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

Uwe Küchler
Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6, Berlin Mitte D-10099, Germany

Michael Sørensen
Department of Applied Mathematics and Statistics
University of Copenhagen
Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark

Abstract

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. Also a more general class of prediction-based estimating functions is investigated. 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. For models where the delay measure is concentrated on a finite set, an estimator obtained by discretization of the continuous-time likelihood function is presented, and its asymptotic properties are investigated. The estimator is very easy to calculate, but it is shown to have a significant bias when the sampling frequency is low. Two examples of affine stochastic delay equation are considered in detail.

Key words: Asymptotic normality, consistency, discrete time observation of continuous time models, prediction-based estimating functions, pseudo-likelihood, stochastic delay differential equation, quasi-likelihood.

AMS-Classification: 62M09, 34K50.

1 Introduction

In the last decade statistical inference for stochastic delay differential equations (SDDEs) has been studied from various view points. Early work on maximum likelihood estimation was done by Küchler & Mensch (1992). Gushchin & Küchler (1999) and Küchler & Kutoyants (2000) determined the non-standard asymptotic properties of the maximum likelihood estimator for SDDEs, while Küchler & Vasil'jev (2005) constructed sequential procedures with a given accuracy in the L_2-sense. Nonparametric estimators for affine SDDEs were investigated by Reiß (2002a). All these studies were concerned with continuous observation of the solution process.

As opposed to the situation for ordinary stochastic differential equation, observations at discrete time points have so far been studied very little for SDDEs. Reiß (2002b) has studied nonparametric estimation. In this paper we make a first attempt at investigating parametric inference for affine stochastic delay equations observed at discrete time points. To apply the methods in practice it is often necessary to be able to simulate solutions of SDDEs. This problem has been studied by e.g. Küchler & Platen (2000) and Buckwar (2000). For a practical application of one of the simplest SDDEs, discussed in Example 2.1 below, see Küchler & Platen (2007).

We consider the model given by the stochastic differential equation

$$dX(t) = \int_{-r}^{0} X(t+s)a_\alpha(ds)dt + \sigma dW(t), \tag{1.1}$$

where a_α is a measure on $[-r, 0]$ such that (1.1) has a unique stationary solution (for a given initial condition). Conditions under which (1.1) has a unique stationary solution were given by Gushchin & Küchler (2000). We assume that the measure a_α depends on a parameter α. The parameter about which inference is to be drawn is $\theta = (\alpha, \sigma) \in \Theta \subseteq \mathbb{R}^p$ ($\sigma > 0$). The process W is a Wiener process. The initial condition is that the distribution of $\{X(s) \,|\, s \in [-r, 0]\}$ is the stationary distribution, which has always expectation zero. The data are observations at discrete time points $X(\Delta), X(2\Delta), \ldots, X(n\Delta)$.

An interesting particular case of (1.1) is

$$dX(t) = \sum_{k=1}^{N} \alpha_k X(t - r_k)dt + \sigma dW(t). \tag{1.2}$$

Here the measure a_α is concentrated in the discrete points $-r_1, \ldots, -r_N$, ($r_i \geq 0$). The vector $(r_1, \cdots, r_N)$ can 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 we propose a pseudo-likelihood function that approximates the likelihood function well and is considerably easier to calculate. Two examples are considered in detail. In Section 3 we present prediction-based estimating functions for affine stochastic delay equations and show that the pseudo-likelihood estimator is a particular case of a prediction-based estimator. Thus 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 to the optimal prediction-based estimating function. This is done in Section 4, where conditions ensuring consistency and asymptotic normality are given. In Section

5 we study the asymptotic properties of the estimator that is obtained by discretizing the likelihood function for continuous time observations of the model (1.2). It is shown to be asymptotically normal with an asymptotic bias of order Δ, where Δ is the time between observations. Finally, properties of the pseudo-likelihood estimator are investigated in a simulation study presented in Section 6.

2 The likelihood and the pseudo-likelihood function

Since the data are in fact a Gaussian time series with expectation zero, the likelihood function can in principle be calculated if we can determine, analytically or numerically, the autocovariances

$$K_\theta(t) = E_\theta(X(0)X(t)), \quad t \geq 0. \tag{2.1}$$

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

$$\partial_t K_\theta(t) = \int_{-r}^{0} K_\theta(t+s) a_\theta(ds), \quad t \geq 0, \tag{2.2}$$

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 & Küchler (2003). The condition $\partial_t K_\theta(0+) = -\frac{1}{2}\sigma^2$ can also be written in the form

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

The equation (2.2) is a continuous time analogue of the Yule-Walker equation known from time-series analysis, and we will refer to (2.2) as the delay Yule-Walker equation of (2.1). In general, this equation must be solved numerically, but below we shall consider two particular examples, where it can be solved explicitly.

To calculate the *likelihood function*, define for every $\ell = 1, \ldots, n$ the ℓ-dimensional vector $\kappa_\ell(\theta) = (K_\theta(\Delta), \ldots, K_\theta(\ell\Delta))^T$ and the $\ell \times \ell$-matrix $\mathcal{K}_\ell(\theta) = \{K_\theta((i-j)\Delta)\}_{i,j=1,\ldots,\ell}$. 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 $X(\Delta), \ldots, X(\ell\Delta)$.

The conditional distribution of the observation $X((i+1)\Delta)$ given the previous observations $X(\Delta), \ldots, X(i\Delta)$ 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 $X_{i:j} = (X(i\Delta), \ldots, X(j\Delta))^T$, $i > j \geq 1$. The vector $\phi_i(\theta) = (\phi_{i,1}(\theta), \ldots, \phi_{i,i}(\theta))^T$ and the conditional variance $v_i(\theta)$ can be found by means of the Durbin-Levinson algorithm, see e.g. p. 169 in Brockwell & Davis (1991). Specifically, $\phi_{1,1}(\theta)) = K_\theta(\Delta)/K_\theta(0)$ and $v_0(\theta) = K_\theta(0)$, while

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

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

and

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

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

$$L_n(\theta) = \frac{1}{\sqrt{2\pi v_0(\theta)}} \exp\left(-\frac{1}{2v_0(\theta)} X(\Delta)^2\right) \tag{2.3}$$
$$\times \prod_{i=1}^{n-1} \left[\frac{1}{\sqrt{2\pi v_i(\theta)}} \exp\left(-\frac{1}{2v_i(\theta)}\left(X((i+1)\Delta) - \phi_i(\theta)^T X_{i:1}\right)^2\right)\right].$$

This expression quickly gets very complicated 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. (1995) (p. 34). It is not difficult to see, using the Durbin-Levinson Algorithm, that this implies that the quantities $\phi_{i,j}(\theta)$ decreases exponentially with j. Hence the conditional distribution of $X((i+1)\Delta)$ given $X(\Delta), \ldots, X(i\Delta)$ depends only very little on observations in the far past.

We therefore propose to use instead a *pseudo-likelihood function* obtained by replacing in the likelihood function the conditional density of $X((i+1)\Delta)$ given $X(\Delta), \ldots, X(i\Delta)$ by the conditional density of $X((i+1)\Delta)$ given $X((i+1-k)\Delta), \ldots, X(i\Delta)$, where k will typically be relatively small. This pseudo-likelihood function was proposed by H. Sørensen (2003) in connection with stochastic volatility models, but the idea is widely applicable. The pseudo-likelihood is given by

$$\tilde{L}_n(\theta) = \prod_{i=k}^{n-1} \left[\frac{1}{\sqrt{2\pi v_k(\theta)}} \exp\left(-\frac{1}{2v_k(\theta)}\left(X((i+1)\Delta) - \phi_k(\theta)^T X_{i:i+1-k}\right)^2\right)\right]. \tag{2.4}$$

We have not included the density of $X_{k:1}$. Note that the computational gain is large because to calculate (2.4) we only need to find $\phi_k(\theta)$ and $v_k(\theta)$ once. The number k will be called the *depth* of the pseudo-likelihood function.

