Ph.D. thesis

Parametric Statistical Inference for Discretely Observed Diffusion Processes: A Review

Part I: Theoretical Results

Part II: Statistical Applications of Gaussian Diffusion Processes in Freshwater Ecology

Asger Roer Pedersen

Department of Theoretical Statistics
Institute of Mathematics
University of Aarhus
DK–8000 Aarhus C, Denmark

National Environmental Research Institute
Department of Freshwater Ecology
Vejlsøvej 25, DK–8600 Silkeborg, Denmark

Part I: Theoretical Results

1 Introduction

Diffusion processes defined by stochastic differential equations are important in many sciences from a modeling point of view. In some applications, e.g. in mathematical finance, the stochastic differential equation emerges directly from theoretical modeling considerations, and in other cases it is derived as a stochastic analogue of a deterministic model given by an ordinary differential equation. The latter type of applications is particularly important in many natural sciences where ordinary differential equations are widely used as theoretical deterministic models for physical phenomena, see e.g. Erlandsen & Thyssen (1983), Guttorp & Kulperger (1984), Thyssen et al (1990), Madsen & Holst (1991), Madsen & Melgaard (1991), Pedersen (1994c) and Jensen et al (1994). In stochastic modeling of continuous-time phenomena it is very natural and desirable to apply continuous-time stochastic processes such as diffusion processes, but the statistical analysis of a diffusion process must of course be based on discrete-time observations of the process. For some special diffusion processes this can be done by means of classical likelihood methods, see e.g. Pedersen (1993b,1994c) for a treatment of Gaussian diffusion processes, but for the vast majority of diffusion processes this is impossible since the transition densities are usually unknown, see section 2.

In this paper we consider general methods for the statistical analysis of discretely observed diffusion processes. Various methods for the estimation of unknown parameters in stochastic differential equations are presented, and corresponding asymptotic results to be used for hypothesis testing and for the calculation of confidence regions for the unknown parameters are also discussed. Finally some general tools for the validation of parametric statistical models given by discretely observed diffusion processes are presented. Section 2 contains a brief historical sketch of the development of statistical methods for the analysis of discretely observed diffusion processes. In section 3 we consider the approximate likelihood inference approach for discretely observed diffusion processes that was developed in Pedersen (1993a,c). The quasi-likelihood methods for discretely observed diffusion processes studied in Pedersen (1994b) are reviewed in section 4, and the general methods proposed in Pedersen (1994a) for the validation of parametric statistical models given by discretely observed diffusion processes are reviewed in section 5.

2 Historical sketch

Statistical inference about the parameter $\theta \in \Theta \subseteq \boldsymbol{R}^p$ in the stochastic differential equation

$$dX_t = b(t, X_t; \theta)dt + \sigma(t, X_t; \theta)dW_t,\ X_0 = x_0,\ t \geq 0, \tag{1}$$

where W is an r-dimensional Wiener process, $b : [0, \infty) \times \boldsymbol{R}^d \mapsto \boldsymbol{R}^d$ and $\sigma : [0, \infty) \times \boldsymbol{R}^d \mapsto M^{d\times r}$ (the set of $d \times r$ matrices), based on discrete observations of X at time-points $0 = t_0 < t_1 < \cdots < t_n$ should ideally be based on the log-likelihood function for θ

$$\ell_n(\theta) = \log p(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i}; \theta),$$

since the corresponding maximum likelihood estimator $\hat{\theta}_n$ for θ is known in many cases to have the usual good properties, see Billingsley (1961) and Dacunha-Castelle & Florens-Zmirou (1986). Here $p(s, x, t, y; \theta)$ denote the transition densities of X. Classical likelihood inference about θ based on $\ell_n(\theta)$ can for instance be performed for Gaussian diffusion processes, see Pedersen (1993b,1994c), but in general this is impossible since the transition densities of X are usually unknown.

When the transition densities of X are unknown the first approach was to perform the inference about θ by means of a discretization of the likelihood function for θ based on continuous observation of X. To understand the developments and drawbacks of this approach it is instructive first to review the theory for continuous observation of X, see also Basawa & Prakasa Rao (1980), Kutoyants (1984) and Prakasa Rao (1985). It is characteristic for the statistical analysis of continuous observation of X that maximum likelihood estimation of unknown parameters in σ is impossible, since the distributions of a continuous path $(X_t)_{t\in[0,T]}$ for different values of θ are non-equivalent when σ depends on θ. This means that σ must be considered known, i.e. $\sigma(t, x; \theta) = \sigma(t, x)$. However, in some cases σ can be calculated in advance by means of the quadratic variation of X, see Le Breton (1974), Brown & Hewitt (1975) and Le Breton & Musiela (1984). This is for instance the case when σ is a constant matrix, in which case

$$a = \sigma\sigma^T = \frac{1}{t}[X]_t = \lim_{k\longrightarrow\infty} \frac{1}{t} \sum_{i=1}^{k} (X_{i\frac{t}{k}} - X_{(i-1)\frac{t}{k}})(X_{i\frac{t}{k}} - X_{(i-1)\frac{t}{k}})^T.$$

Here $[X]$ denotes the quadratic variation of X, and ^{T} denotes matrix transposition. Assuming that σ is a known function, the log-likelihood function for θ based on a continuous observation of X in the time-interval $[0, T]$ is under certain conditions, see Liptser & Shiryayev (1977), given by

$$\ell_T^c(\theta) = \int_0^T b(t, X_t; \theta)^T a(t, X_t)^{-1} dX_t - \frac{1}{2} \int_0^T b(t, X_t; \theta)^T a(t, X_t)^{-1} b(t, X_t; \theta)dt.$$

The corresponding maximum likelihood estimator for θ is denoted by $\hat{\theta}_T^c$. Liptser & Shiryayev (1977) calculated in a special case the bias and mean square error of $\hat{\theta}_T^c$, but in general there are no exact results about $\hat{\theta}_T^c$ for fixed T. Kutoyants (1977,1984) studied the asymptotic behaviour of $\hat{\theta}_T^c$ on a fixed time-interval $[0, T]$ for one-dimensional diffusion processes as $\sigma(t, x; \theta) \equiv \sigma$ (a scalar) tends to zero, and proved in this sense the consistency and asymptotic normality of $\hat{\theta}_T^c$. The majority of asymptotic studies of $\hat{\theta}_T^c$ has however been for T tending to infinity and for a fixed known diffusion coefficient $\sigma(t, x)$. In these studies, $\hat{\theta}_T^c$ is in particular shown to be consistent and asymptotically normally distributed as T tends to infinity. The case where b depends linearly on θ was studied by Le Breton (1974), Taraskin (1974), Brown & Hewitt (1975) and Le Breton & Musiela (1984), whereas Lanska (1979) considered a general dependency of b on θ. The methods used by Lanska (1979) are however essentially only applicable for one-dimensional diffusion processes, since they rely on the fact that $\ell_T^c(\theta)$ can be written entirely in terms of stochastic Lebesgue integrals for one-dimensional diffusion processes. In these papers the asymptotic distribution of $\hat{\theta}_T^c$ is obtained via a non-random normalization proportional to $\sqrt{T}$. Barndorff-Nielsen & Sørensen (1991,1994) discuss in a general context several alternative non-random and random measures of information about θ derived from martingale limit theory, see also Feigin (1976). These martingale tools and results have been applied by Kloeden et al (1992) in the study of exponential families of stochastic processes of diffusion type, which in particular covers continuously observed diffusion processes with b depending linearly on θ, since the distributions of $(X_t)_{t \in [0,T]}$ for all $\theta \in \Theta$ in such cases constitute an exponential family of probability measures for every fixed T. A general treatment of so-called exponential families of stochastic processes is given in Kuchler & Sørensen. In the preceding discussion we have assumed that σ can be considered a known function $\sigma(t, x)$. This condition is not satisfied if σ depends on unknown parameters that can not be identified in advance, i.e. σ depends on θ, and in such cases it is clear from the non-equivalency of the distributions of $(X_t)_{t \in [0,T]}$ for different values of θ that θ must be estimated by some alternative method. Hutton & Nelson (1986) suggested in such cases to use a quasi-score function for θ. When σ does not depend on θ the score function for θ is under certain regularity conditions given by

$$\dot{\ell}_T^c(\theta) = \int_0^T \dot{b}(t, X_t; \theta)^T a(t, X_t)^{-1} dX_t - \int_0^T \dot{b}(t, X_t; \theta)^T a(t, X_t)^{-1} b(t, X_t; \theta) dt,$$

where a dot denotes differentiation with respect to θ. The quasi-score function for θ proposed by Hutton & Nelson (1986) when σ depends on θ is then obtained by replacing $a(t, X_t)$ in this expression by $a(t, X_t; \theta) = \sigma(t, X_t; \theta)\sigma(t, X_t; \theta)^T$. Hutton & Nelson (1986) then showed that the corresponding maximum quasi-likelihood estimator for θ enjoys the optimality properties defined by Godambe & Heyde (1987). For continuous semimartingales, e.g. diffusion processes, the approach in Hutton &

Nelson (1986) coincides with that of Sørensen (1990). Various measures of information about θ attached to such maximum quasi-likelihood estimators for θ can be found in Godambe & Heyde (1987) and Barndorff-Nielsen & Sørensen (1991,1994).

The historical development of the theory for discrete observations of X based on the theory for continuous observation of X is basically an imitation of the developments outlined above, see Prakasa Rao (1985,1988) and Yoshida (1992). The obtained results essentially state that any result that can be proven for continuous observation of X has by proper discretization an analogous version for discrete observations of X as long as the discrete observation time-points become dense in some sense when the number of discrete observation time-points tends to infinity. It may however have diastrous consequences to apply these results in practice if the discrete observation time-points are not sufficiently closely spaced. In particular, the estimators can be strongly biased. Moreover, the developments to be presented inherit some unnatural difficulties regarding the diffusion coefficient σ from the fact that σ must be considered known in likelihood inference based on continuous observation of X. We return to these points at the end of the section. Suppose for a while that σ is known, i.e. $\sigma(t,x;\theta)=\sigma(t,x)$. Then the log-likelihood function for θ based on continuous observation of X in the time-interval $[0,t_n]$ can be approximated by

$$\begin{aligned}\tilde{\ell}_n(\theta) &= \sum_{i=1}^{n} b(t_{i-1},X_{t_{i-1}};\theta)^T a(t_{i-1},X_{t_{i-1}})^{-1}(X_{t_i}-X_{t_{i-1}}) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n} b(t_{i-1},X_{t_{i-1}};\theta)^T a(t_{i-1},X_{t_{i-1}})^{-1} b(t_{i-1},X_{t_{i-1}};\theta)(t_i-t_{i-1}).\end{aligned}$$

The estimator for θ obtained by maximizing $\tilde{\ell}_n(\theta)$ is denoted by $\tilde{\theta}_n$. If the discrete observation time-points $\{t_i\}_{i=0}^n$ become dense in some fixed time-interval $[0,T]$ as n tends to infinity, it is quite generally true that $\tilde{\ell}_n(\theta)$ converges in probability to $\ell_T^c(\theta)$ for every θ as n tends to infinity. In some cases it can furthermore be proven that $\tilde{\theta}_n$ converges to $\hat{\theta}_T^c$ in probability as n tends to infinity, see Le Breton (1976). Genon-Catalot (1990) proved that the small-noise asymptotic results for $\hat{\theta}_T^c$ on a fixed time-interval $[0,T]$ proved by Kutoyants (1977,1984) also hold for $\tilde{\theta}_n$ provided $\sigma(t,x;\theta)\equiv\sigma$ (a scalar) and the distance $\Delta=T/n$ between equidistant discrete observation time-points tend to zero at a certain rate as n tends to infinity. The asymptotic results for $\hat{\theta}_T^c$ as T tends to infinity have also been translated into asymptotic results for $\tilde{\theta}_n$ as the length t_n of the observation period tends to infinity and the discrete observation time-points become dense, simultaneously, as n tends to infinity. Hence, we assume from now on that the discrete observation time-points for each n are given by

$$t_i = i\Delta_n,\ i=0,1,\ldots,n,$$

where $\{\Delta_n\}$ is a sequence of positive real numbers. The asymptotic results that have been proven for $\tilde{\theta}_n$ are then in the limit where n tends to infinity, Δ_n tends to zero and $n\Delta_n$ tends to infinity, simultaneously. The circumstances for these studies of $\tilde{\theta}_n$ are however more complicated than the corresponding studies of $\hat{\theta}_T^c$, because σ can no longer be identified in advance. For one-dimensional diffusion proceses with constant diffusion coefficient σ one may simply consider σ as being fixed, since $\tilde{\ell}_n(\theta)$ in that case is given by

$$\sigma^2\tilde{\ell}_n(\theta) = \sum_{i=1}^n b(t_{i-1}, X_{t_{i-1}};\theta)(X_{t_i} - X_{t_{i-1}}) - \frac{1}{2}\sum_{i=1}^n b(t_{i-1}, X_{t_{i-1}};\theta)^2(t_i - t_{i-1}),$$

and $\tilde{\theta}_n$ is thus independent of the fixed (but unknown) value of σ. This convenient property does however not hold for multi-dimensional diffusion processes, and several authors have assumed that $\sigma(t,x;\theta) \equiv I_d$ (the $d \times d$ identity matrix). Under this assumption, Dorogovcev (1976) proved for one-dimensional diffusion processes and parameter θ that $\tilde{\theta}_n$ is consistent as n tends to infinity, Δ_n tends to zero and $n\Delta_n$ tends to infinity. Assuming this consistency of $\tilde{\theta}_n$, Prakasa Rao (1983) furthermore proved the asymptotic normality and efficiency of $\tilde{\theta}_n$ under the additional condition that $\sqrt{n}\Delta_n$ tends to zero. This condition is usually refered to as the condition of "rapidly increasing experimental design". The results of Prakasa Rao (1983) were extended to multi-dimensional diffusion processes by Penev (1985), see also Prakasa Rao (1988), and Kasonga (1988) proved the consistency of $\tilde{\theta}_n$ for multi-dimensional diffusion processes and parameter θ, also under the assumption of rapidly increasing experimental design. Penev (1985) attacked the problem of estimating σ from discrete observations. Assuming that σ is a constant matrix, an obvious estimator for the diffusion matrix $a = \sigma\sigma^T$ is the (discretized quadratic variation) quadratic variation like estimator

$$\tilde{a}_n = \frac{1}{n\Delta_n}\sum_{i=1}^n (X_{t_i} - X_{t_{i-1}})(X_{t_i} - X_{t_{i-1}})^T.$$

Under the assumption of rapidly increasing experimental design, Penev (1985) then proved that $\tilde{a}_n$ is a consistent and asymptotically normally distributed estimator of a, see also Prakasa Rao (1988). For one-dimensional diffusion processes, Florens-Zmirou (1989) proved this under the weaker assumption that $n\Delta_n^3$ tends to zero. Having studied, marginally, the estimator $\tilde{\theta}_n$ for a fixed known diffusion coefficient and estimators of $a = \sigma\sigma^T$, the next step is then, generally speaking, to insert an estimator of a in $\tilde{\ell}_n$, and then to estimate unknown parameters in b by maximizing the obtained approximation of $\tilde{\ell}_n$. As previously noted this does not alter the estimator $\tilde{\theta}_n$ in case of one-dimensional diffusion processes with constant diffusion coefficient, see Florens-Zmirou (1989), but for multi-dimensional diffusion processes it may affect the estimation of unknown parameters in b. Yoshida (1992) considered this problem in the following general situation, where θ can be divided into a vector ψ of unknown parameters in the drift

$$b(t,x;\theta) = b(x;\psi),$$

and a $q \times r$ matrix η entering the diffusion coefficient in the following way

$$\sigma(t, x; \theta) = \sigma(x)\eta.$$

Here $\sigma(x)$ is a known $d \times q$ matrix function, e.g. $q = d$ and $\sigma(x) \equiv I_d$. For a given matrix η we may estimate ψ by maximizing

$$\begin{aligned}\tilde{\ell}_n(\psi; \eta) &= \sum_{i=1}^{n} b(X_{t_{i-1}}; \psi)^T \left[\sigma(X_{t_{i-1}})\eta\eta^T\sigma(X_{t_{i-1}})\right]^{-1} (X_{t_i} - X_{t_{i-1}}) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n} b(X_{t_{i-1}}; \psi)^T \left[\sigma(X_{t_{i-1}})\eta\eta^T\sigma(X_{t_{i-1}})\right]^{-1} b(X_{t_{i-1}}; \psi)(t_i - t_{i-1})\end{aligned}$$

with respect to ψ. Notice that $\tilde{\ell}_n(\psi; \eta)$ only depends on η through $\Gamma = \eta\eta^T$. We may therefore denote $\tilde{\ell}_n(\psi; \eta)$ by $\tilde{\ell}_n(\psi; \Gamma)$. Yoshida (1992) then suggested to estimate ψ and Γ by the following four-step procedure. First Γ is estimated by the generalized quadratic variation like estimator

$$\begin{aligned}\Gamma_n^0 &= \frac{1}{n\Delta_n}\sum_{i=1}^{n} \delta_i\delta_i^T \\ \delta_i &= \left[\sigma(X_{t_{i-1}})^T\sigma(X_{t_{i-1}})\right]^{-1} \sigma(X_{t_{i-1}})^T(X_{t_i} - X_{t_{i-1}}),\ i = 1, \ldots, n.\end{aligned}$$

Then ψ is estimated by maximizing $\tilde{\ell}_n(\psi; \Gamma_n^0)$ with respect to ψ, and the obtained estimator ψ_n^0 of ψ is then used to improve the estimator Γ_n^0. First by the "corrected sum of squares"

$$\begin{aligned}\overline{\Gamma}_n^0 &= \frac{1}{n\Delta_n}\sum_{i=1}^{n} D_i(\psi_n^0)D_i(\psi_n^0)^T \\ D_i(\psi) &= \left[\sigma(X_{t_{i-1}})^T\sigma(X_{t_{i-1}})\right]^{-1} \sigma(X_{t_{i-1}})^T \\ &\quad (X_{t_i} - X_{t_{i-1}} - (t_i - t_{i-1})b(X_{t_{i-1}}; \psi)),\ i = 1, \ldots, n,\end{aligned}$$

and then by combining ψ_n^0, Γ_n^0 and $\overline{\Gamma}_n^0$ in a complicated way that involves the estimated generator of the diffusion process, see Yoshida (1992). We denote this final estimator of Γ by Γ_n^1, and the final estimator for ψ obtained by maximizing $\tilde{\ell}_n(\psi; \Gamma_n^1)$ with respect to ψ is denoted by ψ_n^1. Yoshida (1992) then first proved that ψ_n^0 and Γ_n^0 are consistent estimators when n tends to infinity, Δ_n tends to zero and $n\Delta_n$ tends to infinity. Moreover he proved that ψ_n^1 and Γ_n^1 are consistent and jointly asymptotically normally distributed under the additional condition that $n\Delta_n^3$ tends to zero, thus avoiding the assumption of rapidly increasing experimental design. Even though the results by Yoshida (1992) at the present stage seem to be the most well-developed in this line of contributions to the theory for discrete observations of X, the proposed estimation procedure is still rather restrictive with respect to the parameter dependence of the diffusion coefficient. In particular the diffusion coefficient is required to depend linearly on unknown parameters and the drift and the diffusion coefficient must not have any unknown parameters in common. An estimation method that is less restrictive with respect to the parameter dependence of b and σ can be obtained by discretization of the quasi-score function proposed by Hutton & Nelson (1986), that is

$$\begin{aligned} H_n(\theta) &= \sum_{i=1}^{n} \dot{b}(t_{i-1}, X_{t_{i-1}};\theta)^T a(t_{i-1}, X_{t_{i-1}};\theta)^{-1}(X_{t_i} - X_{t_{i-1}}) \\ &\quad - \sum_{i=1}^{n} \dot{b}(t_{i-1}, X_{t_{i-1}};\theta)^T a(t_{i-1}, X_{t_{i-1}};\theta)^{-1} b(t_{i-1}, X_{t_{i-1}};\theta)(t_i - t_{i-1}). \end{aligned}$$

The expectation of $H_n(\theta)$ is, however, in general not zero and the estimating function $H(\theta) = \{H_n(\theta)\}_{n=1}^{\infty}$ is not a martingale, so strongly biased estimators are likely to be obtained by using this approximate quasi-score function, see Bibby & Sørensen (1994), Pedersen (1994b) and section 4. The estimation of unknown parameters in the diffusion coefficient is particularly important in mathematical finance where the diffusion coefficient (also called the volatility) is of major interest. In some diffusion models in mathematical finance, e.g. within the context of Black-Scholes option pricing, the drift is in in fact unspecified. This means that the estimation problem (estimation of the volatility) is in fact a non- or semi-parametric estimation problem. This problem has been examined by Flores-Zmirou (1993) who defined a non-parametric estimator of the diffusion coefficient that does not depend on the unspecified drift function. Genon-Catalot & Jacod (1993,1994) treated the same problem in a semi-parametric setting by assuming a parametric form of the diffusion coefficient, and introduced furthermore the concept of random sampling in this context. Both in Florens-Zmirou (1993) and Genon-Catalot & Jacod (1993,1994) the behaviour of the proposed estimator is studied in the limit where an infinite number of discrete observations is made on a fixed finite time-interval.

