Contents

1 Estimating functions for diffusion-type processes 1

by Michael Sørensen

1.1 Introduction 1

1.2 Low frequency asymptotics 3

1.3 Martingale estimating functions 7

1.3.1 Asymptotics 8

1.3.2 Likelihood inference 10

1.3.3 Godambe-Heyde optimality 12

1.3.4 Small Δ-optimality 22

1.3.5 Simulated martingale estimating functions 27

1.3.6 Explicit martingale estimating functions 30

1.3.7 Pearson diffusions 34

1.3.8 Implementation of martingale estimating functions 42

1.4 The likelihood function 45

1.5 Non-martingale estimating functions 49

1.5.1 Asymptotics 49

1.5.2 Explicit non-martingale estimating functions 51

1.5.3 Approximate martingale estimating functions 54

1.6 High-frequency asymptotics 56

1.7 High-frequency asymptotics in a fixed time-interval 63

1.8 Small-diffusion asymptotics 65

1.9 Non-Markovian models 70

1.9.1 Prediction-based estimating functions 70

1.9.2 Asymptotics 76

1.9.3 Measurement errors 77

1.9.4 Integrated diffusions and hypoelliptic stochastic differential equations 78

1.9.5 Sums of diffusions 80

1.9.6 Stochastic volatility models 83

1.9.7 Compartment models 84

1.10 General asymptotics results for estimating functions 85

1.11 Optimal estimating functions: general theory 88

1.11.1 Martingale estimating functions 92

Bibliography 97

CHAPTER 1

Estimating functions for diffusion-type processes

Michael Sørensen
University of Copenhagen

1.1 Introduction

In this chapter we consider parametric inference based on discrete time observations $X_0, X_{t_1}, \ldots, X_{t_n}$ from a d-dimensional stochastic process. In most of the chapter the statistical model for the data will be a diffusion model given by a stochastic differential equation. We shall, however, also consider some examples of non-Markovian models, where we typically assume that the data are partial observations of a multivariate stochastic differential equation. We assume that the statistical model is indexed by a p-dimensional parameter θ.

The focus will be on estimating functions. An *estimating function* is a p-dimensional function of the parameter θ and the data:

$$G_n(\theta; X_0, X_{t_1}, \ldots, X_{t_n}).$$

Usually we suppress the dependence on the observations in the notation and write $G_n(\theta)$. We obtain an estimator by solving the equation

$$G_n(\theta) = 0. \tag{1.1}$$

Estimating functions provide a general framework for finding estimators and studying their properties in many different kinds of statistical models. The estimating function approach has turned out to be very useful for discretely sampled parametric diffusion-type models, where the likelihood function is usually not explicitly known. Estimating functions are typically constructed by combining relationships (dependent on the unknown parameter) between an observation and one or more of the previous observations that are informative about the parameters.

As an example, suppose the statistical model for the data $X_0, X_\Delta, X_{2\Delta}, \ldots,$

$X_{n\Delta}$ is the one-dimensional stochastic differential equation

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

where $\theta > 0$ and W is a Wiener process. The state-space is $(-\pi/2, \pi/2)$. This model will be considered in more detail in Subsection 1.3.6. For this process Kessler & Sørensen (1999) proposed the estimating function

$$G_n(\theta) = \sum_{i=1}^{n} \sin(X_{(i-1)\Delta}) \left[\sin(X_{i\Delta}) - e^{-(\theta+\frac{1}{2})\Delta} \sin(X_{(i-1)\Delta}) \right],$$

which can be shown to be a martingale, when θ is the true parameter. For such martingale estimating functions, asymptotic properties of the estimators as the number of observations tends to infinity can be studied by means of martingale limit theory, see Subsection 1.3.1. An explicit estimator $\hat{\theta}_n$ of the parameter θ is obtained by solving the estimating equation (1.1):

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

provided that

$$\sum_{i=1}^{n} \sin(X_{(i-1)\Delta}) \sin(X_{i\Delta}) > 0. \tag{1.2}$$

If this condition is not satisfied, the estimating equation (1.1) has no solution, but fortunately it can be shown that the probability that (1.2) holds tends to one as n tends to infinity. As illustrated by this example, it is quite possible that the estimating equation (1.1) has no solution. We shall give general conditions that ensure the existence of a unique solution as the number of observations tend to infinity.

The idea of using estimating equations is an old one and goes back at least to Karl Pearson's introduction of the method of moments. The term estimating function may have been coined by Kimball (1946). In the econometric literature, the method was introduced by Hansen (1982) and is known as the generalized method of moments (GMM).

A general asymptotic theory for estimating functions is presented in Section 1.10, and Section 1.11 reviews the theory of optimal estimating functions. Given a collection of relations between observations at different time points that can be used for estimation, this theory clarifies how to combine the relations in an optimal way, i.e. in such a way that the most efficient estimator is obtained. In Section 1.2 we present conditions ensuring that estimators from estimating functions are consistent and asymptotically normal under the so-called low frequency asymptotics, which is the same as usual large sample asymptotics. In Section 1.3 we present martingale estimating functions for diffusion models including asymptotics and two optimality criteria. One of these criteria, small Δ-optimality, is particular of diffusion models. Likelihood inference is included as a particular case. It is also discussed how to implement martingale estimating functions. There are several methods available for calculating approximations to the likelihood function. These a briefly reviewed in Section 1.4, where a particular expansion approach is presented in detail. Non-martingale estimating functions are considered in Section 1.5. In important aspect of the statistical theory of diffusion processes is that a number of alternative asymptotic scenarios particular to diffusions are available to supplement the traditional large sample asymptotics. High frequency asymptotics, high frequency asymptotics in a fixed time-interval and small-diffusion asymptotics are presented in the Sections 1.6, 1.7, and 1.8. A number of non-Markovian models are considered in Section 1.9, including observations with measurement errors, integrated diffusions, sums of diffusions, stochastic volatility models and compartment models. A general tool for these models are prediction-based estimating functions, which generalize the martingale estimating functions and share some of their convenient features.

1.2 Low frequency asymptotics

In this section, we assume that observations have been made at the equidistant time points $i\Delta$, $i = 1, \ldots, n$, and consider the classical large sample asymptotic scenario, where the time between observations, Δ, is fixed, and the number of observations, n, goes to infinity. Since Δ is fixed, we will generally suppress Δ in the notation in this section. We assume that the statistical model is indexed by a p-dimensional parameter $\theta \in \Theta$, which we want to estimate. The corresponding probability measures are denoted by P_θ. The distribution of the data is given by the true probability measure, which we denote by P.

Under the true probability measure, P, it is assumed that $\{X_{i\Delta}\}$ is a stationary process with state space $D \subseteq \mathbb{R}^d$. We study the asymptotic properties of an estimator, $\hat{\theta}_n$, obtained by solving the estimating equation (1.1) when G_n is an estimating function of the general form

$$G_n(\theta) = \frac{1}{n} \sum_{i=r}^{n} g(X_{(i-r+1)\Delta}, \ldots, X_{i\Delta}; \theta), \tag{1.3}$$

where r is a fixed integer smaller than n, and g is a suitable function with values in $\mathbb{R}^p$. All estimators discussed in this chapter can be represented in this way. We shall present several useful examples of how g can be chosen in the subsequent sections. A priori there is no guarantee that a unique solution to (1.1) exists, but conditions ensuring this for large sample sizes are given below. By a G_n–estimator, we mean an estimator, $\hat{\theta}_n$, which solves (1.1) when the data belongs to a subset $A_n \subseteq D^n$, and is otherwise given a value $\delta \notin \Theta$. We give results ensuring that, as $n \to \infty$, the probability of A_n tends to one.

Let Q denote the joint distribution of $(X_\Delta, \ldots, X_{r\Delta})$, and $Q(f)$ the expectation of $f(X_\Delta, \ldots, X_{r\Delta})$ for a function $f : D^r \mapsto \mathrm{I\!R}$. To obtain our asymptotic results about G_n–estimators, we need to assume that a law of large numbers (an ergodic theorem) as well as a central limit theorem hold. Specifically, we assume that as $n \to \infty$

$$\frac{1}{n} \sum_{i=r}^{n} f(X_{(i-r+1)\Delta}, \ldots, X_{i\Delta}) \xrightarrow{P} Q(f) \tag{1.4}$$

for any function $f : D^r \mapsto \mathrm{I\!R}$ such that $Q(|f|) < \infty$, and that the estimating function (1.3) satisfies

$$\frac{1}{\sqrt{n}} \sum_{i=r}^{n} g(X_{(i-r+1)\Delta}, \ldots, X_{i\Delta}; \theta) \xrightarrow{\mathcal{D}} N(0, V(\theta)) \tag{1.5}$$

under P for any $\theta \in \Theta$ for which $Q(g(\theta)) = 0$. Here $V(\theta)$ is a positive definite $p \times p$-matrix. Moreover, $g(\theta)$ denotes the function $(x_1, \ldots, x_r) \mapsto g(x_1, \ldots, x_r; \theta)$, convergence in probability under P is indicated by $\xrightarrow{P}$, and $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

The following condition ensures the existence of a consistent G_n–estimator. We denote transposition of matrices by ^{T}, and $\partial_{\theta^T} G_n(\theta)$ denotes the $p \times p$-matrix, where the ijth entry is $\partial_{\theta_j} G_n(\theta)_i$.

Condition 1.2.1 *A parameter value $\bar{\theta} \in \operatorname{int}\Theta$ and a neighbourhood N of $\bar{\theta}$ in Θ exist such that:*

(1) The function $g(\theta) : (x_1, \ldots, x_r) \mapsto g(x_1, \ldots, x_r; \theta)$ is integrable with respect to Q for all $\theta \in N$, and

$$Q\left(g(\bar{\theta})\right) = 0. \tag{1.6}$$

(2) The function $\theta \mapsto g(x_1, \ldots, x_r; \theta)$ is continuously differentiable on N for all $(x_1, \ldots, x_r) \in D^r$.

(3) The function $(x_1, \ldots, x_r) \mapsto \|\partial_{\theta^T} g(x_1, \ldots, x_r; \theta)\|$ is dominated for all $\theta \in N$ by a function which is integrable with respect to Q.

(4) The $p \times p$ matrix

$$W = Q\left(\partial_{\theta^T} g(\bar{\theta})\right) \tag{1.7}$$

is invertible.

Here and later $Q(g(\theta))$ denotes the vector $\left(Q(g_j(\theta))\right)_{j=1,\ldots,p}$, where g_j is the jth coordinate of g, and $Q\left(\partial_{\theta^T} g(\theta)\right)$ is the matrix $\{Q\left(\partial_{\theta_j} g_i(\theta)\right)\}_{i,j=1,\ldots,p}$.

To formulate the uniqueness result in the following theorem, we need the concept of locally dominated integrability. A function $f : D^r \times \Theta \mapsto \mathrm{I\!R}^q$ is called *locally dominated integrable* with respect to Q if for each $\theta' \in \Theta$ there exists a neighbourhood $U_{\theta'}$ of θ' and a non-negative Q-integrable function $h_{\theta'} : D^r \mapsto \mathbb{R}$ such that $|\, f(x_1, \ldots, x_r; \theta) \,| \leq h_{\theta'}(x_1, \ldots, x_r)$ for all $(x_1, \ldots, x_r, \theta) \in D^r \times U_{\theta'}$.

Theorem 1.2.2 *Assume Condition 1.2.1 and (1.5). Then a $\bar{\theta}$-consistent G_n–estimator, $\hat{\theta}_n$, exists, and*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N_p\left(0, W^{-1} V W^{T^{-1}}\right) \tag{1.8}$$

under P, where $V = V(\bar{\theta})$. If, moreover, the function $g(x_1, \ldots, x_r; \theta)$ is locally dominated integrable with respect to Q and

$$Q(g(\theta)) \neq 0 \ \textit{ for all } \ \theta \neq \bar{\theta}, \tag{1.9}$$

then the estimator $\hat{\theta}_n$ is the unique G_n–estimator on any bounded subset of Θ containing $\bar{\theta}$ with probability approaching one as $n \to \infty$.

Remark: By a $\bar{\theta}$-consistent estimator is meant that $\hat{\theta}_n \xrightarrow{P} \bar{\theta}$ as $n \to \infty$. If the true model belongs to the statistical model, i.e. if $P = P_{\theta_0}$ for some $\theta_0 \in \Theta$, then the estimator $\hat{\theta}_n$ is most useful if Theorem 1.2.2 holds with $\bar{\theta} = \theta_0$. Note that because $\bar{\theta} \in \operatorname{int}\Theta$, a $\bar{\theta}$-consistent estimator G_n–estimator $\hat{\theta}_n$ will satisfy $G_n(\hat{\theta}_n) = 0$ with probability approaching one as $n \to \infty$.

In order to prove Theorem 1.2.2, we need the following uniform law of large numbers.