Example 2.1 Consider the equation

$$dX(t) = [aX(t) + bX(t-r)]\,dt + \sigma dW(t), \tag{2.5}$$

where $r > 0$, $\sigma > 0$. This is a particular case of the model (1.2). The real parameters a and b are chosen such that a stationary solution of (2.5) exists. This is the case exactly when $a < r^{-1}$ and $-a/\cos(\xi(ar)) < b < -a$ if $a \neq 0$, and $-\pi/2 < br < 0$ 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. For details, see Küchler & Mensch (1992). In that paper the covariance function of the stationary solution is explicitly found by solving the Yule-Walker delay differential equation (2.2):

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

It is found that

$$K_\theta(0) = \begin{cases} \dfrac{\sigma^2(b\sinh(\lambda(a,b)r) - \lambda(a,b))}{2\lambda(a,b)[a + b\cosh(\lambda(a,b)r)]} & \text{when } |b| < -a \\ \sigma^2(br-1)/(4b) & \text{when } b = a \\ \dfrac{\sigma^2(b\sin(\lambda(a,b)r) - \lambda(a,b))}{2\lambda(a,b)[a + b\cos(\lambda(a,b)r)]} & \text{when } b < -|a|, \end{cases} \tag{2.6}$$

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

$$K_\theta(t) = \begin{cases} K_\theta(0)\cosh(\lambda(a,b)t) - \sigma^2(2\lambda(a,b))^{-1}\sinh(\lambda(a,b)t) & \text{when } |b| < -a \\ K_\theta(0) - \frac{1}{2}t\sigma^2 & \text{when } b = a \\ K_\theta(0)\cos(\lambda(a,b)t) - \sigma^2(2\lambda(a,b))^{-1}\sin(\lambda(a,b)t) & \text{when } b < -|a|. \end{cases} \tag{2.7}$$

Because $K_\theta(t)$ is known in $[0,r]$, the Yule-Walker equation turns into an ordinary differential equation for $K_\theta(t)$ in $[r,2r]$, which is solved by

$$K_\theta(t) = b\int_r^t e^{a(t-s)}K_\theta(s-r)ds + e^{a(t-r)}K_\theta(r), \quad t \in [r,2r]. \tag{2.8}$$

Thus it is possible to determine $K_\theta(t)$ explicitly in this interval too. In a similar way, $K_\theta(t)$ can be determined iteratively in each of the intervals $t \in [nr,(n+1)r]$, $n \geq 2$. Note that the covariance function depends on σ and r in a simple and smooth way, so these parameters can also be estimated by maximizing the pseudo-likelihood function (2.4).

For $b = 0$ the model (2.5) is the Ornstein-Uhlenbeck process, for which (2.6) and (2.7) simplifies to the well-known result that $K_\theta(t) = -(\sigma^2/(2a))e^{at}$ $(t \geq 0)$ in the stationary case $a < 0$. For $a = 0$, we obtain the model

$$dX(t) = bX(t-r)dt + \sigma dW(t). \tag{2.9}$$

This process is stationary when $br \in (-\pi/2, 0)$, and by (2.6) and (2.7) the autocovariance function is given by

$$K_\theta(t) = -\frac{\sigma^2}{2b}\left(\frac{1-\sin(br)}{\cos(br)}\cos(bt) + \sin(bt)\right) \tag{2.10}$$

when $t \in [0,r]$. By (2.8) we find that

$$K_\theta(t) = -\frac{\sigma^2}{2b}\left[2 + \cos(bt)\{(\tan(bt) - \tan(br))(1 - 2\sin(br)) - 1/\cos(br)\}\right] \tag{2.11}$$

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

□

Example 2.2 Consider the equation

$$dX(t) = -b \int_{-r}^{0} X(t+s) e^{as} ds\, dt + \sigma dW(t), \tag{2.12}$$

where $r > 0$, $\sigma > 0$. The set of values of the parameters a and b for which a unique stationary solution of (2.12) exists was studied by Reiß (2002b). This set is rather complicated and irregular. For instance, it is not convex. However, it contains the region $\{(a,b) \,|\, a \geq 0, b > 0, b(1+e^{-ar}) < \max(\pi^2/r^2, a^2(e^{ar}-1)^2)\}$. 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 this case a stationary solution exists for all $a > 0$ and $b > 0$.

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

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

For $a > 0$, an explicit expression for $K_\theta(t)$ involving trigonometric functions exists too, see p. 41 in Reiß (2002b), but it is somewhat complicated and is therefore omitted here.

□

3 Prediction-based estimating functions

In the following we discuss the pseudo-likelihood estimator in the framework of prediction-based estimating functions. This class of estimating functions was introduced by Sørensen (2000) as a generalization of the martingale estimating functions that is applicable also to non-Markovian processes such as solutions to stochastic delay differential equations. For an application of the methodology to observations of integrated diffusion processes, see Ditlevsen & Sørensen (2004). 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 & Heyde (1987), see Heyde (1997). 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:

$$\begin{aligned} \partial_\theta \tilde{\ell}_n(\theta) &:= \partial_\theta \log(\tilde{L}_n(\theta)) \\ &= \sum_{i=k}^{n-1} \frac{\partial_\theta \phi_k(\theta)^T X_{i:i+1-k}}{v_k(\theta)} \left(X((i+1)\Delta) - \phi_k(\theta)^T X_{i:i+1-k} \right) \\ &\quad + \frac{\partial_\theta v_k(\theta)}{2v_k(\theta)^2} \sum_{i=k}^{n-1} \left[\left(X((i+1)\Delta) - \phi_k(\theta)^T X_{i:i+1-k} \right)^2 - v_k(\theta) \right]. \end{aligned} \tag{3.1}$$

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 that is obtained by differentiating the Durbin-Levinson algorithm:

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

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

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

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

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

Since the minimum mean square error 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, the pseudo score function is a prediction-based estimating function in a slightly more general sense than in the original paper, Sørensen (2000), as here the second predicted function of the data depends on the parameters and previous observations. It is therefore of interest to explore the relations of the pseudo score function to the optimal estimating function based on these predictors. We shall see that neither the expression for the optimal estimating functions nor the asymptotic theory is changed by the fact that the prediction-based estimating functions considered here are slightly more general than those considered previously.

We start by introducing terminology like that in Sørensen (2000). Define the $2(k+1) \times 2$-matrices

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

and the $2(k+1)$-dimensional vectors

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

Then the full class of prediction-based estimating function to which (3.1) belongs is given by

$$G_n(\theta) = A(\theta) \sum_{i=k}^{n-1} H_i(\theta), \tag{3.2}$$

where $A(\theta)$ is a $p \times 2(k+1)$-matrix of weights that can depend on the parameter, but not on the data. The pseudo score function (3.1) is obtained if the weight matrix $A(\theta)$ is chosen as

$$\tilde{A}(\theta) = \begin{pmatrix} 0_{p,1} & \frac{\partial_\theta \phi_k(\theta)^T}{v_k(\theta)} & \frac{\partial_\theta v_k(\theta)}{2v_k(\theta)^2} & O_{p,k} \end{pmatrix},$$

with O_{j_1,j_2} denoting here and later the $j_1 \times j_2$-matrix of zeros. 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 that is closest to the true score function in an L^2-sense, for details see Heyde (1997).

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

$$\bar{M}_n(\theta) = M^{(1)}(\theta) + M_n^{(2)}(\theta), \tag{3.3}$$

where

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

and

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

with

$$\bar{\mathcal{K}}_k(\theta) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \mathcal{K}_k(\theta) & \\ 0 & & & \end{pmatrix},$$

and with $\mathcal{K}_k(\theta)$ denoting the covariance matrix of $X(\Delta), \ldots, X(k\Delta)$ defined in Section 2. To find the optimal estimating function we also need the $p \times 2(k+1)$ *sensitivity-matrix* $S(\theta)$ given by

$$S(\theta)^T = E_\theta(\partial_{\theta^T} H_i(\theta)) = - \begin{pmatrix} 0 \cdots 0 \\ \\ \mathcal{K}_k(\theta) \partial_{\theta^T} \phi_k(\theta) \\ \\ \partial_{\theta^T} v_k(\theta) \\ \\ O_{k,p} \end{pmatrix}. \tag{3.4}$$

For the derivation of the expression for the lower half of $S(\theta)^T$ it is crucial that the model is Gaussian so that $\phi_k(\theta)^T X_{i:i+1-k}$ is the conditional expectation and not just a minimum mean squares linear predictor as it is in the general theory of prediction-based estimating functions.

