \dot{f}_{\theta_0}(x_n)\bar{h}\right| = 0 \tag{15}$$

so $M(h_1) = M(h_2) = m = 0$. This is not possible, though, since for any $x \in I$ at least one of the values $|M'(x) - \dot{f}_{\theta_0}(x)h_1|$ and $|M'(x) - \dot{f}_{\theta_0}(x)h_2|$ is strictly positive.

We conclude that M has a unique minimum P_{θ_0}-almost surely and hence that $\sqrt{n}\big(\hat{\theta}_n - \theta_0\big)$ converges weakly. □

Parts of the above proof could be repeated with M_1 or M_2 substituted for M. If $\bar{h}$ and (x_n) are as above with M replaced by M_1, say, then it would still hold that x_n could be made arbitrarily close to l or r by choosing n large enough. But $M_1'(x)$ does not converge to zero as $x \to r$ so $\lim_{n \to \infty} \left|M_1'(x_n) - \dot{f}_{\theta_0}(x_n)\bar{h}\right|$, corresponding to (15), need not be zero and cannot be rejected as the minimum value of M_1. That is, we cannot rule out the possibility that M_1 has several minimum points. Similarly for M_2.

The distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to the distribution of the minimum of the process $M(h) = \sup_{x \in I} |M'(x) - \dot{f}_{\theta_0}(x)h|$, where M' is the limit of M_n' (which has a quite complicated variance structure due to the temporal dependence in X). The limit distribution cannot easily be described more explicitly than that. In particular, there is no reason to believe that the limit distribution is Gaussian (a small simulation study indicates however that the limit distribution *might be* close to Gaussian, see Section 5.1).

5 Example: the CKLS model

Consider now the so-called CKLS model

$$dX_t = (\alpha + \beta X_t)\, dt + \sigma X_t^{\gamma}\, dW_t \tag{16}$$

named after Chan *et al.* (1992) who first discussed it in this generality. The geometric Brownian motion, the Ornstein-Uhlenback process and the Cox-Ingersoll-Ross model all occur as special cases.

Let $\alpha > 0$, $\beta < 0$ and $\sigma > 0$. If $1/2 < \gamma < 1$ then Assumption 2.1 holds with $I = (0, \infty)$ and the value $f_{(\gamma,\sigma)}(x)$ is for fixed α and β given by

$$K_0(\alpha, \beta, \gamma, \sigma) \exp\left(\frac{2\alpha}{\sigma^2(1-2\gamma)} x^{1-2\gamma} + \frac{\beta}{\sigma^2(1-\gamma)} x^{2-2\gamma}\right), \quad x > 0$$

which converges to zero as $x \to 0$ and $x \to \infty$. Hence, the appropriate estimator of f is $\hat{f}_n$. There is no explicit expression for the normalizing constant K_0, but we can calculate it numerically.

In the following we apply the estimation technique from Section 3 to simulated data from the above model. Related studies may be found in Honoré (1997), Poulsen (1999) and Elerian *et al.* (2000), all investigating estimation techniques far more computationally demanding than those considered here.

5.1 Investigation of the limit distribution in a simple case

First, consider the simple (and unrealistic) case where α, β and σ are known and only γ should be estimated. From Section 4 we know that the estimator $\hat{\gamma}_n$ obtained by minimizing U_n is consistent and converges weakly (when centered and scaled properly).

Figure 2 shows a histogram (to the left) and a QQ-plot (to the right) for 1000 simulated values of $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$. Each value of $\hat{\gamma}_n$ was computed as follows: a path of X was simulated from time zero to time 1000 (by means of the Euler scheme with time-step 1/1000); the values at time-points $1, 2, \ldots, 1000$ (corresponding to $\Delta = 1$) were recorded; and $\hat{\gamma}_n$ was

calculated based on these 1000 observations. The true value of γ was $\gamma_0 = 0.75$ and the known parameters were fixed at $(\alpha, \beta, \sigma) = (0.04, -0.6, 0.2)$.

The histogram and the QQ-plot both indicate that the asymptotic distribution of $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$ is quite close to Gaussian. [Figure 2]

5.2 Comparison with two other methods

Consider now the more realistic case where α, β, γ and σ are all unknown. The method from Section 3 does not apply immediately (as the drift is no longer known). Instead we use the following adjusted strategy.