Lemma 1.2.3 *Consider a function $f : D^r \times K \mapsto \mathbb{R}^q$, where K is a compact subset of Θ. Suppose f is a continuous function of θ for all $(x_1, \ldots, x_r) \in D^r$, and that there exists a Q-integrable function $h : D^r \mapsto \mathbb{R}$ such that $\|f(x_1, \ldots, x_r; \theta)\| \leq h(x_1, \ldots, x_r)$ for all $\theta \in K$. Then the function $\theta \mapsto Q(f(\theta))$ is continuous, and*

$$\sup_{\theta \in K} \| \, \frac{1}{n} \sum_{i=r}^{n} f(X_{(i-r+1)\Delta}, \ldots, X_{i\Delta}; \theta) - Q(f(\theta)) \, \| \ \overset{P}{\to} \ 0. \tag{1.10}$$

Proof: That $Q(f(\theta))$ is continuous follows from the dominated convergence theorem. To prove (1.10), define for $\eta > 0$:

$$k(\eta; x_1, \ldots, x_r) = \sup_{\theta, \theta' \in M : \|\theta' - \theta\| \leq \eta} \|f(x_1, \ldots, x_r; \theta') - f(x_1, \ldots, x_r; \theta)\|,$$

and let $k(\eta)$ denote the function $(x_1, \ldots, x_r) \mapsto k(\eta; x_1, \ldots, x_r)$. Since $k(\eta) \leq 2h$, it follows from the dominated convergence theorem that $Q(k(\eta)) \to 0$ as $\eta \to 0$. Moreover, $Q(f(\theta))$ is uniformly continuous on the compact set K. Hence for any given $\epsilon > 0$, we can find $\eta > 0$ such that $Q(k(\eta)) \leq \epsilon$ and such that $\|\theta - \theta'\| < \eta$ implies that $\|Q(f(\theta)) - Q(f(\theta'))\| \leq \epsilon$ for $\theta, \theta' \in K$.

Define the balls $B_\eta(\theta) = \{\theta' : \|\theta - \theta'\| < \eta\}$. Since K is compact, there exists a finite covering

$$K \subseteq \bigcup_{j=1}^{m} B_\eta(\theta_j),$$

where $\theta_1, \ldots, \theta_m \in K$, so for every $\theta \in K$ we can find θ_ℓ, $\ell \in \{1, \ldots, m\}$, such that $\theta \in B_\eta(\theta_\ell)$. Thus with

$$F_n(\theta) = \frac{1}{n} \sum_{i=r}^{n} f(X_{(i-r+1)\Delta}, \ldots, X_{i\Delta}; \theta)$$

we have

$$\begin{aligned}
&\|F_n(\theta) - Q(f(\theta))\| \\
&\quad\leq \|F_n(\theta) - F_n(\theta_\ell)\| + \|F_n(\theta_\ell) - Q(f(\theta_\ell))\| + \|Q(f(\theta_\ell)) - Q(f(\theta))\| \\
&\quad\leq \frac{1}{n} \sum_{\nu=r}^{n} k(\eta; X_{(\nu-r+1)\Delta}, \ldots, X_{\nu\Delta}) + \|F_n(\theta_\ell) - Q(f(\theta_\ell))\| + \epsilon \\
&\quad\leq \left| \frac{1}{n} \sum_{\nu=r}^{n} k(\eta; X_{(\nu-r+1)\Delta}, \ldots, X_{\nu\Delta}) - Q(k(\eta)) \right| \\
&\qquad\qquad + Q(k(\eta)) + \|F_n(\theta_\ell) - Q(f(\theta_\ell))\| + \epsilon \\
&\quad\leq Z_n + 2\epsilon,
\end{aligned}$$

where

$$\begin{aligned}
Z_n = &\left| \frac{1}{n} \sum_{\nu=r}^{n} k(\eta; X_{(\nu-r+1)\Delta}, \ldots, X_{\nu\Delta}) - Q(k(\eta)) \right| \\
&\qquad + \max_{1 \leq \ell \leq m} \|F_n(\theta_\ell) - Q(f(\theta_\ell))\|.
\end{aligned}$$

By (1.4), $P(Z_n > \epsilon) \to 0$ as $n \to \infty$, so

$$P\left(\sup_{\theta \in K} \|F_n(\theta) - Q(f(\theta))\| > 3\epsilon \right) \to 0$$

for all $\epsilon > 0$. □

Proof (of Theorem 1.2.2): The existence of a $\bar{\theta}$-consistent G_n–estimator $\hat{\theta}_n$ follows from Theorem 1.10.2. Condition (i) follows from (1.4) and (1.6). Define the function $W(\theta) = Q\left(\partial_{\theta^T} g(\theta)\right)$. Then condition (iii) in Theorem 1.10.2 is equal to Condition 1.2.1 (4). Finally, let M be a compact subset of N containing $\bar{\theta}$. Then the conditions of Lemma 1.2.3 are satisfied for $f = \partial_{\theta^T} g$, so

(1.158) is satisfied. The asymptotic normality, (1.20), follows from Theorem 1.10.4 and (1.5).

In order to prove the last statement, let K be a compact subset of Θ containing $\bar{\theta}$. By the finite covering property of a compact set, it follows from the local dominated integrability of g that g satisfies the conditions of Lemma 1.2.3. Hence (1.159) holds with $G(\theta) = Q(g(\theta))$ and $M = K$. From the local dominated integrability of g and the dominated convergence theorem it follows that $G(\theta)$ is a continuous function, so (1.9) implies that

$$\inf_{K \setminus \bar{B}_\epsilon(\bar{\theta})} |G(\theta)| > 0,$$

for all $\epsilon > 0$, where $\bar{B}_\epsilon(\theta)$ is the closed ball with radius ϵ centered at θ. By Theorem 1.10.3 it follows that (1.161) holds with $M = K$ for every $\epsilon > 0$. Let $\hat{\theta}'_n$ be a G_n–estimator, and define a G_n–estimator by $\hat{\theta}''_n = \hat{\theta}'_n 1\{\hat{\theta}'_n \in K\} + \hat{\theta}_n 1\{\hat{\theta}'_n \notin K\}$, where 1 denotes an indicator function, and $\hat{\theta}_n$ is the consistent G_n–estimator we know exists. By (1.161) the estimator $\hat{\theta}''_n$ is consistent, so by Theorem 1.10.2, $P(\hat{\theta}_n \neq \hat{\theta}''_n) \to 0$ as $n \to \infty$. Hence $\hat{\theta}_n$ is eventually the unique G_n–estimator on K.

□

1.3 Martingale estimating functions

In this section we consider observations $X_0, X_{t_1}, \ldots, X_{t_n}$ of a d-dimensional diffusion process given by the stochastic differential equation

$$dX_t = b(X_t; \theta)dt + \sigma(X_t; \theta)dW_t, \tag{1.11}$$

where σ is a $d \times d$-matrix and W a d-dimensional standard Wiener process. We denote the state space of X by D. When $d = 1$, the state space is an interval (ℓ, r), where ℓ could possibly be $-\infty$, and r might be ∞. The drift b and the diffusion matrix σ depend on a parameter θ which varies in a subset Θ of $\mathbb{R}^p$. The equation (1.11) is assumed to have a weak solution, and the coefficients b and σ are assumed to be smooth enough to ensure, for every $\theta \in \Theta$, the uniqueness of the law of the solution, which we denote by P_θ. We denote the true parameter value by θ_0.

We suppose that the transition distribution has a density $y \mapsto p(\Delta, x, y; \theta)$ with respect to the Lebesgue measure on D, and that $p(\Delta, x, y; \theta) > 0$ for all $y \in D$. The transition density is the conditional density under P_θ of $X_{t+\Delta}$ given that $X_t = x$.

We shall, in this section, be concerned with statistical inference based on estimating functions of the form

$$G_n(\theta) = \sum_{i=1}^{n} g(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta). \tag{1.12}$$

where g is a p-dimensional function which satisfies that

$$\int_D g(\Delta, x, y; \theta) p(\Delta, x, y; \theta) dy = 0 \tag{1.13}$$

for all $\Delta > 0$, $x \in D$ and $\theta \in \Theta$. Thus, by the Markov property, the stochastic process $\{G_n(\theta)\}_{n \in \mathbb{N}}$ is a martingale with respect to $\{\mathcal{F}_n\}_{n \in \mathbb{N}}$ under P_θ. Here and later $\mathcal{F}_n = \sigma(X_{t_i} : i \leq n)$. An estimating function with this property is called a *martingale estimating function*.

1.3.1 Asymptotics

In this subsection we give asymptotic results for estimators obtained from martingale estimating functions as the number of observations goes to infinity. To simplify the exposition the observation time points are assumed to be equidistant, i.e., $t_i = i\Delta$, $i = 0, 1, \ldots, n$. Since Δ is fixed, we will in most cases suppress Δ in the notation and write for example $p(x, y; \theta)$ and $g(x, y; \theta)$.

It is assumed that the diffusion given by (1.11) is ergodic, that its invariant probability measure has density function μ_θ for all $\theta \in \Theta$, and that $X_0 \sim \mu_\theta$ under P_θ. Thus the diffusion is stationary.

When the observed process, X, is a one-dimensional diffusion, the following simple conditions ensure *ergodicity*, and an explicit expression exists for the density of the invariant probability measure. The *scale measure* of X has Lebesgue density

$$s(x; \theta) = \exp\left(-2 \int_{x^\#}^{x} \frac{b(y; \theta)}{\sigma^2(y; \theta)} dy \right), \quad x \in (\ell, r), \tag{1.14}$$

where $x^\# \in (\ell, r)$ is arbitrary.

Condition 1.3.1 *The following holds for all $\theta \in \Theta$:*

$$\int_{x^\#}^{r} s(x; \theta) dx = \int_{\ell}^{x^\#} s(x; \theta) dx = \infty$$

and

$$\int_{\ell}^{r} [s(x; \theta) \sigma^2(x; \theta)]^{-1} dx = A(\theta) < \infty.$$

Under Condition 1.3.1 the process X is ergodic with an invariant probability measure with Lebesgue density

$$\mu_\theta(x) = [A(\theta) s(x; \theta) \sigma^2(x; \theta)]^{-1}, \quad x \in (\ell, r); \tag{1.15}$$

for details see e.g. Skorokhod (1989). For general one-dimensional diffusions, the measure with Lebesgue density proportional to $[s(x;\theta)\sigma^2(x;\theta)]^{-1}$ is called the speed measure.

Let Q_θ denote the probability measure on D^2 given by

$$Q_\theta(dx,dy) = \mu_\theta(x)p(\Delta,x,y;\theta)dxdy. \tag{1.16}$$

This is the distribution of two consecutive observations $(X_{\Delta(i-1)}, X_{\Delta i})$. Under the assumption of ergodicity the law of large numbers (1.4) is satisfied for any function $f : D^2 \mapsto \mathbb{R}$ such that $Q(|f|) < \infty$, see e.g. Skorokhod (1989).

We impose the following condition on the function g in the estimating function (1.12)

$$\begin{aligned} Q_\theta\left(g(\theta)^T g(\theta)\right) = \\ \int_{D^2} g(y,x;\theta)^T g(y,x;\theta)\mu_\theta(x)p(x,y;\theta)dydx < \infty, \end{aligned} \tag{1.17}$$

for all $\theta \in \Theta$. By (1.4),

$$\frac{1}{n}\sum_{i=1}^{n} g(X_{\Delta i}, X_{\Delta(i-1)};\theta') \xrightarrow{P_\theta} Q_\theta(g(\theta')), \tag{1.18}$$

for all $\theta, \theta' \in \Theta$. Since the estimating function $G_n(\theta)$ is a square integrable martingale under P_θ, the asymptotic normality in (1.5) follows without further conditions from the central limit theorem for martingales, see Hall & Heyde (1980). This result for stationary processes goes back to Billingsley (1961). In the martingale case the asymptotic covariance matrix $V(\theta)$ in (1.5) is given by

$$V(\theta) = Q_{\theta_0}\left(g(\theta)g(\theta)^T\right). \tag{1.19}$$

Thus we have the following particular case of Theorem 1.2.2.

Theorem 1.3.2 *Assume Condition 1.2.1 is satisfied with $r = 2$, $\bar\theta = \theta_0$, and $Q = Q_{\theta_0}$, where θ_0 is the true parameter value, and that (1.17) holds for $\theta = \theta_0$. Then a θ_0-consistent G_n–estimator $\hat\theta_n$ exists, and*

$$\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{\mathcal{D}} N_p\left(0, W^{-1}VW^{T^{-1}}\right) \tag{1.20}$$

under P_{θ_0}, where W is given by (1.7) with $\bar\theta = \theta_0$, and $V = V(\theta_0)$ with $V(\theta)$ given by (1.19). If, moreover, the function $g(x,y;\theta)$ is locally dominated integrable with respect to Q_{θ_0} and

$$Q_{\theta_0}(g(\theta)) \neq 0 \;\; \textit{for all} \;\; \theta \neq \theta_0, \tag{1.21}$$

then the estimator $\hat\theta_n$ is the unique G_n–estimator on any bounded subset of Θ containing θ_0 with probability approaching one as $n \to \infty$.

In practice we do not know the value of θ_0, so it is necessary to check that the conditions of Theorem 1.3.2 hold for any value of $\theta_0 \in \operatorname{int}\Theta$.

The asymptotic covariance matrix of the estimator $\hat{\theta}_n$ can be estimated consistently by means of the following theorem.