By arguments similar to those in Sørensen (2000), the optimal weight-matrix is given by

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

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

$$\tilde{A}(\theta) = -S(\theta) M^{(1)}(\theta)^{-1},$$

so

$$A_n^*(\theta) = \tilde{A}(\theta)[I - A_n^\#(\theta)].$$

where I denotes the identity matrix and

$$A_n^\#(\theta) = M_n^{(2)}(\theta) \bar{M}_n(\theta)^{-1}.$$

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 $M^{(1)}(\theta)$. Since correlations decrease exponentially with the distance in time, the terms in the sum defining $M_n^{(2)}(\theta)$ can be small compared to the entries of $M^{(1)}(\theta)$, but when this happens and exactly how small the terms are depend on θ, Δ and k.

In the next section we shall see that the limit

$$M^{(2)}(\theta) = \lim_{n\to\infty} M_n^{(2)}(\theta) = \sum_{j=1}^{\infty} \left[E_\theta(H_k(\theta) H_{k+j}(\theta)^T) + E_\theta(H_{k+j}(\theta) H_k(\theta)^T \right] \tag{3.5}$$

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

$$A^*(\theta) = S(\theta)\bar{M}(\theta)^{-1}, \tag{3.6}$$

where

$$\bar{M}(\theta) = \lim_{n\to\infty} \bar{M}_n(\theta) = M^{(1)}(\theta) + M^{(2)}(\theta). \tag{3.7}$$

The estimating function

$$G_n^*(\theta) = A^*(\theta) \sum_{i=k}^{n-1} H_i(\theta), \tag{3.8}$$

is asymptotically optimal and is easier to handle theoretically than $A_n^*(\theta) \sum_{i=k}^{n-1} H_i(\theta)$. In practice, the weight matrix $A_n^*(\theta)$ must often be calculated by simulation, and to reduce the amount of computation it is advisable to use the approximation to $G_n^*(\theta)$ obtained by replacing $A^*(\theta)$ or $A_n^*(\theta)$ by the matrix obtained from (3.6) and (3.7) if $M^{(2)}(\theta)$ is replaced by a suitably truncated version of the series in (3.5). This does not make much difference because the terms in the sum (3.5) decreases exponentially fast.

4 Asymptotics of the pseudo-likelihood estimator

In this section we will study the asymptotic properties of estimators obtained by solving the estimating equation $G(\hat{\theta}) = 0$, where G is given by (3.2). Important particular cases are the maximum pseudo-likelihood estimators obtained by maximizing (2.4) and the optimal prediction-based estimator obtained by solving $G_n^*(\hat{\theta}_n) = 0$ with G_n^* given by (3.8). The asymptotic properties are proven for a solution to the general equation (1.1) under the following assumption.

Condition 4.1 *The functions $K_\theta(t)$ and $A(\theta)$ are twice continuously differentiable with respect to θ.*

If A equals $\tilde{A}$ corresponding to the pseudo score function or A^* corresponding to the optimal prediction-based estimating function, then Condition 4.1 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.

From now on, let θ_0 denote the true parameter value. The expectation of $G_n(\theta_0)$ is zero, and its covariance matrix is

$$V_n(\theta_0) = (n-k)A(\theta_0)\bar{M}_n(\theta_0)A(\theta_0)^T, \tag{4.1}$$

with $\bar{M}_n(\theta)$ given by (3.3). We also need the derivative of $G_n(\theta)$

$$\partial_{\theta^T} G_n(\theta) = \partial_{\theta^T} A(\theta) \sum_{i=k}^{n-1} H_i(\theta) + A(\theta) \sum_{i=k}^{n-1} \partial_{\theta^T} H_i(\theta)$$

and

$$U(\theta_0) = E_{\theta_0}(\partial_\theta G_n(\theta_0)^T)/(n-k) = S(\theta_0)A(\theta_0)^T, \tag{4.2}$$

where $S(\theta)$ is the sensitivity matrix given by (3.4).

Theorem 4.2 *Suppose Condition 4.1 is satisfied, and that the matrices $A(\theta_0)$ and $S(\theta_0)$ have full rank. Then for every n an estimator $\hat{\theta}_n$ exists that solves the estimating equation $G_n(\hat{\theta}_n) = 0$ with a probability tending to one as $n \to \infty$. Moreover,*

$$\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$$

and

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

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 (3.7).

Remark: Note that a necessary condition that $S(\theta_0)$ has full rank is that $k + 1 \geq p$.

Proof: Under our general assumption that X is stationary, Reiß (2002b) has shown that X is exponentially β-mixing (see p. 25), and hence that the process $\{H_i(\theta_0)\}$ is exponentially α-mixing. Since the process X is Gaussian, $H_i(\theta)$ has moments of all orders. It therefore follows from Theorem 1 in Section 1.5 of Doukhan (1994) that (3.5) converges, and that

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

as $n \to \infty$. Now the theorem follows by a proof that is analogous to the proof of Theorem 6.2 in Sørensen (2000), in which conditions in Sørensen (1999) were shown to hold. That $U(\theta_0)$ is invertible follows from the assumptions that $A(\theta_0)$ and $S(\theta_0)$ have full rank. At first glance it seems to be a problem that the estimating functions considered here are more general than those in Sørensen (2000) in that the term $[X((i+1)\Delta) - \phi_k(\theta)^T X_{i:i+1-k}]^2$ appears in the definition of $H_i(\theta)$. However, by the multinomial formula

$$\begin{aligned}
&[X((i+1)\Delta) - \phi_k(\theta)^T X_{i:i+1-k}]^2 = \\
&\quad X((i+1)\Delta)^2 + \sum_{\nu_1,\ldots\nu_k} \binom{2}{\kappa_1, \ldots, \kappa_M} \phi_{k,1}(\theta)^{\nu_1} \cdots \phi_{k,k}(\theta)^{\nu_k} X(i\Delta)^{\nu_1} \cdots X((i-k+1)\Delta)^{\nu_k} \\
&\qquad - 2\sum_{j=1}^{k} \phi_{k,j}(\theta) X((i+1)\Delta) X((i-j+1)\Delta),
\end{aligned}$$

where the first sum is over all $0 \leq \nu_1, \ldots, \nu_k \leq 2$ such that $\nu_1 + \cdots + \nu_k = 2$. Thus $H_i(\theta)$ has a form similar to that in Sørensen (2000), and the ergodic theorem can be applied in the same way as in the proof of Theorem 6.2 in that paper.

□

Corollary 4.3 *Suppose the function $K_\theta(t)$ is three times continuously differentiable with respect to θ, and that the matrices $A^*(\theta_0)$, $\tilde{A}(\theta_0)$ and $S(\theta_0)$ have full rank. Then the asymptotic distribution of the optimal prediction-based estimator, $\hat{\theta}_n^*$, is*

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

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

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

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

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

Proof: The result for $\hat{\theta}_n^*$ follows since

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

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 (1997)) 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, so that 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,

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

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

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

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

$$S(\theta_0)\bar{M}(\theta_0)^{-1}S(\theta_0)^T - [W(\theta_0)^{-1} + W(\theta_0)^{-1}B(\theta_0)W(\theta_0)^{-1}]^{-1} = \tag{4.3}$$

$$[B(\theta_0)^{-1} + W(\theta_0)^{-1}]^{-1} - S(\theta_0)M^{(1)}(\theta_0)^{-1}[M^{(1)}(\theta_0)^{-1} + M^{(2)}(\theta_0)^{-1}]^{-1}M^{(1)}(\theta_0)^{-1}S(\theta_0)^T$$

$$= \left[\left(\tilde{A}(\theta_0)M^{(1)}(\theta_0)\tilde{A}(\theta_0)^T\right)^{-1} + \left(\tilde{A}(\theta_0)M^{(2)}(\theta_0)\tilde{A}(\theta_0)^T\right)^{-1}\right]^{-1} - \tilde{A}(\theta_0)\left[M^{(1)}(\theta_0)^{-1} + M^{(2)}(\theta_0)^{-1}\right]^{-1}\tilde{A}(\theta_0)^T.$$

This is an expression of how much the optimal prediction-based estimator is better than the pseudo-likelihood estimator.

