Bernoulli **19**(2), 2013, 409–425
DOI: 10.3150/11-BEJ411

Statistical inference for discrete-time samples from affine stochastic delay differential equations

UWE KÜCHLER[1] and MICHAEL SØRENSEN[2]

[1]*Institut für Mathematik, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany. E-mail: kuechler@mathematik.hu-berlin.de*
[2]*Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark. E-mail: michael@math.ku.dk*

Statistical inference for discrete time observations of an affine stochastic delay differential equation is considered. The main focus is on maximum pseudo-likelihood estimators, which are easy to calculate in practice. A more general class of prediction-based estimating functions is investigated as well. In particular, the optimal prediction-based estimating function and the asymptotic properties of the estimators are derived. The maximum pseudo-likelihood estimator is a particular case, and an expression is found for the efficiency loss when using the maximum pseudo-likelihood estimator, rather than the computationally more involved optimal prediction-based estimator. The distribution of the pseudo-likelihood estimator is investigated in a simulation study. Two examples of affine stochastic delay equation are considered in detail.

Keywords: asymptotic normality; composite likelihood; consistency; discrete time observation of continuous-time models; prediction-based estimating functions; pseudo-likelihood; stochastic delay differential equation

1. Introduction

In the last decade, statistical inference for stochastic delay differential equations (SDDEs) has been studied from various viewpoints. Early work on maximum likelihood estimation was done by Küchler and Mensch [15]. Gushchin and Küchler [8] and Küchler and Kutoyants [14] determined the non-standard asymptotic properties of the maximum likelihood estimator for SDDEs, and Küchler and Vasil'jev [20] constructed sequential procedures with a given accuracy in the L_2 sense. Nonparametric estimators for affine SDDEs were investigated by Reiß [22] and Reiß [23]. All these studies were concerned with continuous observation of the solution process.

As opposed to the situation for ordinary stochastic differential equations, observations at discrete time points have been little studied for SDDEs. Reiß [21] studied nonparametric estimation. Küchler and Sørensen [19] proposed a simple estimator for the parameters α_k in the particular type of SDDE given by (2) below. This estimator is biased, however, and can only be expected to work well for high-frequency observations. In this paper we report a first attempt at investigating parametric inference for affine stochastic delay equations of the general type (1) observed at discrete time points. We propose a pseudo-likelihood function and study it in the framework of prediction-based estimating functions. Applying the methods proposed here in practice often

1350-7265 © 2013 ISI/BS

requires the ability to simulate solutions of SDDEs. This problem has been studied by, among other, Küchler and Platen [16] and Buckwar [2]. A practical application of one of the simplest SDDEs, discussed in Example 2.1 below, was provided by Küchler and Platen [17].

We consider the model given by the stochastic differential equation

$$\mathrm{d}X(t) = \left(\int_{-r}^{0} X(t+s) a_\alpha(\mathrm{d}s) \right) \mathrm{d}t + \sigma \, \mathrm{d}W(t), \tag{1}$$

where a_α is a measure on $[-r, 0]$ $(0 \le r < \infty)$ such that (1) has a unique stationary solution (for a suitable given initial condition). Conditions under which (1) has a unique stationary solution were given by Gushchin and Küchler [9]. By Theorem 3.1 in this paper, the stationary solution is a Gaussian process. We assume that the measure a_α depends on a parameter α. The parameter about which inference is to be drawn is (α, σ) or (α, σ, r), $(\sigma, r > 0)$. As usual, we denote the parameter space by $\Theta \subseteq \mathbb{R}^p$. The process W is a Wiener process. The initial condition is that the distribution of $\{X(s) \mid s \in [-r, 0]\}$ is the stationary distribution, which always has expectation 0. The data are observations at discrete time points: $X(\Delta), X(2\Delta), \ldots, X(n\Delta)$.

An interesting particular case of (1) is

$$\mathrm{d}X(t) = \sum_{k=1}^{N} \alpha_k X(t - r_k) \, \mathrm{d}t + \sigma \, \mathrm{d}W(t). \tag{2}$$

Here the measure a_α is concentrated in the discrete points $-r_1, \ldots, -r_N$, $(r_i \ge 0)$. The vector $(r_1, \ldots, r_N)$ may be among the parameters to be estimated. The particular case where $N = 2$ and $r_1 = 0$ is considered in detail in Example 2.1.

In Section 2 we discuss how to calculate the likelihood function for discrete time observations, and propose a pseudo-likelihood function that closely approximates the likelihood function and is considerably easier to calculate. We consider two examples in detail. In Section 3 we present prediction-based estimating functions for affine stochastic delay equations, find the optimal estimating function in this class, and show that the pseudo-likelihood estimator is a particular case of a prediction-based estimator. The prediction-based estimating functions provide a good framework for discussing the asymptotics of the pseudo-likelihood estimator and in particular the efficiency loss compared with the optimal prediction-based estimating function. We do this in Section 4, specifying conditions ensuring consistency and asymptotic normality. Finally, in Section 5 we present a simulation study of properties of the pseudo-likelihood estimator.

2. The likelihood and the pseudo-likelihood function

Because the stationary solution to (1) is a zero-mean Gaussian process, Gushchin and Küchler [9], the data are in fact a Gaussian time series with expectation 0. Therefore, in principle the likelihood function can be calculated if we can determine, analytically or numerically, the autocovariances

$$K_\theta(t) = E_\theta(X(0)X(t)), \qquad t \ge 0. \tag{3}$$

The autocovariance function, $K_\theta(t)$, satisfies the differential equation

$$\partial_t K_\theta(t) = \int_{-r}^{0} K_\theta(t+s) a_\theta(\mathrm{d}s), \qquad t \geq 0, \tag{4}$$

with $\partial_t K_\theta(0+) = -\frac{1}{2}\sigma^2$, provided that we define $K_\theta(-t) = K_\theta(t)$ for $t \geq 0$ (see Gushchin and Küchler [10]). The condition $\partial_t K_\theta(0+) = -\frac{1}{2}\sigma^2$ also can be written in the form

$$2\int_{-r}^{0} K_\theta(s) a_\theta(\mathrm{d}s) = -\sigma^2.$$

Equation (4) is a continuous-time analogue of the Yule–Walker equation known from time-series analysis, and hereinafter we refer to (4) as the delay Yule–Walker equation of (3). In general, this equation must be solved numerically, but we consider two particular examples where it can be solved explicitly.

To calculate the *likelihood function*, define for every $\ell = 1, \dots, n$ the ℓ-dimensional vector $\kappa_\ell(\theta) = (K_\theta(\Delta), \dots, K_\theta(\ell\Delta))^T$, and the $\ell \times \ell$-matrix $\mathcal{K}_\ell(\theta) = \{K_\theta((i-j)\Delta)\}_{i,j=1,\dots,\ell}$. Here and later T denotes transposition of vectors and matrices. The matrix $\mathcal{K}_\ell(\theta)$ is the covariance matrix of the vector of the first ℓ observations $X(\Delta), \dots, X(\ell\Delta)$.

