Estimation in Financial Models

RISKLAB

Anja Göing
ETH Zurich
Department of Mathematics
CH-8092 Zurich-Switzerland
Telephone: 41-1-632 34 58
Fax: 41-1-632 10 85
E-Mail: goeing@math.ethz.ch

January 1996

Contents

Introduction

Over the last few years various new derivative instruments have emerged in financial markets leading to a demand for versatile estimation methods for relevant model parameters. Typical examples include volatility, covariances and correlations. In this paper we give a survey on statistical estimation methods for both discrete as well as continuous time stochastic models.

The text is organized as follows: in Chapter 1 we first motivate a model in which volatility of a price process is assumed to follow a stochastic process. Out of the variety of continuous time stochastic volatility models introduced in the literature we choose two empirically relevant ones, that is an arithmetic Ornstein-Uhlenbeck process and a square root diffusion model. Those two models serve as reference models in some of the later chapters. As for discrete time stochastic volatility models, we concentrate on log-AR(p) processes, ARCH(q) and GARCH(p, q) processes all of which will be discussed in detail in Section 3.2.

Approximations of diffusion models by discrete time models and vice versa are described in Chapter 2. The convergence result on which these approximations are based is stated. As applications the diffusion limits of GARCH(1,1)-type processes and of AR(1) E-ARCH processes are derived. In addition, we present strategies for approximating diffusions and briefly compare different discretizations of a diffusion model.

In Section 3.1 estimation of an unknown parameter in a general diffusion process is discussed. For continuously observed processes we develop the classical theory of maximum likelihood estimation including properties like consistency and asymptotic normality. In the case of discrete observations the maximum likelihood estimator retains all the 'good' properties if the transition densities of the process are known. However, in most cases relevant for finance, we do not have explicit expressions for the underlying transition densities and the use of approximate likelihood functions leads to inconsistent estimators when the time between observations is bounded away from zero. We describe alternative estimation methods, as there are martingale estimating functions or methods based on approximating the transition densities. As a result, we are able to obtain consistent and asymptotically normal estimators. The methods introduced will be tested on some examples. In Section 3.2 discrete AR, ARCH and GARCH models and their asymptotic properties are discussed. Finally, a brief description of Bayesian estimation is given.

Concerning nonparametric estimation in diffusions, in Section 4.1 we discuss estimation of a probability density and estimation of an unknown signal. Section 4.2 presents two different nonparametric techniques for discrete models, namely kernel estimators and Fourier type estimators.

In Chapter 5 we show how to solve linear stochastic differential equations explicitly and give some classes of nonlinear stochastic differential equations that can be reduced to linear ones.

Some necessary background material is briefly introduced in the appendices. An extensive bibliography guides the interested reader to further published material.

Acknowledgements:
In working out the material presented, I was fortunate to have been able to discuss with various people aspects of the text. I am especially grateful to Prof. P. Embrechts for his invaluable advice and constant encouragement. I also take great pleasure in thanking A.N. Shiryaev, A. Dassios, G. Parker and H. Schurz for their much appreciated help. I thank M. Kafetzaki Boulamatsis very much for all the helpful discussions and her support during the preparation of the manuscript. The fruitful discussions within Risklab also were a constant source of inspiration.

Chapter 1

Stochastic Volatility Models

1.1 Continuous time

The value of a stock price S is supposed to follow the process

$$dS_t = S_t\,(\mu_t\,dt + \sigma_t\,dW_t), \tag{1.1}$$

where W_t is a Wiener process and μ and σ are functions of t. Considering the logarithm of the stock price $H \equiv \ln S$ and using Itô's formula we derive the process followed by H

$$dH_t = (\mu_t - \frac{\sigma_t^2}{2})\,dt + \sigma_t\,dW_t. \tag{1.2}$$

In the following consider the process H_t instead of the equivalent process S_t. For many purposes this leads to a more tractable differential equation. Also from a statistical point of view do the increments of H_t (i.e. the so-called log-returns) behave in a 'nicer' way as the increments of S_t.

In the case $\sigma_t \equiv \sigma$ and $\mu_t \equiv \mu$, S_t follows geometric Brownian motion. This is assumed in the Black-Scholes option pricing model which has been used as an effective tool for the pricing of options for more than two decades.

When comparing the calculated option values using the Black-Scholes model with the option prices there is usually a difference. Among these biases in model prices the well-known "smile"-effect is important: the Black-Scholes option pricing formula tends to underprice out-of-the-money-options and to overprice at-the-money-options, that means implied volatility changes with the striking price (see Ball [1]). This effect arises from the assumption in the Black-Scholes model that volatility is a known constant ("I sometimes wonder why people still use the Black-Scholes formula, since it is based on such simple assumptions – unrealistically simple assumptions". Black [11]).

In real life volatility is not constant at all. It is non-uniform, that means on days with major economic events volatility is usually higher than on other days (especially on non-trading days). In addition, sometimes stock prices jump which can be thought of as a sudden large increase in the stock's volatility. Furthermore many economic time series have a mean-reversion tendency: when the value of a random variable reaches a very high level then it will rather go down than up and vice versa. Put differently, it tends away from extremely high or low values and reverts to some long-term mean.

All these observations lead to the assumption that volatility is a random variable and a lot of stochastic volatility models have been recently introduced in the literature.

Now the question arises in which way stochastic volatility $\sigma_t^2 = \sigma^2(t,\omega)$ can be modeled. The standard framework assumes the volatility specification

$$dv = m(v)\,dt + k(v)\,d\tilde{W}_t, \tag{1.3}$$

with $v = \sigma_t^2$ or $v = \ln \sigma_t^2$ and $\tilde{W}_t$ a Wiener process. Out of the variety of stochastic volatility models we will consider the following two:

$$\begin{aligned} \text{I.}\quad dH_t &= (\mu_t - \frac{\sigma_t^2}{2})\,dt + \sigma_t\,dW_t \\ dv_t &= \beta\,(\alpha - v_t)\,dt + \gamma\,d\tilde{W}_t, \quad \text{with } \ v_t = \ln \sigma_t^2 \end{aligned} \tag{1.4}$$

$$\begin{aligned} \text{II.}\quad dH_t &= (\mu_t - \frac{\sigma_t^2}{2})\,dt + \sigma_t\,dW_t \\ dv_t &= b\,(a - v_t)\,dt + c\,\sqrt{v_t}\,d\tilde{W}_t\,, \quad \text{with } v_t = \sigma_t^2 \end{aligned} \tag{1.5}$$

where $\alpha, \beta, \gamma, a, b$ and c are fixed constants, W_t and $\tilde{W}_t$ are independent Wiener processes.

In the model (1.4) volatility, or more exactly $\ln \sigma^2$, is governed by an arithmetic Ornstein-Uhlenbeck (or first order autoregressive) process with a mean-reverting tendency. A large and growing literature treats this case as an empirically relevant one (see Bollerslev, Engle and Nelson [14], Stein and Stein [72], Wiggins [75]). In (1.5) a square root diffusion model for stochastic volatility is suggested. For this model see Ball [1] and literature given there.

One further reason why we have restricted our attention to the models (1.4) and (1.5) above is to be able to show explicitly how certain statistical estimation procedures are to be implemented and also compared and contrasted in a specific finance context. Most of the techniques introduced do apply to much more general set-ups.

1.2 Discrete time

In the previous section we dealt with continuous time stochastic volatility models based on systems of stochastic differential equations. In the discrete time approach we consider time series models which are systems of stochastic difference equations. In analogy to the continuous time approach we assume the stock price S_n, $S_n > 0$, to follow the process

$$\Delta S_n = S_n(\mu_n + \sigma_n z_n), \tag{1.6}$$

with μ and σ dependent on time and $z_n \sim \mathcal{N}(0,1)$. We consider the logarithm of the stock price $H_n \equiv \ln S_n$. With the notation

$$H_n = y_1 + \ldots + y_n,$$

we derive the process

$$\Delta H_n = y_n = (\mu_n - \frac{\sigma_n^2}{2}) + \sigma_n z_n. \tag{1.7}$$

We will concentrate on three different discrete stochastic volatility models. First we consider a rather simple model in which the conditional variance of the time series $\{h_n\}$ follows a logarithmic AutoRegressive (log-AR) process (see Jacquier, Polson and Rossi [42]).

I. log-AR(p)/stochastic volatility:

$$v_n = \alpha_0 + \alpha_1 v_{n-1} + \ldots + \alpha_p v_{n-p} + \gamma\, \tilde{z}_n, \tag{1.8}$$

with $v_n = \ln \sigma_n^2$ and $(z_n, \tilde{z}_n) \sim$ independent $\mathcal{N}(0,1)$.

The famous AutoRegressive Conditional Heteroskedastic (ARCH) model proposed by Engle [22] will be the second model to look at. The key insight offered by the ARCH model lies in the distinction between the conditional and the unconditional second order moments. While the unconditional covariance matrix for the variables of interest may be time invariant, the conditional variances and covariances often depend non-trivially on the past.

II. ARCH(q):

$$\sigma_n^2 = \alpha_0 + \alpha_1 y_{n-1}^2 + \ldots + \alpha_q y_{n-q}^2. \tag{1.9}$$

A long lag length q and a large number of parameters are often needed in empirical applications of ARCH(q). To avoid this problem Bollerslev introduced the Generalized ARCH (GARCH) model [12].

III. GARCH(p, q):

$$\sigma_n^2 = \alpha_0 + \alpha_1 y_{n-1}^2 + \ldots + \alpha_q y_{n-q}^2 + \beta_1 \sigma_{n-1}^2 + \ldots + \beta_p \sigma_{n-p}^2. \tag{1.10}$$

For a discussion of AR, ARCH and GARCH models see section 3.2.

Chapter 2

Approximations

2.1 Diffusion models by discrete time models

The theory of finance is mainly treated in terms of stochastic differential equations, whereas in practice observations can be made only at discrete time intervals. We want to bridge this gap by approximating diffusion models by discrete time models and vice versa.

In this section we will deal with discrete time models as diffusion approximations. The advantage of approximating a diffusion model by a sequence of discrete time models mainly lies in estimating and forecasting. Whereas the likelihood function of a discrete time model is easy to compute and to maximize, there may arise problems in deriving the likelihood of a discretely observed nonlinear stochastic differential equation system, e.g. when the diffusion coefficient depends on an unknown parameter or when there are unobservable state variables like conditional variance (see [54], [14]). In the case of a diffusion model with continuous observations, see 3.1.1 for the definition of the maximum likelihood estimator, its properties and the mentioned problems like parameter dependence. When the diffusion model is observed at discrete time points and the transition densities are unknown, discretizing the continuous time log-likelihood leads to the problem of inconsistent estimators. This problem is described in 3.1.2 where in addition three different approaches are given to overcome this problem. The latter also in the case of incomplete observations.

In summary rather than estimating and forecasting with a diffusion model observed at discrete time points it may be much easier to use a discrete time model directly, e.g. an ARCH model, as basic approximation.

First of all, as a simple though important example we approximate a Wiener process using the central limit theorem. Let $\xi_1, \xi_2, \dots$ be a sequence of iid random variables defined on a probability space $(\Omega, \mathcal{B}, P)$. Suppose $\mathrm{E}\,\xi_n = 0$, $\mathrm{Var}\,\xi_n = \sigma^2$ and $S_n = \sum_{i=1}^n \xi_i$. Via the central limit theorem we have that the distribution of $S_n/(\sigma\sqrt{n})$ converges in distribution to the normal distribution as n tends to infinity. By means of the partial sums S_n we define piecewise linear functions X_n on the interval $[0,1]$. For each n and each ω the function $X_n(\cdot,\omega)$ is linear on each interval $[(i-1)/n, i/n]$, $1 \le i \le n$ and has the value $S_i(\omega)/(\sigma\sqrt{n})$ at the point i/n with starting point $S_0(\omega) = 0$. Hence, we construct a function X_n of the form

$$X_n(t,\omega) = \frac{1}{\sigma\sqrt{n}}\, S_{i-1}(\omega) + \frac{t-(i-1)/n}{1/n}\frac{1}{\sigma\sqrt{n}}\,\xi_i(\omega),$$

for $t \in \left[\frac{i-1}{n}, \frac{i}{n}\right]$. For each ω, $X_n(\cdot,\omega)$ is a continuous function on $[0,1]$, i.e. $X_n(\cdot,\omega) \in C[0,1]$. Denote by P_n the distribution of $X_n(\cdot,\omega)$ in $C[0,1]$. Then P_n converges weakly to a Wiener measure W

$$P_n \longrightarrow W$$

(see [10], §2). Note that this convergence result can be viewed in two different ways: on the one hand X_n can be seen as an approximation to a Wiener process, and on the other hand the Wiener process as an approximation to X_n. We will come back to this point later.

The main statement of this chapter is a convergence theorem, where general conditions are given for a sequence of discrete time Markov processes to converge weakly to an Itô process. These conditions were developed by Stroock and Varadhan (see [73], §11.2). As mentioned in the example above, note that given a convergence result like this, we may use this fact to tackle the approximation problem of continuous time by discrete models *and* vice versa (see section 2.2). For a presentation of the convergence theorem and the proof we also refer to Nelson [54].

The Convergence Theorem

For $h > 0$ arbitrarily, consider a discrete time Markov process $X_0, X_h, X_{2h}, \dots, X_{kh}$ denoted by $\{X_{kh}\}$, where X_{kh} takes values in $\mathbb{R}^n$ for all k. Assume that we know the transition probabilities of $\{X_{kh}\}$ and the distribution of the initial random point X_0. We construct the continuous time process $\{X_t^{(h)}\}$ from the discrete time process $\{X_{kh}\}$ by making $X_t^{(h)}$ a step function with jumps at times $h, 2h, 3h \dots$ and values $X_t^{(h)} = X_{kh}$ almost surely for $kh \le t < (k+1)h$.

Hence we have three different kinds of processes to distinguish:

1) $\{X_{kh}\}$: the family of discrete time processes $\{X_{kh}\}$ depending on h and on the discrete time index kh, $k \in \mathbb{N}$.

2) $\{X_t^{(h)}\}$: the family of continuous time processes $\{X_t^{(h)}\}$ that are step functions constructed from the discrete time process $\{X_{kh}\}$ as described above. Observe that $\{X_t^{(h)}\}$ depends both on h and on the continuous time index $t \geq 0$.

3) $\{X_t\}$: the limiting process $\{X_t\}$, to which under some conditions as will be shown in the theorem below, the sequence of processes $\{X_t^{(h)}\}$ for $h \downarrow 0$ weakly converges.

Instead of giving the explicit mathematical conditions needed in the following theorem (for details see Nelson [54], pp.10-15) we briefly describe and interpret them. Functions $a_h(x,t)$ and $b_h(x,t)$ are defined as measures of the second moment and the drift, respectively, and are required to converge uniformly on compact sets to well-behaved continuous functions $a(x,t)$ and $b(x,t)$. Moreover, there shall exist a continuous function $\sigma(x,t)$ with $a(x,t) = \sigma(x,t)\sigma(x,t)^T$. The sample paths of the limit process X_t are assumed to be continuous with probability one. We require the probability measures of the initial points $X_0^{(h)}$ to converge to a limit measure ν_0 as $h \downarrow 0$ and thus have determined the initial distribution ν_0 of X_t. Finally, certain conditions are needed so that ν_0, $a(x,t)$ and $b(x,t)$ uniquely define the distribution of the limit process X_t.

Theorem 1 *Under the assumptions indicated above the family $\{X_t^{(h)}\}$ converges weakly as $h \downarrow 0$ to the process $\{X_t\}$ defined by the stochastic differential equation*

$$X_t = X_0 + \int_0^t b(X_s, s)\, ds + \int_0^t \sigma(X_s, s)\, dW_s^{(n)},$$

where $W_t^{(n)}$ is an n-dimensional Wiener process independent of X_0, and $\{X_t\}$ has initial distribution ν_0. The process $\{X_t\}$ exists, is distributionally unique and remains with probability one finite in finite time intervals.

We remark that the distribution of X_t does not depend on the choice of σ, see assumptions above. Moreover, note that convergence in distribution means convergence regarding the whole sample path, that means the probability laws generating the sample paths $\{X_t^{(h)}\}$ converge to the probability law generating the sample path of $\{X_t,\, 0 \leq t \leq T\}$ for any $0 \leq T < \infty$. Furthermore, we remark that Nelson [54] shows the same result based on simpler conditions for the continuity of the sample paths of X_t and the definitions for a_h and b_h. Later we will refer to these conditions as Nelson-conditions.

Now as an application of Theorem 1 Nelson [54] finds and analyzes the diffusion limit of GARCH(1,1)-type processes.

The GARCH(1,1)-M process of Engle and Bollerslev (see [13]) is defined as

$$\begin{aligned} Y_t &= Y_{t-1} + c\,\sigma_t^2 + \sigma_t\,\varepsilon_t, && (2.1)\\ \sigma_{t+1}^2 &= \gamma + \sigma_t^2\left[\beta + \alpha\,\varepsilon_t^2\right], && (2.2) \end{aligned}$$

where $\{\varepsilon_t\} \sim \mathcal{N}(0,1)$ iid.

Now our purpose is to reduce the length of the time intervals more and more. The parameters α, β and γ of the system may depend on h. The drift term in (2.1) and the variance of ε_t are made proportional to h:

$$\begin{aligned} Y_{kh} &= Y_{(k-1)h} + h\cdot c\,\sigma_{kh}^2 + \sigma_{kh}\varepsilon_{kh}, && (2.3)\\ \sigma_{(k+1)h}^2 &= \gamma_h + \sigma_{kh}^2\left[\beta_h + \frac{\alpha_h}{h}\,\varepsilon_{kh}^2\right], && (2.4) \end{aligned}$$

where $\{\varepsilon_{kh}\} \sim \mathcal{N}(0,h)$ iid and as for the initial distribution ($k=0$) we have

$$P\left[(Y_0, \sigma_0^2) \in A\right] = \nu_h(A)$$

for all $A \in \mathcal{B}(\mathbb{R}^2)$. The continuous time processes $Y_t^{(h)}$ and $\sigma_t^{(h)^2}$ are constructed by

$$Y_t^{(h)} \equiv Y_{kh} \quad \text{and} \quad \sigma_t^{(h)^2} \equiv \sigma_{kh}^2$$

for $kh \leq t < (k+1)h$. The parameters γ, α and β are allowed to depend on h because we want to find the sequences $\{\gamma_h, \alpha_h, \beta_h\}$ that make the process $\{Y_t^{(h)}, \sigma_t^{(h)^2}\}$ converge in distribution to a limit process as $h \downarrow 0$. Furthermore, in order to make the process σ_t^2 staying positive with probability one, γ_h, α_h, β_h are assumed to be nonnegative for all h.

For the Nelson-conditions to be satisfied we at least need the following limits to exist and be finite:

$$\begin{aligned} &\lim_{h\downarrow 0} h^{-1}\gamma_h \equiv \gamma \geq 0,\\ &\lim_{h\downarrow 0} h^{-1}(1-\beta_h-\alpha_h) \equiv \theta,\\ &\lim_{h\downarrow 0} 2h^{-1}\alpha_h^2 \equiv \alpha^2 > 0. \end{aligned}$$

By means of Theorem 1 we now obtain a diffusion limit of the form

$$d\begin{pmatrix} Y_t \\ \sigma_t^2 \end{pmatrix} = \begin{pmatrix} c\,\sigma^2 \\ \gamma - \theta\sigma^2 \end{pmatrix} dt + \begin{pmatrix} \sigma^2 & 0 \\ 0 & \alpha^2\sigma^4 \end{pmatrix} d\begin{pmatrix} W_{1,t} \\ W_{2,t} \end{pmatrix}, \tag{2.5}$$

where $W_{1,t}$ and $W_{2,t}$ are independent Wiener processes, independent of (Y_0, σ_0^2) and with initial distribution

$$P[(Y_0, \sigma_0^2) \in A] = \nu_0(A)$$

for all $A \in \mathcal{B}(\mathrm{I\!R}^2)$, see [54], p.17.

At this point we see that the two sections 2.1 and 2.2 are closely related to each other. There exists no closed form for the stationary distribution of the system (2.1,2.2) in discrete time, but in continuous time we are able to derive the stationary distribution of σ_t^2 in (2.5) by using the results of Wong [76] (see Nelson [54]). The idea is to draw a conclusion concerning the stationary distribution from continuous time to discrete time by means of Theorem 1, that is, we use Theorem 1 in the other direction. This is the main subject of section 2.2 and therefore regarding inference from continuous time to discrete time we refer to the next section, where we also consider another example, a so called E-ARCH model, proposed by Nelson.

Continuing this section, we come to the subject of strategies for approximating diffusions. Approximation schemes like the standard Euler approximation (see Appendix B.1) or the Milstein scheme (see Appendix B.2) are known strategies for approximating diffusions. As an example we apply the standard Euler approximation scheme to the model (1.4), which is written here again for the reader's convenience

$$\begin{aligned} dY_t &= \left(\mu_t - \frac{\sigma_t^2}{2}\right) dt + \sigma_t\, dW_{1,t} \\ d\left[\ln\left(\sigma_t^2\right)\right] &= \beta\left[\alpha - \ln\left(\sigma_t^2\right)\right] dt + \gamma\, dW_{2,t}. \end{aligned} \tag{2.6}$$

The standard Euler approximation scheme for (2.6), respectively (1.4), is given by

$$\begin{aligned} y_{t+h} &= y_t + \left(\mu_t - \frac{\sigma_t^2}{2}\right) h + \sigma_t\sqrt{h}\,\varepsilon_{1,t+h} \\ \ln\left(\sigma_{t+h}^2\right) &= \ln\left(\sigma_t^2\right) + h\,\beta\left[\alpha - \ln\left(\sigma_t^2\right)\right] + \gamma\sqrt{h}\,\varepsilon_{2,t+h}, \end{aligned} \tag{2.7}$$

for $t = h, 2h, 3h, \ldots$, with (y_0, σ_0) fixed and $\varepsilon_{1,t}$, $\varepsilon_{2,t}$ independent, $\sim \mathcal{N}(0,1)$. The diffusion (2.6), respectively (1.4), does not satisfy global Lipschitz conditions and cannot be transformed in order to satisfy them. Thus, the known theorems on strong convergence of the Euler approximation (see Appendix B.1) do not apply. But by means of the convergence theorem above the weak convergence of the system (2.7) to the diffusion model (2.6) can be verified (for the proof we refer to [14]). Hence, with the standard Euler approximation scheme we arrive at a way to approximate model (1.4).

As a further remark, consider a sequence of processes $\{X_{n,t}\}$ defined by

$$dX_{n,t} = b(X_{n,t}, t)\, dt + \sigma(X_{n,t}, t)\, dW_{n,t}.$$

Wong and Hajek [76], §4.5, show that, under reasonable conditions, as n tends to infinity, the processes $\{X_{n,t}\}$ converge to the process $\{X_t\}$

$$dX_t = b(X_t, t)\, dt + \frac{1}{2}\, \sigma(X_t, t)\, \sigma'(X_t, t)\, dt + \sigma(X_t, t)\, dW_t,$$

where $\sigma' = (\partial\sigma)/(\partial x)$. Note the presence of the additional term $\frac{1}{2}\sigma\sigma'$ which results from an application of the Itô formula (see Appendix C). Intuitively, one would expect a limit without the term $\frac{1}{2}\sigma\sigma'$, and note that this is the limit in the case where the diffusion coefficient only depends on time, $\sigma(x,t) \equiv \sigma(t)$.