It is considerably easier to calculate the pseudo-likelihood function (2.4) than the optimal estimating function (3.8) 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 (2.4) than (3.8). Therefore it is in practice preferable to use the maximum pseudo-likelihood estimator. The formula (4.3) can then be used to asses whether the loss of efficiency relative to the optimal estimator is acceptable.

5 Discretization of the continuous-time likelihood function.

In this section we will discuss simpler estimators of the parameters for the model given by (1.2). When the process is observed continuously in the time-interval $[0, t]$, the likelihood function is (see Gushchin & Küchler (2003))

$$L_t^c(\theta) = \exp\left(\theta^T A_t^c - \tfrac{1}{2}\,\theta^T I_t^c \theta\right)$$

with $\theta = (\alpha_1, \ldots, \alpha_N)$,

$$A_t^c = \left(\int_0^t X(s - r_1)dX(s), \ldots, \int_0^t X(s - r_N)dX(s)\right)^T, \tag{5.1}$$

and where I_t^c is the Fisher information matrix

$$I_t^c = \left\{\int_0^t X(s - r_i)X(s - r_j)ds\right\}. \tag{5.2}$$

This is an exponential family of stochastic processes in the sense of Küchler & Sørensen (1997). The maximum likelihood estimator of θ is

$$(I_t^c)^{-1} A_t^c. \tag{5.3}$$

The Fisher information matrix is almost surely invertible, see Reiß (2002b) (p. 75).

When the data are discrete time observations $X(\Delta), X(2\Delta), \ldots, X(n\Delta)$, a simple estimator of θ is obtained by discretizing the integrals in A_t^c and I_t^c, i.e.

$$\check{\theta}_n = I_n^{-1} A_n, \tag{5.4}$$

where

$$A_n = \left(\sum_{m=m_0}^{n-1} X(m\Delta - r_1)\delta X(m\Delta), \ldots, \sum_{m=m_0}^{n-1} X(m\Delta - r_N)\delta X(m\Delta)\right)^T \tag{5.5}$$

and

$$I_n = \left\{\Delta \sum_{m=m_0}^{n-1} X(m\Delta - r_i)X(m\Delta - r_j)\right\} \tag{5.6}$$

with $\delta X(m\Delta) = X((m+1)\Delta) - X(m\Delta)$ and $m_0 = [\max\{r_1, \ldots, r_N\}/\Delta] + 1$. Here $[x]$ denotes the integer part of a real number x. We refer to the estimator $\check{\theta}_n$ as the discretization estimator. As earlier θ_0 denotes the true value of θ.

Theorem 5.1 *The discretization estimator, $\check{\theta}$, tends $P_{\theta_0}-$almost surely to the limit*

$$\theta_0 + \Delta^{-1}\mathcal{I}(\theta_0)^{-1}\mathcal{R}(\theta_0)\theta_0$$

as $n \to \infty$, where $\mathcal{I}(\theta)$ and $\mathcal{R}(\theta)$ are the $N \times N$-matrices with entries

$$\mathcal{I}(\theta)_{ij} = K_\theta(|r_i - r_j|).$$

and

$$\mathcal{R}(\theta)_{ij} = \int_0^\Delta [K_\theta(|s + r_i - r_j|) - K_\theta(|r_i - r_j|)]ds.$$

Proof: By (1.2)

$$A_n = I_n\theta_0 + R_n\theta_0 + Z_n \tag{5.7}$$

where

$$R_n = \left\{ \sum_{m=m_0}^{n-1} X(m\Delta - r_i) \int_{m\Delta}^{(m+1)\Delta} [X(t - r_j) - X(m\Delta - r_j)]dt \right\}.$$

and

$$Z_n = \sigma \left(\sum_{m=m_0}^{n-1} X(m\Delta - r_1)\delta W(m\Delta), \ldots, \sum_{m=m_0}^{n-1} X(m\Delta - r_N)\delta W(m\Delta) \right)^T,$$

with $\delta W(m\Delta) = W((m+1)\Delta) - W(m\Delta)$. By the ergodic theorem (X is strongly mixing)

$$n^{-1}I_n \to \Delta\,\mathcal{I}(\theta_0), \quad n^{-1}R_n \to \mathcal{R}(\theta_0) \text{ and } n^{-1}Z_n \to 0,$$

$P_{\theta_0}-$almost surely. Hence

$$\check{\theta}_n = I_n^{-1}A_n = \theta_0 + I_n^{-1}R_n\theta_0 + I_n^{-1}Z_n \to \theta_0 + \Delta^{-1}\mathcal{I}(\theta_0)^{-1}\mathcal{R}(\theta_0)\theta_0,$$

□

Remark: The asymptotic bias of the discretization estimator $\check{\theta}_n$ is

$$\Delta^{-1}\mathcal{I}(\theta_0)^{-1}\mathcal{R}(\theta_0)\theta_0 = \tfrac{1}{2}\Delta\mathcal{I}(\theta_0)^{-1}\mathcal{I}'(\theta_0)\theta_0 + O(\Delta^2),$$

where

$$\mathcal{I}'(\theta) = \{K'_\theta(|r_i - r_j|)\}.$$

with $K'_\theta(t) = \partial_t K_\theta(t)$. Obviously, in an asymptotic scenario where Δ goes to zero as $n \to \infty$, the estimator $\check{\theta}_n$ is asymptotically unbiased. As one would expect, the estimator $\check{\theta}_n$ works best when Δ is small, whereas the bias can be very considerable when Δ is large.

□

Example 5.2 We simplify the model (2.5) by considering the cases $a = 0$ and $b = 0$ separately. Thus the parameter is a scalar, and $\mathcal{I}(\theta) = K_\theta(0)$ is the variance of the stationary process.

$b_0 = 0$: In this case the process is the Ornstein-Uhlenbeck process with true drift parameter $a_0 < 0$, and from (2.7) we obtain the well-known results

$$K_\theta(t) = -\frac{\sigma^2}{2a}e^{at}, \quad K'_\theta(t) = -\frac{\sigma^2}{2}e^{at}, \quad t \geq 0,$$

so the asymptotic bias of the discretization estimator is

$$\tfrac{1}{2}\Delta a_0^2 + O(\Delta^2),$$

which is a well-known result for the Ornstein-Uhlenbeck process.

$a_0 = 0$: This model is stationarity when $-\pi/2 < rb_0 < 0$, and the covariance function $K_\theta(t)$ is given by (2.10) and

$$K'_\theta(t) = \frac{\sigma^2}{2}\left(\frac{1 - \sin(br)}{\cos(br)} \sin(bt) - \cos(bt) \right) \quad t \in [0, r].$$

Hence

$$K_\theta(0) = -\frac{\sigma^2(1-\sin(br))}{2b\cos(br)}, \quad K'_\theta(0) = -\frac{\sigma^2}{2},$$

and the asymptotic bias of the discretization estimator is given by

$$\tfrac{1}{2}\Delta b_0^2 \frac{\cos(b_0 r)}{1-\sin(b_0 r)} + O(\Delta^2).$$

The case $a \cdot b \neq 0$ can, in principle, be treated analogously, but the expressions for the matrices $\mathcal{I}(\theta)$ and $\mathcal{I}'(\theta)$ are complicated, and there is no point in giving the explicit formula.

□

Finally we consider the asymptotic distribution of the discretization estimator, $\check{\theta}_n$.

Theorem 5.3 *The distribution of*

$$\sqrt{n}\left(\check{\theta}_n - \theta_0 - \Delta^{-1}\mathcal{I}(\theta_0)^{-1}\mathcal{R}(\theta_0)\theta_0\right)$$

tends to a centered regular N-dimensional normal distribution.

