Estimation in the Cox-Ingersoll-Ross Model

Author(s): Ludger Overbeck and Tobias Rydén

Source: *Econometric Theory*, Jun., 1997, Vol. 13, No. 3 (Jun., 1997), pp. 430-461

Published by: Cambridge University Press

Stable URL: http://www.jstor.com/stable/3532742

Econometric Theory, 13, 1997, 430–461. Printed in the United States of America.

ESTIMATION IN THE COX–INGERSOLL–ROSS MODEL

LUDGER OVERBECK AND TOBIAS RYDÉN
University of California

The Cox–Ingersoll–Ross model is a diffusion process suitable for modeling the term structure of interest rates. In this paper, we consider estimation of the parameters of this process from observations at equidistant time points. We study two estimators based on conditional least squares as well as a one-step improvement of these, two weighted conditional least-squares estimators, and the maximum likelihood estimator. Asymptotic properties of the various estimators are discussed, and we also compare their performance in a simulation study.

1. INTRODUCTION

In a seminal paper, Cox, Ingersoll, and Ross (1985) proposed a model, the so-called Cox–Ingersoll–Ross (CIR) model, for the term structure of interest rates. The simplest version of this model describes the dynamics of the interest rate $X(t)$ as the solution of the stochastic differential equation

$$dX(t) = (a + bX(t))\,dt + \sigma\sqrt{X^+(t)}\,dW(t), \tag{1}$$

where W is a standard Brownian motion, $x^+ \equiv \max(x,0)$ is the positive part, and $a > 0$, $b < 0$, and $\sigma > 0$. The appealing properties of this process from an applied point of view are as follows:

(i) The interest rate stays nonnegative.
(ii) It converges to a steady-state law with mean $-a/b$, the so-called *long-term value*, with speed of adjustment b.
(iii) The incremental variance is proportional to the current value of X.

The CIR model is still used in the financial world; for example, there is ongoing research on option pricing in this model (see Abken, 1993; Chen and Scott, 1992). In the latter reference, a multifactor model is considered, but because this model is based on a finite collection of independent one-

We thank the editor and the two referees for several constructive comments that improved the paper. Overbeck was supported by a Fellowship of the Deutsche Forschungsgemeinschaft while on leave from the Universität Bonn. Rydén was supported by the Swedish Natural Science Research Council (contract F-PD 10538-301), the Fulbright Commission, and Kungliga Fysiografiska Sällskapet i Lund. Address correspondence to: Tobias Rydén, Department of Mathematical Statistics, Lund University, Box 118, 221 00 Lund, Sweden; e-mail: tobias@maths.lth.se.

  0266-4666/97 $9.00 + .10

factor models defined as in (1), our results may be carried over to the multifactor case.

The process defined by (1) is a so-called *Bessel process*, because it can be viewed as the modulus of an Ornstein–Uhlenbeck process (cf. Pitman and Yor, 1982). It is also a continuous-space branching process with immigration (see Feller, 1951; Kawazu and Watanabe, 1971).

The objective of the present paper is the estimation of the parameters a, b, and σ^2 from observations of $\{X(t)\}$ at equidistant time points $t_k = k\Delta t$, $k = 0, \ldots, n$. Led by the approach to estimation in discrete-time and -space branching processes with immigration in Wei and Winnicki (1990) and Winnicki (1990), we use as a basic method conditional least squares and derive two estimators, different in the way σ is estimated, from this framework. The least-squares estimators have (at least) two appealing properties. First, they are consistent, asymptotically normal and easy to compute. Second, the estimators of the parameters a and b of the drift term in (1) are robust against misspecification of the diffusion term, a property that is not shared by the maximum likelihood estimator (MLE).

Furthermore, we will prove local asymptotic normality (LAN) of our model, a property that gives a Cramér–Rao-type bound on the variance of estimators of the parameters. Additionally, it yields a one-step improvement of the least-squares estimator, leading to an estimator that is efficient; that is, it attains the Cramér–Rao bound asymptotically.

Other approaches to estimation in the CIR model can be found in Longstaff and Schwartz (1992), Chen and Scott (1993), and Bibby and Sørensen (1995). The first approach is based on an approximation of the CIR model by a discrete-time GARCH model, however, whose parameters may not directly be mapped to the parameters of the continuous-time model, as remarked in Longstaff and Schwartz (1992, footnote 20). Chen and Scott (1993) employed maximum likelihood, which requires some computational efforts, because there is no closed-form expression for the MLE. Indeed, the success of the ML approach is highly dependent on the availability of good starting values for the numerical optimization algorithm, and this point can be chosen as the least-squares estimator of the present paper. Alternatively, one may initialize the optimization at several different starting points, as is done in Longstaff and Schwartz (1992, p. 1274), but this increases the computational burden even more. The log likelihood is not globally concave, although the LAN condition implies that under some conditions it is, asymptotically, locally concave around the true parameter. Instead of computing the full MLE, one may, as already remarked, make a one-step improvement of a least-squares estimator, which gives an estimator that is asymptotically efficient.

Bibby and Sørensen (1995) considered three approximations of the likelihood function but restricted attention to estimation of a and b for known σ, although the two estimators given by (2.5) and (2.10) in their paper can actually be computed without knowledge of σ. These estimators can both be

viewed as weighted least-squares estimators, and they are therefore robust against misspecification of the diffusion term. The latter estimator is the so-called *maximum quasilikelihood estimator* (MQLE) (see Godambe and Heyde, 1987; Wefelmeyer, 1996a). The third estimator considered by Bibby and Sørensen (1995, equation (2.13)) is of little interest for estimation in the CIR model, because it is an approximation of the MQLE.

An interesting feature of the CIR model is that for certain parameter constellations, the LAN property does not hold. More precisely, some elements of the Fisher information matrix of the model may become infinite, which indicates a kind of superefficiency of the MLE.

Finally, we also mention that parameter estimation for stochastic differential equations may be carried out using so-called *indirect inference*, involving simulation (see Gouriéroux, Monfort, and Renault, 1993; Gouriéroux and Monfort, 1995). Such methods are not considered in the present paper, however.

2. PRELIMINARIES

In this section, we give some basic properties of the diffusion process $\{X(t)\}$, defined by (1), and the sampled version $\{X_k\} \equiv \{X(t_k)\}$, where $t_k = k\Delta t$. The sampling interval Δt is assumed to be known and is fixed throughout the paper. The parameters of the model will be denoted by $\theta = (a, b, \sigma^2)$, and this triple will belong to the space $\Theta = (0,\infty) \times (-\infty,0) \times (0,\infty)$.

According to Revuz and Yor (1991, p. 362), stochastic differential equation (1) has a unique strong solution for every parameter $\theta \in \Theta$ and starting point $X_0 = x_0 > 0$. Furthermore, the origin is inaccessible if $2a/\sigma^2 \geq 1$ (cf. Pitman and Yor, 1982; Feller, 1951). If $0 < 2a/\sigma^2 < 1$, then the origin is instantaneously reflecting, and for $a = 0$ (a case that is not considered in our paper) zero is absorbing and is hit in finite time almost surely. Thus, the process is always nonnegative and we could erase the + under the square root in (1).

In Kawazu and Watanabe (1971), the branching structure of $\{X(t)\}$ is explored to prove that its conditional Laplace functional is given by

$$\begin{aligned}\psi(x,\lambda) &\equiv E_\theta[e^{-\lambda X(t)} \mid X(0) = x] \\ &= \left\{1 - \frac{\sigma^2\lambda}{2b}(1 - \exp(bt))\right\}^{-2a/\sigma^2} \\ &\quad \times \exp\left\{-x\lambda e^{bt}\left(1 - \frac{\sigma^2\lambda}{2b}(1 - \exp(bt))\right)^{-1}\right\}. \qquad (2)\end{aligned}$$

Letting $t \to \infty$ in this expression, it follows that $X(t)$ converges weakly to a Gamma distribution with parameters $\alpha = -2b/\sigma^2$ and $p = 2a/\sigma^2$, that is, a distribution with density $\alpha^p x^{p-1} \exp\{-\alpha x\}/\Gamma(p)$. Let us denote this

density by $\gamma(x;\theta)$. The mean of the invariant distribution is $-a/b$, and its variance is $a\sigma^2/2b^2$. The results of this paper will be derived under the assumption that $X(0)$ is distributed according to the stationary law, so that $\{X(t)\}$ is a stationary and ergodic process, but by a fairly simple (continuous-time) coupling argument it can be seen that they are valid for arbitrary initial distributions.

The sampled process $\{X_k\}$ has of course the same stationary law as $\{X(t)\}$, and from (2) one may derive its transition density $p(x,y)$. It is given by

$$p(x,y;\theta) \equiv \frac{\partial}{\partial y} P_\theta(X_k \le y | X_{k-1} = x)$$

$$= c \exp\{-cy - ce^{b\Delta t}x\} \left(\frac{y}{xe^{b\Delta t}}\right)^{q/2} I(2c\sqrt{yxe^{b\Delta t}}, q), \tag{3}$$

where

$$c = -\frac{2b}{\sigma^2(1 - e^{b\Delta t})},$$

$$q = 2a/\sigma^2 - 1,$$

and

$$I(z,q) = \sum_{j=0}^{\infty} \frac{(z/2)^{q+2j}}{j!\Gamma(q + j + 1)}$$

is a modified Bessel function of the first kind of order q (see Revuz and Yor, 1991; Pitman and Yor, 1982; Chen and Scott, 1993).

3. CONDITIONAL LEAST-SQUARES ESTIMATION

The concept of conditional least squares, which is a general approach for estimation of the parameter ϕ involved in the conditional mean function $E[X_k|X_{k-1}]$ of a stochastic process, was given a thorough treatment by Klimko and Nelson (1978). The conditional least-squares estimator $\hat{\phi}_n$ minimizes the sum of squares $\sum_1^n (X_k - E_\phi[X_k|\mathcal{F}_{k-1}])^2$, where $\mathcal{F}_k$ is the σ-field generated by $X_1, \ldots, X_k$. In our case, $\{X_k\}$ is a Markov process, whence $\mathcal{F}_{k-1}$ can be replaced by $\sigma(X_{k-1})$ in this expression. The conditional mean function for the CIR model is easily derived by averaging in (1) and is given by

$$m(x;\theta) \equiv E_\theta[X_k|X_{k-1} = x] = \gamma_0 + \gamma_1 x, \tag{4}$$

with

$$\gamma_0 = \frac{a}{b}(e^{b\Delta t} - 1),$$

$$\gamma_1 = e^{b\Delta t}. \tag{5}$$

Thus, only the parameters a and b appear in $m(x;\theta)$, that is, $\phi = (a,b)$. The discrete-time model may be written as

$$X_k = \gamma_0 + \gamma_1 X_{k-1} + \varepsilon_k,$$

where $\{\varepsilon_k\}$ is a martingale increment sequence with respect to $\{\mathcal{F}_k\}$; that is, ε_k is $\mathcal{F}_k$-measurable and $E_\theta[\varepsilon_k|\mathcal{F}_{k-1}] = 0$. In some simpler models, such as the one with linear drift and diffusion term $\sigma\, dW(t)$, one may solve the stochastic differential equation and hence derive an explicit expression for ε_k (see Theorem 2 in Bergstrom, 1984). In this case, $\{\varepsilon_k\}$ is even independent and identically distributed (i.i.d.). For the CIR model, no such solution exists, although the conditional density of ε_k given $\mathcal{F}_{k-1}$ is easily obtained from (3), and the conditional variance $E_\theta[\varepsilon_k^2|\mathcal{F}_{k-1}]$ is given later.

The conditional least-squares estimator of a and b can be found by minimizing the sum of squares first over a and then over b. The resulting expressions are

$$\hat{b}_n = \frac{1}{\Delta t} \log \left\{ \frac{\frac{1}{n}\sum_{k=1}^{n} (X_k - \bar{X}_n)(X_{k-1} - \bar{X}'_n)}{\frac{1}{n}\sum_{k=1}^{n} (X_{k-1} - \bar{X}'_n)^2} \right\} \tag{6}$$

and

$$\hat{a}_n = \frac{\bar{X}_n - e^{\hat{b}_n \Delta t} \bar{X}'_n}{e^{\hat{b}_n \Delta t} - 1}\, \hat{b}_n, \tag{7}$$

where $\bar{X}_n = n^{-1}\sum_1^n X_k$ and $\bar{X}'_n = n^{-1}\sum_1^n X_{k-1}$.