The conditional distribution of the observation $X((i+1)\Delta)$ given the previous observations $X(\Delta), \dots, X(i\Delta)$ is the Gaussian distribution with expectation $\phi_i(\theta)^T X_{i:1}$ and variance $v_i(\theta)$, where $\phi_i(\theta)$ is the i-dimensional vector given by $\phi_i(\theta) = \mathcal{K}_i(\theta)^{-1}\kappa_i(\theta)$, $v_i(\theta) = K_\theta(0) - \kappa_i(\theta)^T \mathcal{K}_i(\theta)^{-1}\kappa_i(\theta)$, and $X_{i:j} = (X(i\Delta), \dots, X(j\Delta))^T$, $i > j \geq 1$. The vector $\phi_i(\theta) = (\phi_{i,1}(\theta), \dots, \phi_{i,i}(\theta))^T$ and the conditional variance $v_i(\theta)$ can be found using the Durbin–Levinson algorithm (see, e.g., page 169 in Brockwell and Davis [1]). Specifically, $\phi_{1,1}(\theta) = K_\theta(\Delta)/K_\theta(0)$ and $v_0(\theta) = K_\theta(0)$, whereas

$$\phi_{i,i}(\theta) = \left(K_\theta(i\Delta) - \sum_{j=1}^{i-1} \phi_{(i-1),j}(\theta) K_\theta\big((i-j)\Delta\big)\right) v_{i-1}(\theta)^{-1},$$

$$\begin{pmatrix} \phi_{i,1}(\theta) \\ \vdots \\ \phi_{i,i-1}(\theta) \end{pmatrix} = \begin{pmatrix} \phi_{i-1,1}(\theta) \\ \vdots \\ \phi_{i-1,i-1}(\theta) \end{pmatrix} - \phi_{i,i}(\theta) \begin{pmatrix} \phi_{i-1,i-1}(\theta) \\ \vdots \\ \phi_{i-1,1}(\theta) \end{pmatrix}$$

and

$$v_i(\theta) = v_{i-1}(\theta)\big(1 - \phi_{i,i}(\theta)^2\big).$$

The likelihood function based on the data $X(\Delta), \dots, X(n\Delta)$ is

$$\begin{aligned} L_n(\theta) = {} & \frac{1}{\sqrt{2\pi v_0(\theta)}} \exp\left(-\frac{1}{2v_0(\theta)} X(\Delta)^2\right) \\ & \times \prod_{i=1}^{n-1} \left[\frac{1}{\sqrt{2\pi v_i(\theta)}} \exp\left(-\frac{1}{2v_i(\theta)} \big(X\big((i+1)\Delta\big) - \phi_i(\theta)^T X_{i:1}\big)^2\right)\right]. \end{aligned} \tag{5}$$

Calculation of this function quickly becomes very time-consuming as the sample size n increases. In particular, $\phi_i(\theta)$ and $v_i(\theta)$ must be calculated for every observation time point. However, the autocovariances $K_\theta(i\Delta)$ decrease exponentially with i (see Diekmann et al. [3], page 34). Using the Durbin–Levinson algorithm, it is readily apparent that this implies that the quantities $\phi_{i,j}(\theta)$ decrease exponentially with j. Thus the conditional distribution of $X((i+1)\Delta)$ given $X(\Delta), \dots, X(i\Delta)$ depends only very little on observations in the distant past.

Therefore, we propose using instead a *pseudo-likelihood function* obtained by replacing in the likelihood function the conditional density of $X((i+1)\Delta)$ given $X(\Delta), \dots, X(i\Delta)$ with the conditional density of $X((i+1)\Delta)$ given $X((i+1-k)\Delta), \dots, X(i\Delta)$, where k typically is relatively small. This pseudo-likelihood function was proposed by H. Sørensen [24] in connection with stochastic volatility models, but the idea is widely applicable. The pseudo-likelihood is given by

$$\tilde{L}_n(\theta) = \prod_{i=k}^{n-1} \left[\frac{1}{\sqrt{2\pi v_k(\theta)}} \exp\left(-\frac{1}{2v_k(\theta)} \left(X\left((i+1)\Delta\right) - \phi_k(\theta)^T X_{i:i+1-k} \right)^2 \right) \right]. \tag{6}$$

We have not included the density of $X_{k:1}$. Note that the computational gain is large because we calculate (6) using the same values of $\phi_k(\theta)$ and $v_k(\theta)$ for all observation time points. Thus, these quantities must be calculated only once for every value of $\tilde{L}_n(\theta)$. We call the number k the *depth* of the pseudo-likelihood function. We consider the influence of k on the quality of the estimators in the simulation study reported in Section 5. As would be expected, the quality increases with increasing depth. For the model considered in Section 5, the present study indicates that the bias and the variance of the estimators do not depend much on the depth when k is larger than 3–5 times r.

Example 2.1. Consider the equation

$$\mathrm{d}X(t) = [aX(t) + bX(t-r)]\,\mathrm{d}t + \sigma\,\mathrm{d}W(t), \tag{7}$$

where $r > 0$, $\sigma > 0$. This is a particular case of the model (2). The real parameters a and b are chosen such that a stationary solution of (7) exists. This is the case exactly when $a < r^{-1}$ and $-a/\cos(\xi(ar)) < b < -a$ if $a \neq 0$, and when $-\pi/2 < br < 0$ if $a = 0$. Here the function $\xi(u) \in (0, \pi)$ is the root of $\xi(u) = u\tan(\xi(u))$ if $u \neq 0$, and $\xi(0) = \pi/2$. The stationary solution is unique if it exists. Details of this have been provided by Küchler and Mensch [15], who explicitly found the covariance function of the stationary solution by solving the Yule–Walker delay differential equation (4),

$$\partial_t K_\theta(t) = aK_\theta(t) + bK_\theta(t-r), \qquad t \geq 0.$$

They found that

$$K_\theta(0) = \begin{cases} \dfrac{\sigma^2(b\sinh(\lambda(a,b)r) - \lambda(a,b))}{2\lambda(a,b)[a + b\cosh(\lambda(a,b)r)]} & \text{when } |b| < -a, \\ \sigma^2(br-1)/(4b) & \text{when } b = a, \\ \dfrac{\sigma^2(b\sin(\lambda(a,b)r) - \lambda(a,b))}{2\lambda(a,b)[a + b\cos(\lambda(a,b)r)]} & \text{when } b < -|a|, \end{cases} \tag{8}$$

where $\lambda(a,b)=\sqrt{|a^2-b^2|}$, and that for $t\in[0,r]$ the covariance function is

$$K_\theta(t)=\begin{cases}K_\theta(0)\cosh(\lambda(a,b)t)-\sigma^2(2\lambda(a,b))^{-1}\sinh(\lambda(a,b)t) & \text{when } |b|<-a,\\ K_\theta(0)-\frac{1}{2}t\sigma^2 & \text{when } b=a,\\ K_\theta(0)\cos(\lambda(a,b)t)-\sigma^2(2\lambda(a,b))^{-1}\sin(\lambda(a,b)t) & \text{when } b<-|a|.\end{cases}\tag{9}$$

Because $K_\theta(t)$ is known in $[0,r]$, the Yule–Walker equation becomes an ordinary differential equation for $K_\theta(t)$ in $[r,2r]$, which can be easily solved. Similarly, for $t>r$, the autocovariance function $K_\theta(t)$ is given by

$$K_\theta(t)=b\int_{nr}^{t}\mathrm{e}^{a(t-s)}K_\theta(s-r)\,\mathrm{d}s+\mathrm{e}^{a(t-nr)}K_\theta(nr),\qquad t\in[nr,(n+1)r],n\in\mathbb{N}.\tag{10}$$

Thus $K_\theta(t)$ can be determined iteratively in each of the intervals $t\in[nr,(n+1)r]$, $n\in\mathbb{N}$. Note that the covariance function depends on σ and r in a simple and smooth way, so that these parameters also can be estimated by maximizing the pseudo-likelihood function (6).