In addition to the difficulties regarding the diffusion coefficient a major drawback of the estimation methods reviewed in this section is that the performance of the estimators is highly sensitive with respect to the observation frequency. For any of the presented estimators to be reliable it is required that the discrete observation time-points are sufficiently closely spaced. However, simulations clearly show that this condition is not satisfied for the observation frequencies usually encountered in practice, see Pedersen (1993a) and Bibby & Sørensen (1994). Most often the estimators are severely biased. Bearing in mind the origin of these estimators and the asymptotic results that can be proven for them this deficiency might be expected, but obviously it is extremely difficult to formulate it in any rigorous mathematical way, e.g. an expression for the finite sample bias. Furthermore it is impossible to give any general criterion for how close the discrete observations should be made for these estimators to be reliable in practice. Any such critical level would depend on the given estimation problem and various characteristics of the diffusion process. Finally, it is very likely that any such level can not be met in practice, since there is usually a lower limit to how frequent the discrete observations can be made, e.g. in biology where the acquisition of each observation may be a complicated and time consuming task. The number of observations can however often be quite large, see Erlandsen & Thyssen (1983), Madsen & Holst (1991) and Pedersen (1994c). Anyway, it is only natural that the number of observations must be large for any estimator to be close to the true parameter value, and for any given observation frequency it should be the case that the estimators improve as the number of observations increase. Ideally, an estimator should be consistent as the number of discrete observations tends to infinity irrespective of the observation frequency. From these considerations one might argue that the relevant asymptotics from a practical point of view is for a fixed observation frequency (or observation scheme) and an increasing number of observations. For concreteness assume that the discrete observation time-points are given by

$$t_i = i\Delta, \; i = 0, 1, \ldots \tag{2}$$

for some fixed $\Delta > 0$. For this sampling scheme Florens-Zmirou (1989) showed that the quadratic variation like estimator $\tilde{a}_n$ of the squared (constant) diffusion coefficient and the estimator $\tilde{\theta}_n$ are inconsistent as n tends to infinity. This means that the estimators become worse as the sampling with frequency Δ is continued. Simulations furthermore indicate that both $\tilde{a}_n$ and $\tilde{\theta}_n$ have quite small variances, that is they may for large values of n be highly concentrated and severely biased. Consider for instance the one-dimensional ergodic Ornstein-Uhlenbeck process defined by the stochastic differential equation

$$dX_t = \alpha X_t dt + \sigma dW_t, \; X_0 = x_0, \; t \geq 0,$$

where $\alpha < 0, \sigma > 0$ and the discrete observation time-points are given by (2). In this case the estimator for α obtained by maximizing $\tilde{\ell}_n(\alpha)$ is given by

$$\tilde{\alpha}_n = \frac{1}{\Delta}\left(\frac{\sum_{i=1}^n X_{(i-1)\Delta}X_{i\Delta}}{\sum_{i=1}^n X_{(i-1)\Delta}^2} - 1\right),$$

whereas σ^2 might be estimated by the quadratic variation like estimator

$$\tilde{\sigma}_n^2 = \frac{1}{n\Delta}\sum_{i=1}^n (X_{i\Delta} - X_{(i-1)\Delta})^2.$$

For these estimators it is easily seen that

$$\begin{aligned} \tilde{\alpha}_n &\longrightarrow \frac{e^{\alpha_0\Delta} - 1}{\Delta} > \alpha_0 \\ \tilde{\sigma}_n^2 &\longrightarrow \sigma_0^2 \cdot \frac{e^{\alpha_0\Delta} - 1}{\alpha_0\Delta} < \sigma_0^2 \end{aligned}$$

in probability as n tends to infinity. Here α_0 and σ_0^2 denote the true parameter values. Furthermore it is easily seen that

$$\sqrt{n}(\tilde{\alpha}_n - \frac{e^{\alpha_0 \Delta} - 1}{\Delta}) \longrightarrow N(0, \frac{1 - e^{2\alpha_0 \Delta}}{\Delta^2})$$

in distribution as n tends to infinity. Hence, since

$$\begin{aligned} \frac{e^{\alpha_0 \Delta} - 1}{\Delta} &\simeq 0 \\ \frac{1 - e^{2\alpha_0 \Delta}}{\Delta^2} &\simeq 0 \end{aligned}$$

for large values of Δ we see that $\tilde{\alpha}_n$ may be highly concentrated and severely biased for large values of Δ. There is however an explicit estimator for $\theta = (\alpha, \sigma^2)$ in this case which is consistent as n tends to infinity irrespective of the value of Δ, namely the maximum likelihood estimator $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\sigma}_n^2)$ given by

$$\begin{aligned} \hat{\alpha}_n &= \frac{1}{\Delta} \log(\sum_{i=1}^{n} X_{(i-1)\Delta} X_{i\Delta} / \sum_{i=1}^{n} X_{(i-1)\Delta}^2) \\ \hat{\sigma}_n^2 &= \frac{-2\hat{\alpha}_n}{n(1 - e^{2\hat{\alpha}_n \Delta})} \sum_{i=1}^{n} (X_{i\Delta} - e^{\hat{\alpha}_n \Delta} X_{(i-1)\Delta})^2. \end{aligned}$$

Furthermore, $\hat{\theta}_n$ is asymptotically normally distributed with asymptotic covariance given by the inverse of the Fisher-information matrix, see Pedersen (1993c).

3 Approximate likelihood inference

In this section we review the approximate likelihood inference methods developed in Pedersen (1993a,c). The essential difference between the approach in Pedersen (1993a,c) and the earlier approaches considered in section 2 is that the present approach is based on approximations of the likelihood function for discrete observation of the diffusion process whereas the earlier approaches were based on approximations of the likelihood function for continuous observation of the diffusion process. This fundamental difference has two important consequences. First of all we avoid by the present approach the difficulties regarding the diffusion coefficient which the other approaches inherit from the theory for continuous observation, but more importantly, we also avoid the high sensitivity with respect to the observation frequency which the methods considered in section 2 possess, since classical likelihood inference methods based on the true log-likelihood function for discrete observations work effectively for all observation schemes, see Billingsley (1961) and Dacunha-Castelle & Florens-Zmirou (1986). The approximate likelihood inference methods developed in Pedersen (1993a,c) are based on a sequence $\{\ell_n^{(N)}(\theta)\}_{N=1}^{\infty}$ of approximations to the log-likelihood function $\ell_n(\theta)$, cf. section 2, for discrete observations at time-points $0 = t_0 < t_1 < \cdots < t_n$ of the diffusion process X defined by the stochastic differential equation (1). In these approximations the letter N denotes an integer that determines the precision of the approximation and which is chosen by the statistician. The first approximation $\ell_n^{(1)}(\theta)$ is a generalization of $\tilde{\ell}_n(\theta)$, cf. section 2, that is defined without restrictions on the parameter dependence of σ, and the approximations $\ell_n^{(N)}(\theta)$ for $N > 1$ are improvements of $\ell_n^{(1)}(\theta)$ that converge in probability under the true model to $\ell_n(\theta)$ for all θ as N tends to infinity, see Pedersen (1993c). This convergence implies that $\ell_n^{(N)}(\theta)$ for large values of N can be used as an substitute for $\ell_n(\theta)$ in all aspects of classical likelihood inference, see Pedersen (1993c). In particular the approximate maximum likelihood estimator $\hat{\theta}_n^{(N)}$ obtained by maximizing $\ell_n^{(N)}(\theta)$ with respect to θ often converges to the true maximum likelihood estimator $\hat{\theta}_n$ in probability as N tends to infinity, but even if this convergence does not hold it may still be possible to prove that $\hat{\theta}_n^{(N)}$ is consistent and asymptotically normally distributed as n and N tend to infinity with the same asymptotic distribution as $\hat{\theta}_n$, see Pedersen (1993c). The approximate log-likelihood functions can be defined under very natural and weak assumptions, see Pedersen (1993a), which makes the proposed methods quite generally applicable. In practice one simply calculates $\hat{\theta}_n^{(N)}$ for an increasing sequence of values of N until the estimates have converged. The value of $\hat{\theta}_n^{(N)}$ at termination may then be taken as an approximation of $\hat{\theta}_n$. In all considered simulations and applications this method have proven to work quite efficiently, see Pedersen (1993a,1994c). The actual calculation of $\hat{\theta}_n^{(N)}$ must usually be performed by numerical maximization of $\ell_n^{(N)}(\theta)$, see e.g. Fielding (1970). The calculation of $\ell_n^{(N)}(\theta)$ can be done by means of the general method described in Pedersen (1993a). This method involves no more than the data and the functions b and σ themselves, which makes it easy to implement. In the remainder of this section we give more details about the approximate likelihood inference methods developed in Pedersen (1993a,1993c).

Even though the transition densities of X are usually unknown they do in fact exist quite generally, see Friedman (1975) and Stroock & Varadhan (1979), and so it makes sense to approximate them when they are unknown. Indeed a sequence $\{p^{(N)}(s,x,t,y;\theta)\}_{N=1}^{\infty}$ of approximations of the transition densities of X is the basis of the definition of the approximate log-likelihood functions $\{\ell_n^{(N)}(\theta)\}_{N=1}^{\infty}$, in that

$$\ell_n^{(N)}(\theta) = \sum_{i=1}^{n} \log p^{(N)}(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i}; \theta)$$

for each $N \in \boldsymbol{N}$. The approximate transition densities $p^{(N)}(s,x,t,y;\theta)$ can be defined under the following very natural conditions. The stochastic differential equation (1) must of course have a (weak) solution for all $x_0 \in \boldsymbol{R}^d$ and $\theta \in \Theta$, and for statistical inference to be meaningful the solutions must be unique in law, see Stroock

& Varadhan (1979), Rogers & Williams (1987), Karatzas & Shreve (1988) and Revuz & Yor (1991). Furthermore, the diffusion matrix $a(t,x;\theta)=\sigma(t,x;\theta)\sigma(t,x;\theta)^T$ must be positive definite for all $t\geq 0, x\in \boldsymbol{R}^d$ and $\theta\in\Theta$. Recall that $p(s,x,t,\cdot\,;\theta)$ is the density with respect to λ^d (the d-dimensional Lebesgue measure) of the conditional distribution of X_t given $X_s=x$ for the given parameter value θ. Given X_s, the random vector X_t can be approximated by the Euler-Maruyama approximation $Y_t^{(N)}$ obtained by dividing the time-interval $[s,t]$ into N subintervals, that is $Y_t^{(N)}=Y_{\tau_N}^{(N)}$, where

$$\begin{aligned}
\tau_k &= s+k\frac{t-s}{N}\\
Y_s^{(N)} &= X_s\\
Y_{\tau_k}^{(N)} &= Y_{\tau_{k-1}}^{(N)}+\frac{t-s}{N}b(\tau_{k-1},Y_{\tau_{k-1}}^{(N)};\theta)+\sigma(\tau_{k-1},Y_{\tau_{k-1}}^{(N)};\theta)(W_{\tau_k}-W_{\tau_{k-1}})
\end{aligned}$$

for $k=1,\ldots,N$, see Kloeden & Platen (1992) and Pedersen (1993a,c). In fact $Y_t^{(N)}$ converges to X_t in L^1 as N tends to infinity, see Kloeden & Platen (1992). Thus we may approximate $p(s,x,t,\cdot\,;\theta)$ by the density $p^{(N)}(s,x,t,\cdot\,;\theta)$ with respect to λ^d of the conditional distribution of $Y_t^{(N)}$ given $X_s=x$. Indeed it follows from the L^1-convergence of $Y_t^{(N)}$ to X_t as N tends to infinity that

$$\int_A p^{(N)}(s,x,t,y;\theta)dy\longrightarrow\int_A p(s,x,t,y;\theta)dy$$

as N tends to infinity, e.g. for all open sets A in $\boldsymbol{R}^d$. The continuous version of $p^{(1)}(s,x,t,\cdot\,;\theta)$ is given by

$$p^{(1)}(s,x,t,y;\theta)=[2\pi(t-s)]^{-d/2}\cdot|a(s,x;\theta)|^{-1/2}$$

$$\cdot\exp\left(-\frac{1}{2(t-s)}[y-x-(t-s)b(s,x;\theta)]^Ta(s,x;\theta)^{-1}[y-x-(t-s)b(s,x;\theta)]\right),$$

where $|a(s,x;\theta)|$ denotes the determinant of $a(s,x;\theta)$. Moreover, for any version of $p^{(1)}(s,x,t,\cdot\,;\theta)$ a version of $p^{(N)}(s,x,t,\cdot\,;\theta)$ for $N>1$ is given by

$$p^{(N)}(s,x,t,y;\theta)=E_{\theta,s,x}\left(p^{(1)}(\tau_{N-1},Y_{\tau_{N-1}}^{(N)},t,y;\theta)\right), \tag{3}$$

where $E_{\theta,s,x}$ denotes conditional expectation given $X_s=x$ for the given parameter value θ, see Pedersen (1993a,c). From the closed expression for $p^{(1)}(s,x,t,y;\theta)$ it immediately follows that $\ell_n^{(1)}(\theta)$ is a generalization of $\tilde{\ell}_n(\theta)$, cf. section 2. Indeed, if σ does not depend on θ we have that

$$\ell_n^{(1)}(\theta)=K_n+\tilde{\ell}_n(\theta),$$

where K_n is a random variable that does not depend on θ. A notable difference between $\ell_n^{(1)}(\theta)$ and $\tilde{\ell}_n(\theta)$ is however that $\ell_n^{(1)}(\theta)$ is defined with no restrictions on the parameter dependence of σ. For the one-dimensional ergodic Ornstein-Uhlenbeck process considered in section 2 the $\ell_n^{(1)}(\theta)$-estimators are $\hat{\alpha}_n^{(1)} = \tilde{\alpha}_n$ and

$$\hat{\sigma}_n^{2(1)} = \frac{1}{n\Delta} \sum_{i=1}^{n} (X_{i\Delta} - X_{(i-1)\Delta} - \hat{\alpha}_n^{(1)} \Delta X_{(i-1)\Delta})^2.$$

Notice that the estimator $\hat{\sigma}_n^{2(1)}$ in fact is one of the improvements of the quadratic variation like estimator $\tilde{\sigma}_n^2$ suggested by Yoshida (1992). The $\ell_n^{(N)}(\theta)$-estimator $\hat{\theta}_n^{(N)}$ of θ must usually be calculated in practice by numerical maximization of $\ell_n^{(N)}(\theta)$ with respect to θ. This means that it is required to calculate $\ell_n^{(N)}(\theta)$ for a finite number of values of θ, which readily can be done by means of expression (3), see Pedersen (1993a). In fact $p^{(N)}(s, x, t, y; \theta)$ can be calculated for each fixed value of $N \in \boldsymbol{N}, 0 \leq s < t, x, y \in \boldsymbol{R}^d$ and $\theta \in \Theta$ as the average of a large number of independent replicates of $p^{(1)}(\tau_{N-1}, Y_{\tau_{N-1}}^{(N)}, t, y; \theta)$, cf. formula (3). Independent replicates of $Y_{\tau_{N-1}}^{(N)}$ are easily simulated by means of the Euler-Maruyama scheme for $Y_t^{(N)}$, see above. An important quality of this calculation method is that it involves no more than the functions b and σ themselves, which makes it generally applicable and easy to implement. In some cases, $\ell_n^{(N)}(\theta)$ is even known explicitly. This is for instance the case for multi-dimensional Gaussian diffusion processes defined by stochastic differential equations of the form

$$dX_t = (A_t X_t + a_t)dt + B_t dW_t, \ X_0 = x_0, \ t \geq 0,$$

where $A : [0, \infty) \mapsto M^{d \times d}, \ a : [0, \infty) \mapsto \boldsymbol{R}^d$ and $B : [0, \infty) \mapsto M^{d \times r}$ are deterministic continuous functions. Even though the transition densities of Gaussian diffusion processes in principle are known explicitly they can very often not be calculated explicitly in practice, see Pedersen (1993b,1994c). Applying $\ell_n^{(N)}(\theta)$ in such cases then correspond to making approximations of certain non-random vectors and matrices that defines the Gaussian Markov chain for the discrete observations $\{X_{t_i}\}_{i=0}^n$. Besides this convenient property, Gaussian diffusion proceses have several other nice statistical properties, and a general treatment of parametric statistical inference for Gaussian diffusion processes is given in Pedersen (1993b,1994c). In these papers the application of $\ell_n^{(N)}(\theta)$ is introduced in a completely different manner, but it is easily checked that the approaches are equivalent for large values of N.

As a first step in the theoretical study of $\ell_n^{(N)}(\theta)$ and $\hat{\theta}_n^{(N)}$ and as a first justification of the suggestion to use $\ell_n^{(N)}(\theta)$ and $\hat{\theta}_n^{(N)}$ for large values of N as substitutes for $\ell_n(\theta)$ and $\hat{\theta}_n$ when these are unknown, the behaviour of $p^{(N)}(s, x, t, y; \theta)$ as N tends to infinity was studied in Pedersen (1993a). From the very definition of $p^{(N)}(s, x, t, \cdot\,; \theta)$ and $p(s, x, t, \cdot\,; \theta)$ as densities with respect to λ^d it is clear that proving the pointwise convergence of $p^{(N)}(s, x, t, y; \theta)$ to $p(s, x, t, y; \theta)$ as N tends to infinity is a non-trivial task, since it involves choosing definitive versions of the respective densities. This is possible in some special cases where closed expressions for concrete versions of $p^{(N)}(s,x,t,\cdot\,;\theta)$ and $p(s,x,t,\cdot\,;\theta)$ are available, but in general it is a delicate matter. The $L^1(\lambda^d)$-convergence of $p^{(N)}(s,x,t,\cdot\,;\theta)$ to $p(s,x,t,\cdot\,;\theta)$ as N tends to infinity was however established in Pedersen (1993a), thus avoiding the problem of having to choose definitive versions. Moreover, this convergence is sufficient for statistical purposes. An immediate consequence is that

$$\ell_n^{(N)}(\theta) \longrightarrow \ell_n(\theta)$$

in probability under the true model for all θ and n as N tends to infinity, see Pedersen (1993c). The $L^1(\lambda^d)$-convergence of $p^{(N)}(s,x,t,\cdot\,;\theta)$ to $p(s,x,t,\cdot\,;\theta)$ as N tends to infinity was proven in Pedersen (1993a) in two general cases. In the first case σ was assumed to be a constant matrix (allowed to depend on θ) and the convergence could be proved under weak assumptions on b by means of ordinary stochastic calculus. However, when σ is allowed to depend on t or x, more advanced techniques are needed. This can easily be understood by heuristic arguments. Consider for a moment the time-homogeneous case, that is $b(t,x;\theta) = b(x;\theta)$ and $\sigma(t,x;\theta) = \sigma(x;\theta)$. Then

$$p^{(N)}(t,x,y;\theta) = E_{\theta,x}\left(p^{(1)}(t/N, Y^{(N)}_{(N-1)t/N}, y;\theta)\right),$$

and $p^{(1)}(t/N,u,\cdot\,;\theta)$ is the density with respect to λ^d of the d-dimensional normal distribution with mean $u + \frac{t}{N}b(u;\theta)$ and covariance $\frac{t}{N}a(u;\theta)$. Loosely speaking this means that

$$\begin{aligned} p^{(1)}(t/N,u,y;\theta) &\longrightarrow \delta_y(u) \\ Y^{(N)}_{(N-1)t/N} &\longrightarrow X_t \end{aligned}$$

as N tends to infinity, where δ_y denotes the Dirac delta function, and so

$$p^{(N)}(t,x,y;\theta) \longrightarrow E_{\theta,x}\left(\delta_y(X_t)\right)$$

as N tends to infinity, where the expectation of $\delta_y(X_t)$ has to be understood properly. Ikeda & Watanabe (1989) actually prove existence theorems for $p(t,x,y;\theta)$ by showing that it is the generalized expectation of the generalized Wiener functional $\delta_y(X_t)$. Generalized expectations and generalized Wiener functionals are fundamental concepts in Malliavin calculus. In order to enter the framework of Malliavin calculus there are two essential conditions that must be satisfied. Firstly, it is required that the solution to the stochastic differential equation can be realized on the Wiener space in the strong sense of Ikeda & Watanabe (1989), and secondly, the Malliavin covariance matrix must be sufficiently regular. These conditions are satisfied under standard but rather restrictive assumptions on b and σ, see Bell (1987) and Ikeda & Watanabe (1989). The conditions are that b and σ must be bounded with bounded derivatives with respect to x of any order, and that $a = \sigma\sigma^T$ must be strongly positive definite. Under these conditions the $L^1(\lambda^d)$-convergence of $p^{(N)}(s, x, t, \cdot; \theta)$ to $p(s, x, t, \cdot; \theta)$ as N tends to infinity was proven in Pedersen (1993a) by means of Malliavin calculus for non-constant diffusion coefficients.