A natural extension of the basic conditional least-squares approach is to consider weighted versions. Obviously, the least-squares estimator is the unique solution of the system of equations $\sum_1^n (\partial/\partial\phi)(X_k - m(X_{k-1};\phi))^2 = 0$, and the corresponding weighted expression is $\sum_1^n w(X_{k-1};\theta)(\partial/\partial\phi)(X_k - m(X_{k-1};\phi))^2 = 0$. The estimator defined by (2.5) in Bibby and Sørensen (1995), henceforth denoted the BSE, is obtained by letting $w(x;\theta) = 1/(\sigma^2 x)$, and the MQLE is obtained by letting $w(x;\theta) = 1/v(x;\theta)$. Here, $v(x;\theta)$ is the conditional variance of X_k given $X_{k-1} = x$. These weight functions are both proportional to $1/\sigma^2$ (see (8)), so that the corresponding estimates of $\phi = (a,b)$ can be computed without knowledge of σ. There is an explicit expression for the BSE (cf. (2.18) in Bibby and Sørensen, 1995), but not for the MQLE. The latter is less time-consuming to compute than the MLE, however.

We now turn to the estimation of σ^2. For this, we need to compute $v(x;\theta)$. The conditional second moment function can be found either by applying Itô's formula to the function $x \mapsto x^2$ in (1) and averaging or by dif-

ferentiating (2) with respect to λ twice, and from this the conditional variance follows. It is given by

$$v(x;\theta) \equiv E_\theta[(X_k - E_\theta[X_k|X_{k-1} = x])^2|X_{k-1} = x] = \sigma^2(\eta_0 + \eta_1 x), \tag{8}$$

with

$$\eta_0 = \frac{a}{2b^2}(e^{b\Delta t} - 1)^2,$$
$$\eta_1 = \frac{1}{b}e^{b\Delta t}(e^{b\Delta t} - 1). \tag{9}$$

Several natural estimators of σ^2 exist. Inspired by standard linear regression, one may consider

$$\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n}\frac{\{X_k - m(X_{k-1};\hat{a}_n,\hat{b}_n)\}^2}{\hat{\eta}_0 + \hat{\eta}_1 X_{k-1}}, \tag{10}$$

where $\hat{\eta}_0$ is η_0 evaluated at $(\hat{a}_n,\hat{b}_n)$ and so on (possibly one may replace n^{-1} by $(n-2)^{-1}$). This idea is what Carroll and Ruppert (1988) call the *pseudo-likelihood method.*

Another approach is to consider (8) as a regression equation with $(X_k - E_\theta[X_k|X_{k-1}])^2$ being the response, $\eta_0 + \eta_1 X_{k-1}$ being the predictor, and σ^2 being the unknown. The minimizer of the corresponding sum of squares is

$$\sigma_n^2 = \frac{\sum_{k=1}^{n}(\eta_0 + \eta_1 X_{k-1})\{X_k - m(X_{k-1};a,b)\}^2}{\sum_{k=1}^{n}(\eta_0 + \eta_1 X_{k-1})^2}. \tag{11}$$

Here we must of course substitute a and b by the corresponding estimates, which gives us the estimator

$$\hat{\sigma}_n^2 = \frac{\frac{1}{n}\sum_{k=1}^{n}(\hat{\eta}_0 + \hat{\eta}_1 X_{k-1})\{X_k - m(X_{k-1};\hat{a}_n,\hat{b}_n)\}^2}{\frac{1}{n}\sum_{k=1}^{n}(\hat{\eta}_0 + \hat{\eta}_1 X_{k-1})^2}. \tag{12}$$

Carroll and Ruppert (1988) call this method *unweighted least squares based on squared residuals.* Of course, σ^2 can also be estimated as in (10) or (12) with $(\hat{a}_n,\hat{b}_n)$ replaced by either the BSE or the MQLE.

An estimator similar to $\hat{\sigma}_n^2$ was proposed by Winnicki (1990) for estimating variances in branching processes with immigration. In that case, however, the predictors are completely known, which is enough to conclude that the asymptotic normal laws of σ_n^2 and $\hat{\sigma}_n^2$ agree. In our case, these normal laws differ (cf. later). For regression with i.i.d. errors, this phenomenon was

observed by Davidian and Carroll (1987, Corollary 4.1(a)). Remarks analogous to the preceding ones apply to $\tilde{\sigma}_n^2$.

3.1. Consistency

THEOREM 3.1. *The estimator* $(\hat{a}_n, \hat{b}_n)$ *is strongly consistent.*

Proof. By ergodicity, the expression in brackets in (6) converges to the autocorrelation function of $\{X_k\}$ at lag one P_θ-a.s. By (4) and (5), this correlation equals $\exp(b\Delta t)$, and we conclude that $\hat{b}_n$ is strongly consistent.

The fraction in (7) is equal to $-\overline{X}_n + o(1)$, where the notation $Z_n = o(c_n)$ means that $Z_n(\omega) = o(c_n)$ for all ω in some subset of the probability space with P_θ-measure one. Because $\overline{X}_n \to E_\theta[X_0] = -a/b$ P_θ-a.s., it follows that $\hat{a}_n$ is strongly consistent. ■

Remark. It is easy to show that $\hat{a}_n = -\overline{X}_n \hat{b}_n + o(n^{-1/2})$, whence $\hat{a}_n$ can be substituted by $\hat{a}_n' = -\overline{X}_n \hat{b}_n$ in the asymptotic analysis below.

Weak consistency of the BSE and the MQLE follows by Theorem 3.2 in Bibby and Sørensen (1995). For the BSE, however, their Condition 3.2(b)–(c) is violated if $2a/\sigma^2 \leq 2$. A closer analysis of this estimator (see (2.18) in their paper) reveals that the BSE is strongly consistent as long as $E_\theta[1/X(0)]$ is finite, which is the case if $2a/\sigma^2 > 1$. This is essentially the condition that $\{X(t)\}$ must not hit zero.

Because the estimators $\hat{a}_n$ and $\hat{b}_n$ are derived solely from the conditional mean function, they are robust against misspecification of the diffusion term in (1) (cf. Wefelmeyer, 1996a); that is, they are consistent as long as the observed process $\{X_k\}$ is Markovian with finite second moment and satisfies (4) and (5). This property is not shared by the MLE. In our case, (4) and (5) hold as long as the interest rate model has linear drift, but the diffusion term in (1) may be replaced by $\sigma(X(t))\,dW(t)$, where $\sigma(\cdot)$ is an arbitrary function such that the induced stationary distribution of $\{X(t)\}$ has finite second moment. Thus, $\hat{a}_n$ and $\hat{b}_n$ are consistent estimators of a and b in any such model. The BSE and the MQLE are robust against misspecification of the diffusion term, as well.

If the diffusion term is correctly specified, then the MQLE has a certain efficiency property (see Godambe and Heyde, 1987; Wefelmeyer, 1996a), although it may be severely inefficient if this term is misspecified (cf. Crowder, 1987).

THEOREM 3.2. *The estimators* $\tilde{\sigma}_n^2$ *and* $\hat{\sigma}_n^2$ *are strongly consistent.*

Proof. We show consistency only for $\hat{\sigma}_n^2$; the proof for $\tilde{\sigma}_n^2$ is completely similar.

Fix θ, let $h(x, y; \vartheta) = (\eta_0(\vartheta) + \eta_1(\vartheta)x)\{y - m(x;\vartheta)\}^2$, and let $U \subset \Theta$ be a neighborhood of θ such that $E_\theta[\sup_{\vartheta \in U} |h(X_0, X_1; \vartheta)|] < \infty$.

By Theorem 3.1 and ergodicity of $\{X_k\}$, the inequality

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} (\hat{\eta}_0 + \hat{\eta}_1 X_{k-1})\{X_k - m(X_{k-1};\hat{a}_n,\hat{b}_n)\}^2$$

$$\leq \limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \sup_{\vartheta\in U} (\eta_0(\vartheta) + \eta_1(\vartheta)X_{k-1})\{X_k - m(X_{k-1};\vartheta)\}^2$$

$$= E_\theta\left[\sup_{\vartheta\in U} h(X_0, X_1;\vartheta)\right]$$

holds P_θ-a.s. Now, let $U \downarrow \{\theta\}$ and use continuity of η_0, η_1, and m, as functions of ϑ, and dominated convergence to conclude that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} (\hat{\eta}_0 + \hat{\eta}_1 X_{k-1})\{X_k - m(X_{k-1};\hat{a}_n,\hat{b}_n)\}^2$$

$$\leq E_\theta[(\eta_0 + \eta_1 X_0)\{X_1 - m(X_0;\theta)\}^2]$$

$$= \sigma^2 E_\theta[(\eta_0 + \eta_1 X_0)^2] \quad P_\theta\text{-a.s.}$$

An entirely analogous argument with lim inf instead of lim sup proves convergence to the right-hand side. In a similar fashion, the denominator in (12) can be shown in converge to $E_\theta[(\eta_0 + \eta_1 X_0)^2]$ P_θ-a.s., which completes the proof. ■

The argument of the proof can be modified to show that any weakly consistent estimator of (a, b) gives weakly consistent estimators of σ^2. Thus, the BSE and the MQLE may be applied to estimate σ^2 consistently.

3.2. Asymptotic Normality

This section is devoted to an analysis of the asymptotic distributions of the various estimators discussed so far, starting with $\hat{\theta}_n = (\hat{a}_n, \hat{b}_n, \hat{\sigma}_n^2)$.

The estimator $\hat{\theta}_n$ is the unique solution of the estimating equation

$$G_n(\theta) = \sum_{k=1}^{n} \{w_m(X_{k-1};\theta)(X_k - m(X_{k-1};\theta))$$

$$+ w_v(X_{k-1};\theta)[(X_k - m(X_{k-1};\theta))^2 - v(X_{k-1};\theta)]\} = 0, \quad \textbf{(13)}$$

where

$$w_m(x;\theta) = \begin{pmatrix} \dfrac{\partial m(x;a,b)}{\partial a} \\ \dfrac{\partial m(x;a,b)}{\partial b} \\ 0 \end{pmatrix} = \begin{pmatrix} \dfrac{\partial \gamma_0}{\partial a} \\ \dfrac{\partial \gamma_0}{\partial b} + \dfrac{\partial \gamma_1}{\partial b} x \\ 0 \end{pmatrix} \quad \textbf{(14)}$$

and

$$w_v(x;\theta) = \begin{pmatrix} 0 \\ 0 \\ \eta_0 + \eta_1 x \end{pmatrix}. \tag{15}$$

The key to the analysis of $\hat{\theta}_n$ is that $G_n(\theta)$ is a P_θ-martingale. Let

$$\mu_j(x;\theta) = E_\theta[(X_1 - m(x;\theta))^j \mid X_0 = x]$$

denote the jth conditional central moment, and if A is a nonsingular square matrix let $A^{-T} = (A^{-1})^T$, where T denotes transpose.