For $b=0$, the model (7) is the Ornstein–Uhlenbeck process, for which (8) and (9) simplifies to the well-known result $K_\theta(t)=-(\sigma^2/(2a))\mathrm{e}^{at}$ $(t\geq 0)$ in the stationary case $a<0$. For $a=0$, we obtain the model

$$\mathrm{d}X(t)=bX(t-r)\,\mathrm{d}t+\sigma\,\mathrm{d}W(t).\tag{11}$$

This process is stationary if and only if $br\in(-\pi/2,0)$, and in this case, by (8) and (9), the autocovariance function is given by

$$K_\theta(t)=-\frac{\sigma^2}{2b}\left(\frac{1-\sin(br)}{\cos(br)}\cos(bt)+\sin(bt)\right)\tag{12}$$

when $t\in[0,r]$. By (10), we find that

$$K_\theta(t)=-\frac{\sigma^2}{2b}\big[2+\cos(bt)\big\{\big(\tan(bt)-\tan(br)\big)\big(1-2\sin(br)\big)-1/\cos(br)\big\}\big]\tag{13}$$

for $t\in[r,2r]$.

Example 2.2. Consider the equation

$$\mathrm{d}X(t)=-b\left(\int_{-r}^{0}X(t+s)\mathrm{e}^{as}\,\mathrm{d}s\right)\mathrm{d}t+\sigma\,\mathrm{d}W(t),\tag{14}$$

where $r>0$, $\sigma>0$. The set of values of the parameters a and b for which a unique stationary solution of (14) exists was studied by Reiß [21]. This set is rather complicated and irregular; for instance, it is not convex. However, it contains the region $\{(a,b)\mid a\geq 0, b>0, b(1+\mathrm{e}^{-ar})<\max(\pi^2/r^2,a^2(\mathrm{e}^{ar}-1)^2)\}$. For $a=0$, corresponding to a uniform delay measure, a stationary solution exists exactly when $0<b<\frac{1}{2}\pi^2/r^2$. When $r=\infty$, the situation is much simpler. In that case, a stationary solution exists for all $a>0$ and $b>0$.

When $a = 0$ (and r is finite),

$$K_\theta(t) = \frac{\sigma^2 \sin(r\sqrt{2b}(1/2 - t))}{2r\sqrt{2b}\cos(r\sqrt{b/2})} + \frac{\sigma^2}{2br^2}, \qquad 0 \le t \le r.$$

For $a > 0$, an explicit expression for $K_\theta(t)$ involving trigonometric functions exists as well (see Reiß [21], page 41), but it is somewhat complicated, and thus we omit it here.

3. Prediction-based estimating functions

In this section, we discuss the pseudo-likelihood estimator in the framework of prediction-based estimating functions. This class of estimating functions was introduced by Sørensen [25] as a generalization of the martingale estimating functions that is also applicable to non-Markovian processes such as solutions to stochastic delay differential equations. Applications of the methodology to observations of integrated diffusion processes and sums of diffusions have been described by Ditlevsen and Sørensen [4] and Forman and Sørensen [6]. An up-to-date review of the theory of prediction-based estimating functions has been provided by Sørensen [26].

We show that the pseudo-likelihood estimator is a prediction-based estimator, and find the optimal prediction-based estimating function, which turns out to be different from the pseudo-score function. Optimality is in the sense of Godambe and Heyde [7] (see Heyde [11]). We impose the following condition that is satisfied for the models considered in Examples 2.1 and 2.2.

Condition 3.1. *The function $K_\theta(t)$ is continuously differentiable with respect to θ.*

Under this assumption, we find the following expression for the pseudo-score function:

$$\begin{aligned}\partial_\theta \tilde{\ell}_n(\theta) &:= \partial_\theta \log(\tilde{L}_n(\theta)) \\ &= \sum_{i=k}^{n-1} \frac{\partial_\theta \phi_k(\theta)^T X_{i:i+1-k}}{v_k(\theta)} \bigl(X\bigl((i+1)\Delta\bigr) - \phi_k(\theta)^T X_{i:i+1-k}\bigr) \\ &\quad + \frac{\partial_\theta v_k(\theta)}{2v_k(\theta)^2} \sum_{i=k}^{n-1} \bigl[\bigl(X\bigl((i+1)\Delta\bigr) - \phi_k(\theta)^T X_{i:i+1-k}\bigr)^2 - v_k(\theta)\bigr].\end{aligned} \tag{15}$$

The derivatives $\partial_\theta \phi_k(\theta)$ and $\partial_\theta v_k(\theta)$ exist when $K_\theta(t)$ is differentiable and can be found by the following algorithm, which is obtained by differentiating the Durbin–Levinson algorithm:

$$\begin{aligned}\partial_\theta \phi_{i,i}(\theta) = \Biggl[\Biggl(&\partial_\theta K_\theta(i\Delta) - \sum_{j=1}^{i-1}\bigl(\partial_\theta \phi_{(i-1),j}(\theta) K_\theta\bigl((i-j)\Delta\bigr) \\ &+ \phi_{(i-1),j}(\theta)\partial_\theta K_\theta\bigl((i-j)\Delta\bigr)\bigr)\Biggr) v_{i-1}(\theta)\end{aligned}$$

$$+\left(K_\theta(i\Delta)-\sum_{j=1}^{i-1}\phi_{(i-1),j}(\theta)K_\theta\big((i-j)\Delta\big)\right)\partial_\theta v_{i-1}(\theta)\Bigg]v_{i-1}(\theta)^{-2},$$

$$\begin{pmatrix}\partial_{\theta_j}\phi_{i,1}(\theta)\\ \vdots\\ \partial_{\theta_j}\phi_{i,i-1}(\theta)\end{pmatrix}=\begin{pmatrix}\partial_{\theta_j}\phi_{i-1,1}(\theta)\\ \vdots\\ \partial_{\theta_j}\phi_{i-1,i-1}(\theta)\end{pmatrix}-\partial_{\theta_j}\phi_{i,i}(\theta)\begin{pmatrix}\phi_{i-1,i-1}(\theta)\\ \vdots\\ \phi_{i-1,1}(\theta)\end{pmatrix}$$
$$-\phi_{i,i}(\theta)\begin{pmatrix}\partial_{\theta_j}\phi_{i-1,i-1}(\theta)\\ \vdots\\ \partial_{\theta_j}\phi_{i-1,1}(\theta)\end{pmatrix}$$

for $j=1,\dots,p$, and

$$\partial_\theta v_i(\theta)=\partial_\theta v_{i-1}(\theta)\big(1-\phi_{i,i}(\theta)^2\big)-2v_{i-1}(\theta)\phi_{i,i}(\theta)\,\partial_\theta\phi_{i,i}(\theta).$$

The minimum mean squared error linear predictors of $X((i+1)\Delta)$ and $(X((i+1)\Delta)-\phi_k(\theta)^T X_{i:i+1-k})^2$ given $X_{i:i+1-k}$ are $\phi_k(\theta)^T X_{i:i+1-k}$ and $v_k(\theta)$, respectively. This is because for the Gaussian processes considered in this paper, the two conditional expectations are linear in $X_{i:i+1-k}$. Thus the pseudo-score function is a prediction-based estimating function as defined in Sørensen [26], where estimating functions of a slightly more general type than in the original paper (Sørensen [25]) are treated. The generalization allows the predicted function to depend both on the parameter and on the previous observations. Exploring the relation of the pseudo-score function to the optimal estimating function based on these predictors is of interest.