The derivation of the approximate maximum likelihood estimator $\hat{\theta}_n^{(N)}$ is strongly motivated by the good properties of the maximum likelihood estimator $\hat{\theta}_n$. The general idea in the asymptotic study of $\hat{\theta}_n^{(N)}$ in Pedersen (1993c) is therefore to show that $\hat{\theta}_n^{(N)}$ inherits for large values of N whatever good properties $\hat{\theta}_n$ is assumed to possess. Accordingly, the results in Pedersen (1993c) are derived under the general assumption that $\hat{\theta}_n$ is consistent and asymptotically normally distributed as n tends to infinity. In some cases one can then prove that $\hat{\theta}_n^{(N)}$ converges to $\hat{\theta}_n$ in probability as N tends to infinity, and the consistency and asymptotic normality of $\hat{\theta}_n^{(N)}$ as N and n tend to infinity is then an immediate consequence of the consistency and asymptotic normality of $\hat{\theta}_n$ as n tends to infinity. Moreover, $\hat{\theta}_n^{(N)}$ and $\hat{\theta}_n$ have the same asymptotic distribution. In other cases, when this convergence can not be proved, it may still be possible to prove that $\hat{\theta}_n^{(N)}$ is consistent and asymptotically normally distributed as N and n tend to infinity with the same asymptotic distribution as $\hat{\theta}_n$. In such cases the proofs are based on general results on consistency and asymptotic normality, see Billingsley (1961), Sweeting (1980), Dacunha-Castelle & Duflo (1983), Dacunha-Castelle & Florens-Zmirou (1986), Jensen (1986) and Barndorff-Nielsen & Sørensen (1994), in that it is proved that these apply for $\hat{\theta}_n^{(N)}$ provided they apply for $\hat{\theta}_n$. The results in Pedersen (1993c) are first proven in a very general context and then applied to a class of one-dimensional diffusion processes. Furthermore it is indicated how the results can be applied to far more general diffusion processes. By a completely similar approach one may also prove that assumed good properties of $\ell_n(\theta)$ are inherited by $\ell_n^{(N)}(\theta)$ for large values of N. In particular one may prove that the usual asymptotic χ^2-distribution of the likelihood ratio test statistic for point or composite hypothesis about θ also hold for the approximate likelihood ratio test statistic based on $\ell_n^{(N)}(\theta)$ as N and n tend to infinity. For the purpose of hypothesis testing about θ, e.g. for the calculation of confidence regions for θ, it is advisable to use this result in practice if $\ell_n^{(N)}(\theta)$ is calculated by means of the general simulation method described in Pedersen (1993a). The reason for this is that the asymptotic Wald test statistic based on the asymptotic distribution of $\hat{\theta}_n^{(N)}$ (or $\hat{\theta}_n$) usually depends on the (approximate observed information) matrix of second derivatives of $\ell_n^{(N)}(\theta)$ with respect to θ which it may not be possible to calculate with sufficient accuracy by secant approximation of $\ell_n^{(N)}(\theta)$ and the method described in Pedersen (1993a), see also Pedersen (1994b) or section 4.

In order to avoid unnecessary technicalities, and since we have already given the essence of the results in Pedersen (1993c), we shall not repeat any of them here.

Instead we shall illustrate the results by reconsidering the one-dimensional ergodic Ornstein-Uhlenbeck process which also was considered in section 2 and previosly in this section. In this way we also illuminate some points concerning the simultaneous convergence of N and n to infinity that has not been discussed earlier. For the one-dimensional ergodic Ornstein-Uhlenbeck process with equidistant observation time-points $t_i = i\Delta,\ i = 0, 1, \ldots$ for some fixed $\Delta > 0$, the approximate maximum likelihood estimator $\hat{\theta}_n^{(N)} = (\hat{\alpha}_n^{(N)}, \hat{\sigma}_n^{2(N)})$ is for sufficiently large values of n, see Pedersen (1993c), given by

$$\begin{aligned}\hat{\alpha}_n^{(N)} &= \frac{N}{\Delta}(\psi_n^{1/N} - 1)\\ \hat{\sigma}_n^{2(N)} &= \frac{N}{\Delta}\eta_n^2 \frac{1 - \psi_n^{2/N}}{1 - \psi_n^2},\end{aligned}$$

where

$$\begin{aligned}\psi_n &= \sum_{i=1}^n X_{(i-1)\Delta} X_{i\Delta} \Big/ \sum_{i=1}^n X_{(i-1)\Delta}^2\\ \eta_n^2 &= \frac{1}{n}\sum_{i=1}^n X_{(i-1)\Delta}^2 - \frac{(\frac{1}{n}\sum_{i=1}^n X_{(i-1)\Delta} X_{i\Delta})^2}{\frac{1}{n}\sum_{i=1}^n X_{(i-1)\Delta}^2}.\end{aligned}$$

Since $-N(1 - x^{1/N}) \longrightarrow \log x$ as N tends to infinity for all $x > 0$, we have for sufficiently large values of n that

$$\hat{\theta}_n^{(N)} \longrightarrow \hat{\theta}_n$$

in probability as N tends to infinity. Now it is well-known, see e.g. Pedersen (1993c), that $\hat{\theta}_n$ is consistent and asymptotically normally distributed as n tends to infinity. In fact

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \longrightarrow N_2(0, i(\theta_0, \Delta)^{-1})$$

in distribution as n tends to infinity, where θ_0 denotes the true value of θ and $i(\theta_0, \Delta)^{-1}$ is the inverse of the Fisher-information matrix. This means, see Pedersen (1993c), that there exists a subsequence $N(n) \longrightarrow \infty$ such that

$$\hat{\theta}_n^{(N(n))} \longrightarrow \theta_0$$

in probability as $n \longrightarrow \infty$, and such that

$$\sqrt{n}(\hat{\theta}_n^{(N(n))} - \theta_0) \longrightarrow N_2(0, i(\theta_0, \Delta)^{-1})$$

in distribution as $n \longrightarrow \infty$. Moreover, if $N'(n) \longrightarrow \infty$ is a faster subsequence, i.e. $N'(n) \geq N(n)$ for all n, then the same results hold for this subsequence. This is the kind of asymptotic results that are proven for $\hat{\theta}_n^{(N)}$ in general in Pedersen (1993c). It is proved that there exists a subsequence $N(n) \longrightarrow \infty$ such that $\hat{\theta}_n^{(N(n))}$ is consistent and asymptotically normally distributed as $n \longrightarrow \infty$ with the same asymptotic distribution as $\hat{\theta}_n$. Furthermore, if $N'(n) \longrightarrow \infty$ is a faster subsequence then the same results hold for this subsequence. In practice this means that we do not have to worry about choosing the right value of N for a given number n of observations. The message is simply to increase N until the estimates converge. Simulations and actual applications, see Pedersen (1994c), show that moderate values of N are sufficient in practice. Hence, it is inessential from a practical point of view to study the rate at which N and n must converge to infinity, but obviously it would still be interesting to gain more insinght into this simultaneous convergence.

4 Quasi-likelihood inference

In this section we review the quasi-likelihood inference methods studied in Pedersen (1994b). The topic is again parametric statistical inference about the unknown parameter θ in the stochastic differential equation (1) based on discrete observations of X at time-points $0 = t_0 < t_1 < \cdots < t_n$, but the methods can immediately be extended to general Markov processes.

The fact that the transition densities of X are usually unknown is just one aspect of the more general fact that the transition distributions of X are usually unknown. Partial information about the transition distributions of X is however provided by the conditional moments of X given the past, and the general idea in Pedersen (1994b) is to base the inference about θ on these conditional moments. Some or all of the conditional moments of X are in some cases known explicitly even though the transition distributions of X are unknown, see Bibby & Sørensen (1994), but in any case, they, and their derivatives with respect to θ can be calculated in practice by means of the general method introduced in Pedersen (1994b). In rare cases, e.g. for Gaussian diffusion processes, the true log-likelihood function $\ell_n(\theta)$ for θ can be written in terms of a finite number of conditional moments of X, but in general this is obviously not the case. Hence, the conditional moments are employed in the definition of various quasi-likelihood functions for θ. The quasi-likelihood functions are derived in accordance with the general principles in Godambe & Heyde (1987), see also Heyde (1987,1988), Kulkarni & Heyde (1987) and Barndorff-Nielsen & Sørensen (1991,1994). Stated in the present context, the general idea in Godambe & Heyde (1987) is to base the inference about θ on zero-mean square integrable martingale estimating functions for θ, of which the true unknown score martingale $\dot{\ell}(\theta) = \{\dot{\ell}_n(\theta)\}_{n=1}^{\infty}$ is the optimal example. A zero-mean square integrable martingale estimating function for θ is a sequence $\{G_n(\theta)\}_{n=1}^{\infty}$ of functions $G_n(\theta)$ of θ and the data $\{X_{t_i}\}_{i=0}^{n}$ for which $G(\theta) = \{G_n(\theta)\}_{n=1}^{\infty}$ is a zero-mean square integrable martingale with respect to the natural filtration generated by the observations and under the probability measure corresponding to the given parameter value θ. The corresponding estimator for θ is defined for each n as the solution (provided it exists) of the equation $G_n(\theta) = 0$. Such estimators are under standard regularity conditions consistent and asymptotically normally distributed as n tends to infinity, see Godambe & Heyde (1987), Wefelmeyer (1992) and Bibby & Sørensen (1994). For a given class of such estimating functions one may attach to the elements of the class a general measure of information about θ, and the optimal estimating function within the class is then defined as the one that carries the highest amount of information about θ. The optimal estimating function may not be unique, but any estimating function that satisfies this criterion is called a quasi-score function of the class, and the corresponding estimator for θ is called a maximum quasi-likelihood estimator of the class. This optimality criterion is the so-called asymptotic optimality criterion defined by Godambe & Heyde (1987), but all quasi-score functions that are defined by this criterion in Pedersen (1994b) are furthermore optimal in the fixed sample sense of Godambe & Heyde (1987). The fixed sample optimality criterion defines the quasi-score function of a class as the estimating function that in a certain sense is closest to the true unknown score function. Equivalent criteria for optimality in both the fixed sample and the asymptotic sense can be found in Heyde (1988). In Pedersen (1994b), an increasing sequence of classes of estimating functions is defined, and the corresponding quasi-score functions with respect to the martingale information, see Heyde (1987), Godambe & Heyde (1987) and Barndorff-Nielsen & Sørensen (1991,1994), are found. Each class of estimating functions is an extension of the previous one that includes more conditional moments of X in the estimating functions. Interrelations between the derived quasi-score functions, the score function and other quasi-score functions are studied in detail. Finally it is shown that the quasi-score functions introduced in Pedersen (1994b) extend and improve well-known quasi-score functions for ergodic diffusion processes. We now review the developments in Pedersen (1994b) in more detail.

In the following we restrict attention to one-dimensional diffusion processes. This restriction is in no way essential, and is made entirely in order to simplify the exposition. The k'th conditional moment and the k'th conditional centered moment of X_t given $X_s = x$ are denoted by $\nu_k(s, x, t; \theta)$ and $\mu_k(s, x, t; \theta)$, respectively, for $k = 0, 1, \dots$ and all $0 \le s < t$, $x \in \boldsymbol{R}$ and $\theta \in \Theta$. If X is time-homogeneous these conditional moments are denoted by $\nu_k(t - s, x; \theta)$ and $\mu_k(t - s, x; \theta)$, respectively, and if the discrete observation time-points furthermore are equidistant, i.e. $t_i = i\Delta$, $i = 0, 1, \dots$ for some fixed $\Delta > 0$, we denote the conditional moments $\nu_k(\Delta, x; \theta)$ and $\mu_k(\Delta, x; \theta)$ by $\nu_k(x; \theta)$ and $\mu_k(x; \theta)$, respectively. Before considering the classes of estimating functions introduced in Pedersen (1994b) we shall first consider some motivating examples. Estimating functions can for instance be obtained by applying the ideas in Bibby & Sørensen (1994). Suppose $g_n(\theta)$ is a function of θ and the data $\{X_{t_i}\}_{i=0}^n$ for which an estimator for θ may be obtained by maximizing/minimizing

$g_n(\theta)$ with respect to θ, or in particular, by solving the equation $\dot{g}_n(\theta) = 0$. A zero-mean square integrable martingale estimating function for θ can then under certain regularity conditions be derived from $g_n(\theta)$ by subtracting the compensator of $\dot{g}_n(\theta)$ from $\dot{g}_n(\theta)$. If σ does not depend on θ this operation can be applied to $\tilde{\ell}_n(\theta)$, cf. section 2, leading to the estimating function

$$\tilde{G}_n(\theta) = \sum_{i=1}^{n} \frac{\dot{b}(t_{i-1}, X_{t_{i-1}}; \theta)^T}{\sigma(t_{i-1}, X_{t_{i-1}})^2} \delta_i(\theta),$$

where $\delta_i(\theta) = X_{t_i} - \nu_1(t_{i-1}, X_{t_{i-1}}, t_i; \theta)$, $i = 1, \ldots, n$. Bibby & Sørensen (1994) proposed also to use $\tilde{G}$ when σ depends on θ, since $\tilde{G}(\theta)$ in that case remains a martingale. Applying the same operation to the generalization $\ell_n^{(1)}(\theta)$ of $\tilde{\ell}_n(\theta)$, cf. section 3, leads to the estimating function

$$\begin{aligned} G_n^{(1)}(\theta) &= \sum_{i=1}^{n} \frac{\dot{b}(t_{i-1}, X_{t_{i-1}}; \theta)^T}{\sigma(t_{i-1}, X_{t_{i-1}}; \theta)^2} \delta_i(\theta) + \sum_{i=1}^{n} \Big(a_0(t_{i-1}, X_{t_{i-1}}, t_i; \theta) \\ &\quad + a_1(t_{i-1}, X_{t_{i-1}}, t_i; \theta) \delta_i(\theta) + a_2(t_{i-1}, X_{t_{i-1}}, t_i; \theta) \delta_i(\theta)^2 \Big), \end{aligned}$$

where

$$\begin{aligned} a_0(s, x, t; \theta) &= -\frac{\mu_2(s, x, t; \theta) \dot{\sigma}(s, x; \theta)^T}{(t-s)\sigma(s, x; \theta)^3} \\ a_1(s, x, t, ; \theta) &= \frac{2[\nu_1(s, x, t; \theta) - x - (t-s) b(s, x; \theta)] \dot{\sigma}(s, x; \theta)^T}{(t-s)\sigma(s, x; \theta)^3} \\ a_2(s, x, t; \theta) &= \frac{\dot{\sigma}(s, x; \theta)^T}{(t-s)\sigma(s, x; \theta)^3}. \end{aligned}$$

The variables $\delta_i(\theta)^k$, $i = 1, 2, \ldots$ for some integer k are called the k'th order increments. Notice that $\tilde{G} = G^{(1)}$ if σ does not depend on θ. The essential difference between $\tilde{G}$ and $G^{(1)}$ is that $G^{(1)}$ includes both first and second order increments whereas $\tilde{G}$ includes only first order increments, which as a consequence of the compensation operation implies that $G^{(1)}$ uses both first and second order conditional moments of X whereas $\tilde{G}$ uses only first order conditional moments of X. The difference between $\tilde{G}$ and $G^{(1)}$ is particularly apparent in estimation of unknown parameters in σ. If for instance b does not depend on θ then the estimator corresponding to $\tilde{G}$ does not exist whereas the estimator corresponding to $G^{(1)}$ still exists. Similar conclusions hold if σ depends on parameters that do not enter the drift b. Such models are particularly of interest in mathematical finance, where parameters appearing only in σ (the volatility) are of major interest, see also the examples in Bibby & Sørensen (1994) and Pedersen (1994b). The differences between $\tilde{G}$ and $G^{(1)}$ can furthermore be quantified in terms of the information about θ that they carry, or alternatively, in terms of the information about θ that is carried by the quasi-score functions of the two classes of estimating functions which $\tilde{G}$ and $G^{(1)}$ naturally belong to. The estimating function $G^{(1)}$ naturally belongs to the class of estimating functions of the form

$$\begin{aligned} G_n(\theta) &= \sum_{i=1}^{n} \Big(a_0(t_{i-1}, X_{t_{i-1}}, t_i; \theta) \\ &\quad + a_1(t_{i-1}, X_{t_{i-1}}, t_i; \theta)\delta_i(\theta) + a_2(t_{i-1}, X_{t_{i-1}}, t_i; \theta)\delta_i(\theta)^2 \Big) \end{aligned} \tag{4}$$

with $a_0 = -a_2\mu_2$ to ensure the martingale property of $G(\theta)$, whereas $\tilde{G}$ naturally belongs to the subclass defined by the restriction $a_0 = a_2 = 0$. These two classes of estimating functions were studied by Wefelmeyer (1992) for time-homogeneous ergodic diffusion processes and time-equidistant observations, for which one may attach the expected information as a measure of information about θ to the elements of the classes. Doing so, Wefelmeyer (1992) found that the respective quasi-score functions are given by

$$\begin{aligned} \tilde{G}_n^*(\theta) &= \sum_{i=1}^{n} \frac{\dot{\nu}_1(x;\theta)^T}{\mu_2(x;\theta)} \delta_i(\theta) \\ G_n^{(1)*}(\theta) &= \sum_{i=1}^{n} \sum_{k=0}^{2} a_k^*(X_{t_{i-1}};\theta)\delta_i(\theta)^k, \end{aligned}$$

where $(a_0^* = -a_2^*\mu_2)$

$$\begin{aligned} a_1^* &= \frac{(\mu_4 - \mu_2^2)\dot{\nu}_1^T - \mu_3\dot{\mu}_2^T}{(\mu_4 - \mu_2^2)\mu_2 - \mu_3^2} \\ a_2^* &= \frac{-\mu_3\dot{\nu}_1^T + \mu_2\dot{\mu}_2^T}{(\mu_4 - \mu_2^2)\mu_2 - \mu_3^2}. \end{aligned}$$

Moreover, he showed that the expected information for $G^{(1)*}$ is strictly larger than the expected information for $\tilde{G}^*$ unless $\mu_2\dot{\mu}_2 = \mu_3\dot{\nu}_1$. This means that even if σ does not depend on θ then there is still a loss in information in using $\tilde{G}^*$ compared to using $G^{(1)*}$, see the examples in Pedersen (1994b). Again the essential difference between $\tilde{G}^*$ and $G^{(1)*}$ is that $G^{(1)*}$ uses more conditional moments of X as a consequence of being defined in terms of more increments $\delta_i(\theta)^k$ and the martingale property. Exactly the same results hold for far more general diffusion processes, including non-ergodic and time-inhomogeneous diffusion processes for which the expected information may not be well-defined, provided one uses the martingale information proposed by Heyde (1987) as a measure of information about θ, see Pedersen (1994b). This indicates that the martingale information may be an appropriate choice of a general measure of information about θ for martingale estimating functions. Indeed such a general measure is needed outside the class of diffusion processes for which the expected information can be used. In addition, the martingale information can in general be calculated in practice whereas the expected information is usually not known, see Pedersen (1994b).