Another interesting problem is to find the in some sense best discretization of a continuous time model. In the case of the centered Cox-Ingersoll-Ross model

$$dr_t = -kr_t dt + \sigma\sqrt{r_t + \gamma}\, dW_t, \tag{2.8}$$

(see also model (1.5)), we refer to Deelstra and Parker [20] for a discussion of two different discretizations. Besides the 'simple' discretization of (2.8)

$$r_t = \phi r_{t-1} + \sigma_\varepsilon \sqrt{r_{t-1} + \gamma}\, \varepsilon_t, \tag{2.9}$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d. and $\phi, \sigma_\varepsilon > 0$, Deelstra and Parker [20] consider another discrete representation of (2.8)

$$r_t = \phi r_{t-1} + \sigma_\varepsilon \sqrt{\frac{2\phi}{1+\phi} r_{t-1} + \gamma}\; \varepsilon_t, \tag{2.10}$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$ i.i.d. and $\phi, \sigma_\varepsilon > 0$. They establish the parametric relations between the continuous time model and the discrete models. Whereas in the simple model (2.9), the mean and the stationary variance are the same as in the continuous time model (2.8), the discretization (2.10) has in addition the same covariance as model (2.8) at all sampling times, that is both the first and the second moment are equal. Therefore the covariance equivalent discretization (2.10) is suggested in [20] to be used as a discrete representation of the continuous time model (2.8). It is also illustrated in [20] that although model (2.9) looks a lot like model (2.10) both models can produce significantly different estimates for the parameters.

2.2 Discrete time models by diffusion models

Approximation of discrete time models by diffusion models is a way to simplify the analysis of discrete models. For instance, the properties of discrete time models such as consistency and asymptotic normality of maximum likelihood estimates are rather difficult, and often distributional results are available for the diffusion limit of a sequence of discrete processes that are not available for the discrete models themselves. In such cases we may be able to use a convergence theorem as in section 2.1 and approximate discrete time processes, especially ARCH processes, by diffusion processes.

We pick up again the GARCH(1,1) model (2.1,2.2) dealt with in the previous section, where the diffusion limit (2.5) of the system (2.1,2.2) was obtained via the convergence theorem (with Nelson-conditions). As mentioned, there exists no closed form for the stationary distribution of the system (2.1,2.2) in discrete time, but we are able to derive the stationary distribution of σ_t^2 in the diffusion limit (2.5) by using the results of Wong [76] (see Nelson [54]). Nelson shows that the stationary distribution of σ_t^2 is an inverted gamma and uses this knowledge in continuous time to obtain distributional results for the discrete time. While the innovation process $\sigma_{kh} \cdot \varepsilon_{kh}$, see (2.3), is conditionally normal distributed, we obtain that it is (unconditionally) approximately distributed as a Student t, in the case when the time length between the observations, i.e. h, is small and kh is large (see [54], p.18f).

Consider a (slightly different) model (1.4)

$$\begin{aligned} dY_t &= \theta\, \sigma_t^2 dt + \sigma_t\, dW_{1,t} \\ d\left[\ln\left(\sigma_t^2\right)\right] &= \beta\, \left[\alpha - \ln\left(\sigma_t^2\right)\right]\, dt + dW_{2,t}, \end{aligned} \tag{2.11}$$

where $W_{1,t}$ and $W_{2,t}$ are Brownian motions with

$$\begin{pmatrix} dW_{1,t} \\ dW_{2,t} \end{pmatrix} (dW_{1,t}\ dW_{2,t}) = \begin{pmatrix} 1 & C_{12} \\ C_{12} & C_{22} \end{pmatrix} dt,$$

and $C_{22} \geq C_{12}^2$. As mentioned in the example in the previous section (see p.12) the system (2.11) does not satisfy global Lipschitz conditions and hence the standard convergence theorems of the Euler approximation do not apply. Nelson developes a class of discrete time models based on the Exponential ARCH (E-ARCH) model that converge weakly to a diffusion, see [54] pp.20-23. We will consider such a diffusion approximation for the model (2.11). Since $\ln(\sigma_t^2)$ follows a continuous time AR(1) process, in the discrete time model the conditional variance process is also assumed to be an AR(1) process, that means we assume that it follows an AR(1) E-ARCH process:

$$
\begin{aligned}
\ln\left[Y_{kh}\right] &= \ln\left[Y_{(k-1)h}\right] + h\,\theta_h\,\sigma_{kh}^2 + \sigma_{kh}\cdot\varepsilon_{kh}, \\
\ln\left[\sigma_{(k+1)h}^2\right] &= \ln\left[\sigma_{kh}^2\right] + \beta\,\left[\alpha - \ln\left(\sigma_{kh}^2\right)\right]\,h + C_{12}\cdot\varepsilon_{kh} \\
&\quad +\gamma\,\left[|\varepsilon_{kh}| - (2h/\pi)^{1/2}\right],
\end{aligned}
\tag{2.12}
$$

where $\gamma \equiv [(C_{22} - C_{12}^2)/(1-2/\pi)]^{1/2}$ and ε_{kh} iid, $\sim \mathcal{N}(0,h)$.

If $\ln(\sigma_0^2)$ is normally distributed, then the $\ln(\sigma_t^2)$ process in (2.11) is Gaussian. Otherwise, if $\beta > 0$, a Gaussian stationary limit distribution for $\ln(\sigma_t^2)$ exists. Thus, the conditional variance in continuous time is lognormal, and as in the GARCH(1,1) example above, from this information Nelson [54] infers the distribution of the innovation process in the discrete time model (2.12). He shows that in the discrete time model (with short time intervals) the distribution of the innovations is approximately a normal-lognormal mixture.

Hence, we derived the approximate distributions of GARCH(1,1) and AR(1) E-ARCH models for small sampling intervals by using the distributional results available for the diffusion limit.

Chapter 3

Parameter Estimation

3.1 Diffusion models

Consider the general type of a diffusion process $X = (X_t)_{t \geq 0}$ defined as the solution to the stochastic differential equation

$$dX_t = b(t, X_t; \theta)\, dt + \sigma(t, X_t; \theta)\, dW_t, \quad X_0 = x_0, \quad t \geq 0, \tag{3.1}$$

where W is an r-dimensional Wiener process, $\theta \in \Theta \subseteq \mathbb{R}^p$, $b(\cdot, \cdot; \theta) : [0, \infty) \times \mathbb{R}^d \mapsto \mathbb{R}^d$ and $\sigma(\cdot, \cdot; \theta) : [0, \infty) \times \mathbb{R}^d \mapsto M^{d \times r}$ are "nice"[1] functions where $M^{d \times r}$ denotes the set of real $d \times r$ matrices.

The situation will be discussed where a realization of the process X is observed, but the parameter θ is unknown to the observer. Hence, we have to construct sufficiently good estimators of θ and examine their properties. Deriving estimates of θ there are two different kinds of observations to distinguish: continuous observations of X as considered in section 3.1.1, and discrete observations that will be considered in section 3.1.2.

3.1.1 Continuous observations

Suppose that X satisfies

$$dX_t = b(\theta, X_t)dt + \sigma(\theta, X_t)dW_t, \quad X_0 = x_0, \quad t \geq 0, \tag{3.2}$$

where for convenience X and W are one-dimensional processes, b and σ are smooth functions, and the parameter $\theta \in \Theta \subseteq \mathbb{R}^p$ is to be estimated by

[1] b and σ are Lipschitz continuous and satisfy a growth condition. Then there exists a unique strong solution of (3.1), see [65], p.128 and p.136.

continuous observations of X. X is a Markov process. P_θ denotes the law of the process X on the canonical space $\Omega = C(\mathbb{R}_+, \mathbb{R})$ with the canonical filtration $\mathcal{F}_t = \sigma(X_s | s \leq t)$. Denote by $P_{\theta,t} := P_\theta | \mathcal{F}_t$ the restriction of P_θ to $\mathcal{F}_t$.

First consider the case where $\sigma(\theta, x)$ does not depend on θ. If σ does not vanish, the measures $P_{\theta,t}$ and $P_{\theta_1,t}$ for any $\theta, \theta_1 \in \Theta$, $t < \infty$, are equivalent, denoted by $P_{\theta,t} \sim P_{\theta_1,t}$, (see [39] and [51] §7, see also Appendix D). Then we are able to introduce the so called likelihood process

$$L_t^{\theta,\theta_1} = \frac{dP_{\theta,t}}{dP_{\theta_1,t}}, \tag{3.3}$$

the 'density process' of P_θ, relative to P_{θ_1}, or the 'Radon-Nikodym derivative' (see Appendix D). The process $\left(L_t^{\theta,\theta_1}\right)$ is a $\mathcal{F}_t$-martingale (see [65], §4.17).

First we will concentrate on the case with a drift term linear in θ

$$dX_t = \theta b(X_t)dt + \sigma(X_t)dW_t, \; X_0 = x_0, \tag{3.4}$$

where b is possibly nonlinear, $\sigma > 0$, and the process (X_t) is observed in the time interval $[0, T]$. Then the likelihood process (3.3) in T with $\theta_1 = 0$ equals

$$L_T^{\theta,0} \equiv L_T(\theta) = \exp\left[\int_0^T \frac{\theta b(X_s)}{\sigma^2(X_s)} dX_s - \frac{1}{2}\int_0^T \frac{\theta^2 b^2(X_s)}{\sigma^2(X_s)} ds\right], \tag{3.5}$$

see e.g. [51], §17, or [43], p.76, or [39]. In the following consider a constant diffusion term σ, say $\sigma \equiv 1$. For convenience we take the logarithm in (3.5) and obtain the logarithmic likelihood (log-likelihood) function

$$\ln L_T(\theta) = \int_0^T \theta b(X_s) dX_s - \frac{1}{2}\int_0^T \theta^2 b^2(X_s) ds. \tag{3.6}$$

Maximizing L_T, or equivalently $\ln L_T$, with respect to θ we obtain the so called Maximum Likelihood Estimator (MLE) $\hat{\theta}_T$ based on continuous observations of X in the interval $0 \leq t \leq T$. We remark that by definition of the likelihood function the measure P_{θ_1}, that dominates the measure P_θ (that is $P_\theta << P_{\theta_1}$, see Appendix D), must not depend on an unknown parameter. Such a measure P_{θ_1} cannot be found, if besides the drift coefficient also the diffusion coefficient σ depends on an unknown parameter, that is this parameter cannot be estimated by the ML method.

The derivative with respect to θ of the log-likelihood is called the score function

$$\frac{\partial}{\partial \theta} \ln L_T(\theta) = \frac{L_T'(\theta)}{L_T(\theta)},$$

which in the case (3.6) has the form

$$\frac{\partial}{\partial \theta}(\ln L_T(\theta)) = \int_0^T b(X_s)dX_s - \theta \int_0^T b^2(X_s)ds.$$

Solving $\frac{\partial}{\partial \theta}(\ln L_T(\theta)) = 0$, we obtain the MLE

$$\hat{\theta}_T = \frac{\int_0^T b(X_s)dX_s}{\int_0^T b^2(X_s)ds}$$

and hence, using equality (3.4), we have

$$\hat{\theta}_T = \theta_0 + \frac{\int_0^T b(X_s)dW_s}{\int_0^T b^2(X_s)ds}, \tag{3.7}$$

where θ_0 denotes the true value of the parameter.

Now we derive some properties of the MLE $\hat{\theta}_T$.

First, we treat the problem of the bias of an estimator. The bias in $\hat{\theta}_T$ as an estimator of θ_0 is defined as

$$b_{\hat{\theta}_T}(\theta_0) = \mathrm{E}_{\theta_0}(\hat{\theta}_T - \theta_0),$$

where E_{θ_0} denotes the expectation with respect to P_{θ_0}. By (3.7) we have

$$E_{\theta_0}\hat{\theta}_T = \theta_0 + E_{\theta_0}\left[\frac{\int_0^T b(X_s)dW_s}{\int_0^T b^2(X_s)ds}\right]. \tag{3.8}$$

Note that in the case of a constant drift coefficient $b(X_t) \equiv b$

$$E_\theta \frac{bW_T}{b^2 T} = \frac{1}{bT} E_\theta W_T = 0,$$

thus $E_{\theta_0}\hat{\theta}_T = \theta_0$, so that $\hat{\theta}_T$ is unbiased. In all other cases the estimator is biased and in the case (3.6) we have under some regularity conditions the equality

$$b_{\hat{\theta}_T}(\theta_0) = \frac{d}{d\theta_0}\mathrm{E}_{\theta_0}\left(\int_0^T b^2(X_t)dt\right)^{-1},$$

(see [51], §17.2 and §17.3, Theorem 17.2 and 17.3).

Now we turn to the properties of consistency and asymptotic normality. Assuming some natural regularity conditions (see [38], p.83 and p.90, or [17], p.500f, or [48], p.95), the stochastic differential equation (3.4) has an ergodic solution. For the ergodic theory we refer to e.g. [32], §18. Then the stationary density $\tilde{p}$ of the process solves the (deterministic) stationary Fokker-Planck equation, which in the above case reduces to

$$\frac{d}{dx}\left[\theta b(x)\tilde{p}(x)\right] - \frac{1}{2}\frac{d^2}{dx^2}\tilde{p}(x) = 0.$$

Ergodicity implies

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T b(X_t)dW_t = 0 \quad P_{\theta_0} \text{ a.s.}$$

and

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T b^2(X_t)dt = \int_{-\infty}^{\infty} b^2(x)\tilde{p}(x)dx \quad P_{\theta_0} \text{ a.s.}$$

Hence we conclude that it is possible to achieve ultimate arbitrary precision of the MLE $\hat{\theta}_T$ in (3.7) by infinitely increasing T, that means the MLE $\hat{\theta}_T$ is weakly consistent:

$$\lim_{T\to\infty} P_{\theta_0}\left[|\hat{\theta}_T - \theta_0| > \varepsilon\right] = 0 \tag{3.9}$$

for all $\varepsilon > 0$.

Furthermore, under these regularity conditions the MLE is asymptotically normal, that means as $T \to \infty$ the difference $\sqrt{T}(\hat{\theta}_T - \theta_0)$ is asymptotically normal with parameters $(0, \sigma^2(\theta_0))$, where

$$\sigma^2(\theta_0) = \left[\int_{-\infty}^{\infty} b^2(x)\tilde{p}(x)dx\right]^{-1},$$

(see e.g. [48], p.95f, or [45], p.243).

We remark that this convergence is uniformly for all θ on a compact set $K \subseteq \Theta$, (see e.g. [48], p.94f). Moreover, note that under the same regularity conditions the MLE is also consistent and asymptotic normal if the drift term is nonlinear (again see e.g. [48], p.94f). Finally, we remark that $\sigma^2(\theta_0)$ equals $I^{-1}(\theta_0)$, where I denotes the so called Fisher information

$$\begin{aligned} I(\theta) &= \mathrm{Var}\left[\frac{\partial}{\partial\theta}\ln L_T(\theta)\right] \\ &= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln L_T(\theta)\right)^2\right], \end{aligned}$$

which can alternatively be computed by

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln L_T(\theta)\right],$$

(see [49], p.249f).

Asymptotic normality can be used to determine confidence intervals for θ and to determine a 'suitable' value of T.

Example 1. a) The regularity conditions are satisfied in the case of observations coming from

$$dX_t = -\theta X_t dt + dW_t, \quad X_0 = 0,\ 0 \le t \le T, \tag{3.10}$$

where $\theta \in (\alpha, \beta)$, $\alpha > 0$, that is $-\theta < 0$ (see [48], p.99). Thus the MLE $\hat{\theta}_T$ is consistent and asymptotically normal. The stationary density is

$$\tilde{p}(x) = \sqrt{\tfrac{\theta}{\pi}} e^{-\theta x^2},$$

and we have

$$\sigma^2(\theta_0) = \left[\sqrt{\tfrac{\theta_0}{\pi}} \int_{-\infty}^{\infty} x^2 e^{-\theta_0 x^2} dx\right]^{-1} = 2\theta_0,$$

and hence the MLE $\hat{\theta}_T$ is asymptotically normal with parameters $(0, 2\theta_0)$.

b) In the case $\theta \in (\alpha, \beta)$ and $\beta < 0$, that is $-\theta > 0$, the process has no stationary distribution. Nevertheless we can show that the MLE $\hat{\theta}_T$ is consistent. Furthermore, $e^{\theta_0 T}(\hat{\theta}_T - \theta_0)$ is asymptotically normal with parameters $(0, 4\theta_0^2)$ (see [48], p.100).

c) In the case $\theta = 0$ the MLE $\hat{\theta}_T$ is consistent, but not asymptotically normal (see [48], p.100).

We remark that in the discrete case (section 3.2) an analogous distinction regarding parameter values is made, see p.43.

After having treated the special case of a drift term linear in θ we give a short remark to the general case

$$dX_t = b(\theta, X_t)dt + dW_t,$$

where $b(\theta, x)$ may be a nonlinear function. By Taylor expansion for θ we have

$$b(\theta, x) = b(\theta_0, x) + (\theta - \theta_0)\frac{d}{d\theta} b(\theta_0, x) + O((\theta - \theta_0)^2),$$

where θ_0 again denotes the true value of parameter θ. Hence $b(\theta, x)$ can be approximated by a term linear in θ. Since in practice a 'good guess' θ of θ_0 often is available, then with the above considerations it is sufficient only to consider the already treated case: a drift term linear in θ. This closes our considerations about ML estimation.

In the following we give some remarks to the case where the diffusion term depends on an unknown parameter θ. As mentioned before in this case we cannot estimate θ by using Maximum Likelihood theory (see p.17).

As a first estimating problem we deal with the process

$$dX_t = \sigma(\theta)dW_t, \quad 0 \le t \le T. \tag{3.11}$$

In order to estimate θ we discretize the process and derive its limit. Consider partitions $\Delta_n = \left(t_0^{(n)}, \ldots, t_{m(n)}^{(n)}\right)$ of $[0, T]$ constructed in such a way that $\sum_{n=1}^{\infty} \sup_k |t_{k+1}^{(n)} - t_k^{(n)}| < \infty$ for $n \longrightarrow \infty$. Then for the discretized Wiener process W we have the convergence result

$$\sum_{k=1}^{m} |W_{t_k^{(n)}} - W_{t_{k-1}^{(n)}}|^2 \longrightarrow T,$$

in probability for $n \longrightarrow \infty$ (see [15], p.262f). Hence for (3.11) we obtain

$$\sum_{k=1}^{m} |X_{t_k^{(n)}} - X_{t_{k-1}^{(n)}}|^2 \longrightarrow T\sigma^2(\theta),$$

in probability for $n \longrightarrow \infty$, thus we are able to give an arbitrarily precise estimate of $\sigma^2(\theta)$. Notice that we do not obtain a direct estimate of θ.

Next, if the unknown parameter θ splits into two parts (θ_1, θ_2) in the following way

$$dX_t = b(t, \theta_1)dt + \sigma(\theta_2)dW_t, \tag{3.12}$$

we want to estimate the part θ_2 in the diffusion coefficient. Denoting $\Delta X_k^{(n)} \equiv X_{t_k^{(n)}} - X_{t_{k-1}^{(n)}}$ and $\Delta W_k^{(n)} \equiv W_{t_k^{(n)}} - W_{t_{k-1}^{(n)}}$ we obtain by discretizing (3.12)

$$\begin{aligned} \sum_{k=1}^{m} \left[\left(\Delta X_k^{(n)}\right)^2 - \left(\sigma(\theta_2)\Delta W_k^{(n)}\right)^2 \right] &= \sum_{k=1}^{m} b^2(t_k, \theta_1) \left(\Delta t_k^{(n)}\right)^2 \\ &\quad + 2\sum_{k=1}^{m} b(t_k, \theta_1)\sigma(\theta_2)\Delta W_k^{(n)} \Delta t_k^{(n)}, \end{aligned}$$

which tends to 0 in probability as $n \longrightarrow \infty$. Thus the drift term is insignificant as $n \longrightarrow \infty$ and we are able to estimate $\sigma(\theta_2)$ just as in the previous case. Then by treating θ_2 as known we can estimate θ_1 via the ML method.

Finally, for the general problem with a diffusion coefficient also depending on X

$$dX_t = \sigma(\theta, X_t)dW_t, \tag{3.13}$$

a convergence in probability result is given by

$$\sum_{k=1}^{m} |X_{t_k^{(n)}} - X_{t_{k-1}^{(n)}}|^2 \longrightarrow \int_0^T \sigma^2(\theta, X_s)ds, \tag{3.14}$$

for $n \longrightarrow \infty$ (see [24], [51], §4). A hint towards the correctness of (3.14) is given by the well-known equality

$$\mathrm{E}(X_t^2) = \int_0^t \mathrm{E}(\sigma^2(\theta, X_s))ds.$$

Again note that we cannot estimate θ directly but are only able to estimate $\int_0^T \sigma^2(\theta, X_t)dt$.

3.1.2 Discrete observations

In this section we deal with the (more realistic) situation where the diffusion process X (3.1) has only been observed at not necessarily equally spaced discrete time points $0 = t_0 < t_1 < \ldots < t_n$. In our discussion below, we basicly follow Kloeden, Schurz, Platen and Sørensen [46], Bibby [4, 5, 6, 7], Bibby and Sørensen [8] and Pedersen [58, 59, 60, 61, 62, 63].

For the estimation of θ from discrete observations of X we have to distinguish between two cases: the transition densities of X are known or unknown.

If the transition densities $p(s, x, t, y; \theta)$ of X are known, e.g. in the case of an Ornstein-Uhlenbeck process (see p.36, Example 1), an obvious choice of an estimator for θ is the Maximum Likelihood Estimator (MLE) $\hat{\theta}_n$ which maximizes the likelihood function

$$L_n(\theta) = \prod_{i=1}^{n} p(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i}; \theta),$$

or equivalently the log-likelihood function

$$l_n(\theta) \equiv \ln L_n(\theta) = \sum_{i=1}^{n} \log\left(p(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i}; \theta)\right) \tag{3.15}$$

for θ, see e.g. [2], p.14. In the case of time-equidistant observations ($t_i = i\Delta$, $i = 0, 1 \ldots, n$ for some fixed $\Delta > 0$) Dacunha-Castelle and Florens-Zmirou [19] prove consistency and asymptotic normality of $\hat{\theta}_n$ as $n \to \infty$, independent of the value of Δ. Unfortunately in general the transition densities of X are unknown.

When the transition densities of X are unknown, the usual alternative estimator is defined by approximating the log-likelihood function for θ based on continuous observations of X. For this log-likelihood function to be defined, the diffusion coefficient $\sigma(t,x;\theta)=\sigma(t,x)$ has to be known (see Section 3.1.1, p.17). Under some assumptions (see [51] §7) the log-likelihood function for θ based on continuous observations of X in $[0,t_n]$ can be written in terms of integrals

$$\begin{aligned} l^c_{t_n}(\theta) \; = \; & \int_0^{t_n} b\left(s,X_s;\theta\right)^T \left(\sigma(s,X_s)\sigma(s,X_s)^T\right)^{-1} dX_s \\ & -\frac{1}{2}\int_0^{t_n} b\left(s,X_s;\theta\right)^T \left(\sigma(s,X_s)\sigma(s,X_s)^T\right)^{-1} b\left(s,X_s;\theta\right) ds, \end{aligned} \tag{3.16}$$

and the usual approximation of these integrals leads to the approximate log-likelihood function for θ based on discrete observations of X

$$\begin{aligned} \tilde{l}_n(\theta) \; = \; & \sum_{i=1}^n b\left(t_{i-1},X_{t_{i-1}};\theta\right)^T \sigma_{i-1}\left(X_{t_i}-X_{t_{i-1}}\right) \\ & -\frac{1}{2}\sum_{i=1}^n b\left(t_{i-1},X_{t_{i-1}};\theta\right)^T \sigma_{i-1}\, b\left(t_{i-1},X_{t_{i-1}};\theta\right)(t_i-t_{i-1}), \end{aligned} \tag{3.17}$$

with the notation

$$\sigma_{i-1} \equiv \left(\sigma(t_{i-1},X_{t_{i-1}})\,\sigma(t_{i-1},X_{t_{i-1}})^T\right)^{-1}.$$

When the diffusion coefficient also depends on an unknown parameter, the question arises how the parameter is to be estimated. If θ divides into two parts $\theta=(\theta_1,\theta_2)$, such that $b(\cdot,\cdot;\theta)=b(\cdot,\cdot;\theta_1)$ and $\sigma(\cdot,\cdot;\theta)$ is known up to the scalar factor θ_2, that means $\sigma(\cdot,\cdot;\theta)=\theta_2\tilde{\sigma}(\cdot,\cdot)$, we may avoid the problem of parameter dependence. In this case θ_2^2 can be estimated in advance by a quadratic variance-like formula (see Florens-Zmirou [25]), and by inserting this estimate in $\sigma(\cdot,\cdot;\theta)=\theta_2\tilde{\sigma}(\cdot,\cdot)$, the diffusion term can be assumed "known". Then the estimate of the approximate log-likelihood function $\tilde{l}_n$ can be used to estimate θ_1.