We start by defining a class of prediction-based estimating functions. Define the $(k+1)\times 2$-matrices

$$Z^{(i)}=\begin{pmatrix}X_{i:i+1-k}^T & 0\\ 0\cdots 0 & 1\end{pmatrix}^T,\qquad i=k,\dots,n-1,$$

and the $(k+1)$-dimensional vectors

$$H_i(\theta)=Z^{(i)}\begin{pmatrix}X\big((i+1)\Delta\big)-\phi_k(\theta)^T X_{i:i+1-k}\\ \big(X\big((i+1)\Delta\big)-\phi_k(\theta)^T X_{i:i+1-k}\big)^2-v_k(\theta)\end{pmatrix},\qquad i=k,\dots,n-1.$$

Then the class of prediction-based estimating functions to which (15) belongs is given by

$$G_n(\theta)=A(\theta)\sum_{i=k}^{n-1}H_i(\theta),\tag{16}$$

where $A(\theta)$ is a $p\times(k+1)$ matrix of weights that can depend on the parameter, but not on the data. The pseudo-score function (15) is obtained if the weight matrix $A(\theta)$ is chosen as

$$\tilde{A}(\theta)=\left(\frac{\partial_\theta\phi_k(\theta)^T}{v_k(\theta)}\quad\frac{\partial_\theta v_k(\theta)}{2v_k(\theta)^2}\right).$$

Within the class of estimators obtained by solving the estimating equation $G_n(\theta)=0$ for some choice of $A(\theta)$, the estimator with the smallest asymptotic variance is obtained by choosing the optimal weight matrix $A^*(\theta)$. The optimal estimating function is the one closest to the true score function in an L^2-sense (for details, see Heyde [11]).

We now can find the optimal weight matrix $A^*(\theta)$. The covariance matrix of the $(k+1)$-dimensional random vector $\sum_{i=k}^{n-1} H^{(i)}(\theta)/\sqrt{n-k}$ is

$$\bar{M}_n(\theta) = M^{(1)}(\theta) + M_n^{(2)}(\theta), \tag{17}$$

where

$$M_n^{(2)}(\theta) = \sum_{j=1}^{n-k-1} \frac{(n-k-j)}{(n-k)} [E_\theta(H_k(\theta)H_{k+j}(\theta)^T) + E_\theta(H_{k+j}(\theta)H_k(\theta)^T]$$

and

$$M^{(1)}(\theta) = E_\theta(H_k(\theta)H_k(\theta)^T) = \begin{pmatrix} v_k(\theta)\mathcal{K}_k(\theta) & O_{k,1} \\ O_{1,k} & 2v_k(\theta)^2 \end{pmatrix},$$

with O_{j_1,j_2} denoting here and later the $j_1 \times j_2$-matrix of 0s, and with $\mathcal{K}_k(\theta)$ denoting the covariance matrix of $(X(k\Delta), \dots, X(\Delta))$ defined in Section 2. We have used $E_\theta(H_i(\theta)) = 0$, which is a general property of prediction-based estimating functions (see Sørensen [26]). In this particular case, this is easily seen by finding the conditional expectation of $H_i(\theta)$ given $X_{i:i+1-k}$, which is 0.

To find the optimal estimating function, we also need the $p \times (k+1)$ *sensitivity-matrix* $S(\theta)$, given by

$$S(\theta)^T = E_\theta(\partial_{\theta^T} H_i(\theta)) = -\begin{pmatrix} \mathcal{K}_k(\theta)\partial_{\theta^T}\phi_k(\theta) \\ \partial_{\theta^T} v_k(\theta) \end{pmatrix}. \tag{18}$$

For the derivation of $M^{(1)}(\theta)$ and $S(\theta)$, we use that the model is Gaussian and, in particular, we use that $\phi_k(\theta)^T X_{i:i+1-k}$ is the conditional expectation of $X((i+1)\Delta)$ and not just the minimum mean squared linear predictor as in the general theory of prediction-based estimating functions. The optimal weight matrix is given by

$$A_n^*(\theta) = -S(\theta)\bar{M}_n(\theta)^{-1}, \tag{19}$$

see Sørensen [26].

The class of estimating functions considered above is not the full class of prediction-based estimating functions to which (15) belongs, as defined by Sørensen [25] and Sørensen [26]. The full class is obtained by replacing in (16) $A(\theta)$ with a $p \times 2(k+1)$ matrix and $H_i(\theta)$ with the $2(k+1)$-dimensional vectors $\breve{H}_i(\theta)$ obtained when $Z^{(i)}$ is replaced by the $2(k+1) \times 2$ matrix,

$$\breve{Z}^{(i)} = \begin{pmatrix} X_{i:i+1-k}^T & 0 & 1 & O_{1,k} \\ O_{1,k} & 1 & 0 & X_{i:i+1-k}^T \end{pmatrix}^T$$

in the definition of $H_i(\theta)$. In this way, $H_i(\theta)$ is extended by $k+1$ extra coordinates. Because the moments of an odd order of a centered multivariate Gaussian distribution equal 0, we see that the extra $k+1$ coordinates of $\breve{H}_i(\theta)$ have expectation 0 under the true probability measure irrespective of the value of the parameter θ; therefore, they cannot be expected to be a useful addition to $H_i(\theta)$. However, the extra coordinates might be correlated with the coordinates of $H_i(\theta)$, and thus might be used to reduce the variance of the estimating function. To see that this is not the case, the optimal estimating function based on $\breve{H}_i(\theta)$ can be calculated. The covariance matrix of the random vector $\sum_{i=k}^{n-1} \breve{H}^{(i)}(\theta)/\sqrt{n-k}$ can be shown to be a block-diagonal matrix with two $(k+1) \times (k+1)$-blocks, the first of which equals $\bar{M}_n(\theta)$. This follows from the fact that moments of an odd order of a centered multivariate Gaussian distribution equal 0. Moreover, the sensitivity matrix corresponding to $\breve{H}_i(\theta)$ is

$$\breve{S}(\theta)^T = E_\theta(\partial_{\theta^T} \breve{H}_i(\theta)) = -\begin{pmatrix} \mathcal{K}_k(\theta)\,\partial_{\theta^T}\phi_k(\theta) \\ \partial_{\theta^T} v_k(\theta) \\ O_{k+1,p} \end{pmatrix}.$$

Therefore, the optimal weight matrix is

$$\breve{A}_n^*(\theta) = (\,A_n^*(\theta) \quad O_{p,k+1}\,),$$

and thus the optimal prediction-based estimating function obtained from $\breve{H}_i(\theta)$ equals the optimal estimating function obtained from $H_i(\theta)$. It is therefore sufficient to consider the aforementioned smaller class of prediction-based estimating functions, which we do in the rest of the paper.