Two conclusions may be drawn from the considerations above. Firstly, more information about θ is evidently carried by estimating functions that use more conditional moments of X. Moreover, conditional moments of X appear in a natural way in an estimating function given as a sum of affine combinations of increments $\delta_i(\theta)^k$ when the estimating function is required to be a martingale. Secondly, the martingale information seems to be an appropriate choice of a general measure of information about θ for martingale estimating functions. Further justification for using the martingale information is given below. With this motivation the following increasing sequence of classes of zero-mean square integrable martingale estimating functions was defined in Pedersen (1994b). A zero-mean square integrable martingale estimating function G, satisfying certain regularity conditions, belongs to the class $\mathcal{C}_N$, where $N \in \boldsymbol{N}$, if it can be written on the form

$$G_n(\theta) = \sum_{i=1}^{n} \sum_{k=0}^{N} a_k(t_{i-1}, X_{t_{i-1}}, t_i; \theta) \delta_i(\theta)^k. \tag{5}$$

Elements of $\mathcal{C}_N$ are denoted by G_N and their coefficients are denoted by $\{a_{N,k}\}_{k=0}^N$. Notice that the martingale property of $G_N \in \mathcal{C}_N$ implies that

$$a_{N,0} = -\sum_{k=1}^{N} a_{N,k}\mu_k,$$

that is the presence of conditional moments of X up to and including order N is implicit for $G_N \in \mathcal{C}_N$. Other regularity conditions are also implicit in the definition of $\mathcal{C}_N$, see Pedersen (1994b). The form (5) of the elements in $\mathcal{C}_N$ can also be motivated by a formal truncated Taylor-expansion of the score function, see Pedersen (1994b). By viewing the class $\mathcal{C}_N$ in that way it furthermore becomes apparent how the definition of $\mathcal{C}_N$ should be extended to multi-dimensional diffusion processes, see Pedersen (1994b).

For a zero-mean square integrable martingale estimating function G, satisfying certain regularity conditions, the martingale information is defined as

$$I_n(\theta) = \bar{G}_n(\theta)^T < G(\theta) >_n^{-1} \bar{G}_n(\theta),$$

where $\bar{G}(\theta) = \{\bar{G}_n(\theta)\}_{n=1}^\infty$ is the compensator of $\dot{G}(\theta) = \{\dot{G}_n(\theta)\}_{n=1}^\infty$ and $< G(\theta) >= \{< G(\theta) >_n\}_{n=1}^\infty$ is the quadratic characteristic of $G(\theta)$. By attaching this martingale information to the elements of the classes $\{\mathcal{C}_N\}_{N=1}^\infty$ the corresponding quasi-score functions were found in Pedersen (1994b). Under certain regularity conditions the estimating function $G_N^* \in \mathcal{C}_N$ with coefficients

$$a_{N,k}^* = \sum_{l=1}^{N} A_N^{kl} \xi_l, \;\; k = 1, \ldots, N,$$

where $\xi_k = \dot{\mu}_k^T + k \dot{\nu}_1^T \mu_{k-1}$ for $k = 1, \ldots, N$ and A_N^{kl} denotes the kl'th element of the inverse of the positive definite $N \times N$ matrix $A_N = \{\mu_{k+l} - \mu_k \mu_l\}_{k,l=1}^N$, is optimal within $\mathcal{C}_N$ in both the fixed sample and the asymptotic sense of Godambe & Heyde (1987). The (optimal) martingale information corresponding to G_N^* can be written in the four equivalent forms

$$\begin{aligned} I_{N,n}^*(\theta) &= \bar{G}_{N,n}^*(\theta)^T < G_N^*(\theta) >_n^{-1} \bar{G}_{N,n}^*(\theta) \\ &= -\bar{G}_{N,n}^*(\theta) \\ &= < G_N^*(\theta) >_n \\ &= \sum_{i=1}^{n} \sum_{k=1}^{N} \sum_{l=1}^{N} \psi_{N,k,l}(t_{i-1}, X_{t_{i-1}}, t_i; \theta), \end{aligned}$$

where

$$\psi_{N,k,l} = A_N^{kl} \xi_k \xi_l^T, \;\; k, l = 1, \ldots, N.$$

The quasi-score function G_N^* of $\mathcal{C}_N$ is not necessarily unique for multi-dimensional parameters θ, but for one-dimensional parameters it was proven in Pedersen (1994b) to be unique up to a constant non-zero multiple. The four equivalent forms of the optimal martingale information within $\mathcal{C}_N$ further justify the use of the martingale information, since the two information measures $-\bar{G}_N^*(\theta)$ and $< G_N^*(\theta) >$ are well-accepted measures of information, see Barndorff-Nielsen (1991,1994) for a discuusion of various general measures of information. From the closed expressions for the coefficients of G_N^* and the fourth equivalent form of I_N^* it is evident that quasi-likelihood inference about θ can easily be performed in practice by means of G_N^* and I_N^*, provided the conditional moments of X and their derivatives with respect to θ can be calculated in practice. This can be done by means of the general method derived in Pedersen (1994b), see also later in this section. However, first one must decide which quasi-score function to use, and in this respect it is important to be able to assess whether further information about θ can be obtained by including increments of one higher order in the quasi-score function. In Pedersen (1994b), three theorems are given that characterizes when this is the case. Firstly, a result is proven for general classes of zero-mean square integrable martingale estimating functions that characterizes the difference between the optimal martingale information within a given class and the optimal martingale information within a subclass in terms of the quadratic characteristics of the respective quasi-score functions and their mutual quadratic characteristic. A more practicable result that is specific for the classes $\{\mathcal{C}_N\}_{N=1}^{\infty}$ states that if one finds that $a_{N,N}^* = 0$ then

$$I_{N,n}^*(\theta) = I_{N-1,n}^*(\theta)$$

almost surely for all n and θ. Finally it is proven that if the true score function belongs to $\mathcal{C}_N$ then it is equivalent to G_N^* in the sense that they are both optimal within $\mathcal{C}_N$ in both the fixed sample and the asymptotic sense of Godambe & Heyde (1987), and no further information about θ can be obtained by including higher order increments in the quasi-score function. In fact

$$E_\theta(-\ddot{\ell}_n(\theta)) = E_\theta(I_{N,n}^*(\theta))$$

for all n and θ if $\dot{\ell} \in \mathcal{C}_N$.

In practice it is most often the case that the conditional moments of X and their derivatives with respect to θ are unknown. In such cases they can be calculated by means of the general method derived in Pedersen (1994b). According to the binomial formula it is sufficient to be able to calculate ν_k and $\dot{\nu}_k$ for $k = 0, 1, \ldots$. The calculation of $\nu_k(s, x, t; \theta)$ can for given values of $0 \leq s < t, x \in \boldsymbol{R}$ and $\theta \in \Theta$ be done by means of simulation, in that

$$\nu_k(s, x, t; \theta) = E_\theta(X_t^k | X_s = x)$$

can be calculated as the average of a large number of independent replicates of X_t given $X_s = x$, see Kloeden & Platen (1992). This approach works very efficiently in practice, but the calculation of $\dot{\nu}_k(s, x, t; \theta)$ for given values of $0 \leq s < t, x \in \boldsymbol{R}$ and $\theta \in \Theta$ is more problematic, since the usual secant approximation of $\dot{\nu}_k(s, x, t; \theta)$ may be very poor if ν_k is calculated by means of simulations as above, see Pedersen (1994b). Instead one can calculate $\nu_k(s, x, t; \theta)$ and $\dot{\nu}_k(s, x, t; \theta)$ simultaneously by means of simulations by exploiting some analytical properties of strong solutions to stochastic differential equation in the sense of Friedman (1975). Assume for simplicity that θ is one-dimensional, and suppose $(X_t^{\theta,s,x})_{t \geq s}$ is a strong solution (on some probability space) to (1) with initial condition $X_s = x$. Then the two-dimensional process

$$Y_t^{\theta,s,x} = \begin{pmatrix} X_t^{\theta,s,x} \\ \theta \end{pmatrix}, \ t \geq s$$

is also a strong solution to a stochastic differential equation, and under some additional conditions, $Y_t^{\theta,s,x}$ is furthermore differentiable in the L^2-sense with respect to its initial values $(x, \theta)^T$. This implies that $X_t^{\theta,s,x}$ is differentiable in the L^2-sense with respect to θ, and that

$$\begin{aligned} \dot{\nu}_k(s, x, t; \theta) &= E(\frac{\partial}{\partial\theta}((X_t^{\theta,s,x})^k)) \\ &= kE((X_t^{\theta,s,x})^{k-1} \dot{X}_t^{\theta,s,x}), \end{aligned}$$

where $\dot{X}_t^{\theta,s,x}$ denotes the derivative in the L^2-sense of $X_t^{\theta,s,x}$ with respect to θ. Hence, $\dot{\nu}_k(s,x,t;\theta)$ can be calculated as the average of a large number of independent replicates of

$$k(X_t^{\theta,s,x})^{k-1}\dot{X}_t^{\theta,s,x}.$$

The simulations of $X_t^{\theta,s,x}$ and $\dot{X}_t^{\theta,s,x}$ needed for the calculation of $\nu_k(s,x,t;\theta)$ and $\dot{\nu}_k(s,x,t;\theta)$ can be obtained simultaneously, since the two-dimensional process

$$Z_t^{\theta,s,x} = \begin{pmatrix} X_t^{\theta,s,x} \\ \dot{X}_t^{\theta,s,x} \end{pmatrix}, \; t \geq s$$

satisfies the two-dimensional stochastic differential equation

$$dZ_t = B(t,Z_t;\theta)dt + \Sigma(t,Z_t;\theta)dW_t, \; Z_s = \begin{pmatrix} x \\ 0 \end{pmatrix}, \; t \geq s, \tag{6}$$

where W is a one-dimensional Wiener process, and

$$\begin{aligned} B(t,z;\theta) &= \begin{pmatrix} b(t,x;\theta) \\ b'(t,x;\theta)y + \dot{b}(t,x;\theta) \end{pmatrix} \\ \Sigma(t,z;\theta) &= \begin{pmatrix} \sigma(t,x;\theta) \\ \sigma'(t,x;\theta)y + \dot{\sigma}(t,x;\theta) \end{pmatrix} \end{aligned}$$

for $z = (x,y)^T$, see Pedersen(1994b). Independent replicates of $Z_t^{\theta,s,x}$ can thus easily be simulated by means of any stochastic Itô-Taylor scheme for (6), see Kloeden & Platen (1992). This calculation method was applied in a numerical example in Pedersen (1994b), where independent replicates of $Z_t^{\theta,s,x}$ were simulated by means the Milstein scheme for (6).

5 Model validation

In this section we review the model validation techniques introduced in Pedersen (1994a). The methods apply for general discrete-time stochastic processes, but special attention is given to discretely observed diffusion processes.

In parametric statistical inference it is essential to have a tool for validating the parametric model. Classical concepts in this respect are test for goodness of fit, the Kolmogorov-Smirnov test and analysis of residuals, but several other specialzed techniques are available in concrete models. Outside the class of Gaussian diffusion processes, see Pedersen (1993b,1994c), such a tool has however been missing for discretely observed diffusion processes. In Pedersen (1994a) some uniform residuals for discretely observed diffusion processes are defined, that is a sequence of functions of the data that are stochastically independent and uniformly distributed if the parametric model given by the stochastic differential equation defining the diffusion process is valid. These uniform residuals can be calculated in practice by means of simulations as described in Pedersen (1994a), and the validation is then performed by analyzing them by standard techniques (goodness of fit, Kolmogorov-Smirnov, histogram, quantile plot etc.). The uniform residuals are closely related to classical standard normal residuals when these apply. Moreover, the general method described in Pedersen (1994a) for calculating the uniform residuals for discretely observed diffusion processes also enables us to calculate some forecasts, forecast error covariances and standardized forecast errors which generalize similar concepts for Gaussian diffusion processes, see Pedersen (1994c).

Suppose we observe the $\boldsymbol{R}^d$-valued random vectors $X_0, X_1, \ldots, X_n$, and that we want to validate the parametric model given by the parametrized family $\{P_\theta : \theta \in \Theta\}$ of probability measures. Let $X_i^{(k)}$ denote the k'th coordinate of X_i and let $F_i^{(k)}(\cdot\,|x_0, x_1, \ldots, x_{i-1}; \theta)$ denote the conditional distribution function under P_θ of $X_i^{(k)}$ given $X_0 = x_0, X_1 = x_1, \ldots, X_{i-1} = x_{i-1}$. Then the validation can be based on the following simple observation, see Pedersen (1994a). Suppose that the conditional distribution function $F_i^{(k)}(\cdot\,|x_0, x_1, \ldots, x_{i-1}; \theta)$ is absolutely continuous for all $\{x_j\}_{j=1}^{i-1}, k \in \{1, \ldots, d\}$ and $i \in \{1, \ldots, n\}$. Then we have for each $k \in \{1, \ldots, d\}$ that

$$U_i^{(k)}(\theta) = F_i^{(k)}(X_i^{(k)}|X_0, X_1, \ldots, X_{i-1}; \theta), \ i = 1, \ldots, n$$

is a sequence of stochastically independent and uniformly distributed random variables under P_θ. This result can be extended in several directions, see Pedersen (1994a), but the present formulation suits the purpose of discretely observed diffusion processes. Usually the parameter θ is unknown, and so the validation must be based on the estimated uniform residuals $\{U_i^{(k)}(\hat{\theta})\}_{i=1}^n$ for $k = 1, \ldots, d$, where $\hat{\theta}$ is some estimator for θ, e.g. the maximum likelihood estimator for θ in the parametric model.

Suppose that $X_0, X_1, \ldots, X_n$ in the general setting is a sample from a d-dimensional discrete-time Markov process with absolutely continuous marginal transition distribution functions $F_i^{(k)}(\cdot\,|x_{i-1}; \theta), \ i = 1, \ldots, n, \ k = 1, \ldots, d$ under P_θ. Then the uniform residuals are given by

$$U_i^{(k)}(\theta) \quad = \quad F_i^{(k)}(X_i^{(k)}|X_{i-1}; \theta), \ i = 1, \ldots, n, \ k = 1, \ldots, d.$$

As a particular case we have discrete observations at time-points $0 = t_0 < t_1 < \cdots < t_n$ of the diffusion process X defined by (1). Let $F^{(k)}(s, x, t, \cdot; \theta)$ denote the conditional distribution function of $X_t^{(k)}$ given $X_s = x$, which is absolutely continuous provided the transition densities of X exist. The uniform residuals are then given by

$$U_{t_i}^{(k)}(\theta) = F^{(k)}(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i}; \theta), \ i = 1, \ldots, n, \ k = 1, \ldots, d.$$

As a supplement to the validation of the parametric model given by (1) based on the uniform residuals one may also examine the standardized forecast erros

$$R_{t_i}(\theta) = V(t_{i-1}, X_{t_{i-1}}, t_i; \theta)^{-1/2} \left(X_{t_i} - E(t_{i-1}, X_{t_{i-1}}, t_i; \theta) \right), \ i = 1, \ldots, n,$$

where

$$\begin{aligned} E(s,x,t;\theta) &= E_\theta(X_t | X_s = x) \\ V(s,x,t;\theta) &= E_\theta([X_t - E(s,x,t;\theta)][X_t - E(s,x,t;\theta)]^T | X_s = x). \end{aligned}$$

Notice that the standardized forecast errors are stochastically independent and standard d-dimensional normally distributed if X is Gaussian, see Pedersen (1993b, 1994c). For non-Gaussian diffusion processes the usefulness of the standardized forecast errors is however severely limited by the fact that their distributional properties are unknown. Alternatively, one might compare the data with the l-step-ahead forecasts $E(t_{i-1}, X_{t_{i-1}}, t_{i-1+l}; \theta)$, see Jazwinski (1970) and Åström (1970), and with the l-step-ahead forecast 95%-variation bounds given by the limits

$$E(t_{i-1}, X_{t_{i-1}}, t_{i-1+l}; \theta)_k \pm 1.96 V(t_{i-1}, X_{t_{i-1}}, t_{i-1+l}; \theta)_{kk}, \ i = 1, \ldots, n-l+1$$

for $k = 1, \ldots, d$, see Pedersen (1994c). The uniform residuals, the l-step-ahead forecasts, the l-step-ahead forecast error covariances and the standardized forecast errors can all be calculated in practice if one can calculate $F^{(k)}(s,x,t,y;\theta)$, $E(s,x,t;\theta)$ and $V(s,x,t;\theta)$ for $k = 1, \ldots, d$ and all $0 \leq s < t, x, y \in \boldsymbol{R}^d$ and $\theta \in \Theta$. This can straightforwardly be done by means of simulations as described in Pedersen (1994a). The proposed method is also applied in a numerical example in Pedersen (1994a).

6 References

Barndorff-Nielsen, O.E. & Sørensen, M. (1991): Information Quantities in a Non-classical Setting. *Comp. Stat. & Data Anal.* **12**, 143–158.

Barndorff-Nielsen, O.E. & Sørensen, M. (1994): A Review of Some Aspects of Asymptotic Likelihood Theory for Stochastic Processes. *Int. Stat. Rev.* **62**.

Basawa, I.V. & Prakasa Rao, B.L.S. (1980): *Statistical Inference for Stochastic Processes.* Academic Press, London.

Bell, D.R. (1987): *The Malliavin Calculus.* Wiley, New York.

Bibby, B.M. & Sørensen, M. (1994): Martingale Estimation Functions for Discretely Observed Diffusion Processes. *Research Report* **277**, Dept. Theor. Statist., University of Aarhus.

BILLINGSLEY, P. (1961): *Statistical Inference for Markov Processes.* Chicago Press, Chicago.

BROWN, B.M. & HEWITT, J.I. (1975): Asymptotic Likelihood Theory for Diffusion Processes. *J. Appl. Prob.* **12**, 228–238.

DACUNHA-CASTELLE, D. & DUFLO, M. (1983): *Probabilités et Statistiques.* Masson, Paris.

DACUNHA-CASTELLE, D. & FLORENS-ZMIROU, D. (1986): Estimation of the Coefficients of a Diffusion from Discrete Observations. *Stochastics* **19**, 263–284.

DOROGOVCEV, A.J. (1976): The Consistency of an Estimate of a Parameter of a Stochastic Differential Equation. *Theor. Prob. and Math. Statist.* **10**, 73–82.

ERLANDSEN, M. & THYSSEN, N. (1983): Modelling the Community Oxygen Concentration in Lowland Streams Dominated by Submerged Macrophytes. *Analysis of Ecological Systems: State-of-the-Art in Ecological Modelling*, 855-860. Elsevier Scientific Publishing Company.

FEIGIN, P.D. (1976): Maximum Likelihood Estimation for Continuous-time Stochastic Processes. *Adv. Appl. Prob.* **8**, 712–736.

FIELDING, K. (1970): Function Minimization and Linear Search. *Communications of the ACM* **13** 8, 509–510.

FLORENS-ZMIROU, D. (1989): Approximate Discrete-Time Schemes for Statistics of Diffusion Processes. *Statistics* **20** 4, 547–557.

FLORENS-ZMIROU, D. (1993): On Estimating the Diffusion Coefficient from Discrete Observations. *J. Appl. Prob.* **30** 4, 790–804.