Theorem 1.3.3 *Under Condition 1.2.1 (2) – (4) (with $r = 2$, $\bar{\theta} = \theta_0$, and $Q = Q_{\theta_0}$),*

$$W_n = \frac{1}{n} \sum_{i=1}^{n} \partial_{\theta^T} g(X_{(i-1)\Delta}, X_{i\Delta}; \hat{\theta}_n) \xrightarrow{P_{\theta_0}} W, \tag{1.22}$$

where $\hat{\theta}_n$ is a θ_0-consistent estimator. The probability that W_n is invertible approaches one as $n \to \infty$. If, moreover, the function $(x, y) \mapsto \|g(x, y; \theta)\|$ is dominated for all $\theta \in N$ by a function which is square integrable with respect to Q_{θ_0}, then

$$V_n = \frac{1}{n} \sum_{i=1}^{n} g(X_{(i-1)\Delta}, X_{i\Delta}; \hat{\theta}_n) g(X_{(i-1)\Delta}, X_{i\Delta}; \hat{\theta}_n)^T \xrightarrow{P_{\theta_0}} V. \tag{1.23}$$

Proof: Let C be a compact subset of N such that $\theta_0 \in \operatorname{int} C$. By Lemma 1.2.3, $\frac{1}{n}\sum_{i=1}^{n} \partial_{\theta^T} g(X_{(i-1)\Delta}, X_{i\Delta}; \theta)$ converges to $Q_{\theta_0}(\partial_{\theta^T} g(\theta))$ in probability uniformly for $\theta \in C$. This implies (1.22) because $\hat{\theta}_n$ converges in probability to θ_0. The result about invertibility follows because W is invertible. Also the uniform convergence in probability for $\theta \in C$ of $\frac{1}{n}\sum_{i=1}^{n} g(X_{(i-1)\Delta}, X_{i\Delta}; \theta)$ $g(X_{(i-1)\Delta}, X_{i\Delta}; \theta)^T$ to $Q_{\theta_0}(g(\theta)g(\theta)^T)$ follows from Lemma 1.2.3.

□

1.3.2 *Likelihood inference*

The diffusion process X is a Markov process, so the likelihood function based on the observations $X_0, X_{t_1}, \cdots, X_{t_n}$, conditional on X_0, is

$$L_n(\theta) = \prod_{i=1}^{n} p(t_i - t_{i-1}, X_{t_{i-1}}, X_{t_i}; \theta), \tag{1.24}$$

where $y \mapsto p(s, x, y; \theta)$ is the transition density and $t_0 = 0$. Under weak regularity conditions the maximum likelihood estimator is efficient, i.e. it has the smallest asymptotic variance among all estimators. The transition density is only rarely explicitly known, but several numerical approaches and accurate approximations make likelihood inference feasible for diffusion models. We shall return to the problem of calculating the likelihood function in Subsection 1.4.

The vector of partial derivatives of the log-likelihood function with respect to the coordinates of θ,

$$U_n(\theta) = \partial_\theta \log L_n(\theta) = \sum_{i=1}^{n} \partial_\theta \log p(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta), \tag{1.25}$$

where $\Delta_i = t_i - t_{i-1}$, is called the *score function* (or score vector). Here it is obviously assumed that the transition density is a differentiable function of θ. The maximum likelihood estimator usually solves the estimating equation $U_n(\theta) = 0$. The score function is a martingale with respect to $\{\mathcal{F}_n\}_{n \in \mathbb{N}}$ under P_θ, which is easily seen provided that the following interchange of differentiation and integration is allowed:

$$\begin{aligned} & \mathrm{E}_\theta \left(\partial_\theta \log p(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta) \middle| \, X_{t_1}, \ldots, X_{t_{i-1}} \right) \\ = & \int_D \frac{\partial_\theta p(\Delta_i, X_{t_{i-1}}, y; \theta)}{p(\Delta_i, X_{t_{i-1}}, y; \theta)} p(\Delta_i, X_{t_{i-1}}, y, \theta) dy \\ = & \ \partial_\theta \int_D p(\Delta_i, X_{t_{i-1}}, y; \theta) dy = 0. \end{aligned}$$

Since the score function is a martingale estimating function, the asymptotic results in the previous subsection applies to the maximum likelihood estimator. Asymptotic results for the maximum likelihood estimator in the low frequency (fixed Δ) asymptotic scenario considered in that subsection were established by Dacunha-Castelle & Florens-Zmirou (1986). Asymptotic results when the observations are made at random time points were obtained by Aït-Sahalia & Mykland (2003).

In the case of likelihood inference, the function $Q_{\theta_0}(g(\theta))$ appearing in the identifiability condition (1.21) is related to the Kullback-Leibler divergence between the models. Specifically, if the following interchange of differentiation and integration is allowed,

$$Q_{\theta_0}(\partial_\theta \log p(x, y, \theta)) = \partial_\theta Q_{\theta_0}(\log p(x, y, \theta)) = -\partial_\theta \bar{K}(\theta, \theta_0),$$

where $\bar{K}(\theta, \theta_0)$ is the average Kullback-Leibler divergence between the transition distributions under P_{θ_0} and P_θ given by

$$\bar{K}(\theta, \theta_0) = \int_D K(\theta, \theta_0; x) \, \mu_{\theta_0}(dx),$$

with

$$K(\theta, \theta_0; x) = \int_D \log[p(x, y; \theta_0)/p(x, y; \theta)] p(x, y; \theta_0) \, dy.$$

Thus the identifiability condition can be written in the form $\partial_\theta \bar{K}(\theta, \theta_0) \neq 0$ for all $\theta \neq \theta_0$. The quantity $\bar{K}(\theta, \theta_0)$ is sometimes referred to as the Kullback-

Leibler divergence between the two Markov chain models for the observed process $\{X_{i\Delta}\}$ under P_{θ_0} and P_θ.

A simple approximation to the likelihood function is obtained by approximating the transition density by a Gaussian density with the correct first and second conditional moments. For a one-dimensional diffusion we get

$$p(\Delta,x,y;\theta) \approx q(\Delta,x,y;\theta) = \frac{1}{\sqrt{2\pi\phi(\Delta,x;\theta)}} \exp\left[-\frac{(y-F(\Delta,x;\theta))^2}{2\phi(\Delta,x;\theta)}\right]$$

where

$$F(\Delta,x;\theta) = \mathrm{E}_\theta(X_\Delta|X_0=x) = \int_\ell^r yp(\Delta,x,y;\theta)dy. \tag{1.26}$$

and

$$\begin{aligned}\phi(\Delta,x;\theta) = & \qquad\qquad (1.27)\\ \mathrm{Var}_\theta(X_\Delta|X_0=x) &= \int_\ell^r [y-F(\Delta,x;\theta)]^2 p(\Delta,x,y;\theta)dy.\end{aligned}$$

In this way we obtain the *quasi-likelihood*

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

and by differentiation with respect to the parameter vector, we obtain the quasi-score function

$$\begin{aligned}\partial_\theta \log QL_n(\theta) = \sum_{i=1}^n \Bigg\{ & \frac{\partial_\theta F(\Delta_i, X_{t_{i-1}};\theta)}{\phi(\Delta_i, X_{t_{i-1}};\theta)}[X_{t_i} - F(\Delta_i, X_{t_{i-1}};\theta)] \qquad (1.28)\\ &+ \frac{\partial_\theta \phi(\Delta_i, X_{t_{i-1}};\theta)}{2\phi(\Delta_i, X_{t_{i-1}};\theta)^2}\left[(X_{t_i} - F(\Delta_i, X_{t_{i-1}};\theta))^2 - \phi(\Delta_i, X_{t_{i-1}};\theta)\right]\Bigg\}.\end{aligned}$$

It is clear from (1.26) and (1.27) that $\{\partial_\theta \log QL_n(\theta)\}_{n\in\mathbb{N}}$ is a martingale with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ under P_θ. This quasi-score function is a particular case of the quadratic martingale estimating functions considered by Bibby & Sørensen (1995) and Bibby & Sørensen (1996). Maximum quasi-likelihood estimation for diffusions was considered by Bollerslev & Wooldridge (1992).

1.3.3 Godambe-Heyde optimality

In this section we present a general way of approximating the score function by means of martingales of a similar form. Suppose we have a collection of real valued functions $h_j(x,y,;\theta)$, $j=1,\ldots,N$, satisfying

$$\int_D h_j(x,y;\theta)p(x,y;\theta)dy = 0 \tag{1.29}$$

for all $x \in D$ and $\theta \in \Theta$. Each of the functions h_j could be used separately to define an estimating function of the form (1.3) with $g = h_j$, but a better approximation to the score function, and hence a more efficient estimator, is obtained by combining them in an optimal way. Therefore we consider estimating functions of the form

$$G_n(\theta) = \sum_{i=1}^{n} a(X_{(i-1)\Delta}, \theta) h(X_{(i-1)\Delta}, X_{i\Delta}; \theta), \qquad (1.30)$$

where $h = (h_1, \ldots, h_N)^T$, and the $p \times N$ weight matrix $a(x, \theta)$ is a function of x such that (1.30) is P_θ-integrable. It follows from (1.29) that $G_n(\theta)$ is a martingale estimating function, i.e., it is a martingale under P_θ for all $\theta \in \Theta$.

The matrix a determines how much weight is given to each of the h_js in the estimation procedure. This weight matrix can be chosen in an optimal way using the theory of optimal estimating functions reviewed in Section 1.11. The *optimal weight matrix*, a^*, gives the estimating function of the form (1.30) that provides the best possible approximation to the score function (1.25) in a mean square sense. Moreover, the optimal $g^*(x, y; \theta) = a^*(x; \theta) h(x, y; \theta)$ is obtained from $\partial_\theta \log p(x, y; \theta)$ by projection in a certain space of square integrable functions, for details see Section 1.11.

The choice of the functions h_j, on the other hand, is an art rather than a science. The ability to tailor these functions to a given model or to particular parameters of interest is a considerable strength of the estimating functions methodology. It is, however, also a source of weakness, since it is not always clear how best to choose the h_js. In the following and in the Subections 1.3.6 and 1.3.7, we shall present ways of choosing these functions that usually work well in practice. Also the theory in Subsection 1.3.4 and Section 1.6 casts interesting light on this problem.

Example 1.3.4 The martingale estimating function (1.28) is of the type (1.30) with $N = 2$ and

$$\begin{aligned} h_1(x, y; \theta) &= y - F(\Delta, x; \theta), \\ h_2(x, y; \theta) &= (y - F(\Delta, x; \theta))^2 - \phi(\Delta, x, \theta), \end{aligned}$$

where F and ϕ are given by (1.26) and (1.27). The weight matrix is

$$\left(\frac{\partial_\theta F(\Delta, x; \theta)}{\phi(\Delta, x; \theta)}, \; \frac{\partial_\theta \phi(\Delta, x; \theta)}{2\phi^2(\Delta, x; \theta)\Delta} \right), \qquad (1.31)$$

which we shall see is an approximation to the optimal weigth matrix. □

In the econometrics literature, a popular way of using functions like $h_j(x, y, ; \theta)$, $j = 1, \ldots, N$, to estimate the parameter θ is the *generalized method of moments* (GMM) of Hansen (1982). In practice, the method is often implemented as follows, see e.g. Campbell, Lo & MacKinlay (1997). Consider

$$F_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} h(X_{(i-1)\Delta}, X_{i\Delta}; \theta).$$

Under weak conditions, cf. Theorem 1.3.3, a consistent estimator of the asymptotic covariance matrix M of $\sqrt{n}F_n(\theta_0)$ is

$$M_n = \frac{1}{n}\sum_{i=1}^{n} h(X_{(i-1)\Delta}, X_{i\Delta}; \tilde{\theta}_n) h(X_{(i-1)\Delta}, X_{i\Delta}; \tilde{\theta}_n)^T,$$

where $\tilde{\theta}_n$ is a θ_0-consistent estimator (for instance obtained by minimizing $F_n(\theta)^T F_n(\theta)$). The GMM-estimator is obtained by minimizing the function

$$H_n(\theta) = F_n(\theta)^T M_n^{-1} F_n(\theta).$$

The corresponding estimating function is obtained by differentiation with respect to θ

$$\partial_\theta H_n(\theta) = D_n(\theta) M_n^{-1} F_n(\theta),$$

where by (1.4)

$$D_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \partial_\theta h(X_{(i-1)\Delta}, X_{i\Delta}; \theta)^T \xrightarrow{P_{\theta_0}} Q_{\theta_0}\left(\partial_\theta h(\theta)^T\right).$$

Hence the estimating function $\partial_\theta H_n(\theta)$ is asymptotically equivalent to an estimating function of the form (1.30) with a weight matrix that does not depend on the data

$$a(x, \theta) = Q_{\theta_0}\left(\partial_\theta h(\theta)^T\right) M^{-1}.$$

We see that the GMM-estimators described here are covered by the theory for martingale estimating functions presented in this section.

We now return to the problem of finding the optimal estimating function $G_n^*(\theta)$, i.e. the estimating functions of the form (1.30) with the optimal weight matrix. We assume that the functions h_j satisfy the following condition.

Condition 1.3.5

(1) The functions $y \mapsto h_j(x, y; \theta)$, $j = 1, \ldots N$, are linearly independent for all $x \in D$ and $\theta \in \Theta$.

(2) The functions $y \mapsto h_j(x, y; \theta)$, $j = 1, \ldots N$, are square integrable with respect to $p(x, y; \theta)$ for all $x \in D$ and $\theta \in \Theta$.

(3) $h(x, y; \theta)$ is differentiable with respect to θ.