THEOREM 3.3. $n^{1/2}(\hat{\theta}_n - \theta) \to \mathcal{N}(0, V^{-1}WV^{-T})$ P_θ-*weakly, where*

$$\begin{aligned} W = E_\theta[&v(X_0;\theta)w_m(X_0;\theta)w_m^T(X_0;\theta) + \mu_3(X_0;\theta)w_m(X_0;\theta)w_v^T(X_0;\theta) \\ &+ \mu_3(X_0;\theta)w_v(X_0;\theta)w_m^T(X_0;\theta) \\ &+ \{\mu_4(X_0;\theta) - v^2(X_0;\theta)\}w_v(X_0;\theta)w_v^T(X_0;\theta)] \end{aligned} \tag{16}$$

and

$$V = E_\theta\left[w_m(X_0;\theta)\left(\frac{\partial m(X_0;\theta)}{\partial\theta}\right)^T + w_v(X_0;\theta)\left(\frac{\partial v(X_0;\theta)}{\partial\theta}\right)^T\right]. \tag{17}$$

Proof. The proof is essentially the same as that of Theorem 6.4.1 in Lehmann (1990). Fix θ. By Theorems 3.1 and 3.2, $\hat{\theta}_n$ is strongly consistent. Making a Taylor expansion of G_n about θ, it is enough to show that

$$\frac{1}{n}\frac{\partial}{\partial\theta_j}G_{n,i}(\theta) + \sum_{k=1}^{3}\frac{1}{2n}(\hat{\theta}_{n,k} - \theta_k)\frac{\partial^2}{\partial\theta_j\,\partial\theta_k}G_{n,i}(\bar{\theta}_n) \to V_{ij} \quad P_\theta\text{-a.s.} \tag{18}$$

for each $1 \le i, j \le 3$, and that

$$-n^{-1/2}G_n(\theta) \to \mathcal{N}(0, W) \quad P_\theta\text{-weakly}, \tag{19}$$

where $\bar{\theta}_n$ is a point on the line segment between θ and $\hat{\theta}_n$. Let $g(X_{k-1}, X_k; \cdot)$ denote the terms in (13). It is easy to see that there is a function $h(x,y)$ and a neighborhood U of θ such that (i) $|\partial^2 g(x,y;\vartheta)/\partial\vartheta_i\,\partial\vartheta_j| \le h(x,y)$ for all $1 \le i, j \le 3$, uniformly in $\vartheta \in U$, and (ii) $E_\theta h(X_0, X_1) < \infty$. This shows that we can neglect the sum on the left-hand side of (18). Now, (18) follows from ergodicity and (19) follows from the martingale central limit theorem (see, e.g., Durrett, 1991, Ch. VII, Theorem 7.5). That W and V can be written as in (16) and (17), respectively, follows by some simple algebra. ■

Remark. The preceding result is a multidimensional version of (1.7) in Wefelmeyer (1996b).

Expressions for W and V are given in Appendix A. Of course, W and V can also be estimated from data by

$$\hat{W}_n = \frac{1}{n}\sum_{k=1}^{n} g(X_{k-1}, X_k; \hat{\theta}_n) g^T(X_{k-1}, X_k; \hat{\theta}_n) \tag{20}$$

and

$$\hat{V}_n = \frac{1}{n}\sum_{k=1}^{n} \frac{\partial}{\partial\theta} g(X_{k-1}, X_k; \hat{\theta}_n), \tag{21}$$

where $g(x, y; \theta)$ is as defined in the proof of Theorem 3.3 or by computing the sample means corresponding to (16) and (17), respectively.

The BSE and the MQLE also solve estimating equations of form (13), with $w_m(x;\theta)$ replaced by $x^{-1}w_m(x;\theta)$ and $(\eta_0 + \eta_1 x)^{-1}w_m(x;\theta)$, respectively. For the BSE, the matrices W and V may be given explicitly, but for the MQLE there are no closed-form expressions, although they may of course be evaluated by numerical integration, estimated from data, as earlier, or estimated by bootstrap. For the BSE, $2a/\sigma^2 > 2$ is a necessary condition for asymptotic normality.

We close this section with a short remark on the asymptotic analysis of $\tilde{\theta}_n = (\hat{a}_n, \hat{b}_n, \tilde{\sigma}_n^2)$. This estimator solves an estimating equation of form (13) with $w_m(x;\theta)$ as in (14) and with

$$w_v(x;\theta) = \begin{pmatrix} 0 \\ 0 \\ 1/(\eta_0 + \eta_1 x) \end{pmatrix}.$$

Again, there are no closed-form expressions for W and V. This is the case also if $(\hat{a}_n, \hat{b}_n)$ is replaced by either the BSE or the MQLE.

3.3. Estimation in Continuous Time

For many financial assets, the yield may be observed almost in continuous time, whence it is natural to look for continuous-time analogs of the least-squares estimator. This also quantifies the information loss, in terms of asymptotic variance of the estimators, that is caused by the discrete-time sampling.

Define

$$\hat{b}^c(t) = \frac{\frac{1}{t}\int_0^t X(s)\,dX(s) - \frac{1}{t}(X(t) - X(0))\bar{X}(t)}{\frac{1}{t}\int_0^t (X(s) - \bar{X}(t))^2\,ds} \tag{22}$$

and

$$\hat{a}^c(t) = -\bar{X}(t)\hat{b}^c(t) + \frac{1}{t}(X(t) - X(0)), \tag{23}$$

where $\bar{X}(t) = (1/t)\int_0^t X(s)\,ds$. The following result motivates us to name $(\hat{a}^c(t), \hat{b}^c(t))$ the *continuous-time conditional least-squares estimator* of (a, b).

THEOREM 3.4. *Fix $t > 0$ and let the sampling interval be $\Delta t = t/n$. Then, $(\hat{a}_n, \hat{b}_n) \to (\hat{a}^c(t), \hat{b}^c(t))$ in probability as $n \to \infty$.*

Proof. Let $\Sigma_n^{(1)} = \sum_1^n (X_k - \bar{X}_n)(X_{k-1} - \bar{X}_n')$ and $\Sigma_n^{(2)} = \sum_1^n (X_{k-1} - \bar{X}_n')^2$, so that

$$\hat{b}_n = \frac{1}{\Delta t} \log \frac{\Sigma_n^{(1)}}{\Sigma_n^{(2)}} = \frac{1}{\Delta t} \log\left(1 + \frac{\Sigma_n^{(1)} - \Sigma_n^{(2)}}{\Sigma_n^{(2)}}\right).$$

Now,

$$\begin{aligned}\Sigma_n^{(1)} - \Sigma_n^{(2)} &= \sum_{k=1}^n (X_k - \bar{X}_n - X_{k-1} + \bar{X}_n')(X_{k-1} - \bar{X}_n') \\ &= \sum_{k=1}^n (X_k - X_{k-1})(X_{k-1} - \bar{X}_n') \\ &= \sum_{k=1}^n X_{k-1}(X_k - X_{k-1}) - (X_n - X_0)\bar{X}_n'.\end{aligned}$$

By the definitions of the Riemann and Itô integrals, this expression tends to

$$\int_0^t X(s)\,dX(s) - (X(t) - X(0))\bar{X}(t)$$

in mean square as $n \to \infty$. Let $J_1(t) = \int_0^t X(s)\,dX(s)$. Moreover, $n^{-1}\Sigma_n^{(2)}$ converges a.s. to the variance of $\{X(t)\}$ as $n \to \infty$. Therefore,

$$\begin{aligned}\hat{b}_n &= \frac{1}{\Delta t}\left\{\frac{\Sigma_n^{(1)} - \Sigma_n^{(2)}}{\Sigma_n^{(2)}} + o_P(n^{-1})\right\} \\ &= \frac{\Sigma_n^{(1)} - \Sigma_n^{(2)}}{\Sigma_n^{(2)}\Delta t} + o_P(1) \to \frac{J_1(t) - (X(t) - X(0))\bar{X}(t)}{J_2(t)},\end{aligned}$$

where $J_2(t) = \int_0^t (X(s) - \bar{X}(t))^2\,ds$ and with convergence in probability. This is the expression given in (22).

We now turn to the estimate of a. It may be written as

$$\begin{aligned}\hat{a}_n &= \frac{\bar{X}_n - (\Sigma_n^{(1)}/\Sigma_n^{(2)})\bar{X}_n'}{\Sigma_n^{(1)}/\Sigma_n^{(2)} - 1}\,\hat{b}_n = \left\{-\bar{X}_n + \frac{(\bar{X}_n - \bar{X}_n')\Sigma_n^{(1)}}{\Sigma_n^{(1)} - \Sigma_n^{(2)}}\right\}\hat{b}_n \\ &= \left\{-\bar{X}_n + \frac{t^{-1}(X_n - X_0)\Sigma_n^{(1)}\Delta t}{\Sigma_n^{(1)} - \Sigma_n^{(2)}}\right\}\hat{b}_n.\end{aligned}$$

One can show that $\Sigma_n^{(1)}\Delta t \to J_2(t)$ a.s. as $n \to \infty$, whence $\hat{a}_n$ converges in probability to

$$\left\{-\bar{X}(t) + \frac{1}{t}(X(t) - X(0))\frac{1}{\hat{b}^c(t)}\right\}\hat{b}^c(t)$$
$$= -\bar{X}(t)\hat{b}^c(t) + \frac{1}{t}(X(t) - X(0)),$$

that is, the expression given in (23). ■

The asymptotic properties of the continuous-time least-squares estimator are given in the following two theorems.

THEOREM 3.5. *The estimator* $(\hat{a}^c(t), \hat{b}^c(t))$ *is strongly consistent.*

Proof. First, we observe that by ergodicity,

$$\frac{1}{t}\int_0^t (X(s) - \bar{X}(t))^2\, ds \to V_\theta[X(0)] \quad P_\theta\text{-a.s.},$$

where $V_\theta[X(0)]$ is the variance of the invariant distribution. Furthermore, by (1) we have

$$\int_0^t X(s)\, dX(s) = a\int_0^t X(s)\, ds + b\int_0^t X^2(s)\, ds + \sigma\int_0^t X^{3/2}(s)\, dW(s).$$

The quadratic variation of the stochastic integral on the right-hand side is $\langle \int X^{3/2}\, dW\rangle_t = \int_0^t X^3(s)\, ds$, whence we obtain

$$\frac{1}{t}\int_0^t X(s)\, dX(s) = \frac{a}{t}\int_0^t X(s)\, ds + \frac{b}{t}\int_0^t X^2(s)\, ds$$
$$+ \frac{\sigma}{t}\int_0^t X^3(s)\, ds \times \frac{\int_0^t X^{3/2}(s)\, dW(s)}{\left\langle \int X^{3/2}\, dW\right\rangle_t}.$$

Because $E_\theta[X^3(0)] < \infty$, and because for any continuous martingale $\{M(t)\}$ with $\lim_{t\to\infty}\langle M\rangle_t = \infty$ a.s. and $M(0) = 0$ we have the law of large numbers $M(t)/\langle M\rangle_t \to 0$ a.s. (see Revuz and Yor, 1991, p. 175), the last equality implies

$$\frac{1}{t}\int_0^t X(s)\, dX(s) \to aE_\theta[X(0)] + bE_\theta[X^2(0)] = bV_\theta[X(0)] \quad P_\theta\text{-a.s.},$$

where we have used $E_\theta[X(0)] = -a/b$. In a similar manner, one may show that

$$\frac{1}{t}X(t) = a + \frac{b}{t}\int_0^t X(s)\,ds + \frac{\sigma}{t}\int_0^t X^{1/2}(s)\,dW(s)$$

$$\to a + bE_\theta[X(0)] = 0 \quad P_\theta\text{-a.s.},$$

so that $\hat{b}^c(t) \to b$ P_θ-a.s. Obviously, $\hat{a}^c(t)$ converges to $-E[X(0)]\,b = a$ P_θ-a.s. ■

Remark. As in discrete time, $(\hat{a}^c(t), \hat{b}^c(t))$ is robust against misspecification of the diffusion term. More precisely, this estimator is strongly consistent for any diffusion term $\sigma(X(t))\,dW(t)$ such that $X^2(0)\sigma^2(X(0))$ is integrable under the corresponding stationary distribution. This result follows along the lines of the preceding proof.

THEOREM 3.6.

$$\sqrt{t}\begin{pmatrix} \hat{a}^c(t) - a \\ \hat{b}^c(t) - b \end{pmatrix} \to \mathcal{N}(0, V_c W_c V_c^T) \quad P_\theta\textit{-weakly},$$

where

$$W_c = \begin{pmatrix} -\dfrac{a\sigma^2}{b^3} & \dfrac{a\sigma^2}{b^4}(2a + \sigma^2) \\ \dfrac{a\sigma^2}{b^4}(2a + \sigma^2) & -\dfrac{a\sigma^2}{4b^5}(2a + \sigma^2)(8a + 5\sigma^2) \end{pmatrix}, \tag{24}$$

$$V_c = \frac{2b^4}{a^2\sigma^2}\begin{pmatrix} 2\nu_1\nu_2 & -\nu_1^2 \\ -(\nu_2 + \nu_1^2) & \nu_1 \end{pmatrix}, \tag{25}$$

and $\nu_j = E_\theta[X^j(0)]$. ■

The proof is given in Appendix B. An informal approach is to derive $V_c W_c V_c^T$ as the limit of $V^{-1}WV^{-T}\Delta t$ as $\Delta t \to 0$.