Proof: By (5.7)

$$\begin{aligned}
&\sqrt{n}\left(\hat{\theta}_n - \theta_0 - \Delta^{-1}\mathcal{I}(\theta_0)^{-1}\mathcal{R}(\theta_0)\theta_0\right) \\
&= \sqrt{n}\left((I_n^{-1}R_n - \Delta^{-1}\mathcal{I}(\theta_0)^{-1}\mathcal{R}(\theta_0))\theta_0 + I_n^{-1}Z_n\right) \\
&= (I_n/n)^{-1}\frac{R_n - n\mathcal{R}(\theta_0)}{\sqrt{n}}\theta_0 + \sqrt{n}\left[(I_n/n)^{-1} - \Delta^{-1}\mathcal{I}(\theta_0)^{-1}\right]\mathcal{R}(\theta_0)\theta_0 + (I_n/n)^{-1}\frac{Z_n}{\sqrt{n}} \\
&= (I_n/n)^{-1}\frac{R_n - n\mathcal{R}(\theta_0)}{\sqrt{n}}\theta_0 - \Delta^{-1}\mathcal{I}(\theta_0)^{-1}\frac{I_n - n\Delta\mathcal{I}(\theta_0)}{\sqrt{n}}(I_n/n)^{-1}\mathcal{R}(\theta_0)\theta_0 + (I_n/n)^{-1}\frac{Z_n}{\sqrt{n}},
\end{aligned}$$

where we have used that $B^{-1} = A^{-1} - A^{-1}(B-A)B^{-1}$ for two matrices A and B. This random variable has the same asymptotic distribution as

$$\Delta^{-1}\mathcal{I}(\theta_0)^{-1}\left[\frac{R_n - n\mathcal{R}(\theta_0)}{\sqrt{n}}\theta_0 - \frac{I_n - n\Delta\mathcal{I}(\theta_0)}{\sqrt{n}}\Delta^{-1}\mathcal{I}(\theta_0)^{-1}\mathcal{R}(\theta_0)\theta_0 + \frac{Z_n}{\sqrt{n}}\right].$$

Since the process X is exponentially β-mixing (Reiß (2002b)), $R_n - n\mathcal{R}(\theta_0)$, $I_n - n\Delta\mathcal{I}(\theta_0)$, and Z_n are sums of centered exponentially β-mixing sequences. All moments are finite, so it follows from the central limit theorem for mixing sequences (see e.g. Theorem 1 in Section 1.5 of Doukhan (1994)) that the estimator $\hat{\theta}_n$ is asymptotically normal. The process Z_n is a martingale, so the asymptotic covariance matrix of the term $Z_n/\sqrt{n}$ is proportional to the covariance matrix of $(X(m_0\Delta - r_1), \ldots, X(m_0\Delta - r_N))$, which is regular ($r_1, \ldots, r_N$ are assumed to be different). The other two terms in the expression above are linearly independent of Z_n, so the regularity of the limiting covariance matrix of $\check{\theta}_n$ follows from the regularity of asymptotic covariance matrix of the term $Z_n/\sqrt{n}$.

□

The expectation of the asymptotic distribution differs from the true parameter value θ_0 by the bias found previously. The expression for the asymptotic covariance matrix is extremely complicated and is best determined by simulation. Methods for simulating solutions of SDDEs can be found in the references mentioned in the introduction.

6 Simulation study

In this section we shall in a simulation study investigate 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 the parameters a and b. The delay time r is chosen equal to one, and σ^2 is fixed at one. This is not intended as a complete simulation study, rather the intention is to illustrate some important properties of the estimator. The simulations 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 θ. The simulations have been done for three values of θ: $\theta = (-1, 0.95)$ near the upper boundary of the domain of stationarity, $\theta = (-1, -1/e^2) = (-1, -0.1353)$ 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 four sampling frequencies were considered with the same number of observation time points, 200. Specifically, the observation time points were $i\Delta$, $i = 1, \ldots, 200$ with $\Delta = 0.05, 0.1, 0.5, 1$. The simulations of the SDDE were made 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 tables below report the mean values, standard deviations and empirical correlations of the simulated estimates of a and b. The dependence of the standard deviations on the time between observations and the depth of the pseudo-likelihood function is also summarized in the plots below.

The following observations from the simulations seem remarkable.

- For a fixed number of observation time-points, the bias and standard deviation of the estimators get worse 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 it depends on the parameter value whether the bias and variance increases or decreases with Δ.

- The smaller Δ is, 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 goes down as Δ increases. Here there is, however, a clearer tendency that the estimators 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 for 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 effect of the lagged and the non-lagged term in the drift, so 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 being crescent shaped.

	k						
Δ	1	3	5	7	9	13	20
0.05	-3.07	-1.39	-1.11	-1.07	-1.05	-1.03	-1.02
	4.96	1.85	0.30	0.22	0.19	0.16	0.15
0.1	-1.94	-1.10	-1.04	-1.03	-1.03	-1.02	-1.02
	2.95	0.29	0.16	0.16	0.15	0.15	0.14
0.5	-1.04	-1.01	-1.01	-1.02	-1.01	-1.12	-1.01
	0.18	0.12	0.12	0.12	0.11	0.11	0.12
1.0	-1.02	-1.02	-1.02	-1.02	-1.02	-1.02	-1.02
	0.12	0.11	0.12	0.11	0.11	0.12	0.11
2.0	-1.04	-1.02	-1.02	-1.02	-1.02	-1.02	-1.02
	0.25	0.13	0.15	0.13	0.13	0.13	0.13

Table 6.1: Mean and standard deviation of the pseudo-likelihood estimator of a for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = -0.1353$.

	k						
Δ	1	3	5	7	9	13	20
0.05	1.60	0.17	-0.08	- 0.11	-0.12	-0.13	0.14
	5.41	2.06	0.55	0.39	0.32	0.23	0.15
0.1	0.62	-0.09	-0.15	-0.13	-0.13	-0.14	-0.14
	3.28	0.50	0.29	0.18	0.17	0.15	0.14
0.5	-0.12	-0.14	- 0.14	-0.13	-0.13	-0.14	-0.13
	0.34	0.12	0.13	0.13	0.12	0.13	0.13
1.0	-0.13	-0.14	-0.13	-0.13	-0.14	- 0.13	-0.13
	0.16	0.14	0.14	0.14	0.13	0.14	0.14
2.0	-0.17	-0.14	-0.15	-0.14	-0.14	-0.13	-0.14
	0.38	0.18	0.20	0.17	0.18	0.16	0.16

Table 6.2: Mean and standard deviation of the pseudo-likelihood estimator of b for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = -0.1353$.

	k						
Δ	1	3	5	7	9	13	20
0.05	-0.96	-0.96	-0.72	-0.63	-0.54	-0.39	-0.28
0.1	-0.96	-0.73	-0.47	-0.46	-0.38	-0.32	-0.29
0.5	-0.70	-0.38	-0.39	-0.35	-0.36	-0.32	-0.40
1.0	-0.55	-0.44	-0.51	-0.52	-0.44	-0.47	-0.45
2.0	0.77	-0.27	0.03	-0.25	-0.11	-0.51	-0.37

Table 6.3: Empirical correlation between the the pseudo-likelihood estimators of a and b for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = -0.1353$.

