Simulated Likelihood Approximations for Stochastic Volatility Models*

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

September 19, 2000

Abstract

The objective of this paper is parametric inference for stochastic volatility models. We consider a two-dimensional diffusion process (X,V) where V is ergodic and X has drift and diffusion coefficient completely determined by V. The drift and the diffusion coefficient for V depend on an unknown parameter θ, and our concern is estimation of θ from discrete-time observations of X. The volatility process V remains unobserved. We consider approximate maximum likelihood estimation: for the k'th order approximation we pretend that the observations form a k'th order Markov chain, find the corresponding approximate log-likelihood function, and maximize it with respect to θ. The approximate log-likelihood function is not known analytically but can easily be calculated by simulation. For each k the method yields consistent and asymptotically normal estimators. Simulations from the model where V is a Cox-Ingersoll-Ross model are used for illustration.

Keywords: Approximate maximum likelihood; Cox-Ingersoll-Ross process; discretely-time observations; stochastic volatility models.

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

1 Introduction

Our concern is approximate maximum likelihood estimation for continuous-time stochastic volatility models. By stochastic volatility models we will mean models for a pair of processes (X,V) where V is a latent, positive diffusion process and the observable process X solves a stochastic differential equation with diffusion term $\sqrt{V}$ and drift determined by V as well. The process V is called the *volatility process*. We consider parametric specifications of the drift and the diffusion function for V, and the objective is estimation from *discrete-time observations* of X.

For a start, consider the classical Black-Scholes model (or geometric Brownian motion)

$$dP_t = \alpha P_t\,dt + \tau P_t\,dW_t \tag{1}$$

which is (or rather was) often used to model stock prices. The classical option pricing formula was derived under the assumption that the price of the underlying stock evolved according to this model (Black & Scholes 1973). If P solves (1) then $\log P$ has constant volatility (squared diffusion) and independent, normally distributed increments. It is well-known that these properties are inconsistent with empirical findings: studies have revealed that stock returns (and other financial data) most often are dependent, have strongly leptokurtic marginal distributions and exhibit signs of randomly varying variance over time.

In the discrete-time setting ARCH-type models and discrete-time stochastic volatility models have been used for modeling such phenomena; see Shephard (1996) for an overview of both model types. However, for derivative pricing (and related problems) it may be advantageous to use diffusion-type models, retaining the Itô calculus at our disposal. Also, irregularly sampled data are in general easier handled for continuous-time models than for discrete-time models.

Of course one could generate the above features by simply allowing for certain non-linear drift and diffusion functions for the price process. In the stochastic volatility framework, however, the linear structure of the equation for P is retained, but an additional source of variability is introduced: the constant τ in (1) is replaced by the value of a latent diffusion process $\sqrt{V}$. The modified equation for P is thus

$$dP_t = \alpha P_t\,dt + \sqrt{V_t}P_t\,dW_t. \tag{2}$$

In this paper we shall consider models given by

$$dX_t = \xi(V_t)\,dt + \sqrt{V_t}\,dW_t \tag{3}$$

$$dV_t = b(V_t,\theta)\,dt + \sigma(V_t,\theta)\,d\tilde{W}_t. \tag{4}$$

With $P = e^X$ it follows by Itô's formula that this model is equivalent to (2) if $\xi(v) = \alpha - v/2$. Hence, a possible application of the model is for the logarithm of stock prices. The drift and diffusion for V are parameter dependent, and we shall be interested in estimation of θ from equidistant observations $X_0, X_\Delta, \ldots, X_{n\Delta}$ of X. The volatility process V remains unobserved.

For the above model to make sense, V must be a positive process. Various models were suggested in the late eighties and early nineties: V was modeled as a geometric Brownian motion (Hull & White 1987), as a Cox-Ingersoll-Ross process (Hull & White 1988, Heston 1993), as the exponential of a Ornstein-Uhlenbeck process (Wiggins 1987, Chesney & Scott 1989) and as a squared Ornstein-Uhlenbeck process (Scott 1987, Stein & Stein 1991).

All these papers focus on pricing of a European call option written on a stock with price process $P = e^X$. Pricing is investigated for fixed value of the parameter θ in the equation for V, and the majority of the papers pay no or little attention to estimation of θ. Only Scott (1987) and Chesney & Scott (1989) address the problem seriously and derive moment-like estimators for the parameters. More recently, several estimation approaches have been suggested in the statistics literature. Below we briefly review the basic ideas; see Sørensen (2000, Section 3.4) for a more detailed survey. Some of the methods have been applied earlier for discrete-time versions of the model (see the surveys by Shephard (1996) and Ghysels, Harvey & Renault (1996)), but the continuous-time case is more troublesome, and in general the methods do not carry over immediately from discrete time to continuous time.

Genon-Catalot, Jeantheau & Laredo (1999) consider the approximation that the increments $Z_1, \ldots, Z_n$ are independent and identically distributed with conditional distribution of Z_1 given V equal to $N(\Delta\xi(V_0), \Delta V_0)$. The estimators are consistent as $n \to \infty$ only if the time-step Δ decreases to zero as n increases. For (large) fixed values of Δ the bias may be considerable. Also, only estimation of parameters from the stationary distribution of V is possible. In another paper, Genon-Catalot, Jeantheau & Laredo (1998) consider mean-reverting models for V. Then calculation of various moments of the joint distribution of the increments is possible, and estimation is carried out by matching theoretical and empirical moments. For any fixed Δ the estimators so obtained are consistent and asymptotically normal as n increases. However, the simulation study in Section 7 indicates that there may be serious existence problems in practice.

The two above methods require no hard numerical computations or simulations and are thus quick in practice. As opposed to this most other methods (including the one suggested in this paper) are quite computationally intensive. Nielsen, Vestergaard & Madsen (2000) use a filtering approach where values of V are estimated together with the parameter. This requires that n (that is, the number of observed increments) five-dimensional differential equation are solved by numerical methods. Eraker (1998) uses a Bayesian approach which requires Markov Chain Monte Carlo simulation of values of θ as well as of V and X at a number of time-points in between those where X is observed; see also Elerian, Chib & Shephard (2000). The so-called efficient method of moments (Gallant & Tauchen 1996) is applied to a stochastic volatility model by Andersen & Lund (1997). Finally, Sørensen (1999) studies prediction-based estimating functions. Particular attention is paid to the case where, for a function f and an integer k, each term in the estimating function is given in terms of the value $f(Z_i)$ and its projection on some space determined by the previous k increments $Z_{i-k+1}, \ldots, Z_i$.

Typically, the projections must be calculated by simulation.

The method suggested in this paper is somewhat related to the approach just mentioned since we also choose a number $k \geq 0$ and base inference on k lags of the increments. For a given value of k the idea is to *pretend that* $(Z_1, Z_2, \ldots)$ *is k'th order Markov*, find the corresponding approximate likelihood function, and maximize it with respect to θ. In particular $k = 0$ corresponds to pretending that observations are independent, drawn from the stationary distribution (and may thus be interpreted as an improvement of the method by Genon-Catalot *et al.* (1999) who use an approximation to the stationary density), and $k = 1$ corresponds to pretending that observations are Markov.

There is no explicit expression for the k-lag conditional density, but it can be expressed in terms of expectations with respect to the distribution of $(V_t)_{0 \leq t \leq (k+1)\Delta}$ and therefore calculated by simulation of V on the interval from zero to $(k+1)\Delta$. For any fixed Δ and any $k \geq 0$ the corresponding approximate score function in unbiased and (under regularity conditions, of course) the estimator is consistent and asymptotically normal as the number of observations increases. We use the model where $\xi \equiv 0$ and V is a Cox-Ingersoll-Ross process as an example and apply the method to simulated data. In the simple (but unrealistic) case with only one parameter unknown we obtain satisfactory estimates even for $k = 0$, whereas we for all three parameters unknown must use a larger k, say 4, to get reasonable estimates.

The paper is organized as follows. In Section 2 we discuss the model and some of its probabilistic properties. We introduce the likelihood approximations and the estimation method in Section 3 and discuss computational aspects in Section 4. Asymptotic results are proved in Section 5 and efficiency of the estimators is briefly discussed in Section 6. We discuss the Cox-Ingersoll-Ross model in Section 7. Conclusions are drawn in Section 8.

2 Preliminaries

Let $(W, \tilde{W}) = \{(W_t, \tilde{W}_t)\}_{t \geq 0}$ be a standard two-dimensional Brownian motion defined on a filtered probability space $(\Omega, \mathscr{F}, \mathscr{F}_t, Pr)$ satisfying the usual conditions, and let $U_X : \Omega \to \mathbb{R}$ and $U_V : \Omega \to (0, \infty)$ be $\mathscr{F}_0$-measurable random variables, mutually independent and independent of $(W, \tilde{W})$. Furthermore, let θ be an unknown p-dimensional parameter varying in $\Theta \subseteq \mathbb{R}^p$ and consider the stochastic differential equations (3)–(4). We shall assume that U_X, U_V and the functions $\xi : (0, \infty) \to \mathbb{R}$, $b : (0, \infty) \times \Theta \to \mathbb{R}$ and $\sigma : (0, \infty) \times \Theta \to (0, \infty)$ are such that there exists a solution (X, V) with V positive, stationary and ergodic:[1]

Assumption 2.1 For any value of $\theta \in \Theta$

(A1) there exists a unique strong solution (X, V) to (3)–(4) with state space $\mathbb{R} \times (0, \infty)$ and initial value $(X_0, V_0) = (U_X, U_V)$;

[1] Simple integral conditions ensuring stationarity can be found in Karlin & Taylor (1981, Section 15.6) or Karatzas & Shreve (1991, Section 5.5), for example.