(4) The functions $y \mapsto \partial_{\theta_i} h_j(x, y; \theta)$ are integrable with respect to $p(x, y; \theta)$ for all $x \in D$ and $\theta \in \Theta$.

The class of estimating functions considered here is a particular case of the class treated in detail in Example 1.11.4. By (1.182), the optimal choice of the weight matrix a is given by

$$a^*(x;\theta) = B_h(x;\theta)\, V_h(x;\theta)^{-1}, \tag{1.32}$$

where

$$B_h(x;\theta) = \int_D \partial_\theta h(x,y;\theta)^T p(x,y;\theta)dy \tag{1.33}$$

and

$$V_h(x;\theta) = \int_D h(x,y;\theta)h(x,y;\theta)^T p(x,y;\theta)dy. \tag{1.34}$$

The matrix $V_h(x;\theta)$ is invertible because the functions h_j, $j = 1, \ldots N$, are linearly independent. Compared to (1.182), we have omitted a minus here. This can be done because an optimal estimating function multiplied by an invertible $p \times p$-matrix is also an optimal estimating function and yields the same estimator.

The asymptotic variance of an optimal estimator, i.e. a G_n^*–estimator, is simpler than the general expression in (1.20) because in this case the matrices W and V given by (1.7) and (1.19) are identical and given by (1.35). This is a general property of optimal estimating functions as discussed in Section 1.11. The result can easily be verified under the assumption that $a^*(x;\theta)$ is a differentiable function of θ. By (1.29)

$$\int_D \left[\partial_{\theta_i} a^*(x;\theta)\right] h(x,y;\theta)p(x,y;\theta)dy = 0,$$

so that

$$\begin{aligned} W &= \int_{D^2} \partial_{\theta^T}[a^*(x;\theta_0)h(x,y;\theta_0)]Q_{\theta_0}(dx,dy) \\ &= \mu_{\theta_0}(a^*(\theta_0)B_h(\theta_0)^T) = \mu_{\theta_0}\left(B_h(\theta_0)V_h(\theta_0)^{-1}B_h(\theta_0)^T\right), \end{aligned}$$

and by direct calculation

$$\begin{aligned} V &= Q_{\theta_0}(a^*(\theta_0)h(\theta_0)h(\theta_0)^T a^*(\theta_0)^T) \\ &= \mu_{\theta_0}\left(B_h(\theta_0)V_h(\theta_0)^{-1}B_h(\theta_0)^T\right). \end{aligned} \tag{1.35}$$

Thus we have, as a corollary to Theorem 1.3.2, that if $g^*(x,y,\theta) = a^*(x;\theta)h(x,y;\theta)$ satisfies the conditions of Theorem 1.3.2, then the sequence $\hat{\theta}_n$ of consistent G_n^*–estimators has the asymptotic distribution

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N_p\left(0, V^{-1}\right). \tag{1.36}$$

Example 1.3.6 Consider the martingale estimating function of form (1.30) with $N = 2$ and with h_1 and h_2 as in Example 1.3.4, where the diffusion is one-dimensional. The optimal weight matrix has columns given by

$$\begin{aligned}
a_1^*(x;\theta) &= \frac{\partial_\theta \phi(x;\theta)\eta(x;\theta) - \partial_\theta F(x;\theta)\psi(x;\theta)}{\phi(x;\theta)\psi(x;\theta) - \eta(x;\theta)^2} \\
a_2^*(x;\theta) &= \frac{\partial_\theta F(x;\theta)\eta(x;\theta) - \partial_\theta \phi(x;\theta)\phi(x;\theta)}{\phi(x;\theta)\psi(x;\theta) - \eta(x;\theta)^2},
\end{aligned}$$

where

$$\eta(x;\theta) = \mathrm{E}_\theta([X_\Delta - F(x;\theta)]^3 | X_0 = x)$$

and

$$\psi(x;\theta) = \mathrm{E}_\theta([X_\Delta - F(x;\theta)]^4 | X_0 = x) - \phi(x;\theta)^2.$$

For the square-root diffusion (the a.k.a. the CIR-model)

$$dX_t = -\beta(X_t - \alpha)dt + \tau\sqrt{X_t}dW_t, \quad X_0 > 0, \tag{1.37}$$

where $\alpha, \beta, \tau > 0$, the optimal weights can be found explicitly. For this model

$$\begin{aligned}
F(x;\theta) &= xe^{-\beta\Delta} + \alpha(1 - e^{-\beta\Delta}) \\
\phi(x;\theta) &= \frac{\tau^2}{\beta}\left((\tfrac{1}{2}\alpha - x)e^{-2\beta\Delta} - (\alpha - x)e^{-\beta\Delta} + \tfrac{1}{2}\alpha\right) \\
\eta(x;\theta) &= \frac{\tau^4}{2\beta^2}\left(\alpha - 3(\alpha - x)e^{-\beta\Delta} + 3(\alpha - 2x)e^{-2\beta\Delta}\right. \\
&\qquad\qquad\qquad\qquad \left. - (\alpha - 3x)e^{-3\beta\Delta}\right) \\
\psi(x;\theta) &= \frac{3\tau^6}{4\beta^3}\left((\alpha - 4x)e^{-4\beta\Delta} - 4(\alpha - 3x)e^{-3\beta\Delta}\right. \\
&\qquad \left. + 6(\alpha - 2x)e^{-2\beta\Delta} - 4(\alpha - x)e^{-\beta\Delta} + \alpha\right) + 2\phi(x;\theta)^2.
\end{aligned}$$

We give a method to derive these expression in Subsection 1.3.6.

The expressions for a_1^* and a_2^* can for general diffusions be simplified by the approximations

$$\eta(t,x;\theta) \approx 0 \quad \text{and} \quad \psi(t,x;\theta) \approx 2\phi(t,x;\theta)^2, \tag{1.38}$$

which would be exactly true if the transition density were a Gaussian density function. If we insert these Gaussian approximations into the expressions for a_1^* and a_2^*, we obtain the weight functions in (1.28). When Δ is not large this can be justified, because the transition distribution is not far from Gaussian. In Section 1.4 we present a data transformation after which the transition distribution is close to a normal distribution. □

In Subsections 1.3.6 and 1.3.7 we present martingale estimating functions for which the matrices $B_h(x;\theta)$ and $V_h(x;\theta)$ can be found explicitly, but for most models these matrices must be found by simulation, a problem considered in Subsection 1.3.5. In situations where a^* must be determined by a relatively time consuming numerical method, it might be preferable to use the estimating function

$$G_n^\bullet(\theta) = \sum_{i=1}^n a^*(X_{(i-1)\Delta}; \tilde{\theta}_n) h(X_{(i-1)\Delta}, X_{i\Delta}; \theta), \qquad (1.39)$$

where $\tilde{\theta}_n$ is a θ_0-consistent estimator, for instance obtained by some simple choice of the weight matrix a. In this way a^* needs to be calculated only once per observation point, whereas the weight matrix must be recalculated for every call of $G_n^*(\theta)$ given by (1.30) with $a = a^*$. Typically, $G_n^*(\theta)$ will be called many times in a numerical search procedure in order to find the G_n^*-estimator. Under weak regularity conditions, the $G_n^\bullet$-estimator has the same efficiency as the optimal G_n^*-estimator; see e.g. Jacod & Sørensen (2009).

Most martingale estimating functions proposed in the literature are of the form (1.30) with

$$h_j(x, y; \theta) = f_j(y; \theta) - \pi_\Delta^\theta(f_j(\theta))(x), \qquad (1.40)$$

or more specifically,

$$G_n\,(\theta) = \sum_{i=1}^n a(X_{(i-1)\Delta}, \theta) \left[f(X_{i\Delta}; \theta) - \pi_\Delta^\theta(f(\theta))(X_{(i-1)\Delta}) \right]. \qquad (1.41)$$

Here $f = (f_1, \ldots, f_N)^T$ maps $D \times \Theta$ into IR^N, and π_Δ^θ denotes the *transition operator*

$$\pi_s^\theta(f)(x) = \int_D f(y) p(s, x, y; \theta) dy = \mathrm{E}_\theta(f(X_s) \,|\, X_0 = x), \qquad (1.42)$$

applied to each coordinate of f. The polynomial estimating functions given by $f_j(y) = y^j$, $j = 1, \ldots, N$, are an example. For martingale estimating functions of the special form (1.41), the expression for the optimal weight matrix simplifies a bit because

$$B_h(x; \theta)_{ij} = \pi_\Delta^\theta(\partial_{\theta_i} f_j(\theta))(x) - \partial_{\theta_i} \pi_\Delta^\theta(f_j(\theta))(x), \qquad (1.43)$$

$i = 1, \ldots p$, $j = 1, \ldots, N$, and

$$V_h(x; \theta)_{ij} = \pi_\Delta^\theta(f_i(\theta) f_j(\theta))(x) - \pi_\Delta^\theta(f_i(\theta))(x) \pi_\Delta^\theta(f_j(\theta))(x), \qquad (1.44)$$

$i, j = 1, \ldots, N$. If the functions f_j are chosen to be independent of θ, then (1.43) simplifies to

$$B_h(x; \theta)_{ij} = -\partial_{\theta_i} \pi_\Delta^\theta(f_j)(x). \qquad (1.45)$$

A useful *approximation to the optimal weight matrix* can be obtained by applying the expansion of conditional moments given in the following lemma. The expansion is expressed in terms of the *generator* of the diffusion, which is defined by

$$A_\theta f(x) = \sum_{k=1}^{d} b_k(x;\theta)\partial_{x_k} f(x) + \tfrac{1}{2}\sum_{k,\ell=1}^{d} C_{k\ell}(x;\theta)\partial^2_{x_k x_\ell} f(x), \qquad (1.46)$$

where $C = \sigma\sigma^T$. By A^i_θ we mean i-fold application of A_θ, and in particular, $A^0_\theta f = f$. For an ergodic diffusion with invariant measure with Lebesgue density μ_θ, let Φ_θ be the class of real functions f defined on the state space D that are twice continuously differentiable, square intergrable with respect to μ_θ, and satisfy that

$$\int_D (A_\theta f(x))^2 \mu_\theta(x)dx < \infty$$
$$\sum_{i,j=1}^{d} \int_D \partial_{x_i} f(x)\partial_{x_j} f(x) C_{i,j}(x;\theta)\mu_\theta(x)dx < \infty.$$

Lemma 1.3.7 *Suppose that the diffusion process (1.11) is ergodic with invariant measure μ_θ, that f is $2(k+1)$ times continuously differentiable, and that $A^i_\theta f \in \Phi_\theta$, $i = 0, \ldots, k$. Then*

$$\pi^\theta_t(f)(x) = \sum_{i=0}^{k} \frac{t^i}{i!} A^i_\theta f(x) + O(t^{k+1}). \qquad (1.47)$$

Proof: We sketch the proof of (1.47), and consider only $d = 1$ to simplify the exposition. First consider $k = 1$. By Ito's formula

$$\begin{aligned} f(X_t) &= f(X_0) + \int_0^t A_\theta f(X_s)ds + \int_0^t f'(X_s)\sigma(X_s;\theta)dW_s \\ A_\theta f(X_s) &= A_\theta f(X_0) + \int_0^s A^2_\theta f(X_u)du + \int_0^s \partial_x A_\theta f(X_u)\sigma(X_u;\theta)dW_u, \end{aligned}$$

and by inserting the expression for $A_\theta f(X_s)$ given by the second equation into the Lebesgue integral in the first equation, we find that

$$\begin{aligned} f(X_t) = f(X_0) + tA_\theta f(X_0) + \int_0^t\int_0^s A^2_\theta f(X_u)duds \qquad (1.48) \\ + \int_0^t\int_0^s \partial_x A_\theta f(X_u)\sigma(X_u;\theta)dW_u ds + \int_0^t f'(X_s)\sigma(X_s;\theta)dW_s. \end{aligned}$$

Because $A^i_\theta f \in \Phi_\theta$, $i = 0, 1$, the Ito-integrals are proper P_θ-martingales. Hence by taking the conditional expectation given $X_0 = x$, we obtain

$$\pi^\theta_t(f)(x) = f(x) + tA_\theta f(x) + O(t^2).$$

The result for $k = 2$ is obtained by applying Ito's formula to $A^2_\theta f(X_t)$, inserting the result into the first Lebesgue integral in (1.48), and finally taking the conditional expectation given $X_0 = x$. The result for $k \geq 3$ is obtained similarly by applying Ito's formula to $A^i_\theta f(X_t)$, $i = 3, \ldots, k$. □

Note that (1.47) is an expansion result, so the corresponding power series does not necessarily converge. For a fixed k, the sum is a good approximation to the conditional expectation when t is small. The remainder term depends on k and θ. The explicit sufficient conditions in Lemma 1.3.7 for (1.47) to hold for ergodic diffusions was given in Jacobsen (2001). The expansion holds under mild regularity conditions for non-ergodic diffusions too. In a proof similar to that of Lemma 1.3.7, such conditions must essentially ensure, that Ito-integrals are proper martingales and that the remainder term can be controlled.