FRIEDMAN, A. (1975): *Stochastic Differential Equations and Applications*, Volume 1. Academic Press, New York.

GENON-CATALOT, V. (1990): Maximum Contrast Estimation for Diffusion Processes from Discrete Observations. *Statistics* **21**, 99–116.

GENON-CATALOT, V. & JACOD, J. (1993): On the Estimation of the Diffusion Coefficient for Multi-dimensional Diffusion Processes. *Ann. Inst. Henri Poincaré* **29** 1, 119–151.

GENON-CATALOT, V. & JACOD, J. (1994): Estimation of the Diffusion Coefficient for Diffusion Processes: Random Sampling. *Scand. J. Statist.* **21**, 193–221.

Godambe, V.P. & Heyde, C.C. (1987): Quasi-likelihood and Optimal Estimation. *Int. Stat. Rev.* **55** 3, 231–244.

Guttorp, P. & Kulperger, R. (1984): Statistical Inference for some Volterra Population processes in a Random Environment. *The Canadian Journal of Statistics* **12** 4, 289–302.

Heyde, C.C. (1987): On Combining Quasi-likelihood Estimating Functions. *Stoch. Proc. Appl.* **25**, 281–287.

Heyde, C.C. (1988): Fixed Sample and Asymptotic Optimality for Classes of Estimating Functions. *Contemporary Mathematics* **80**, 241–247.

Hutton, J.E. & Nelson, P.I. (1986): Quasi-Likelihood Estimation for Semimartingales. *Stochastic Processes and their Applications* **22**, 245–257.

Ikeda, N. & Watanabe, S. (1989): *Stochastic Differential Equations and Diffusion Processses*, Second Edition. North-Holland, Amsterdam.

Jazwinski, A.H. (1970): *Stochastic Processes and Filtering Theory.* Academic Press, New York.

Jensen, J.L. (1986): *Nogle Asymptotiske Resultater.* University of Aarhus.

Jensen, J.P., Bøgestrand, J., Jeppesen, E., Pedersen, A.R., Søndergaard, M., Windolf, J., Sortkjær, L. (1994): *Ferske Vandområder – Søer.* Faglig rapport fra Danmarks Miljøundersøgelser, Afdeling for Ferskvandsøkologi. Vandmiljøplanens Overvågningsprogram, Miljøministeriet.

Karatzas, I. & Shreve, S.E. (1988): *Brownian Motion and Stochastic Calculus.* Springer, New York.

Kasonga, R.A. (1988): The Consistency of a Non-linear Least Squares Estimator from Diffusion Processes. *Stoch. Proc. Appl.* **30**, 263–275.

Kloeden, P.E. & Platen, E. (1992): *Numerical Solution of Stochastic Differential Equations.* Springer, New York.

Kloeden, P.E., Schurz, H., Platen, E., Sørensen, M. (1992): On Effects of Discretization on Estimates of Drift Parameters for Diffusion Processes. *Research Report No.* **249**, Dept. Theor. Statist., University of Aarhus.

Kuchler, U. & Sørensen, M. : *Exponential Families of Stochastic Processes.* In preparation.

KULKARNI, P.M. & HEYDE, C.C. (1987): Optimal Robust Estimation for Discrete-time Stochastic Processes. *Stoch. Proc. Appl.* **26**, 267–276.

KUTOYANTS, Y.A. (1977): Estimation of a Trend Parameter of a Diffusion Process in the Smooth Case. *Theor. Prob. and Appl.* **22**, 399–405.

KUTOYANTS, Y.A. (1984): *Parameter Estimation for Stochastic Processes.* Research and Exposition in Math. 6, Heldermann Verlag, Berlin.

LANSKA, V. (1979): Minimum Contrast Estimation in Diffusion Processes. *J. Appl. Prob.* **16**, 65–75.

LE BRETON, A. (1974): Parameter Estimation in a Linear Stochastic Differential Equation. *Trans. of 7th Prague Conf. and 1974 E.M.S.,* 353–366.

LE BRETON, A. (1976): On Continuous and Discrete Sampling for Parameter Estimation in Diffusion Type Proceses. *Math. Programming Study* **5**, 124–144.

LE BRETON, A. & MUSIELA, M. (1984): A Study of a One-dimensional Bilinear Differential Model for Stochastic Processes. *Prob. and Math. Statist.* **4** 1, 91–107.

LIPTSER, R.S. & SHIRYAYEV, A.N. (1977): *Statistics of Random Processes I: General Theory.* Springer, New York.

MADSEN, H. & HOLST, J. (1991): Estimation of Continuous Time Models for the Heat Dynamics of a Building. *Tech. rep.* **4**, University of Lund.

MADSEN, H. & MELGAARD, H. (1991): The Mathematical and Numerical Methods used in CTLSM — a Program for ML-estimation in Stochastic, Continuous Time Dynamical Models. *Tech. rep.* **7**, Technical University of Denmark, Lyngby, Denmark.

PEDERSEN, A.R. (1993a): A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations. *Research Report No.* **264**, Dept. Theor. Statist., University of Aarhus. To appear in *Scand. J. Statist.*

PEDERSEN, A.R. (1993b): Maximum Likelihood Estimation Based on Incomplete Observations for a Class of Discrete Time Stochastic Processes by Means of the Kalman Filter. *Research Report No.* **272**, Dept. Theor. Statist., University of Aarhus.

PEDERSEN, A.R. (1993c): Consistency and Asymptotic Normality of an Approximate Maximum Likelihood Estimator for Discretely Observed Diffusion Processes. *Research Report No.* **273**, Dept. Theor. Statist., University of Aarhus.

PEDERSEN, A.R. (1994a): Uniform Residuals for Discretely Observed Diffusion Processes. *Research Report No.* **292**, Dept. Theor. Statist., University of Aarhus.

PEDERSEN, A.R. (1994b): Quasi-likelihood Inference for Discretely Observed Diffusion Processes. *Research Report No.* **295**, Dept. Theor. Statist., University of Aarhus.

PEDERSEN, A.R. (1994c): Statistical Analysis of Gaussian Diffusion Processes Based on Incomplete Discrete Observations. *Research Report No.* **297**, Dept. Theor. Statist., University of Aarhus.

PENEV, S.I. (1985): Parametric Statistical Inference for Multivariate Diffusion Processes Using Discrete Observations. *Proc. Fourteenth Spring Conf. Bulg. Math.* , 501–510.

PRAKASA RAO, B.L.S. (1983): Asymptotic Theory for Non-linear Least Squares Estimator for Diffusion Processes. *Math. Oper. Statist., Ser. Statist., Berlin* **14** 2, 195–209.

PRAKASA RAO, B.L.S. (1985): Estimation of the Drift for Diffusion Processes. *Statistics* **16** 2, 263–275.

PRAKASA RAO, B.L.S. (1988): Statistical Inference from Sampled Data for Stochastic Processes. *Contemporary Mathematics* **80**, 249–284.

REVUZ, D. & YOR, M. (1991): *Continuous Martingales and Brownian Motion.* Springer, New York.

ROGERS, L.C.G. & WILLIAMS, D. (1987): *Diffusions, Markov Processes, and Martingales, Volume 2: Itô Calculus.* Wiley, Chichester.

STROOCK, D.W. & VARADHAN, S.R.S. (1979): *Multidimensional Diffusion Processes.* Springer, Berlin.

SWEETING, T.J. (1980): Uniform Asymptotic Normality of the Maximum Likelihood Estimator. *Ann. Stat.* **8** 6, 1375–1381.

SØRENSEN, M. (1990): On Quasi Likelihood for Semimartingales. *Stoch. Proc. Appl.* **35**, 331–346.

TARASKIN, A.F. (1974): On the Asymptotic Normality of Vector-valued Stochastic Integrals and Estimates of Drift Parameters of a Multi-dimensional Diffusion Process. *Theor. Prob. and Math. Statist.* **2**, 209–224.

THYSSEN, N., ERLANDSEN, M., KRONVANG, B., SVENDSEN, L.M. (1990): *Vandløbsmodeller — Biologisk Struktur og Stofomsætning.* Miljøministeriet, Miljøstyrelsen.

WEFELMEYER, W. (1992): Quasi-likelihood Models and Optimal Inference. *Preprints in Statistics* **138**, University of Cologne.

YOSHIDA, N. (1992): Estimation for Diffusion Processes from Discrete Observations. *J. Mult. Anal.* **41** 2, 220–242.

ÅSTRÖM, K.J. (1970): *Introduction to Stochastic Control Theory.* Academic Press, New York.

Part II: Statistical Applications of Gaussian Diffusion Processes in Freshwater Ecology

1 Introduction

Many Danish lakes have very high concentrations of nitrogen and phosphorus mainly due to large loadings of poorly treated waste water in the past and leackage from agricultivated soil. During the seventies and eighties various actions were undertaken to improve and intensify the efforts at the Danish sewage plants, and in 1989 the Monitoring Program of the Danish Action Plan for the Aquatic Environment was effectuated, see Jensen et al (1994). In particular a standard data acquisition procedure for 37 representative Danish lakes was initiated. The data are reported to the National Environmental Research Institute, Department of Freshwater Ecology. One of the purposes of this data acquisition is to employ the collected data in the development of dynamic statistical models that can predict the effects of reducing the external loading of nitrogen and phosphorus to a lake, see Jensen et al (1994), Pedersen (1994c) and section 2 and 3. Such prognosis models can be very helpful in the strategic planning of future actions of sewage plants and other environmental efforts. Historically, the environmental efforts in Danish lakes have primarily been concerned with the reduction of the external loading of nitrogen and phosphorus. However, a reduced external loading of nitrogen or phosphorus to a lake often implies an increased internal loading of the same nutrient due to an increased release of the nutrient from the large amount in the sediment that has accumulated in the past, see Jensen et al (1992), Søndergaard et al (1993) and Jensen et al (1994). This means that the effect of such environmental efforts on the water quality of the lake can be seriously delayed. Accordingly, one of the important topics in present applied lake research is to study the possibilities for reducing the internal loading of nitrogen and phosphorus. For this purpose there is a need for detailed dynamic statistical models that describe the dynamics and interactions of nitrogen, phosphorus and other biological and chemical processes in a lake, see Pedersen (1994c) and section 2 and 3. More generally, such research models can be very useful in the development of alternative means of environmental control.

Ordinary differential equations are widely used in biology as theoretical models for the evolution and interaction of biological and chemical processes, see e.g. Odum (1956), Erlandsen & Thyssen (1983) and Thyssen et al (1990), and provide an excellent foundation for the development of dynamic models of the types discussed above. One of the advantages of models that are derived in this way is that they to a very high degree can be based directly on biological theory and prior knowledge.

In particular they can be used for evaluating theoretical models and hypothesis. Once a deterministic model given by a multi-dimensional differential equation has been established the first step in the statistical analysis is to make some stochastic interpretation of the deterministic model in order to describe the system noise which is inevitably present in idealized models for physical phenomena. The system noise represents the discrepancy between the deterministic model and the true physical processes, and accounts in particular for the accumulated influence of factors not included in the model and for inaccuracies of exogeneous variables in the deterministic model. This system noise can readily be modeled by converting the multi-dimensional differential equation into a multi-dimensional stochastic differential equation, see section 2, 3 and 4. The physical phenomena under consideration is then modeled by a multi-dimensional diffusion process, and the statistical analysis should ideally be based on discrete-time observations of the diffusion process, see Pedersen (1993a,c), Bibby & Sørensen (1994) and Pedersen (1994a,b). However, a common characteristic of such models is that they very often contain dynamic variables that can not be observed, that is the discrete observations of the diffusion process are incomplete. In fact the possibility for extracting information about unobservable processes from observable processes and a model that relates them is often a major motivation for considering such models. In addition the partial observations may contain measurement errors. This incompleteness of the discrete observations of the diffusion process complicates the statistical analysis considerably, but for Gaussian diffusion processes it is still possibly to perform maximum likelihood estimation of unknown parameters, model validation, forecasting, prediction etc., see Madsen & Holst (1991), Madsen & Melgaard (1991), Melgaard et al (1992), Pedersen (1993b,1994c) and section 5. Furthermore, it is possible to reconstruct the unobserved diffusion variables in an optimal way, see Jazwinski (1970), Åström (1970), Pedersen (1994c) and section 5. Gaussian diffusion processes are for instance derived from linear differential equations by the method described above, that is by adding a system noise to the differential equation in such a way that a stochastic differential equation is obtained.

The statistical interpretation of the multi-dimensional differential equation outlined above assumes that the differential equation may be regarded as an idealized model for each path of a diffusion process. This means that the differential equation to a very high degree must be causal and specified in detail. Such models are usually intended to be used for scientific purposes. However, in some cases the differential equation is rather to be interpreted as a model for the average evolution of some system. Such models are often relevant for prognosis purposes, see Jensen et al (1994) and section 2 and 3. In the latter case the differential equation should be interpreted as a model for the evolution of the expectation of a stochastic process, and again the interpretation may readily be performed within the framework of diffusion processes, see section 4. The relevant statistical interpretation of the differential equation is thus highly dependent on the scientific nature of the model, and it may affect the statistical possibilities considerably, see section 4. However, for linear differential equations the pathwise interpretation of the differential equation outlined in the previous paragraph implies in addition that the expectations of the derived diffusion process solves the differential equation. The pathwise interpretation of linear differential equations also possesses other desirable properties, see section 4. Hence, Gaussian diffusion processes derived from linear differential equations seem attractive both from a modeling point of view and from a statistical point of view.

In the present paper we review a number of applications of Gaussian diffusion processes in freshwater ecology. The Gaussian diffusion process is in all cases derived from a multi-dimensional linear differential equation, and the observations are in all cases incomplete in the sense that one of the coordinates of the diffusion process is unobservable. In section 2 we consider two Gaussian diffusion models for the dynamics and interactions of phosphorus processes in a lake. The first model is quite detailed and is based on four related differential equations for the evolution of four phosphorus processes in a lake. It is intended to be used for scientific purposes as discussed in this section. The second model is an application of the first model. It is a simple prognosis model that is designed for making predictions about the effects of reducing the external loading of phosphorus to a lake. Both models are developed by means of an extensive data set from Lake Søbygaard, and the latter model is furthermore applied to some of the lakes in the monitoring program. A detailed account of these models can be found in Jensen et al (1994) and Pedersen (1994c). In section 3 we consider two Gaussian diffusion models for the interactions and dynamics of nitrogen processes in a lake. From a statistical point of view these models are very similar to the phosphorus models in section 2. Section 4 contains a discussion of the two statistical interpretations mentioned above. In particular it is illustrated how the statistical interpretation may affect the statistical possibilities considerably when the differential equation is non-linear and the discrete data are incomplete. Finally, a brief review of the available statistical methods for the analysis of incompletely discretely observed Gaussian diffusion processes is given in section 5, see also Madsen & Holst (1991), Madsen & Melgaard (1991), Melgaard et al (1992) and Pedersen (1993b,1994c).

2 Phosphorus models

Basically the amount of phosphorus in a lake can be divided into four parts. In the water, the amount of phosphorus is the sum of phosphorus in a dissolved form (P_d) and in an organic form (P_o). In the sediment, the amount of phosphorus can be divided into a relatively small but biologically very active part ($\overline{P}_s$) located in the upper layers of the sediment and a relatively large but biologically less active part ($\underline{P}_s$) located in the lower layers of the sediment. Under certain biological constraints the dynamics and interactions of $P_d, P_o, \overline{P}_s$ and $\underline{P}_s$ can be described by a four-dimensional differential equation, see Pedersen (1994c). However, since the available data for $\overline{P}_s$ and $\underline{P}_s$ are very scarce it is impossible to treat them both as dynamic variables in the model, see Pedersen (1994c). Treating instead $\underline{P}_s$ as an exogeneous variabel the following three-dimensional differential equation for the dynamics and interactions of P_d, P_o and $\overline{P}_s$ was derived in Pedersen (1994c)

$$\begin{aligned}
\frac{dP_d(t)}{dt} &= \frac{Q(t)}{V}(P_{d,in}(t) - P_d(t)) - P(t) + R(t) + \overline{F}(t) + \underline{F}(t) - \underline{S}(t) \\
\frac{dP_o(t)}{dt} &= \frac{Q(t)}{V}(P_{o,in}(t) - P_o(t)) + P(t) - R(t) - \overline{S}(t) \\
\frac{d\overline{P}_s(t)}{dt} &= \overline{S}(t) - \overline{F}(t),
\end{aligned}$$

where $P_{d,in}$ is the amount of dissolved phosphorus at the inlet, $P_{o,in}$ is the amount of organic phosphorus at the inlet, Q is the water flow at the inlet, V is the water volume of the lake (a known constant), P is the rate of organic phosphorus production, R is the rate of organic phosphorus respiration, $\overline{F}$ is the rate of phosphorus release from the upper sediment, $\underline{F}$ is the rate of phosphorus release from the lower sediment, $\underline{S}$ is the rate of dissolved phosphorus sedimentation to the lower sediment and $\overline{S}$ is the rate of organic phosphorus sedimentation to the upper sediment. From the available data set for the statistical analysis of this model one may derive, see Pedersen (1994c), daily measurements from Lake Søbygaard from the 1st of January 1985 until the 16th of December 1991 of the dynamic variables P_d and P_o, the exogeneous variables $P_{d,in}, P_{o,in}$ and Q, the rate P, the water temperature (T_w), the zooplankton biomass (Z) in the water, the pH value (pH) in the water, the amount of nitrate (Ni) in the water and the amount of iron (Fe) at the inlet. The observed rate P is treated as an exogeneous variable in the model whereas the remaning unobserved rates are related to observed exogeneous variables by

$$\begin{aligned}
R(t) &= \alpha_R \cdot 1.08^{T_w(t)-20} \cdot P_o(t) && (1)\\
\overline{F}(t) &= \alpha_{\overline{F}} \cdot 1.08^{T_w(t)-20} \cdot \overline{P}_s(t) \cdot K(t) && (2)\\
\underline{F}(t) &= \alpha_{\underline{F}} \cdot 1.08^{T_w(t)-20} \cdot \underline{P}_s(t) \cdot K(t) && (3)\\
\overline{S}(t) &= \alpha_{\overline{S}} \cdot (P_o(t) - Z(t)) && (4)\\
\underline{S}(t) &= \alpha_{\underline{S}} \cdot (\frac{Q(t)}{V} Fe(t))^{\beta_{Fe}} \cdot P_d(t), && (5)
\end{aligned}$$

where

$$K(t) = \exp(\beta_{pH} \cdot pH(t) - \beta_{Ni} \cdot Ni(t))$$

and $\alpha_R, \alpha_{\overline{F}}, \alpha_{\underline{F}}, \alpha_{\overline{S}}, \alpha_{\underline{S}}, \beta_{Fe}, \beta_{pH}$ and β_{Ni} are unknown parameters. We shall not discuss the biological justifications of these relations in any detail, but merely emphasize that they are all to a high degree in accordance with biological theory, prior knowledge and experience. Hence the model should be interpreted as an idealized model for the pathwise evolution of some three-dimensional stochastic process, cf. the discussion about statistical interpretations in section 1. Since all rates in the deterministic model obtained by inserting the relations (1)–(5) in the three-dimensional differential equation above are linear in P_d, P_o and $\overline{P}_s$ we can write the model as

$$\frac{dx_t}{dt} = A_t x_t + a_t, \tag{6}$$

where $x_t = (P_d(t), P_o(t), \overline{P}_s(t))^T$, A is a 3×3 matrix function of exogeneous variables and unknown parameters and a is a 3×1 vector function of exogeneous variables and unknown parameters, see Pedersen (1994c) for details. Here ^{T} denotes matrix transposition. Adding to this deterministic model a simple one-parameter system noise, see Pedersen (1994c), we obtain the following stochastic differential equation