In the case where the diffusion coefficient $\sigma(\cdot,\cdot;\theta)$ depends on θ more generally, Hutton and Nelson [37] show that the discretized score function corresponding to $l^c_{t_n}$ can under certain regularity conditions still be used to estimate θ. That means in this case we are able to estimate θ based on discrete observations of X as well.

However, estimation methods for discrete observations that arise from the theory of continuous observations have the undesirable property that the estimators are strongly biased unless $\max_{1\le i\le n}|t_i-t_{i-1}|$ is "small". If the time between observations is bounded away from zero, especially in the case of time-equidistant observations with Δ fixed, Florens-Zmirou [25] shows that the estimator $\tilde{\theta}_n$ of θ obtained by maximizing the approximate log-likelihood function $\tilde{l}_n(\theta)$ is inconsistent.

To overcome the difficulties regarding parameter dependence and the dependence of $\tilde{l}_n(\theta)$ on $\max_{1\leq i\leq n}|t_i - t_{i-1}|$, we will propose three different estimation approaches. The basic idea of the first two approaches is to find good approximations to the transition densities. In the third approach we construct martingale estimating functions.

Approximation to the transition densities of X by a sequence of transition densities of approximating Markov processes

The following approach is proposed by Pedersen [58, 60].

We shall derive a sequence $(l_{n,N}(\theta))_{N=1}^{\infty}$ of approximations to $l_n(\theta)$, that builds a connection between $\tilde{l}_n(\theta)$ (see (3.17)) and $l_n(\theta)$ in the following sense: the approximation $l_{n,1}(\theta)$ is a generalization of $\tilde{l}_n(\theta)$ with no restrictions on $\sigma(\cdot,\cdot;\theta)$ regarding parameter dependence, each $l_{n,N}(\theta)$ for $N \geq 2$ is an improvement of $l_{n,1}(\theta)$ and as N tends to infinity $l_{n,N}(\theta)$ converges for each θ in probability to $l_n(\theta)$.

The crucial point of the approach is to approximate the transition densities $p(s,x,t,y;\theta)$ of X by a sequence of transition densities $(p_N(s,x,t,y;\theta))_{N=1}^{\infty}$ of approximating Markov processes which converges to $p(s,x,t,y;\theta)$ as N tends to infinity, and then to define the approximate log-likelihood functions

$$l_{n,N}(\theta) = \sum_{i=1}^{n} \log\left(p_N(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i};\theta)\right). \tag{3.18}$$

In the following we will derive the approximating densities $p_N(s,x,t,y;\theta)$. We remark that we may relax the Lipschitz assumption made in (3.1) for b and σ a little bit, namely we only assume b and σ to be locally Lipschitz continuous as defined below.

Under the following conditions (A1), (A2) and (A3) which must hold for all $\theta \in \Theta$, the stochastic differential equation (3.1) has a weak solution for all x_0 and θ and has the pathwise-uniqueness property which implies the uniqueness in law (see [65], p.132 and p.151; in this context see also [73], §5-§8).

(A1) b and σ are continuous in t for all $x \in \mathbb{R}^d$.

(A2) b and σ are local Lipschitz continuous:
for all $0 < R < \infty$ there exists $0 < K_R < \infty$ such that

$$\begin{aligned} ||\sigma(t,x;\theta) - \sigma(t,y;\theta)|| &\leq K_R||x-y|| \,, \\ ||b(t,x;\theta) - b(t,y;\theta)|| &\leq K_R||x-y|| \,, \end{aligned}$$

for all $0 \leq t \leq R$ and $x, y \in \mathbb{R}^d$ with $||x|| \leq R$, $||y|| \leq R$.

(A3) Growth condition:
for all $0 < T < \infty$ there exists $0 < C_T < \infty$ such that

$$||\sigma(t,x;\theta)|| + ||b(t,x;\theta)|| \leq C_T(1 + ||x||)$$

for all $0 \leq t \leq T$ and $x \in \mathbb{R}^d$.

We remark that the stochastic differential equation (3.1) has a strong solution, if the coefficients b and σ satisfy the (global) Lipschitz condition and for each $T > 0$ there exists some C_T such that

$$|\sigma(s,0;\theta)| + |b(s,0;\theta)| \leq C_T$$

for all $s \leq T$, see [65], p.136.
In addition we assume for all $\theta \in \Theta$

(A4) $a(t,x;\theta) \equiv \sigma(t,x;\theta)\,\sigma(t,x;\theta)^T$ is positive definite for all $t \geq 0$ and $x \in \mathbb{R}^d$.

Under the assumptions (A1)-(A4), any solution to (3.1) is also a solution to

$$dX_t = b\left(t, X_t;\theta\right)dt + a\left(t, X_t;\theta\right)^{1/2} d\tilde{W}_t, \; X_0 = x_0, t \geq 0, \tag{3.19}$$

where $a\left(t, X_t;\theta\right)^{1/2}$ denotes the positive square root of $a\left(t, X_t;\theta\right)$, and

$$\tilde{W}_t = \int_0^t a\left(s, X_s;\theta\right)^{-1/2} d\left(X_s - x_0 - \int_0^s b\left(u, X_u;\theta\right)du\right), \; t \geq 0, \tag{3.20}$$

is a d-dimensional Wiener process.
The solutions to (3.1) (or (3.19)) for $t \geq s$ with initial conditions $X_s = x$ induce for each $\theta \in \Theta$ a unique family $(P_{\theta,s,x})_{s\geq 0, x\in\mathbb{R}^d}$ of probability measures on $(\Omega, \mathcal{F}) = (C([0,\infty), \mathbb{R}^d), \mathcal{B})$, the space of continuous trajectories from $[0,\infty)$ into $\mathbb{R}^d$ with its Borel σ-field. These probability measures have the important property that they determine the transition function $P(s,x,t,A;\theta)$ of X under P_θ for $0 \leq s < t$, $x \in \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$:

$$P(s,x,t,A;\theta) = P_{\theta,s,x}(X_t \in A).$$

For fixed $0 \leq s < t$, $x \in \mathbb{R}^d$, $\theta \in \Theta$ and $N \in \mathbb{N}$, we consider for $k = 0, 1, \ldots, N$ the Euler approximation (see Appendix B.1):

$$\begin{aligned}
\tau_k &= s + k\frac{t-s}{N} \\
Y_s^{(N)} &= x \\
Y_{\tau_k}^{(N)} &= Y_{\tau_{k-1}}^{(N)} + \frac{t-s}{N}\, b\left(\tau_{k-1}, Y_{\tau_{k-1}}^{(N)}; \theta\right) + a\left(\tau_{k-1}, Y_{\tau_{k-1}}^{(N)}; \theta\right)^{1/2}\left(W_{\tau_k}^{\theta,s} - W_{\tau_{k-1}}^{\theta,s}\right).
\end{aligned}$$

Under (A1)-(A4) we have

$$Y_{\tau_N}^{(N)} \equiv Y_t^{(N)} \longrightarrow X_t \tag{3.21}$$

in $L^1(P_{\theta,s,x})$ as $N \longrightarrow \infty$ (see [45], §10.2).

Theorem 2 *For fixed* $0 \leq s < t$, $x \in \mathbb{R}^d$, $\theta \in \Theta$ *and* $N \in \mathbb{N}$ *the distribution of* $Y_t^{(N)}$ *under* $P_{\theta,s,x}$ *has a density* $p_N(s,x,t,\cdot;\theta)$ *with respect to the d-dimensional Lebesgue measure* λ^d. *For N=1*

$$\begin{aligned}
p_1(s,x,t,y;\theta) = {} & (2\pi(t-s))^{-d/2}[\det(a(s,x;\theta))]^{-1/2} \\
& \cdot \exp\left\{-\frac{1}{2(t-s)}[y - x - (t-s)b(s,x;\theta)]^T\right. \\
& \left.\cdot\, a(s,x;\theta)^{-1}[y - z - (t-s)b(s,x;\theta)]\right\},
\end{aligned} \tag{3.22}$$

and for $N \geq 2$

$$p_N(s,x,t,y;\theta) = E_{P_{\theta,s,x}}\left(p_1(\tau_{N-1}, Y_{\tau_{N-1}}^{(N)}, t, y; \theta)\right). \tag{3.23}$$

For the proof we refer to Pedersen [58].

Our aim is to show that $l_{n,N}(\theta)$ will indeed be close to $l_n(\theta)$ for large values of N, that is we give a result on the convergence of the approximating densities $p_N(s,x,t,y;\theta)$ to $p(s,x,t,y;\theta)$ as $N \longrightarrow \infty$.

Theorem 3 *In addition to (A2) and (A3) assume for all* $\theta \in \Theta$ *that*

(1) b *is continuous in* t *and in* x, *and*

(2) $a(t,x;\theta) \equiv a(\theta)$ *is positive definite.*

Then $p(s,x,t,y;\theta)$ exists, and for all $0 \le s < t$, $x \in \mathbb{R}^d$ and $\theta \in \Theta$ we have

$$p_N(s,x,t,\cdot\,;\theta) \longrightarrow p(s,x,t,\cdot\,;\theta)$$

in $L^1(\lambda^d)$ as $N \longrightarrow \infty$.

The proof is given in Pedersen [58]. We remark that in addition Pedersen [58] proves a convergence theorem, where a may also depend on x.

The $L^1(\lambda^d)$-convergence of $p_N(s,x,t,\cdot\,;\theta)$ to $p(s,x,t,\cdot\,;\theta)$ as $N \longrightarrow \infty$ has the following important consequence:

Corollary 1 *If $p_N(s,x,t,\cdot\,;\theta) \longrightarrow p(s,x,t,\cdot\,;\theta)$ in $L^1(\lambda^d)$ as $N \longrightarrow \infty$ for all $0 \le s < t$, $x \in \mathbb{R}^d$ and $\theta \in \Theta$, then $l_{n,N}(\theta) \longrightarrow l_n(\theta)$ in probability under P_{θ_0} as $N \longrightarrow \infty$ for all $\theta \in \Theta$ and $n \in \mathbb{N}$, where θ_0 denotes the true parameter value.*

For the proof see Pedersen [60].

We give some remarks:

(1) As shown in Pedersen [60] the ML estimator $\hat{\theta}_{n,N}$ obtained by maximizing $l_{n,N}(\theta)$ is consistent and asymptotically normal as n and N tend to infinity.

(2) In the case where $\sigma(\cdot,\cdot\,;\theta) = \sigma(\cdot,\cdot)$ is independent of θ one obtains

$$l_{n,1}(\theta) = K + \tilde{l}_n(\theta)\,,$$

where K is some constant. Therefore $l_{n,1}$ can in this case be seen as a generalization of $\tilde{l}_n(\theta)$.

(3) We have derived the approximate log-likelihood functions $l_{n,N}(\theta)$ in (3.18) under the assumption that each time-interval $[t_{i-1}, t_i]$ is divided into N intervals. But if the observation points are not equally spaced, it might be favourable to choose a larger N for wider time-intervals $[t_{i-1}, t_i]$. That means in general we may choose an $N_i \in \mathbb{N}$ for each time-interval $[t_{i-1}, t_i]$ and we obtain

$$l_{n,(N_1,\dots,N_n)} = \sum_{i-1}^{n} \log\left(p_{N_i}(t_{i-1}, X_{t_{i-1}}, t_i, X_{t_i};\theta)\right).$$

After these theoretical considerations we deal with the actual calculation of $l_{n,N}(\theta)$ for large values of N.

In order to maximize $l_{n,N}(\theta)$ numerical algorithms usually require the value of $l_{n,N}(\theta)$ in a finite number of points θ. For $N = 1$ $l_{n,N}(\theta)$ is explicitly given, because p_1 has a closed expression, but for $N \geq 2$ this is in general not the case.

For calculating $l_{n,N}(\theta)$ for $N \geq 2$, we have to know how to calculate $p_N(s, x, t, y; \theta)$ for all $0 \leq s < t$, $x, y \in \mathbb{R}^d$ and $\theta \in \Theta$. Considering again equation (3.23)

$$p_N(s, x, t, y; \theta) = E_{P_{\theta,s,x}} \left(p_1(\tau_{N-1}, Y^{(N)}_{\tau_{N-1}}, t, y; \theta) \right),$$

the idea is to calculate $p_N(s, x, t, y; \theta)$ by finding a good approximation to the right hand side. Denote by $(U_k^m)_{k=1,m=1}^{N-1,M}$ an i.i.d. sample from the r-dimensional standard normal distribution. Then $(Y^m)_{m=1}^M = (Y_{N-1}^m)_{m=1}^M$ given by the Euler approximation

$$\begin{aligned} Y_0^m &= x, \ m = 1, \ldots, M \\ Y_k^m &= Y_{k-1}^m + \frac{t-s}{N} b(\tau_{k-1}, Y_{k-1}^m; \theta) + \sqrt{\frac{t-s}{N}}\, \sigma(\tau_{k-1}, Y_{k-1}^m; \theta)\, U_k^m \end{aligned}$$

for $k = 1, \ldots, N-1$ and $m = 1, \ldots, M$, has the same distribution as an i.i.d. sample of $Y^{(N)}_{\tau_{N-1}}$ under $P_{\theta,s,x}$. Thus we are able to approximate the right hand side of (3.23), while we calculate

$$\frac{1}{M} \sum_{m=1}^{M} p_1\left(\tau_{N-1}, Y^m, t, y; \theta\right) \tag{3.24}$$

by means of the sequence $(U_k^m)_{k=1,m=1}^{N-1,M}$, for M chosen sufficiently large, and thereby we are able to calculate $p_N(s, x, t, y; \theta)$ with any given accuracy.

From a practical point of view it is convenient to simulate the sample $(U_k^m)_{k=1,m=1}^{N-1,M}$ once (see Kloeden and Platen [45]) and store it. Then it can be used to calculate $p_N(s, x, t, y; \theta)$ for all values of $0 \leq s < t$, $x, y \in \mathbb{R}^d$ and $\theta \in \Theta$. As for the numerical maximization of $l_{n,N}(\theta)$, we obtain appropriate starting points by maximizing $l_{n,1}(\theta)$.

Example 1

For the one-dimensional stochastic differential equation

$$dX_t = -\theta X_t dt + \theta \sqrt{1 + \frac{X_t^2}{1 + X_t^2}}\, dW_t, \tag{3.25}$$

where $X_0 = 0$ and $t \geq 0$, the log-likelihood function $l_n(\theta)$ is unknown. The approximate log-likelihood functions $(l_{n,N}(\theta))_{N=1}^{\infty}$ can be used to estimate θ, because (3.25) has a weak solution for all x_0 and for all $\theta > 0$ which is unique in law and (A4) is satisfied for all $\theta > 0$.

With $\theta_0 = 5$, $n = 1000$ and $\Delta = 0.1$ the Milstein-scheme (see Appendix B.2) with time-step $\Delta/1000$ is used to simulate $(X_{i\Delta})_{i=0}^{n}$. For such a simulation the approximate log-likelihood functions $l_{n,N}(\theta)$ are calculated for $N = 1, 25, 50, 1000$. The estimates that are obtained are shown in Table 3.1.

N	1	25	50	1000
$\hat{\theta}_{n,N}$	3.98	4.76	5.20	5.04

Table 3.1: Example 1. The estimates corresponding to the functions $l_{n,N}(\theta)$.

Approximation of the transition densities of X by means of the Kalman filter

The following approach is proposed by Pedersen [59, 63].

We consider the stochastic system

$$X_i = D_i X_{i-1} + S_i + \varepsilon_i, \ i = 1, \ldots, n, \tag{3.26}$$

where $(X_i)_{i=0}^{n}$ are random $d \times 1$ vectors, $(D_i)_{i=0}^{n}$ are non-random $d \times d$ matrices, $(S_i)_{i=1}^{n}$ are non-random $d \times 1$ vectors, $X_0 \sim \mathcal{N}_d(x_0, V_0)$, $\varepsilon_i \sim \mathcal{N}_d(0, V_i)$, $i = 1, \ldots, n$ and X_0, $\varepsilon_1, \ldots, \varepsilon_n$ are stochastically independent. The non-random elements in (3.26) including x_0 and $(V_i)_{i=0}^{n}$ are given up to the parameter $\theta \in \Theta \subseteq \mathbb{R}^p$.

Though this chapter is about diffusion models, we nevertheless look here at stochastic difference equations of the type (3.26), since as a particular case of (3.26) we will consider later discretely observed diffusion processes given as solutions to linear stochastic differential equations in the narrow sense. These are equations of the following kind

$$dX_t = (A_t X_t + a_t)dt + B_t dW_t, \ X_0 = x_0, \ t \geq 0, \tag{3.27}$$

where $A : [0, \infty) \mapsto M^{d \times d}$, $a : [0, \infty) \mapsto \mathbb{R}^d$ and $B : [0, \infty) \mapsto M^{d \times m}$ $(d \leq m)$ are deterministic functions of t and W is an m-dimensional Wiener process. Why (3.27) is a particular case of (3.26), will be explained below.

Considering the system (3.26) we distinguish two cases depending on whether X can be observed completely or only partially, i.e. wether all or only a few coordinates of X can be observed. First, if the complete observations of $X_0, X_1, \ldots, X_n$ are given, we can use the log-likelihood function to estimate the parameter θ, see (3.15).

However, the case where $X_0, X_1, \ldots, X_n$ can only be observed partially and possibly with measurement errors often arises in practice. As mentioned, with 'partially observed' we do not mean partially in time but in coordinates of X_i. We assume the observable quantities are $Y_0, Y_1, \ldots, Y_n$ given by

$$Y_i = T_i X_i + U_i + e_i, \; i = 0, 1, \ldots, n, \tag{3.28}$$

where $(T_i)_{i=0}^n$ are non-random $k \times d$ matrices $(k \leq d)$, $(U_i)_{i=0}^n$ are non-random $k \times 1$ vectors, $e_i \sim \mathcal{N}_k(0, W_i)$ and $X_0, \varepsilon_1, \ldots, \varepsilon_m, e_0, e_1, \ldots, e_n$ are stochastically independent, $i = 0, 1, \ldots, n$. The matrices (T_i) specify the observable parts of (X_i), the vectors (U_i) are other inputs and the vectors (e_i) are measurement errors. It is obvious that both the case of complete observations, $Y_i = X_i$, and the case of partial observations without measurement errors are contained in the general case (3.28).

As an application to the case of incomplete observations think of a stochastic volatility model that can be seen as a multi-dimensional process with the volatility process as the unobservable coordinates.

In the case where all non-random elements in (3.26) and (3.28) including x_0, $(V_i)_{i=0}^n$ and $(W_i)_{i=0}^n$ are known, we want to obtain the, in some sense, best predictions of $X_0, X_1, \ldots, X_n$ from given observations of $Y_0, Y_1, \ldots, Y_n$. The conditional expectations $(\mathrm{E}(X_i|Y^i))_{i=0}^n$, respectively $(\mathrm{E}(X_i|Y^{i-1}))_{i=1}^n$, are the best predictors of X_i given $Y^i \equiv (Y_0^T, \ldots, Y_i^T)^T$, respectively given $Y^{i-1} \equiv (Y_0^T, \ldots, Y_{i-1}^T)^T$, in the sense of minimal variance (see Appendix A.1). They can be calculated by means of an iterative procedure called the Kalman-Bucy filter.

The Kalman-Bucy filter and filtering theory in general are well-studied for continuous time stochastic processes with continuous observations (see Liptser and Shiryaev [51], Kallianpur [43], Øksendal [56] and Appendix A). Here we consider discrete time stochastic processes, and discretely observed continuous time stochastic processes in particular. Below we give the iterative procedure used for calculating $(\mathrm{E}(X_i|Y^i))$ and $(\mathrm{E}(X_i|Y^{i-1}))$.

The important point is that the Kalman-Bucy filter gives besides the predictors $(\mathrm{E}(X_i|Y^i))$ and $(\mathrm{E}(X_i|Y^{i-1}))$ the density $p_{i|i-1}$ for the conditional distribution of Y_i given Y^{i-1}. That means we are able to calculate the transition densities of Y and thus the log-likelihood function for θ based on observations of $Y_0, Y_1, \ldots, Y_n$, from which an estimation of θ can be obtained. As Pedersen [59] points out, the unknown vector $x_0 \in \mathbb{R}^d$ should be treated as a part of θ and is not to be chosen at random. The independence of X_0 and e_0 implies $Y_0 \sim \mathcal{N}_k(T_0 x_0 + U_0, T_0 V_0 T_0^T + W_0)$. Hence we can calculate x_0 by $x_0 = T_0^{-1}(Y_0 - U_0)$ a.s. if and only if $k = d$, $V_0 = W_0 = 0$ and T_0 is of full rank. Furthermore, the starting points of the iterative procedure used to calculate the densities $p_{i|i-1}$ depend on x_0.

The Kalman-Bucy filter

Under some technical assumptions (see Pedersen [59], pp. 4–5), we have for given observations $y_0, y_1, \ldots, y_n$ of $Y_0, Y_1, \ldots, Y_n$

$$
\begin{aligned}
X_i | Y^i = y^i \quad &\sim \quad \mathcal{N}_d\left(\mu_i(y^i), \Sigma_i\right), && (3.29)\\
X_i | Y^{i-1} = y^{i-1} \quad &\sim \quad \mathcal{N}_d\left(D_i \mu_{i-1}(y^{i-1}) + S_i, R_i\right), && (3.30)\\
Y_i | Y^{i-1} = y^{i-1} \quad &\sim \quad \mathcal{N}_d\left(T_i(D_i \mu_{i-1}(y^{i-1}) + S_i) + U_i, T_i R_i T_i^T + W_i\right), && (3.31)
\end{aligned}
$$

where $R_i = D_i \Sigma_{i-1} D_i^T + V_i$ is positive definite, and where

$$
\begin{aligned}
\mu_0(y^0) &= x_0 + V_0 T_0^T \left(T_0 V_0 T_0^T + W_0\right)^{-1} (y_0 - T_0 x_0 - U_0), && (3.32)\\
\Sigma_0 &= V_0 - V_0 T_0^T \left(T_0 V_0 T_0^T + W_0\right)^{-1} T_0 V_0, && (3.33)\\
\mu_i(y^i) &= D_i \mu_{i-1}(y^{i-1}) + S_i + R_i T_i^T \left(T_i R_i T_i^T + W_i\right)^{-1} \\
&\quad \left(y_i - T_i \left(D_i \mu_{i-1}(y^{i-1}) + S_i\right) - U_i\right), && (3.34)\\
\Sigma_i &= R_i - R_i T_i^T (T_i R_i T_i^T + W_i)^{-1} T_i R_i. && (3.35)
\end{aligned}
$$

By means of the Kalman-Bucy filter we are now able to calculate $l_n(\theta)$ for every fixed $\theta \in \Theta$ for given observations $y_0, y_1, \ldots, y_n$ of $Y_0, Y_1, \ldots, Y_n$.

The iterative procedure (by means of the Kalman-Bucy filter)

(0) Calculate $\mu_0(y^0)$ and Σ_0 by means of formula (3.32) and (3.33).