First, the drift parameters, α and β, are estimated by a least squares (LS) approach as suggested by Aït-Sahalia (1996). Specifically, estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ are obtained by minimization of $\sum_{i=2}^{n}(X_{i\Delta} - \varphi(X_{(i-1)\Delta}, \alpha, \beta))^2$ where

$$\varphi(x, \alpha, \beta) = \mathrm{E}_{\alpha,\beta,\gamma,\sigma}\big(X_\Delta | X_0 = x\big) = e^{\beta\Delta}\Big(x + \frac{\alpha}{\beta}\Big) - \frac{\alpha}{\beta} \tag{17}$$

is the conditional expectation one step ahead (which does not depend on γ and σ). This is equivalent to estimation via a martingale estimating function (Bibby and Sørensen, 1995). Next, the diffusion parameters, γ and σ, are estimated as described in Section 3 — except that the true drift function is now replaced by the estimated version $\hat{\alpha}_n + \hat{\beta}_n x$ (in both $\hat{f}_n$ and f). Note that we have no evidence that these estimators for γ and σ have nice asymptotic properties since the proofs in Section 4 do not take into account the errors introduced by estimation of the drift parameters.

Below we compare the above estimation technique to two other simple methods in a small simulation study. In the first approach α and β are estimated as above, and γ and σ are thereafter estimated by maximizing $l_n(\gamma, \sigma) = \sum_{i=1}^{n} \log \mu(X_{i\Delta}, \hat{\alpha}_n, \hat{\beta}_n, \gamma, \sigma)$ which would be the log-likelihood *if* the observations were independent and identically distributed with density $\mu(\cdot, \hat{\alpha}_n, \hat{\beta}_n, \gamma, \sigma)$.

The second method is the one suggested by Chan *et al.* (1992) themselves — based on rough approximations of the conditional moments one step ahead. To be specific, define $\varepsilon_i = X_{i\Delta} - X_{(i-1)\Delta} - (\alpha + \beta X_{(i-1)\Delta})\Delta$ and solve the equation

$$\sum_{i=2}^{n}\Big(\varepsilon_i, \varepsilon_i X_{(i-1)\Delta}, \varepsilon_i^2 - \Delta\sigma^2 X_{(i-1)\Delta}^{2\gamma}, \varepsilon_i^2 X_{(i-1)\Delta} - \Delta\sigma^2 X_{(i-1)\Delta}^{1+2\gamma}\Big) = 0.$$

wrt. $(\alpha, \beta, \gamma, \sigma)$. The estimating function — and thus the estimators — can be considerably biased when Δ is not "small".

Table 1 reports empirical means and standard errors of simulated values of the above-mentioned estimators (with obvious names). The estimators have been computed for 100 simulated datasets, each of length 500 and

with the same parameter values as in Section 5.1. The CKLS-estimates are clearly biased (in fact half the γ-estimates are less than $1/2$ and should strictly speaking have been excluded). Of course, 100 simulations are far too few too draw any final conclusions on the two remaining estimators, but the study indicates that they are both quite reasonable. See Sørensen (2000, Section II.7) for further details on the study, in particular for a discussion on some numerical problems related to the optimizations. [Table 1]

Informal studies indicate that the non-parametric procedure suggested by Aït-Sahalia (1996) yields quite reasonable estimators of the diffusion function in the central area of the distribution but the estimator is of course extremely variable in areas with few observations.

6 Concluding remarks

In this paper we have discussed a method for estimation of parameters in the diffusion function. It provides consistent and in some cases also weakly convergent estimators. The usual limit theory does not apply; instead we used empirical process theory for proving the asymptotic results. We applied (an adjusted version of) the method to simulated data from the difficult CKLS model and obtained satisfactory (though presumably not efficient) estimators. From a theoretical point of view the application of empirical process theory is perhaps most interesting.

Acknowledgements Thanks to Søren Feodor Nielsen for help with the asymptotic results, to my advisor Martin Jacobsen, to Michael Sørensen, and to two referees for valuable comments.

References

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

Aït-Sahalia, Y. (1998), Maximum likelihood estimation of discretely sampled diffusions: a closed-form approach, Revised version of Working Paper 467, Graduate School of Business, University of Chicago.

Arcones, M. A. and Yu, B. (1994), Central limit theorems for empirical and U-processes of stationary mixing sequences, *Journal of Theoretical Probability* **7**, 47–71.

Banon, G. (1978), Nonparametric identification for diffusion processes, *Siam J. Control and Optimization* **16**, 380–395.

Bibby, B. M. and Sørensen, M. (1995), Martingale estimation functions for discretely observed diffusion processes, *Bernoulli* **1**, 17–39.

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

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

Dohnal, G. (1987), On estimating the diffusion coefficient, *Journal of Applied Probability* **24**, 105–114.

Elerian, O., Chib, S. and Shephard, N. (2000), Likelihood inference for discretely observed non-linear diffusions, Economics discussion paper 146, Nuffield College, Oxford. To appear in *Econometrica*.