(A2) the process V is stationary and ergodic with invariant measure μ_θ and $V_0 = U_V \sim \mu_\theta$. □

It is natural to consider increments of X. Define for $i \in \mathbb{N}$ the random variables Z_i, M_i and S_i by

$$Z_i = X_{i\Delta} - X_{(i-1)\Delta}; \quad M_i = \int_{(i-1)\Delta}^{i\Delta} \xi(V_s)\, ds; \quad S_i = \int_{(i-1)\Delta}^{i\Delta} V_s\, ds$$

and let furthermore $H_i = (M_i, S_i)$. The sequence $Z = (Z_1, Z_2, \dots)$ takes values in the space $\mathbb{R}^\infty$ of real sequences. Denote by P_θ the distribution of Z when $V_0 \sim \mu_\theta$. It is not possible to characterize P_θ explicitly, but the following properties are well-known: *Assume that condition (A1) holds. Then, conditional on $\{V_t\}_{t\geq 0}$, the increments $Z_1, Z_2, \dots$ of X are independent and the conditional distribution of Z_i is Gaussian with expectation M_i and variance S_i. If furthermore conditions (A2) and (A3) hold then $H = (H_1, H_2, \dots)$ and $Z = (Z_1, Z_2, \dots)$ are strictly stationary and ergodic.*

The random variables $Z_1, Z_2, \dots$ are not (marginally) independent; neither is Z Markov. However, P_θ defines a *hidden Markov model* (Genon-Catalot *et al.* 1998): Let $\tilde{H}_i = (V_{i\Delta}, M_i, S_i)$. Then $\tilde{H} = (\tilde{H}_1, \tilde{H}_2, \dots)$ is stationary Markov (because V is stationary Markov and $\tilde{H}_i$ is a function of $(V_t)_{(i-1)\Delta \leq t \leq i\Delta}$), and conditionally on $\tilde{H}$ the increments $Z_1, Z_2, \dots$ are independent with conditional distribution of Z_i depending on $(i, \tilde{H})$ via $\tilde{H}_i$ only. Note that the hidden chain $\tilde{H}$ has continuous state space. Also note that Z is reversible: for any $n \in \mathbb{N}$, $(Z_1, \dots, Z_n)$ and $(Z_n, \dots, Z_1)$ have same distribution (Sørensen 2000, Proposition III.2.3).

For later use we introduce some further notation on the distribution of Z: let $p_\theta^k(z_1, \dots, z_k)$ denote the density at $(z_1, \dots, z_k)$ of the simultaneous distribution of $Z_1, \dots, Z_k$, $k \in \mathbb{N}$. Then $p_\theta^k > 0$ so the k-lag conditional density

$$p_\theta^{c,k}(z_{k+1} | z_1, \dots, z_k) = \frac{p_\theta^{k+1}(z_1, \dots, z_{k+1})}{p_\theta^k(z_1, \dots, z_k)}$$

at z_{k+1} of Z_{k+1} given $(Z_1, \dots, Z_k) = (z_1, \dots, z_k)$ is well-defined and positive for all $z_1, \dots, z_{k+1}$, $k \in \mathbb{N}$. For $k = 0$ we let $p_\theta^{c,0} = p_\theta^1$. Furthermore, let z_i^j be short for the vector $(z_i, \dots, z_j)$, $i \leq j$.

Finally some comments on possible generalizations of the model. The increments of X would be independent and Gaussian even if ξ and the diffusion function for X were allowed to depend on an unknown parameter η. The mean and variance of the Gaussian distributions would of course depend on η, and estimation of η is easily built into the estimation method below. However, if the Brownian motions W and $\tilde{W}$ are correlated or if ξ or the diffusion term for X depend on X, then the conditional distribution result above is no longer true, and estimation is very difficult. Such models cannot be handled by the method in this paper.

3 Approximations to the likelihood function

We aim at estimation of θ from discrete-time observations $X_0, X_\Delta, \dots, X_{n\Delta}$ while the volatility process V remains completely unobserved. In this section

we introduce a class of approximations to the likelihood function. Later we discuss computational aspects (Section 4) and show that maximization of *any* of the approximations leads to a consistent and asymptotically normal estimator of θ (Section 5).

Motivated by the distributional result in Section 2 we consider the vector of increments $(Z_1,\dots,Z_n)$. For an observation $(z_1,\dots,z_n)$ the likelihood function is given by

$$L_n(\theta)=\int\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi s_i}}\exp\Big(-\frac{(z_i-m_i)^2}{2s_i}\Big)\,d\pi_\theta^n(h^n)=\mathrm{E}_{\pi_\theta^n}\prod_{i=1}^{n}\varphi(z_i,M_i,S_i) \tag{5}$$

where h^n is short for $(h_1,\dots,h_n)=\big((m_1,s_1),\dots,(m_n,s_n)\big)$, π_θ^n is the distribution of H^n and $\varphi(\cdot,m,s)$ is the density of $N(m,s)$.

The likelihood (5) is the expectation with respect to the distribution of H^n of a certain functional. In principle, this expectation could be calculated to any precision as follows: (*i*) simulate a number of paths V up to time $n\Delta$ according to (4); (*ii*) calculate for each simulation (approximations to) the integrals M_i and S_i and the product in (5); (*iii*) calculate the average of the simulated product values. Finally the (simulated) likelihood function should be maximized in order to obtain an estimator of θ. However, this approach is not feasible in practice because one needs a *huge* number of simulated paths of V just to calculate the likelihood function for a *single* parameter value. This is not strange since two paths of V over a large time interval can be very different.

Our approach will be to maximize suitable approximations to L_n rather than L_n itself. The approximations under consideration are easier to simulate, but of course this is at the expense of loss of efficiency.

With the notation from Section 2 the likelihood can be rewritten as

$$L_n(\theta)=p_\theta^n(z)=\prod_{i=0}^{n-1}p_\theta^{c,i}(z_{i+1}|z_1,\dots,z_i)=\prod_{i=0}^{n-1}p_\theta^{c,i}(z_{i+1}|z_1^i). \tag{6}$$

The idea is to approximate the conditional densities in (6) by k-lag conditional densities for some k large enough. This makes sense because the dependence between Z_i and $(Z_1,\dots,Z_j)$ is weak when i is much larger than j (at least under certain mixing conditions). To be specific, leave for $k\in\{0,\dots,n-1\}$ fixed the first $k+1$ terms in (6) unchanged but approximate for $i=k+1,\dots,n-1$ the conditional density $p_\theta^{c,i}(z_{i+1}|z_1^i)$ by $p_\theta^{c,k}(z_{i+1}|z_{i-k+1}^i)$ — recall that Z is strictly stationary. The corresponding approximation of the likelihood is

$$\begin{aligned}L_n^k(\theta)&=\prod_{i=0}^{k}p_\theta^{c,i}(z_{i+1}|z_1,\dots,z_i)\prod_{i=k+1}^{n-1}p_\theta^{c,k}(z_{i+1}|z_{i-k+1},\dots,z_i)\\&=p_\theta^{k+1}(z_1,\dots,z_{k+1})\prod_{i=k+1}^{n-1}p_\theta^{c,k}(z_{i+1}|z_{i-k+1},\dots,z_i).\end{aligned}$$

In particular $k=1$ corresponds to a Markov approximation:

$$L_n^1(\theta)=p_\theta^1(z_1)\prod_{i=1}^{n-1}p_\theta^{c,1}(z_{i+1}|z_i)$$

and $k = 0$ corresponds to independence of $Z_1, \ldots, Z_n$:

$$L_n^0(\theta) = \prod_{i=1}^{n} p_\theta^1(z_i).$$

No approximation is made for $k = n-1$, but the idea is to use a relatively small k to make computations feasible in practice. Note that L_n^k would be the true likelihood function *if* Z was k'th order Markov.

The estimator $\hat{\theta}_n^k$ is of course defined as the value that maximizes L_n^k (for the moment implicitly assuming that it exists and is unique). In practice we shall minimize $U_n^k = -\log L_n^k/n$ rather than maximize L_n^k. If $u_\theta^k = -\log p_\theta^k$ and $u_\theta^{c,k} = -\log p_\theta^{c,k}$ then

$$U_n^k(\theta) = -\frac{1}{n}\log L_n^k(\theta) = \frac{1}{n}u_\theta^{k+1}(z_1^{k+1}) + \frac{1}{n}\sum_{i=k+1}^{n-1} u_\theta^{c,k}(z_{i+1}|z_{i-k+1}^i) \tag{7}$$

$$= \frac{1}{n}\sum_{i=k}^{n-1} u_\theta^{k+1}(z_{i-k+1}^{i+1}) - \frac{1}{n}\sum_{i=k+1}^{n-1} u_\theta^k(z_{i-k+1}^i). \tag{8}$$

It is important to realize that, although we use approximations of the likelihood function, *no bias is introduced* and the estimators are consistent (Section 5). The reason is that we use the *true* k-lag conditional densities rather than approximations. For example, we use the true stationary density p_θ^1 for $k = 0$. This is a crucial difference from the approach taken by Genon-Catalot *et al.* (1998): they apply an approximation to p_θ^1 and the corresponding estimators are thus biased (unless $\Delta \to 0$).

It is of course crucial that the parameter is identifiable from the conditional distribution of Z_{k+1} given Z_1^k:

$$\mathscr{L}_\theta(Z_{k+1}|Z_1^k) \neq \mathscr{L}_{\theta'}(Z_{k+1}|Z_1^k), \quad \theta \neq \theta'.$$

The distribution $\{V_t\}_{0\leq t\leq\Delta}$ depends on all parameters (otherwise the model is overparametrized). Typically, this implies that the distributions of H_1 and Z_1 depend on all parameters as well, such that the identifiability condition is satisfied even for $k = 0$ (that is, where the above conditional distributions above are replaced by marginals).

Another important property is that the k'th order approximate maximum likelihood estimator is *invariant to data transformations*: if g is a bijective function from $\mathbb{R}$ to some subset of $\mathbb{R}$ then the estimator based on $g(Z_1), \ldots, g(Z_n)$ is the same as that based on $Z_1, \ldots, Z_n$.

Finally some remarks on how to choose k (see also Section 6). Since for increasing k, U_n^k takes more of the dependence structure of the model into account, it might be useful to plot the autocorrelation functions for various transformations of the data (like the data squared and the absolute values of the data). If the empirical autocorrelation coefficients from lag k_0 and onwards are negligible then it seems reasonable not to use k much larger than k_0. If we for some k_0 have caught the important features of the distribution then U_n^k should be close to $U_n^{k_0}$ for $k > k_0$. Hence, so should the corresponding estimates and one may try increasing values of k until the parameter estimates and the minimal values of U_n^k stabilize.