It is often enough to use the approximation $\pi^\theta_\Delta(f_j)(x) \approx f_j(x) + \Delta A_\theta f_j(x)$. When f does not depend on θ this implies that for $d = 1$

$$B_h(x;\theta) \approx \Delta \left[\partial_\theta b(x;\theta) f'(x) + \tfrac{1}{2}\partial_\theta \sigma^2(x;\theta) f''(x)\right] \tag{1.49}$$

and for $d = 1$ and $N = 1$

$$V_h(x;\theta) \approx \Delta \left[A_\theta(f^2)(x) - 2f(x)A_\theta f(x)\right] = \Delta\, \sigma^2(x;\theta) f'(x)^2. \tag{1.50}$$

We will refer to estimating functions obtained by approximating the optimal weight-matrix a^* in this way as *approximately optimal estimating functions*. Use of this approximation will save computer time and improve the numerical performance of the estimation procedure. The approximation will not affect the consistency of the estimators, and if Δ is not too large, it will just lead to a relatively minor loss of efficiency. The magnitude of this loss of efficiency can be calculated by means of (1.47).

Example 1.3.8 If we simplify the optimal weight matrix found in Example 1.3.6 by the expansion (1.47) and the Gaussian approximation (1.38), we obtain the approximately optimal quadratic martingale estimating function

$$\begin{aligned} G^\circ_n(\theta) = \sum_{i=1}^n \Bigg\{ & \frac{\partial_\theta b(X_{(i-1)\Delta};\theta)}{\sigma^2(X_{(i-1)\Delta};\theta)} [X_{i\Delta} - F(X_{(i-1)\Delta};\theta)] \\ & + \frac{\partial_\theta \sigma^2(X_{(i-1)\Delta};\theta)}{2\sigma^4(X_{(i-1)\Delta};\theta)\Delta} \left[(X_{i\Delta} - F(X_{(i-1)\Delta};\theta))^2 - \phi(X_{(i-1)\Delta};\theta)\right] \Bigg\}. \end{aligned} \tag{1.51}$$

As in Example 1.3.6 the diffusion is assumed to be one-dimensional.

Consider a diffusion with *linear drift*, $b(x;\theta) = -\beta(x - \alpha)$. Diffusion models with linear drift and a given marginal distribution were studied in Bibby, Skovgaard & Sørensen (2005). If $\int \sigma^2(x;\theta)\mu_\theta(x)dx < \infty$, then the Ito-integral in

$$X_t = X_0 - \int_0^t \beta(X_s - \alpha)ds + \int_0^t \sigma(X_s;\theta)dW_s$$

is a proper martingale with mean zero, so the function $f(t) = \mathrm{E}_\theta(X_t \mid X_0 = x)$ satisfies that

$$f(t) = x - \beta \int_0^t f(s)ds + \beta\alpha t$$

or

$$f'(t) = -\beta f(t) + \beta\alpha, \quad f(0) = x.$$

Hence

$$f(t) = xe^{-\beta t} + \alpha(1 - e^{-\beta t})$$

or

$$F(x; \alpha, \beta) = xe^{-\beta\Delta} + \alpha(1 - e^{-\beta\Delta})$$

If only estimates of drift parameters are needed, we can use the linear martingale estimating function of the form (1.30) with $N = 1$ and $h_1(x, y; \theta) = y - F(\Delta, x; \theta)$. If $\sigma(x; \theta) = \tau\kappa(x)$ for $\tau > 0$ and κ a positive function, then the approximately optimal estimating function of this form is

$$G_n^\circ(\alpha, \beta) = \begin{pmatrix} \displaystyle\sum_{i=1}^n \frac{1}{\kappa^2(X_{(i-1)\Delta})} \left[X_{i\Delta} - X_{(i-1)\Delta}e^{-\beta\Delta} - \alpha(1 - e^{-\beta\Delta})\right] \\ \\ \displaystyle\sum_{i=1}^n \frac{X_{(i-1)\Delta}}{\kappa^2(X_{(i-1)\Delta})} \left[X_{i\Delta} - X_{(i-1)\Delta}e^{-\beta\Delta} - \alpha(1 - e^{-\beta\Delta})\right] \end{pmatrix},$$

where multiplicative constants have been omitted. To solve the estimating equation $G_n^\circ(\alpha, \beta) = 0$ we introduce the weights

$$w_i^\kappa = \kappa(X_{(i-1)\Delta})^{-2} / \sum_{j=1}^n \kappa(X_{(j-1)\Delta})^{-2},$$

and define $\bar{X}^\kappa = \sum_{i=1}^n w_i^\kappa X_{i\Delta}$ and $\bar{X}_{-1}^\kappa = \sum_{i=1}^n w_i^\kappa X_{(i-1)\Delta}$. These two quantities are conditional precision weighted sample averages of $X_{i\Delta}$ and $X_{(i-1)\Delta}$, respectively. The equation $G_n^\circ(\alpha, \beta) = 0$ has a unique explicit solution provided that the weighted sample autocorrelation

$$r_n^\kappa = \frac{\sum_{i=1}^n w_i^\kappa (X_{i\Delta} - \bar{X}^\kappa)(X_{(i-1)\Delta} - \bar{X}_{-1}^\kappa)}{\sum_{i=1}^n w_i^\kappa (X_{(i-1)\Delta} - \bar{X}_{-1}^\kappa)^2}$$

is positive. By the law of large numbers (1.4) for ergodic processes, the probability that $r_n^\kappa > 0$ tends to one as n tends to infinity. Specifically, we obtain the explicit estimators

$$\begin{aligned} \hat{\alpha}_n &= \frac{\bar{X}^\kappa - r_n^\kappa \bar{X}_{-1}^\kappa}{1 - r_n^\kappa} \\ \hat{\beta}_n &= -\frac{1}{\Delta} \log\left(r_n^\kappa\right), \end{aligned}$$

see Christensen & Sørensen (2008). A slightly simpler and asymptotically equivalent estimator may be obtained by substituting $\bar{X}^{\kappa}$ for $\bar{X}^{\kappa}_{-1}$ everywhere, in which case α is estimated by the precision weighted sample average $\bar{X}^{\kappa}$. For the square-root process (CIR-model) given by (1.37), where $\kappa(x) = \sqrt{x}$, a simulation study and an investigation of the asymptotic variance of these estimators in Bibby & Sørensen (1995) show that they are not much less efficient than the estimators from the optimal estimating function. See also the simulation study in Overbeck & Rydén (1997), who find that these estimators are surprisingly efficient, a finding that can be explained by the results in Section 1.6.

To obtain an explicit approximately optimal quadratic estimating function, we need an expression for the conditional variance $\phi(x;\theta)$. As we saw in Example 1.3.6, $\phi(x;\theta)$ is explicitly known for the *square-root process (CIR-model)* given by (1.37). For this model the approximately optimal quadratic martingale estimating function is

$$\begin{pmatrix} \displaystyle\sum_{i=1}^{n} \frac{1}{X_{(i-1)\Delta}} \left[X_{i\Delta} - X_{(i-1)\Delta} e^{-\beta\Delta} - \alpha(1 - e^{-\beta\Delta}) \right] \\ \displaystyle\sum_{i=1}^{n} \left[X_{i\Delta} - X_{(i-1)\Delta} e^{-\beta\Delta} - \alpha(1 - e^{-\beta\Delta}) \right] \\ \displaystyle\sum_{i=1}^{n} \frac{1}{X_{(i-1)\Delta}} \Big[\left(X_{i\Delta} - X_{(i-1)\Delta} e^{-\beta\Delta} - \alpha(1 - e^{-\beta\Delta}) \right)^2 \\ \displaystyle\quad - \frac{\tau^2}{\beta} \left\{ \left(\alpha/2 - X_{(i-1)\Delta} \right) e^{-2\beta\Delta} - (\alpha - X_{(i-1)\Delta}) e^{-\beta\Delta} + \alpha/2 \right\} \Big] \end{pmatrix}.$$

This expression is obtained from (1.51) after multiplication by an invertible non-random matrix to obtain a simpler expression. This does not change the estimator. From this estimating function explicit estimators can easily be obtained:

$$\hat{\alpha}_n = \frac{1}{n} \sum_{i=1}^{n} X_{i\Delta} + \frac{e^{-\hat{\beta}_n \Delta}}{n \left(1 - e^{-\hat{\beta}_n \Delta} \right)} (X_{n\Delta} - X_0),$$

essentially the sample mean when n is large, and

$$\begin{aligned} e^{-\hat{\beta}_n \Delta} &= \frac{n \sum_{i=1}^{n} X_{i\Delta}/X_{(i-1)\Delta} - (\sum_{i=1}^{n} X_{i\Delta})(\sum_{i=1}^{n} X_{(i-1)\Delta}^{-1})}{n^2 - (\sum_{i=1}^{n} X_{(i-1)\Delta})(\sum_{i=1}^{n} X_{(i-1)\Delta}^{-1})} \\ \hat{\tau}_n^2 &= \frac{\sum_{i=1}^{n} X_{(i-1)\Delta}^{-1} \left(X_{i\Delta} - X_{(i-1)\Delta} e^{-\hat{\beta}_n \Delta} - \hat{\alpha}_n (1 - e^{-\hat{\beta}_n}) \right)^2}{\sum_{i=1}^{n} X_{(i-1)\Delta}^{-1} \psi(X_{(i-1)\Delta}; \hat{\alpha}_n, \hat{\beta}_n)}, \end{aligned}$$

where

$$\psi(x;\alpha,\beta) = \left((\tfrac{1}{2}\alpha - x) e^{-2\beta\Delta} - (\alpha - x) e^{-\beta\Delta} + \tfrac{1}{2}\alpha \right) / \beta.$$

It is obviously necessary for this solution to the estimating equation to exist that the expression for $e^{-\hat{\beta}_n \Delta}$ is strictly positive, an event that happens with a probability tending to one as $n \to \infty$. Again this follows from the law of large numbers (1.4) for ergodic processes. □

When the optimal weight matrix is approximated by means of (1.47), there is a certain loss of efficiency, which as in the previous example is often quite small; see Bibby & Sørensen (1995) and Section 1.6 on high frequency asymptotics below. Therefore the relatively simple estimating function (1.51) is often a good choice in practice.

It is tempting to go on to approximate $\pi_\Delta^\theta(f_j(\theta))(x)$ in (1.41) by (1.47) in order to obtain an explicit estimating function, but as will be demonstrated in Subsection 1.5.3, this can be a dangerous procedure. In general the conditional expectation in π_Δ^θ should therefore be approximated by simulations. Fortunately, Kessler & Paredes (2002) have established that, provided the simulation is done with sufficient accuracy, this does not cause any bias, only a minor loss of efficiency that can be made arbitrarily small; see Subsection 1.3.5. Moreover, as we shall see in the Subsections 1.3.6 and 1.3.7, $\pi_\Delta^\theta(f_j(\theta))(x)$ can be found explicitly for a quite flexible class of diffusions.

1.3.4 *Small Δ-optimality*

The Godambe-Heyde optimal estimating functions discussed above are optimal within a certain class of estimating functions. In this subsection we present the concept of small Δ-optimality, introduced and studied by Jacobsen (2001) and Jacobsen (2002). A small Δ-optimal estimating function is optimal among all estimating functions satisfying weak regularity conditions, but only for high sampling frequencies, i.e. when the time between observations is small. Thus the advantage of the concept of small Δ-optimality is that the optimality is global, while the advantage of the concept of Godambe-Heyde optimality is that the optimality holds for all sampling frequencies. Fortunately, we do not have to choose between the two, because it turns out that Godambe-Heyde optimal martingale estimating functions of the form (1.30) and (1.40) are small Δ-optimal.

Small Δ-optimality was originally introduced for general estimating functions for multivariate diffusion models, but to simplify the exposition we will concentrate on martingale estimating functions and on one-dimensional diffusions of the form

$$dX_t = b(X_t; \alpha)dt + \sigma(X_t; \beta)dW_t, \tag{1.52}$$

where $\theta = (\alpha, \beta) \in \Theta \subseteq \mathbb{R}^2$. This is the simplest model type for which the essential features of the theory appear. Note that the drift and the diffusion coefficient depend on different parameters. It is assumed that the diffusion is ergodic, that its invariant probability measure has density function μ_θ for all $\theta \in \Theta$, and that $X_0 \sim \mu_\theta$ under P_θ. Thus the diffusion is stationary.

Throughout this subsection, we shall assume that the observation times are equidistant, i.e. $t_i = i\Delta$, $i = 0, 1, \ldots, n$, where Δ is fixed, and that the martingale estimating function (1.12) satisfies the conditions of Theorem 1.3.2, so that we know that (eventually) a G_n-estimator $\hat{\theta}_n$ exists, which is asymptotically normal with covariance matrix $M(g) = W^{-1}VW^{T^{-1}}$, where W is given by (1.7) with $\bar{\theta} = \theta_0$ and $V = V(\theta_0)$ with $V(\theta)$ given by (1.19).

The main idea of small Δ-optimality is to expand the asymptotic covariance matrix in powers of Δ

$$M(g) = \frac{1}{\Delta} v_{-1}(g) + v_0(g) + o(1). \tag{1.53}$$

Small Δ-optimal estimating functions minimize the leading term in (1.53). Jacobsen (2001) obtained (1.53) by Ito-Taylor expansions, see Kloeden & Platen (1999), of the random matrices that appear in the expressions for W and V under regularity conditions that will be given below. A similar expansion was used in Aït-Sahalia & Mykland (2003) and Aït-Sahalia & Mykland (2004).