The pseudo-score function, $\partial_\theta \tilde{\ell}_n(\theta)$, is not equal to the optimal prediction-based estimating function. In fact,

$$\tilde{A}(\theta) = -S(\theta)M^{(1)}(\theta)^{-1}. \tag{20}$$

The magnitude of the difference between the two estimating functions depends on how small the entries of $M_n^{(2)}(\theta)$ are relative to the entries of $M^{(1)}(\theta)$. Because correlations decrease exponentially with the distance in time (see Reiß [21], page 26), the terms in the sum defining $M_n^{(2)}(\theta)$ can be small compared with the entries of $M^{(1)}(\theta)$; however, under what conditions this occurs and exactly how small the terms are depend on θ, Δ, and k.

In the next section we show that the limit

$$M^{(2)}(\theta) = \lim_{n\to\infty} M_n^{(2)}(\theta) = \sum_{j=1}^{\infty} [E_\theta(H_k(\theta)H_{k+j}(\theta)^T) + E_\theta(H_{k+j}(\theta)H_k(\theta)^T)] \tag{21}$$

exists. Therefore, we can define the following weight matrix, which does not depend on n:

$$A^*(\theta) = -S(\theta)\bar{M}(\theta)^{-1}, \tag{22}$$

where

$$\bar{M}(\theta) = \lim_{n\to\infty} \bar{M}_n(\theta) = M^{(1)}(\theta) + M^{(2)}(\theta). \tag{23}$$

The estimating function

$$G_n^*(\theta) = A^*(\theta) \sum_{i=k}^{n-1} H_i(\theta) \tag{24}$$

is asymptotically optimal and theoretically is easier to handle than $A_n^*(\theta)\sum_{i=k}^{n-1} H_i(\theta)$. In practice, the optimal weight matrices $A_n^*(\theta)$ or $A^*(\theta)$ usually must be calculated by simulation. The amount of computation can be reduced by using the approximation to $G_n^*(\theta)$ obtained by replacing $A^*(\theta)$ or $A_n^*(\theta)$ with the matrix obtained from (22) and (23) when $M^{(2)}(\theta)$ is replaced by a suitably truncated version of the series in (21). This does not make much difference, because the terms in the sum (21) decrease exponentially fast.

4. Asymptotics of the pseudo-likelihood estimator

In this section, we present the asymptotic properties of estimators obtained by solving the estimating equation $G_n(\hat{\theta}_n)=0$, where G_n is given by (16). Important particular cases of this are the maximum pseudo-likelihood estimator obtained by maximizing (6) and the optimal prediction-based estimator obtained by solving $G_n^*(\hat{\theta}_n)=0$ with G_n^* given by (24). The depth, k, of G_n is assumed fixed. The asymptotic properties are proven for a solution to the general equation (1) under the following assumption:

Condition 4.1.

(a) *The functions $K_\theta(t)$ and $A(\theta)$ are twice continuously differentiable with respect to θ.*
(b) *The $p\times(k+1)$ matrix $(\partial_\theta\phi_k^T(\theta)\ \ \partial_\theta v_k(\theta))$ has rank p (in particular, $k+1\geq p$).*
(c) *$A(\theta)\bar{\mathcal{K}}(\bar{\phi}_k(\theta_0)-\bar{\phi}_k(\theta))=0$ if and only if $\theta=\theta_0$.*

Here

$$\bar{\mathcal{K}}=\begin{pmatrix}\mathcal{K}_k(\theta_0) & O_{k,1}\\ O_{1,k} & 1\end{pmatrix}, \tag{25}$$

and

$$\bar{\phi}_k(\theta)=\begin{pmatrix}\phi_k(\theta)\\ v_k(\theta)+2\phi_k(\theta)^T\kappa_k(\theta_0)-\phi_k(\theta)^T\mathcal{K}_k(\theta_0)\phi_k(\theta)\end{pmatrix}.$$

If A equals $\tilde{A}$ (corresponding to the pseudo-score function) or A equals A^* (corresponding to the optimal prediction-based estimating function), then Condition 4.1(a) is satisfied if $K_\theta(t)$ is three times continuously differentiable, which is the case for the models considered in Examples 2.1 and 2.2. Condition 4.1(a) ensures that the functions $\phi_k(\theta)$ and $v_k(\theta)$ are continuously differentiable. Condition 4.1(c) is an identifiability condition that ensures eventual uniqueness of the estimator.

Theorem 4.2. *Assume that the true parameter value θ_0 belongs to the interior of the parameter space Θ. Suppose that Condition 4.1 is satisfied, and that the matrix $A(\theta_0)$ has full rank p. Then a consistent estimator $\hat{\theta}_n$ that solves the estimating equation $G_n(\hat{\theta}_n)=0$ exists and is unique in any compact subset of Θ containing θ_0 with a probability tending to 1 as $n\to\infty$. Moreover,*

$$\sqrt{n}(\hat{\theta}_n-\theta_0)\xrightarrow{\mathcal{D}}N_p(0,U(\theta_0)^{-1}V(\theta_0)(U(\theta_0)^{-1})^T)$$

as $n \to \infty$, where $V(\theta_0) = A(\theta_0)\bar{M}(\theta_0)A(\theta_0)^T$ with $\bar{M}(\theta_0)$ given by (23), and

$$U(\theta_0) = E_{\theta_0}(\partial_\theta G_n(\theta_0)^T)/(n-k) = S(\theta_0)A(\theta_0)^T. \tag{26}$$

Here $S(\theta)$ is the sensitivity matrix given by (18).

Note that it follows from (19) and (20) that $A^*(\theta_0)$ and $\tilde{A}(\theta_0)$ have rank p if $S(\theta_0)$ has rank p, because $\bar{M}_n$ and $M^{(1)}$ are non-singular covariance matrices. That $S(\theta_0)$ has rank p follows from Condition 4.1(b) by (27) below.

Proof of Theorem 4.2. The theorem follows from general asymptotic statistical results for stochastic processes (see, e.g., Jacod and Sørensen [13]). We need to establish that a law of large numbers and a central limit theorem hold and to check regularity conditions.

Under our general assumption that X is stationary, Reiß [21] (page 25) showed that X is exponentially β-mixing. Therefore, a law of large numbers holds for sums of the form $n^{-1}\sum_{i=1}^{n} f(X_{i+k:i})$. The process $\{H_i(\theta_0)\}$ is exponentially α-mixing, and because the process X is Gaussian, $H_i(\theta)$ has moments of all orders. Therefore, it follows from Theorem 1 in Section 1.5 of Doukhan [5] that (21) converges, and that

$$\frac{G_n(\theta_0)}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, V(\theta_0))$$

as $n \to \infty$.