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

Florens-Zmirou, D. (1993), On estimating the diffusion coefficient from discrete observations, *Journal of Applied Probability* **30**, 790–804.

Genon-Catalot, V. and Jacod, J. (1993), On the estimation of the diffusion coefficient for multi-dimensional diffusion processes, *Ann. Inst. Henri Poincaré* **29**, 119–151.

Genon-Catalot, V. and Jacod, J. (1994), Estimation of the diffusion coefficient for diffusion processes: random sampling, *Scandinavian Journal of Statistics* **21**, 193–221.

Genon-Catalot, V., Laredo, C. and Picard, D. (1992), Non-parametric estimation of the diffusion coefficient by wavelets methods, *Scandinavian Journal of Statistics* **19**, 317–335.

Hansen, L. P., Scheinkman, J. A. and Touzi, N. (1998), Spectral methods for identifying scalar diffusions, *Journal of Econometrics* **86**, 1–32.

Hoffmann, M. (1997), Minimax estimation of the diffusion coefficient through irregular samplings, *Statistics and Probability Letters* **32**, 11–24.

Hoffmann, M. (1999*a*), Adaptive estimation in diffusion processes, *Stochastic Processes and their Applications* **79**, 135–163.

Hoffmann, M. (1999*b*), L_p estimation of the diffusion coefficient, *Bernoulli* **5**, 447–481.

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

Jacod, J. (1993), Random sampling in estimation problems for continuous Gaussian processes with independent increments, *Stochastic Processes and their Applications* **44**, 181–204.

Jacod, J. (2000), Non-parametric kernel estimation of the coefficient of a diffusion, *Scandinavian Journal of Statistics* **27**, 83–96.

Jiang, G. J. and Knight, J. L. (1997), A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model, *Econometric Theory* **13**, 615–645.

Karatzas, I. and Shreve, S. E. (1991), *Brownian Motion and Stochastic Calculus*, 2nd edn, Springer-Verlag, New York.

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

Loève, M. (1963), *Probability Theory*, 3rd edn, D Van. Nostrand, Princeton.

Pedersen, A. R. (1995), A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations, *Scandinavian Journal of Statistics* **22**, 55–71.

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

Soulier, P. (1998), Non parametric estimation of the diffusion coefficient of a diffusion process, *Stochastic Analysis and Applications* **16**, 185–200.

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

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer-Verlag, New York.