4 Computational aspects

In this section we discuss how to compute $U_n^k(\theta)$ in practice for a fixed but arbitrary value of θ. Let us first focus on calculation of $p_\theta^{k+1}(\tilde{z}_1^{k+1})$ for arbitrary $\tilde{z}_1,\dots,\tilde{z}_{k+1}\in\mathbb{R}$. An expression for $U_n^k(\theta)$ follows almost immediately.

Replace n in formula (5) by $k+1$ in order to write $p_\theta^{k+1}(\tilde{z}_1^{k+1})$ as an expectation

$$p_\theta^{k+1}(\tilde{z}_1^{k+1}) = \mathrm{E}_{\pi_\theta^{k+1}} \prod_{j=1}^{k+1} \varphi(\tilde{z}_j, M_j, S_j) \tag{9}$$

with respect to the distribution of $\{(M_j,S_j)\}_{j=1,\dots,k+1}$. As above $\varphi(\cdot,m,s)$ is the density of $N(m,s)$. We compute (9) as an average

$$\frac{1}{R}\sum_{r=1}^{R}\prod_{j=1}^{k+1}\varphi\left(\tilde{z}_j, M_j^{(r)}, S_j^{(r)}\right) \tag{10}$$

of R simulated product values, where

$$\left(\left(M_1^{(r)},S_1^{(r)}\right),\dots,\left(M_{k+1}^{(r)},S_{k+1}^{(r)}\right)\right),\quad r=1,\dots,R$$

are independent simulations of $\{(M_j,S_j)\}_{j=1,\dots,k+1}$. By choosing R large enough, (9) can be computed to any accuracy in this way.

The r'th simulation is calculated via a simulation, $V^{(r)}$, of the volatility process V from time zero to time $(k+1)\Delta$ as follows. First, the initial value of $V^{(r)}$ is chosen according to the stationary distribution, $V_0^{(r)}\sim\mu_\theta$. Next, split the interval $[0,(k+1)\Delta]$ into $N(k+1)\Delta$ subintervals of length $\delta=1/N$ and calculate values $V_{l\delta}^{(r)}$, $1\le l\le N(k+1)\Delta$ recursively by the Millstein scheme (say),

$$\begin{aligned}V_{l\delta}^{(r)} &= V_{(l-1)\delta}^{(r)} + b\left(V_{(l-1)\delta}^{(r)},\theta\right)\delta + \sigma\left(V_{(l-1)\delta}^{(r)},\theta\right)\varepsilon_l^{(r)}\\ &\quad+\frac{1}{2}\sigma\left(V_{(l-1)\delta}^{(r)},\theta\right)\sigma'\left(V_{(l-1)\delta}^{(r)},\theta\right)\left(\left(\varepsilon_l^{(r)}\right)^2-\delta\right),\quad 1\le l\le N(k+1)\Delta\end{aligned}$$

where $\sigma'=\partial\sigma/\partial v$ is the derivative of σ with respect to the state variable and the innovations $\varepsilon_1^{(r)},\dots,\varepsilon_{(k+1)N}^{(r)}$ are independent, identically $N(0,\delta)$-distributed random variables. Finally, let

$$M_j^{(r)} = \frac{1}{\delta}\sum_{l=(j-1)N}^{jN-1}\xi\left(V_{l\delta}^{(r)}\right),\qquad S_j^{(r)}=\frac{1}{\delta}\sum_{l=(j-1)N}^{lN-1}V_{l\delta}^{(r)}$$

be the simple left Riemann approximations of M_j and S_j, $j=1,\dots,k+1$.[2]

The *same simulations* of M_j and S_j can be used to calculate $p_\theta^k(\tilde{z}_1^k)$; simply replace the product in (10) from 1 to $k+1$ by the product from 1 to k. Even

[2]Of course, one could use better approximations to the integrals. It would probably not improve the calculation much though, since (*i*) the simple approximation introduces no systematic error, and (*ii*) we do not know how the simulated path would behave had we simulated it at points in between the $l\delta$'s.

more important, we can use the same simulations for *all arguments* $\tilde{z}_1^{k+1}$ and thus calculate $U_n^k(\theta)$ as

$$-\frac{1}{n}\sum_{i=k}^{n-1}\log\frac{1}{R}\sum_{r=1}^{R}\prod_{j=1}^{k+1}\varphi_{i-k+j,j}^{(r)}+\frac{1}{n}\sum_{i=k+1}^{n-1}\log\frac{1}{R}\sum_{r=1}^{R}\prod_{j=1}^{k}\varphi_{i-k+j,j}^{(r)} \tag{11}$$

where $\varphi_{i,j}^{(r)}$ is short for $\varphi(z_i, M_j^{(r)}, S_j^{(r)})$, see (8). In particular, we need to simulate V *up to time* $(k+1)\Delta$ *only* in order to compute U_n^k.

There are several "parameters" to choose: the number R of repetitions, the number N of subintervals per Δ-interval, and of course the number of lags k. We already commented on how to choose k in the end of Section 3. The parameters N and (in particular) R determine how accurately the values of U_n^k are determined and must be large enough that the calculation of $U_n^k(\theta)$ is suitably stable.

The number of calculations needed to compute one single value of $U_n^k(\theta)$ increases approximately linearly in both R and $k+1$, and if computing time is limited one must compromize between stability and the number of lags involved. Note that it might be necessary to increase R as k increases since we must simulate longer paths of V and thus might need more simulations to obtain numerical stability.

So-called *antithetic variables* may increase computational stability. Here it means that we make simulations of V in pairs where we in one simulation use the randomly generated ε's in the Millstein scheme and in the other one use *minus* the ε's. For R sets of randomly generated ε's we thus compute $2R$ simulated paths of V, compute the $\varphi^{(r)}$-values in (11) for each of the $2R$ simulated paths of V, and average over all $2R$ simulations. The two $\varphi^{(r)}$-values corresponding to the same set of ε's (plus and minus) tend to be negatively correlated. The computing time is approximately doubled when we use antithetic variables, but hopefully we need R less than half as big as without antithetic variables in order to obtain same precision.

It is indeed possible to compute suitably accurate values of U_n^k in reasonable time: for $n = 500$ observations from the model where $\xi \equiv 0$ and V is a Cox-Ingersoll-Ross process, it takes somewhat less than a minute to compute a value of U_n^4 with $N = 10$ and $R = 10.000$ on a Digital alpha running at 500 MHz. This is only to give an idea of the computational burden — no attempts have been made as to optimize the routine.

Finally a very important remark: As always when criterion functions (or estimating functions) are simulated, it is crucial to use the *same random numbers for different values of* θ. Otherwise R must be chosen extremely large for the simulated criterion function to behave continuously.

5 Asymptotic results

In this section we prove consistency and asymptotic normality (as $n \to \infty$) of the estimator $\hat{\theta}_n^k$ satisfying $U_n^k(\hat{\theta}_n^k) = \inf_{\theta\in\Theta} U_n^k(\theta)$. The results hold for *any fixed values of* k *and* Δ. The true parameter is denoted by θ_0, and all results are with respect to P_{θ_0}.

Note that the first term in (7) is negligible as n increases so we can focus on the sum $\frac{1}{n}\sum_{i=k+1}^{n-1} u_\theta^{c,k}(z_{i+1}|z_{i-k+1}^i)$. In the following we let $||\cdot||$ denote

the usual Eucledian norm on $\mathbb{R}^p$. Also, for $d \geq 1$ we let P_θ^d denote the P_θ-distribution of $(Z_1, \ldots, Z_d)$ and for a function $g : \mathbb{R}^d \to \mathbb{R}$ and a probability Q on $\mathbb{R}^d$ we write $Q(g)$ for the integral $\int g\, dQ$.

5.1 Consistency

Let $k \in \{0, 1, \ldots\}$ and $\Delta > 0$ be fixed. Apart from Assumption 2.1 we need the following regularity conditions for consistency of $\hat{\theta}_n^k$.

Assumption 5.1 The following conditions hold:

(B1) the parameter space Θ is a compact subset of $\mathbb{R}^p$;

(B2) for all $\theta \in \Theta$ there exist a constant $\delta_\theta > 0$ and a function $\bar{u}_\theta : \mathbb{R}^{k+1} \to \mathbb{R}$ in $L^1(P_{\theta_0}^{k+1})$ such that $\sup_{\theta' \in T_{\theta,\delta_\theta}} \left| u_{\theta'}^{c,k}(z_{k+1}|z_1^k) \right| \leq \bar{u}_\theta(z_1^{k+1})$ for all states $z_1, \ldots, z_{k+1} \in \mathbb{R}$ where $T_{\theta,\delta} = \{\theta' \in \Theta : ||\theta - \theta'|| \leq \delta\}$;

(B3) the function $\theta \to u_\theta^{c,k}(z_{k+1}|z_1, \ldots, z_k)$ from Θ to $\mathbb{R}$ is continuous for all $z_1, \ldots, z_{k+1} \in \mathbb{R}$;

(B4) For any $z_1, \ldots, z_k \in \mathbb{R}$ the conditional distributions of Z_{k+1} given $Z_1^k = z_1^k$ with respect to P_θ^{k+1} and $P_{\theta'}^{k+1}$ are different for $\theta \neq \theta'$. □

Note that conditions (B1) and (B3) ensure that a minimum of U_n^k exists, but the minimum could be attained at the boundary of Θ and it need not be unique. Condition (B2) expresses that $u_\theta^{c,k}$ is locally dominated integrable wrt. $P_{\theta_0}^{k+1}$. Condition (B4) is an identifiability condition ensuring that the limit (in P_{θ_0}-probability) of U_n^k has unique minimum at θ_0; see the proof below.

Theorem 5.2 *Under Assumptions 2.1 and 5.1, $\hat{\theta}_n^k$ is consistent for θ_0, that is, $\hat{\theta}_n^k \to \theta_0$ in probability wrt. P_{θ_0} as $n \to \infty$.*

Proof The proof is quite standard and follows Dacunha-Castelle & Duflo (1986, Chapter 3), for example.