Next, we need to check regularity conditions that ensure the asymptotic results. The estimating function satisfies that $E_{\theta_0}(G_n(\theta_0)) = 0$. Furthermore, it is obvious from the continuous differentiability of the functions $\phi_k(\theta)$ and $v_k(\theta)$ that the derivatives $\partial_{\theta_j} A(\theta)H_i(\theta) + A(\theta)\,\partial_{\theta_j} H_i(\theta)$ are locally dominated integrable. Finally, the matrix $U(\theta_0)$ is invertible because $A(\theta_0)$ has full rank and

$$S(\theta_0) = -\left(\,\partial_\theta \phi_k^T(\theta_0) \quad \partial_\theta v_k(\theta_0)\,\right)\bar{\mathcal{K}}, \tag{27}$$

with $\bar{\mathcal{K}}$ given by (25). The first matrix has full rank by Condition 4.1(b), and $\bar{\mathcal{K}}$ is invertible because $\mathcal{K}_k(\theta_0)$ is the covariance matrix of $X(\Delta), \ldots, X(k\Delta)$, which is not degenerate. Now the existence and consistency of $\hat{\theta}_n$, as well as the eventual uniqueness of a consistent estimator on any compact subset of Θ containing θ_0, follow (see Jacod and Sørensen [13]). The locally dominated integrability of $A(\theta)H_i(\theta)$ (which follows from Condition 4.1(a)) implies that $n^{-1}G_n(\theta)$ converges uniformly to $A(\theta)E_{\theta_0}(H_i(\theta)) = A(\theta)\bar{\mathcal{K}}(\bar{\phi}_k(\theta_0) - \bar{\phi}_k(\theta))$ for θ in a compact set. The fact that the limit is a continuous functions of θ and satisfies $A(\theta)E_{\theta_0}(H_i(\theta)) \neq 0$ for $\theta \neq \theta_0$ implies that any non-consistent solution to the estimating equation will eventually leave any compact subset of Θ containing θ_0. The asymptotic normality follows by standard arguments (see, e.g., Jacod and Sørensen [13]). □

A simpler estimator with the same asymptotic distribution as in the estimator from (16) is obtained from the estimating function

$$G_n^\circ(\theta) = A(\theta_n^\circ)\sum_{i=k}^{n-1} H_i(\theta),$$

where θ_n° is some consistent estimator of θ, obtained, for instance, by simply using p suitably chosen coordinates of $H_i(\theta)$. For this estimating function, the identifiability condition Condition 4.1(c) can be replaced by the following condition:

Condition 4.3.

(a) *The function* $(\phi_k(\theta), v_k(\theta))$ *is one-to-one.*
(b) $\bar{\phi}_k(\theta_0) - \bar{\phi}_k(\theta) \in N^\perp$ *for all* $\theta \in \Theta$, *where* N *is the null space of the matrix* $A(\theta_0)\bar{\mathcal{K}}$.

This readily follows from the fact that the limit of $n^{-1}G_n^\circ(\theta)$ is $A(\theta_0)\bar{\mathcal{K}}(\bar{\phi}_k(\theta_0) - \bar{\phi}_k(\theta))$. In the case of the pseudo-likelihood function, we have the simple expression

$$\tilde{A}(\theta_0)\bar{\mathcal{K}} = \left(v_k(\theta_0)^{-1}\, \partial_\theta \phi_k^T(\theta_0)\mathcal{K}_k(\theta_0) \quad \tfrac{1}{2} v_k(\theta_0)^{-2}\, \partial_\theta v_k(\theta_0) \right).$$

Condition 4.3(a) is a basic assumption without which there is no hope of estimating θ using the pseudo-score function (15). The condition must be checked for individual models. Obviously, it is not always satisfied, as demonstrated by the following examples. Consider the model in Example 2.1 with the restriction that $b = a$. For this model, the autocovariance function depends on r and b only through $r - b^{-1} > 0$ for $t \in [0, r]$. In Example 2.2 with the restriction that $a = 0$, the autocovariance function depends on r and b only through $r\sqrt{b}$ for $t \in [0, r]$.

Theorem 4.2 implies that the asymptotic distribution of the optimal prediction-based estimator, $\hat{\theta}_n^*$, is

$$\sqrt{n}(\hat{\theta}_n^* - \theta_0) \xrightarrow{\mathcal{D}} N_p(0, (S(\theta_0)\bar{M}(\theta_0)^{-1}S(\theta_0)^T)^{-1}), \tag{28}$$

and the asymptotic distribution of the pseudo-likelihood estimator, $\tilde{\theta}_n$, is

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N_p\big(0, W(\theta_0)^{-1} + W(\theta_0)^{-1}B(\theta_0)W(\theta_0)^{-1}\big), \tag{29}$$

where

$$W(\theta) = S(\theta)M^{(1)}(\theta)^{-1}S(\theta)^T = \frac{\partial_\theta \phi_k(\theta)^T \mathcal{K}_k(\theta)\, \partial_{\theta^T}\phi_k(\theta)}{v_k(\theta)} + \frac{\partial_\theta v_k(\theta)\, \partial_{\theta^T} v_k(\theta)}{2v_k(\theta)^2}$$

and

$$B(\theta) = \tilde{A}(\theta)M^{(2)}(\theta)\tilde{A}(\theta)^T = S(\theta)M^{(1)}(\theta)^{-1}M^{(2)}(\theta)M^{(1)}(\theta)^{-1}S(\theta)^T.$$

The result for $\hat{\theta}_n^*$ follows because

$$-S(\theta_0)A^*(\theta_0)^T = A^*(\theta_0)\bar{M}(\theta_0)A^*(\theta_0)^T = S(\theta_0)\bar{M}(\theta)^{-1}S(\theta_0)^T,$$

and the result for $\tilde{\theta}_n$ follows because

$$-S(\theta_0)\tilde{A}(\theta_0)^T = S(\theta_0)M^{(1)}(\theta_0)^{-1}S(\theta_0)^T$$

and

$$\tilde{A}(\theta_0)\bar{M}(\theta_0)\tilde{A}(\theta_0)^T = S(\theta_0)M^{(1)}(\theta_0)^{-1}S(\theta_0)^T + \tilde{A}(\theta_0)M^{(2)}(\theta_0)\tilde{A}(\theta_0)^T.$$

According to the general theory of estimating functions (see, e.g., Heyde [11]), the matrix $S(\theta_0)\bar{M}(\theta_0)^{-1}S(\theta_0)^T - (W(\theta_0)^{-1} + W(\theta_0)^{-1}B(\theta_0)W(\theta_0)^{-1})^{-1}$ is positive definite; that is, the asymptotic covariance matrix of $\tilde{\theta}_n$ is larger than that of $\hat{\theta}_n^*$ (in the usual ordering of positive semi-definite matrices). Thus the asymptotic variance of $f(\tilde{\theta}_n)$ is larger than that of $f(\hat{\theta}_n^*)$ for any differentiable function $f:\mathbb{R}^p \mapsto \mathbb{R}$. If $B(\theta_0)$ is invertible, then

$$[W(\theta_0)^{-1} + W(\theta_0)^{-1}B(\theta_0)W(\theta_0)^{-1}]^{-1} = W(\theta_0) - [B(\theta_0)^{-1} + W(\theta_0)^{-1}]^{-1},$$

and if $M^{(2)}(\theta_0)$ is invertible, then

$$\bar{M}(\theta_0)^{-1} = M^{(1)}(\theta_0)^{-1} - M^{(1)}(\theta_0)^{-1}\big[M^{(1)}(\theta_0)^{-1} + M^{(2)}(\theta_0)^{-1}\big]^{-1}M^{(1)}(\theta_0)^{-1},$$

where we have used twice that $(I+A)^{-1} = I - A(I+A)^{-1}$ for a matrix A. Thus the difference between the two inverse asymptotic covariance matrices can be expressed as