To formulate the conditions, we define the differential operator $\mathcal{A}_\theta$, $\theta \in \Theta$. Its domain, Γ, is the set of continuous real-valued functions $(s, x, y) \mapsto \varphi(s, x, y)$ of $s \geq 0$ and $(x, y) \in (\ell, r)^2$ that are continuously differentiable with respect to s and twice continuously differentiable with respect to y. The operator $\mathcal{A}_\theta$ is given by

$$\mathcal{A}_\theta \varphi(s, x, y) = \partial_s \varphi(s, x, y) + A_\theta \varphi(s, x, y), \tag{1.54}$$

where A_θ is the generator (1.46), which for fixed s and x is applied to the function $y \mapsto \varphi(s, x, y)$. The operator $\mathcal{A}_\theta$ acting on functions in Γ that do not depend on x is the generator of the space-time process $(t, X_t)_{t \geq 0}$. We also need the probability measure Q_θ^Δ given by (1.16). Note that in this section the dependence on Δ is explicit in the notation.

Condition 1.3.9 *The function φ belongs to Γ and satisfies that*

$$\int_\ell^r \int_\ell^r \varphi^2(s, x, y) Q_{\theta_0}^s(dx, dy) < \infty$$

$$\int_\ell^r \int_\ell^r (\mathcal{A}_{\theta_0} \varphi(s, x, y))^2 Q_{\theta_0}^s(dx, dy) < \infty$$

$$\int_\ell^r \int_\ell^r (\partial_y \varphi(s, x, y))^2 \sigma^2(y; \beta_0) Q_{\theta_0}^s(dx, dy) < \infty$$

for all $s \geq 0$.

As usual $\theta_0 = (\alpha_0, \beta_0)$ denotes the true parameter value. We will say that a function with values in $\mathbb{R}^k$ or $\mathbb{R}^{k\times\ell}$ satisfies Condition 1.3.9 if each component of the functions satisfies this condition.

Suppose φ satisfies Condition 1.3.9. Then by Ito's formula

$$\begin{aligned}\varphi(t,X_0,X_t) = \varphi(0,X_0,X_0) + \int_0^t \mathcal{A}_{\theta_0}\varphi(s,X_0,X_s)ds \qquad (1.55)\\ + \int_0^t \partial_y\varphi(s,X_0,X_s)\sigma(X_s;\beta_0)dW_s\end{aligned}$$

under P_{θ_0}. A significant consequence of Condition 1.3.9 is that the Ito-integral in (1.55) is a true P_{θ_0}-martingale, and thus has expectation zero under P_{θ_0}. If the function $\mathcal{A}_{\theta_0}\varphi$ satisfies Condition 1.3.9, a similar result holds for this functions, which we can insert in the Lebesgue integral in (1.55). By doing so and then taking the conditional expectation given $X_0 = x$ on both sides of (1.55), we obtain

$$\pi_t^{\theta_0}(\varphi)(t,x) = \varphi(0,x,x) + t\mathcal{A}_{\theta_0}\varphi(0,x,x) + O(t^2), \qquad (1.56)$$

where

$$\pi_t^{\theta}(\varphi)(t,x) = \mathrm{E}_\theta\left(\varphi(t,X_0,X_t)|X_0 = x\right).$$

If the functions $\mathcal{A}_{\theta_0}^i\varphi,\ i = 0,\ldots,k$ satisfy Condition 1.3.9, where $\mathcal{A}_{\theta_0}^i$ denotes i-fold application of the operator $\mathcal{A}_{\theta_0}$, we obtain by similar arguments that

$$\pi_t^{\theta_0}(\varphi)(t,x) = \sum_{i=0}^{k} \frac{s^i}{i!}\mathcal{A}_{\theta_0}^i\varphi(0,x,x) + O(t^{k+1}). \qquad (1.57)$$

Note that $\mathcal{A}_\theta^0$ is the identity operator, $\mathcal{A}_\theta^0\varphi = \varphi$. The previously used expansion (1.47) is a particular case of (1.57). In the case where φ does not depend on x (or y) the integrals in Condition 1.3.9 are with respect to the invariant measure μ_{θ_0}. If, moreover, φ does not depend on time s, the conditions do not depend on s.

Theorem 1.3.10 *Suppose that the function $g(\Delta,x,y;\theta_0)$ in (1.12) is such that g, $\partial_{\theta^T}g$, gg^T and $\mathcal{A}_{\theta_0}g$ satisfy Condition 1.3.9. Assume, moreover, that we have the expansion*

$$g(\Delta,x,y;\theta_0) = g(\Delta,x,y;\theta_0) + \Delta\partial_\Delta g(0,x,y;\theta_0) + o_{\theta_0,x,y}(\Delta).$$

If the matrix

$$S = \int_\ell^r B_{\theta_0}(x)\mu_{\theta_0}(x)dx \qquad (1.58)$$

is invertible, where

$$B_\theta(x) = \qquad (1.59)$$
$$\begin{pmatrix} \partial_\alpha b(x;\alpha)\partial_y g_1(0,x,x;\theta) & \frac{1}{2}\partial_\beta\sigma^2(x;\beta)\partial_y^2 g_1(0,x,x;\theta) \\ \partial_\alpha b(x;\alpha)\partial_y g_2(0,x,x;\theta) & \frac{1}{2}\partial_\beta\sigma^2(x;\beta)\partial_y^2 g_2(0,x,x;\theta) \end{pmatrix},$$

then (1.53) holds with

$$v_{-1}(g) \geq \begin{pmatrix} \left(\int_\ell^r (\partial_\alpha b(x;\alpha_0))^2 / \sigma^2(x;\beta_0)\mu_{\theta_0}(x)dx\right)^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{1.60}$$

There is equality in (1.60) if

$$\begin{aligned} \partial_y g_1(0,x,x;\theta_0) &= \partial_\alpha b(x;\alpha_0)/\sigma^2(x;\beta_0), & (1.61) \\ \partial_y g_2(0,x,x;\theta_0) &= 0 & (1.62) \end{aligned}$$

for all $x \in (\ell, r)$. In this case, the second term in (1.53) satisfies that

$$v_0(g)_{22} \geq 2\left(\int_\ell^r \left(\partial_\beta \sigma^2(x;\beta_0)\right)^2 / \sigma^4(x;\beta_0)\mu_{\theta_0}(x)dx\right)^{-1}$$

with equality if

$$\partial_y^2 g_2(0,x,x;\theta_0) = \partial_\beta \sigma^2(x;\beta_0)/\sigma^4(x;\beta_0), \tag{1.63}$$

for all $x \in (\ell, r)$.

By $\partial_y g_i(0,x,x;\theta)$ we mean $\partial_y g_i(0,y,x;\theta)$ evaluated at $y = x$, and similarly for second order partial derivatives. Thus the conditions for small Δ-optimality are (1.61), (1.62) and (1.63). For a proof of Theorem 1.3.10, see Jacobsen (2001). The condition (1.62) ensures that all entries of $v_{-1}(g)$ involving the diffusion coefficient parameter, β, are zero. Since $v_{-1}(g)$ is the Δ^{-1}-order term in the expansion (1.53) of the asymptotic covariance matrix, this dramatically decreases the asymptotic variance of the estimator of β when Δ is small. We refer to the condition (1.62) as *Jacobsen's condition*.

The reader is reminded of the trivial fact that for any non-singular 2×2 matrix, M_n, the estimating functions $M_n G_n(\theta)$ and $G_n(\theta)$ give exactly the same estimator. We call them *versions* of the same estimating function. The matrix M_n may depend on Δ_n. Therefore a given version of an estimating function needs not satisfy (1.61) – (1.63). The point is that a version must exist which satisfies these conditions.

Example 1.3.11 Consider a quadratic martingale estimating function of the form

$$g(\Delta, y, x; \theta) = \begin{pmatrix} a_1(x,\Delta;\theta)[y - F(\Delta,x;\theta)] \\ a_2(x,\Delta;\theta)\left[(y - F(\Delta,x;\theta))^2 - \phi(\Delta,x;\theta)\right] \end{pmatrix}, \tag{1.64}$$

where F and ϕ are given by (1.26) and (1.27). By (1.47), $F(\Delta,x;\theta) = x + O(\Delta)$ and $\phi(\Delta,x;\theta) = O(\Delta)$, so

$$g(0,y,x;\theta) = \begin{pmatrix} a_1(x,0;\theta)(y-x) \\ a_2(x,0;\theta)(y-x)^2 \end{pmatrix}. \tag{1.65}$$

Since $\partial_y g_2(0, y, x; \theta) = 2a_2(x, \Delta; \theta)(y - x)$, the Jacobsen condition (1.62) is satisfied for all quadratic martingale estimating functions. Using again (1.47), it is not difficult to see that the two other conditions (1.61) and (1.63) are satisfied in three particular cases: the optimal estimating function given in Example 1.3.6 and the approximations (1.28) and (1.51). □

The following theorem gives conditions ensuring, for given functions $f_1, \ldots,$ f_N, that a small Δ-optimal estimating function of the form (1.30) and (1.40) exists. This not always the case. We assume that the functions $f_1(\cdot\,; \theta), \ldots,$ $f_N(\cdot\,; \theta)$ are of full affine rank for all θ, i.e., for any $\theta \in \Theta$, the identity

$$\sum_{j=1}^{N} a_j^\theta f_j(x; \theta) + a_0^\theta = 0, \quad x \in (\ell, r),$$

for constants a_j^θ, implies that $a_0^\theta = a_1^\theta = \cdots = a_N^\theta = 0$.

Theorem 1.3.12 *Suppose that $N \geq 2$, that the functions f_j are twice continuously differentiable and satisfies that the matrix*

$$D(x) = \begin{pmatrix} \partial_x f_1(x; \theta) & \partial_x^2 f_1(x; \theta) \\ \partial_x f_2(x; \theta) & \partial_x^2 f_2(x; \theta) \end{pmatrix} \tag{1.66}$$

is invertible for μ_θ-almost all x. Moreover, assume that the coefficients b and σ are continuously differentiable with respect to the parameter. Then a specification of the weight matrix $a(x; \theta)$ exists such that the estimating function (1.30) satisfies the conditions (1.62), (1.61) and (1.63). When $N = 2$, these conditions are satisfy for

$$a(x; \theta) = \begin{pmatrix} \partial_\alpha b(x; \alpha)/v(x; \beta) & c(x; \theta) \\ 0 & \partial_\beta v(x; \beta)/v(x; \beta)^2 \end{pmatrix} D(x)^{-1} \tag{1.67}$$

for any function $c(x; \theta)$.

For a proof of Theorem 1.3.12, see Jacobsen (2002). In Section 1.6, we shall see that the Godambe-Heyde optimal choice (1.32) of the weight-matrix in (1.30) gives an estimating function which has a version that satisfies the conditions for small Δ-optimality, (1.61) – (1.63).

We have focused on one-dimensional diffusions to simplify the exposition. The situation becomes more complicated for multi-dimensional diffusions, as we shall now briefly describe. Details can be found in Jacobsen (2002). For a d-dimensional diffusion, $b(x; \alpha)$ is d-dimensional and $v(x; \beta) = \sigma(x; \beta)\sigma(x; \beta)^T$ is a $d \times d$-matrix. The Jacobsen condition is unchanged (except that $\partial_y g_2(0, x, x; \theta_0)$ is now a d-dimensional vector). The other two conditions for small Δ-optimality are

$$\partial_y g_1(0, x, x; \theta_0) = \partial_\alpha b(x; \alpha_0)^T v(x; \beta_0)^{-1}$$

and

$$\text{vec}\left(\partial_y^2 g_2(0,x,x;\theta_0)\right) = \text{vec}\left(\partial_\beta v(x;\beta_0)\right)\left(v^{\otimes 2}(x;\beta_0)\right)^{-1}.$$

In the latter equation, vec(M) denotes, for a $d \times d$ matrix M, the d^2-dimensional row vector consisting of the rows of M placed one after the other, and $M^{\otimes 2}$ is the $d^2 \times d^2$-matrix with $(i',j'),(ij)$th entry equal to $M_{i'i}M_{j'j}$. Thus if $M = \partial_\beta v(x;\beta)$ and $M^\bullet = (v^{\otimes 2}(x;\beta))^{-1}$, then the (i,j)th coordinate of vec$(M)\, M^\bullet$ is $\sum_{i'j'} M_{i'j'} M^\bullet_{(i'j'),(i,j)}$.

For a d-dimensional diffusion process, the conditions analogous to those in Theorem 1.3.12 ensuring the existence of a small Δ-optimal estimating function of the form (1.30) is that $N \geq d(d+3)/2$, and that the $N \times (d+d^2)$-matrix

$$\begin{pmatrix} \partial_{x^T} f(x;\theta) & \partial^2_{x^T} f(x;\theta) \end{pmatrix}$$

has full rank $d(d+3)/2$.

1.3.5 Simulated martingale estimating functions

The conditional moments that appear in the martingale estimating functions can for most diffusion models not be calculated explicitly. For a versatile class of one-dimensional diffusions, optimal martingale estimating functions can be found explicitly; see Subsections 1.3.6 and 1.3.7. Estimation and inference is dramatically simplified by using a model for which an explicit optimal martingale estimating function is available. However, if for some reason a diffusion from this class is not a suitable model, the conditional moments must be determined by simulation.