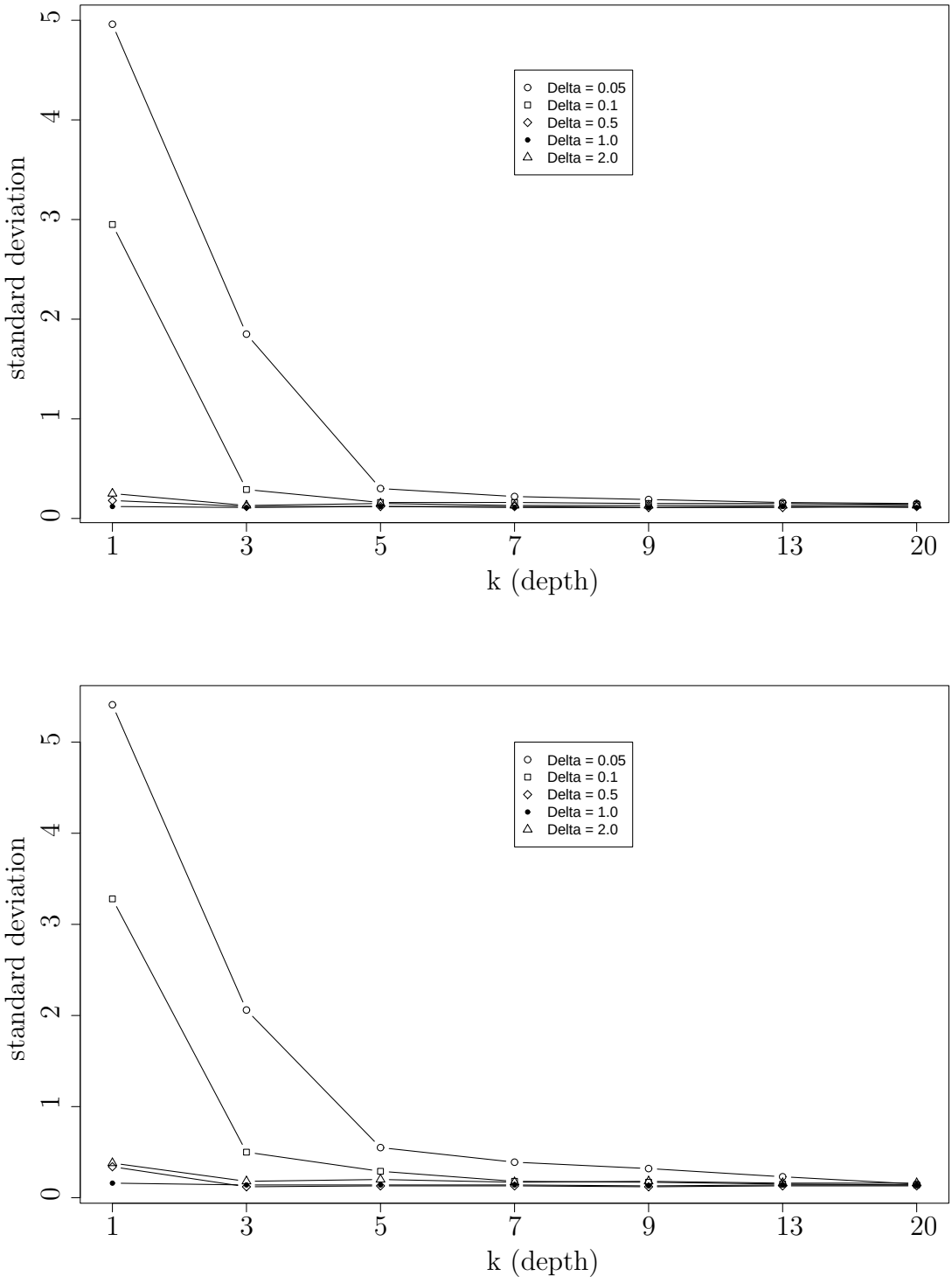

Figure 6.1: Standard deviation of the pseudo-likelihood estimator of a (upper) and b (lower) for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = -0.1353$.

	k						
Δ	1	3	5	7	9	13	20
0.05	-1.94	-1.54	-1.32	-1.14	-1.10	-1.03	-1.03
	2.88	1.87	1.24	0.59	0.47	0.25	0.15
0.1	-1.90	-1.23	-1.09	-1.04	-1.03	-1.03	-1.03
	2.53	0.98	0.40	0.22	0.16	0.14	0.14
0.5	-1.09	-1.02	-1.02	-1.03	-1.02	-1.02	-1.02
	0.58	0.11	0.14	0.16	0.14	0.16	0.15
1.0	-1.02	-1.03	-1.09	-1.09	-1.08	-1.07	-1.09
	0.18	0.17	0.26	0.25	0.24	0.24	0.25
2.0	-1.02	-1.01	-1.00	-1.01	-1.01	-1.01	-1.01
	0.15	0.13	0.13	0.14	0.13	0.13	0.13

Table 6.4: Mean and standard deviation of the pseudo-likelihood estimator of a for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = 0.95$.

	k						
Δ	1	3	5	7	9	13	20
0.05	1.89	1.46	1.23	1.06	1.02	0.95	0.95
	2.91	1.89	1.26	0.93	0.49	0.26	0.14
0.1	1.81	1.15	1.01	0.97	0.95	0.96	0.95
	2.55	1.00	0.41	0.22	0.16	0.13	0.14
0.5	1.01	0.95	0.96	0.96	0.96	0.96	0.96
	0.60	0.14	0.15	0.16	0.15	0.16	0.15
1.0	0.95	0.97	1.04	1.03	1.02	1.02	1.03
	0.18	0.18	0.27	0.27	0.25	0.25	0.27
2.0	0.96	0.95	0.95	0.96	0.95	0.95	0.95
	0.15	0.13	0.13	0.14	0.14	0.14	0.14

Table 6.5: Mean and standard deviation of the pseudo-likelihood estimator of b for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = 0.95$.

	k						
Δ	1	3	5	7	9	13	20
0.05	-0.999	-0.999	-0.987	-0.992	-0.988	-0.964	-0.891
0.1	-0.999	-0.997	-0.988	-0.966	-0.904	-0.907	-0.885
0.5	-0.996	-0.942	-0.954	-0.956	-0.949	-0.961	-0.957
1.0	-0.984	-0.982	-0.993	-0.993	-0.991	-0.991	-0.990
2.0	-0.987	-0.983	-0.982	-0.984	-0.984	-0.982	-0.982

Table 6.6: Empirical correlation between the the pseudo-likelihood estimators of a and b for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = 0.95$.