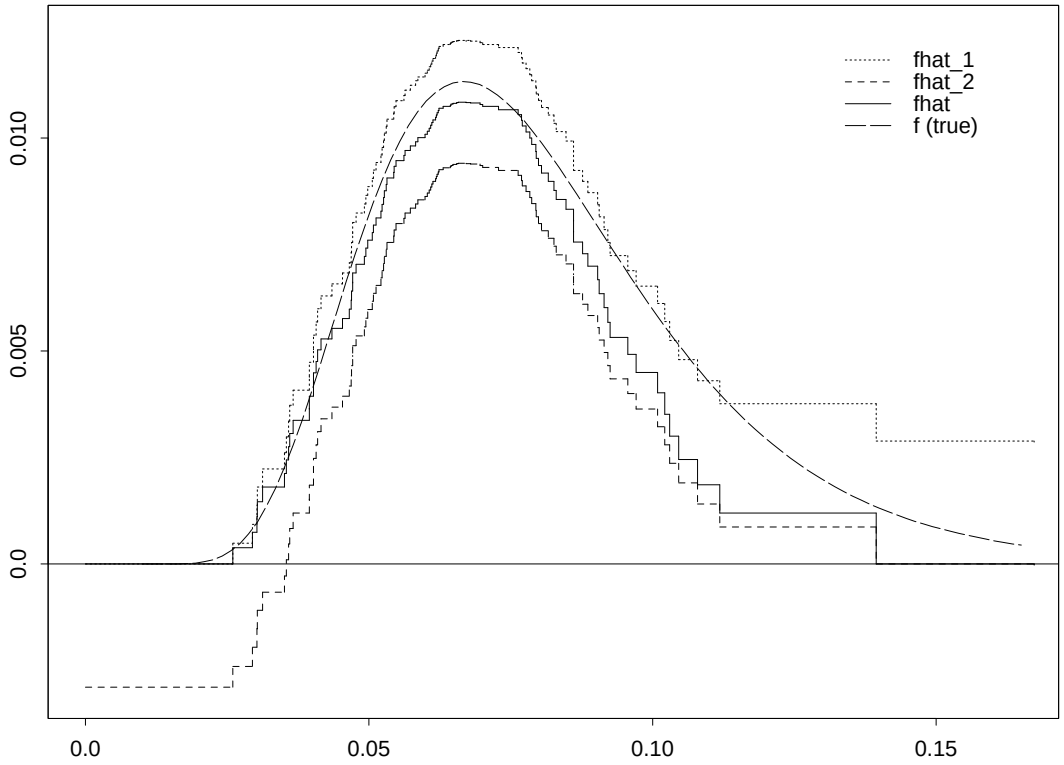

Figure 1: Graphs for the estimators $\hat{f}_{1,n}$, $\hat{f}_{2,n}$ and $\hat{f}_n$ for 100 simulated data from the model $dX_t = (0.04 - 0.6X_t)\,dt + 0.2X_t^{\gamma}\,dW_t$ with true value $\gamma_0 = 0.75$ together with the graph of f corresponding to the true parameter value. The value of Δ is 1.

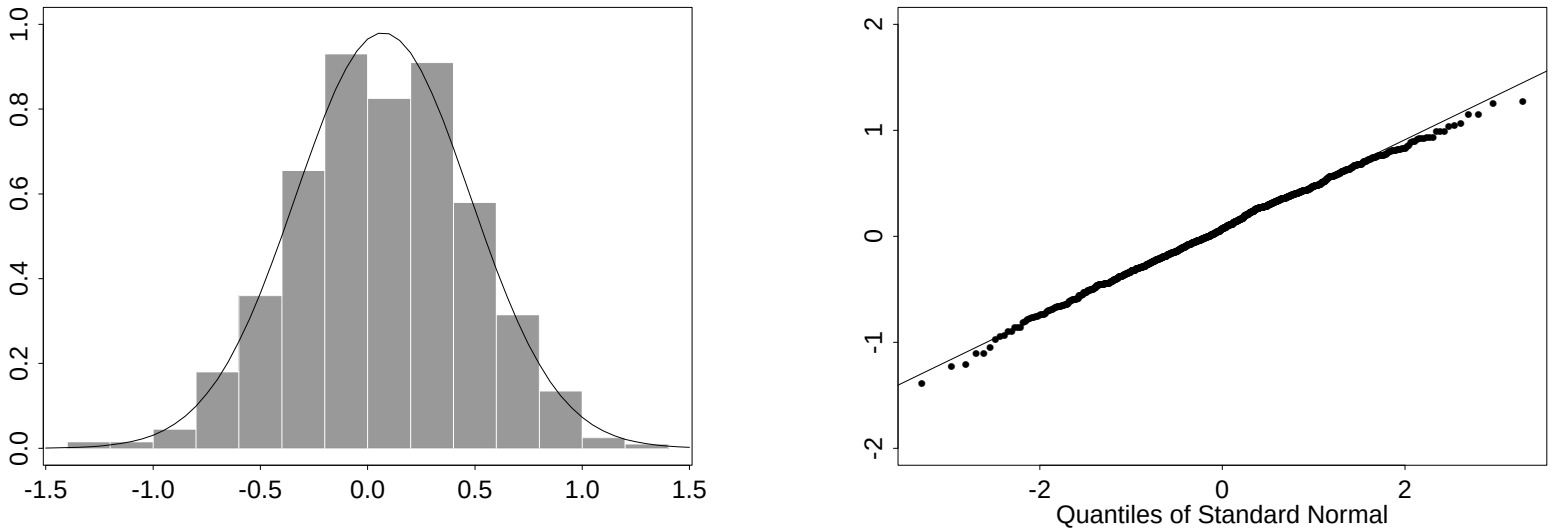

Figure 2: Histogram (to the left) and Q-plot (to the right) for 1000 simulated values of $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$. The values of α, β and σ are considered as known. The curve on the left is the density for normal distribution with mean and standard error equal to those of the empirical distribution of the estimates (0.0716 and 0.4071 respectively).

Method	Failures	$\hat{\alpha}_n$ (0.04)		$\hat{\beta}_n$ (-0.60)		$\hat{\gamma}_n$ (0.75)		$\hat{\sigma}_n$ (0.20)	
		mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
LS-min U_n	6	0.0411	0.0050	-0.6166	0.0785	0.7386	0.0958	0.2009	0.0531
LS-IID	7	0.0411	0.0050	-0.6166	0.0785	0.7467	0.0800	0.2039	0.0439
CKLS	(49)	0.0306	0.0027	-0.4586	0.0422	0.5076	0.1328	0.0862	0.0352

Table 1: Empirical means and standard errors of various estimators for 100 realizations of the CKLS model. The true parameters are given in the top line, $n = 500$, and $\Delta = 1$. In the two top lines 'Failures' reports the number of simulations for which the optimization problem did not have a solution; in the bottom line it reports the number of γ-estimates less than 1/2 (note that we have averaged over all 100 values anyway).