The parameter σ^2 may be estimated perfectly in continuous-time; the quadratic variation of $\{X(t)\}$ is $\langle X\rangle_t = \sigma^2 \int_0^t X(s)\,ds$ and $\langle X\rangle_t$ is the limit (in probability) of $\sum(X(\tau_k^n) - X(\tau_{k-1}^n))^2$ as $n \to \infty$, where τ^n is a division of $[0,t]$ such that $\max_k|\tau_k^n - \tau_{k-1}^n| \to 0$ as $n \to \infty$ (see Revuz and Yor, 1991, p. 27).

4. LOCAL ASYMPTOTIC NORMALITY

In this section, we give a result on LAN of our statistical experiment and comment on some implications of this property. Let $p_n(x_0, \ldots, x_n; \theta)$ be the joint density of $(X_0, \ldots, X_n)$ under θ and consider the log likelihood locally around θ, that is,

$$\ell_n(\theta + \delta_n u, \theta) \equiv \log\frac{p_n(X_0, \ldots, X_n; \theta + \delta_n u)}{p_n(X_0, \ldots, X_n; \theta)}, \tag{26}$$

where δ_n is a positive real number and $u \in \mathbb{R}^3$. By the Markov property, we obtain

$$\ell_n(\theta + \delta_n u, \theta) = \sum_{k=1}^{n} \{\log p(X_{k-1}, X_k; \theta + \delta_n u) - \log p(X_{k-1}, X_k; \theta)\} + \log \gamma(X_0; \theta + u\delta_n) - \log \gamma(X_0; \theta),$$

where $\gamma(x;\theta)$ and $p(x,y;\theta)$ were defined in Section 2.

According to the definition of LAN as in Pfanzagl (1994, p. 264) or Ibragimov and Hasminskii (1981, p. 120), we have to show that for some norming sequence $\delta_n \to 0$ and each $u \in \mathbb{R}^3$,

$$\ell_n(\theta + \delta_n u, \theta) = u^T S_n - \tfrac{1}{2} u^T K_n u + o_{P_\theta}(1) \tag{27}$$

for some sequences of random vectors S_n and symmetric random matrices K_n such that K_n converges in probability to a finite, positive semidefinite nonrandom matrix K, and S_n converges P_θ-weakly to $\mathcal{N}(0,K)$.

THEOREM 4.1. *If $\theta = (a,b,\sigma^2)$ satisfies $2a/\sigma^2 > 1$, then LAN holds at θ with norming sequence $\delta_n = n^{-1/2}$ and asymptotic information $K = (K_{ij})$ given by*

$$K_{ij} = E_\theta\left[\frac{\partial}{\partial\theta_i} \log p(X_0, X_1; \theta) \times \frac{\partial}{\partial\theta_j} \log p(X_0, X_1; \theta)\right]. \tag{28}$$

The proof is given in Appendix C.

Remark. Notice that the condition $2a/\sigma^2 > 1$ essentially amounts to saying that $\{X(t)\}$ must not hit zero (see Section 2).

Remark. Asymptotic statistical properties of processes similar to the present one are considered in Dohnal (1987). It is proved that if one observes a diffusion process at discrete time points over a fixed time interval and the division of the interval gets finer and finer, then the resulting statistical experiment is local asymptotic mixed normal. The assumptions in Dohnal (1987) on the drift and diffusion coefficients are not satisfied in our situation; however, in particular, we allow the diffusion coefficient to be zero. Also in the papers by Dacunha-Castelle and Florens-Zmirou (1986) and Florens-Zmirou (1993), the assumptions are too restrictive. Florens-Zmirou (1993) assumed the diffusion coefficient to be bounded from below and above by positive constants, and Dacunha-Castelle and Florens-Zmirou (1986) only considered constant diffusion coefficients.

The most important implication of the LAN property is that, provided K is nonsingular, a large class of "reasonable" estimators, so-called *regular* estimators, cannot converge to the true parameter θ at a rate faster than $n^{-1/2}$, and the asymptotic covariance matrix of any regular $n^{1/2}$-consistent estimator is bounded from below by K (see, e.g., Ibragimov and Hasminskii, 1981,

p. 161). For the definition of the term *regular estimator*, we refer to Ibragimov and Hasminskii (1981, p. 151). Moreover, using essentially the same techniques as in the proof of Theorem 4.1 in this paper and Theorem 6.4.1 in Lehmann (1990), one may show the following result on ML estimation.

THEOREM 4.2. *Assume that $K = K(\theta)$ is nonsingular. Then, with probability tending to one there exist solutions θ_n^* of the likelihood equation $\partial \log p_n(X_0, \ldots, X_n; \theta)/\partial\theta = 0$ such that*

(i) *θ_n^* is consistent and*
(ii) *$\sqrt{n}(\theta_n^* - \theta) \to \mathcal{N}(0, K^{-1})$ P_θ-weakly.*

Thus, the estimator θ_n^* referred to in the theorem is asymptotically efficient.

Remark. We give a heuristic argument to make plausible that K is nonsingular for any θ. Because K is a covariance matrix, it is symmetric and nonnegative definite, and thus there exists an orthogonal matrix H such that $\Lambda = HKH^T$, where Λ is the diagonal matrix of eigenvalues of K. But Λ is also the covariance matrix of $H\nabla_\theta \log p(X_0, X_1; \theta)$, where ∇_θ is the gradient operator. Hence, if any of the eigenvalues of K is zero, there exists a linear combination $\sum_1^3 \alpha_i(\partial/\partial\theta_i)\log p(X_0, X_1)$, with at least one nonzero α_i, which is zero in mean square sense and thus also P_θ-a.s. The P_θ-distribution of (X_0, X_1) has a positive density on $(0,\infty)^2$, so that $\sum_1^3 \alpha_i(\partial/\partial\theta_i)\log p(x, y) = 0$ a.e. with respect to Lebesgue measure and, by continuity, we conclude that $\sum_1^3 \alpha_i(\partial/\partial\theta_i)\log p(x, y) = 0$ for all $x, y > 0$. Now, because the functions $(\partial/\partial\theta_i)\log p(x, y)$ are nonlinear, no such linear combination can exist. We have not carried out a strict proof of this fact, however, but such a proof would include, for example, an exact computation of several coefficients in a series expansion of $(\partial/\partial\theta_i)\log p(x, y)$, $i = 1,2,3$, and then checking linear independence of these.

The LAN property also enables us, again provided K is nonsingular, to construct asymptotically efficient estimators by a one-step improvement of any $\sqrt{n}$-consistent estimator. Hence, we may obtain with less computational effort an estimator that has the same efficiency as the MLE. The improvement is essentially carried out by performing a quadratic fit of the log likelihood surface about the first estimator and then maximizing the resulting quadratic, that is, taking a Newton–Raphson step (see LeCam and Yang, 1990, Sect. 5.3, for details).

If $2a/\sigma^2 < 1$, then some elements of the asymptotic information matrix K are not finite, which implies that some elements of the MLE may converge to the corresponding true elements at a faster rate than $n^{-1/2}$. Further research is needed to clarify this. In Overbeck (1995), maximum likelihood estimation in the CIR model is considered, based on continuous-time data. In the case $2a/\sigma^2 < 1$, we then have perfect information about the parameter

a in finite time. More precisely, the estimate of a is a function of a stopped path $\{X_{t\wedge T}\}_{t\geq 0}$, where T is the first time such that the information with respect to a is infinite. Informally, T is such that zero has been hit sufficiently often in $[0,T]$. For $2a/\sigma^2 < 1$, the stopping time T is finite.

5. NUMERICAL EXAMPLES

5.1. Data from the CIR Model

We applied the conditional least-squares estimators (LSE's) $\hat{\theta}_n$ and $\tilde{\theta}_n$, the MLE, a one-step improvement (OSE) of $\hat{\theta}_n$, the BSE, and the MQLE to three different cases of simulated data, henceforth denoted cases A–C. For the two latter estimators, the pseudolikelihood approach (cf. (10)) was used to estimate σ. Despite the complexity of transition density (3), generating simulated samples from this (conditional) density is remarkably simple (see Devroye, 1986, p. 468). The initial value X_0 was simulated from the stationary Gamma distribution of the process (see Section 2).

The true parameters are shown in Tables 1–3. The ratio $2a/\sigma^2$ equals 9.375, 5.0, and 0.8 for cases A, B, and C, respectively, so that zero is inaccessible and LAN holds for cases A and B. The sampling interval Δt was set to unity, for each case 250 replicates of sample size 2,500 were simulated, and for each replicate estimates were computed for $n = 300$, $n = 1{,}000$, and $n = 2{,}500$.

The OSE was computed by evaluating the log likelihood at the points $\hat{\theta}_n + d_i + d_j$, $i,j = 0,1,2,3$, where $d_0 = 0$, $d_i = 0.01\theta_{n,i}e_i$ for $i = 1,2,3$, and e_i is the ith unit vector, then fitting a quadratic through these points, and finding the maximum of the approximation.

For cases A and B, an approximation of the asymptotic information matrix K (see Theorem 4.1) was computed by simulating 500,000 independent replicates of (X_0, X_1), obtained as outlined earlier, and then approximating the expectation in (28) with the corresponding sample mean. The gradient of $\log p(x,y;\cdot)$ was obtained by numerical differentiation. For all estimators, the standard deviations given by asymptotic theory when $n = 2{,}500$ are reported in Table 4. For the MLE, these standard deviations were derived from the simulated information matrices, and for the other estimators they were computed as in Theorem 3.3. As already remarked, only $\hat{\theta}_n$ admits explicit expressions for W and V (because the pseudolikelihood approach is used to estimate σ elsewhere), and these matrices were evaluated by numerical integration for $\tilde{\theta}_n$, the BSE, and the MQLE.

Table 5 shows the percentage of the 95% marginal confidence intervals, constructed assuming normality of the estimators, that covered the true parameter. By "marginal interval," we mean that a separate interval was constructed for each of a, b, and σ; simultaneous coverage was not considered. For the MLE and the OSE, an approximation to the Fisher information was found by a numerical evaluation of the observed information (i.e., the neg-

TABLE 1. Results from the simulation study of case A

Parameter:		a			b			σ		
True values:		0.03			−0.5			0.08		
		Mean	Bias	s	Mean	Bias	s	Mean	Bias	s
					$n = 300$					
$\hat{\theta}_n$	(250)	0.0310	0.0010	0.0049	−0.5286	−0.0286	0.0876	0.0848	0.0048	0.0056
$\tilde{\theta}_n$	(250)	0.0310	0.0010	0.0049	−0.5286	−0.0286	0.0876	0.0835	0.0035	0.0051
MLE	(250)	0.0313	0.0013	0.0044	−0.5339	−0.0339	0.0809	0.0841	0.0041	0.0051
OSE	(250)	0.0311	0.0011	0.0044	−0.5305	−0.0305	0.0807	0.0837	0.0037	0.0051
BSE	(250)	0.0305	0.0005	0.0045	−0.5198	−0.0198	0.0814	0.0831	0.0031	0.0048
MQLE	(250)	0.0307	0.0007	0.0045	−0.5221	−0.0221	0.0813	0.0832	0.0032	0.0048
					$n = 1{,}000$					
$\hat{\theta}_n$	(250)	0.0301	0.0001	0.0027	−0.5124	−0.0124	0.0467	0.0844	0.0044	0.0031
$\tilde{\theta}_n$	(250)	0.0301	0.0001	0.0027	−0.5124	−0.0124	0.0467	0.0829	0.0029	0.0028
MLE	(250)	0.0306	0.0006	0.0025	−0.5211	−0.0211	0.0436	0.0836	0.0036	0.0028
OSE	(250)	0.0305	0.0005	0.0025	−0.5194	−0.0194	0.0435	0.0834	0.0034	0.0028
BSE	(250)	0.0297	−0.0003	0.0025	−0.5067	−0.0067	0.0434	0.0827	0.0027	0.0026
MQLE	(250)	0.0299	−0.0001	0.0025	−0.5087	−0.0087	0.0433	0.0828	0.0028	0.0027
					$n = 2{,}500$					
$\hat{\theta}_n$	(250)	0.0298	−0.0002	0.0018	−0.5076	−0.0076	0.0315	0.0844	0.0044	0.0019
$\tilde{\theta}_n$	(250)	0.0298	−0.0002	0.0018	−0.5076	−0.0076	0.0315	0.0830	0.0030	0.0018
MLE	(250)	0.0304	0.0004	0.0016	−0.5186	−0.0186	0.0293	0.0837	0.0037	0.0018
OSE	(250)	0.0304	0.0004	0.0016	−0.5174	−0.0174	0.0292	0.0836	0.0036	0.0018
BSE	(250)	0.0297	−0.0003	0.0016	−0.5057	−0.0057	0.0285	0.0829	0.0029	0.0017
MQLE	(250)	0.0297	−0.0003	0.0016	−0.5065	−0.0065	0.0288	0.0829	0.0029	0.0017

Note: For each estimator and sample size, the figure in parentheses shows for how many replicates (out of 250) the estimator produced a valid estimate, and to the right the sample means, bias, and standard deviations over these replicates are shown.

ative of the Hessian of the log likelihood) at the estimate in question. For the row "$\hat{\theta}_n(e)$", the matrices W and V were evaluated as in Appendix A at $\hat{\theta}_n$. For the remaining rows in the table, W was estimated as in (20) and V was estimated by the sample mean corresponding to (17), at the estimate in question. Confidence intervals were not computed for invalid estimates (see later).