The conditional moment $\pi_\theta^\Delta f(x) = E_\theta(f(X_\Delta)\,|\,X_0 = x)$ can be found straightforwardly. Simply fix θ and simulate numerically M independent trajectories $X^{(i)}$, $i = 1,\ldots,M$ of $\{X_t : t \in [0,\Delta]\}$ with $X_0 = x$. Of course, a trajectory cannot be simulated continuously for all $t \in [0,\Delta]$. In practice values of $\{X_{j\delta} : j = 0,\ldots,K\}$ are simulated, where K is a large integer and $\delta = \Delta/K$. By the law of large numbers,

$$\pi_\theta^\Delta f(x) \doteq \frac{1}{M}\sum_{i=1}^{M} f(X_\Delta^{(i)}).$$

The variance of the error can be estimated in the traditional way, and by the cental limit theorem, the error is approximately normal distributed. This simple approach can be improved by applying variance reduction methods, for instance methods that take advantage of the fact that $\pi_\theta^\Delta f(x)$ can be approximated by (1.47). Methods for numerical simulation of diffusion models can be found in Chapter xxxx and Kloeden & Platen (1999).

The approach just described is sufficient when calculating the conditional expectation appearing in (1.40). Note, however, that when using a search algorithm to find a solution to the estimating equation, it is important to use the same random numbers (the same seed) when calculating the estimating functions for different values of the parameter θ. More care is needed if the optimal weight functions are calculated numerically. The problem is that the optimal weight matrix typically contains derivatives with respect to θ of functions that must be determined numerically, see e.g. Example 1.3.6. Pedersen (1994) proposed a procedure for determining $\partial_\theta \pi_\theta^\Delta f(x;\theta)$ by simulations based on results in Friedman (1975). However, it is often preferable to use an approximation to the optimal weight matrix obtained by using (1.47), possibly supplemented by Gaussian approximations, as explained in Subsection 1.3.3. This is not only much simpler, but also avoids potentially serious problems of numerical instability, and by results in Section 1.6 the loss of efficiency is often very small. The approach outlined here, where martingale estimating functions are approximated by simulation, is closely related to the simulated method of moments, see Duffie & Singleton (1993) and Clement (1997).

One might be worried that when approximating a martingale estimating function by simulation of conditional moments, the resulting estimator might have considerably smaller efficiency or even be inconsistent. The asymptotic properties of the estimators obtained when the conditional moments are approximated by simulation were investigated by Kessler & Paredes (2002), who found that if the simulations are done with sufficient care, there is no need to worry. However, their results also show that care is needed: if the discretization used in the simulation method is too crude, the estimator behaves badly. Kessler & Paredes (2002) considered martingale estimating functions of the general form

$$G_n(\theta) = \sum_{i=1}^{n} \left[f(X_{i\Delta}, X_{(i-1)\Delta}; \theta) - F(X_{(i-1)\Delta}; \theta) \right], \tag{1.68}$$

where f is a p-dimensional function, and

$$F(x;\theta) = E_\theta(f(X_\Delta, x; \theta))|X_0 = x).$$

As previously, X is the unique solution of the stochastic differential equation (1.11). For simplicity X is assumed to be one-dimensional, but Kessler & Paredes (2002) point out that similar results hold for multivariate diffusions. Below the dependence of X on the initial value $X_0 = x$ and θ is, when needed, emphasized in the notation by writing $X(x,\theta)$.

Let $Y(\delta,\theta,x)$ be an approximation to the solution $X(\theta,x)$, which is calculated at discrete time points with a step size δ that is much smaller than Δ, and which satisfies that $Y_0(\delta,\theta,x) = x$. A simple example is the Euler scheme

$$Y_{i\delta} = Y_{(i-1)\delta} + b(Y_{(i-1)\delta};\theta)\delta + \sigma(Y_{(i-1)\delta};\theta)Z_i, \qquad Y_0 = x, \tag{1.69}$$

where the Z_is are independent and $Z_i \sim N(0,\delta)$.

If the conditional expectation $F(x;\theta)$ is approximated by the simple method described above, we obtain the following approximation to the estimating function (1.68)

$$G_n^{M,\delta}(\theta) = \sum_{i=1}^{n} \left[f(X_{i\Delta}, X_{(i-1)\Delta};\theta) - \frac{1}{M}\sum_{j=1}^{M} f(Y_\Delta^{(j)}(\delta,\theta,X_{(i-1)\Delta}), X_{(i-1)\Delta};\theta) \right], \tag{1.70}$$

where $Y^{(j)}(\delta,\theta,x)$, $j = 1,\ldots,M$, are independent copies of $Y(\delta,\theta,x)$.

Kessler & Paredes (2002) assume that the approximation scheme $Y(\delta,\theta,x)$ is of weak order $\beta > 0$ in the sense that

$$|E_\theta(g(X_\Delta(x,\theta),x;\theta)) - E(g(Y_\Delta(\delta,\theta,x),x;\theta))| \le R(x;\theta)\delta^\beta \tag{1.71}$$

for all $\theta \in \Theta$, for all x in the state space of X, and for δ sufficiently small. Here $R(x;\theta)$ is of polynomial growth in x uniformly for θ in compact sets, i.e., for any compact subset $K \subseteq \Theta$, there exist constants $C_1, C_2 > 0$ such that $\sup_{\theta\in K} |R(x;\theta)| \le C_1(1+|x|^{C_2})$ for all x in the state space of the diffusion. The inequality (1.71) is assumed to hold for any function $g(y,x;\theta)$ which is $2(\beta+1)$ times differentiable with respect to x, and satisfies that g and its partial derivatives with respect to x up to order $2(\beta+1)$ are of polynomial growth in x uniformly for θ in compact sets. This definition of weak order is stronger than the definition in Kloeden & Platen (1999) in that control of the polynomial order with respect to the initial value x is added, but Kessler & Paredes (2002) point out that theorems in Kloeden & Platen (1999) that give the order of approximation schemes can be modified in a tedious, but straightforward, way to ensure that the schemes satisfy the stronger condition (1.71). In particular, the Euler scheme (1.69) is of weak order one if the coefficients of the stochastic differential equation (1.11) are sufficiently smooth.

Under a number of further regularity conditions, Kessler & Paredes (2002) showed the following results about a $G_n^{M,\delta}$-estimator, $\hat\theta_n^{M,\delta}$, with $G_n^{M,\delta}$ given by (1.70). We shall not go into these rather technical conditions. Not surprisingly, they include conditions that ensure the eventual existence of a consistent and asymptotically normal G_n-estimator, cf. Theorem 1.3.2. If δ goes to zero sufficiently fast that $\sqrt{n}\delta^\beta \to 0$ as $n \to \infty$, then

$$\sqrt{n}\left(\hat\theta_n^{M,\delta} - \theta_0\right) \xrightarrow{\mathcal{D}} N\left(0,(1+M^{-1})\Sigma\right),$$

where Σ denotes the asymptotic covariance matrix of a G_n-estimator, see Theorem 1.3.2. Thus for δ sufficiently small and M sufficiently large, it does not matter much that the conditional moment $F(x;\theta)$ has been determined by simulation in (1.70). Moreover, we can control the loss of efficiency by our choice of M. However, when $0 < \lim_{n\to\infty} \sqrt{n}\delta^\beta < \infty$,

$$\sqrt{n}\left(\hat{\theta}_n^{M,\delta} - \theta_0\right) \xrightarrow{\mathcal{D}} N\left(m(\theta_0), (1 + M^{-1})\Sigma\right),$$

and when $\sqrt{n}\delta^\beta \to \infty$,

$$\delta^{-\beta}\left(\hat{\theta}_n^{N,\delta} - \theta_0\right) \to m(\theta_0)$$

in probability. Here the p-dimensional vector $m(\theta_0)$ depends on f and is generally different from zero. Thus it is essential that a sufficiently small value of δ is used. The discretization problems caused by the choice of δ can be avoided by using the exact simulation methods introduced by Beskos & Roberts (2005) and Beskos, Papaspiliopoulos & Roberts (2006), see Chapter xxxx.

1.3.6 Explicit martingale estimating functions

In this section we consider one-dimensional diffusion models for which estimation is particularly easy because an explicit martingale estimating function exists.

Kessler & Sørensen (1999) proposed estimating functions of the form (1.41) where the functions f_j, $i = 1, \ldots, N$ are *eigenfunctions* for the generator (1.46), i.e.

$$A_\theta f_j(x;\theta) = -\lambda_j(\theta) f_j(x;\theta),$$

where the real number $\lambda_j(\theta) \geq 0$ is called the *eigenvalue* corresponding to $f_j(x;\theta)$. Under weak regularity conditions, f_j is also an eigenfunction for the transition operator π_t^θ defined by (1.42), i.e.

$$\pi_t^\theta(f_j(\theta))(x) = e^{-\lambda_j(\theta)t} f_j(x;\theta).$$

for all $t > 0$. Thus the function h_j given by (1.40) is explicit.

The following result holds for ergodic diffusions. The density of the stationary distribution is, as usual, denoted by μ_θ.

Theorem 1.3.13 *Let $\phi(x;\theta)$ be an eigenfunction for the generator (1.46) with eigenvalue $\lambda(\theta)$, and suppose that*

$$\int_\ell^r [\partial_x \phi(x;\theta)\sigma(x;\theta)]^2 \mu_\theta(dx) < \infty \tag{1.72}$$

for all $t > 0$. Then

$$\pi_t^\theta(\phi(\theta))(x) = e^{-\lambda(\theta)t}\phi(x;\theta). \tag{1.73}$$

for all $t > 0$.

Proof: Define $Y_t = e^{\lambda t}\phi(X_t)$. We suppress θ in the notation. By Ito's formula

$$\begin{aligned} Y_t &= Y_0 + \int_0^t e^{\lambda s}[A\phi(X_s) + \lambda\phi(X_s)]ds + \int_0^t e^{\lambda s}\phi'(X_s)\sigma(X_s)dW_s \\ &= Y_0 + \int_0^t e^{\lambda s}\phi'(X_s)\sigma(X_s)dW_s, \end{aligned}$$

so by (1.72), Y is a true martingale, which implies (1.73). □

Note that if $\sigma(x;\theta)$ and $\partial_x\phi(x;\theta)$ are bounded functions of $x \in (\ell, r)$, then (1.72) holds. If ϕ is a polynomial of order k and $\sigma(x) \leq C(1 + x^m)$, then (1.72) holds if the $2(k + m - 1)$'th moment of the invariant distribution μ_θ is finite.

Example 1.3.14 For the square-root model (CIR-model) defined by (1.37) with $\alpha > 0$, $\beta > 0$, and $\tau > 0$, the eigenfunctions are $\phi_i(x) = L_i^{(\nu)}(2\beta x\tau^{-2})$ with $\nu = 2\alpha\beta\tau^{-2} - 1$, where $L_i^{(\nu)}$ is the ith order Laguerre polynomial

$$L_i^{(\nu)}(x) = \sum_{m=0}^{i}(-1)^m \left(\begin{array}{c} i+\nu \\ i-m \end{array} \right) \frac{x^m}{m!},$$

and the eigenvalues are $\{i\beta \,:\, i = 0, 1, \cdots\}$. It is easily seen by direct calculation that $L_i^{(\nu)}$ solves the differential equation

$$\tau x f''(x) - \beta(x - \alpha)f'(x) + i\beta f(x) = 0.$$

By Theorem 1.3.13, (1.73) holds, so we can calculate all conditional polynomial moments, of which the first four were given in Example 1.3.6. Thus all polynomial martingale estimating functions are explicit for the square-root model.

□

Example 1.3.15 The diffusion given as the solution of

$$dX_t = -\theta\tan(X_t)dt + dW_t, \tag{1.74}$$

is an ergodic diffusion on the interval $(-\pi/2, \pi/2)$ provided that $\theta \geq 1/2$, which implies that Condition 1.3.1 is satisfied. This process was introduced by Kessler & Sørensen (1999), who called it an Ornstein-Uhlenbeck process on $(-\pi/2, \pi/2)$ because $\tan x \sim x$ near zero. The generalization to other finite intervals is obvious. The invariant measure has a density proportional to $\cos(x)^{2\theta}$.

The eigenfunctions are

$$\phi_i(x;\theta) = C_i^\theta(\sin(x)), \quad i = 1, 2, \ldots,$$

where C_i^θ is the Gegenbauer polynomial of order i, and the eigenvalues are $i(\theta + i/2)$, $i = 1, 2, \ldots$. This follows because the Gegenbauer polynomial C_i^θ

solves the differential equation

$$f''(y) + \frac{(2\theta+1)y}{y^2-1} f'(y) - \frac{i(2\theta+i)}{y^2-1} f(y) = 0,$$

so that $\phi_i(x;\theta)$ solves the equation

$$\frac{1}{2}\phi_i''(x;\theta) - \theta \tan(x)\phi_i'(x;\theta) = -i(\theta + i/2)\phi_i(x;\theta).$$

Hence ϕ_i is an eigenfunction for the generator of the model with eigenvalue $i(\theta + i/2)$. From equation 8.934-2 in Gradshteyn & Ryzhik (1965) it follows that

$$\phi_i(x;\theta) = \sum_{m=0}^{i} \left(\begin{array}{c} \theta - 1 + m \\ m \end{array} \right) \left(\begin{array}{c} \theta - 1 + i - m \\ i - m \end{array} \right) \cos[(2m - i)(\pi/2 - x)].$$

Condition (1.72) in Theorem 1.3.13 is obviously satisfied because the state space is bounded, so (1.73) holds.