$$dX_t = (A_t X_t + a_t)dt + \sigma I_3 dW_t, \tag{7}$$

where σ is a positive constant, I_3 is the three-dimensional identity matrix and W is a three-dimensional Wiener process. In this way the processes P_d, P_o and $\overline{P}_s$ are modeled by a three-dimensional Gaussian diffusion process $X_t = (P_d(t), P_o(t), \overline{P}_s(t))^T$ with short-term pathwise evolution given by (6) and with expectations $(E(X_t))_{t \geq 0}$ satisfying the differential equation (6), see section 4. Notice that the discrete observations of X given by the measurements of P_d and P_o are incomplete in the sense that the third coordinate of X is unobserved. This three-dimensional Gaussian diffusion model with incomplete discrete observations given by the daily data from Lake Søbygaard of P_d and P_o has been analyzed in Pedersen (1994c) by means of the statistical methods described in Pedersen (1993b,1994c), see also section 5 for a brief review. Two important qualities of these statistical methods are that they do not rely on any restrictions on A or the dimension of the Gaussian diffusion process, see e.g. Arnold (1974), Madsen & Holst (1991), Madsen & Melgaard (1991), Melgaard et al (1992) and Pedersen (1993b,1994c), and that they do not require the calculation of the exponential of matrices, see e.g. Madsen & Holst (1991), Madsen & Melgaard (1991) and Melgaard et al (1992). In particular this means that the same statistical methods can be applied to all Gaussian diffusion proceses satisfying the regularity conditions in section 5. As emphasized in Pedersen (1993b,1994c) the inital value of $\overline{P}_s$ is treated as an unknown parameter. On a daily basis the estimated model for the expectations of P_d and P_o has difficulties in capturing the high values and fast fluctuations of P_d and P_o in the summer time, see Pedersen (1994c). Some of the explanation for this deficiency of the model may be that the observation frequency of one day is too low in the summer time to capture the fluctuations of P_d and P_o more accurately than accomplished by the present model, since most biological processes are highly active in the summer time and may vary considerably during a single day. Future developments of the model will clarify the significance of this point, since there are still several biologically justifiable ways of altering the present model. Both the estimated values of the unknown parameters and the reconstructed values of the unobserved phosphorus process $\overline{P}_s$ in the sediment, see Pedersen (1994c), are in accordance with prior biological expectations. The most important quality of the model for the present purposes, see Jensen et al (1994), is however that the model for the expectations of P_d and P_o essentially captures the seasonal variations of P_d and P_o over a long period of time, which to a certain extend qualifies the model for making predictions, see Pedersen (1994c). As an example where predictions are relevant, assume that the model is basically correct and that the exogeneous variable pH to a certain degree can be controlled. Then the model claims that a reduced value of pH in the water *ceteris paribus* reduces the internal loading $\overline{F} + \underline{F}$ of phosphorus to the lake, cf. relation (2) and (3). As discussed in section 1 it would certainly be of interest to be able to predict the effects of such and intervention. However, predictions in the presented model requires that reliable future values of the exogeneous variables $P_{d,in}, P_{o,in}, Q, T_w, P, Z, pH, Ni, Fe$ and $\underline{P}_s$ can be obtained. Concerning Q and T_w this is no problem, since they exhibit more or less the same seasonal variations every year and are independent of all other biological processes in the model. Also the exogeneous variables $P_{d,in}$ and $P_{o,in}$ are independent of all biological processes in the lake, and the total external loading $P_{d,in} + P_{o,in}$ of phosphorus to the lake can to a certain degree be controlled at the sewage plants and by other environmental efforts. Hence, the exogeneous variable $P_{d,in} + P_{o,in}$ can to some extend be regarded as a future environmental control variable. The future evolution of the remaining exogeneous variables in the model is however closely related to the evolution of the processes P_d, P_o and $\overline{P}_s$ and should therefore, ideally, be included in a simultaneous model, see also section 3.

In the remainder of this section we shall consider a simple Gaussian diffusion model that is designed for making predictions about the effects of reducing the external loading of phosphorus to a lake. The model was derived in Pedersen (1994c) as an application of the model presented above, and it is intended to describe the average evolution of the amount of phosphorus in the water and in the sediment. To ensure the possibility for making predictions it uses only the exogeneous variables Q, T_w and $P_{d,in} + P_{o,in}$, cf. the discussion preceeding this paragraph. Hence, the model is based on rather crude approximations of the rates (1)–(5), and it does not distinguish between the two forms of phosphorus in the water nor between the two phosphorus processes in the sediment. The model should not be considered as a model for the pathwise evolution of the involved processes, but rather as a biologically interpretable model that roughly captures the trend or average evolution of the processes. Put

$$\begin{aligned}
P_w &= P_d + P_o \\
P_s &= \overline{P}_s + \underline{P}_s \\
P_{w,in} &= P_{d,in} + P_{o,in} \\
S &= \overline{S} + \underline{S} \\
F &= \overline{F} + \underline{F}.
\end{aligned}$$

Then the following two-dimensional deterministic model for the dynamics and interactions of the phosphorus processes P_w and P_s was derived in Pedersen (1994c)

$$\frac{dP_w(t)}{dt} = \frac{Q(t)}{V}(P_{w,in}(t) - P_w(t)) + F(t) - S(t) \tag{8}$$

$$\frac{dP_s(t)}{dt} = S(t) - F(t), \tag{9}$$

where the sedimentation rate S and the release rate F are approximated by

$$S = \alpha_S \cdot P_w \tag{10}$$

$$F = \alpha_F \cdot \theta_F^{T_w - 20} \cdot P_s \tag{11}$$

for some unknown constants α_S, α_F and θ_F. By inserting the relations (10) and (11) in the equations (8) and (9) we see that the dynamics and interactions of P_w and P_s in this deterministic model are described by a two-dimensional linear differential equation of the form (6), where A is a 2×2 matrix function of Q, T_w and unknown parameters and a is a 2×1 vector function of $Q, P_{w,in}$ and unknown parameters, see Pedersen (1994c) for details. Taking into account that the relation (11) is by far the most noisy of the relations (10) and (11), the following stochastic differential equation for $X = (P_w, P_s)^T$ was derived in Pedersen (1994c)

$$dX_t = (A_t X_t + a_t)dt + B_t dW_t, \tag{12}$$

where W is a two-dimensional Wiener process and B satisfies the matrix equation

$$B_t B_t^T = \begin{bmatrix} \sigma^2 + f(T_w(t))^2 & -f(T_w(t))^2 \\ -f(T_w(t))^2 & \sigma^2 + f(T_w(t))^2 \end{bmatrix}, \; f(T_w) = \sigma_F \cdot \psi_F^{T_w - 20}$$

for all $t \geq 0$. Here σ, σ_F and ψ_F are unknown positive constants. The discrete observations of the Gaussian diffusion process X are incomplete, since the second coordinate (P_s) of X is unobserved. This model was also analyzed in Pedersen (1994c) by means of the data from Lake Søbygaard and the statistical methods described in Pedersen (1993b,1994c) and section 5. The performance of the model was found to be quite satisfactory for the purpose of making predictions. For illustration the predicted future evolutions of P_w and P_s in response to different proportional future reductions of the external loading $P_{w,in}$ to the lake was calculated in Pedersen (1994c). These predictions basically show that the evolution of P_w and P_s will continue as in the observation period, that is an extremely slow decrease in P_w due to a large releasement of phosphorus from the sediment, and a steady decrease in P_s due to the reduction of $P_{w,in}$. These conclusions are completely in accordance with biological knowledge and expectations, saying that no dramatic positive short-term effect on the amount of phosphorus in the water can be obtained by further reduction of the external loading of phosphorus to the lake, due to the high internal loading of phosphorus from the large amount of phosphorus in the sediment that has accumulated in the past. Accordingly, one of the important topics in present applied lake research is to study the possibilities for reducing the internal loading of phoshorus in a lake.

One of the motivations for developing simple prognosis models as the one just considered is the possibility for applying them to the lakes in the Monitoring Program of the Danish Action Plan for the Aquatic Environment, cf. section 1. In this monitoring program, monthly measurements of several quantities including $P_w, P_{w,in}, Q$ and T_w (but not P_s) are made in 37 Danish lakes, and in the following we shall briefly discuss some of the problems encountered when diffusion models are applied to these lakes and data. For illustration we apply the prognosis model above to four lakes in the monitoring program which resemble Lake Søbygaard with respect to a number of characteristics, see also Jensen et al (1994). The most serious problem in applying diffusion models to the lakes in the monitoring program is that the data are very crude (monthly measurements, P_s unobserved). In particular this implies that only very few unknown parameters can be estimated for each lake. As a first step in reducing the number of unknown parameters in the prognosis model above we simplify it by fixing the noise parameters $\sigma_F^2 = \psi_F = 0$. Reestimating the remaining parameters except for α_S and $P_s(0)$, see Pedersen (1994c), on the data set from Lake Søbygaard implies no dramatic changes, and for the purpose of prediction the model is essentially unaltered. The estimates of the unknown parameters are shown in table 1. For the lakes in the monitoring program we should ideally estimate the unknown parameters $P_s(0), \alpha_S, \alpha_F, \theta_F$ and σ^2, but unfortunately this is only possible for a very small number of lakes. Instead we fix the values of α_S, α_F and θ_F at the estimated values from Lake Søbygaard shown in table 1, and estimate only $P_s(0)$ and σ^2 for each lake. Obviously this constrains the application of the model to lakes which are similar to Lake Søbygaard with respect to a number of characteristics, see Jensen et al (1994). The results for the four lakes in the monitoring program considered here are shown in table 1. In this table, N denotes the value of the iterated Euler-approximation parameter used in the estimation procedure, see Pedersen (1993b,1994c) or section 5. The value of N for the four lakes in the monitoring program is 30 times the value of N for Lake Søbygaard, which seems reasonable since the observation frequency for the lakes in the monitoring program is 30 times the observation frequency for Lake Søbygaard. Plots of the observed values of P_w and the estimated expectations of P_w are shown in figure 1 for each of the four lakes in the monitoring program, and corresponding plots of the reconstructed values of P_s and the estimated expectations of P_s are shown in figure 2. Quantile plots of the standardized residuals for each of the four lakes in the monitoring program are shown in figure 3. Even though the performance of the estimated expectations in figure 1 and the behaviour of the standardized residuals in figure 3 is quite satisfactory, considering the crudeness of the model and the data, it is disturbing that there is a dramatic difference between the reconstructed values of P_s and the estimated expectations of P_s shown in figure 2. Theoretically, P_s and the reconstruction of P_s have the same expectation, see Pedersen (1994c) or section 5. The explanation for this deficiency may very well be that the values of α_S, α_F and θ_F taken from Lake Søbygaard are inappropriate, since the problem in figure 2 disappears when all parameters are estimated for the few lakes in the monitoring program for which all parameters can be estimated, even though the dramatic differences in figure 2 also appear for these lakes when the model is applied as above. Compare for instance the plots in figure 1–3 for the lakes entitled Borup and Kilen with the corresponding plots in figure 4, wherein the values of all unknown parameters are estimated seperately for each lake. The reconstructed values of P_s in figure 4 are however rather unrealistic from a biological point of view.

Lake	N	$P_s(0)$	$\alpha_S \cdot 10^2$	$\alpha_F \cdot 10^4$	θ_F	$\sigma^2 \cdot 10^4$
Søbygaard	25	151	9.00	6.46	1.11	6.87
Borup	750	35	-	-	-	2.41
Fugle	-	140	-	-	-	96.8
Jels	-	107	-	-	-	15.5
Kilen	-	171	-	-	-	62.3

Table 1: Estimates of the unknown parameters in the statistical model given by incomplete discrete observations of the Gaussian diffusion process defined by (12) with the values of σ_F^2 and ψ_F fixed at zero.

3 Nitrogen models

In analogy with the development of the phosphorus models in the previous section, incompletely discretely observed Gaussian diffusion processes also appear in a natural way in the development of statistical models for the interactions and dynamics of nitrogen processes in a lake. In fact such models can be based on multi-dimensional

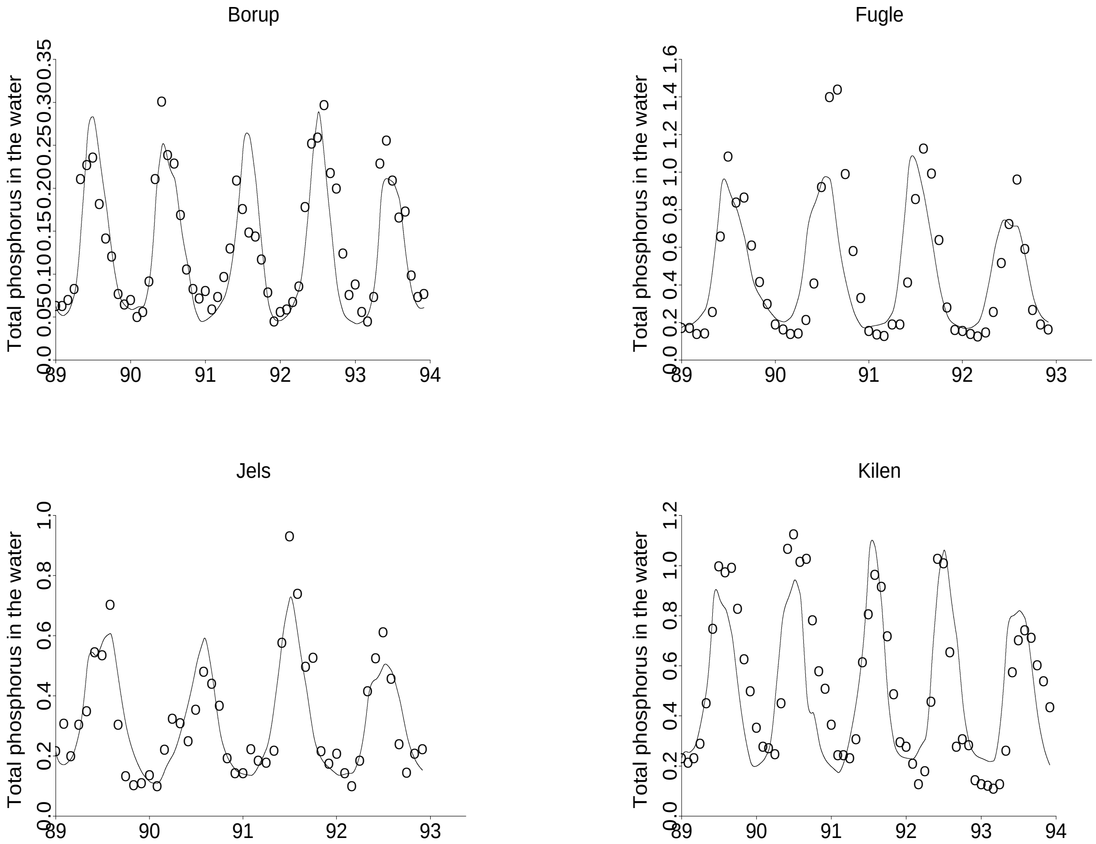

Figure 1: The observed values (O) of P_w from four of the lakes in the monitoring program, and the corresponding estimated expectations (solid line) of P_w for the statistical model given by incomplete discrete observations of the Gaussian diffusion process defined by (12) with the respective estimates in table 1.

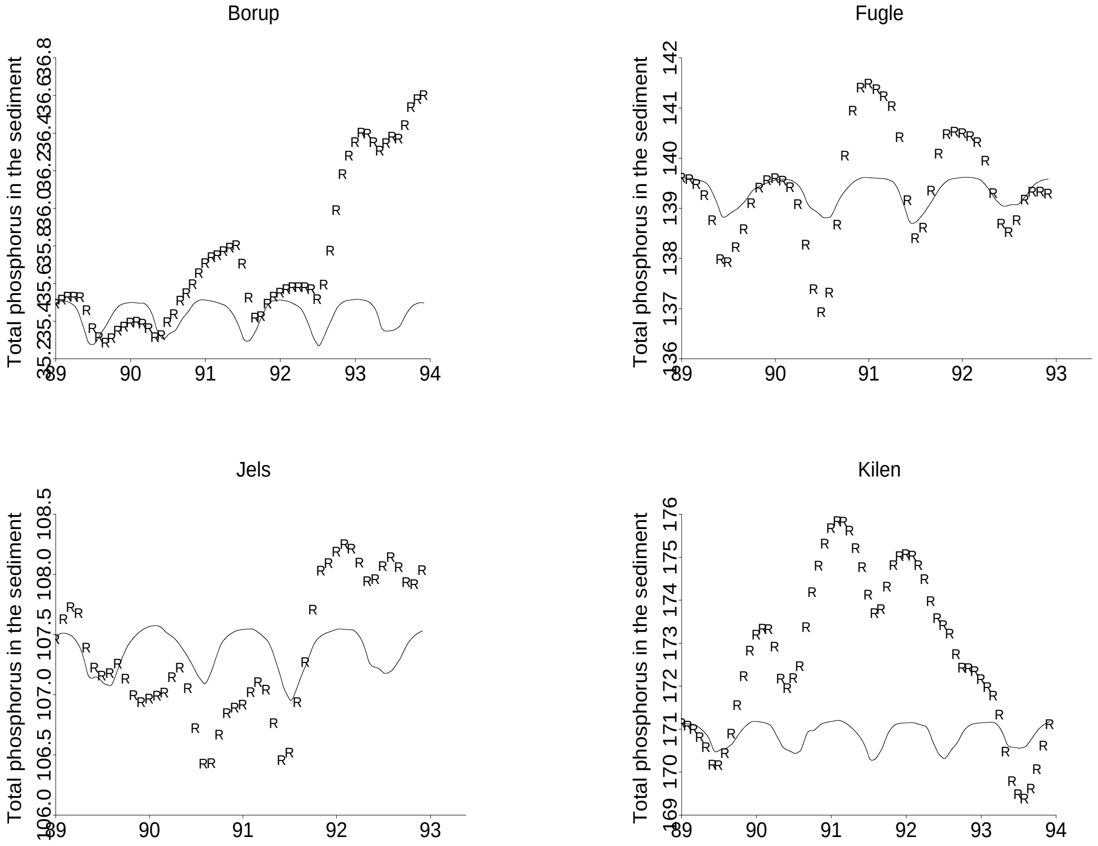

Figure 2: The reconstructed values (R) of P_s in four of the lakes in the monitoring program, and the corresponding estimated expectations (solid line) of P_s for the statistical model given by incomplete discrete observations of the Gaussian diffusion process defined by (12) with the respective estimates in table 1.

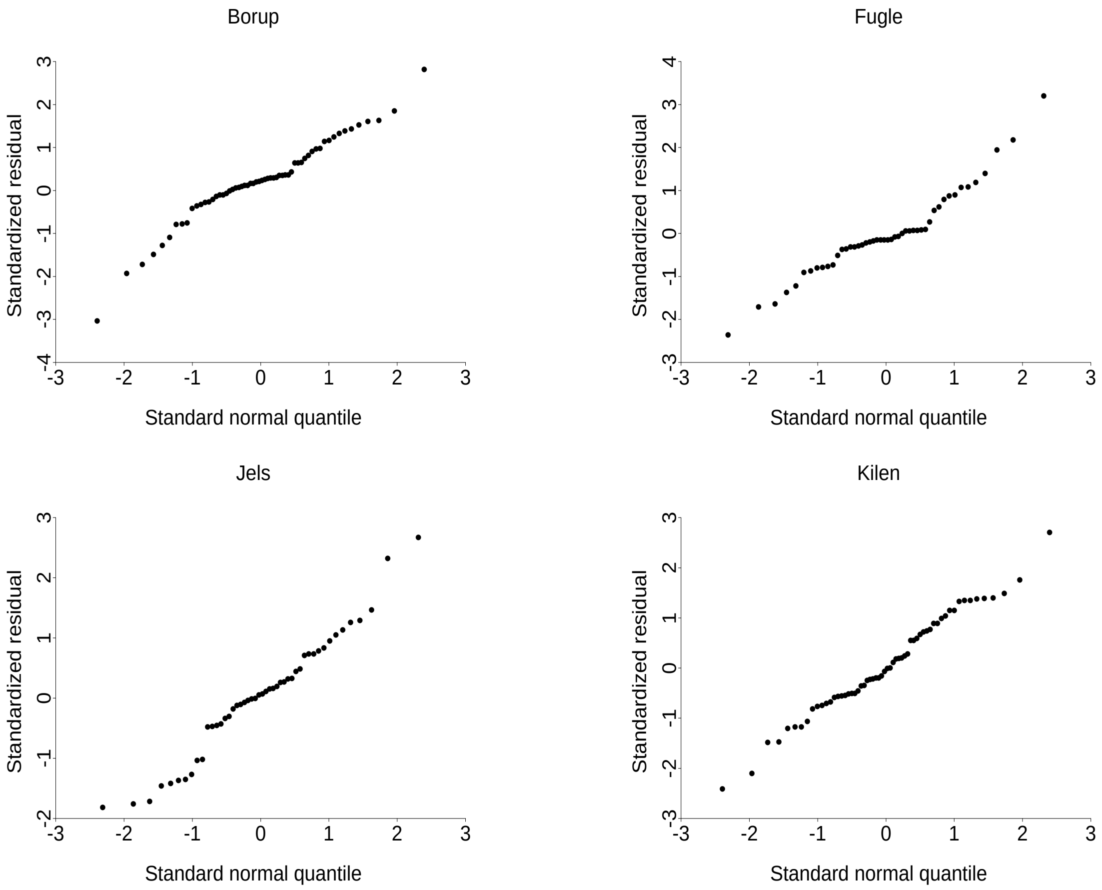

Figure 3: Quantile plots for four of the lakes in the monitoring program of the standardized residuals in the statistical model given by incomplete discrete observations of the Gaussian diffusion process defined by (12) with the respective estimates in table 1.