Case A is approximately the same as the one-factor model in Chen and Scott (1993, p. 20), who obtained these parameters by ML estimation in a time series of real data. All estimators perform about equally well, although the BSE and MQLE have somewhat smaller bias and, quite surprisingly, the MLE has the largest bias for a and b. For $n = 2{,}500$, all standard deviations are in reasonable agreement with the figures in Table 4, and this table also shows that the asymptotic performance of all estimators is very similar. The asymptotic standard deviations of the continuous-time LSE are about 20% smaller than those of the discrete-time LSE's. The coverage of the confidence

TABLE 2. Results from the simulation study of case B

Parameter:		a			b			σ		
True values:		0.1			−2.5			0.2		
		Mean	Bias	s	Mean	Bias	s	Mean	Bias	s
					$n = 300$					
$\hat{\theta}_n$	(225)	0.1077	0.0077	0.0336	−2.7125	−0.2125	0.8493	0.2059	0.0059	0.0309
$\tilde{\theta}_n$	(225)	0.1077	0.0077	0.0336	−2.7125	−0.2125	0.8493	0.2058	0.0058	0.0310
MLE	(250)	0.1645	0.0645	0.1747	−4.1456	−1.6456	4.4088	0.2365	0.0365	0.0975
OSE	(219)	0.1072	0.0072	0.0436	−2.7018	−0.2018	1.1062	0.2041	0.0041	0.0397
BSE	(218)	0.1062	0.0062	0.0392	−2.6677	−0.1677	0.9654	0.2032	0.0032	0.0341
MQLE	(250)	0.1800	0.0800	0.2236	−4.5362	−2.0362	5.6397	0.2419	0.0419	0.1118
					$n = 1{,}000$					
$\hat{\theta}_n$	(248)	0.1084	0.0084	0.0213	−2.7275	−0.2275	0.5309	0.2079	0.0079	0.0195
$\tilde{\theta}_n$	(248)	0.1084	0.0084	0.0213	−2.7275	−0.2275	0.5309	0.2078	0.0078	0.0196
MLE	(250)	0.1127	0.0127	0.0497	−2.8335	−0.3335	1.2474	0.2102	0.0102	0.0328
OSE	(243)	0.1116	0.0116	0.0764	−2.8060	−0.3060	1.9013	0.2103	0.0103	0.0729
BSE	(239)	0.1087	0.0087	0.0279	−2.7337	−0.2337	0.6993	0.2077	0.0077	0.0249
MQLE	(250)	0.1133	0.0133	0.0631	−2.8489	−0.3489	1.5892	0.2103	0.0103	0.0368
					$n = 2{,}500$					
$\hat{\theta}_n$	(250)	0.1050	0.0050	0.0114	−2.6382	−0.1382	0.2857	0.2055	0.0055	0.0111
$\tilde{\theta}_n$	(250)	0.1050	0.0050	0.0114	−2.6382	−0.1382	0.2857	0.2054	0.0054	0.0111
MLE	(250)	0.1051	0.0051	0.0114	−2.6411	−0.1411	0.2862	0.2055	0.0055	0.0111
OSE	(246)	0.1016	0.0016	0.0250	−2.5544	−0.0544	0.6228	0.2020	0.0020	0.0238
BSE	(249)	0.1055	0.0055	0.0163	−2.6515	−0.1515	0.4097	0.2057	0.0057	0.0150
MQLE	(250)	0.1047	0.0047	0.0113	−2.6320	−0.1320	0.2829	0.2052	0.0052	0.0110

Note: For each estimator and sample size, the figure in parentheses shows for how many replicates (out of 250) the estimator produced a valid estimate, and to the right the sample means, bias, and standard deviations over these replicates are shown.

intervals is as expected for a and b (except possibly for the MLE and the OSE when $n = 2{,}500$), but for σ the coverage is decreasing with n and very low when $n = 2{,}500$. The reason for this is the bias of the estimates of σ. As an example, when $n = 2{,}500$, if the sample standard deviation of $\hat{\theta}_n$ (which is 0.0019 for σ; see Table 1) is used for constructing confidence intervals for this estimator, still only 36.8% of them cover the true σ. The corresponding figure for the MQLE is 62.0%.

In case B, the correlation coefficient at lag one is $\exp(b\Delta t) \approx 0.08$, so that estimation of b is more difficult than in case A. Here we also encounter a new problem, namely, that the LSE's, the OSE, and the BSE do not always produce valid parameter estimates (parameters in Θ, that is). For the LSE's, this is the case when the bracketed expression in (6) is negative. For the BSE, the reason is similar, and for the OSE the minimum point of the approximating quadratic may fall outside Θ.

TABLE 3. Results from the simulation study of case C

Parameter:		a			b			σ		
True values:		0.025			−0.5			0.25		
		Mean	Bias	s	Mean	Bias	s	Mean	Bias	s
					$n = 300$					
$\hat{\theta}_n$	(250)	0.0283	0.0033	0.0056	−0.5972	−0.0972	0.1446	0.2534	0.0034	0.0242
$\tilde{\theta}_n$	(250)	0.0283	0.0033	0.0056	−0.5972	−0.0972	0.1446	0.2590	0.0090	0.0187
MLE	(250)	0.0267	0.0017	0.0038	−0.5609	−0.0609	0.1073	0.2553	0.0053	0.0152
OSE	(242)	0.0263	0.0013	0.0435	−0.5497	−0.0497	0.8718	0.2498	−0.0002	0.1131
BSE	(240)	0.0268	0.0018	0.0160	−0.5616	−0.0616	0.3436	0.2645	0.0145	0.0335
MQLE	(250)	0.0271	0.0021	0.0043	−0.5707	−0.0707	0.1201	0.2566	0.0066	0.0178
					$n = 1{,}000$					
$\hat{\theta}_n$	(250)	0.0269	0.0019	0.0029	−0.5682	−0.0682	0.0693	0.2542	0.0042	0.0142
$\tilde{\theta}_n$	(250)	0.0269	0.0019	0.0029	−0.5682	−0.0682	0.0693	0.2565	0.0065	0.0105
MLE	(250)	0.0261	0.0011	0.0019	−0.5509	−0.0509	0.0533	0.2543	0.0043	0.0084
OSE	(250)	0.0252	0.0002	0.0022	−0.5318	−0.0318	0.0587	0.2501	0.0001	0.0094
BSE	(239)	0.0282	0.0032	0.0172	−0.5922	−0.0922	0.3535	0.2669	0.0169	0.0325
MQLE	(250)	0.0263	0.0013	0.0021	−0.5542	−0.0542	0.0579	0.2554	0.0054	0.0100
					$n = 2{,}500$					
$\hat{\theta}_n$	(250)	0.0267	0.0017	0.0020	−0.5627	−0.0627	0.0473	0.2546	0.0046	0.0097
$\tilde{\theta}_n$	(250)	0.0267	0.0017	0.0020	−0.5627	−0.0627	0.0473	0.2559	0.0059	0.0072
MLE	(250)	0.0260	0.0010	0.0014	−0.5500	−0.0500	0.0366	0.2540	0.0040	0.0057
OSE	(250)	0.0256	0.0006	0.0014	−0.5410	−0.0410	0.0375	0.2519	0.0019	0.0061
BSE	(240)	0.0275	0.0025	0.0168	−0.5800	−0.0800	0.3549	0.2665	0.0165	0.0327
MQLE	(250)	0.0261	0.0011	0.0016	−0.5522	−0.0522	0.0400	0.2551	0.0051	0.0068

Note: For each estimator and sample size, the figure in parentheses shows for how many replicates (out of 250) the estimator produced a valid estimate, and to the right the sample means, bias, and standard deviations over these replicates are shown.

TABLE 4. Standard deviations of estimators for $n = 2{,}500$ as given by asymptotic theory; for the continuous-time least-squares estimator $\hat{\theta}^c$, $t = n\Delta t = 2{,}500$

	Case A			Case B			Case C		
Parameter	a	b	σ	a	b	σ	a	b	σ
$\hat{\theta}_n$	0.0016	0.0282	0.0016	0.0100	0.2501	0.0101	0.0019	0.0446	0.0091
$\tilde{\theta}_n$	0.0016	0.0282	0.0015	0.0100	0.2501	0.0101	0.0019	0.0446	0.0062
$\hat{\theta}^c(t)$	0.0013	0.0220	–	0.0020	0.0529	–	0.0015	0.0374	–
MLE	0.0015	0.0251	0.0012	0.0099	0.2469	0.0099	–	–	–
BSE	0.0016	0.0271	0.0015	0.0121	0.3038	0.0121	–	–	–
MQLE	0.0016	0.0268	0.0015	0.0099	0.2489	0.0101	0.0014	0.0350	0.0060

TABLE 5. Percentage of marginal confidence intervals, constructed assuming normality of estimators, that covered the true parameter

	Case A			Case B			Case C		
Parameter	a	b	σ	a	b	σ	a	b	σ
				$n = 300$					
$\hat{\theta}_n$ (e)	96.0%	95.6%	85.2%	94.2%	94.2%	93.3%	96.4%	94.0%	98.4%
$\hat{\theta}_n$	95.2%	96.0%	90.0%	94.2%	95.1%	93.3%	93.2%	88.4%	94.8%
$\tilde{\theta}_n$	95.2%	96.0%	90.4%	94.2%	95.1%	93.8%	93.2%	88.4%	94.8%
MLE	96.0%	94.8%	85.6%	80.4%	80.8%	81.6%	94.4%	96.6%	94.8%
OSE	95.6%	94.0%	86.8%	78.5%	79.4%	79.9%	78.5%	84.3%	80.2%
BSE	96.0%	94.8%	90.4%	89.9%	91.3%	90.8%	74.2%	83.3%	92.1%
MQLE	95.2%	96.4%	90.0%	92.4%	92.0%	92.0%	94.0%	92.0%	94.8%
				$n = 1{,}000$					
$\hat{\theta}_n$ (e)	93.6%	94.0%	62.0%	94.4%	95.6%	96.4%	93.6%	89.2%	94.4%
$\hat{\theta}_n$	94.0%	94.8%	70.8%	94.4%	95.6%	95.6%	91.2%	88.4%	93.6%
$\tilde{\theta}_n$	94.0%	94.8%	83.6%	94.4%	95.6%	96.4%	91.2%	88.4%	92.0%
MLE	96.0%	94.4%	70.8%	93.2%	93.6%	95.2%	94.4%	87.2%	92.8%
OSE	95.6%	94.4%	74.0%	84.8%	86.0%	86.8%	91.6%	90.8%	90.4%
BSE	92.8%	95.6%	84.0%	94.6%	95.4%	95.8%	73.2%	76.8%	90.4%
MQLE	93.6%	95.2%	82.0%	95.6%	95.2%	96.4%	96.4%	88.8%	94.0%
				$n = 2{,}500$					
$\hat{\theta}_n$ (e)	92.4%	92.0%	25.6%	97.2%	97.6%	98.4%	88.0%	76.4%	92.4%
$\hat{\theta}_n$	92.4%	92.4%	31.6%	96.8%	97.2%	98.4%	85.2%	74.4%	92.4%
$\tilde{\theta}_n$	92.4%	92.4%	58.8%	96.8%	97.2%	98.4%	85.2%	74.4%	88.4%
MLE	93.2%	87.6%	32.4%	97.2%	97.6%	98.8%	85.6%	70.8%	84.8%
OSE	92.8%	88.4%	35.6%	89.0%	89.8%	91.1%	89.2%	77.6%	86.0%
BSE	94.0%	94.0%	58.4%	97.6%	97.6%	98.0%	74.2%	76.7%	89.2%
MQLE	93.6%	92.4%	58.4%	98.0%	98.0%	98.4%	88.0%	73.2%	86.0%

Note: Intervals were only computed for valid estimates. See text for how the covariance matrices of the estimators were computed.