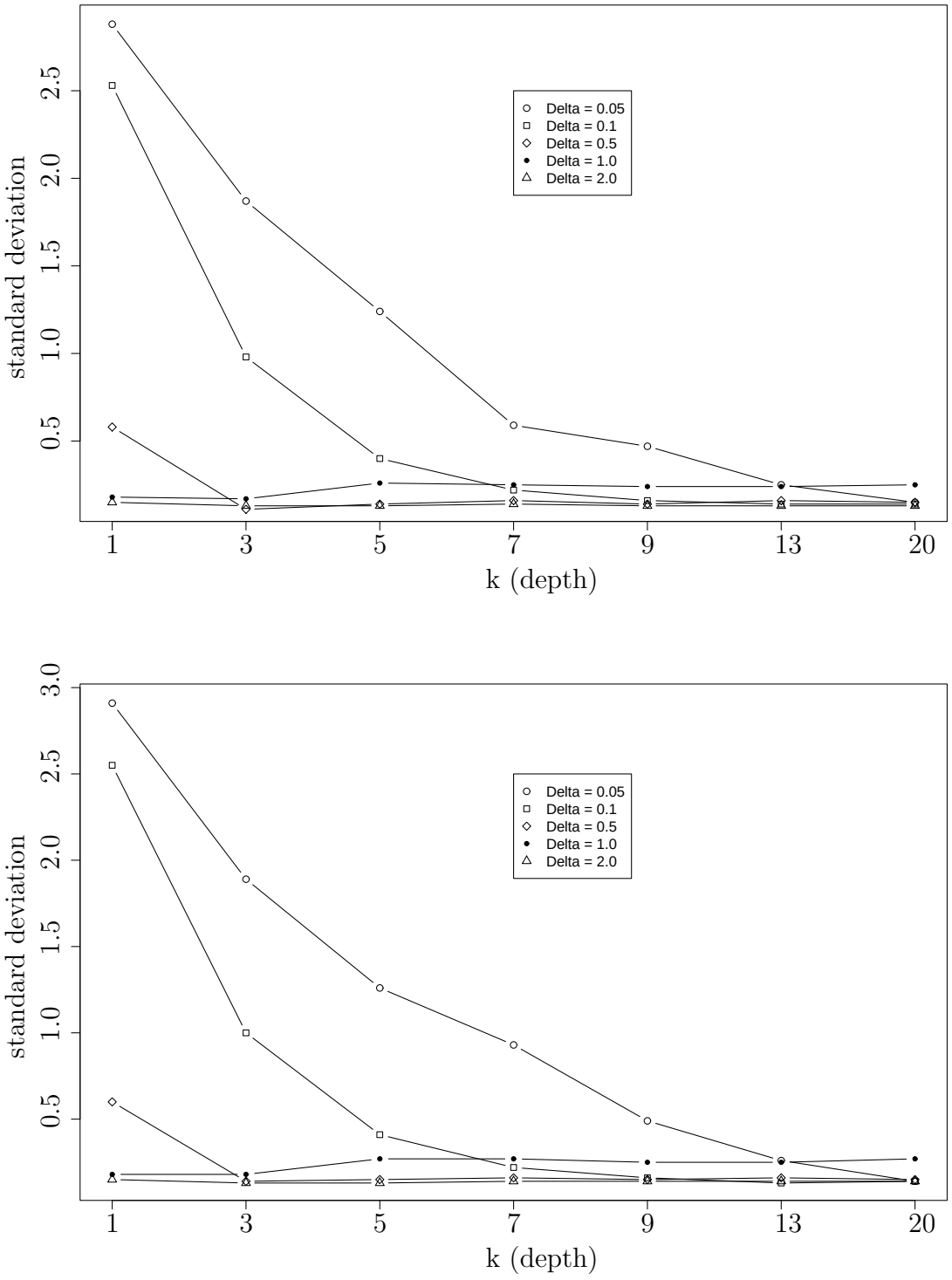

Figure 6.2: Standard deviation of the pseudo-likelihood estimator of a (upper) and b (lower) for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = 0.95$.

	k						
Δ	1	3	5	7	9	13	20
0.05	-1.02	-1.01	-1.01	-1.00	-1.01	-1.01	-1.01
	0.26	0.12	0.09	0.08	0.08	0.08	0.07
0.1	-1.02	-1.01	-1.01	-1.01	-1.01	-1.01	-1.01
	0.19	0.09	0.08	0.07	0.07	0.07	0.07
0.5	-1.00	-1.00	-1.00	-1.01	-1.00	-1.01	-1.00
	0.06	0.06	0.05	0.05	0.05	0.05	0.05
1.0	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
	0.05	0.04	0.05	0.05	0.04	0.05	0.04
2.0	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
	0.08	0.06	0.04	0.04	0.04	0.04	0.04

Table 6.7: Mean and standard deviation of the pseudo-likelihood estimator of a for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = -2.1$.

	k						
Δ	1	3	5	7	9	13	20
0.05	-2.09	-2.08	-2.08	-2.08	-2.08	-2.08	-2.09
	0.22	0.11	0.09	0.09	0.09	0.08	0.07
0.1	-2.10	-2.09	-2.08	-2.09	-2.08	-2.09	-2.09
	0.16	0.09	0.08	0.08	0.07	0.07	0.07
0.5	-2.09	-2.09	-2.09	-2.09	-2.09	-2.09	-2.09
	0.06	0.05	0.05	0.05	0.05	0.05	0.05
1.0	-2.10	-2.10	-2.09	-2.09	-2.09	-2.09	-2.09
	0.04	0.04	0.05	0.05	0.04	0.05	0.05
2.0	-2.09	-2.09	-2.10	-2.10	-2.10	-2.10	-2.10
	0.08	0.05	0.04	0.04	0.04	0.04	0.04

Table 6.8: Mean and standard deviation of the pseudo-likelihood estimator of b for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = -2.1$.

	k						
Δ	1	3	5	7	9	13	20
0.05	0.92	0.69	0.53	0.49	0.49	0.39	0.32
0.1	0.88	0.57	0.39	0.44	0.37	0.25	0.24
0.5	0.51	0.41	0.34	0.33	0.40	0.30	0.40
1.0	0.44	0.39	0.43	0.40	0.44	0.40	0.40
2.0	0.90	0.72	0.51	0.50	0.51	0.51	0.46

Table 6.9: Empirical correlation between the the pseudo-likelihood estimators of a and b for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = -2.1$.

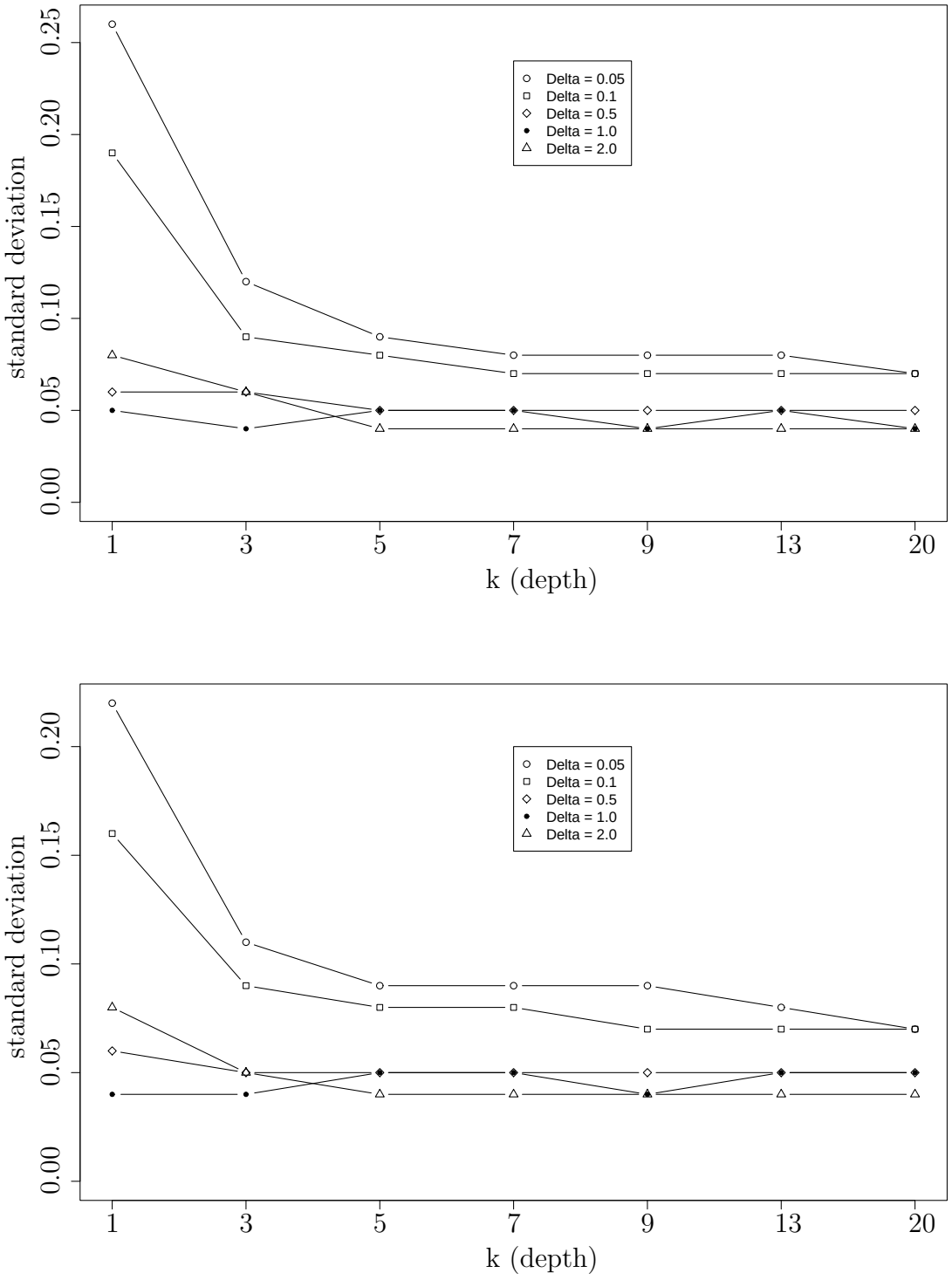

Figure 6.3: Standard deviation of the pseudo-likelihood estimator of a (upper) and b (lower) for various values of the the depth k and the time between observations Δ. The number of observations is 200, and the true parameter values are $a = -1$ and $b = -2.1$.