(1) Given $\mu_{i-1}(y^{i-1})$ and Σ_{i-1} the conditional distribution of Y_i given Y^{i-1} is known, and so we can calculate the transition density $p_{i|i-1}(y_i | y^{i-1}; \theta)$.

(2) Calculate $\mu_i(y^i)$ and Σ_i by means of formula (3.34) and (3.35) and return to (1).

An Application

As an application of the above theory, we consider the linear stochastic differential equation (3.27). The functions A, a, and B are assumed to be continuous and given up to the unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. Under these assumptions the stochastic differential equation (3.27) has for every fixed $x_0 \in \mathbb{R}^d$ and $\theta \in \Theta$ a unique solution given by

$$X_t = \Phi_t \left(x_0 + \int_0^t \Phi_u^{-1} a_u du + \int_0^t \Phi_u^{-1} B_u dW_u \right) \;,\; t \geq 0 \,, \tag{3.36}$$

where Φ is the deterministic $d \times d$ matrix process solving

$$d\Phi_t = A_t \Phi_t dt, \;\Phi_0 = I_d, \;t \geq 0. \tag{3.37}$$

If

$$A_t \left[\int_0^t A_s ds \right] = \left[\int_0^t A_s ds \right] A_t$$

for all $t > 0$ then

$$\Phi_t = \exp \left[\int_0^t A_s ds \right]$$

is the unique solution to (3.37). We want to estimate θ from possibly incomplete discrete observations of X at time-points $0 = t_0 < t_1 < \ldots < t_n$. From (3.36) we obtain

$$X_t = \Phi_t \Phi_s^{-1} X_s + \int_s^t \Phi_t \Phi_u^{-1} a_u du + \int_s^t \Phi_t \Phi_u^{-1} B_u dW_u$$

for $0 \leq s < t$, and hence the Markov chain $\{X_{t_i}\}_{i=0}^n$ can be represented by the stochastic system

$$X_{t_i} = \Phi_{t_i} \Phi_{t_{i-1}}^{-1} X_{t_{i-1}} + \int_{t_{i-1}}^{t_i} \Phi_{t_i} \Phi_s^{-1} a_s ds + \varepsilon_{t_i} \;,\; i = 1, \ldots, n, \tag{3.38}$$

where $\varepsilon_{t_i}, \ldots, \varepsilon_{t_n}$ is a sequence of stochastically independent random $d \times 1$ vectors with $\varepsilon_{t_i} \sim \mathcal{N}_d(0, V_{t_i})$ and

$$V_{t_i} = \int_{t_{i-1}}^{t_i} (\Phi_{t_i} \Phi_s^{-1} B_s)(\Phi_{t_i} \Phi_s^{-1} B_s)^T ds \;,\; i = 1, \ldots, n. \tag{3.39}$$

All assumptions in Pedersen [59], pp. 4–5, for the Kalman-Bucy filter (see (3.29)–(3.35)) can be made to be satisfied for (3.38). Thus (3.38) is a particular case of (3.26) with

$$D_{t_i} = \Phi_{t_i} \Phi_{t_{i-1}}^{-1} \,, \tag{3.40}$$

$$S_{t_i} = \int_{t_{i-1}}^{t_i} \Phi_{t_i} \Phi_s^{-1} a_s ds. \tag{3.41}$$

We can use the previous results to estimate θ from possibly incomplete observations of $\{X_{t_i}\}_{i=0}^n$ of the type given by (3.28).

Note that in many applications, the non-random elements D_{t_i} (3.40), S_{t_i} (3.41) and V_{t_i} (3.39), can not be calculated exactly. The solution to (3.37) may be unknown. In that case $\{X_{t_i}\}_{i=0}^n$ can be approximated by the Euler approximation (see Appendix B.1). But even if the solution to (3.37) is known, S_{t_i} and V_{t_i} often can not be calculated exactly, and hence have to be approximated; this can be done by different methods depending on the concrete application.

Martingale estimating functions

The following approach is proposed by Bibby and Sørensen [8].

We consider one-dimensional diffusion processes defined by the stochastic differential equations

$$dX_t = b(X_t;\theta)\,dt + \sigma(X_t;\theta)\,dW_t, \tag{3.42}$$

where $X_0 = x_0$ and $t \geq 0$. Besides the usual assumptions on b and σ in (3.1), such that (3.42) has a unique solution for all θ in an open subset $\Theta \subseteq \mathbb{R}$, the functions b and σ are supposed to be twice continuously differentiable with respect to both arguments and σ is assumed to be positive. In contrast to (3.1), for convenience here we only consider the time-homogeneous case. To simplify the exposition further, assume that we can observe $\{X_t\}$ at discrete equidistant time points, say $\Delta, 2\Delta, \ldots, n\Delta$. Later we will give an extension to the case where X and θ are multi-dimensional.

Our goal is to estimate the parameter θ from these discrete observations $X_\Delta, X_{2\Delta}, \ldots, X_{n\Delta}$ of $\{X_t\}$. Inference from discrete time observations can be based on an approximation of the score function of the continuous log-likelihood function $l_t(\theta)$. Denote this approximation by $\dot{\tilde{l}}_n(\theta)$. For the definition of the continuous log-likelihood see 3.1.1, p.17. In the case where σ does not depend on θ the continuous time log-likelihood function is

$$l_t(\theta) = \int_0^t \frac{b(X_s;\theta)}{\sigma^2(X_s)}\,dX_s - \frac{1}{2}\int_0^t \frac{b^2(X_s;\theta)}{\sigma^2(X_s)}\,ds, \tag{3.43}$$

see also (3.5). If we replace the Lebesgue integrals and the Itô integrals by Riemann-Itô sums and differentiate with respect to θ we get the approximate score function

$$\dot{\tilde{l}}_n(\theta) = \sum_{i=1}^n \frac{\dot{b}(X_{(i-1)\Delta};\theta)}{\sigma^2(X_{(i-1)\Delta})}(X_{i\Delta} - X_{(i-1)\Delta}) - \Delta\sum_{i=1}^n \frac{b(X_{(i-1)\Delta};\theta)\,\dot{b}(X_{(i-1)\Delta};\theta)}{\sigma^2(X_{(i-1)\Delta})}\,. \tag{3.44}$$

If σ depends on θ we use the same estimating function

$$\dot{\tilde{l}}_n(\theta) = \sum_{i=1}^{n} \frac{\dot{b}(X_{(i-1)\Delta};\theta)}{\sigma^2(X_{(i-1)\Delta};\theta)}(X_{i\Delta} - X_{(i-1)\Delta}) - \Delta \sum_{i=1}^{n} \frac{b(X_{(i-1)\Delta};\theta)\,\dot{b}(X_{(i-1)\Delta};\theta)}{\sigma^2(X_{(i-1)\Delta};\theta)}\,. \tag{3.45}$$

By using this approach for the estimation of θ the problem of inconsistency arises as already mentioned in the introduction of this section, p.23. To avoid this problem we can use martingale estimating functions of which we will construct four different types. The idea is to modify the discretized score-function $\dot{\tilde{l}}_n(\theta)$ in such a way that a zero-mean P_θ-martingale is obtained. Then the estimator can be shown to be consistent and asymptotically normal.

(1) Our first approach is to **compensate** $\dot{\tilde{l}}_n$, so that a martingale $\tilde{G}_n$ is obtained.

By substracting from $\dot{\tilde{l}}_n(\theta)$ its compensator we get a zero-mean P_θ-martingale with respect to the filtration defined by $\mathcal{F}_i = \sigma(X_\Delta, \ldots, X_{i\Delta})$, $i = 1, 2, \ldots$. The compensator is:

$$\begin{aligned}\sum_{i=1}^{n} E_\theta\left(\dot{\tilde{l}}_i(\theta) - \dot{\tilde{l}}_{i-1}(\theta)|\mathcal{F}_{i-1}\right) \quad &= \quad \sum_{i=1}^{n} \frac{\dot{b}(X_{(i-1)\Delta};\theta)}{\sigma^2(X_{(i-1)\Delta};\theta)}(F(X_{(i-1)\Delta};\theta) - X_{(i-1)\Delta}) \\ &\quad -\Delta \sum_{i=1}^{n} \frac{b(X_{(i-1)\Delta};\theta)\,\dot{b}(X_{(i-1)\Delta};\theta)}{\sigma^2(X_{(i-1)\Delta};\theta)},\end{aligned} \tag{3.46}$$

where

$$F(X_{(i-1)\Delta};\theta) \equiv E_\theta(X_{i\Delta}|X_{(i-1)\Delta}). \tag{3.47}$$

Thus we obtain a zero-mean martingale estimating function of the form

$$\tilde{G}_n(\theta) = \sum_{i=1}^{n} \frac{\dot{b}(X_{(i-1)\Delta};\theta)}{\sigma^2(X_{(i-1)\Delta};\theta)} \left(X_{i\Delta} - F(X_{(i-1)\Delta};\theta)\right)\,. \tag{3.48}$$

(2) Alternatively we consider the **general class of zero-mean P_θ-martingale estimating functions**

$$G_n(\theta) = \sum_{i=1}^{n} g_{i-1}(X_{(i-1)\Delta};\theta)\left(X_{i\Delta} - F(X_{(i-1)\Delta};\theta)\right)\,, \tag{3.49}$$

where for $i = 1, \ldots, n$, the function g_{i-1} is $\mathcal{F}_{i-1}$-measurable and continuously differentiable in θ. The optimal estimating function within the class (3.49) in the asymptotic sense of Godambe and Heyde [33] is

$$G_n^*(\theta) = \sum_{i=1}^{n} \frac{\dot{F}(X_{(i-1)\Delta};\theta)}{\phi(X_{(i-1)\Delta};\theta)}\left(X_{i\Delta} - F(X_{(i-1)\Delta};\theta)\right)\,, \tag{3.50}$$

where

$$\phi(X_{(i-1)\Delta};\theta) = E_\theta\left[(X_{i\Delta} - F(X_{(i-1)\Delta};\theta))^2 | X_{(i-1)\Delta}\right],\ i = 1,\dots,n. \quad (3.51)$$

The function $G_n^*(\theta)$ is within the class (3.49) in some sense "closest" to the score function based on the usually unknown exact likelihood function.

It should be mentioned here that for small Δ the martingale estimating function $\tilde{G}_n(\theta)$, defined in (3.48), is a first order approximation in Δ of G_n^*, that means $\tilde{G}_n(\theta)$ is approximately optimal.

(3) In some cases there are possibly **numerical problems** in calculating $\dot{F}(x,\theta)$ in (3.50). One way to solve these problems involves approximating $\dot{F}(x,\theta)$ up to the order $O(\Delta^2)$. That leads to a **third martingale estimating function**

$$\begin{aligned} G_n^+(\theta) &= \sum_{i=1}^n \Big[\dot{b}(X_{(i-1)\Delta};\theta)\Delta + \frac{1}{2}\Delta^2\left(\dot{b}(X_{(i-1)\Delta};\theta)b'(X_{(i-1)\Delta};\theta)\right. \\ &\quad + b(X_{(i-1)\Delta};\theta)\dot{b}'(X_{(i-1)\Delta};\theta) + \frac{1}{2}(\dot{\sigma}^2(X_{(i-1)\Delta};\theta)b''(X_{(i-1)\Delta};\theta) \\ &\quad \left. + \sigma^2(X_{(i-1)\Delta};\theta)\dot{b}''(X_{(i-1)\Delta};\theta))\right)\Big] \frac{(X_{i\Delta} - F(X_{(i-1)\Delta};\theta))}{\phi(X_{(i-1)\Delta};\theta)}. \quad (3.52) \end{aligned}$$

Altogether we have now found expressions for three different zero-mean P_θ-martingale estimating functions.

As for the multi-dimensional case, suppose θ is k-dimensional, $\{X_t\}$ and $b(X_t;\theta)$ are d-dimensional, σ is a $d \times m$-dimensional matrix with $\sigma\sigma^T$ positive definite and the Wiener process $\{W_t\}$ is m-dimensional. Then the $k \times 1$-dimensional martingale estimating functions $\tilde{G}_n$ and G_n^* have the form

$$\begin{aligned} \tilde{G}_n(\theta) &= \sum_{i=1}^n \dot{b}(X_{(i-1)\Delta};\theta)^T \left(\sigma(X_{(i-1)\Delta};\theta)(\sigma(X_{(i-1)\Delta};\theta)^T)\right)^{-1} \\ &\quad \cdot \left(X_{i\Delta} - F(X_{(i-1)\Delta};\theta)\right) \end{aligned}$$

and

$$G_n^*(\theta) = \sum_{i=1}^n \dot{F}(X_{(i-1)\Delta};\theta)^T \phi(X_{(i-1)\Delta};\theta)^{-1}\left(X_{i\Delta} - F(X_{(i-1)\Delta};\theta)\right),$$

where ϕ is assumed to be positive definite and $\dot{b}$ and $\dot{F}$ denote the $d \times k$-dimensional matrices of partial derivatives with respect to the components of θ.

The asymptotic properties of the estimator $\hat{\theta}_n$ we obtain from the martingale estimating functions (3.48), (3.50) and (3.52), or more generally from the class of martingale estimating functions G_n of the form (3.49), are discussed by Bibby and Sørensen [8]. Under natural regularity conditions (see Bibby and Sørensen [8], pp. 7–9) we have

Theorem 4 *An estimator $\hat{\theta}_n$, which solves the equation*

$$G_n(\hat{\theta}_n) = 0,$$

exists with probability tending to one as $n \longrightarrow \infty$ under P_{θ_0}. Moreover, as $n \longrightarrow \infty$,

$$\hat{\theta}_n \longrightarrow \theta_0$$

in probability under P_{θ_0} and $\hat{\theta}_n$ is asymptotically normal in distribution under P_{θ_0}.

For the proof we refer to [8].

As a first example we consider the Ornstein-Uhlenbeck process where the transition densities are well-known.

Example 1
The Ornstein-Uhlenbeck process is the solution of the stochastic differential equation

$$dX_t = \theta X_t \, dt + \sigma \, dW_t, \tag{3.53}$$

with $X_0 = x_0$. In this case the drift coefficient is $b(x;\theta) = \theta x$ and the diffusion coefficient $\sigma(x,\theta) \equiv \sigma$ is assumed to be known. The transition probability is normal with mean $F(x;\theta) = x\, e^{\theta\Delta}$ and variance $\phi(\theta) = \frac{\sigma^2}{2\theta}(e^{2\theta\Delta} - 1)$. Hence the estimating function $\tilde{G}_n$ has the form

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

and we obtain $\hat{\theta}_n$ as solution of $\tilde{G}_n(\theta) = 0$:

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

The estimators $\hat{\theta}_n$ we obtain from the martingale estimating functions G^+ and G^* are the same because G^+ and G^* are proportional to $\tilde{G}$.

In the next example we consider a wider class of stochastic processes.

Example 2
The solutions of the stochastic differential equation

$$dX_t = (\alpha + \theta X_t)\, dt + \psi(X_t)\, dW_t, \tag{3.54}$$

where $X_0 = x_0$ and the function ψ takes positive values in $\mathbb{R}$, are called mean-reverting processes (see also model (1.5)). The unknown parameters are α and θ. Our aim is to be able to calculate the martingale estimating functions $\tilde{G}_n$ and G_n^*.

Lemma 1 *The function*

$$f(t) \equiv E_{\alpha,\theta}(X_t|X_0)$$

solves

$$f'(t) = \alpha + \theta f(t). \tag{3.55}$$

Proof: Write (3.54) in integral form

$$X_t = X_0 + \int_0^t (\alpha + \theta X_s) ds + \int_0^t \psi(X_s) dW_s.$$

Conditioning on X_0 we have

$$\begin{aligned} \mathrm{E}_{\alpha,\theta}(X_t|X_0) &= \mathrm{E}_{\alpha,\theta}(X_0|X_0) + \mathrm{E}_{\alpha,\theta}\left[\int_0^t (\alpha + \theta X_s) ds | X_0\right] \\ &\quad + \underbrace{\mathrm{E}_{\alpha,\theta}\left[\int_0^t \psi(X_s) dW_s | X_0\right]}_{=0}, \end{aligned}$$

and equivalently

$$\mathrm{E}_{\alpha,\theta}(X_t|X_0) = X_0 + \alpha t + \theta \int_0^t \mathrm{E}_{\alpha,\theta}(X_s|X_0) ds.$$

We conclude

$$\frac{d\mathrm{E}_{\alpha,\theta}(X_t|X_0)}{dt} = \alpha + \theta \mathrm{E}_{\alpha,\theta}(X_t|X_0),$$

and the claim follows. Note that the function $f_r(t) = \mathrm{E}_{\alpha,\theta}(X_t|X_r)$, $0 \le r \le t$, also solves (3.55), for the proof stays the same apart from

$$\mathrm{E}_{\alpha,\theta}\left[\int_0^t \psi(X_s) dW_s | X_r\right] = \int_0^r \psi(X_s) dW_s,$$

which is independent of t and thus plays no role in the derivative. □

Corollary 2 *For*

$$F(X_{(i-1)\Delta};\alpha,\theta)=E_{\alpha,\theta}\left(X_{i\Delta}|X_{(i-1)\Delta}\right)$$

we have

$$F(x;\alpha,\theta)=xe^{\theta\Delta}+\frac{\alpha}{\theta}(e^{\theta\Delta}-1). \tag{3.56}$$

Proof: The solution of

$$f'(t)=\alpha+\theta f(t),\quad f(t_0)=f_0,\ t\geq t_0,$$

is

$$f(t)=f_0\,e^{\theta(t-t_0)}+\frac{\alpha}{\theta}(e^{\theta(t-t_0)}-1).$$

Hence we have for $\mathrm{E}_{\alpha,\theta}(X_{t_i}|X_{t_{i-1}})\equiv f(t_i)$ with constant $\Delta\equiv t_i-t_{i-1}$ for all i

$$\begin{aligned}\mathrm{E}_{\alpha,\theta}(X_{t_i}|X_{t_{i-1}}) &= \mathrm{E}(X_{t_{i-1}}|X_{t_{i-1}})\,e^{\theta\Delta}+\frac{\alpha}{\theta}(e^{\theta\Delta}-1)\\ &= X_{t_{i-1}}e^{\theta\Delta}+\frac{\alpha}{\theta}(e^{\theta\Delta}-1),\end{aligned}$$

and the claim follows. □

With (3.56) we are now able to calculate $\tilde{G}_n$ and G_n^* in the following way

$$\begin{aligned}\tilde{G}_n(\alpha,\theta)=\Bigg[&\sum_{i=1}^{n}\frac{1}{\psi^2(X_{(i-1)\Delta})}\left(X_{i\Delta}-X_{(i-1)\Delta}e^{\theta\Delta}+\frac{\alpha}{\theta}(1-e^{\theta\Delta})\right),\\ &\sum_{i=1}^{n}\frac{X_{(i-1)\Delta}}{\psi^2(X_{(i-1)\Delta})}\left(X_{i\Delta}-X_{(i-1)\Delta}e^{\theta\Delta}+\frac{\alpha}{\theta}(1-e^{\theta\Delta})\right)\Bigg]^T,\end{aligned}$$

$$\begin{aligned}G_n^*(\alpha,\theta)=\Bigg[&\sum_{i=1}^{n}\frac{e^{\theta\Delta}-1}{\theta\phi(X_{(i-1)\Delta};\alpha,\theta)}\left(X_{i\Delta}-X_{(i-1)\Delta}e^{\theta\Delta}+\frac{\alpha}{\theta}(1-e^{\theta\Delta})\right),\\ &\sum_{i=1}^{n}\frac{\Delta e^{\theta\Delta}(X_{(i-1)\Delta}+\frac{\alpha}{\theta})+\frac{\alpha}{\theta^2}(1-e^{\theta\Delta})}{\phi(X_{(i-1)\Delta};\alpha,\theta)}\left(X_{i\Delta}-X_{(i-1)\Delta}e^{\theta\Delta}+\frac{\alpha}{\theta}(1-e^{\theta\Delta})\right)\Bigg]^T.\end{aligned}$$

Considering G^+ is not interesting since F is known.

The estimation equation $\tilde{G}_n(\alpha,\theta)=0$ can be solved explicitly. Abbreviating $\psi_{i-1}^2\equiv\psi^2(X_{(i-1)\Delta})$ we obtain

$$e^{\tilde{\theta}_n\Delta}=\frac{\left(\sum_{i=1}^n\frac{X_{(i-1)\Delta}}{\psi_{i-1}^2}\right)\left(\sum_{i=1}^n\frac{X_{i\Delta}}{\psi_{i-1}^2}\right)-\left(\sum_{i=1}^n\frac{X_{(i-1)\Delta}X_{i\Delta}}{\psi_{i-1}^2}\right)\left(\sum_{i=1}^n\frac{1}{\psi_{i-1}^2}\right)}{\left(\sum_{i=1}^n\frac{X_{(i-1)\Delta}}{\psi_{i-1}^2}\right)^2-\left(\sum_{i=1}^n\frac{X_{(i-1)\Delta}^2}{\psi_{i-1}^2}\right)\left(\sum_{i=1}^n\frac{1}{\psi_{i-1}^2}\right)} \tag{3.57}$$

and

$$\tilde{\alpha}_n = \frac{\tilde{\theta}_n}{1 - e^{\tilde{\theta}_n \Delta}} \frac{\left(\sum_{i=1}^n \frac{X_{(i-1)\Delta}}{\psi_{i-1}^2}\right) e^{\tilde{\theta}_n \Delta} - \left(\sum_{i=1}^n \frac{X_{i\Delta}}{\psi_{i-1}^2}\right)}{\left(\sum_{i=1}^n \frac{1}{\psi_{i-1}^2}\right)}. \tag{3.58}$$

As regarding the martingale estimating function G_n^* we are able to find a closed expression for ϕ only in a few cases where the diffusion coefficient is rather simple. For instance, if $\psi(x) = \sigma\sqrt{x}$ (as in the square root diffusion model (1.5), the Cox Ingersoll Ross model) we have with a similar argument as for F (that is, the conditional second moment solves an ordinary differential equation)

$$\phi(x; \alpha, \theta) = \frac{\sigma^2}{2\theta^2}\left((\alpha + 2\theta x)e^{2\theta\Delta} - 2(\alpha + \theta x)e^{\theta\Delta} + \alpha\right).$$

As an extension we consider the mean-reverting process where $\theta > 0$ enters the diffusion coefficient

$$dX_t = -\theta X_t dt + \sqrt{\theta + X_t^2}\, dW_t.$$

With the same argument as above we obtain the conditional variance

$$\phi(x; \alpha, \theta) = x^2 e^{-2\theta\Delta}(e^{\Delta} - 1) + \frac{\theta}{2\theta - 1}(1 - e^{(1-2\theta)\Delta}).$$

We remark that in practical applications the estimating equations corresponding to G_n^* can be solved using a generalization of Newton's method.

Furthermore, if no explicit expressions for the conditional mean F and the conditional variance ϕ are known, then F and ϕ can be approximated by the sample mean and sample variance of a large number of simulated realizations of the diffusion process at the relevant time point.

(4) As an extension we shall **combine martingale estimating functions** in an 'optimal' way.

The following approach is proposed by Bibby [5].