Because all invalid estimates are ignored, the estimators are not that easy to compare for $n = 300$. In particular, this is so because sample paths for which the LSE's and/or the BSE produce invalid estimates tend to be "bad" for estimation. For example, if the 25 replicates for which the LSE's are invalid are removed, the performance of the MLE is given by $\text{bias}_a = 0.0076$, $s_a = 0.0335$, $\text{bias}_b = -0.2099$, $s_b = 0.8462$, $\text{bias}_\sigma = 0.0057$, and $s_\sigma = 0.0310$, which is as good as for the LSE's. For the MQLE, the corresponding figures are larger ($\text{bias}_a = 0.0111$, $s_a = 0.0587$, $\text{bias}_b = -0.2995$, $s_b = 1.5182$, $\text{bias}_\sigma = 0.0075$, $s_\sigma = 0.0386$). A useful heuristic is thus that an invalid LSE and/or BSE indicates that any estimator may be unreliable.

Going to $n = 1{,}000$, the bias increases for the LSE's, the OSE, and the BSE. This is because many of those replicates that gave invalid estimates for

$n = 300$ now give valid, but poor, estimates. Removing the two replicates with invalid LSE's, the figures for the MLE are $\text{bias}_a = 0.0086$, $s_a = 0.0211$, $\text{bias}_b = -0.2322$, $s_b = 0.5255$, $\text{bias}_\sigma = 0.0079$, and $s_\sigma = 0.0192$ and for the MQLE they are $\text{bias}_a = 0.0082$, $s_a = 0.0212$, $\text{bias}_b = -0.2226$, $s_b = 0.5289$, $\text{bias}_\sigma = 0.0076$, and $s_\sigma = 0.0196$. Taking this into account, the LSE's, the MLE, and the MQLE are the best estimators, followed by the BSE, which has similar bias but slightly larger standard deviations. The BSE is also more likely to produce an invalid estimate than are the LSE's. Also for $n = 2{,}500$, the LSE's, the MLE, and the MQLE are the best estimators, again followed by the BSE, which has larger standard deviations. The OSE has even larger standard deviations than has the BSE. For $n = 2{,}500$, its standard deviations are about twice as large as those of the best estimators, probably because the asymptotic information matrix K is ill conditioned; the condition number of the simulated approximation was about 75,000.

When $n = 2{,}500$, the standard deviations agree reasonably with what is suggested by asymptotic theory, except for the OSE and the BSE. Comparing the figures in Table 4, we see that all estimators except the BSE have asymptotic variances that are very close to optimal. The continuous-time LSE has asymptotic standard deviations that are about 80% smaller than those of the discrete-time LSE's, showing that the sampling causes a substantial information loss in this case, which is not surprising, as $b\Delta t$ is large.

The confidence interval coverage is as expected, except for the MLE and the BSE when $n = 300$ and for the OSE. For $n = 300$, removing the same replicates as above, the figures for the MLE become 89.3%, 89.8%, and 90.1%, respectively, that is, a significant improvement. That the OSE confidence intervals have somewhat low coverage is not surprising, as the overall performance of the OSE is poor.

In case C, LAN does not hold, so that we do not know what to expect from the OSE and, in view of the discussion in Section 3.1, from the BSE. The LSE's have somewhat larger bias and standard deviations than have the MLE and the MQLE, at least for a and b. The MLE and the MQLE have the smallest standard deviations. The OSE has small bias and large standard deviations for $n = 300$ but performs very well for $n = 1{,}000$ and $n = 2{,}500$, which is a little surprising. The BSE, finally, has reasonable bias for a and b but not for σ, and its standard deviations do not decrease with increasing sample size.

For $n = 2{,}500$, the standard deviations of $\hat{\theta}_n$, $\tilde{\theta}_n$, and the MQLE agree quite well with the figures in Table 4, and we see that the MQLE has about 25% smaller standard deviations for a and b than have the LSE's. In this case, there is a significant performance difference between the two methods to estimate σ, the pseudolikelihood approach being the better one. The asymptotic standard deviations of the continuous-time LSE are about 20% smaller than those of the discrete-time LSE's. As LAN does not hold, we have no firm theoretical basis for forming confidence intervals for the MLE and the

OSE. The coverage of these is not much worse than for the LSE's and the MQLE, however, except for the OSE when $n = 300$. The BSE is not asymptotically normal, and its confidence interval coverage is a little lower. For $n = 300$ and $n = 1{,}000$, the LSE's and the MQLE have coverages for b that are lower than expected, and for $n = 2{,}500$ the coverage of both a and b, and sometimes σ, for these estimators is low. The reason is, again, bias; using sample standard deviations (from Table 3) for computing the intervals does not improve coverage significantly; it actually decreases in some cases.

5.2. Data from Other Models

As already remarked, all estimators of a and b discussed in this paper, except for the MLE and the OSE, are robust against misspecification of the diffusion term. Chan et al. (1992) considered the model

$$dX(t) = (a + bX(t))\,dt + \sigma\{X^+(t)\}^d\,dW(t), \tag{29}$$

which obviously comprises the CIR model as a special case ($d = \frac{1}{2}$). Brennan and Schwartz (1980) studied the model obtained by setting $d = 1$ in (29), and in this section we use this model to examine robustness properties of the estimators.

Three different cases, denoted cases D–F, were studied. Case D corresponds to case A in the way that a and b were the same for both cases, and σ_D was chosen so that $\sigma_A\sqrt{m} = \sigma_D m$, where $m = -a/b$ is the mean of the processes. Thus, the processes are in some sense "equally noisy." Similarly, cases E and F correspond to cases B and C, respectively. Data were simulated from (29) with $d = 1$ using an order 1.5 strong Taylor scheme (see Kloeden and Platen, 1992, p. 351) with step length 0.001. The approximation of the continuous-time stochastic differential equation by a discrete-time simulation scheme of course affects the estimators, but we did not attempt to analyze this influence. Estimates, except the OSE, were computed for the sample sizes $n = 300$ and $n = 1{,}000$, and the results are found in Tables 6–8.

In case D, the BSE and the MQLE are the best estimators, with respect to both bias and standard deviation. The LSE and the MLE have similar standard deviations but larger bias, in particular, the MLE. Also note that the bias of the MLE of b decreases much less than the corresponding bias of the other estimators when going from $n = 300$ to $n = 1{,}000$.

In case E for $n = 300$, the LSE and the BSE again produce invalid estimates for many replicates. Removing all replicates for which the LSE and/or BSE are invalid, the figures for the MQLE become $\text{bias}_a = 0.0305$, $s_a = 0.1458$, $\text{bias}_b = -0.7576$, and $s_b = 3.5744$. As in case B, the new standard deviations are larger than for the LSE and BSE, this time much larger. For $n = 1{,}000$, it turns out that the invalid estimates are caused by the same three replicates. If these are removed, the figures for the MQLE are $\text{bias}_a = 0.0054$, $s_a = 0.0232$, $\text{bias}_b = -0.1376$, and $s_b = 0.5918$, which together with

Table 6. Results from the simulation study of case D

Parameter:		a			b		
True values:		0.03			−0.5		
		Mean	Bias	s	Mean	Bias	s
				$n = 300$			
$\hat{\theta}_n$	(250)	0.0320	0.0020	0.0060	−0.5359	−0.0359	0.1067
MLE	(250)	0.0343	0.0043	0.0051	−0.5742	−0.0742	0.0924
BSE	(250)	0.0311	0.0011	0.0053	−0.5215	−0.0215	0.0957
MQLE	(250)	0.0313	0.0013	0.0053	−0.5240	−0.0240	0.0957
				$n = 1{,}000$			
$\hat{\theta}_n$	(250)	0.0304	0.0004	0.0032	−0.5089	−0.0089	0.0570
MLE	(250)	0.0334	0.0034	0.0027	−0.5576	−0.0576	0.0492
BSE	(250)	0.0302	0.0002	0.0028	−0.5047	−0.0047	0.0507
MQLE	(250)	0.0302	0.0002	0.0028	−0.5054	−0.0054	0.0507

Note: For each estimator and sample size, the figure in parentheses shows for how many replicates (out of 250) the estimator produced a valid estimate, and to the right the sample means, bias, and standard deviations over these replicates are shown.

the LSE is the best performance. The BSE has somewhat larger bias and standard deviations, while the MLE has about the same standard deviations but larger bias ($\text{bias}_a = 0.0147$, $s_a = 0.0225$, $\text{bias}_b = -0.3688$, and $s_b = 0.5719$ without the three "bad" replicates).

Table 7. Results from the simulation study of case E

Parameter:		a			b		
True values:		0.1			−2.5		
		Mean	Bias	s	Mean	Bias	s
				$n = 300$			
$\hat{\theta}_n$	(217)	0.1101	0.0101	0.0364	−2.7593	−0.2593	0.9336
MLE	(250)	0.1939	0.0939	0.1984	−4.8424	−2.3424	4.9316
BSE	(219)	0.1062	0.0062	0.0346	−2.6575	−0.1575	0.8689
MQLE	(250)	0.2183	0.1183	0.2731	−5.4435	−2.9435	6.7503
				$n = 1{,}000$			
$\hat{\theta}_n$	(247)	0.1056	0.0056	0.0239	−2.6438	−0.1438	0.6086
MLE	(250)	0.1212	0.0212	0.0638	−3.0346	−0.5346	1.6158
BSE	(247)	0.1069	0.0069	0.0278	−2.6752	−0.1752	0.7048
MQLE	(250)	0.1124	0.0124	0.0681	−2.8142	−0.3142	1.7235

Note: For each estimator and sample size, the figure in parentheses shows for how many replicates (out of 250) the estimator produced a valid estimate, and to the right the sample means, bias, and standard deviations over these replicates are shown.

TABLE 8. Results from the simulation study of case F

Parameter:		a			b		
True values:		0.025			−0.5		
		Mean	Bias	s	Mean	Bias	s
				$n = 300$			
$\hat{\theta}_n$	(250)	0.0437	0.0187	0.0175	−0.9062	−0.4062	0.3145
MLE	(250)	0.0532	0.0282	0.0126	−1.0989	−0.5989	0.2316
BSE	(250)	0.0283	0.0033	0.0069	−0.6134	−0.1134	0.2232
MQLE	(250)	0.0299	0.0049	0.0081	−0.6501	−0.1501	0.2459
				$n = 1{,}000$			
$\hat{\theta}_n$	(250)	0.0389	0.0139	0.0117	−0.7882	−0.2882	0.2273
MLE	(250)	0.0519	0.0269	0.0073	−1.0493	−0.5493	0.1284
BSE	(250)	0.0261	0.0011	0.0039	−0.5391	−0.0391	0.1324
MQLE	(250)	0.0267	0.0017	0.0046	−0.5527	−0.0527	0.1457

Note: For each estimator and sample size, the figure in parentheses shows for how many replicates (out of 250) the estimator produced a valid estimate, and to the right the sample means, bias, and standard deviations over these replicates are shown.