First, note that condition (B2) implies that $u_\theta^{c,k}$ is in $L^1(P_{\theta_0}^{k+1})$ for all $\theta \in \Theta$. The ergodic theorem thus yields

$$U_n^k(\theta) \to P_{\theta_0}^{k+1}(u_\theta^{c,k}) = \mathrm{E}_{\theta_0} u_\theta^{c,k}\big(Z_{k+1}|Z_1, \ldots, Z_k\big)$$

as $n \to \infty$ in P_{θ_0}-probability (even P_{θ_0}-a.s and in $L^1(P_{\theta_0})$). Denote the limit by $J^k(\theta)$. By conditions (B2) and (B3), J^k is continuous. Furthermore, condition (B4) implies that J^k has unique minimum at θ_0. Indeed, by definition of J^k and Jensen's inequality

$$\begin{aligned} J^k(\theta_0) - J^k(\theta) &\leq \log \mathrm{E}_{\theta_0}\left(\frac{p_\theta^{c,k}(Z_{k+1}|Z_1^k)}{p_{\theta_0}^{c,k}(Z_{k+1}|Z_1^k)}\right) \\ &= \log \int_{\mathbb{R}^k} p_{\theta_0}^k(z_1^k) \int_{\mathbb{R}} p_\theta^{c,k}(z_{k+1}|z_1^k)\, dz_{k+1}\, dz_1^k = 0 \end{aligned} \tag{12}$$

for all $\theta \in \Theta$ with equality if and only if $\theta = \theta_0$.

Next, define $W_n(\eta) = \sup_{||\theta-\theta'||\leq\eta} \left|U_n^k(\theta) - U_n^k(\theta')\right|$, $\eta > 0$. By the triangle inequality

$$\begin{aligned} W_n(\eta) &\leq \sup_{||\theta-\theta'||\leq\eta} \Big(\left|U_n^k(\theta) - J^k(\theta)\right| + \left|J^k(\theta) - J^k(\theta')\right| + \left|U_n^k(\theta') - J^k(\theta')\right| \Big) \\ &\leq 2 \sup_{\theta\in\Theta} \left|U_n^k(\theta) - J^k(\theta)\right| + \sup_{||\theta-\theta'||\leq\eta} \left|J^k(\theta) - J^k(\theta')\right|. \end{aligned}$$

Here, the second term is deterministic and converges to zero as $\eta \to 0$ since J^k is continuous and defined on a compact set (condition (B1)). The first term converges to zero in P_{θ_0}-probability. The proof of this is almost identical to that of Lemma 3.3 in Bibby & Sørensen (1995); see Sørensen (2000, Lemma III.5.3) for further details. Consistency of $\hat{\theta}_n^k$ now follows from standard arguments. □

5.2 Asymptotic normality

In the following we shall assume that the criterion function U_n^k is twice continuously differentiable with derivative $\dot{U}_n^k$ which takes values in $\mathbb{R}^p$. A minimizer of U_n^k is then either on the boundary of Θ or solves the equation $\dot{U}_n^k(\theta) = 0$. Below we give a result on existence and asymptotic normality of a solution to the estimating equation.

First, let us be more specific about the differentiability assumption and introduce some further notation.

Assumption 5.3 Let Θ° denote the inner of Θ and assume that $\theta_0 \in \Theta^\circ$. Assume furthermore that the function $\theta \to u_\theta^{c,k}(z_{k+1}|z_1^k)$ is twice continuously differentiable on Θ° for all $z_1, \ldots, z_{k+1} \in \mathbb{R}$. □

Let $\dot{u}_\theta^{c,k} = (\dot{u}_{\theta,j}^{c,k})_{j=1,\ldots,p} = (\partial u_\theta^{c,k}/\partial\theta_j)_{j=1,\ldots,p}$ denote the p-vector of first derivatives and $\ddot{u}_\theta^{c,k} = (\ddot{u}_{\theta,jl}^{c,k})_{j,l=1,\ldots,p} = (\partial^2 u_\theta^{c,k}/\partial\theta_j\partial\theta_l)_{j,l=1,\ldots,p}$ be the symmetric $p\times p$-matrix of second derivatives of $u_\theta^{c,k}$ wrt. the parameter. With this notation (and without the first, negligible term) the approximate score function $\dot{U}_n^k$ is given by $\frac{1}{n}\sum_{i=k+1}^{n-1} \dot{u}_\theta^{c,k}(z_{i+1}|z_{i-k+1}^i)$.

Note that $\dot{U}_n^k$ is an unbiased estimating function, that is, $\mathrm{E}_\theta \dot{U}_n^k(\theta) = 0$ for all $\theta \in \Theta^\circ$. Indeed,

$$\mathrm{E}_\theta\, \dot{u}_{\theta,j}^{c,k}(Z_{k+1}|Z_1^k) = \mathrm{E}_\theta\, \mathrm{E}_\theta \Big(\dot{u}_{\theta,j}^{c,k}(Z_{k+1}|Z_1^k) \big| Z_1^k \Big)$$

and, with obvious notation for the derivatives of $p_\theta^{c,k}$ (and if differentiation wrt. θ_j and integration wrt. z_{k+1} are interchangeable),

$$\mathrm{E}_\theta \Big(\dot{u}_{\theta,j}^{c,k}(Z_{k+1}) \big| Z_1^k = z_1^k \Big) = -\int \dot{p}_{\theta,j}^{c,k}(z|z_1^k)\, dz = -\frac{\partial}{\partial\theta_j} \int p_\theta^{c,k}(z|z_1^k)\, dz = 0$$

for all $z_1, \ldots, z_k \in \mathbb{R}$ and all $j = 1, \ldots, p$.

It is essential that the estimating function $\dot{U}_n^k$, scaled properly and evaluated at the true parameter, converges in distribution. We shall impose a version of the central limit theorem based on α-mixing (Hall & Heyde 1980)

which was also used by Genon-Catalot *et al.* (1998). For a stochastic process $Y = \{Y_t\}_{t\in T}$ in discrete time ($T = \mathbb{N}\cup\{0\}$) or continuous time ($T = [0,\infty)$), define the *α-mixing coefficients* by

$$\alpha_Y(t) = \sup_{t'\geq 1}\sup_{A,B}\big|Pr(A\cap B) - Pr(A)Pr(B)\big|, \quad t\in T$$

where the second supremum is taken over sets A and B from the σ-algebras generated by $(Y_s)_{s\leq t'}$ and $(Y_s)_{s\geq t'+t}$ respectively. We can think of the α-mixing coefficients as measures of the temporal dependence in Y, and Y is said to be α-mixing if $\alpha_Y(t)\to 0$ as $t\to\infty$. See Doukhan (1994) for an exposition on the general theory of mixing.

If the α-mixing coefficients for Z decrease to zero fast enough and if $\dot{u}^{c,k}_{\theta_0}$ has moments of sufficiently high order (condition (C1) below) then $n^{1/2}\dot{U}^k_n(\theta_0)$ converges in distribution (Hall & Heyde 1980). Note that the α-mixing coefficients for V and Z satisfy $\alpha_Z(m)\leq\alpha_V((m-1)\Delta)$ for all $m\geq 1$ (Genon-Catalot *et al.* 1998), so it is sufficient to show that $\alpha_V(m\Delta)$ decrease at a geometric rate, say. There are well-known, sufficient conditions for this to hold (Genon-Catalot *et al.* 1998).

We need some further regularity conditions: an identifiability assumption (condition (C3)) and locally dominated integrability of $\ddot{u}^{c,k}_{\theta}$ (condition (C2).

Assumption 5.4 Assume that

(C1) there exists an $\eta>0$ such that $\dot{u}^{c,k}_{\theta_0,j}$ is in $L^{2+\eta}(P^{k+1}_{\theta_0})$ for all $j=1,\ldots,p$ and such that the α-mixing coefficients for Z corresponding to θ_0 satisfy the condition $\sum_{m=1}^{\infty}\alpha_Z(m)^{2/(2+\eta)}<\infty$;

(C2) there is a neighbourhood T_0 of θ_0 such that for all $\theta\in T_0$ and all $j,l=1,\ldots,p$ there is a constant $\delta_{\theta,jl}>0$ and a function $\bar{u}_{\theta,jl}:\mathbb{R}^{k+1}\to\mathbb{R}$ in $L^1(P^{k+1}_{\theta_0})$ such that for all $z_1,\ldots,z_{k+1}\in\mathbb{R}$, $\sup_{\theta'\in T_{\theta,\delta_{\theta,jl}}}\big|\ddot{u}^{c,k}_{\theta',jl}(z_{k+1}|z_1^k)\big|\leq \bar{u}_{\theta,jl}(z_1^{k+1})$ where, as before, $T_{\theta,\delta}=\{\theta'\in\Theta:||\theta-\theta'||\leq\delta\}$;

(C3) the symmetric $p\times p$ matrix $A^k(\theta_0)=P^{k+1}_{\theta_0}\big(\ddot{u}^{c,k}_{\theta_0}\big)=\mathrm{E}_{\theta_0}\ddot{u}^{c,k}_{\theta_0}(Z_{k+1}|Z_1^k)$ is positive definite. □

Theorem 5.5 *Suppose that Assumptions 2.1, 5.3 and 5.4 hold. Then a solution $\hat{\theta}^k_n$ to $\dot{U}^k_n(\theta)=0$ exists with a probability tending to 1 as $n\to\infty$. Moreover*

$$\sqrt{n}\big(\hat{\theta}^k_n-\theta_0\big)\to N\big(A^k(\theta_0)^{-1}\Gamma^k(\theta_0)A^k(\theta_0)^{-1}\big) \tag{13}$$

in distribution wrt. P_{θ_0} where the $p\times p$ matrix $\Gamma^k(\theta_0)$ is given by

$$\begin{aligned}\Gamma^k(\theta_0) = {}& \mathrm{E}_{\theta_0}\dot{u}^{c,k}_{\theta_0}(Z_{k+1}|Z_1^k)\dot{u}^{c,k}_{\theta_0}(Z_{k+1}|Z_1^k)^T\\ &+\sum_{m=1}^{\infty}\mathrm{E}_{\theta_0}\dot{u}^{c,k}_{\theta_0}(Z_{k+1}|Z_1^k)\dot{u}^{c,k}_{\theta_0}(Z_{k+m+1}|Z_{m+1}^{m+k})^T\\ &+\sum_{m=1}^{\infty}\mathrm{E}_{\theta_0}\dot{u}^{c,k}_{\theta_0}(Z_{k+m+1}|Z_{m+1}^{m+k})\dot{u}^{c,k}_{\theta_0}(Z_{k+1}|Z_1^k)^T\end{aligned}$$

and $A^k(\theta_0)$ is given in condition (C5).