If the unknown parameter θ is multi-dimensional and a part of θ is only found in the diffusion coefficient, then using the martingale estimating functions generated by the conditional mean, $\tilde{G}_n$ and G_n^*, leads to fewer estimation equations than parameters. That is in this case neither G^* nor $\tilde{G}$ can be used to estimate the part of θ that only enters the diffusion coefficient. This disadvantage of $\tilde{G}_n$ and G_n^* motivates us to consider **martingale estimating functions generated by higher order conditional moments** for example by the conditional variance:

$$H_n(\theta) = \sum_{i=1}^{n} h(X_{(i-1)\Delta};\theta)\left[\left(X_{i\Delta} - F(X_{(i-1)\Delta};\theta)\right)^2 - \phi(X_{(i-1)\Delta};\theta)\right], \quad (3.59)$$

with ϕ given by (3.51) and F given by (3.47). In analogy to (3.50) we obtain the optimal estimating function within the class (3.59), in the sense of Godambe and Heyde [33]. This optimal function takes the form

$$H_n^*(\theta) = \sum_{i=1}^{n} \frac{\dot{\phi}(X_{(i-1)\Delta};\theta)}{\psi(X_{(i-1)\Delta};\theta)}\left[\left(X_{i\Delta} - F(X_{(i-1)\Delta};\theta)\right)^2 - \phi(X_{(i-1)\Delta};\theta)\right], \quad (3.60)$$

where ϕ is assumed to be differentiable in θ and ψ is the fourth conditional cumulant

$$\psi(X_{(i-1)\Delta};\theta) = E_\theta\left[\left(X_{i\Delta} - F(X_{(i-1)\Delta};\theta)\right)^4 |X_{(i-1)\Delta}\right] - \phi(X_{(i-1)\Delta};\theta)^2,$$

where $i = 1,\dots,n$.

Combining the functions G_n and H_n leads to further martingale estimating functions that may have better properties. Following the optimal way of combining G_n and H_n described in Heyde [35], the function K_n^* is obtained:

$$\begin{aligned} K_n^*(\theta) &= \sum_{i=1}^{n}\left[\frac{\dot{\phi}(\theta)\eta(\theta) - F(\theta)\psi(\theta)}{\phi(\theta)\psi(\theta) - \eta^2(\theta)}(X_{i\Delta} - F(\theta))\right. \\ &\quad \left. + \frac{\dot{F}(\theta)\eta(\theta) - \dot{\phi}(\theta)\phi(\theta)}{\phi(\theta)\psi(\theta) - \eta^2(\theta)}\left((X_{i\Delta} - F(\theta))^2 - \phi(\theta)\right)\right], \quad (3.61)\end{aligned}$$

where for abbreviation the first argument of all functions on the right hand side, that is $X_{(i-1)\Delta}$, has been left out and where η denotes the third conditional central moment

$$\eta(X_{(i-1)\Delta};\theta) = E_\theta\left[(X_{i\Delta} - F(X_{(i-1)\Delta};\theta))^3 |X_{(i-1)\Delta}\right], \; i = 1,\dots,n.$$

For K_n^*, as for G_n, it can be shown that an estimator obtained from the estimating equation $K_n^*(\theta) = 0$ exists, is consistent and asymptotically normal. Under some regularity conditions (see Bibby [5], pp. 4–6) we have

Theorem 5 *An estimator $\hat{\theta}_n$ exists for every n, which on a set C_n solves the equation*

$$K_n(\hat{\theta}_n) = 0,$$

where $P_{\theta_0}(C_n) \longrightarrow 1$ *as* $n \longrightarrow \infty$. *Moreover, as* $n \longrightarrow \infty$,

$$\hat{\theta}_n \longrightarrow \theta_0$$

in probability under P_{θ_0} *and* $\hat{\theta}_n$ *is asymptotically normal in distribution under* P_{θ_0}.

For the proof we refer to [5, 8].

3.2 Discrete models

In section 1.2 we dealt with three kinds of stochastic volatility models, following an AR(p)-process, an ARCH or even a GARCH process. Now we discuss Maximum Likelihood (ML) estimation for these models and derive the properties of the estimators. Finally, we treat the Bayesian analysis.

3.2.1 AR models

In our representation of estimation theory for the AR(1)-process we follow [67] §5. The parameter estimation problem for linear time series (i.e. ARMA processes) is to be found in many textbooks, see for instance [16], [34], [64]. Below we discuss a slightly different approach which will turn out to be useful in more general models. The AR(1) case should therefore be seen as a pedagogic example with respect to this more general methodology.

Consider the autoregressive AR(1) model

$$X_n = \theta X_{n-1} + \varepsilon_n, \quad n \geq 1, \tag{3.62}$$

where $\theta \in \Theta \subseteq \mathbb{R}$ is the parameter to be estimated, X_0 is a random variable with $\mathrm{E}(X_0) = 0$ and $\{\varepsilon_n, n \geq 1\}$ is 'noise', where $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ i.i.d. with known σ^2.

From equation (3.62) we conclude

$$X_n = \varepsilon_n + \theta\varepsilon_{n-1} + \ldots + \theta^{n-1}\varepsilon_1 + \theta^n X_0,$$

and thus the probabilistic properties of $\{X_n\}$ essentially depend on the joint distribution of $X_0, \varepsilon_1, \varepsilon_2, \ldots$. We have

$$\mathrm{Var}_\theta X_n = \sigma^2(1 + \theta^2 + \ldots + \theta^{2(n-1)}) + \theta^{2n}\mathrm{Var}_\theta X_0 \tag{3.63}$$

and

$$\mathrm{E}_\theta X_n X_{n-k} = \sigma^2\theta^k(1 + \theta^2 + \ldots + \theta^{2(n-k-1)}) + \theta^k\theta^{2(n-k)}\mathrm{Var}_\theta X_0. \tag{3.64}$$

In the case $|\theta| < 1$, we obtain from (3.63) and (3.64) by choosing $X_0 \sim \mathcal{N}(0, \frac{\sigma^2}{1-\theta^2})$

$$\mathrm{E}_\theta X_n = 0, \quad \mathrm{Var}_\theta X_n = \frac{\sigma^2}{1-\theta^2}, \quad \mathrm{E}_\theta X_n X_{n-k} = \frac{\sigma^2\theta^k}{1-\theta^2},$$

and $\{X_n\}$ has a stationary distribution.

In the case $|\theta| \geq 1$ we have $\mathrm{Var}_\theta X_n \longrightarrow \infty$ as $n \longrightarrow \infty$, that is the process explodes.

In both cases $\theta = \pm 1$, $\{X_n\}$ reduces to a random walk.

From these considerations we suppose that as for probabilistic properties of estimators of θ, e.g. asymptotic behaviour, we have to distinguish between the cases $|\theta| > 1$, $|\theta| < 1$ and $|\theta| = 1$.

In the following we assume for the AR(1) model (3.62) $X_0 = 0$ and $\sigma^2 = 1$ for convenience.

Denoting by $p_\theta(X_1, \ldots, X_n)$ the joint density of $X_1, \ldots, X_n$, the Maximum Likelihood Estimator (MLE) $\hat{\theta}_n$ is defined to be a value θ such that the joint density p_θ reaches a maximum that is

$$\hat{\theta}_n = \operatorname{argmax} p_\theta(X_1, \ldots, X_n),$$

(for the definition of the MLE in continuous time see §3.1.1, p.17). We know the joint density p_θ of $X_1, \ldots, X_n$

$$p_\theta(X_1, \ldots, X_n) = (2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \theta X_{i-1})^2\right],$$

and hence are able to calculate $\hat{\theta}_n$ by solving $\frac{d}{d\theta} p_\theta = 0$

$$\hat{\theta}_n = \frac{\sum_{i=1}^n X_{i-1} X_i}{\sum_{i=1}^n X_{i-1}^2}.$$

Inserting (3.62) for X_i we obtain

$$\hat{\theta}_n = \theta + \frac{\sum_{i=1}^n X_{i-1}\varepsilon_i}{\sum_{i=1}^n X_{i-1}^2} \quad P_\theta \; a.s. \tag{3.65}$$

Denoting

$$M_n = \sum_{i=1}^n X_{i-1}\varepsilon_i,$$

we see immediately that the process M_n is a martingale with respect to the filtration $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ under P_θ for any θ. The martingale M_n is square integrable, i.e. $\mathrm{E}M_n^2 < \infty$, $n \geq 0$, and we know that the stochastic sequence M_n^2 is a submartingale (see [66] §7.1, p.455). By means of the Doob decomposition (see e.g. [66], p.454) there is a martingale m_n and a predictable[2] increasing[3] sequence $\langle M \rangle_n$ such that

$$M_n^2 = m_n + \langle M \rangle_n.$$

[2] A process X_n is predictable if X_n is $\mathcal{F}_{n-1}$ measurable.

[3] A process X_n is increasing if $X_0 = 0$, $X_n \leq X_{n+1}$ P a.s.

The sequence $\langle M\rangle_n$ is called the square characteristic or the quadratic variation of M_n. Here $\langle M\rangle_n$ is

$$\langle M\rangle_n = \sum_{i=1}^{n} X_{i-1}^2,$$

see [66], p.455.

Hence, equation (3.65) can be written as

$$\hat{\theta}_n - \theta = \frac{M_n}{\langle M\rangle_n}. \tag{3.66}$$

Recall the definition of the Fisher information in the continuous time case (see 3.1.1, p.19). Here in discrete time the Fisher information equals

$$I_n(\theta) = \mathrm{E}_\theta\left[-\frac{\partial^2 \ln p_\theta(x_1,\ldots,x_n)}{\partial\theta^2}\right].$$

Direct calculation shows

$$I_n(\theta) = \mathrm{E}_\theta \sum X_{i-1}^2,$$

hence

$$I_n(\theta) = \mathrm{E}_\theta\langle M\rangle_n,$$

and therefore $\langle M\rangle_n$ is often called the stochastic Fisher information. Calculations based on (3.63) show that for large n the Fisher information is approximately

$$I_n(\theta) \sim \begin{cases} \frac{n}{1-\theta^2}, & |\theta| < 1, \\ \frac{n^2}{2}, & |\theta| = 1, \\ \frac{\theta^{2n}}{(\theta^2-1)^2}, & |\theta| > 1. \end{cases}$$

Since

$$\langle M\rangle_n \longrightarrow \infty \quad P_\theta \ a.s.,$$

we can apply the 'law of large numbers for square integrable martingales' and obtain

$$\frac{M_n}{\langle M\rangle_n} \longrightarrow 0 \quad P_\theta \ a.s.,$$

see [66], p.487, Theorem 4. Thus, with (3.66) we conclude that the estimator is strongly consistent, that is

$$\hat{\theta}_n \longrightarrow \theta \qquad P_\theta \text{ a.s.}$$

as n tends to infinity.

As for asymptotic behaviour of the estimator we only state the results and refer to [67] §5 for a detailed treatment.

Theorem 6 *Depending on the value of θ we have the following asymptotic behaviour for the normalized deviation $\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta)$:*

$$\lim_{n \longrightarrow \infty} P_\theta \left[\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta) \leq z \right] = \begin{cases} \Phi(z), & |\theta| < 1, \\ H_\theta(z), & |\theta| = 1, \\ Ch(z), & |\theta| > 1, \end{cases}$$

where $\Phi(z)$ is the standard normal distribution, $Ch(z)$ is the Cauchy distribution and $H_\theta(z)$ is the distribution of the random variable

$$\theta \frac{W^2(1) - 1}{2^{\frac{3}{2}} \int_0^1 W^2(s) ds},$$

where W denotes a Wiener process.

Hence, in the stationary case $|\theta| < 1$ the normalized deviation $\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta)$ is asymptotically normally distributed, whereas in the other cases it has as limit distribution the Cauchy distribution or the quite unexpected H_θ distribution. The question arises wether we can reduce the number of limit distributions by modifying the normalizing factor $\sqrt{I_n(\theta)}$. Indeed, instead of using the Fisher information $I_n(\theta) = \mathrm{E}\langle M \rangle_n$, choosing the stochastic Fisher information $\langle M \rangle_n$ as normalizing factor we obtain

Theorem 7

$$\lim_{n \longrightarrow \infty} P_\theta \left[\sqrt{\langle M \rangle_n}\, (\hat{\theta}_n - \theta) \leq z \right] = \begin{cases} \Phi(z), & |\theta| \neq 1, \\ H_\theta(z), & |\theta| = 1. \end{cases}$$

Summarizing: the MLE $\hat{\theta}_n$ is (strongly) consistent and the normalized deviations $\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta)$ and $\sqrt{\langle M \rangle_n}(\hat{\theta}_n - \theta)$ are asymptotically distributed as shown in both theorems.

3.2.2 ARCH and GARCH models

In the following we concentrate on ARCH and GARCH models and especially on estimation in these models. We refer to the fundamental papers by Engle [22] and Bollerslev [12] and to Bollerslev, Chou and Kroner [13] and Bollerslev, Engle and Nelson [14].

Conventional econometric time series models assume constant variance. However, over a decade ago risk and uncertainty considerations lead to the development of new econometric time series models that allow for the modeling of time varying variances and covariances. Engle [22] introduced such a new class of stochastic processes, called the class of AutoRegressive Conditional Heteroscedastic (ARCH) processes where the conditional variances and covariances depend on the past. These are discrete time stochastic processes $\{\varepsilon_t\}$ of the form

$$\varepsilon_t = z_t \sigma_t \tag{3.67}$$

with z_t i. i. d. , E(z_t)=0, Var(z_t)=1, and with time-varying positive σ_t which is time $(t-1)$ measurable. Assume in the following $z_t \sim \mathcal{N}(0,1)$ and ε_t is a scalar process. The extension to the multivariate case is straightforward, see e.g. [13].

First we focus directly on the process $\{\varepsilon_t\}$ (3.67) and assume that ε_t is itself observable. However, note that in many applications ε_t corresponds to the innovations for some other stochastic process y_t, where $y_t = f(b, x_t) + \varepsilon_t$ with ε_t conditionally distributed $\mathcal{N}(0, \sigma^2)$, f a function of x_t which is time $(t-1)$ measurable and of a parameter vector b. In this context we will consider later the ARCH regression model, see (3.75), where $f(b, x_t) \equiv x_t' b$, that is the conditional mean of y_t is given as $x_t' b$.

The conditional variance of ε_t equals σ_t^2. We will concentrate in the following on some frequently used models for σ_t^2. Engle [22] suggests one possible parametrization for σ_t^2 as a linear function of the past q squared values of the process

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2, \tag{3.68}$$

where for the model to be well-defined $\alpha_0 > 0$ and $\alpha_i \geq 0$ for $i = 1, \ldots, q$. This model is known as the linear ARCH(q) model.

Now we discuss Maximum Likelihood (ML) estimation in the linear ARCH(q) model. For a discussion of the ML estimation see section 3.1.1. Apart from some constants, the log-likelihood of the tth observation is

$$l_t = -\frac{1}{2} \log \sigma_t^2 - \frac{1}{2} \varepsilon_t^2 / \sigma_t^2, \tag{3.69}$$

where the term $-\frac{1}{2} \log \sigma_t^2$ arises from the transformation from z_t to ε_t. The log-likelihood function for the full sample $\varepsilon_T, \varepsilon_{T-1}, \ldots, \varepsilon_1$ is

$$L = \frac{1}{T} \sum_{t=1}^{T} l_t. \tag{3.70}$$

To estimate the unknown parameters $\alpha' \equiv (\alpha_0, \alpha_1, \ldots, \alpha_q)$ we maximize the log-likelihood function, that is the first order derivatives are set equal to zero.

Denoting $u_t' \equiv (1, \varepsilon_{t-1}^2, \ldots, \varepsilon_{t-q}^2)$ and $h_t \equiv \sigma_t^2$ we abbreviate (3.68) by

$$h_t = u_t'\alpha.$$

With this notation the first order derivatives are

$$\frac{\partial l_t}{\partial \alpha} = \frac{1}{2h_t} u_t \left(\frac{\varepsilon_t^2}{h_t} - 1 \right). \tag{3.71}$$

In order to obtain the limiting distributions for the normalized deviations of the estimators in (3.76) below, we have to estimate the Fisher information matrix $\mathcal{F}$, see §3.1.1, p.19, which is the negative expectation of the Hessian averaged over all observations. Therefore we calculate the Hessian

$$\frac{\partial^2 l_t}{\partial \alpha \partial \alpha'} = -\frac{1}{2h_t^2} u_t \, u_t' \left(\frac{\varepsilon_t^2}{h_t} \right) + \left[\frac{\varepsilon_t^2}{h_t} - 1 \right] \frac{\partial}{\partial \alpha'} \left[\frac{1}{2h_t} u_t \right]. \tag{3.72}$$

Since the conditional expectation of the factor ε_t^2/h_t is one and of the second term in (3.72) is zero, the Fisher information matrix is given by

$$\mathcal{F} = \sum_t \frac{1}{2T} E \left[\frac{1}{h_t^2} u_t \, u_t' \right],$$

and consistently estimated by

$$\hat{\mathcal{F}} = \frac{1}{T} \sum_t \left[\frac{1}{2h_t^2} u_t \, u_t' \right].$$

Later we will discuss the properties of the estimators, see p.49.

In many applications with the linear ARCH(q) model a large number of parameters and a long lag length q are needed. These problems are avoided by an alternative, more general parametrization of h_t introduced by Bollerslev [12]. This more general model is called the Generalized ARCH, or GARCH(p, q), model

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2,$$

where $q > 0$, $p \geq 0$, $\alpha_0 > 0$, $\alpha_i \geq 0$ for $i = 1, \ldots, q$, and $\beta_i \geq 0$ for $i = 1, \ldots, p$. The generalization in the GARCH(p, q) model in comparison to the ARCH(q) model is that beside past values of the process, also past conditional variances enter.

Denote $\gamma' \equiv (\alpha_0, \alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p)$, $h_t \equiv \sigma_t^2$ and $v_t' \equiv (1, \varepsilon_{t-1}^2, \ldots, \varepsilon_{t-q}^2, h_{t-1}, \ldots, h_{t-p})$. As for the ARCH model we estimate γ by differentiating the log-likelihood with respect to γ

$$\frac{\partial l_t}{\partial \gamma} = \frac{1}{2h_t} \frac{\partial h_t}{\partial \gamma} \left(\frac{\varepsilon_t^2}{h_t} - 1 \right).$$

The Hessian is

$$\frac{\partial^2 l_t}{\partial\gamma\partial\gamma'} = -\frac{1}{2h_t^2}\frac{\partial h_t}{\partial\gamma}\frac{\partial h_t}{\partial\gamma'}\left(\frac{\varepsilon_t^2}{h_t}\right) + \left[\frac{\varepsilon_t^2}{h_t} - 1\right]\frac{\partial}{\partial\gamma'}\left[\frac{1}{2h_t}\frac{\partial h_t}{\partial\gamma}\right], \tag{3.73}$$

where

$$\frac{\partial h_t}{\partial\gamma} = v_t + \sum_{i=1}^{p}\beta_i\frac{\partial h_{t-i}}{\partial\gamma}. \tag{3.74}$$

The difference from the ARCH model is the inclusion of the recursive part in (3.74). As for the ARCH model, the Fisher information matrix is consistently estimated by the sample analogue of the first term in (3.73).

As mentioned earlier we consider a generalization of model (3.67), (3.68), the ARCH regression model

$$\begin{aligned} \varepsilon_t &= y_t - x_t'b, \\ h_t &= u_t'\alpha, \end{aligned} \tag{3.75}$$

where ε_t is conditionally distributed as $\mathcal{N}(0, h_t)$, b is a parameter vector and x_t is time $(t-1)$ measurable. As in the linear ARCH(q) model the conditional mean of ε_t is zero, but the conditional mean of y_t is given as $x_t'b$. The conditional variance for both ε_t and y_t is h_t.

As for the linear ARCH(q) model we consider ML estimation for the ARCH regression model. In addition to the parameter α estimated as in the linear ARCH(q) model we have to estimate parameter b. The derivative with respect to b is given by

$$\frac{\partial l_t}{\partial b} = \frac{\varepsilon_t x_t'}{h_t} + \frac{1}{2h_t}\frac{\partial h_t}{\partial b}\left(\frac{\varepsilon_t^2}{h_t} - 1\right).$$

The Hessian is

$$\begin{aligned} \frac{\partial^2 l_t}{\partial b\partial b'} &= -\frac{x_t'x_t}{h_t} - \frac{1}{2h_t^2}\frac{\partial h_t}{\partial b}\frac{\partial h_t}{\partial b'}\left(\frac{\varepsilon_t^2}{h_t}\right) \\ &\quad -\frac{2\varepsilon_t x_t'}{h_t^2}\frac{\partial h_t}{\partial b} + \left(\frac{\varepsilon_t^2}{h_t} - 1\right)\frac{\partial}{\partial b'}\left[\frac{1}{2h_t}\frac{\partial h_t}{\partial b}\right]. \end{aligned}$$

Taking conditional expectations, the last two terms of the Hessian vanish and ε_t^2/h_t becomes one, so that the part of the Fisher information matrix corresponding to b is given by

$$\mathcal{F}_{bb} = \frac{1}{T}\sum_t E\left[\frac{x_t'x_t}{h_t} + \frac{1}{2h_t^2}\frac{\partial h_t}{\partial b}\frac{\partial h_t}{\partial b'}\right],$$

and by substituting the linear variance function, $\mathcal{F}_{bb}$ is consistently estimated by

$$\hat{\mathcal{F}}_{bb} = \frac{1}{T}\sum_t \left[\frac{x_t' x_t}{h_t} + 2\sum_j \alpha_j^2 \frac{\varepsilon_{t-j}^2}{h_t^2} x_{t-j}' x_{t-j}\right].$$

Finally we give a remark to the ML estimation for the GARCH regression model

$$\varepsilon_t = y_t - x_t' b,$$
$$h_t = v_t' \gamma,$$

with v_t and γ as above. In order to estimate the mean parameters b we differentiate with respect to b as shown for the ARCH regression model with the single difference

$$\frac{\partial h_t}{\partial b} = -2\sum_{j=1}^{q} \alpha_j x_{t-j}\varepsilon_{t-j} + \sum_{j=1}^{p} \beta_j \frac{\partial h_{t-j}}{\partial b}.$$

Paying attention to this difference, the part of the Fisher information matrix corresponding to b is consistently estimated as in the ARCH regression model.

We denote by $\mathcal{F}_{\alpha\alpha}$, respectively $\mathcal{F}_{\gamma\gamma}$, the part of the Fisher information matrix corresponding to α, respectively γ, and the elements in the off-diagonal block of the information matrix by $\mathcal{F}_{\alpha b}$, respectively by $\mathcal{F}_{\gamma b}$. The elements $\mathcal{F}_{\alpha b}$, respectively $\mathcal{F}_{\gamma b}$, may be shown to be zero. Because of this asymptotic independence α, respectively γ, can be estimated without loss of asymptotic efficiency based on a consistent estimate of b and vice versa. Using this fact Engle [22], §6, formulates a simple scoring algorithm for the ML estimation of the parameters α and b.

As for the properties of the estimators, Bollerslev [14] remarks that for the general ARCH class of models the verification of sufficient regularity conditions for the MLE to be consistent and asymptotically normally distributed is very difficult. A detailed proof is only worked out in a few cases. Normally one assumes that these regularity conditions are satisfied, such that the ML estimators $\hat{\alpha}$, respectively $\hat{\gamma}$, and $\hat{b}$ are consistent and asymptotically normally distributed with limiting distribution

$$\sqrt{T}(\hat{\alpha} - \alpha) \longrightarrow \mathcal{N}(0, \mathcal{F}_{\alpha\alpha}^{-1}), \text{ resp. } \sqrt{T}(\hat{\gamma} - \gamma) \longrightarrow \mathcal{N}(0, \mathcal{F}_{\gamma\gamma}^{-1}),$$
$$\text{and } \sqrt{T}(\hat{b} - b) \longrightarrow \mathcal{N}(0, \mathcal{F}_{bb}^{-1}). \tag{3.76}$$

Closing our considerations about estimation in ARCH/GARCH models we remark that Geweke [31] developes Bayesian inference procedures for ARCH models by using Monte Carlo methods to determine the a posteriori distribution. The Bayesian analysis is discussed in the next section.

3.2.3 The Bayesian estimation method

Suppose we have the data $x = (x_1, \ldots, x_n)$ with distribution $p(x|\theta)$ where θ is the unknown parameter we want to estimate. The basic idea of the Bayesian approach is to treat the parameter θ as a random variable and to use a guess or an a priori knowledge of the distribution $\pi(\theta)$ of θ and then to estimate θ by calculating the a posteriori distribution $\pi(\theta|x)$ of θ. For details about the Bayesian theory we refer to [3] and [55]. First of all the Bayesian method will be described in the case where the parameter θ is one-dimensional. Furthermore, the k-dimensional case and the hierarchical model will be discussed.