In case F, the BSE is the best estimator, followed by the MQLE. The LSE and the MLE have considerably larger bias and also larger standard deviations. The bias of the MLE decreases very little when going from $n = 300$ to $n = 1{,}000$.

5.3. Summarizing the Simulations

When the model is correctly specified, the LSE's seem to be almost as good as the MLE if LAN holds and $\exp(b\Delta t)$ is not too small. The MLE is the best estimator, besides having the largest bias for a and b in case A. The OSE is of limited use in our examples, because either the LSE's are as good as the MLE (case A), it performs poorly (case B), or LAN does not hold (case C). It should be remarked, though, that the OSE can apparently perform well despite lack of the LAN property. The BSE performs well in case A, not equally well in case B, and has unacceptable performance in case C. The MQLE performs well in all cases, possibly except when $n = 300$ in case B. If σ is estimated after estimating a and b, pseudolikelihood approach (10) seems preferable to regression approach (12). The differences are negligible, though, except in case C. There is no estimator that generally has better confidence interval coverage than the other ones. What Table 5 does show, though, is that the estimation of the covariance matrices of the estimators, done as described earlier, works well; problems with low coverage are in most cases caused by bias rather than by poor estimates of the covariance matrix.

In the cases where the model is misspecified, the BSE does very well, as does the MQLE, except in case D when $n = 300$. The MLE often has standard deviations similar to those of the other estimators, but its bias is larger. The latter is not surprising, as we do not expect the MLE to be consistent.

It is not trivial to extract an overall best estimator. If the model is correctly specified, it could more or less be any of the estimators, expect for the OSE and the BSE. If misspecification cannot be ruled out, the MQLE is a good choice; it performs satisfactorily except only in case D for $n = 300$. The asymptotic results in Table 4 are also in favor of the MQLE. The LSE is a quick alternative, and it also provides excellent initial values for MQLE computations.

REFERENCES

Abken, P.A. (1993) Innovations in modeling the term structure of interest rates. In *Financial Derivatives: New Instruments and Their Uses*, pp. 107–128. Atlanta: Federal Reserve Bank of Atlanta.

Bergstrom, A.R. (1984) Continuous time stochastic models and issues of aggregation over time. In Z. Griliches & M.D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, pp. 1145–1212. Amsterdam: North-Holland.

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

Brennan, M.J. & E.S. Schwartz (1980) Analyzing convertible bonds. *Journal of Financial and Quantitative Analysis* 15, 907–929.

Carroll, R.J. & D. Ruppert (1988) *Transformation and Weighting in Regression.* New York: Chapman and Hall.

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

Chen, R.-R. & L. Scott (1992) Pricing interest rate options in a two-factor Cox-Ingersoll-Ross model of the term structure. *The Review of Financial Studies* 5, 613–636.

Chen, R.-R. & L. Scott (1993) Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates. *The Journal of Fixed Income* 3 (3), 14–31.

Cox, J.C., J.E. Ingersoll, & S.A. Ross (1985) A theory of the term structure of interest rates. *Econometrica* 53, 385–407.

Crowder, M. (1987) On linear and quadratic estimating functions. *Biometrika* 74, 591–597.

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

Davidian, M. & R.J. Carroll (1987) Variance function estimation. *Journal of the American Statistical Association* 82, 1079–1091.

Devroye, L. (1986) *Non-Uniform Random Variate Generation.* New York: Springer-Verlag.

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

Durrett, R. (1991) *Probability: Theory and Examples.* Pacific Grove, California: Wadsworths & Brooks/Cole.

Feller, W. (1951) Two singular diffusion problems. *Annals of Mathematics* 54, 173–182.

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

Godambe, V.P. & C.C. Heyde (1987) Quasi-likelihood and optimal estimation. *International Statistical Review* 55, 231–244.

Gouriéroux, C. & A. Monfort (1995) Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory* 11, 195–228.
Gouriéroux, C., A. Monfort, & E. Renault (1993) Indirect inference. *Journal of Applied Econometrics* 8, S85–S118.
Gradshteyn, I.S. & I.M. Ryzhik (1980) *Table of Integrals, Series, and Products.* New York: Academic Press.
Ibragimov, I.A. & R.Z. Hasminskii (1981) *Statistical Estimation.* New York: Springer-Verlag.
Jacod, J. & A.N. Shiryaev (1987) *Limit Theorems for Stochastic Processes.* Berlin: Springer-Verlag.
Kawazu, K. & S. Watanabe (1971) Branching processes with immigration and related limit theorems. *Theory of Probability and Its Applications* 16, 36–54.
Klimko, L.A. & P.I. Nelson (1978) On conditional least squares estimation for stochastic processes. *Annals of Statistics* 6, 629–642.
Kloeden, P.E. & E. Platen (1992) *Numerical Solution of Stochastic Differential Equations.* Berlin: Springer-Verlag.
LeCam, L. & G. Yang (1990) *Asymptotics in Statistics.* New York: Springer-Verlag.
Lehmann, E.L. (1990) *Theory of Point Estimation.* Pacific Grove, California: Wadsworths & Brooks/Cole.
Longstaff, F.A. & E.S. Schwartz (1992) Interest rate volatility and the term structure: A two-factor general equilibrium model. *The Journal of Finance* 47, 1259–1282.
Overbeck, L. (1995) Estimation for Continuous-State Branching Processes. Preprint.
Pfanzagl, J. (1994) *Parametric Statistical Theory.* Berlin: Walter de Gruyter.
Pitman, J. & M. Yor (1982) A decomposition of Bessel bridges. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 59, 425–457.
Revuz, D. & M. Yor (1991) *Continuous Martingales and Brownian Motion.* Berlin: Springer-Verlag.
Wefelmeyer, W. (1996a) Adaptive estimators for the parameters of the autoregression function of a Markov chain. *Journal of Statistical Planning and Inference*, forthcoming.
Wefelmeyer, W. (1996b) Quasi-likelihood models and optimal inference. *Annals of Statistics* 24, 405–422.
Wei, C.Z. & J. Winnicki (1990) Estimation of the means in the branching process with immigration. *Annals of Statistics* 18, 1757–1773.
Winnicki, J. (1990) Estimation of the variances in the branching process with immigration. *Probability Theory and Related Fields* 88, 77–106.

APPENDIX A: EXPRESSIONS FOR V AND W

In this appendix, we sketch the derivation of the expressions for the matrices V and W for the estimator $\hat{\theta}_n$.

One may show that

$$\mu_3(x;\theta) = \mu_{30} + \mu_{31}x, \qquad \mu_4(x;\theta) = \mu_{40} + \mu_{41}x + \mu_{42}x^2,$$

with

$$\mu_{30} = \frac{1}{2b^3}\, a\sigma^4(e^{b\Delta t} - 1)^3, \qquad \mu_{40} = \frac{3a\sigma^4(a + \sigma^2)}{4b^4}\,(e^{b\Delta t} - 1)^4,$$

$$\mu_{31} = \frac{3\sigma^4}{2b^2}\, e^{b\Delta t}(e^{b\Delta t} - 1)^2, \qquad \mu_{41} = \frac{3\sigma^4(a + \sigma^2)}{b^3}\, e^{b\Delta t}(e^{b\Delta t} - 1)^3,$$

$$\mu_{42} = \frac{3\sigma^4}{b^2}\, e^{2b\Delta}(e^{b\Delta t} - 1)^2.$$

Let also $\nu_j = E_\theta[X_0^j]$. Straightforward but rather tedious calculations show that

$$V_{11} = \left(\frac{\partial \gamma_0}{\partial a}\right)^2,$$

$$V_{12} = \frac{\partial \gamma_0}{\partial a}\,\frac{\partial \gamma_0}{\partial b} + \frac{\partial \gamma_0}{\partial a}\,\frac{\partial \gamma_1}{\partial b}\,\nu_1,$$

$$V_{13} = 0,$$

$$V_{21} = V_{12},$$

$$V_{22} = \left(\frac{\partial \gamma_0}{\partial b}\right)^2 + 2\,\frac{\partial \gamma_0}{\partial b}\,\frac{\partial \gamma_1}{\partial b}\,\nu_1 + \left(\frac{\partial \gamma_1}{\partial b}\right)^2 \nu_2,$$

$$V_{23} = 0,$$

$$V_{31} = \sigma^2\left\{\eta_0\,\frac{\partial \eta_0}{\partial a} + \eta_1\,\frac{\partial \eta_0}{\partial a}\,\nu_1\right\},$$

$$V_{32} = \sigma^2\left\{\eta_0\,\frac{\partial \eta_0}{\partial b} + \eta_0\,\frac{\partial \eta_1}{\partial b}\,\nu_1 + \eta_1\,\frac{\partial \eta_0}{\partial b}\,\nu_1 + \eta_1\,\frac{\partial \eta_1}{\partial b}\,\nu_2\right\},$$

$$V_{33} = \eta_0^2 + 2\eta_0\eta_1\nu_1 + \eta_1^2\nu_2,$$

and

$$W_{11} = \sigma^2\left\{\eta_0\left(\frac{\partial \gamma_0}{\partial a}\right)^2 + \eta_1\left(\frac{\partial \gamma_0}{\partial a}\right)^2 \nu_1\right\},$$

$$W_{21} = \sigma^2\left\{\eta_0\,\frac{\partial \gamma_0}{\partial a}\,\frac{\partial \gamma_0}{\partial b} + \eta_0\,\frac{\partial \gamma_0}{\partial a}\,\frac{\partial \gamma_1}{\partial b}\,\nu_1 + \eta_1\,\frac{\partial \gamma_0}{\partial a}\,\frac{\partial \gamma_0}{\partial b}\,\nu_1 + \eta_1\,\frac{\partial \gamma_0}{\partial a}\,\frac{\partial \gamma_1}{\partial b}\,\nu_2\right\},$$

$$W_{22} = \sigma^2\left\{\eta_0\left(\frac{\partial \gamma_0}{\partial b}\right)^2 + 2\eta_0\,\frac{\partial \gamma_0}{\partial b}\,\frac{\partial \gamma_1}{\partial b}\,\nu_1 + \eta_0\left(\frac{\partial \gamma_1}{\partial b}\right)^2 \nu_2 + \eta_1\left(\frac{\partial \gamma_0}{\partial b}\right)^2 \nu_1 + 2\eta_1\,\frac{\partial \gamma_0}{\partial b}\,\frac{\partial \gamma_1}{\partial b}\,\nu_2 + \eta_1\left(\frac{\partial \gamma_1}{\partial b}\right)^2 \nu_3\right\},$$

$$W_{31} = \eta_0\mu_{30}\,\frac{\partial \gamma_0}{\partial a} + \eta_1\mu_{30}\,\frac{\partial \gamma_0}{\partial a}\,\nu_1 + \eta_0\mu_{31}\,\frac{\partial \gamma_0}{\partial a}\,\nu_1 + \eta_1\mu_{31}\,\frac{\partial \gamma_0}{\partial a}\,\nu_2,$$

$$W_{32} = \eta_0\mu_{30}\frac{\partial\gamma_0}{\partial b} + \eta_0\mu_{30}\frac{\partial\gamma_1}{\partial b}\nu_1 + \eta_1\mu_{30}\frac{\partial\gamma_0}{\partial b}\nu_1 + \eta_1\mu_{30}\frac{\partial\gamma_1}{\partial b}\nu_2$$
$$+ \eta_0\mu_{31}\frac{\partial\gamma_0}{\partial b}\nu_1 + \eta_0\mu_{31}\frac{\partial\gamma_1}{\partial b}\nu_2 + \eta_1\mu_{31}\frac{\partial\gamma_0}{\partial b}\nu_2 + \eta_1\mu_{31}\frac{\partial\gamma_1}{\partial b}\nu_3,$$
$$W_{33} = (\mu_{40} - \sigma^4\eta_0^2)(\eta_0^2 + 2\eta_0\eta_1\nu_1 + \eta_1^2\nu_2)$$
$$+ (\mu_{41} - 2\sigma^4\eta_0\eta_1)(\eta_0^2\nu_1 + 2\eta_0\eta_1\nu_2 + \eta_1^2\nu_3)$$
$$+ (\mu_{42} - \sigma^4\eta_1^2)(\eta_0^2\nu_2 + 2\eta_0\eta_1\nu_3 + \eta_1^2\nu_4),$$

W being symmetric.