Proof It follows from Corollary 2.5 and Theorem 2.8 in Sørensen (1998) that it suffices to show

$$n^{1/2}\dot{U}_n^k(\theta_0) \to N\big(0,\Gamma^k(\theta_0)\big) \tag{14}$$

in distribution wrt. P_{θ_0} as $n \to \infty$ and

$$\sup_{\theta \in T_{\theta_0,\eta/\sqrt{n}}} \left|\ddot{U}_{n,jl}^k(\theta) - A^k(\theta_0)\right| \to 0 \tag{15}$$

in probability wrt. P_{θ_0} as $n \to \infty$ for all $\eta > 0$ and all $j,l \in \{1,\dots,p\}$.

As already noted (14) follows from condition (C1) and a version of the central limit theorem (Hall & Heyde 1980). In particular the sums in $\Gamma^k(\theta_0)$ are finite so that $\Gamma^k(\theta_0)$ is well-defined.

In order to show (15), define $A^k(\theta) = P_{\theta_0}^{k+1}\big(\ddot{u}_\theta^{c,k}\big)$ for $\theta \in T_0$ (well-defined because of condition (C2)) and let $j,l \in \{1,\dots,p\}$ and $\eta > 0$ be fixed. By the triangle inequality

$$\left|\ddot{U}_{n,jl}^k(\theta) - A^k(\theta_0)\right| \le \left|\ddot{U}_{n,jl}^k(\theta) - A^k(\theta)\right| + \left|A^k(\theta) - A^k(\theta_0)\right|.$$

Choose N large enough that $T_{\theta_0,\eta/\sqrt{N}} \subseteq T_0$. Then, for $n \ge N$, $A^k(\theta)$ is well-defined for all $\theta \in T_{\theta_0,\eta/\sqrt{n}}$. One can now show that

$$\sup_{\theta \in T_{\theta_0,\eta/\sqrt{N}}} \left|\ddot{U}_{n,jl}^k(\theta) - A^k(\theta)\right| \to 0$$

in P_{θ_0}-probability as $n \to \infty$. The proof of this is almost identical to that of Lemma 3.3 in Bibby & Sørensen (1995); recall that $T_{\theta_0,\eta/\sqrt{N}}$ is compact. Also, A^k is continuous in θ_0. The convergence (15) follows immediately. This proves both the existence assertion and the weak convergence result in (13). □

Note that although asymptotic normality is indeed a nice property of the estimator, it is difficult to use in practice as we are not able to compute the asymptotic variance. Also, the above conditions are all expressed in terms of the distribution of Z and thus in general difficult (if possible at all) to check. As noted above, condition (C3) is an exception.

Finally, let us stress that the above results hold for *fixed* value of k (and Δ) as $n \to \infty$. In particular, the above results do *not* imply nice asymptotic behaviour of the maximum likelihood estimator (which corresponds to $k = k(n) = n-1$). The problem is of course that the terms in the log-likelihood function U_n^{n-1} originate from *different* functions ($p_\theta^{c,i}$ for observation z_{i+1}) such that the usual limit theorems do not apply. As noted in Section 2 we can think of the model as a hidden Markov model with continuous, unbounded state space of the hidden chain $\tilde{H}$ given by $\tilde{H}_i = (V_{i\Delta}, M_i, S_i)$. Asymptotic results for the maximum likelihood estimator have been proved for hidden Markov models for which the state space of the hidden chain is either finite (Bickel & Ritov 1996, Bickel, Ritov & Rydén 1998) or compact (Jensen & Petersen 1999). Neither approach can be applied in our setting and there are in fact no results in the literature concerning asymptotic properties of the maximum likelihood estimator for the models considered in this paper.

6 Efficiency considerations

In this section we briefly discuss how the number of lags k influence the quality of the estimators. The subject is essential but unfortunately we have not been able to prove any very powerful results.

With the asymptotic results in mind, note that the number of lags, k, is a question of efficiency rather than bias. Intuitively we would expect the estimators to improve as k increases since further characteristics of the distribution are taken into account. Put differently, we would expect the asymptotic variance of $\hat{\theta}_n^k$ to decrease with k (with the usual definition of "$\leq$" for symmetric matrices: $A \leq B$ if and only if the difference $B - A$ is positive semi-definite). We have not been able to prove results like this! The problem is of course that the expression for the asymptotic variance is so complicated that comparison between different k's is impossible, even for a one-dimensional parameter. Anyway, even *if* efficiency increases with k, it should in applications be taken into account that computation time increases with k as well.

The simulation study in Section 7 indicates that minimization of U_n^k in practice may give rise to identification problems even if the k-lag conditional distribution uniquely determines the parameter (theoretically). The problem seems to diminish as we use larger values of k suggesting that estimation becomes easier (and improves in this particular sense) as k increases. On the other hand: in a simpler situation with no identification problems for any value of k we did not find any substantial differences among the estimators for different values of k.

In principle we could improve estimation by introducing weight functions as follows. Consider estimating functions on the form

$$D_n^k(\theta) = \frac{1}{n} \sum_{i=k}^{n-1} d_i(Z_{i-k+1}^i, \theta)\, \dot{u}_\theta^{c,k}(Z_{i+1}|Z_{i-k+1}^i)$$

where $d_k, \ldots, d_{n-1}$ are function from $\mathbb{R}^k \times \Theta$ to $\mathbb{R}$. Note that we for simplicity have left out the contribution from the first k observations and that $\dot{U}_n^k$ (except for the first k observations) corresponds to $d_i \equiv 1$, $i = k, \ldots, n-1$. The estimating function D_n^k is unbiased since for each $i = k \ldots, n-1$

$$\begin{aligned} &\mathrm{E}_{\theta_0} d_i(Z_{i-k+1}^i, \theta_0) u_{\theta_0}^{c,k}(Z_{i+1}|Z_{i-k+1}^i) \\ &\quad = \mathrm{E}_{\theta_0} d_i(Z_{i-k+1}^i) \left(\mathrm{E}_{\theta_0} u_{\theta_0}^{c,k}(Z_{i+1}|Z_{i-k+1}^i)|Z_{i-k+1}^i \right) = 0. \end{aligned}$$

Under regularity conditions similar to those of Assumptions 5.3 and 5.4 the solution to $D_n^k(\theta) = 0$ is a consistent and asymptotically normal estimator of θ. By choosing the functions d_i cleverly we can obtain smaller asymptotic variance than is the case for $\hat{\theta}_n^k$, see Sørensen (1999) for similar considerations. This is only of theoretical interest, though, since it even in simple cases is impossible to determine the optimal weights!

Finally, we prove a result concerning the approximate log-likelihood functions U_n^k rather than the corresponding estimators: the limit, in probability, of $U_n^k(\theta_0)$ is decreasing in k. It holds for U_n^k evaluated at the true parameter only and is thus not very useful in practice. Nevertheless it tells us that the approximations of the likelihood improve in this sense.

Proposition 6.1 *Let $0 \le k' \le k''$ and assume that Condition (B2) is satisfied for $\theta = \theta_0$ and $k = k'$ and $k = k''$. Then*

$$\mathrm{E}_{\theta_0} u_{\theta_0}^{c,k''}\big(Z_{k''+1}|Z_1^{k''}\big) \le \mathrm{E}_{\theta_0} u_{\theta_0}^{c,k'}\big(Z_{k'+1}|Z_1^{k'}\big).$$

Consequently, $\mathrm{E}_{\theta_0} U_n^{k''}(\theta_0) \le \mathrm{E}_{\theta_0} U_n^{k'}(\theta_0)$ *and* $\lim_{n\to\infty} U_n^{k''}(\theta_0) \le \lim_{n\to\infty} U_n^{k'}(\theta_0)$ *where convergence means convergence in P_{θ_0}-probability.*

Proof It will suffice to consider $k' = k$ and $k'' = k+1$ for $k \ge 0$ arbitrary. By stationarity it follows that it is sufficient to show that

$$\mathrm{E}_{\theta_0} u_{\theta_0}^{c,k+1}\big(Z_{k+2}|Z_1^{k+1}\big) \le \mathrm{E}_{\theta_0} u_{\theta_0}^{c,k}\big(Z_{k+2}|Z_2^{k+1}\big). \tag{16}$$

By definition,

$$u_{\theta_0}^{c,k+1}\big(Z_{k+2}|Z_1^{k+1}\big) - u_{\theta_0}^{c,k}\big(Z_{k+1}|Z_1^{k}\big) = \log \frac{p_{\theta_0}^{c,k}\big(Z_{k+2}|Z_2^{k+1}\big)}{p_{\theta_0}^{c,k+1}\big(Z_{k+2}|Z_1^{k+1}\big)}$$

so Jensen's equality yields

$$\mathrm{E}_{\theta_0}\Big(u_{\theta_0}^{c,k+1}\big(Z_{k+2}|Z_1^{k+1}\big) - u_{\theta_0}^{c,k}\big(Z_{k+2}|Z_2^{k+1}\big)\Big) \le \log \mathrm{E}_{\theta_0} \frac{p_{\theta_0}^{c,k}\big(Z_{k+2}|Z_2^{k+1}\big)}{p_{\theta_0}^{c,k+1}\big(Z_{k+2}|Z_1^{k+1}\big)}.$$

Calculations similar to those leading to (12) show that the latter expectation is one, which yields (16). The expectation assertion follows immediately by

$$U_n^{k+1}(\theta_0) - U_n^{k}(\theta_0) = \frac{1}{n}\sum_{k=1}^{n-1}\Big(u_{\theta_0}^{c,k+1}\big(Z_{i+1}|Z_{i-k}^{i}\big) - u_{\theta_0}^{c,k}\big(Z_{i+1}|Z_{i-k+1}^{i}\big)\Big)$$

and the convergence result follows by the ergodic theorem. □

7 Example: The Cox-Ingersoll-Ross model

We now discuss a particular model and present a tiny simulation study based on ten simulated datasets. Of course, the results from such a small experiment are by no means conclusive, but at best indicative, of the properties of the estimators. The study demonstrates, however, that the approximate likelihood method is indeed applicable in practice.