$$\begin{aligned} &S(\theta_0)\bar{M}(\theta_0)^{-1}S(\theta_0)^T - [W(\theta_0)^{-1} + W(\theta_0)^{-1}B(\theta_0)W(\theta_0)^{-1}]^{-1} \\ &\quad = [B(\theta_0)^{-1} + W(\theta_0)^{-1}]^{-1} \\ &\qquad - S(\theta_0)M^{(1)}(\theta_0)^{-1}\big[M^{(1)}(\theta_0)^{-1} + M^{(2)}(\theta_0)^{-1}\big]^{-1}M^{(1)}(\theta_0)^{-1}S(\theta_0)^T \qquad (30) \\ &\quad = \big[\big(\tilde{A}(\theta_0)M^{(1)}(\theta_0)\tilde{A}(\theta_0)^T\big)^{-1} + \big(\tilde{A}(\theta_0)M^{(2)}(\theta_0)\tilde{A}(\theta_0)^T\big)^{-1}\big]^{-1} \\ &\qquad - \tilde{A}(\theta_0)\big[M^{(1)}(\theta_0)^{-1} + M^{(2)}(\theta_0)^{-1}\big]^{-1}\tilde{A}(\theta_0)^T. \end{aligned}$$

It is considerably easier to calculate the pseudo-likelihood function (6) than the optimal estimating function (24), because the latter involves derivatives with respect to θ of the covariance function and higher-order moments of X. In particular, in cases where the covariance function is not explicitly known and must be determined by simulation, it is much easier to calculate (6) than (24). Thus the maximum pseudo-likelihood estimator is preferred in practice. The formula (30) then can be used to assess whether the loss of efficiency relative to the optimal estimator is acceptable.

Example 4.4. As an example, we calculated the efficiency loss for the model (7) in Example 2.1 in a number of cases. When k is sufficiently large, the pseudo-likelihood function is almost efficient, and thus the information loss (30) is necessarily small. Therefore, it is most interesting to calculate the efficiency loss when k is small. We calculate the relative information loss, that is, the information loss (30) relative to the information for the optimal estimator given by (28). The main problem is to calculate the matrix (21). However, for $k = 1$, a simple expression for each term in the sum (21) can be obtained using the formula of Isserlis [12], and so a suitably truncated version of the sum can be easily calculated. For $k \geq 2$, the matrix (21) can be determined by simulation using (23). Specifically, we determine the covariance matrix $\bar{M}_n(\theta_0)$, with n suitably large, by simulation. This is computationally more demanding.

We first considered the efficiency loss for the parameter b in the case $k = 1$. The parameters σ^2 and r were fixed at a value of 1, and a was chosen equal to -1. For $b = -\mathrm{e}^{-2} = -0.1353$

(the value for which the mixing rate is maximal), the relative information loss was found to be very small, less than 0.1 percent for $\Delta = 0.1$, $\Delta = 0.5$, and $\Delta = 1$.

Next, we calculated the information loss for a number of values of b with $\Delta = 1$. For $b = -0.3, -0.06, 0.05, 0.1, 0.2, 0.3$, and 0.5, the mixing rate is relatively high, and the information loss is less than 0.1 percent. For $b = -0.5, -0.4, 0.7$, and 0.9, the information loss is between 0.1 and 1 percent, whereas for $b = -0.6, -0.7$, and -0.9, it is 1.4 percent, 3.0 percent, and 9.8 percent, respectively.

Finally, we calculated the relative the information loss for $k = 3$ and $k = 5$. In this case, information loss was calculated for both a and b. The parameters a, σ^2, and r had the same value as before, and $\Delta = 1$. For $b = -\mathrm{e}^{-2}$, the relative information loss for both a and b is less than 0.1 percent for both values of k. For $b = -0.5$ the information loss is less than 0.1 percent for b and 0.2 percent for a.

In most cases, the relative information loss is so tiny that in practice it is preferable to use the maximum pseudo-likelihood estimator. The information loss increases as the mixing rate decreases. Only for $k = 1$ and $b = -0.9$ is the information loss large enough to justify the use of the more complicated optimal estimator.

5. Simulation study

In this section we report the results of a simulation study in which we investigated some properties of the pseudo-likelihood estimator introduced in Section 2. We restrict ourselves to the model considered in Example 2.1 and to estimating $\theta = (a, b)$. The delay time r is chosen equal to 1, and σ^2 is fixed at 1. This study was not intended to serve as a complete simulation study; rather, the intention was to illustrate some properties of the estimator and give a first impression of how the joint distribution of the two-dimensional estimator $\tilde{\theta}_n = (\tilde{a}_n, \tilde{b}_n)$ depends on the time between observations Δ, the depth k of the pseudo-likelihood function, and the true parameter value. We performed simulations for three values of θ: $\theta = (-1, 0.95)$ near the upper boundary of the domain of stationarity, $\theta = (-1, -1/\mathrm{e}^2) = (-1, -0.1353)$ which is the parameter value with the highest possible mixing rate for the stationary solution X when $a = -1$, and $\theta = (-1, -2.1)$ near the lower boundary of the domain of stationarity. For each parameter value, we considered four sampling frequencies with the same number of observation time points (200); specifically, the observation time points were $i\Delta$, $i = 1, \dots, 200$, with $\Delta = 0.05, 0.1, 0.5, 1$. The simulations of the SDDE were conducted with a step size of 0.001. In all cases, 1000 data sets were simulated, and thus 1000 estimates were generated. For each data set, a new trajectory of the driving Wiener process was generated. The full simulation study is reported in Küchler and Sørensen [18].

Table 1 reports the mean values and standard deviations of the simulated estimates of a and b for $(a, b) = (-1, -0.1353)$. For $(a, b) = (-1, 0.95)$, the estimators are more biased and have a larger standard deviation for small values of Δ and k, whereas for $(a, b) = (-1, -2.1)$, the bias is small and the standard deviations are comparable in all cases. For $(a, b) = (-1, 0.95)$, the estimators of a and b are highly correlated, whereas this is the case only for small values of Δ and k for the other parameter values.

The most remarkable observations from our simulation study can be summarized as follows:

Table 1. Mean and standard deviation of the pseudo-likelihood estimator of a (upper part of the table) and b (lower part of the table) for various values of depth k, time between observations Δ, and number of observations n. In all cases $n\Delta$, the length of the observation interval, is 200, and the true parameter values are $a = -1$ and $b = -0.1353$

		k						
Δ	n	1	3	5	7	9	13	20
0.05	4000	−1.73	−1.07	−1.04	−1.02	−1.02	−1.01	−1.01
		2.32	0.21	0.14	0.11	0.11	0.10	0.09
0.1	2000	−1.27	−1.03	−1.03	−1.01	−1.01	−1.01	−1.01
		0.83	0.13	0.10	0.09	0.09	0.09	0.09
0.5	400	−1.04	−1.01	−1.01	−1.01	−1.01	−1.01	−1.01
		0.14	0.10	0.09	0.09	0.10	0.09	0.10
1.0	200	−1.02	−1.01	−1.01	−1.02	−1.02	−1.01	−1.02
		0.12	0.11	0.11	0.11	0.11	0.11	0.12
2.0	100	−1.10	−1.04	−1.03	−1.03	−1.04	−1.04	−1.04
		0.43	0.21	0.19	0.18	0.21	0.19	0.17
0.05	4000	0.42	−0.09	−0.11	−0.15	−0.13	−0.14	−0.14
		2.66	0.44	0.31	0.23	0.19	0.14	0.09
0.1	2000	0.08	−0.13	−0.14	−0.14	−0.14	−0.13	−0.13
		1.14	0.27	0.18	0.13	0.11	0.09	0.09
0.5	400	−0.12	−0.14	−0.14	−0.13	−0.14	−0.14	−0.14
		0.28	0.10	0.11	0.11	0.11	0.11	0.10
1.0	200	−0.14	−0.14	−0.14	−0.13	−0.14	−0.13	−0.13
		0.16	0.13	0.13	0.13	0.13	0.13	0.14
2.0	100	−0.26	−0.15	−0.14	−0.13	−0.15	−0.14	−0.14
		0.65	0.29	0.27	0.25	0.30	0.27	0.24