APPENDIX B: PROOF OF THEOREM 3.6

By Itô's formula,

$$X^2(t) - X^2(0) = 2\int_0^t X(s)\,dX(s) + \sigma^2\int_0^t X(s)\,ds,$$

so that

$$\hat{b}^c(t) = -\frac{\sigma^2}{2}\frac{\bar{X}(t)}{\frac{1}{t}\int_0^t (X(s) - \bar{X}(t))^2\,ds} + o_{P_\theta}(t^{-1/2})$$
$$= -\frac{\sigma^2}{2}\frac{\bar{X}(t)}{\overline{X^2}(t) - \bar{X}^2(t)} + o_{P_\theta}(t^{-1/2}) \qquad \textbf{(B.1)}$$

and

$$\hat{a}^c(t) = \frac{\sigma^2}{2}\frac{\bar{X}^2(t)}{\overline{X^2}(t) - \bar{X}^2(t)} + o_{P_\theta}(t^{-1/2}), \qquad \textbf{(B.2)}$$

where $\overline{X^2}(t) = (1/t)\int_0^t X^2(s)\,ds$.

One may show that

$$\mathrm{Cov}_\theta(X(s), X(t)) = \frac{a\sigma^2}{2b^2}e^{b|t-s|}, \qquad \textbf{(B.3)}$$

$$\mathrm{Cov}_\theta(X(s), X^2(t)) = -\frac{a\sigma^2}{2b^3}(2a + \sigma^2)e^{b|t-s|}, \qquad \textbf{(B.4)}$$

and

$$\mathrm{Cov}_\theta(X^2(s), X^2(t)) = \frac{a\sigma^2}{4b^4}(2a + \sigma^2)\{\sigma^2 e^{2b|t-s|} + (4a + 2\sigma^2)e^{b|t-s|}\}. \qquad \textbf{(B.5)}$$

By Theorem VIII.3.79 in Jacod and Shiryaev (1987) and the Cramér–Wold device, we obtain

$$\sqrt{t}\begin{pmatrix} \bar{X}(t) - \nu_1 \\ \overline{X^2}(t) - \nu_2 \end{pmatrix} \to \mathcal{N}(0, W_c) \quad P_\theta\text{-weakly}, \tag{B.6}$$

where

$$W_c = 2\int_0^\infty \begin{pmatrix} \mathrm{Cov}_\theta(X(0), X(t)) & \mathrm{Cov}_\theta(X(0), X^2(t)) \\ \mathrm{Cov}_\theta(X^2(0), X(t)) & \mathrm{Cov}_\theta(X^2(0), X^2(t)) \end{pmatrix} dt;$$

one may use $p = q = 2$ in the theorem and their condition (3.80) is then readily verified. From this and (B.3)–(B.5), (24) follows.

Finally, letting

$$h(u,v) = \frac{\sigma^2}{2}\begin{pmatrix} u^2/(v-u^2) \\ -u/(v-u^2) \end{pmatrix},$$

we may write (B.1) and (B.2) as $(\hat{a}^c(t), \hat{b}^c(t))^T = h(\bar{X}(t), \overline{X^2}(t)) + o_{P_\theta}(t^{-1/2})$. The result now follows by Slutsky's theorem, with V_c given by

$$V_c = \left.\frac{\partial h}{\partial(u,v)}\right|_{u=\nu_1, v=\nu_2}$$

Straightforward algebra shows that this expression equals the right-hand side of (25). ∎

APPENDIX C: PROOF OF THEOREM 4.1

Because $\gamma(x;\vartheta)$ is continuous in ϑ for each x, $\log\gamma(X_0;\theta + u\delta_n) - \log\gamma(X_0;\theta)$ tends to zero surely. We can therefore put it into "$o_{P_\theta}(1)$."

A Taylor expansion of $\log p(x,y;\cdot)$ about θ yields

$$\begin{aligned}\log p(x,y;\theta+\delta_n u) - \log p(x,y;\theta) = {} & n^{-1/2}\sum_{i=1}^{3} u_i \frac{\partial}{\partial\theta_i}\log p(x,y;\theta) \\ & + \frac{1}{2}n^{-1}\sum_{i,j=1}^{3} u_i u_j \frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log p(x,y;\theta) \\ & + \frac{1}{6}n^{-3/2}\sum_{|\alpha|=3} u^\alpha D^\alpha \log p(x,y;\theta+h\delta_n u)\end{aligned} \tag{C.1}$$

for some $h \in [0,1]$. Here, α denotes a three-dimensional multi-index, and D^α and u^α the corresponding derivative and product, respectively.

Hence, it is sufficient to prove that

$$\limsup_{n\to\infty} n^{-1}\left|\sum_{k=1}^{n}\sum_{|\alpha|=3} D^{\alpha}\log p(X_{k-1},X_k;\theta+h_k\delta_n u)\right| < \infty \quad P_\theta\text{-a.s.} \tag{C.2}$$

for each sequence $\{h_k\}$ of random variables in [0,1], that

$$\lim_{n\to\infty} n^{-1}\sum_{k=1}^{n}\sum_{i,j=1}^{3} u_i u_j \frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log p(X_{k-1},X_k;\theta) = u^T K u \quad P_\theta\text{-a.s.}, \tag{C.3}$$

and, by the Cramér-Wold device, that

$$n^{-1/2}\sum_{k=1}^{n}\sum_{i=1}^{3} u_i \frac{\partial}{\partial\theta_i}\log p(X_{k-1},X_k;\theta) \to \mathcal{N}(0,u^T K u) \quad P_\theta\text{-weakly}. \tag{C.4}$$

The assertion (C.2) follows by ergodicity if the random variable

$$\sum_{|\alpha|=3}\sup_{\vartheta\in U(\theta)} |D^{\alpha}\log p(X_0,X_1;\vartheta)|,$$

where $U(\theta)$ is some arbitrarily small neighborhood of θ, is integrable for the stationary process. Hence, we have to examine the third derivatives of $\log p(x,y;\theta)$. Let us rewrite the logarithm of the transition density (3),

$$\begin{aligned}\log p(x,y;\theta) &= \log c(\theta) - c(\theta)y - c(\theta)e^{b\Delta t}x \\ &\quad + \tfrac{1}{2}q(\theta)(\log y - \log x - b\Delta t) + q(\theta)\log\bigl(c(\theta)\sqrt{yxe^{b\Delta t}}\bigr) \\ &\quad + \log B\bigl(c(\theta)\sqrt{yxe^{b\Delta t}}, q(\theta)\bigr),\end{aligned} \tag{C.5}$$

with

$$B(z,q) = \sum_{j=0}^{\infty}\frac{z^{2j}}{j!\Gamma(q+j+1)}.$$

We need to differentiate $\log B(z,q)$ with respect to both arguments. The derivative with respect to z is

$$\frac{\partial}{\partial z}B(z,q) = 2zB(z,q+1),$$

and because $B(z,q+1) \le B(z,q)$ we obtain

$$\frac{\partial}{\partial z}\log B(z,q) \le 2z.$$

Taking into account the derivative of the argument $c(\theta)\sqrt{y\exp(b\Delta t)x}$, we obtain, for example, that

$$\frac{\partial}{\partial b}\log B\bigl(c(\theta)\sqrt{ye^{b\Delta t}x}, q(\theta)\bigr) \le xyc_b(\theta)$$

for some continuous function c_b of the parameter θ.

For the differentiation with respect to a and σ^2, we need the partial derivative of $B(z,q)$ with respect to q. It is given by

$$\frac{\partial}{\partial q} B(z,q) = -\sum_{j=0}^{\infty} \frac{(z^2)^j \Gamma'(q+j+1)}{j!\Gamma^2(q+j+1)} = -\sum_{j=0}^{\infty} \frac{(z^2)^j \psi(q+j+1)}{j!\Gamma(q+j+1)},$$

where ψ is the Riemannian ψ-function. Elementary properties of ψ give that ψ increases sublinearly (cf. Gradshteyn and Ryzhik, 1980, p. xxxii), that is,

$$\psi(x) \le K_0 x,$$

whence by $\Gamma(j+q+1) = (j+q)\Gamma(j+q)$ we have for $j \ge 1$

$$\frac{\psi(q+j+1)}{j!\Gamma(q+j+1)} \le 2K_0 \frac{1}{(j-1)!\Gamma(q+j)}$$

if $q \ge 0$. For $j = 0$, we have $\psi(q+1)/\Gamma(q+1) \le K_1/\Gamma(q)$, where $K_1 \ge K_0$ is a constant, independent of q as long as q is bounded away from zero, say $q \ge \varepsilon$, which in turn is guaranteed by the assumption $2a/\sigma^2 > 1$. Using these inequalities, we obtain after a resummation that

$$\left| \frac{\partial}{\partial q} B(z,q) \right| \le K_2(z^2 B(z,q) + 1),$$

and because $B(z,q) \ge 1$, this yields

$$\left| \frac{\partial}{\partial q} \log B(z,q) \right| \le K_2(z^2 + 1).$$

Higher order derivatives of $\log B(z,q)$ with respect to q are easier to analyze because ψ' has the series expansion $\sum_{i=0}^{\infty} (z+i)^{-2}$ (see, e.g., Gradshteyn and Ryzhik, 1980, p. xxxii), which gives uniform boundedness of ψ', ψ'', and ψ'''.

Using these bounds, we obtain by some tedious but straightforward algebra that for each three-dimensional multi-index α summing up to three

$$|D^\alpha \log p(x,y;\theta)| \le c_x^\alpha(\theta)x + c_y^\alpha(\theta)y + c_{\log}^\alpha(\theta)\log y + \sum_{i=1}^{4} (xy)^i c_i^\alpha(\theta)$$

for continuous functions c_x^α, c_y^α, $c_{\log}^\alpha$, and c_i^α of the parameter θ. From this estimate, it is once more clear why we restrict ourselves to the case of $2a/\sigma^2 > 1$, because under this condition the invariant distribution integrates $\log y$. It is also easy to see that $(X_0 X_1)^4$ is integrable under the stationary distribution.

By similar calculations, we obtain that

$$\sum_{i,j=1}^{3} u_i u_j \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(x,y;\theta)$$

can be bounded by a polynomial in xy of order three plus a term involving $\log y$. Therefore, again by ergodicity,

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \sum_{i,j=1}^{3} u_i u_j \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(X_{k-1}, X_k;\theta)$$
$$= E_\theta\left[\sum_{i,j=1}^{3} u_i u_j \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(X_0, X_1;\theta) \right] \quad P_\theta\text{-a.s.}$$

Rewriting this in matrix notation, we obtain that K_n converges almost surely to the matrix with $E_\theta[\partial^2 \log p(X_0, X_1; \theta)/\partial\theta_i\,\partial\theta_j]$ as its ijth entry.

We now show that

$$E_\theta\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log p(X_0, X_1; \theta)\right] = E_\theta\left[\frac{\partial}{\partial\theta_i}\log p(X_0, X_1; \theta)\,\frac{\partial}{\partial\theta_j}\log p(X_0, X_1; \theta)\right]. \quad \textbf{(C.6)}$$

As for the third moments, one may prove that $\sup_{\vartheta \in U(\theta)} |\partial^2 \log p(X_0, X_1; \vartheta)/\partial\theta_i\,\partial\theta_j|$ is integrable with respect to P_θ. Hence, we can interchange differentiation with respect to θ_i and θ_j with integration with respect to P_θ. This together with the fact that $p(x, y; \theta)$ is a density with respect to Lebesgue measure implies

$$E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\,p(X_0, X_1; \theta)\Big/p(X_0, X_1; \theta)\right] = 0,$$

which eventually implies (C.6). Thus, (C.3) is proved.

Finally, one may show that the second moment of $\partial \log p(X_0, X_1; \theta)/\partial\theta_i$ exists, whence we can apply the martingale central limit theorem (see, e.g., Durrett, 1991, Ch. VII, Theorem 7.5) to obtain (C.4). ■