Consider the model where the observable process X has no drift and the volatility process V is a *Cox-Ingersoll-Ross process*. This specification of the volatility process was first considered by Hull & White (1987) and later by Heston (1993). The model is given by the stochastic differential equations

$$\begin{aligned} dX_t &= \sqrt{V_t}\, dW_t \\ dV_t &= \alpha(\beta - V_t)\, dt + \sigma\sqrt{V_t}\, d\tilde{W}_t \end{aligned}$$

with parameter $\theta = (\alpha, \beta, \sigma)$ varying in $\Theta = \{(\alpha,\beta,\sigma) : \alpha,\beta,\sigma > 0,\, \sigma^2 \le 2\alpha\beta\}$. The parameter β is simply the mean value of V whereas the "mean reverting parameter" α can be interpreted as the size of the force pulling

the process back to its mean. It is well-known that for any $\theta \in \Theta$, V is positive, stationary, ergodic with geometrically decreasing α-mixing coefficients. The invariant distribution is the Gamma distribution with shape parameter $2\alpha\beta/\sigma^2$ and scale parameter $\sigma^2/(2\alpha)$, and we assume that V is started according to this distribution.

It is easy to calculate various moments of the distribution of Z: for any $i, j \in \mathbb{N}$ with $j > i$ we have $\mathrm{E}_\theta Z_i = 0$, $\mathrm{Var}_\theta Z_i = \beta\Delta$, $\mathrm{Cov}_\theta(Z_i, Z_j) = 0$, and

$$\mathrm{Var}_\theta Z_i^2 = 2\beta^2\Delta^2 + \frac{3\beta\sigma^2}{\alpha^3}\big(\alpha\Delta - 1 + e^{-\alpha\Delta}\big)$$
$$\mathrm{Cov}_\theta(Z_i^2, Z_j^2) = \frac{\beta\sigma^2}{2\alpha^3} e^{-\alpha\Delta(j-i-1)}\big(1 - e^{-\alpha\Delta}\big)^2.$$

From these expressions it follows that the correlation between Z_i^2 and Z_j^2 is at most 1/5 for all $j > i$ and that the excess kurtosis of the invariant distribution of Z is at most 3. Hence, the model is not appropriate for data with very heavy tails or with large correlations between squared observations.

Now, let us turn to the simulation study. It consists of ten simulated datasets from the above model, each consisting of $n = 500$ observations. The model parameter is $(\alpha, \beta, \sigma) = (\alpha_0, \beta_0, \sigma_0) = (0.1, 1, 0.35)$ and the value of Δ is 1. For all computations of U_n^k below we have used $N = 10$ and $R = 10.000$, cf. Section 4. The top of Figure 1 shows one of the simulated datasets of increments. The bottom of the figure shows the corresponding path of the volatility V from time 0 to time 500. Clearly the increments are more volatile in periods with relatively large values of the volatility process V than in periods with low values of V. Figure 2 is a QQ-plot of the increments; they are clearly too heavy-tailed to be Gaussian. We shall use the data from Figure 1 as example throughout the section.

In the following we consider two different set-ups, namely the one where only one parameter, say α, is unknown whereas the two others are known (Section 7.1) and the one where all three parameters are unknown (Section 7.2). The first situation is of course not realistic but it provides insight to the behaviour of the estimators. We refer to Sørensen (2000, Section III.7) for further details on the experiment and the results, and also for comments on the set-up where two parameters must are unknown and one is known (essentially the conclusions are as for estimation of α solely when α or σ is the known parameter and as for estimation of all parameters when β is the known parameter).

7.1 Estimation of one parameter only

We choose α as the unknown parameter and consider $\beta = \beta_0 = 1$ and $\sigma = \sigma_0 = 0.35$ known. Recall that the true value of α is $\alpha_0 = 0.1$.

Figure 3 shows the graphs of U_n^k, $k = 0, \ldots, 4$ on the interval from 0.06 to 0.16 for the data from Figure 1. For this particular dataset the curvatures of U_n^3 and U_n^4 are almost identical, and very similar to the curvature of U_n^2 and U_n^1. The minimum points are thus close. The function U_n^0 has different curvature and minimum for a somewhat lower value of α. Note that $U_n^4 \leq U_n^3 \leq U_n^2 \leq U_n^0 \leq U_n^1$ at θ_0 — almost in line with Proposition 6.1.

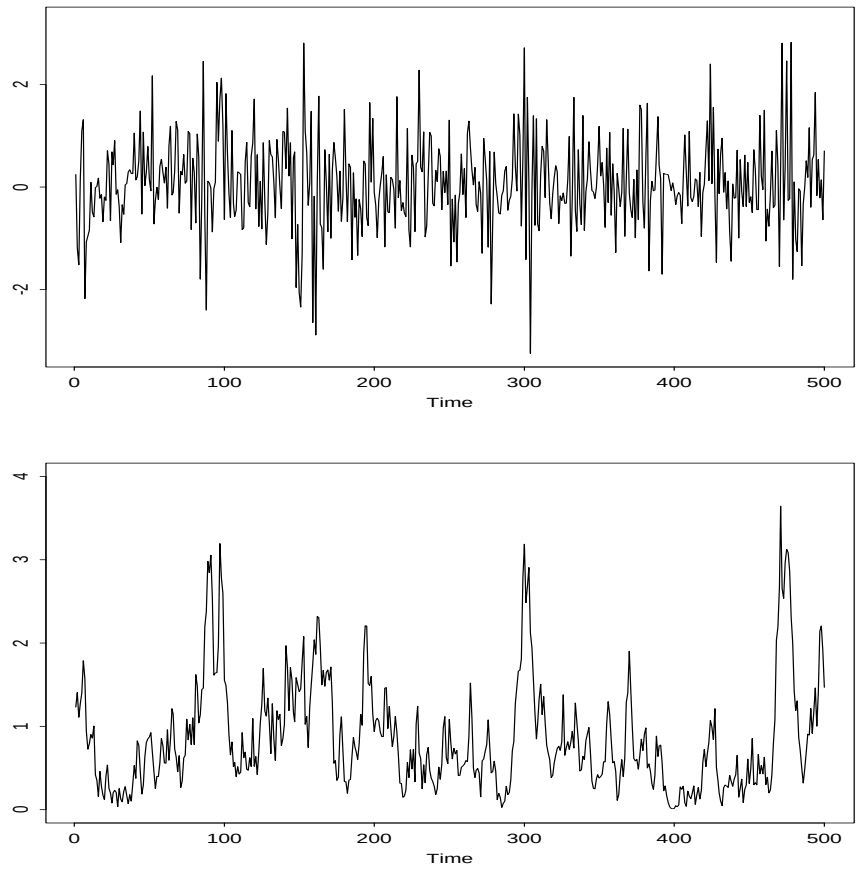

Figure 1: Simulated values of $Z_i = X_{i\Delta} - X_{(i-1)\Delta}$ (top) and $V_{i\Delta}$ (bottom) from the Cox-Ingersoll-Ross model for $\Delta = 1$ and $i = 1, \ldots, n$ where $n = 500$. The model parameter is $(\alpha, \beta, \sigma) = (0.1, 1, 0.35)$.

The estimation results are illustrated in the first five "columns" of Figure 4, where each circle denotes a value of the estimator. All five values of k yield reasonable estimates, with averages from 0.1027 ($k = 1$) to 0.1101 ($k = 0$) and standard errors from 0.0169 ($k = 1$) to 0.0281 ($k = 0$). In particular, the estimator $\hat{\alpha}_n^1$ is the best — and $\hat{\alpha}_n^0$ the worst — in this study with respect to both bias and variance. The difference between the five estimators is not substantial, though, and it is difficult to recognize any pattern in the differences.

For comparison we have also calculated two different moment estimators, that is, estimators obtained by matching theoretical and empirical moments (Genon-Catalot *et al.* 1998). Note that neither the first three moments of Z nor $\mathrm{E}_\theta Z_1 Z_j$, $j \geq 2$ can be used since they do not depend on α. Instead we have used the fourth order moments $\mathrm{E}_\theta Z_1^4$ and $\mathrm{E}_\theta Z_1^2 Z_2^2$ respectively. There is a considerable bias and the standard errors are huge. Also, the estimating equations have no solution for two of the datasets.

Finally, we have estimated α from the volatility data $V_0, V_\Delta, \ldots, V_{n\Delta}$. Maximum likelihood estimation is in principle possible since the transition probabilities are known (non-central χ^2-distributions), but for simplicity we have

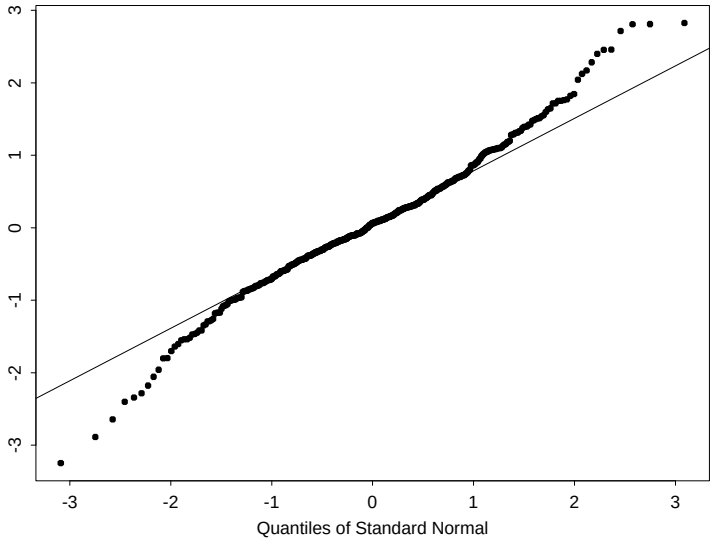

Figure 2: QQ-plot for the data in the top of Figure 1; quantiles of the standard normal distribution at the x-axis, quantiles of data at the y-axis.

used the (optimal) martingale estimating function based on the conditional expectation one step ahead (Bibby & Sørensen 1995, Sørensen 1997). The martingale estimates are plotted in the last "column" of Figure 4. The average of $\hat{\alpha}_n^V$ is 0.1097. As one would expect, $\hat{\alpha}_n^V$ has smaller standard error (0.0154) than the estimators based on Z. It is slightly surprising, though, that the standard error is only roughly 10% lower than that of $\hat{\alpha}_n^1$.