- For a fixed number of observation time points, the bias and standard deviation of the estimators worsen as the time between observations Δ decreases, at least when $\Delta \leq r$. For $\Delta > r$, the quality does not change much with Δ, and whether the bias and variance increase or decrease with Δ depends on the parameter value.
- The smaller the Δ value, the more the choice of the depth k of the pseudo-likelihood functions influences the quality of the estimators when $\Delta \leq r$. For $\Delta > r$, the importance of k increases again for some parameter values.
- It is surprising that a similar pattern is seen when the length of the observation interval $n\Delta$ is fixed so that the sample size decreases as Δ increases. However, here there is a clearer tendency for the estimators to deteriorate when $\Delta > r$, so that there is an optimal value of Δ, which seems to be around r.
- The absolute value of the correlation between $\tilde{a}$ and $\tilde{b}$ decreases with increasing depth k to a limit, which is strongly dependent on the true parameter value. Near the upper boundary

of the stability region, the estimators are highly correlated. A high absolute value of the correlation indicates that it is difficult to distinguish between the effects of the lagged term and the nonlagged term in the drift; thus, it is not surprising that the absolute correlation is large when the depth is small.

- For small values of the depth k, the joint distribution of the estimators of a and b can deviate from a two-dimensional normal distribution by having crescent-shaped contours.

Acknowledgements

We thank Adrian M. Kämmler for assisting with the efficiency loss calculations, Katja Krol for writing the PC program used in the simulation study, and Daniel Skodlerack for providing an earlier version of that program. We also thank the referees, the associate editor, and the editor for their constructive criticism that led to considerable improvement of the manuscript. The research of Michael Sørensen was supported by the Danish Center for Accounting and Finance, funded by the Danish Social Science Research Council; by the Center for Research in Econometric Analysis of Time Series, funded by the Danish National Research Foundation; and by a grant from the University of Copenhagen Excellence Programme.

References

[1] Brockwell, P.J. and Davis, R.A. (1991). *Time Series*: *Theory and Methods*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR1093459

[2] Buckwar, E. (2000). Introduction to the numerical analysis of stochastic delay differential equations. *J. Comput. Appl. Math.* **125** 297–307. MR1803198

[3] Diekmann, O., van Gils, S.A., Verduyn Lunel, S.M. and Walther, H.O. (1995). *Delay Equations*: *Functional-*, *Complex-*, *and Nonlinear Analysis*. *Applied Mathematical Sciences* **110**. New York: Springer. MR1345150

[4] Ditlevsen, S. and Sørensen, M. (2004). Inference for observations of integrated diffusion processes. *Scand. J. Statist.* **31** 417–429. MR2087834

[5] Doukhan, P. (1994). *Mixing*: *Properties and Examples*. *Lecture Notes in Statistics* **85**. New York: Springer. MR1312160

[6] Forman, J.L. and Sørensen, M. (2008). The Pearson diffusions: A class of statistically tractable diffusion processes. *Scand. J. Statist.* **35** 438–465. MR2446729

[7] Godambe, V.P. and Heyde, C.C. (1987). Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.* **55** 231–244. MR0963141

[8] Gushchin, A.A. and Küchler, U. (1999). Asymptotic inference for a linear stochastic differential equation with time delay. *Bernoulli* **5** 1059–1098. MR1735785

[9] Gushchin, A.A. and Küchler, U. (2000). On stationary solutions of delay differential equations driven by a Lévy process. *Stochastic Process. Appl.* **88** 195–211. MR1767844

[10] Gushchin, A.A. and Küchler, U. (2003). On parametric statistical models for stationary solutions of affine stochastic delay differential equations. *Math. Methods Statist.* **12** 31–61. MR1990513

[11] Heyde, C.C. (1997). *Quasi-Likelihood and Its Application*: *A General Approach to Optimal Parameter Estimation*. *Springer Series in Statistics*. New York: Springer. MR1461808

[12] Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika* **12** 134–139.

[13] Jacod, J. and Sørensen, M. (2012). Asymptotic statistical theory for stochastic processes: A review. Preprint, Dept. Mathematical Sciences, Univ. Copenhagen.

[14] Küchler, U. and Kutoyants, Y.A. (2000). Delay estimation for some stationary diffusion-type processes. *Scand. J. Statist.* **27** 405–414. MR1795773

[15] Küchler, U. and Mensch, B. (1992). Langevin's stochastic differential equation extended by a time-delayed term. *Stochastics Stochastics Rep.* **40** 23–42. MR1275126

[16] Küchler, U. and Platen, E. (2000). Strong discrete time approximation of stochastic differential equations with time delay. *Math. Comput. Simulation* **54** 189–205. MR1800113

[17] Küchler, U. and Platen, E. (2007). Time delay and noise explaining cyclical fluctuations in prices of commodities. Preprint, Inst. of Mathematics, Humboldt-Univ. Berlin.

[18] Küchler, U. and Sørensen, M. (2007). Statistical inference for discrete-time samples from affine stochastic delay differential equations. Preprint, Dept. Mathematical Sciences, Univ. Copenhagen.

[19] Küchler, U. and Sørensen, M. (2010). A simple estimator for discrete-time samples from affine stochastic delay differential equations. *Stat. Inference Stoch. Process.* **13** 125–132. MR2653983

[20] Küchler, U. and Vasiliev, V. (2005). Sequential identification of linear dynamic systems with memory. *Stat. Inference Stoch. Process.* **8** 1–24. MR2122699

[21] Reiß, M. (2002). Nonparametric estimation for stochastic delay differential equations. Ph.D. thesis, Institut für Mathematik, Humboldt-Universität zu Berlin.

[22] Reiß, M. (2002). Minimax rates for nonparametric drift estimation in affine stochastic delay differential equations. *Stat. Inference Stoch. Process.* **5** 131–152. MR1917289

[23] Reiss, M. (2005). Adaptive estimation for affine stochastic delay differential equations. *Bernoulli* **11** 67–102. MR2121456

[24] Sørensen, H. (2003). Simulated likelihood approximations for stochastic volatility models. *Scand. J. Statist.* **30** 257–276. MR1983125

[25] Sørensen, M. (2000). Prediction-based estimating functions. *Econom. J.* **3** 123–147. MR1820411

[26] Sørensen, M. (2011). Prediction-based estimating functions: Review and new developments. *Braz. J. Probab. Stat.* **25** 362–391.

Received April 2009