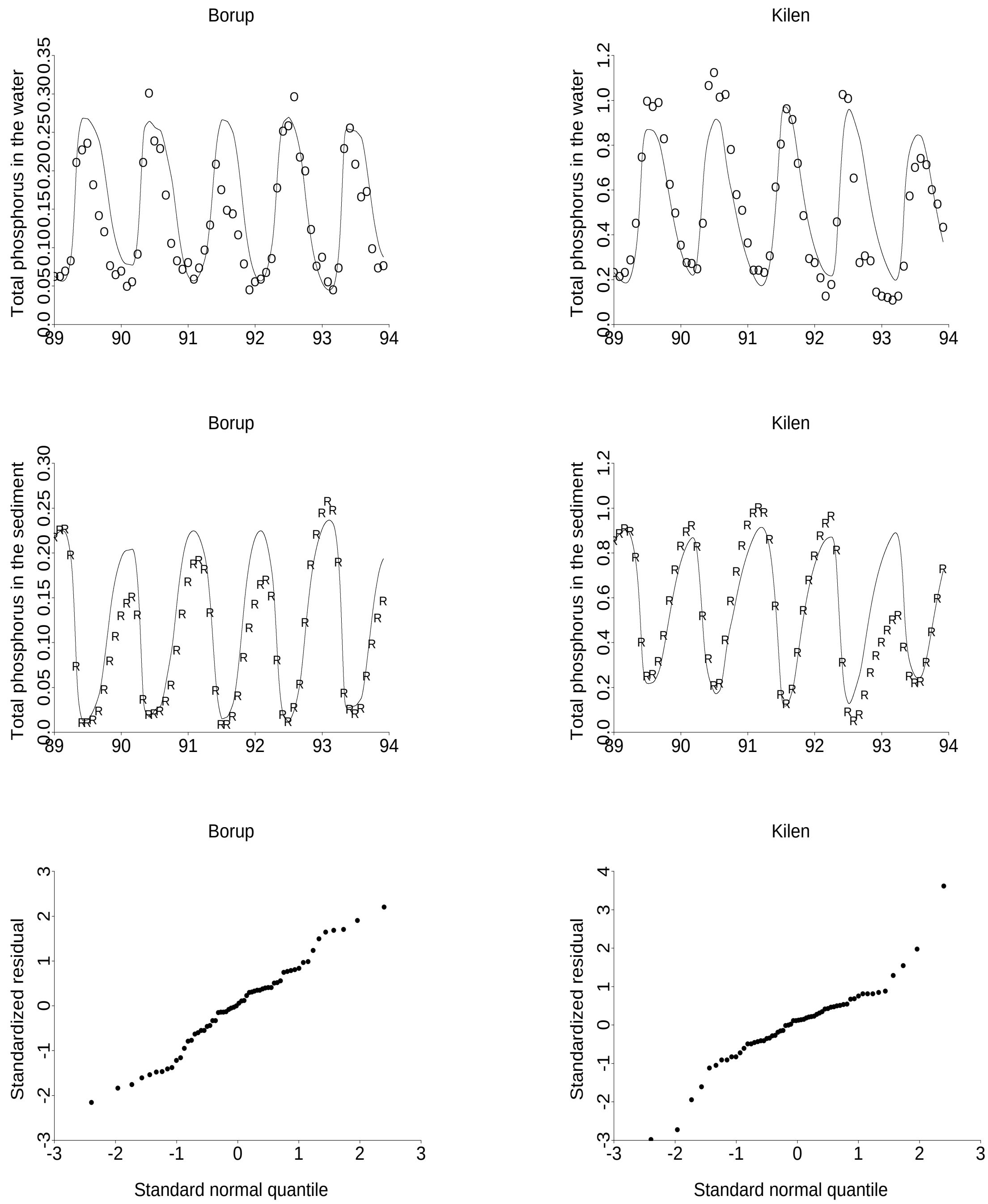

Figure 4: Plots as in figure 1–3 for two lakes in the monitoring program for which lake specific estimates can be obtained for all unknown parameters in the statistical model given by incomplete discrete observations of the Gaussian diffusion process defined by (12).

ordinary differential equations that are very similar to those considered in section 2, and from a statistical point of view the derived Gaussian diffusion models are therefore "identical" to those derived in section 2. Hence, in this section we shall be very brief and confine ourselves to modeling aspects, thus presenting no data analysis.

The amount of nitrogen in a lake can be divided into three parts. In the water, the amount of nitrogen is divided into nitrogen in a dissolved form (N_d) and in an organic form (N_o), wheras the amount of nitrogen in the sediment (N_s) is undivided. Under certain biological constraints the interactions and dynamics of these three nitrogen processes can be described by

$$\begin{aligned}
\frac{dN_d(t)}{dt} &= \frac{Q(t)}{V}(N_{d,in}(t) - N_d(t)) - P(t) + R(t) - D_d(t) + F(t) \\
\frac{dN_o(t)}{dt} &= \frac{Q(t)}{V}(N_{o,in}(t) - N_o(t)) + P(t) - R(t) - S(t) \\
\frac{dN_s(t)}{dt} &= S(t) - D_s(t) - F(t),
\end{aligned}$$

where $N_{d,in}$ is the amount of dissolved nitrogen at the inlet, $N_{o,in}$ is the amount of organic nitrogen at the inlet, Q is the water flow at the inlet, V is the water volume of the lake (a known constant), P is the rate of organic nitrogen production, R is the rate of organic nitrogen respiration, D_d is the rate of dissolved nitrogen denitrification, F is the rate of nitrogen release from the sediment, S is the rate of organic nitrogen sedimentation and D_s is the rate of denitrification of nitrogen from the sediment. The statistical analysis of this model has been based on a data set from Lake Søbygaard which contains measurements of the rate P. This rate has therefore been treated as an exogeneous variable in the model, whereas the remaining rates are modeled by

$$\begin{aligned}
R(t) &= \alpha_R \cdot 1.08^{T_w(t)-20} \cdot N_o(t) &(13)\\
F(t) &= \alpha_F \cdot 1.08^{T_w(t)-20} \cdot N_s(t) &(14)\\
S(t) &= \alpha_S \cdot (N_o(t) - Z(t)) &(15)\\
D_d(t) &= \delta_d \cdot 1.08^{T_w(t)-20} \cdot N_d(t) &(16)\\
D_s(t) &= \delta_s \cdot 1.08^{T_w(t)-20} \cdot N_s(t), &(17)
\end{aligned}$$

where the exogeneous variables T_w and Z are the water temperature and the zooplankton biomass in the water, respectively. The unknown parameters of interest are $\alpha_R, \alpha_F, \alpha_S, \delta_d$ and δ_s. The structure of this deterministic model is quite similar to that of the three-dimensional model considered in section 2. Another similarity between these two models is that the data are incomplete, since N_s can not be observed. However, the multi-dimensional differential equation obtained by combining the relations (13)–(17) with the three-dimensional differential equation above is linear in the dynamic variables N_d, N_o and N_s, and the statistical analysis can thus be performed as in section 2 by means of the methods described in Pedersen (1993b,1994c) and section 5.

The development of the presented nitrogen model is still at a preliminary stage, exactly as the development of the three-dimensional phosphorus model in section 2. The two models are however close related. In fact the exogeneous variable Ni in the phosphorus model is equal to the dynamic variable N_o in the nitrogen model. Also some of the rates in the two models are biologically related. Therefore the two models should ideally be embedded in a simultaneous model. It is however necessary first to understand the presented marginal models in more detail.

As in section 2 it is also of interest to develop simple prognosis models for the interactions and dynamics of the amount of nitrogen $N_w = N_d + N_o$ in the water and the amount of nitrogen in the sediment. Such a model can be obtained by adding the differential equations for N_d and N_o above. In this way the following model is obtained

$$\begin{aligned}\frac{dN_w(t)}{dt} &= \frac{Q(t)}{V}(N_{w,in}(t) - N_w(t)) - D_d(t) + F(t) - S(t) \\ \frac{dN_s(t)}{dt} &= S(t) - D_s(t) - F(t)\end{aligned}$$

with $N_{w,in} = N_{d,in} + N_{o,in}$. For the purposes of predictions, the rates in this model must be specified exclusively in terms of the dynamic variables N_w and N_s and the exogeneous variables Q, T_w and $N_{w,in}$, analogously to the prognosis model in section 2.

4 Statistical interpretations of ordinary differential equations

The first step in the statistical analysis of a deterministic model given by a multidimensional differential equation is to make a statistical interpretation of the model. This interpretation depends on the scientific nature of the model. If the deterministic model is highly causal and carefully specified in detail the appropriate interpretation seems to be that the differential equation is an idealized model for the pathwise evolution of some stochastic process. Such models are particularly relevant for scientific purposes, cf. the three-dimensional models in section 2 and 3. However, for other purposes the differential equation should rather be regarded as a specification of a function that roughly captures the trend or average evolution of some processes, but still in a scientifically interpretable way. This is particularly relevant for prognosis purposes where the models can not be specified in detail for the reasons discussed in section 2, cf. the two-dimensional prognosis models in section 2 and 3. In such cases the appropriate interpretation is that the differential equation is an idealized model for the evolution of the expectation of some stochastic process. In this section we discuss the implications with respect to modeling aspects and statistical possibilities of this distinction between statistical interpretations.

Consider a deterministic model given by the d-dimensional differential equation

$$\frac{dx_t}{dt} = b(t, x_t; \psi), \tag{18}$$

where ψ is a vector of unknown parameters. For the present purposes we assume that (18) can be solved either explicitly or by means of some numerical procedure for each fixed value of ψ, see e.g. Zwillinger (1992). The statistical interpretation of this deterministic model then amounts to defining some d-dimensional stochastic process X with some relation to (18) that depends on the scientific nature of (18). From the scientific problem under study it is usually clear that X should have continuous paths, cf. section 2 and 3, and we shall furthermore assume that X possesses the Markov property. In the pathwise interpretation one might argue that the Markov property is an appropriate stochastic interpretation of the fact that an ordinary differential equation is a model for first order reactions. Thus X is a diffusion process and can accordingly be defined by specifying some stochastic differential equation. Formally, a stochastic differential equation corresponding to a pathwise interpretation of (18) can be derived by adding a weighted "continuous-time white noise process" to (18), that is

$$\frac{dX_t}{dt} = b(t, X_t; \psi) + \sigma(t, X_t; \psi, \eta)\frac{dW_t}{dt},$$

where W is an r-dimensional Wiener process, σ is a $d \times r$ matrix function depending on ψ and some vector η of noise parameters. The precise interpretation of this stochastic equation is the stochastic differential equation

$$dX_t = b(t, X_t; \psi)dt + \sigma(t, X_t; \psi, \eta)dW_t. \tag{19}$$

Essential properties of (19) from a modeling point of view are that

$$E_{\psi,\eta}(X_t | X_s = x) \simeq x + (t - s)b(s, x; \psi) \tag{20}$$

for t close to s and that

$$\frac{dE_{\psi,\eta}(X_t | X_s = x)}{dt} = E_{\psi,\eta}(b(t, X_t; \psi) | X_s = x), \ t \geq s \tag{21}$$

for all $s \geq 0$ and $x \in \boldsymbol{R}^d$. The property (20) says that (18) describes the short-term average evolution of X starting from any point at any time, which is the essential property of X in a pathwise interpretation. As a special case of property (21) we see that the expectations $(E(X_t))_{t\geq 0}$ of X do not satisfy (18) unless b is linear in the state variable. As discussed earlier, cf. also section 1, this may in some cases be the essential requirement for X. Insisting on that the expectations $(E(X_t))_{t\geq 0}$ of X should satisfy (18) one may instead consider the stochastic differential equation

$$dX_t = b(t, \mu(t;\psi); \psi)dt + \sigma(t, X_t; \psi)dW_t, \tag{22}$$

where W, σ and η are as before, and μ is a deterministic function defined by

$$\frac{d\mu(t;\psi)}{dt} = b(t, \mu(t;\psi); \psi).$$

In this interpretation the expectations $(E(X_t))_{t\geq 0}$ satisfy (18) by construction even if b is non-linear in the state variable, since

$$E_{\psi,\eta}(X_t) = \mu(t;\psi).$$

Thus if b is non-linear in the state variable, and the scientific nature of (18) implies that (18) should be satisfied by $(E(X_t))_{t\geq 0}$, then (22) seems to be an appropriate interpretation of (18). However, this interpretation does not possess the property that X evolves on the average according to the same pattern, defined by b, whenever it is reinitialized at some time s. Indeed

$$\frac{dE_{\psi,\eta}(X_t|X_s = x)}{dt} = b(t, E_{\psi,\eta}(X_t); \psi)$$

for all $0 \leq s < t$ and $x \in \boldsymbol{R}^d$. Notice also that X defined by (22) does not satisfy (19) if b is linear in the state variable. In conclusion, when b is linear in the state variable the pathwise interpretation (19) seems in any case most attractive from a modeling point of view. However, when the differential equation (18) is to be interpreted as a specification of the average evolution of a stochastic process and b is non-linear in the state variable then the interpretation given by (22) seems more appropriate. From now on we denote by X^p and X^e the diffusion processes defined by (19) and (22), respectively.

As an example of an application where the interpretation given by X^e is most appropriate recall the prognosis model in section 2 given by the two-dimensional differential equation

$$\begin{aligned}
\frac{dP_w(t)}{dt} &= \frac{Q(t)}{V}(P_{w,in}(t) - P_w(t)) + F(t) - S(t) \\
\frac{dP_s(t)}{dt} &= S(t) - F(t),
\end{aligned}$$

where the sedimentation rate S and the release rate F are approximated by

$$\begin{aligned}
S &= \alpha_S \cdot P_w \\
F &= \alpha_F \cdot \theta_F^{T_w - 20} \cdot P_s.
\end{aligned}$$

The appropriate interpretation of this deterministic model is that it is a model for the evolution of the expectation of some two-dimensional stochastic process. As noted in section 2 this approximate relation for the release rate F is rather crude. However, in order to preserve the possibility for making predictions it is important not to include additional exogeneous variables in the specification of F. Instead one might try to improve the specifcation of F by allowing it to depend non-linearly on P_s. In Jensen et al (1994) the following improved relation was proposed

$$F = \alpha_F \cdot \theta_F^{T_w - 20} \cdot P_s \cdot g(P_s),$$

where g is some function that within a certain range of values of P_s is given by

$$g(P_s) = \exp(-\lambda / P_s)$$

for some known/unknown lake specific parameter λ. This modification of the model implies that the obtained differential equation is non-linear in the state variable $(P_w, P_s)^T$.

Concerning the choice of the diffusion coefficient σ it is instructive to see how a given choice of σ implies quite different probabilistic properties of X^p and X^e. Consider the one-dimensional homogeneous case where b is linear in the state variable and σ is constant, that is

$$\begin{aligned} b(t, x; \psi, \eta) &= Ax + a \\ \sigma(t, x; \psi, \eta) &= \sigma > 0 \end{aligned}$$

for some constants A, a and σ. We assume that $A < 0$. In this case it is possible to see how the two interpretations imply different evolutions of the variance of the diffusion processes. Indeed

$$V(X_t^p) = \sigma^2 \cdot \frac{e^{2At} - 1}{2A} \longrightarrow \frac{\sigma^2}{-2A}$$

as $t \longrightarrow \infty$ whereas

$$V(X_t^e) = \sigma^2 t \longrightarrow \infty$$

as $t \longrightarrow \infty$. The influence of b on the variance of X^p thus ensures in this case that the variance does not explode as time tends to infinity, whereas this must be ensured for X^e entirely by the specification of σ, e.g. by putting

$$\sigma(t, x; \psi, \eta) = \sigma e^{-\alpha t}$$

for some positive constant α. On the other hand it is in the general case very often impossible to deduce the implications on the variance of X^p of a given choice of σ, since the variance is usually unknown and depends in all cases in a complicated way on b. For X^e the implications on the variance of a given choice of σ are usually easier to determine. Assume for instance in the general case that σ has been chosen independent of the state variable, that is $\sigma(t, x; \psi, \eta) = \sigma(t; \psi, \eta)$. Then the variance of X^p is usually unknown when b is non-linear in the state variable, whereas the variance of X^e is given by

$$V(X_t^e) = \int_0^t \sigma(s; \psi, \eta)\sigma(s; \psi, \eta)^T ds$$

for "all" drift functions b.

The possibilities for the statistical analysis of incomplete discrete observations of X^p and X^e are also quite different. From now on we assume that σ is independent of the state variable and that the incomplete discrete observations are given by the linear observation equations considered in Pedersen (1993b,1994c), see also section 5. If b is linear in the state variable then the statistical analysis of incomplete discrete observations of X^p can be performed by means of the methods described in Pedersen (1993b,1994c) and section 5, but if b is non-linear in the state variable then the statistical analysis of incomplete discrete observations of X^p becomes extremely difficult. The usual approach is to perform an approximate analysis based on the theory for approximate non-linear filtering, see Jazwinski (1970) and Åström (1970), or alternatively by approximating b by some linear function as in Melgaard et al (1992). In contrast, the statistical analysis of incomplete discrete observations of X^e can be performed by means of the methods described in Pedersen (1993b,1994c) and section 5 irrespective of whether b is linear or non-linear in the state variable, since X^e is in any case Gaussian. In fact X^e is for "all" drift functions b a Gaussian diffusion process with independent increments.

As an alternative to X^e it may in some cases be sufficient or desirable to consider discrete-time models such as

$$\begin{aligned} X_{t_i} &= \mu(t_i; \psi) + \varepsilon_{t_i}, \ i = 0, 1, \ldots, n \\ X_{t_i} &= \mu(t_i; \psi) \cdot \varepsilon_{t_i}, \ i = 0, 1, \ldots, n, \end{aligned} \tag{23}$$

where $0 = t_0 < t_1 < \cdots < t_n$ are the discrete observation time-points and $\{\varepsilon_{t_i}\}_{i=0}^n$ are some $d \times 1$ random vectors with zero or unit expectation. Apart from being less satisfactory from a modeling point of view such models are of course entirely specific for the given set of discrete observation time-points, and they can usually not in any natural way be seen as discretized versions of continuous-time models. The fact that a discrete-time model is specific for the given discrete observation time-points also restricts the applications of the model. Prediction errors can for instance only be calculated with the same time-steps as in the data set. However, in a preliminary analysis or if no satisfactory continuous-time model can be found such discrete-time models may be an appropriate alternative. As an example, consider the case where the errors in the discrete-time model given by (23) are stochastically independent with marginal distributions

$$\varepsilon_{t_i} \sim N_d(0, \Sigma),\ i = 0, 1, \ldots, n \tag{24}$$

for some positive definite $d \times d$ matrix Σ. It is of course intrinsic in dynamic models that the random vectors $\{X_{t_i}\}_{i=0}^n$ can not be stochastically independent, but if the distance between the discrete observation time-points is large it may still be a reasonable approximation. In addition, if Σ is assumed not to depend on ψ then the maximum likelihood estimator of ψ minimizes the Euclidian distance between the incomplete discrete data and the corresponding model for the expectations, which may be regarded as a justifying quality of the model. This discrete-time model has been applied to the non-linear modification of the two-dimensional prognosis model in section 2 that was made earlier in this section. In this case the distance between the discrete observation time-points is thirty days, and it seems a reasonable approximation to assume that the corresponding two-dimensional random vectors are stochastically independent. The model was applied in Jensen et al (1994) with

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix},$$

and the second coordinate of the estimated two-dimensional expectation function μ was accordingly regarded as a curve for the average evolution of the unobserved phosphorus process P_s. The results were found to be quite satisfactory from a biological point of view, see Jensen et al (1994).