7.2 Estimation of all three parameters

Estimation of α was succesful even for $k = 0$ and $k = 1$. At first glance it seems promising to use $k = 1$ for estimation of all three parameters as well: by the moment considerations above it follows that the three-dimensional parameter is uniquely determined by the distribution of the pair (Z_1, Z_2) — and thereby presumably also by the conditional distribution of Z_2 given Z_1. In practive it turns out that U_n^1 has severe difficulties distinguishing between parameter values for which the invariant distribution of V is the same. This distribution is determined by two parameters only. Hence, we must use a larger value of k in order to obtain reasonable estimates.

We choose $k = 4$. It is important to find good initial points for the numerical minimization routine. Moment estimators (Genon-Catalot *et al.* 1998) are extremely easy to compute but unfortunately there are serious existence problems: the equations requiring that the theoretical and empirical versions of $\mathrm{E}_\theta Z_1^2$, $\mathrm{E}_\theta Z_1^4$ and $\mathrm{E}_\theta Z_1^2 Z_2^2$ agree, have no solution for eight of the ten datasets. (Also, we know from Section 7.1 that moment estimators may be quite bad.) We are thus forced to come up with better alternatives.

Inspired by Genon-Catalot *et al.* (1999) we approximate the invariant

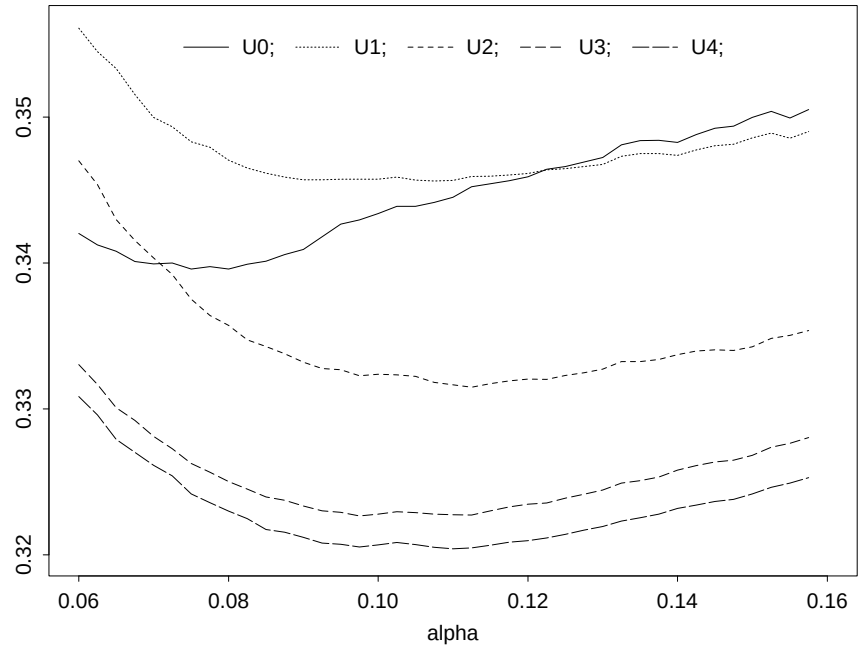

Figure 3: Graphs of $\alpha \to U_n^k(\alpha, \beta_0, \sigma_0)$ for the data from Figure 1, $k = 0, \ldots, 4$, $\beta_0 = 1$ and $\sigma_0 = 0.35$. The true value of α is $\alpha_0 = 0.1$.

distriution of S with a Γ-distribution. Denote by λ and τ the shape and scale parameters. Since the P_θ-expectation of S_1 is $\beta\Delta$ we let $\tau = \beta\Delta/\lambda$. If we furthermore require the variance of $\Gamma(\lambda, \beta\Delta/\lambda))$ to equal the P_θ-variance of S_1, then we obtain σ^2 as a function of α, β and λ:

$$\sigma^2 = \sigma^2(\alpha, \beta, \lambda) = \frac{\alpha^3 \beta \Delta}{\lambda\left(\alpha\Delta - 1 + e^{-\alpha\Delta}\right)}. \tag{17}$$

Estimation is now performed as follows: (*i*) let $\tilde{\beta}_n = \sum_{i=1}^n Z_i^2/\Delta$ be a preliminary estimate of β; (*ii*) estimate λ by pretending that $Z_1, \ldots, Z_n$ are independent and identically distributed with distribution equal to that of $\sqrt{\gamma}\varepsilon$ where γ and ε are independent with $\gamma \sim \Gamma(\lambda, \tilde{\beta}_n\Delta/\lambda)$ and $\varepsilon \sim N(0,1)$; (*iii*) minimize $\alpha \to U_n^4(\alpha, \tilde{\beta}_n, \sigma(\alpha, \tilde{\beta}_n, \tilde{\lambda}_n))$ in order to get preliminary estimates $\tilde{\alpha}_n$ and $\tilde{\sigma}_n^2 = \sigma^2(\tilde{\alpha}_n, \tilde{\beta}_n, \tilde{\lambda}_n)$ of α and σ^2; (*iv*) finally minimize U_n^4 over Θ with initial values $(\tilde{\alpha}_n, \tilde{\beta}_n, \tilde{\sigma}_n)$.

The two last steps are illustrated in Figure 5 which shows the level curves of $(\alpha, \sigma^2) \to U_n^4(\alpha, \tilde{\beta}_n, \sigma)$ for the dataset from Figure 1 together with the dashed curve $(\alpha, \sigma^2(\alpha, \tilde{\beta}_n, \tilde{\lambda}_n))$. Here $\tilde{\beta}_n = 0.7528$ and $\tilde{\lambda}_n = 1.6732$. The minimum along the curve is attained for $\tilde{\alpha}_n = 0.1631$, and the corresponding value of σ is $\tilde{\sigma}_n = \sqrt{0.1549} = 0.3936$. The point $(0.1631, 01549)$ is denoted by a solid circle in Figure 5. In step (*iv*) the minimization routine moves from the initial point (0.1631,0.7528,0.3936) to the global minimum point

$$\left(\hat{\alpha}_n^4, \hat{\beta}_n^4, \hat{\sigma}_n^4\right) = (0.1040, 0.7441, 0.2571).$$

Note that the estimate of β changes (slightly) in step (*iv*), too. The point $(\hat{\alpha}_n^4, (\hat{\sigma}_n^4)^2)$ is shown with a circle in Figure 5.

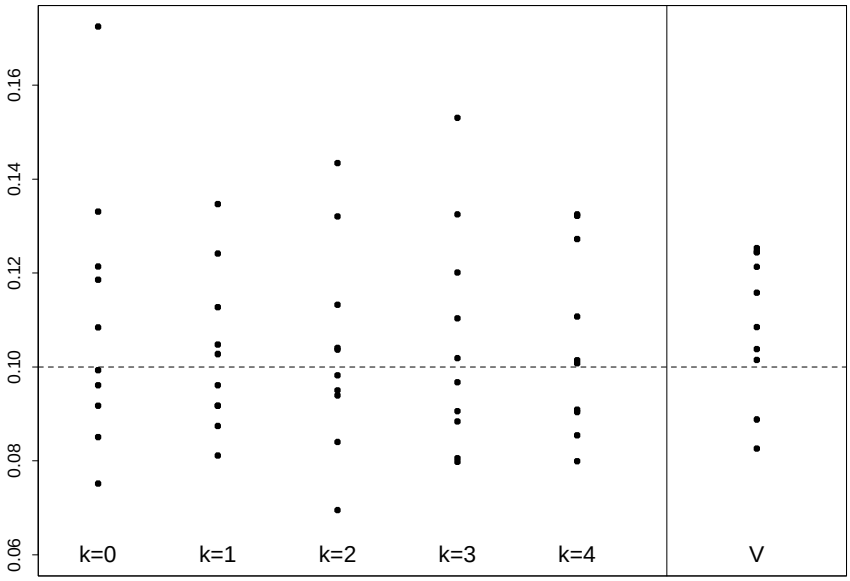

Figure 4: The estimators $\hat{\alpha}_n^k$ for $k = 0, \ldots, 4$ (the first five "columns") and the martingale estimator $\hat{\alpha}_n^V$ based on V (the last "column"). The true value of α is $\alpha_0 = 0.1$ (shown by the dashed line).

The estimators $\hat{\alpha}_n^4$, $\hat{\beta}_n^4$ and $\hat{\sigma}_n^4$ are shown in the first, third and fifth "columns" of Figure 6. The averages are 0.1113, 1.0037 and 0.3036 respectively. This is not too bad. However, for three of the datasets, the estimators of α and σ are very bad; this is reflected in huge standard errors.

In conclusion, $k = 4$ works reasonably well for seven of the ten datasets. Of course we could have used other values of k, and informal studies indicate that $k = 3$ would have worked fairly well for three of the simulations and $k = 2$ for two simulations. In other words estimation seems to improve as k increases. This leaves us with some hope that estimation would improve even further if we used more than four lags. The hope is strengthened by inspection of the correlograms (not shown here) of the squared observations for the three datasets that behaved badly for $k = 4$: all three datasets have relatively large correlations (compared to the other datasets) on several lags larger than four, indicating that U_n^4 does not capture all information in data.

Finally, it is easy to compute estimators based on the volatility process as solutions to simple martingale estimating equations (Sørensen 1997). These estimators are superior to the approximate maximum likelihood estimators based on Z; see Figure 6. Recall however that V would not be observed in applications so martingale estimation based on V would not be an option! Rather than giving up the idea of approximate maximum likelihood estimation, we conclude that quite a lot of information is lost when Z is observed instead of V.