The one-dimensional case

In the one-dimensional case the a posteriori distribution $\pi(\theta|x)$ of θ is calculated by the so called Bayes formula using the a priori distribution $\pi(\theta)$ as follows

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)d\theta}, \tag{3.77}$$

where the denominator is a proportionality constant making the total a posteriori probability equal to one. Now by using the a posteriori distribution $\pi(\theta|x)$ the parameter θ can be estimated for example via the modus of the a posteriori distribution

$$\hat{\theta} = \text{arg max } \pi(\theta|x)$$

or by the mean

$$\hat{\theta} = \mathrm{E}\left[\pi(\theta|x)\right]$$

or by the median

$$\hat{\theta} = \text{median}\left[\pi(\theta|x)\right].$$

For the asymptotic properties of the estimator we refer to [3], §5.3 Asymptotic Analysis.

Example (One-dimensional case)

Suppose that $x = (x_1, \ldots, x_n)$ and $x_i|\theta \sim \mathcal{N}(\theta, \sigma^2)$ i. i. d., where σ^2 is known. Choose an a priori distribution of $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$. With the Bayes formula

$$\begin{aligned}
\pi(\theta|x_i) &\propto \exp\left\{\frac{-(x_i-\theta)^2}{2\sigma^2}\right\}\exp\left\{\frac{-(\theta-\mu_0)^2}{2\sigma_0^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left[\theta^2 - 2\theta\,\frac{x_i\sigma_0^2+\mu_0\sigma^2}{\sigma^2+\sigma_0^2}+\ldots\right]\right\},
\end{aligned}$$

where the terms and factors not written out do not involve θ, we have

$$\theta|x_i \sim \mathcal{N}\left(\frac{x_i\sigma_0^2+\mu_0\sigma^2}{\sigma^2+\sigma_0^2}, \frac{1}{\sigma^{-2}+\sigma_0^{-2}}\right),$$

and with

$$\begin{aligned}\pi(\theta|x) &\propto \prod_{i=1}^{n} p(x_i|\theta)\,\pi(\theta)\\ &\propto \exp\left\{\frac{-(\bar{x}-\theta)^2}{2\sigma^2/n}\right\}\exp\left\{\frac{-(\theta-\mu_0)^2}{2\sigma_0^2}\right\},\end{aligned}$$

we have

$$\theta|x \sim \mathcal{N}\left(\frac{\bar{x}n\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}, \frac{1}{n\sigma^{-2}+\sigma_0^{-2}}\right),$$

where $\bar{x} = \frac{1}{n}\sum x_i$.

The multi-dimensional case

In the multi-dimensional case $\theta = (\theta_1, \ldots, \theta_k)$, the a posteriori distribution of θ can be calculated by the Bayes formula as follows

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int \ldots \int p(x|\theta)\pi(\theta)\,d\theta_1 \ldots d\theta_k}. \tag{3.78}$$

By using the marginal distributions $\pi(\theta_i|x)$ of the joint a posteriori distribution $\pi(\theta|x)$

$$\pi(\theta_i|x) = \int \ldots \int \pi(\theta|x)\,d\theta_1 \ldots d\theta_{i-1}d\theta_{i+1} \ldots d\theta_k, \tag{3.79}$$

we are able to estimate θ by the ways described in the one-dimensional case above.

Usually problems arise in calculating the integrals in (3.79) which require approximation techniques. We will discuss simulations of distributions using so called Markov Chain Monte Carlo (MCMC) methods. The key idea of the MCMC methods is described in the following. For details we refer to [71], [3] p.353 or [55]. Suppose we want to generate a sample from an a posteriori distribution $\pi(\theta|x)$ for $\theta \in \Theta \subseteq \mathbb{R}^k$, but cannot directly do this. However, suppose we are able to construct a Markov chain with state space Θ and with equilibrium distribution $\pi(\theta|x)$. Then under suitable regularity conditions asymptotic results exist, showing in which sense the sample output from such a chain with equilibrium distribution $\pi(\theta|x)$ can be used to mimic a random sample from $\pi(\theta|x)$ or to estimate the expected value of a function

$f(\theta)$ with respect to $\pi(\theta|x)$. If $\theta^1, \theta^2, \ldots, \theta^t, \ldots$ is a realization from a suitable chain then

$$\theta^t \longrightarrow \theta$$

in distribution as t tends to infinity, $\theta \sim \pi(\theta|x)$ and

$$\frac{1}{t}\sum_{i=1}^{t} f(\theta^i) \longrightarrow \mathrm{E}_{\theta|x}\left[f(\theta)\right] \text{ a.s.}$$

as t tends to infinity. Now we need algorithms to construct such chains with specified equilibrium distributions. We will discuss two particular forms of Markov chain schemes, the Gibbs sampling algorithm and the Metropolis–Hastings algorithm.

The Gibbs sampling algorithm

Suppose $\theta = (\theta_1, \ldots, \theta_k)$ is the vector of unknown quantities. We want to simulate $\pi(\theta_i|\, x)$ for $i = 1, \ldots, k$. Denote by $\pi(\theta|\, x) = \pi(\theta_1, \ldots, \theta_k|\, x)$ the joint density and by $\pi(\theta_i|\, x, \theta_j, j \neq i)$ the so called induced full conditional densities for each of the components θ_i given values of the other components θ_j, $j \neq i$, for $i = 1, \ldots, k$. Suppose that we are able to sample from each of these one-dimensional distributions.

The Gibbs sampling algorithm (see Geman and Geman [29]):

1) choose arbitrary starting points $\theta^0 = (\theta_1^0, \ldots, \theta_k^0)$.

2) make random drawings from the full conditional distribution as follows

$$\begin{array}{lll} \theta_1^1 & \text{from} & \pi(\theta_1|x, \theta_j^0, j \neq 1) \\ \theta_2^1 & \text{from} & \pi(\theta_2|x, \theta_1^1, \theta_3^0, \ldots, \theta_k^0) \\ \theta_3^1 & \text{from} & \pi(\theta_3|x, \theta_1^1, \theta_2^1, \theta_4^0, \ldots, \theta_k^0) \\ & \ldots & \\ \theta_k^1 & \text{from} & \pi(\theta_k|x, \theta_j^i, j \neq k) \end{array}$$

   This completes a transition from $\theta^0 = (\theta_1^0, \ldots, \theta_k^0)$ to $\theta^1 = (\theta_1^1, \ldots, \theta_k^1)$.

3) iterating step 2) produces a sequence $\theta^0, \theta^1, \ldots, \theta^t, \ldots$ which is a realization of a Markov chain with transition probability K from θ^t to θ^{t+1}

$$K(\theta^t, \theta^{t+1}) = \prod_{l=1}^{k} \pi\left(\theta_l^{t+1} \;\middle|\; x, \theta_j^t, j > l, \theta_j^{t+1}, j < l\right).$$

The important property of the Gibbs sampling algorithm is that we only sample from the full conditional distributions. As t tends to infinity, $(\theta_1^t, \ldots, \theta_k^t)$ tends in distribution to a random vector with joint density $\pi(\theta|x)$, see e.g. [29]. In particular, θ_i^t tends in distribution to a random quantity with density $\pi(\theta_i|x)$. The above algorithm is assumed to be replicated m times independently, that means we have m replicates of $\theta^t = (\theta_1^t, \ldots, \theta_k^t)$. Then for large t the replicates $(\theta_{i1}^t, \ldots, \theta_{im}^t)$ are approximately a random sample from $\pi(\theta_i|x)$. For m sufficiently large we obtain an estimate $\hat{\pi}(\theta_i|x)$ for $\pi(\theta_i|x)$ from

$$\hat{\pi}(\theta_i|x) = \frac{1}{m} \sum_{l=1}^{m} \pi(\theta_i|x, \theta_{jl}^t, j \neq i).$$

For more details we refer to [55], p.226ff.

The Metropolis–Hastings algorithm

In order to construct a Markov chain $\theta^1, \ldots, \theta^t, \ldots$ with state space Θ and equilibrium distribution $\pi(\theta|x)$ we find the transition probability from $\theta^t \equiv \theta$ to the next state θ^{t+1} via the Metropolis–Hastings algorithm as follows: Denote by $q(\theta, \theta')$ a transition probability function such that in the case $\theta^t = \theta$, θ' drawn from $q(\theta, \theta')$ is considered as a proposed possible value for θ^{t+1}. However, a further randomization takes place. With probability $\alpha(\theta, \theta')$ we accept $\theta^{t+1} = \theta'$, otherwise we reject the value generated from $q(\theta, \theta')$ and set $\theta^{t+1} = \theta$.

This construction defines a Markov chain with transition probability given by

$$\pi(\theta, \theta') = \begin{cases} q(\theta, \theta')\alpha(\theta, \theta'), & \text{if } \theta' \neq \theta, \\ 1 - \sum_{\theta''} q(\theta, \theta'')\alpha(\theta, \theta''), & \text{if } \theta' = \theta. \end{cases}$$

With the definition

$$\alpha(\theta, \theta') = \begin{cases} \min\left\{ \dfrac{\pi(\theta'|x)q(\theta', \theta)}{\pi(\theta|x)q(\theta, \theta')}, 1 \right\}, & \text{if } \pi(\theta|x)\, q(\theta, \theta') > 0, \\ 1, & \text{if } \pi(\theta|x)\, q(\theta, \theta') = 0, \end{cases}$$

provided that $q(\theta, \theta')$ is chosen to be irreducible and aperiodic on a suitable state space, we have that $\pi(\theta|x)$ is the equilibrium distribution of the constructed chain. For more details see [71].

Hierarchical models

An hierarchical model forms a structure in the following way: the distribution of the data x is written conditionally on parameters θ_1 as $p(x|\theta_1)$, and the distribution of θ_1 is written conditionally on 'hyperparameters' θ_2 as

$p(\theta_1|\theta_2)$ and we have the a priori distribution of θ_2, $\pi(\theta_2)$. This is the so called three-stage hierarchical model which is considered in many applications. The three stages are x, θ_1 and θ_2 as above, that means a three-stage model has the structure $p(x|\theta_1)$, $p(\theta_1|\theta_2)$, $\pi(\theta_2)$. We could continue this process and write the distribution of θ_2 conditionally on other hyperparameters θ_3 as $p(\theta_2|\theta_3)$ and so on. We obtain the k-stage hierarchical model $p(x|\theta_1)$, $p(\theta_1|\theta_2)$, $p(\theta_2|\theta_3), \ldots, p(\theta_{k-2}|\theta_{k-1}), \pi(\theta_{k-1})$.

The model includes the assumption of conditionally independence: conditioning on θ_j, the parameters $x, \theta_1, \theta_2, \ldots, \theta_{j-1}$ and the parameters $\theta_{j+1}, \ldots, \theta_{k-1}$ are independent. Especially consider the three-stage model with $p(x|\theta_1)$, $p(\theta_1|\theta_2)$, $\pi(\theta_2)$. The distribution $p(x|\theta_1)$ is formally the distribution of x given θ_1 *and* θ_2. However, if we know θ_1 then knowing θ_2 would not add any information about x, this means x and θ_2 are independent given θ_1.

By the Bayesian analysis of such a three-stage hierarchical model the special structure allows us to write the a posteriori distribution $\pi(\theta_1, \theta_2|x)$ in the form

$$\begin{aligned} \pi(\theta_1, \theta_2|x) &\propto p(x|\theta_1, \theta_2)p(\theta_1, \theta_2) \\ &\propto p(x|\theta_1)p(\theta_1|\theta_2)\pi(\theta_2), \end{aligned}$$

and inference for θ_2 is given by its marginal distribution

$$\pi(\theta_2|x) \propto p(x|\theta_2)\pi(\theta_2).$$

The marginal distributions to be used for inference for the parameters can be calculated by MCMC methods.

Example (Hierarchical model)

An example of an hierarchical three-stage stochastic model can be found in [42], where the conditional variance follows a log-AR(1) process:

$$y_t = \sqrt{h_t}\, u_t$$

$$\ln h_t = \alpha + \delta \ln h_{t-1} + \sigma_\nu \nu_t,$$

where $(u_t, v_t) \sim \mathcal{N}(0,1)$ independent, or more generally we can consider the log-AR(p)-model

$$y_t = \sqrt{h_t}\, u_t$$

$$\ln h_t = \alpha + \delta_1 \ln h_{t-1} + \delta_2 \ln h_{t-2} + \ldots + \delta_p \ln h_{t-p} + \sigma_\nu \nu_t,$$

where $(u_t, v_t) \sim \mathcal{N}(0,1)$ independent. In this model the time series of the data y is generated from a probability model $p(y|h)$ where h denotes the vector of volatilities. The volatilities h are unobserved and are assumed to be generated by $p(h|\gamma)$ with $\gamma^T = (\alpha, \delta, \sigma_\nu)$. Denote $\theta_1 = h$, $\theta_2 = \gamma$, $\delta^T = (\delta_1, \dots, \delta_p)$. Via the Bayes formula we have

$$\pi(\theta_1, \theta_2|y) \propto p(y|\theta_1)p(\theta_1|\theta_2)\pi(\theta_2).$$

The marginal distributions $p(\theta_1|y)$, $p(\theta_2|y)$ can be calculated by using MCMC methods as Gibbs sampling or Metropolis–Hastings algorithm.

Chapter 4

Nonparametric Estimation

4.1 Diffusion models

Nonparametric estimation in general deals with estimating functionals in situations where the latter are not determined by a finite number of parameters, e.g. estimating a probability density at some fixed point, derivatives of a density, and others. For a thorough treatment of nonparametric estimation we refer to Ibragimov and Has'minskii [38], §4 and §7.

We consider estimation of a probability density at a point based on observations in $\mathbb{R}$. Suppose $X_1, \ldots, X_n$ is a sample from a population random variable X taking values in $\mathbb{R}$ with unknown density $f(x)$. If we only know that $f(x)$ belongs to a class $\mathcal{F}$ of functions, estimating $f(x)$ typically is an infinite dimensional nonparametric problem. Using the function

$$\chi(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0, \end{cases}$$

the number of observations smaller than x can be written as

$$\sum_{k=1}^{n} \chi(x - X_k).$$

Thus, with this notation the empirical distribution function of the data $X_1, \ldots, X_n$ is

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} \chi(x - X_k), \tag{4.1}$$

which is a well-known estimator for the distribution function $F(x)$. For n sufficiently large we know that $F_n(x)$ is close to the actual distribution function

$$F(x) = \int_{-\infty}^{x} f(y)dy.$$

The latter statement can be made precise for instance through the Borel-Cantelli Lemma and its various refinements (see [68]) .

In order to construct a density estimator $f_n(x)$ starting from the empirical distribution function $F_n(x)$ in (4.1), we have to smooth F_n. One possibility is by using so-called kernel density estimators

$$f_n(x) = \frac{1}{nh_n} \sum_{k=1}^{n} V\left(\frac{x - X_k}{h_n}\right),$$

where the kernel $V : \mathbb{R} \longrightarrow \mathbb{R}^+$ satisfies

$$\int_{-\infty}^{\infty} V(x)dx = 1,$$

and where the so-called bandwidth sequence (h_n) is such that

$$h_n \longrightarrow 0, \qquad nh_n \longrightarrow \infty.$$

This class of estimators is referred to as the Parzen-Rosenblatt estimators. Under certain restrictions on $f(x)$ the sequence $f_n(x)$ converges in some probabilistic sense to $f(x)$. See [38], §4.4, and [69] for a detailed discussion.

Ibragimov and Has'minskii [38], §7, present some interesting approaches to and examples of nonparametric estimators of a signal $S(t)$, belonging to a certain functional space. Two models for signals are considered. A first one has the form

$$dX(t) = S(t)dt + \varepsilon db(t), \tag{4.2}$$

with $0 \leq t \leq 1$, $(b(t))$ a Wiener process, ε small (typically $\varepsilon \downarrow 0$) and $X(t)$ is an observed signal on $[0, 1]$. A second model is of the form

$$dX(t) = S(t)dt + db(t), \tag{4.3}$$

with $0 \leq t \leq n$, $S(t)$ a one-periodic function and $X(t)$ is observed over n time periods.

As a basic example consider estimation of the functional F on $S(t)$

$$F(S) = \int_0^1 f(t)S(t)dt,$$

with $S, f \in L_2(0, 1)$. The functional F has to be estimated based on observations of (4.2), respectively (4.3). The estimator

$$\hat{F}_1 = \int_0^1 f(t)dX(t)$$

for model (4.2) and the estimator

$$\hat{F}_2 = \frac{1}{n}\int_0^n f(t)dX(t)$$

for model (4.3) both can be shown to have good properties. Both estimators are normally distributed with mean $F(S)$ and variance $\varepsilon^2||f||^2$, respectively $||f||^2/n$. We refer to [38], §7.4 and §7.5, for more details.

4.2 Discrete models

In section 3.2 we dealt with parametrizations of volatility, e.g. with the ARCH(q) process $\{\varepsilon_t\}$ (3.67), where $\varepsilon_t|\mathcal{F}_{t-1} \sim \mathcal{N}(0,\sigma_t^2)$, $\mathcal{F}_t$ being the information σ-algebra at time t, and

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2. \tag{4.4}$$

Instead of considering a parametric approach as (4.4) we now discuss nonparametric techniques for the estimation of the volatility σ_t^2.

Basicly there are two different nonparametric approaches: kernel estimators and Fourier type estimators. For details we refer to Pagan and Schwert [57].

In order to estimate $\sigma_s^2 = \mathrm{E}\,(\varepsilon_s^2|\mathcal{F}_{s-1})$, kernel methods essentially use a weighted sum over ε_j^2, $j = 1,\ldots,T$, $j \neq s$:

$$\hat{\sigma}_s^2 = \sum_{\substack{j=1\\ j\neq s}}^{T} w_j^{(s)} \varepsilon_j^2, \tag{4.5}$$

where the sum over all $(T-1)$ weights $w_j^{(s)}$, $j = 1,\ldots,T$, $j \neq s$, equals one. The aim is to obtain an estimate of $\mathrm{E}\,(\varepsilon_s^2|\varepsilon_{s-1},\varepsilon_{s-2},\ldots,\varepsilon_{s-m})$ for m suitably chosen. If the preceding values of ε_j, that is $\varepsilon_{j-1},\varepsilon_{j-2},\ldots,\varepsilon_{j-m}$, are similar to the preceding values of ε_s, that is $\varepsilon_{s-1},\varepsilon_{s-2},\ldots,\varepsilon_{s-m}$, then ε_j^2 is expected to give useful information about $\mathrm{E}\,(\varepsilon_s^2|\varepsilon_{s-1},\varepsilon_{s-2},\ldots,\varepsilon_{s-m})$. In this case, the weight $w_j^{(s)}$ is large. If the values that preceded ε_j differ 'substantially' from the preceding values of ε_s, then ε_j^2 is expected to give only little information about $\mathrm{E}\,(\varepsilon_s^2|\varepsilon_{s-1},\varepsilon_{s-2},\ldots,\varepsilon_{s-m})$ and therefore $w_j^{(s)}$ is close to zero. We remark that in the sum (4.5) the element $w_s^{(s)}\varepsilon_s^2$ is left out to avoid the situation where 'outliers' in the data lead to a weight $w_s^{(s)}$ close to unity, while all the other $w_j^{(s)}$ are close to zero, so that only ε_s^2 determines $\hat{\sigma}_s^2$.

There are many possible weighting schemes. A popular one uses a Gaussian kernel. With the notation $z_j' = (\varepsilon_{j-1}, \varepsilon_{j-2}, \ldots, \varepsilon_{j-m})$ we have

$$w_j^{(s)} = \frac{K(z_s - z_j)}{\sum_{r=1}^{T} K(z_r - z_s)},$$

where $K(\cdot)$ is the Gaussian kernel

$$K(z_s - z_j) = \frac{1}{\sqrt{2\pi|H|}} \exp\left\{-\frac{1}{2}(z_s - z_j)' H\,(z_s - z_j)\right\},$$

with $H = \mathrm{diag}(h_1, \ldots, h_m)$ containing the bandwidths that were set to $\hat{\sigma}_k T^{-1/(4+m)}$, where $\hat{\sigma}_k$ is the sample standard deviation of ε_{s-k}, the kth component of z_s, $k = 1, \ldots, m$.

Another approach to nonparametric estimation of volatility is an approximation using series expansion. The most frequently used series expansion in economics is the Flexible Fourier Form (FFF) introduced by Gallant [28], which leads to a volatility estimate of the form

$$\hat{\sigma}_t^2 = \alpha_0 + \sum_{j=1}^{m} \left\{ \left(\alpha_j \varepsilon_{t-j} + \beta_j \varepsilon_{t-j}^2\right) + \sum_{k=1}^{2} (\gamma_{jk} \cos(k\,\varepsilon_{t-j}) + \delta_{jk} \sin(k\,\varepsilon_{t-j})) \right\},$$

that means σ_t^2 is estimated by a sum of a low-order polynomial and trigonometric terms based on ε_{t-j}, $j = 1, \ldots, m$. Note the disadvantage of the FFF that the estimates of σ_t^2 may be negative. We refer to [57] for more details.

Chapter 5

Some Diffusion Models with Explicit Solutions

In this chapter our representation follows [45], §4.2-4.4.

5.1 Linear stochastic differential equations

We first consider the class of linear stochastic differential equations. To simplify the exposition we concentrate on the scalar case and afterwards give the extension to the multivariate case.

The general form of a scalar linear stochastic differential equation is

$$dX_t = (a_1(t)X_t + a_2(t))\,dt + (b_1(t)X_t + b_2(t))\,dW_t, \tag{5.1}$$

where a_1, a_2, b_1, b_2 are "nice"[1] functions of time t or constants. As in the case of ordinary differential equations, the method of solution also involves a fundamental solution of an associated homogeneous differential equation. Here the homogeneous linear equation belonging to the general equation (5.1) is

$$dX_t = a_1(t)X_t\,dt + b_1(t)X_t dW_t. \tag{5.2}$$

In order to find its fundamental solution Φ_{t,t_0}, that means the solution of (5.2) satisfying $\Phi_{t_0,t_0} = 1$, we first look at the special case $b_1(t) \equiv 0$, that means at the ordinary differential equation

$$\frac{dX_t}{dt} = a_1(t)X_t. \tag{5.3}$$

[1] They are Lebesgue measurable and bounded on an interval $0 \leq t \leq T$. Then a unique (strong) solution X_t on $t_0 \leq t \leq T$ exists for each $0 \leq t_0 < T$.