5 Statistical analysis of Gaussian diffusion processes based on incomplete discrete observations

In this section we consider statistical analysis of incomplete discrete observations of the Gaussian diffusion proces X defined by the stochastic differential equation

$$dX_t = (A_t X_t + a_t)dt + B_t dW_t,\ X_0 = \xi_0,\ t \geq 0,$$

where $\xi_0 \in \boldsymbol{R}^d$ is a non-random vector, W is an r-dimensional Wiener process ($d \leq r$), A is a continuous $d \times d$ deterministic matrix function, a is a $d \times 1$ deterministic vector function and B is a $d \times r$ deterministic matrix function which satisfies that $B_t B_t^T$ is positive definite for all $t \geq 0$. The incomplete discrete observations of X are assumed to be given by

$$Y_{t_i} = T_{t_i} X_{t_i} + U_{t_i} + e_{t_i},\ i = 0, 1, \ldots, n,$$

where $0 = t_0 < t_1 < \cdots < t_n$ are the discrete observation time-points, $\{T_{t_i}\}_{i=0}^n$ are non-random $m \times d$ matrices ($m \leq d$) specifying the observable linear combinations of the coordinates of $\{X_{t_i}\}_{i=0}^n$, $\{U_{t_i}\}_{i=0}^n$ are non-random $m \times 1$ vectors representing additional inputs and $\{e_{t_i}\}_{i=0}^n$ are stochastically independent random $m \times 1$ vectors that account for measurement errors. The measurement errors are assumed to be stochastically independent of X and to have the marginal distributions

$$e_{t_i} \sim N_m(0, M_{t_i}),\ i = 0, 1, \ldots, n.$$

All unknown parameters in the functions A, a and B and in the non-random matrices and vectors $\xi_0, \{T_{t_i}\}_{i=0}^n, \{U_{t_i}\}_{i=0}^n$ and $\{M_{t_i}\}_{i=0}^n$ are collected in the p-dimensional parameter θ, and all assumptions about these functions, matrices and vectors are assumed to hold for all values of θ in a set $\Theta \subseteq \boldsymbol{R}^p$. We shall however omit θ in the notation for convenience. The statistical model for the random vectors $\{X_{t_i}\}_{i=0}^n$ and $\{Y_{t_i}\}_{i=0}^n$ may equivalently be represented by the stochastic dynamical system

$$\begin{aligned} X_{t_i} &= D_{t_i} X_{t_{i-1}} + S_{t_i} + \varepsilon_{t_i},\ X_{t_0} = \xi_0,\ i = 1, \ldots, n && (25) \\ Y_{t_i} &= T_{t_i} X_{t_i} + U_{t_i} + e_{t_i},\ i = 0, 1, \ldots, n, && (26) \end{aligned}$$

where $\{\varepsilon_{t_i}\}_{i=1}^n$ are stochastically independent random $d \times 1$ vectors with marginal distributions

$$\varepsilon_{t_i} \sim N_d(0, V_{t_i}),\ V_{t_i} = \int_{t_{i-1}}^{t_i} (\Phi_{t_i}\Phi_s^{-1}) B_s B_s^T (\Phi_{t_i}\Phi_s^{-1})^T ds.$$

Here Φ denotes the $d \times d$ deterministic matrix function that solves the differential equation

$$d\Phi_t = A_t \Phi_t dt,\ \Phi_0 = I_d,\ t \geq 0.$$

The solution to this matrix differential equation exists uniquely, and under certain conditions on A, see e.g. Pedersen (1994c), the solution is known explicitly. However, these conditions on A are often not satisfied in practice, cf. the examples in section 2 and 3. Instead we apply the approximation method proposed in Pedersen (1993b,1994c). In this approximation method an iterated Euler-approximation of Φ is used to approximate the non-random matrices and vectors $\{D_{t_i}\}_{i=1}^n, \{S_{t_i}\}_{i=1}^n$ and $\{V_{t_i}\}_{i=1}^n$ in the stochastic dynamical system (25). The approximations depend on an integer N that is chosen by the statistician, and they converge for N tending to infinity to the true matrices and vectors. By replacing the unknown matrices and vectors in the stochastic dynamical system (25) by these approximations the obtained approximate dynamical system is for large values of N a special case of the following more general stochastic dynamical system

$$\begin{aligned} X_i &= D_i X_{i-1} + S_i + \varepsilon_i,\ i = 1, \ldots, n \\ Y_i &= T_i X_i + U_i + e_i,\ i = 0, 1, \ldots, n, \end{aligned}$$

where

- $X_0 \sim N_d(\xi_0, V_0)$
- $\varepsilon_i \sim N_d(0, V_i),\ i = 1, \ldots, n$
- $e_i \sim N_m(0, M_i),\ i = 0, 1, \ldots, n$

and

(A1) $X_0, \varepsilon_1, \ldots, \varepsilon_n, e_0, e_1, \ldots, e_n$ are stochastically independent

(A2) $T_1, \ldots, T_n$ are of full rank m

(A3) $V_1, \ldots, V_n$ are positive definite.

In some cases we also assume either of the following two conditions.

(A4) T_0 is of full rank m and V_0 is positive definite

(A4') M_0 is positive definite.

The statistical analysis of incomplete discrete observations of the Gaussian diffusion process X is then performed within this framework for an increasing sequence of values of N. In particular the (approximate) maximum likelihood estimator of θ is calculated for each value of N, and the calculations are continued until this sequence of estimators has converged. The approximate maximum likelihood estimator of θ at termination is then a close approximation to the true maximum likelihood estimator of θ, see Pedersen (1993b), and the value of N at termination determines the approximation of the matrices and vectors $\{D_{t_i}\}_{i=1}^n$, $\{S_{t_i}\}_{i=1}^n$ and $\{V_{t_i}\}_{i=1}^n$ which is to be used in the further calculations of residuals, expectations, covariances, forecast, reconstructions and predictions, see Pedersen (1994c) and later in this section.

The maximum likelihood estimator of θ or the conditional maximum likelihood estimator of θ based on observations of $\{Y_i\}_{i=0}^n$ in the general stochastic dynamical system above is the maximum point of

$$\ell(\theta) = \log p_0(Y_0; \theta) + \sum_{i=1}^n \log p_{i|i-1}(Y_i | Y^{i-1}; \theta)$$

or

$$\ell | Y_0(\theta) = \sum_{i=1}^n \log p_{i|i-1}(Y_i | Y^{i-1}; \theta),$$

resepctively, where $Y^i = (Y_0^T, Y_1^T, \ldots, Y_i^T)^T,\ i = 0, 1, \ldots, n$ and p_0 is the density with respect to λ^m (the m-dimensional Lebesgue measure) of the distribution of

$$Y_0 \sim N_m(T_0 \xi_0 + U_0, T_0 V_0 T_0^* + M_0),$$

whereas $p_{i|i-1}$ is the density with respect to λ^m of the conditional distribution of Y_i given Y^{i-1} for $i = 1, \ldots, n$. These conditional densities can for a given set $\{y_i\}_{i=0}^n$ of observations of $\{Y_i\}_{i=0}^n$ be calculated iteratively for each fixed value of θ by means of the Kalman filter $\{E(X_i|Y^i), V(X_i|Y^i)\}_{i=0}^n$. From the joint normality of $\{X_i\}_{i=0}^n$ and $\{Y_i\}_{i=0}^n$, see Pedersen (1994c), we have that $X_i|Y^i = y^i$ is normally distributed for $i = 0, 1, \ldots, n$, and we can introduce the notation

$$X_i|Y^i = y^i \sim N_d(\mu_i^i(y^i), \Sigma_i^i), \; i = 0, 1, \ldots, n.$$

The following iterative formulas then show how the Kalman filter $\{\mu_i^i(y^i), \Sigma_i^i\}_{i=0}^n$ and the values $\{p_{i|i-1}(y_i|y^{i-1}; \theta)\}_{i=1}^n$ of the conditional densities can be calculated for each fixed value of θ. The assumptions are (A1)–(A3), see Pedersen (1993b,1994c). For $i = 1, \ldots, n$ we have that

$$\begin{aligned}
\mu_i^i(y^i) &= D_i\mu_{i-1}^{i-1}(y^{i-1}) + S_i + \Sigma_i^{i-1}T_i^T(T_i\Sigma_i^{i-1}T_i^T + M_i)^{-1} \\
&\quad (y_i - T_i(D_i\mu_{i-1}^{i-1}(y^{i-1}) + S_i) - U_i) \\
\Sigma_i^i &= \Sigma_i^{i-1} - \Sigma_i^{i-1}T_i^T(T_i\Sigma_i^{i-1}T_i^T + M_i)^{-1}T_i\Sigma_i^{i-1} \\
Y_i|Y^{i-1} = y^{i-1} &\sim N_m(T_i(D_i\mu_{i-1}^{i-1}(y^{i-1}) + S_i) + U_i, T_i\Sigma_i^{i-1}T_i^T + M_i),
\end{aligned}$$

where $\Sigma_i^{i-1} = D_i\Sigma_{i-1}^{i-1}D_i^T + V_i$ is positive definite for $i = 1, \ldots, n$, and the calculations are initiated by

$$\begin{aligned}
\mu_0^0(y^0) &= \xi_0 + V_0T_0^T(T_0V_0T_0^T + M_0)^{-1}(y_0 - T_0\xi_0 - U_0) \\
\Sigma_0^0 &= V_0 - V_0T_0^T(T_0V_0T_0^T + M_0)^{-1}T_0V_0.
\end{aligned}$$

Hence, the maximum likelihood estimator of θ or the conditional maximum likelihood estimator of θ can be calculated by means of some iterative procedure, and likelihood ratio confidence regions for θ can be calculated as usual.

Once an estimate of θ has been calculated, the estimated statistical model can be evaluated by calculating the stochastically independent standardized residuals

$$R_i = (T_i\Sigma_i^{i-1}T_i^T + M_i)^{-1/2}(Y_i - T_i(D_i\mu_{i-1}^{i-1}(Y^{i-1}) + S_i) - U_i) \sim N_m(0, I_m)$$

for $i = 1, \ldots, n$. Here $C^{-1/2}$ denotes the inverse of any square root of the positive definite matrix C, e.g. the lower triangular matrix in the Cholesky decomposition of C. Notice that these standardized residuals also can be calculated by means of the Kalman filter. As another means of evaluating the estimated statistical model one might calculate the estimated expectations and covariances

$$\begin{aligned}
E(X_i) &= D_iE(X_{i-1}) + S_i, \; i = 1, \ldots, n, \; E(X_0) = \xi_0 \\
V(X_i) &= D_iV(X_{i-1})D_i^T + V_i, \; i = 1, \ldots, n, \; V(X_0) = V_0 \\
E(Y_i) &= T_iE(X_i) + U_i, \; i = 0, 1, \ldots, n \\
V(Y_i) &= T_iV(X_i)T_i^T + M_i, \; i = 0, 1, \ldots, n
\end{aligned}$$

and the corresponding estimated 95%-variation bounds given by the limits

$$E(Y_i)_j \pm 1.96V(Y_i)_{jj},\ j=1,\dots,m,\ i=1,\dots,n.$$

For the purpose of forecasting we introduce the notation

$$X^i_{i+k}|Y^i=y^i \sim N_d(\mu^i_{i+k}(y^i),\Sigma^i_{i+k}),\ i=0,1,\dots,n-k,\ k=0,1,\dots,n.$$

The k-step ahead forecasts $\{\mu^i_{i+k}(y^i)\}_{i=0}^{n-k}$ for the unobservable vectors $\{X_{i+k}\}_{i=0}^{n-k}$ and the corresponding forecast error covariances $\{\Sigma^i_{i+k}\}_{i=0}^{n-k}$ can be calculated iteratively by

$$\begin{aligned}\mu^i_{i+j}(y^i) &= D_{i+j}\mu^i_{i+j-1}(y^i)+S_{i+j},\ j=1,\dots,k\\ \Sigma^i_{i+j} &= D_{i+j}\Sigma^i_{i+j}D^T_{i+j}+V_{i+j},\ j=1,\dots,k\end{aligned}$$

for $i=0,1,\dots n-k$, and the corresponding quantities for the observable quantities $\{Y_{i+k}\}_{i=0}^{n-k}$ are then eventually calculated by

$$\begin{aligned}E(Y_{i+k}|Y^i=y^i) &= T_{i+k}\mu^i_{i+k}(y^i)+U_{i+k},\ i=0,1,\dots,n-k\\ V(Y_{i+k}|Y^i=y^i) &= T_{i+k}\Sigma^i_{i+k}T^T_{i+k}+M_{i+k},\ i=0,1,\dots,n-k.\end{aligned}$$

The corresponding 95%-variation bounds are given by the limits

$$E(Y_{i+k}|Y^i=y^i)_j \pm 1.96V(Y_{i+k}|Y^i=y^i)_{jj},\ j=1,\dots,m,\ i=0,1,\dots,n-k.$$

Again we see that the Kalman filter plays a key role in the calculations. The Kalman filter is also the intial value in the calculation of the prediction $\mu^n_{n+k}(y^n)$ of the future value of X_{n+k} and the corresponding prediction error covariance matrix Σ^n_{n+k}. These quantities can be calculated inductively by

$$\begin{aligned}\mu^n_{n+j}(y^n) &= D_{n+j}\mu^n_{n+j-1}(y^n)+S_{n+j},\ j=1,\dots,k\\ \Sigma^n_{n+j} &= D_{n+j}\Sigma^n_{n+j-1}D^T_{n+j}+V_{n+j},\ j=1,\dots,k,\end{aligned}$$

and the corresponding 95%-variation bounds are given by the limits

$$\mu^n_{n+k}(y^n)_j \pm 1.96(\Sigma^n_{n+k})_{jj},\ j=1,\dots,d.$$

Finally, the optimal reconstructions $\{\mu^n_i(y^n)\}_{i=0}^n$ of the unobservable vectors $\{X_i\}_{i=0}^n$ and the corresponding reconstruction error covariances $\{\Sigma^n_i\}_{i=0}^n$ can be calculated recursively as modifications of the Kalman filter reconstructions and reconstruction error covariances. The recursive formulas are valid under the assumptions (A1)–(A3) and either (A4) or (A4'), see Pedersen (1994c). For $i=n-1,n-2,\dots,0$ we have that

$$\begin{aligned}\mu^n_i(y^n) &= \mu^i_i(y^i)+\Sigma^i_iD^T_{i+1}(\Sigma^i_{i+1})^{-1}(\mu^n_{i+1}(y^n)-D_{i+1}\mu^i_i(y^i)-S_{i+1})\\ \Sigma^n_i &= \Sigma^i_i+\Sigma^i_iD^T_{i+1}(\Sigma^i_{i+1})^{-1}(\Sigma^n_{i+1}-\Sigma^i_{i+1})(\Sigma^i_iD^T_{i+1}(\Sigma^i_{i+1})^{-1})^T.\end{aligned}$$

The corresponding 95%-variation bounds are given by the limits

$$\mu^n_i(y^n)_j \pm 1.96(\Sigma^n_i)_{jj},\ j=1,\dots,d,\ i=0,1\dots,n.$$

6 References

ARNOLD, L. (1974): *Stochastic Differential Equations: Theory and Applications.* Wiley, New York.

BIBBY, B.M. & SØRENSEN, M. (1994): Martingale Estimation Functions for Discretely Observed Diffusion Processes. *Research Report* **277**, Dept. Theor. Statist., University of Aarhus.

ERLANDSEN, M. & THYSSEN, N. (1983): Modelling the Community Oxygen Concentration in Lowland Streams Dominated by Submerged Macrophytes. *Analysis of Ecological Systems: State-of-the-Art in Ecological Modelling*, 855-860. Elsevier Scientific Publishing Company.

JAZWINSKI, A.H. (1970): *Stochastic Processes and Filtering Theory.* Academic Press, New York.

JENSEN, J.P., JEPPESEN, E., KRISTENSEN, P., CHRISTENSEN, P.B., SØNDERGAARD, M. (1992): Nitrogen Loss and Denitrification as Studied in Relation to Reductions in Nitrogen Loading in a Shallow, Hypertrophic Lake (Lake Søbygård, Denmark). *Int. Revue ges. Hydrobiol.* **77**, 29–42.

JENSEN, J.P., BØGESTRAND, J., JEPPESEN, E., PEDERSEN, A.R., SØNDERGAARD, M., WINDOLF, J., SORTKJÆR, L. (1994): *Ferske Vandområder – Søer.* Faglig rapport fra Danmarks Miljøundersøgelser, Afdeling for Ferskvandsøkologi. Vandmiljøplanens Overvågningsprogram, Miljøministeriet.

MADSEN, H. & HOLST, J. (1991): Estimation of Continuous Time Models for the Heat Dynamics of a Building. *Tech. rep.* **4**, University of Lund.

MADSEN, H. & MELGAARD, H. (1991): The Mathematical and Numerical Methods used in CTLSM — a Program for ML-estimation in Stochastic, Continuous Time Dynamical Models. *Tech. rep.* **7**, Technical University of Denmark, Lyngby, Denmark.

MELGAARD, H., MADSEN, H., HOLST, J. (1992): Methods for Parameter Estimation in Stochastic Differential Equations. Preprint.

ODUM, H.T. (1956): Primary Production in Flowing Waters. *Limnol. Oceanogr.* **1**, 102–117.

PEDERSEN, A.R. (1993a): A New Approach to Maximum Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations. *Research Re-*

port No. **264**, Dept. Theor. Statist., University of Aarhus. To appear in *Scand. J. Statist.*

PEDERSEN, A.R. (1993b): Maximum Likelihood Estimation Based on Incomplete Observations for a Class of Discrete Time Stochastic Processes by Means of the Kalman Filter. *Research Report No.* **272**, Dept. Theor. Statist., University of Aarhus.

PEDERSEN, A.R. (1993c): Consistency and Asymptotic Normality of an Approximate Maximum Likelihood Estimator for Discretely Observed Diffusion Processes. *Research Report No.* **273**, Dept. Theor. Statist., University of Aarhus.

PEDERSEN, A.R. (1994a): Uniform Residuals for Discretely Observed Diffusion Processes. *Research Report No.* **292**, Dept. Theor. Statist., University of Aarhus.

PEDERSEN, A.R. (1994b): Quasi-likelihood Inference for Discretely Observed Diffusion Processes. *Research Report No.* **295**, Dept. Theor. Statist., University of Aarhus.

PEDERSEN, A.R. (1994c): Statistical Analysis of Gaussian Diffusion Processes Based on Incomplete Discrete Observations. *Research Report No.* **297**, Dept. Theor. Statist., University of Aarhus.

SØNDERGAARD, M., KRISTENSEN, P., JEPPESEN, E. (1993): Eight Years of Internal Phosphorus Loading and Changes in the Sediment Phosphorus Profile of Lake Søbygaard, Denmark. *Hydrobiologia* **253**, 345–356.

THYSSEN, N., ERLANDSEN, M., KRONVANG, B., SVENDSEN, L.M. (1990): *Vandløbsmodeller — Biologisk Struktur og Stofomsætning.* Miljøministeriet, Miljøstyrelsen.

ZWILLINGER, D. (1992): *Handbook of Differential Equations.* Academic Press, London.

ÅSTRÖM, K.J. (1970): *Introduction to Stochastic Control Theory.* Academic Press, New York.