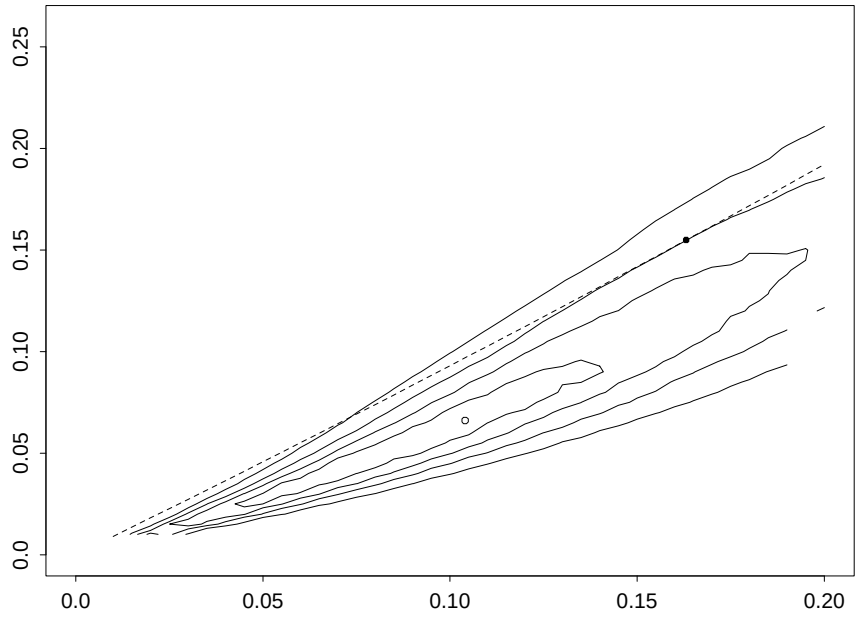

Figure 5: Level curves of $(\alpha, \sigma^2) \to U_n^4(\alpha, \tilde{\beta}_n, \sigma)$ for the data from Figure 1; α on the x-axis and σ^2 on the y-axis. The dashed curve is given by (17) with $\beta = \tilde{\beta}_n = 0.7528$ and $\lambda = \tilde{\lambda}_n = 1.6732$. The solid circle denotes the minimum point $(\tilde{\alpha}_n, \tilde{\sigma}_n^2) = (0.1631, 0.1549)$ along the dashed curve, and the circle denotes the global minimum — when β varies as well — $(\hat{\alpha}_n^4, (\hat{\sigma}_n^4)^2) = (0.1040, 0.0661)$. The true value of (α, σ^2) is $(\alpha_0, \sigma_0^2) = (0.1, 0.1225)$.

8 Concluding remarks

We have discussed approximate maximum likelihood estimation for increments $Z_1, \ldots, Z_n$ from a certain class of stochastic volatility models. The k'th order approximation to the likelihood function was obtained by pretending that $Z_1, \ldots, Z_n$ are independent ($k = 0$) or k'th order Markov ($k \geq 1$). The corresponding estimator is consistent and asymptotically normal for any $k \geq 0$, essentially because we use the *true* conditional densities given the k previous observations.

The estimation procedure is applicable to other data types with a complicated dependence structure, in particular for (other) hidden Markov models. The essential properties making simulation of the approximate likelihood functions easy are the following: (*i*) given the values of an unobservable process, the observations $Z_1, \ldots, Z_n$ are independent with a known distribution (up to some parameter) determined by the latent process; (*ii*) the unobserved process is easy to simulate for all values of the parameter. Note that the first property does not hold for stochastic volatility models if the Brownian motions driving X and V, respectively, are correlated or if the drift or the diffusion term for X depends on X itself. Hence, such models cannot be handled by the approach from this paper.

Finally, let us stress that there are other possible approximations to the

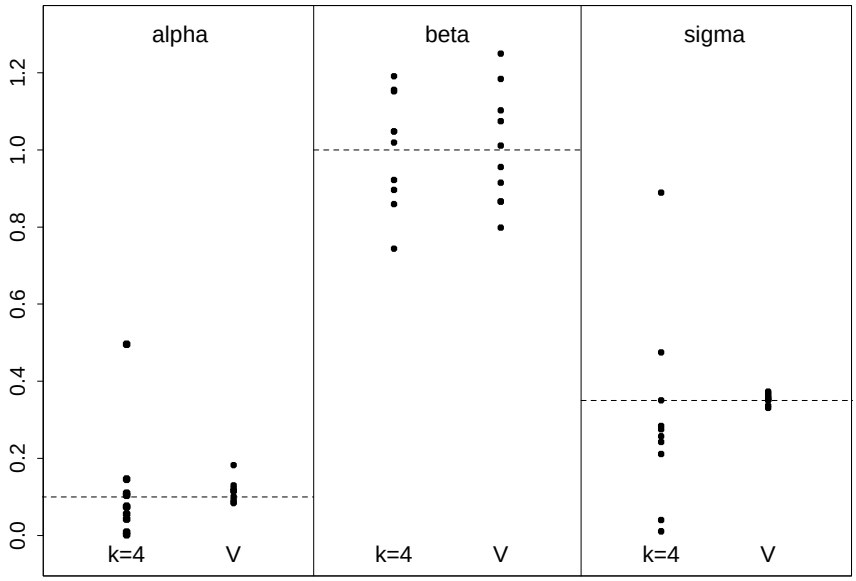

Figure 6: The approximate maximum likelihood estimators $\hat{\alpha}_n^4$, $\hat{\beta}_n^4$, $\hat{\sigma}_n^4$ in "columns" 1, 3 and 5, and the martingale estimators $\hat{\alpha}_n^V$, $\hat{\beta}_n^V$, $\hat{\sigma}_n^V$ in "columns" 2, 4 and 6. The true values (0.1, 1 and 0.35) are shown with the dashed lines.

likelihood function than those based on the k-lag conditional densities. For example, one could split data into tuples of some length, and pretend that the tuples are independent (Rydén 1994). Or one could both condition forwards and backwards in time, *i.e.* base estimation on the conditional densities $p_\theta^{c,k}(Z_i|Z_{i-k},\dots,Z_{i-1},Z_{i+1},\dots,Z_{i+k})$ given the k previous and the k subsequent observations. We would get asymptotically well-behaved estimators by these approximations as well. However, since time runs forward, we feel that the approximations based on conditioning backwards in time only, are the most natural.

Acknowledgements I wish to thank my advisor Martin Jacobsen for many valuable discussions and for many helpful comments on the manuscript.

References

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

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

Bickel, P. J. & Ritov, Y. (1996), Inference in hidden Markov models I: Local asymptotic normality in the stationary case, *Bernoulli* **2**, 199–228.

Bickel, P. J., Ritov, Y. & Rydén, T. (1998), Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models, *Ann. Statist.* **26**, 1614–1635.

Black, F. & Scholes, M. (1973), The pricing of options and corporate liabilities, *J. Political Economy* **81**, 637–654.

Chesney, M. & Scott, L. (1989), Pricing European currency options: a comparison of the modified Black-Scholes model and a random variance model, *J. Financial and Quantitative Analysis* **24**, 267–284.

Dacunha-Castelle, D. & Duflo, M. (1986), *Probability and statistics*, Vol. 2, Springer-Verlag, New York.

Doukhan, P. (1994), *Mixing: properties and examples*, Lecture Notes in Statistics **85**, Springer-Verlag, New York.

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

Eraker, B. (1998), MCMC analysis of diffusion models with application to finance, Discussion paper 1998-5, Department of Finance and Management Science, Norwegian School of Economics and Business Administration.

Gallant, A. R. & Tauchen, G. (1996), Which moments to match?, *Econometric Theory* **12**, 657–681.

Genon-Catalot, V., Jeantheau, T. & Laredo, C. (1998), Stochastic volatility models as hidden Markov models and statistical applications, Preprint 1998-22, Equipe d'Analyse et de Mathématiques Appliquées, Université de Marne-la-Vallée.

Genon-Catalot, V., Jeantheau, T. & Laredo, C. (1999), Parameter estimation for discretely observed stochastic volatility models, *Bernoulli* **5**, 855–872.

Ghysels, E., Harvey, A. C. & Renault, E. (1996), Stochastic volatility, *in* G. S. Maddala & C. R. Rao, eds, *Statistical methods in finance*, Vol. 14 of *Handbook of Statistics*, North-Holland, Amsterdam, pp. 119–191.

Hall, P. & Heyde, C. C. (1980), *Martingale limit theory and its application*, Academic Press, New York.

Heston, S. L. (1993), A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Rev. Financial Studies* **6**, 327–343.

Hull, J. & White, A. (1987), The pricing of options on assets with stochastic volatilities, *J. Finance* **42**, 281–300.

Hull, J. & White, A. (1988), An analysis of the bias in option pricing caused by a stochastic volatility, *Advances in Futures and Options Research* **3**, 29–61.

Jensen, J. L. & Petersen, N. V. (1999), Asymptotic normality of the maximum likelihood estimator in state space models, *Ann. Statist.* **27**, 514–535.

Karatzas, I. & Shreve, S. E. (1991), *Brownian motion and stochastic calculus*, 2nd edn, Springer-Verlag, New York.

Karlin, S. & Taylor, H. M. (1981), *A second course in stochastic processes*, Academic Press, New York.

Nielsen, J. N., Vestergaard, M. & Madsen, H. (2000), Estimation in continuous-time stochastic volatility models using nonlinear filters. To appear in *Intl. J. Theoret. Appl. Finance*.

Rydén, T. (1994), Consistent and asymptotically normal parameter estimates for hidden Markov models, *Ann. Statist.* **22**, 1884–1895.

Scott, L. O. (1987), Option pricing when the variance changes randomly: theory, estimation and an application, *J. Financial and Quantitative analysis* **22**, 419–438.

Shephard, N. (1996), Statistical aspects of ARCH and stochastic volatility, *in* D. R. Cox, D. V. Hinkley & O. E. Barndorff-Nielsen, eds, *Time series models in econometrics, finance and other fields*, Chapman & Hall, London, pp. 1–67.

Sørensen, H. (2000), Inference for diffusion processes and stochastic volatility models, PhD thesis, Department of Statistics and Operations Research, University of Copenhagen.

Sørensen, M. (1997), Estimating functions for discretely observed diffusions: A review, *in* I. V. Basawa, V. P. Godambe & R. L. Taylor, eds, *Selected proceedings of the symposium on estimating functions*, Vol. 32, IMS Lecture notes, pp. 305–325.

Sørensen, M. (1998), On asymptotics of estimating functions, Preprint 1998-6, Department of Theoretical Statistics, University of Copenhagen. To appear in *Brazilian J. Probability and Statistics*.

Sørensen, M. (1999), Prediction-based estimating functions, Preprint 1999-5, Department of Theoretical Statistics, University of Copenhagen.

Stein, E. M. & Stein, J. C. (1991), Stock price distributions with stochastic volatility: an analytic approach, *Rev. Financial Studies* **4**, 727–752.

Wiggins, J. B. (1987), Options values under stochastic volatility, *J. Financial Economics* **19**, 351–372.