		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

Table 6.10: Mean and standard deviation of the pseudo-likelihood estimator of a for various values of the the depth k, the time between observations Δ, and the 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	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

Table 6.11: Mean and standard deviation of the pseudo-likelihood estimator of b for various values of the the depth k, the time between observations Δ, and the 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	-0.96	-0.76	-0.72	-0.55	-0.55	-0.43	-0.27
0.1	2000	-0.91	-0.64	-0.50	-0.39	-0.33	-0.32	-0.33
0.5	400	-0.69	-0.32	-0.39	-0.37	-0.35	-0.34	-0.31
1.0	200	-0.56	-0.48	-0.47	-0.44	-0.44	-0.50	-0.51
2.0	100	0.86	0.18	0.10	-0.11	0.18	-0.10	-0.23

Table 6.12: Empirical correlation between the the pseudo-likelihood estimators of a and b for various values of the the depth k and the time between observations Δ, and the 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 $b = -0.1353$.

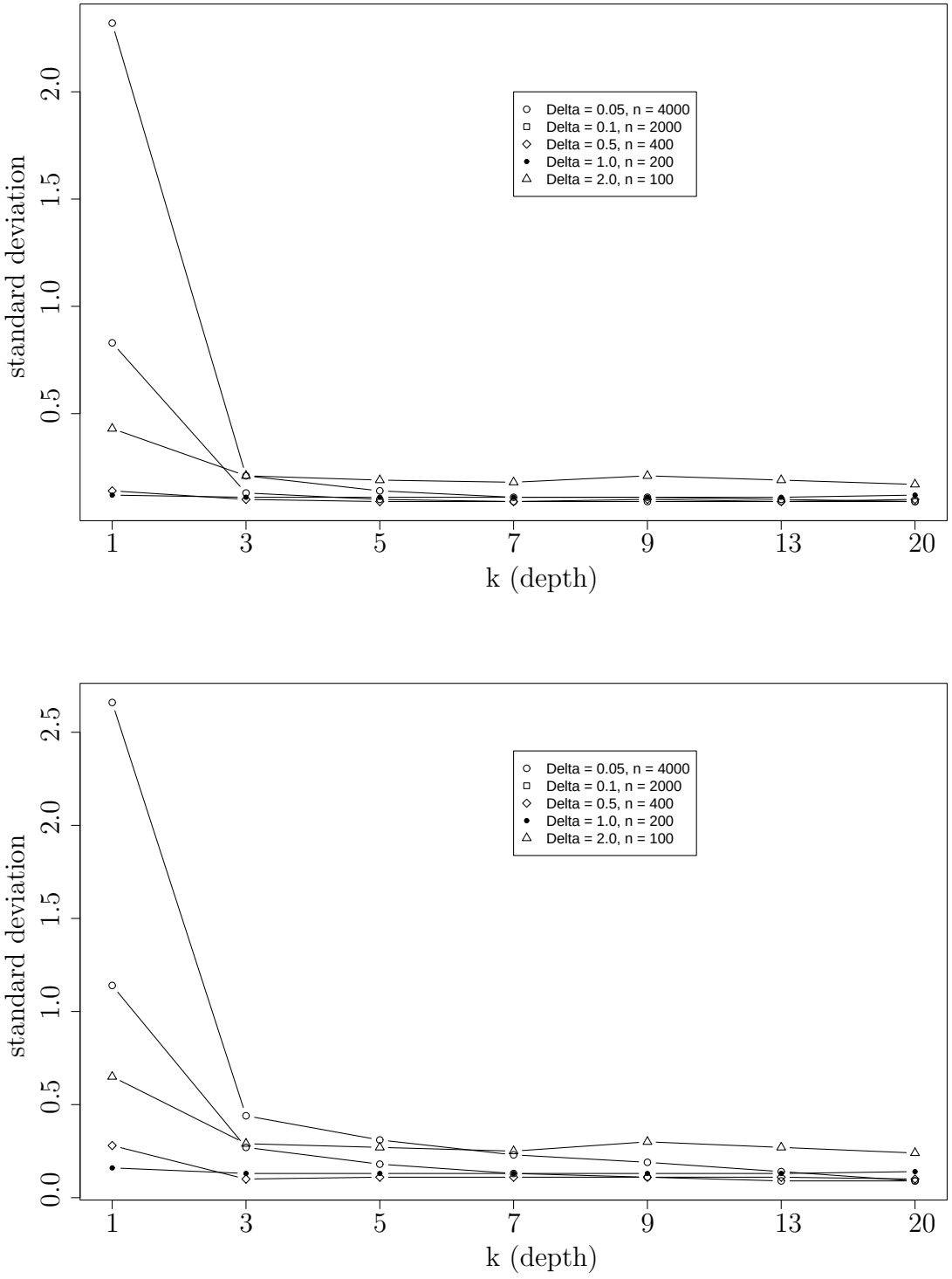

Figure 6.4: Standard deviation of the pseudo-likelihood estimator of a (upper) and b (lower) for various values of the the depth k, the time between observations Δ, and the number of observations n. In all cases $n\Delta$, the length of the observation interval, is 200, and the true parameter values are $a = -1$ and $b = -0.1353$.

Acknowledgement

The authors are grateful to stud. math. Katja Krol for writing the PC-program used in the simulation study and to Daniel Skodlerack for an earlier version of this program.

References

Brockwell, P. J. & Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer-Verlag, New York.

Buckwar, E. (2000). "Introduction to the numerical analysis of stochastic delay differential equations". *Journal of Computational and Applied Mathematics*, 125:297–307.

Diekmann, O.; van Gils, S. A.; Lunel, S. M. V. & Walther, H.-O. (1995). *Delay Equations: Functional-, Complex-, and Nonlinear Analysis*. Springer-Verlag, New York.

Ditlevsen, S. & Sørensen, M. (2004). "Inference for observations of integrated diffusion processes.". *Scand. J. Statist.*, 31:417–429.

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

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

Gushchin, A. A. & Küchler, U. (1999). "Asymptotic properties of maximum-likelihood-estimators for a class of linear stochastic differential equations with time delay". *Bernoulli*, 5:1059–1098.

Gushchin, A. A. & Küchler, U. (2000). "On stationary solutions of delay differential equations driven by a Lévy process". *Stochastic Proc. Appl.*, 88:195–211.

Gushchin, A. A. & Küchler, U. (2003). "On parametric statistical models for stationary solutions of affine stochastic delay differential equations". *Mathematical Methods in Statistics*, 12:31–61.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application*. Springer-Verlag, New York.

Küchler, U. & Kutoyants, Y. (2000). "Delay estimation for some stationary diffusion-type processes". *Scand. J. Statist.*, 27:405–414.

Küchler, U. & Mensch, B. (1992). "Langevin's stochastic differential equation extended by a time-delayed term". *Stochastics and Stochastics Reports*, 40:23–42.

Küchler, U. & Platen, E. (2000). "Strong discrete time approximation of stochastic differential equations with time delay". *Mathematics & Computer Simulation*, 54:189–205.

Küchler, U. & Platen, E. (2007). "Time delay and noise explaining cyclical fluctuations in prices of commodities". Preprint, Inst. of Mathematics, Humboldt-University of Berlin.

Küchler, U. & Sørensen, M. (1997). *Exponential Families of Stochastic Processes*. Springer, New York.

Küchler, U. & Vasil'jev, V. A. (2005). "Sequential identification of linear dynamic systems with memory". *Statistical Inference for Stochastic Processes*, 8:1–24.

Reiß, M. (2002a). "Minimax rates for nonparametric drift estimation in affine stochastic delay differential equations". *Statistical Inference for Stochastic Processes*, 5:131–152.

Reiß, M. (2002b). *Nonparametric Estimation for Stochastic Delay Differential Equations*. PhD thesis, Institut für Mathematik, Humboldt-Universität zu Berlin.

Sørensen, H. (2003). "Simulated Likelihood Approximations for Stochastic Volatility Models". *Scand. J. Statist.*, 30:257–276.

Sørensen, M. (1999). "On asymptotics of estimating functions". *Brazilian Journal of Probability and Statistics*, 13:111–136.

Sørensen, M. (2000). "Prediction-based estimating functions". *Econometrics Journal*, 3:123–147.