We know the fundamental solution of (5.3)

$$\Phi_{t,t_0} = \exp\left[\int_{t_0}^{t} a_1(s)ds\right], \tag{5.4}$$

or equivalently

$$d\left[\ln \Phi_{t,t_0}\right] = a_1(t)dt. \tag{5.5}$$

Equation (5.5) motivates us to consider $\ln \Phi_{t,t_0}$ also in the general case with Φ_{t,t_0} solving equation (5.2). Applying the Itô formula[2] on $\ln \Phi_{t,t_0}$ we obtain

$$d\left[\ln \Phi_{t,t_0}\right] = \left(a_1(t) - \frac{1}{2}\, b_1^2(t)\right) dt + b_1(t)dW_t, \tag{5.6}$$

thus, with $\Phi_{t_0,t_0} = 1$,

$$\ln \Phi_{t,t_0} = \int_{t_0}^{t} \left(a_1(s) - \frac{1}{2}\, b_1^2(s)\right) ds + \int_{t_0}^{t} b_1(s)dW_s,$$

or

$$\Phi_{t,t_0} = \exp\left[\int_{t_0}^{t} \left(a_1(s) - \frac{1}{2}\, b_1^2(s)\right) ds + \int_{t_0}^{t} b_1(s)dW_s\right]. \tag{5.7}$$

Now that we derived the solution Φ_{t,t_0} of the homogeneous linear equation (5.2) we want to obtain the general solution X_t of (5.1). In an analogous way to (5.6) we derive by means of the Itô formula[3]

$$d[\Phi_{t,t_0}^{-1}] = \left(-a_1(t) + b_1^2(t)\right) \Phi_{t,t_0}^{-1} dt - b_1(t)\Phi_{t,t_0}^{-1} dW_t. \tag{5.8}$$

As we will see later we have to consider the process $\Phi_{t,t_0}^{-1} X_t$ in order to find an explicit expression for X_t. The processes Φ_{t,t_0}^{-1} and X_t both include the same Wiener process W_t, and (5.2) and (5.8) can be seen as a two-dimensional stochastic differential equation. In order to calculate $d(\Phi_{t,t_0}^{-1} X_t)$ we therefore have to use the Itô formula for vector valued processes[4] with the transformation $U(X_t, \Phi_{t,t_0}^{-1}) = \Phi_{t,t_0}^{-1} X_t$. We obtain

$$d[\Phi_{t,t_0}^{-1} X_t] = \left(a_2(t) - b_1(t)b_2(t)\right) \Phi_{t,t_0}^{-1} dt + b_2(t)\Phi_{t,t_0}^{-1} dW_t.$$

Hence with $\Phi_{t_0,t_0} = 1$ we get by integration

$$\Phi_{t,t_0}^{-1} X_t = X_{t_0} + \int_{t_0}^{t} \left(a_2(s) - b_1(s)b_2(s)\right) \Phi_{s,t_0}^{-1} ds + \int_{t_0}^{t} b_2(s)\Phi_{s,t_0}^{-1} dW_s,$$

[2]see Appendix C.1
[3]see Appendix C.1
[4]see Appendix C.2

and thus

$$X_t = \Phi_{t,t_0} \left[X_{t_0} + \int_{t_0}^{t} (a_2(s) - b_1(s)b_2(s))\, \Phi_{s,t_0}^{-1} ds + \int_{t_0}^{t} b_2(s)\Phi_{s,t_0}^{-1} dW_s \right]. \quad (5.9)$$

Summarizing: the solution X_t of the general scalar linear stochastic differential equation

$$dX_t = (a_1(t)X_t + a_2(t))\, dt + (b_1(t)X_t + b_2(t))\, dW_t, \quad (5.10)$$

is

$$X_t = \Phi_{t,t_0} \left[X_{t_0} + \int_{t_0}^{t} (a_2(t) - b_1(t)b_2(t))\, \Phi_{t,t_0}^{-1} ds + \int_{t_0}^{t} b_2(t)\Phi_{s,t_0}^{-1} dW_s \right], \quad (5.11)$$

with the fundamental solution

$$\Phi_{t,t_0} = \exp \left[\int_{t_0}^{t} \left(a_1(s) - \frac{1}{2} b_1^2(s) \right) ds + \int_{t_0}^{t} b_1(s) dW_s \right]. \quad (5.12)$$

All the special cases – as only additive or only multiplicative noise, homogeneous or inhomogeneous equations, constant or variable coefficients – are contained in the general case (5.10).

At this point we give a short extension of the scalar to the multidimensional case. The general form of a d-dimensional linear stochastic differential equation is

$$dX_t = (A(t)X_t + a(t))dt + \sum_{i=1}^{m} (B^i(t)X_t + b^i(t))dW_t^i, \quad (5.13)$$

where $A(t), B^1(t), B^2(t), \ldots, B^m(t)$ are $d \times d$-matrix functions, $a(t)$, $b^1(t)$, $b^2(t) \ldots, b^m(t)$ are d-dimensional vector functions and $W = \{W_t, t \geq 0\}$ is an m-dimensional Wiener process with components $W_t^1, W_t^2, \ldots, W_t^m$, which are independent scalar Wiener processes. Using the same arguments as for the scalar case above we find the solution of (5.13)

$$X_t = \Phi_{t,t_0} \left[X_{t_0} + \int_{t_0}^{t} \Phi_{s,t_0}^{-1} \left(a(s) - \sum_{i=1}^{m} B^i(s)b^i(s) \right) ds + \sum_{i=1}^{m} \int_{t_0}^{t} \Phi_{s,t_0}^{-1} b^i(s) dW_s^i \right], \quad (5.14)$$

where Φ_{t,t_0} is the $d \times d$ fundamental matrix with $\Phi_{t_0,t_0} = I$ and satisfying the homogeneous matrix stochastic differential equation

$$d\Phi_{t,t_0} = A(t)\Phi_{t,t_0} dt + \sum_{i=1}^{m} B^i(t)\Phi_{t,t_0} dW_t^i, \quad (5.15)$$

which can be seen column vector by column vector as d vector stochastic differential equations. We remark that unlike in the case of scalar homogeneous linear equations we cannot generally solve (5.15) for Φ_{t,t_0} explicitly.

5.2 Nonlinear stochastic differential equations

In the previous section we saw how to solve a scalar linear stochastic differential equation (5.10) explicitly. Under special conditions also some classes of nonlinear stochastic differential equations can be solved explicitly by a reduction to a linear equation. In the following we will examine those classes and their reductions in the scalar case. We remark that the two models (1.4) and (1.5) considered in chapter 1 do not belong to these classes.

Certain nonlinear stochastic differential equations

$$dX_t = a(t, X_t)\, dt + b(t, X_t)\, dW_t, \tag{5.16}$$

with $a(t, X_t), b(t, X_t)$ differentiable functions, can be reduced under special conditions to linear stochastic differential equations

$$dY_t = (a_1(t)Y_t + a_2(t))dt + (b_1(t)Y_t + b_2(t))\, dW_t \tag{5.17}$$

by a transformation $U(t, X_t) = Y_t$. We know by means of the Inverse Function Theorem that in case of $\frac{\partial U}{\partial x}(t, x) \neq 0$ a local inverse $x = V(t, y)$ of $y = U(t, x)$ exists.

Our purpose is to determine conditions for a and b in (5.16), so that such a suitable transformation $U(t, x) = y$ (with $\frac{\partial U}{\partial x}(t, x) \neq 0$) and coefficient functions $a_1(t), a_2(t), b_1(t)$ and $b_2(t)$ can be found. Then we solve the linear equation (5.17) explicitly (by means of (5.11)) and with $X_t = V(t, Y_t)$ we get a solution of (5.16).

In order to determine conditions for the reducibility of (5.16) to (5.17), that means conditions for a and b, we apply the Itô formula[5] to the transformation $U(t, x)$

$$dU(t, X_t) = \left[\frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} + \frac{1}{2}b^2\frac{\partial^2 U}{\partial x^2}\right] dt + b\frac{\partial U}{\partial x} dW_t. \tag{5.18}$$

A transformation U exists, if equations (5.17) and (5.18) are the same, that means if the following conditions for the coefficients are satisfied:

$$\frac{\partial U}{\partial t}(t, x) + a\frac{\partial U}{\partial x}(t, x) + \frac{1}{2}b^2(t, x)\frac{\partial^2 U}{\partial x^2}(t, x) = a_1(t)U(t, x) + a_2(t) \tag{5.19}$$

and

$$b(t, x)\frac{\partial U}{\partial x}(t, x) = b_1(t)U(t, x) + b_2(t). \tag{5.20}$$

[5] see Appendix C.1

If we do not specialize at this point, we are not able to calculate much more further and therefore we restrict ourselves to two cases.

First, we consider the case $a_1(t) \equiv b_1(t) \equiv 0$ and denote $a_2(t) =: \alpha(t)$ and $b_2(t) =: \beta(t)$. In order to obtain conditions for a and b it is useful to differentiate equation (5.19) with respect to x and equation (5.20) with respect to t:

$$\frac{\partial^2 U}{\partial t \partial x}(t,x) + \frac{\partial}{\partial x}\left(a(t,x)\frac{\partial U}{\partial x}(t,x) + \frac{1}{2}b^2(t,x)\frac{\partial^2 U}{\partial x^2}(t,x)\right) = 0 \tag{5.21}$$

and

$$b(t,x)\frac{\partial^2 U}{\partial t \partial x}(t,x) + \frac{\partial b}{\partial t}(t,x)\frac{\partial U}{\partial x}(t,x) = \beta'(t). \tag{5.22}$$

By means of (5.21) and (5.22) we obtain

$$\beta'(t) = \beta(t)\left[\frac{1}{b(t,x)}\frac{\partial b(t,x)}{\partial t}(t,x) - b(t,x)\frac{\partial}{\partial x}\left(\frac{a(t,x)}{b(t,x)} - \frac{1}{2}\frac{\partial b}{\partial x}(t,x)\right)\right]. \tag{5.23}$$

In (5.21) and hence in (5.23) we assumed that the left hand side of (5.19) is independent of x, in other words, we assumed that $\alpha(t)$ can be determined. Since β is independent of x we now follow the condition for the determination of β from equation (5.23)

$$\frac{\partial f}{\partial x}(t,x) = 0, \tag{5.24}$$

where

$$f(t,x) = \frac{1}{b(t,x)}\frac{\partial b(t,x)}{\partial t}(t,x) - b(t,x)\frac{\partial}{\partial x}\left(\frac{a(t,x)}{b(t,x)} - \frac{1}{2}\frac{\partial b}{\partial x}(t,x)\right). \tag{5.25}$$

If condition (5.24) is satisfied we are able to determine β and α. Thus condition (5.24) is sufficient for the reducibility of equation (5.16) to the linear equation

$$dX_t = \alpha(t)dt + \beta(t)dW_t \tag{5.26}$$

by means of a transformation U. Integrating equations (5.20) and (5.23) we obtain the explicit expression

$$U(t,x) = C\exp\left[\int_0^t f(s,x)ds\right]\int_0^x \frac{1}{b(t,z)}dz, \tag{5.27}$$

with an arbitrary constant C.

The second case to consider is the time-independent case. We want to reduce the nonlinear autonomous stochastic differential equation

$$dX_t = a(X_t)dt + b(X_t)dW_t \tag{5.28}$$

to the linear stochastic differential equation

$$dY_t = (a_1 Y_t + a_2)dt + (b_1 Y_t + b_2)dW_t, \tag{5.29}$$

by means of a suitable transformation $Y_t = U(X_t)$.

Equations (5.19) and (5.20) simplify to

$$a(x)\frac{dU}{dx}(x) + \frac{1}{2}b^2(x)\frac{d^2U}{dx^2}(x) = a_1 U(x) + a_2 \tag{5.30}$$

and

$$b(x)\frac{dU}{dx}(x) = b_1 U(x) + b_2. \tag{5.31}$$

We assume $b(x) \not\equiv 0$ and $b_1 \neq 0$ and conclude from (5.31)

$$U(x) = C\exp(b_1 h(x)) - \frac{b_2}{b_1} \tag{5.32}$$

with an arbitrary constant C and

$$h(x) = \int_{x_0}^{x} \frac{ds}{b(s)}. \tag{5.33}$$

As in the previous case we want to find conditions for a and b. Therefore we substitute (5.32) in (5.30), differentiate, multiply with $\left(\frac{dU}{dx}\right)^{-1} = \frac{b(x)}{b_1}\exp(-b_1 h(x))$ and obtain an expression with only a_1 on the right hand side. Differentiating again we get the following condition

$$b_1 \frac{dA}{dx}(x) + \frac{d}{dx}\left(b(x)\frac{dA}{dx}(x)\right) = 0 \tag{5.34}$$

with

$$A(x) = \frac{a(x)}{b(x)} - \frac{1}{2}\frac{db}{dx}(x). \tag{5.35}$$

This condition is satisfied in the trivial case $\frac{dA}{dx} = 0$ for any b_1, or in the case

$$\frac{d}{dx}\left(\frac{\frac{d}{dx}\left(b\frac{dA}{dx}\right)}{\frac{dA}{dx}}\right) = 0, \tag{5.36}$$

where we assume the choice of b_1

$$b_1 = -\frac{\frac{d}{dx}\left(b\frac{dA}{dx}\right)}{\frac{dA}{dx}}. \tag{5.37}$$

If $b_1 \neq 0$ then we may choose $b_2 = 0$ and use the transformation

$$U(x) = C\exp(b_1 h(x)). \tag{5.38}$$

If $b_1 = 0$ we use

$$U(x) = b_2 h(x) + C. \tag{5.39}$$

Now we apply the derived theory to three groups of examples.

Example 1 The stochastic differential equation

$$dX_t = \frac{1}{2}g(X_t)g'(X_t)dt + g(X_t)dW_t, \tag{5.40}$$

where g is a given differentiable function, is reducible with the general solution

$$X_t = h^{-1}(W_t + h(X_0)), \tag{5.41}$$

with

$$h(x) = \int_{x_0}^{x} \frac{ds}{g(s)}. \tag{5.42}$$

In the following we show how to obtain the general solution (5.41). With the notation of the theory above we see that $A(x) = 0$, and hence (5.34) is satisfied for all b_1. We choose $b_1 = 0$, $b_2 = 1$ and obtain

$$U(x) = h(x) + C.$$

Inserting U in (5.30) gives $a_1(h(x) + C) + a_2 = 0$, and with $a_1 = 0$, $a_2 = 0$, (5.29) reduces to $dY_t = dW_t$ with the solution $Y_t = Y_0 + W_t$. With $C = 0$ we have $Y_t = h(X_t)$, especially $Y_0 = h(X_0)$, and hence obtain (5.41).

We remark that in the special case (5.40) we may find a solution in another more pleasant way. Equation (5.40) is equivalent to the Stratonovich stochastic differential equation (for the Stratonovich integral see e.g. [56], p.16, or [45], §4.9)

$$dX_t = b(X_t) \circ dW_t.$$

That means, instead of reducing (5.40) to a linear stochastic differential equation we may integrate the Stratonovich stochastic differential equation directly as well, obtaining (5.41) at once.

Applications of Example 1:

$$\begin{aligned} dX_t &= \frac{1}{2}a^2 X_t dt + aX_t dW_t, \\ X_t &= X_0 \exp(aW_t). \end{aligned} \tag{5.43}$$

$$\begin{aligned} dX_t &= \frac{1}{2}a(a-1)X_t^{1-2/a} dt + aX_t^{1-1/a} dW_t, \\ X_t &= \left(W_t + X_0^{1/a}\right)^a. \end{aligned} \tag{5.44}$$

$$\begin{aligned} dX_t &= 1dt + 2\sqrt{X_t}\, dW_t, \\ X_t &= \left(W_t + \sqrt{X_0}\right)^2. \end{aligned} \tag{5.45}$$

$$\begin{aligned} dX_t &= \frac{1}{2}a^2 m X_t^{2m-1} dt + aX_t^m dW_t\,, \quad m \neq 1, \\ X_t &= \left(X_0^{1-m} - a(m-1)W_t\right)^{1/(1-m)}. \end{aligned} \tag{5.46}$$

$$\begin{aligned} dX_t &= \frac{1}{2}X_t dt + \sqrt{X_t^2 + 1}\, dW_t, \\ X_t &= \sinh(W_t + \mathrm{arcsinh} X_0). \end{aligned} \tag{5.47}$$

Example 2 The stochastic differential equation

$$dX_t = \left[\alpha g(X_t) + \frac{1}{2}g(X_t)g'(X_t)\right] dt + g(X_t) dW_t, \tag{5.48}$$

with a given differentiable function g, is reducible with the general solution

$$X_t = h^{-1}(\alpha t + W_t + h(X_0)), \tag{5.49}$$

where h is given by (5.42).

Again we want to show how (5.49) can be derived. We see $A(x) = \alpha$, and hence b_1 can be chosen arbitrarily. We choose $b_1 = 0$ and $b_2 = 1$ and obtain

$$U(x) = h(x) + C.$$

Inserting U in (5.30) gives $\alpha = h(x)a_1 + a_2$, and hence $a_1 = 0$, $a_2 = \alpha$. We have $dY_t = \alpha dt + dW_t$ with its solution $Y_t = \alpha t + W_t + Y_0$. With $C = 0$ we have $Y_t = h(x_t)$, especially $Y_0 = h(X_0)$, and we obtain (5.49).

As in the previous example we remark that (5.48) is equivalent to the Stratonovich stochastic differential equation

$$dX_t = \alpha b(X_t)dt + b(X_t) \circ dW_t,$$

which can be integrated directly in order to obtain the general solution (5.49).

The applications of example 1 we gave can all be modified to represent applications to example 2. For instance, consider

$$\begin{aligned} dX_t &= \left(\frac{1}{2}X_t + \sqrt{X_t^2+1}\right)dt + \sqrt{X_t^2+1}\,dW_t, \\ X_t &= \sinh(t + W_t + \mathrm{arcsinh} X_0). \end{aligned} \tag{5.50}$$

Example 3 The stochastic differential equation

$$dX_t = \left[\beta g(X_t)h(X_t) + \frac{1}{2}g(X_t)g'(X_t)\right]dt + g(X_t)dW_t, \tag{5.51}$$

where g is a given differentiable function and h is given by (5.42), can be reduced to the Langevin equation of the form

$$dY_t = -aY_t dt + bdW_t,$$

with $a = -\beta$ and $b = 1$, and has the general solution

$$X_t = h^{-1}\left(e^{\beta t}h(X_0) + e^{\beta t}\int_0^t e^{-\beta s}dW_s\right). \tag{5.52}$$

Again let us show how we derive the solution (5.52). We see $A(x) = \beta h(x)$ and by (5.37) we obtain $b_1 = 0$. With $b_2 = 1$ we have

$$U(x) = h(x) + C.$$

Inserting U in (5.30) gives $\beta h(x) = a_1 h(x) + a_2$ and hence $a_1 = \beta$ and $a_2 = 0$. We have $dY_t = \beta Y_t dt + dW_t$ with its solution

$$Y_t = e^{\beta t} Y_0 + e^{\beta t} \int_0^t e^{-\beta s} dW_s.$$

With $C = 0$ we have $Y_t = h(Y_t)$, especially $Y_0 = h(X_0)$, and we obtain (5.52).

The applications of example 1 we gave can all be modified to represent applications to example 3. For instance, consider

$$\begin{aligned} dX_t &= \left(2a\sqrt{X_t}X_t + 1\right) dt + 2\sqrt{X_t}\, dW_t, \\ X_t &= e^{at}\left(X_0 + \int_0^t e^{-as} dW_s\right). \end{aligned} \tag{5.53}$$

We remark that some examples of nonlinear reducible stochastic differential equations that are not included in the preceding three cases are listed in [45], p.124-126.

Finally, we again put stress on the fact that the two models (1.4) and (1.5) considered in chapter 1 are not reducible and cannot be solved explicitly.

Appendix A

The Kalman-Bucy Filter

A.1 The continuous case

For the Kalman-Bucy filter in the continuous case we refer to e.g. Kallianpur [43], §10. In our notation we follow Øksendal [56], §4, pp. 46–68.

A.1.1 The general filtering problem

Consider

$$\begin{aligned} dX_t &= b(t, X_t)\,dt + \sigma(t, X_t)\,dU_t\,, \quad \text{(System)} && \text{(A.1)} \\ dZ_t &= c(t, X_t)\,dt + \gamma(t, X_t)\,dV_t\,, \quad \text{(Observations)} && \text{(A.2)} \end{aligned}$$

where U and V are independent Brownian Motions, U_t p-dimensional, V_t r-dimensional.

The filtering problem is the following. Given the observations Z_s satisfying (A.2) for $0 \le s \le t$, what is the best estimate $\hat{X}_t$ of the state X_t of (A.1) based on these observations?

To solve this problem we formulate it mathematically. Let $\mathcal{G}_t$ be the σ-algebra generated by $\{Z_s(\cdot),\, s \le t\}$, $(\Omega, \mathcal{F}, P)$ the probability space corresponding to the $(p+r)$-dimensional Brownian motion (U_t, V_t) and

$$\mathcal{K} = \{S : \Omega \mapsto \mathrm{I\!R}^n,\ S \in L^2(\Omega, \mathcal{F}, P),\ S\ \mathcal{G} - \text{measurable}\}\,.$$

"Based on the observations Z_s" means that the estimate $\hat{X}_t$ is $\mathcal{G}_t$-measurable. $\hat{X}_t$ shall be "the best estimate based on the observations Z_s" which means that

$$\mathrm{E}[|X_t - \hat{X}_t|^2] = \inf_{S \in \mathcal{K}} \{\mathrm{E}[|X_t - S|^2]\}.$$

Let $P_{\mathcal{K}_t}(X_t)$ denote the orthogonal projection from $L^2(\Omega, \mathcal{F}, P)$ onto the subspace $\mathcal{K}_t$. The following statement holds

$$\hat{X}_t = P_{\mathcal{K}_t}(X_t) = \mathrm{E}(X_t|\mathcal{G}) .$$

We will concentrate on the linear case, which allows an explicit solution in terms of a stochastic differential equation for $\hat{X}_t$. This method is called the Kalman-Bucy filter.

A.1.2 The linear filtering problem

To focus on the main ideas consider only the one-dimensional case

$$\begin{aligned} dX_t &= F(t)X_t\,dt + C(t)\,dU_t\,, \quad \text{(System)} &\text{(A.3)}\\ dZ_t &= G(t)X_t\,dt + D(t)\,dV_t\,, \quad \text{(Observations)} &\text{(A.4)}\end{aligned}$$

where the real functions $F(t)$, $G(t)$, $C(t)$, $D(t)$ are bounded on bounded intervals. Furthermore assume that $D(t)$ is bounded away from 0 on bounded intervals, $Z_0 = 0$, X_0 is normally distributed and independent of $\{U_t\}$, $\{V_t\}$ and $\mathrm{E}[X_0] = 0$ (see Øksendal [56], p. 48f). The ideas and techniques in the one-dimensional case can be extended in an analogous way to the multi-dimensional case.

By means of some technical transformations of $P_{\mathcal{K}_t}(X_t)$ and calculations we obtain

The Kalman-Bucy Filter

The solution

$$\hat{X}_t = P_{\mathcal{K}_t}(X_t) = \mathrm{E}(X_t|\mathcal{G})$$

of the linear filtering problem (A.3), (A.4) satisfies the stochastic differential equation

$$d\hat{X}_t = \left[F(t) - \frac{G^2(t)S(t)}{D^2(t)}\right]\hat{X}_t\,dt + \frac{G(t)S(t)}{D^2(t)}\,dZ_t\,,\ \hat{X}_0 = \mathrm{E}[X_0]\,, \tag{A.5}$$

where $S(t) = \mathrm{E}[(X_t - \hat{X}_t)^2]$ satisfies the "Riccati-equation"

$$\frac{dS}{dt} = 2F(t)S(t) - \frac{G^2(t)}{D^2(t)}S^2(t) + C^2(t)\,, \tag{A.6}$$

with $S(0) = \mathrm{E}[(X_0 - \mathrm{E}[X_0])^2]$.

A.2 The discrete case

For reasons of completeness we again note here the Kalman-Bucy filter for the discrete case, as already given in section 3.1.2.

Consider the stochastic system

$$X_i = D_i X_{i-1} + S_i + \varepsilon_i,\ i = 1, \ldots, n, \ \text{(System)} \tag{A.7}$$

where $\{X_i\}_{i=0}^n$ are random $d \times 1$ vectors, $\{D_i\}_{i=0}^n$ are non-random $d \times d$ matrices, $\{S_i\}_{i=1}^n$ are non-random $d \times 1$ vectors, $X_0 \sim \mathcal{N}_d(x_0, V_0)$, $\varepsilon_i \sim \mathcal{N}_d(0, V_i)$, $i = 1, \ldots, n$ and X_0, $\varepsilon_1, \ldots, \varepsilon_n$ are stochastically independent. Assume the observable quantities are $Y_0, Y_1, \ldots, Y_n$ given by

$$Y_i = T_i X_i + U_i + e_i,\ i = 0, 1, \ldots, n, \ \text{(Observations)} \tag{A.8}$$

where $\{T_i\}_{i=0}^n$ are non-random $k \times d$ matrices ($k \leq d$), $\{U_i\}_{i=0}^n$ are non-random $k \times 1$ vectors, $e_i \sim \mathcal{N}_k(0, W_i)$ and $X_0, \varepsilon_1, \ldots, \varepsilon_m, e_0, e_1, \ldots, e_n$ are stochastically independent, $i = 0, 1, \ldots, n$.

The Kalman-Bucy filter

Under some assumptions (see Pedersen [59], pp. 4–5), we have for given observations $y_0, y_1, \ldots, y_n$ of $Y_0, Y_1, \ldots, Y_n$

$$\begin{aligned}
X_i | Y^i = y^i &\sim \mathcal{N}_d\left(\mu_i(y^i), \Sigma_i\right), && \text{(A.9)}\\
X_i | Y^{i-1} = y^{i-1} &\sim \mathcal{N}_d\left(D_i \mu_{i-1}(y^{i-1}) + S_i, R_i\right), && \text{(A.10)}\\
Y_i | Y^{i-1} = y^{i-1} &\sim \mathcal{N}_d\left(T_i(D_i \mu_{i-1}(y^{i-1}) + S_i) + U_i, T_i R_i T_i^T + W_i\right), && \text{(A.11)}
\end{aligned}$$

where $R_i = D_i \Sigma_{i-1} D_i^T + V_i$ is positive definite, and where

$$\begin{aligned}
\mu_0(y^0) &= x_0 + V_0 T_0^T \left(T_0 V_0 T_0^T + W_0\right)^{-1} (y_0 - T_0 x_0 - U_0), && \text{(A.12)}\\
\Sigma_0 &= V_0 - V_0 T_0^T \left(T_0 V_0 T_0^T + W_0\right)^{-1} T_0 V_0, && \text{(A.13)}\\
\mu_i(y^i) &= D_i \mu_{i-1}(y^{i-1}) + S_i + R_i T_i^T \left(T_i R_i T_i^T + W_i\right)^{-1} \\
&\quad \left(y_i - T_i \left(D_i \mu_{i-1}(y^{i-1}) + S_i\right) - U_i\right), && \text{(A.14)}\\
\Sigma_i &= R_i - R_i T_i^T (T_i R_i T_i^T + W_i)^{-1} T_i R_i. && \text{(A.15)}
\end{aligned}$$

Appendix B

Numerical Methods

In our representation we follow [45], §9 and §10.

B.1 The Euler Scheme

The Euler approximation is a basic discrete time method to approximate an Itô process. Consider an Itô process $X = \{X(t), t_0 \leq t \leq T\}$ following the scalar stochastic differential equation

$$dX_t = a(t, X_t)dt + b(t, X_t)dW_t,$$

with $t_0 \leq t \leq T$ and the initial condition $X_{t_0} = X_0$. A discretization $t_0 = t_0 < t_1 < \ldots < t_n < \ldots < t_N = T$ of the time interval $[0, T]$ may be given. Then a continuous time stochastic process $Y = \{Y(t), t_0 \leq t \leq T\}$ with the initial condition

$$Y_0 = X_0,$$

satisfying the stochastic iterative scheme

$$Y_{n+1} = Y_n + a(t_n, Y_n)(t_{n+1} - t_n) + b(t_n, Y_n)(W_{t_{n+1}} - W_{t_n}), \tag{B.1}$$

for $n = 0, 1, \ldots, N-1$, where we denote $Y_n = Y(t_n)$, is called an **Euler approximation** of X. The scheme (B.1) is called the **Euler scheme**.

With the notations

$$\Delta_n = t_{n+1} - t_n, \quad \Delta W_n = W_{t_{n+1}} - W_{t_n}$$

and

$$a = a(t_n, Y_n), \quad b = b(t_n, Y_n),$$

we can write the Euler scheme (B.1) in the impressive form

$$Y_{n+1} = Y_n + a\Delta_n + b\Delta W_n \tag{B.2}$$

for $n = 0, 1, \ldots, N-1$.

In order to compute the sequence $\{Y_n, n = 0, 1, \ldots, N-1\}$ of values of the Euler approximation we have to generate the random increments ΔW_n for $n = 0, 1, \ldots, N-1$ of the Wiener process $W = \{W_t, t \geq 0\}$. These increments are independent Gaussian random variables with $\mathrm{E}(\Delta W_n) = 0$ and $\mathrm{Var}(\Delta W_n) = \Delta_n$ and can be generated by a random number generator (see e.g. [45], §1.3).

For the multi-dimensional case of the Euler scheme see e.g. [45], §10.2.

Note that when the diffusion coefficient b is identically zero the stochastic iterative scheme (B.2) reduces to the well-known deterministic Euler scheme for the ordinary differential equation $x' = a(t, x)$.

We introduce the notion of strong convergence.

Definition 1 A time discrete approximation Y^δ with maximum step size δ **converges strongly** to X at time T if

$$\lim_{\delta \downarrow 0} \mathrm{E}(|X_T - Y^\delta(T)|) = 0.$$

The rate of strong convergence is crucial if we want to compare different time discrete approximation methods.

Definition 2 A time discrete approximation Y^δ **converges strongly with order $\boldsymbol{\gamma}$** > 0 at time T, if there exists a constant $C > 0$, independent of δ, and a $\delta_0 > 0$ such that

$$\mathrm{E}(|X_T - Y^\delta(T)|) \leq C\, \delta^\gamma$$

for each $\delta \in (0, \delta_0)$.

Under some regularity assumptions **the Euler scheme converges strongly with order $\boldsymbol{\gamma = 0.5}$** :

Theorem 8 *Suppose*

$$E(|X_0|^2) < \infty,$$
$$E\left(|X_0 + Y_0^\delta|^2\right)^{\frac{1}{2}} \leq C_1\, \delta^{\frac{1}{2}},$$
$$|a(t,x) - a(t,y)| + |b(t,x) - b(t,y)| \leq C_2 |x - y|,$$
$$|a(t,x)| + |b(t,x)| \leq C_3(1 + |x|)$$

and
$$|a(s,x) - a(t,x)| + |b(s,x) - b(t,x)| \leq C_4(1+|x|)|s-t|^{\frac{1}{2}}$$
for all $s, t \in [0,T]$ *and* $x, y \in \mathbb{R}^d$. *Then*
$$E\Big(|X_T - Y^\delta(T)|\Big) \leq C_5\,\delta^{\frac{1}{2}}$$
for the Euler approximation Y^δ.

For the proof see [45], §10.2. We will compare the Euler method with the Milstein method described in the next section.

B.2 The Milstein Scheme

By adding to the Euler scheme (B.2) the term
$$\frac{1}{2} b\, b' \,[(\Delta W)^2 - \Delta]$$
we obtain the **Milstein scheme**
$$Y_{n+1} = Y_n + a\Delta + b\Delta W + \frac{1}{2} b\, b' \,[(\Delta W)^2 - \Delta]. \qquad \text{(B.3)}$$

If a and b are smoothly enough **the Milstein scheme** can be shown to have **order of strong convergence** $\boldsymbol{\gamma = 1.0}$. In comparison with the Euler method the strong convergence order is increased from $\gamma = 0.5$ to $\gamma = 1.0$ and we conclude that the Milstein method is an improvement of the Euler method.

We remark that in the case of the diffusion term $b = 0$, that is in the deterministic case, the (deterministic) Euler scheme has strong convergence order $\gamma = 1.0$. Therefore, as to the order of strong convergence, the Milstein scheme can be seen as a generalization of the deterministic Euler scheme.

Appendix C

The Itô Formula

C.1 The one-dimensional case

The function $U : [0, T] \times \mathrm{I\!R} \mapsto \mathrm{I\!R}$ have continuous partial derivatives $\frac{\partial U}{\partial t}, \frac{\partial U}{\partial x}$ and $\frac{\partial^2 U}{\partial x^2}$ and X_t satisfy the one-dimensional Itô stochastic differential equation

$$dX_t = a(t, \omega)dt + b(t, \omega)dW_t,$$

where $\sqrt{|a|}$ and b are in the space L^2. Define a process Y_t by $Y_t = U(t, X_t)$ for $0 \leq t \leq T$. Then

$$\begin{aligned} dY_t &= \left[\frac{\partial U}{\partial t}(t, X_t) + a_t \frac{\partial U}{\partial x}(t, X_t) + \frac{1}{2}\, b_t^2 \frac{\partial^2 U}{\partial x^2}(t, X_t)\right] dt \\ &\quad + b_t \, \frac{\partial U}{\partial x}(t, X_t)\, dW_t, \end{aligned}$$

w. p. 1 for $0 \leq t \leq T$, and with the notation $a_t = a(t, \omega), b_t = b(t, \omega)$.

C.2 The multi-dimensional case

The process X_t satisfy the d-dimensional Itô stochastic differential equation

$$dX_t = a(t, \omega)dt + B(t, \omega)dW_t,$$

where $\{W_t, t \geq 0\}$ is an m-dimensional Wiener process with independent components, $W_t = (W_t^1, W_t^2, \ldots, W_t^m)$, and $a : [0, T] \times \Omega \mapsto \mathrm{I\!R}^d$, $B : [0, T] \times \Omega \mapsto \mathrm{I\!R}^{d \times m}$, satisfying $\sqrt{|a^k|}$ and $B^{k,j} \in L^2$ for $k = 1, \ldots, d$, $j = 1, \ldots, m$.

The function $U : [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}$ have continuous partial derivatives $\frac{\partial U}{\partial t}, \frac{\partial U}{\partial x_k}$ and $\frac{\partial^2 U}{\partial x_k \partial x_i}$ for $k, i = 1, 2, \ldots, d$. Define a scalar process $\{Y_t, 0 \leq t \leq T\}$ by $Y_t = U(t, X_t) = U(t, X_t^1, X_t^2, \ldots, X_t^d)$ w. p. 1. Then

$$\begin{aligned} dY_t &= \left[\frac{\partial U}{\partial t} + \sum_{k=1}^{d} a_t^k \frac{\partial U}{\partial x_k} + \frac{1}{2} \sum_{j=1}^{m} \sum_{i,k=1}^{d} B_t^{i,j} B_t^{k,j} \frac{\partial^2 U}{\partial x_i \partial x_k} \right] dt \\ &\quad + \sum_{j=1}^{m} \sum_{i=1}^{d} B_t^{i,j} \frac{\partial U}{\partial x_i} dW_t^j, \end{aligned}$$

w. p. 1 for $0 \leq t \leq T$, where the partial derivatives are evaluated at (t, X_t) and where we denote $a_t^k = a^k(t, \omega)$, $B_t^{i,j} = B^{i,j}(t, \omega)$.

For an extension of the Itô formula to a wider class of non-smooth functions we refer to the recently published paper by Föllmer, Protter and Shiryaev [26].

Appendix D

The Radon-Nikodym Theorem

In our notation we follow Billingsley [9], §32.

1) A probability measure μ on a field $\mathcal{F}$ in Ω is called **finite**, if $\mu(\Omega) < \infty$.

2) If $\Omega = A_1 \cup A_2 \cup \ldots$ for some finite or countable sequence of $\mathcal{F}$-sets satisfying $\mu(A_k) < \infty$, then μ is $\boldsymbol{\sigma}$**-finite**.

3) A measure ν is **absolutely continuous** with respect to μ if for each A in $\mathcal{F}$

$$\mu(A) = 0 \Rightarrow \nu(A) = 0.$$

This relation is indicated by $\nu << \mu$. If $\nu << \mu$ and $\mu << \nu$, the measures are **equivalent**, indicated by $\nu \sim \mu$.

The Radon-Nikodym Theorem

If μ and ν are σ-finite measures such that $\nu << \mu$, then there exists a nonnegative f, a density, such that

$$\nu(A) = \int_A f d\mu$$

for all A in $\mathcal{F}$.

The density f is called the **Radon-Nikodym derivative** of ν with respect to μ and is denoted by $\frac{d\nu}{d\mu}$.

Note that in the σ-finite case there is a countable decomposition of Ω into $\mathcal{F}$-sets A_n for which $\mu(A_n)$ and $\nu(A_n)$ are finite. Because of this argument it is sufficient to treat finite μ and ν (see [9], p.444).

Bibliography

[1] Ball, C.A., 1993, "A Review of Stochastic Volatility Models with Application to Option Pricing", *Financial Markets, Institutions and Instruments*, Vol. 2, No. 5, pp. 55-71.

[2] Basawa, I.V. and B.L.S. Prakasa Rao, 1980, "Statistical Inference for Stochastic Processes", *Academic Press Inc.*

[3] Bernardo, J.M. and A.F.M. Smith, 1994, "Bayesian Theory", *Wiley, New York.*

[4] Bibby, B.M, 1994, "A Two-Compartment Model with Additive White Noise", *Research Report No.290, Dept. Theor. Statist., University of Aarhus.*

[5] Bibby, B.M, 1994, "Optimal Combination of Martingale Estimating Functions for Discretely Observed Diffusion Processes", *Research Report No.298, Dept. Theor. Statist., University of Aarhus.*

[6] Bibby, B.M., 1995, "Analysis of a Tracer Experiment Based on Compartmental Diffusion Models", *Research Report No.303, Dept. Theor. Statist., University of Aarhus.*

[7] Bibby, B.M., 1995, "On Estimation in Compartmental Diffusion Models", *Research Report No.305, Dept. Theor. Statist., University of Aarhus.*

[8] Bibby, B.M and M. Sørensen, 1995, "Martingale Estimation Functions for Discretely Observed Diffusion Processes", *Bernoulli* Vol. 1 No.(1/2), pp. 17-39.

[9] Billingsley, P., 1986, "Probability and Measure", 2nd ed., *Wiley, New York.*

[10] Billingsley, P., 1968, "Convergence of Probability Measures", *Wiley, New York.*

[11] Black, F., 1990, "Living up to the model", *Risk Magazine*, Mar 1990.

[12] Bollerslev, T., 1986, "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics* 31, pp. 307-327.

[13] Bollerslev, T., R.Y. Chou and K.F. Kroner, 1992, "ARCH Modeling in Finance", *Journal of Econometrics* 52, pp. 5-59.

[14] Bollerslev, T., R.F. Engle and D.B. Nelson, 1994, "ARCH Models", *The Handbook of Econometrics*, Vol. 4.

[15] Breiman, L., 1968, "Probability", *Addison-Wesley Publishing Company.*

[16] Brockwell, P.J. and R.A. Davis, 1987, "Time Series: Theory and Methods", *Springer-Verlag, New York.*

[17] Cramér, H., 1946, "Mathematical Methods of Statistics", *Princeton University Press.*

[18] Dacunha-Castelle, D. and M. Duflo, 1993, "Probabilités et Statistiques", *Masson, Paris.*

[19] Dacunha-Castelle, D. and D. Florens-Zmirou, 1986, "Estimation of the Coefficients of a Diffusion from Discrete Observations", *Stochastics* 19, pp. 263-284.

[20] Deelstra, G. and G. Parker, 1995, "A Covariance Equivalent Discretisation of the CIR Model", *Proceedings of the 5th AFIR International Colloquium,* Vol.2, pp. 731-747.

[21] Duffie, D. and K.J. Singleton, 1995, "An Econometric Model of the Term Structure of Interest Rate Swap Yields", *Working paper, Graduate School of Business, Stanford University.*

[22] Engle, R.F., 1982, "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation", *Econometrica* 50, pp. 987-1008.

[23] Engle, R.F. and M. Rothschild, 1992, "ARCH Models in Finance", *Journal of Econometrics, Annals* 1992-1, Vol. 52.

[24] Feigin, P.D., 1976, "Maximum Likelihood Estimation for Continuous-time Stochastic Processes', *Adv. Appl. Probab.*, Vol. 8, pp. 712-736.

[25] Florens-Zmirou, D., 1989, "Approximate Discrete-Time Schemes for Statistics of Diffusion Processes", *Statistics* 20, pp. 547-557.

[26] Föllmer, H., P. Protter and A.N. Shiryayev, 1995, "Quadratic covariation and an extension of Itô's formula", *Bernoulli* Vol. 1 No.(1/2), pp. 149-169.

[27] Fournie, E. and D. Talay, "Application de la Statistique des Diffusions a un Modèle de Taux d'Intérêt", *INRIA*, *forthcoming in Finance.*

[28] Gallant, A.R., 1981, "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form", *Journal of Economics* 15, pp. 211-244.

[29] Geman, S. and D. Geman, 1984, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Trans. Pattern Anal. Machine Intell.,* Vol. PAMI-6, No. 6., pp. 721-741.

[30] Genon-Catalot, V. and D. Picard, 1993, "Eléments de statistique asymptotique", *Springer-Verlag France, Paris.*

[31] Geweke, J., 1989, "Bayesian Inference in Econometric models using Monte Carlo integration", *Econometrica* 57, pp. 1317-1339.

[32] Gihman, I.I. and A.V. Skorokhod, 1972, "Stochastic Differential Equations", *Springer-Verlag, Berlin.*

[33] Godambe, V.P. and C.C. Heyde, 1987, "Quasi-likelihood and Optimal Estimation", *Int. Statist. Rev.* 55, pp. 231-244.

[34] Hamilton, J.D., 1994, "Time Series Analysis", *Princeton University Press.*

[35] Heyde, C.C., 1987, "On Combining Quasi-likelihood Estimating Functions", *Stochastic Processes and their Applications* 25, pp. 281-287.

[36] Heyde, C.C., 1989, "Quasi-Likelihood and Optimality for Estimating Functions: Some Current Unifying Themes", *Fisher Lecture, I.S.I, 47th Session, Paris Aug 29 – Sep 6, 1989.*

[37] Hutton, J.E. and P.I. Nelson, 1986, "Quasi-Likelihood Estimation for Semimartingales", *Stochastic Processes and their Applications* 22, pp. 245-257.

[38] Ibragimov, I.A. and R.Z. Has'minskii, 1981, "Statistical Estimation", *Springer-Verlag, New York.*

[39] Jacod, J., 1994, "Statistics of Diffusion Processes: Some Elements", *Istituto per le Applicazioni della Matematica e dell'Informatica* 94.17.

[40] Jacod, J., "La variation quadratique du brownien en pr'esence d'erreurs d'arrondi", *Working Paper, Laboratoire de Probabilités, Université Pierre et Marie Curie, Paris.*

[41] Jacod, J. and A.N. Shiryaev, 1987, "Limit Theorems for Stochastic Processes", *Springer-Verlag, Berlin.*

[42] Jacquier, E., N.G. Polson and P.E. Rossi, 1994, "Bayesian Analysis of Stochastic Volatility Models", *Journal of Business and Economic Statistics*, Oct 1994, Vol. 12, No. 4.

[43] Kallianpur, G., 1980, "Stochastic Filtering Theory", *Springer-Verlag, New York.*

[44] Kallianpur, G. and R.L. Karandikar, 1988, "White Noise Theory of Prediction, Filtering and Smoothing", *Gordon and Breach Science Publishers.*

[45] Kloeden, P.E. and E. Platen, 1992, "Numerical Solution of Stochastic Differential Equations", *Springer-Verlag, New York.*

[46] Kloeden, P.E., H. Schurz, E. Platen and M. Sørensen, 1992, "On Effects of Discretization on Estimators of Drift Parameters for Diffusion Processes", *Research Report No.249, Dept. Theor. Statist., University of Aarhus.*

[47] Kutoyants, Yu.A., 1994, "Identification of Dynamical Systems with Small Noise", *Kluwer Academic Publishers.*

[48] Kutoyants, A.Y., 1984, "Parameter Estimation for Stochastic Processes", *Heldermann-Verlag, Berlin.*

[49] Lindgren, B.W., 1976, "Statistical Theory", 3rd ed., *Macmillan Publishing Co.*

[50] Linton, O., 1993, "Adaptive Estimation in ARCH Models", *Econometric Theory* 9, pp. 539-569.

[51] Liptser, R.S. and A.N. Shiryaev, 1977, "Statistics of Random Processes I, II", *Springer-Verlag, New York*.

[52] Luschgy, H. and A.L. Rukhin, 1993, "Asymptotic Properties of Tests for a Class of Diffusion Processes: Optimality and Adaption", *Mathematical Methods of Statistics, Allerton Press, Inc.*, Vol. 2, No. 1, pp. 42-51.

[53] Mills, T.C., 1993, "The Econometric Modelling of Financial Time Series", *Cambridge University Press*.

[54] Nelson, D.B., 1990, "ARCH Models as Diffusion Approximations", *Journal of Econometrics* 45, pp. 7-38.

[55] O'Hagan, A., 1994, "Bayesian inference", *Kendall's Advanced Theory of Statistics, Arnold*, Vol. 2B.

[56] Øksendal, B., 1989, "Stochastic Differential Equations", Second Edition, *Springer-Verlag, Berlin*.

[57] Pagan, A.R. and G.W. Schwert, 1990, "Alternative Models for Conditional Stock Volatility", *Journal of Econometrics* 45, pp. 267-290.

[58] Pedersen, A.R., 1993, "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*.

[59] Pedersen, A.R., 1993, "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*.

[60] Pedersen, A.R., 1995, "Consistency and Asymptotic Normality of an Approximate Maximum Likelihood Estimator for Discretely Observed Diffusion Processes", *Bernoulli* Vol.1 No.3, pp. 257-279.

[61] Pedersen, A.R., 1994, "Uniform Residuals for Discretely Observed Diffusion Processes", *Research Report No.290, Dept. Theor. Statist., University of Aarhus*.

[62] Pedersen, A.R., 1994, "Quasi-Likelihood Inference for Discretely Observed Diffusion Processes", *Research Report No.295, Dept. Theor. Statist., University of Aarhus*.

[63] Pedersen, A.R., 1994, "Statistical Analysis of Gaussian Diffusion Processes Based on Incomplete Discrete Observations", *Research Report No.297, Dept. Theor. Statist., University of Aarhus.*

[64] Priestley, M.B., 1981, "Spectral analysis and time series", Vol.1 and 2, *Academic Press.*

[65] Rogers, L.C.G. and D. Williams, 1987, "Diffusions, Markov Processes and Martingales, Vol. 2: Itô Calculus", *Wiley, Chichester.*

[66] Shiryayev, A.N., 1984, "Probability", *Springer-Verlag, New York.*

[67] Shiryayev, A.N. and V.G. Spokoiny, "Statistical Experiments and Decisions: Asymptotic Theory", forthcoming in *Springer-Verlag.*

[68] Shorack, G.R. and J.A. Wellner, 1986, "Empirical Processes with Applications to Statistics", *Wiley, New York.*

[69] Silverman, B.W., 1986, "Density Estimation for Statistics and Data Analysis", *Chapman and Hall, London.*

[70] Skorokhod, A.V., 1989, "Asymptotic Methods in the Theory of Stochastic Differential Equations", *American Mathematical Society, Translations of Mathematical Monographs* Vol. 78.

[71] Smith, A.F.M. and G.O.Roberts, 1993, "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods", *J.R.Statist.Soc.* B, 55, No. 1, pp. 3-23.

[72] Stein, E.M. and J.C. Stein, 1991, "Stock Price Distributions with Stochastic Volatility: An Analytic Approach", *The Review of Financial Studies*, Vol. 4, No. 4, pp. 727-752.

[73] Stroock, D.W. and S.R.S. Varadhan, 1979, "Multidimensional Diffusion Processes", *Springer-Verlag, Berlin.*

[74] Taylor, S., 1986, "Modelling Financial Time Series", *Wiley, Chichester.*

[75] Wiggins, J.B., 1987, "Option Values Under Stochastic Volatility: Theory and Empirical Estimates", *Journal of Financial Economics*, 19, pp. 351-372.

[76] Wong, E. and B. Hajek, 1985, "Stochastic Processes in Engineering Systems", *Springer-Verlag, New York.*