The first non-trivial eigenfunction is $\sin(x)$ (a constant is omitted) with eigenvalue $\theta + 1/2$. From the martingale estimating function

$$\check{G}_n(\theta) = \sum_{i=1}^{n} \sin(X_{(i-1)\Delta})[\sin(X_{i\Delta})) - e^{-(\theta+1/2)\Delta} \sin(X_{(i-1)\Delta}))], \quad (1.75)$$

we obtain the simple estimator for θ

$$\check{\theta}_n = -\Delta^{-1} \log \left(\frac{\sum_{i=1}^n \sin(X_{(i-1)\Delta}) \sin(X_{i\Delta})}{\sum_{i=1}^n \sin^2(X_{(i-1)\Delta})} \right) - 1/2, \quad (1.76)$$

which is defined when the numerator is positive.

An asymmetric generalization of (1.74) was proposed in Larsen & Sørensen (2007) as a model of the logarithm of an exchange rate in a target zone. The diffusion solves the equation

$$dX_t = -\rho \frac{\sin\left(\frac{1}{2}\pi(X_t - m)/z\right) - \varphi}{\cos\left(\frac{1}{2}\pi(X_t - m)/z\right)} dt + \sigma dW_t,$$

where $\rho > 0$, $\varphi \in (-1, 1)$, $\sigma > 0$ $z > 0$, $m \in \mathbb{R}$. The process (1.74) is obtained is when $\varphi = 0$, $m = 0$, $z = \pi/2$, and $\sigma = 1$. The state space is $(m - z, m + z)$, and the process is ergodic if $\rho \geq \frac{1}{2}\sigma^2$ and $-1 + \sigma^2/(2\rho) \leq \varphi \leq 1 - \sigma^2/(2\rho)$. The eigenfunctions are

$$\phi_i(x; \rho, \varphi, \sigma, m, z) = P_i^{(\rho(1-\varphi)\sigma^{-2} - \frac{1}{2},\, \rho(1+\varphi)\sigma^{-2} - \frac{1}{2})} \left(\sin(\tfrac{1}{2}\pi x/z - m) \right),$$

with eigenvalues $\lambda_i(\rho, \varphi, \sigma) = i\left(\rho + \frac{1}{2}n\sigma^2\right)$, $i = 1, 2, \ldots$. Here $P_i^{(a,b)}(x)$ denotes the Jacobi polynomial of order i.

□

For most diffusion models where explicit expressions for eigenfunctions can be found, including the examples above, the eigenfunctions are of the form

$$\phi_i(y;\theta) = \sum_{j=0}^{i} a_{i,j}(\theta)\,\kappa(y)^j \tag{1.77}$$

where κ is a real function defined on the state space and is independent of θ. For martingale estimating functions based on eigenfunctions of this form, the optimal weight matrix (1.32) can be found explicitly too.

Theorem 1.3.16 *Suppose $2N$ eigenfunctions are of the form (1.77) for $i = 1, \ldots, 2N$, where the coefficients $a_{i,j}(\theta)$ are differentiable with respect to θ. If a martingale estimating functions is defined by (1.40) using the first N eigenfunctions, then*

$$B_h(x,\theta)_{ij} = \sum_{k=0}^{j} \Big(\partial_{\theta_i} a_{j,k}(\theta)\nu_k(x;\theta) - \partial_{\theta_i}[e^{-\lambda_j(\theta)\Delta}\phi_j(x;\theta)] \Big) \tag{1.78}$$

and

$$V_h(x,\theta)_{i,j} = \tag{1.79}$$

$$\sum_{r=0}^{i}\sum_{s=0}^{j} \Big(a_{i,r}(\theta)a_{j,s}(\theta)\nu_{r+s}(x;\theta) - e^{-[\lambda_i(\theta)+\lambda_j(\theta)]\Delta}\,\phi_i(x;\theta)\phi_j(x;\theta) \Big),$$

where $\nu_i(x;\theta) = \pi^\theta_\Delta(\kappa^i)(x)$, $i = 1, \ldots, 2N$, solve the following triangular system of linear equations

$$e^{-\lambda_i(\theta)\Delta}\phi_i(x;\theta) = \sum_{j=0}^{i} a_{i,j}(\theta)\nu_j(x;\theta) \quad i = 1, \ldots, 2N, \tag{1.80}$$

with $\nu_0(x;\theta) = 1$.

Proof: The expressions for B_h and V_h follow from (1.43) and (1.44) when the eigenfunctions are of the form (1.77), and (1.80) follows by applying the transition operator π^θ_Δ to both sides of (1.77).

□

Example 1.3.17 Consider again the diffusion (1.74) in Example 1.3.15. We will find the optimal martingale estimating function based on the first non-trivial eigenfunction, $\sin(x)$ (where we have omitted a non-essential multiplicative function of θ) with eigenvalue $\theta + 1/2$. It follows from (1.45) that

$$B_h(x;\theta) = \Delta e^{-(\theta+1/2)\Delta}\sin(x)$$

because $\sin(x)$ does not depend on θ. To find V_h we need Theorem 1.3.16.

The second non-trivial eigenfunction is $2(\theta+1)\sin^2(x)-1$ with eigenvalue $2(\theta+1)$, so

$$\nu_2(x;\theta) = e^{-2(\theta+1)\Delta}[\sin^2(x) - \frac{1}{2}(\theta+1)^{-1}] + \frac{1}{2}(\theta+1)^{-1}.$$

Hence the optimal estimating function is

$$G_n^*(\theta) = \sum_{i=1}^n \frac{\sin(X_{(i-1)\Delta})[\sin(X_{i\Delta}) - e^{-(\theta+\frac{1}{2})\Delta}\sin(X_{(i-1)\Delta})]}{\frac{1}{2}(e^{2(\theta+1)\Delta}-1)/(\theta+1) - (e^{\Delta}-1)\sin^2(X_{(i-1)\Delta})}$$

where a constant has been omitted. When Δ is small, it is a good idea to multiply $G_n^*(\theta)$ by Δ because the denominator is then of order Δ.

Note that when Δ is sufficiently small, we can expand the exponential function in the numerator to obtain (after multiplication by Δ) the approximately optimal estimating function

$$\tilde{G}_n(\theta) = \sum_{i=1}^n \frac{\sin(X_{(i-1)\Delta})[\sin(X_{i\Delta}) - e^{-(\theta+\frac{1}{2})\Delta}\sin(X_{(i-1)\Delta})]}{\cos^2(X_{(i-1)\Delta})},$$

which has the explicit solution

$$\tilde{\theta}_n = -\Delta^{-1}\log\left(\frac{\sum_{i=1}^n \tan(X_{(i-1)\Delta})\sin(X_{i\Delta}))/\cos(X_{(i-1)\Delta})}{\sum_{i=1}^n \tan^2(X_{(i-1)\Delta})}\right) - \frac{1}{2}.$$

The explicit estimator $\tilde{\theta}$ can, for instance, be used as a starting value when finding the optimal estimator by solving $G_n^*(\theta) = 0$ numerically. Note, however, that for $\tilde{G}_n$ the square integrability (1.17) under Q_{θ_0} required in Theorem 1.3.2 (to ensure the central limit theorem) is only satisfied when $\theta_0 > 1.5$. This problem can be avoided by replacing $\cos^2(X_{(i-1)\Delta})$ in the numerator by 1, to which it is close when the process is not near the boundaries. In that way we arrive at the simple estimating function (1.75), which is thus approximately optimal too.

□

1.3.7 Pearson diffusions

A widely applicable class of diffusion models for which explicit polynomial eigenfunctions are available is the class of Pearson diffusions, see Wong (1964) and Forman & Sørensen (2008). A Pearson diffusion is a stationary solution to a stochastic differential equation of the form

$$dX_t = -\beta(X_t-\alpha)dt + \sqrt{2\beta(aX_t^2+bX_t+c)}dW_t, \tag{1.81}$$

where $\beta > 0$, and a, b and c are such that the square root is well defined when X_t is in the state space. The parameter $\beta > 0$ is a scaling of time that determines how fast the diffusion moves. The parameters α, a, b, and c determine the state space of the diffusion as well as the shape of the invariant distribution. In particular, α is the expectation of the invariant distribution. We define $\theta = (\alpha, \beta, a, b, c)$.

In the context of martingale estimating functions, an important property of the Pearson diffusions is that the generator (1.46) maps polynomials into polynomials. It is therefore easy to find eigenfunctions among the polynomials

$$p_n(x) = \sum_{j=0}^{n} p_{n,j} x^j .$$

Specifically, the polynomial $p_n(x)$ is an eigenfunction if an eigenvalue $\lambda_n > 0$ exist satisfying that

$$\beta(ax^2 + bx + c)p_n''(x) - \beta(x - \alpha)p_n'(x) = -\lambda_n p_n(x),$$

or

$$\sum_{j=0}^{n} \{\lambda_n - a_j\} p_{n,j} x^j + \sum_{j=0}^{n-1} b_{j+1} p_{n,j+1} x^j + \sum_{j=0}^{n-2} c_{j+2} p_{n,j+2} x^j = 0.$$

where $a_j = j\{1 - (j-1)a\}\beta$, $b_j = j\{\alpha + (j-1)b\}\beta$, and $c_j = j(j-1)c\beta$ for $j = 0, 1, 2, \ldots$. Without loss of generality, we assume $p_{n,n} = 1$. Thus by equating the coefficients, we find that the eigenvalue is given by

$$\lambda_n = a_n = n\{1 - (n-1)a\}\beta. \tag{1.82}$$

If we define $p_{n,n+1} = 0$, then the coefficients $\{p_{n,j}\}_{j=0,\ldots,n-1}$ solve the linear system of equations

$$(a_j - a_n)p_{n,j} = b_{j+1}p_{n,j+1} + c_{j+2}p_{n,j+2}. \tag{1.83}$$

Equation (1.83) is equivalent to a simple recursive formula if $a_n - a_j \neq 0$ for all $j = 0, 1, \ldots, n-1$. Note that $a_n - a_j = 0$ if and only if there exists an integer $n - 1 \leq m < 2n - 1$ such that $a = m^{-1}$ and $j = m - n + 1$. In particular, $a_n - a_j = 0$ cannot occur if $a < (2n-1)^{-1}$. It is important to notice that λ_n is positive if and only if $a < (n-1)^{-1}$. We shall see below that this is exactly the condition ensuring that $p_n(x)$ is integrable with respect to the invariant distribution. If the stronger condition $a < (2n-1)^{-1}$ is satisfied, the first n eigenfunctions belong to the space of functions that are square integrable with respect to the invariant distribution, and they are orthogonal with respect to the usual inner product in this space. The space of functions that are square integrable with respect to the invariant distribution (or a subset of this space) is often taken as the domain of the generator. Obviously, the eigenfunction $p_n(x)$ satisfies the condition (1.72) if $p_n(x)$ is square integrable with respect to the invariant distribution, i.e. if $a < (2n-1)^{-1}$. By Theorem 1.3.13 this implies that the transition operator satisfies (1.73), so that $p_n(x)$ can be used to construct *explicit optimal martingale estimating functions* as explained in Subsection 1.3.6. For Pearson diffusions with $a \leq 0$, $a < (2n-1)^{-1}$ is automatically satisfied, so there are infinitely many polynomial eigenfunctions. In these cases the eigenfunctions are well-known families of orthogonal polynomials. When $a > 0$, there are only finitely many square integrable polynomial eigenfunctions. In these cases more complicated eigenfunctions defined in terms of special functions exist too, see Wong (1964). It is of some historical interest that Hildebrandt (1931) derived the polynomials above from the viewpoint of Gram-Charlier expansions associated with the Pearson system. Some special cases had previously been derived by Romanovsky (1924).

From a modeling point of view, it is important that the class of stationary distributions equals the full Pearson system of distributions. Thus a very wide spectrum of marginal distributions is available ranging from distributions with compact support to very heavy-tailed distributions with tails of the Pareto-type. To see that the invariant distributions belong to the Pearson system, note that the scale measure has density

$$s(x) = \exp\left(\int_{x_0}^{x} \frac{u-\alpha}{au^2+bu+c} du \right),$$

where x_0 is a point such that $ax_0^2 + bx_0 + c > 0$, cf. (1.14). Since the density of the invariant probability measure is given by

$$\mu_\theta(x) \propto \frac{1}{s(x)(ax^2+bx+c)},$$

cf. (1.15), it follows that

$$\mu_\theta'(x) = -\frac{(2a+1)x - \alpha + b}{ax^2+bx+c}\mu_\theta(x).$$

The Pearson system is defined as the class of probability densities obtained by solving a differential equation of this form, see Pearson (1895).

In the following we present a full classification of the ergodic Pearson diffusions, which shows that all distributions in the Pearson system can be obtained as invariant distributions for a model in the class of Pearson diffusions. We consider six cases according to whether the squared diffusion coefficient is constant, linear, a convex parabola with either zero, one or two roots, or a concave parabola with two roots. The classification problem can be reduced by first noting that the Pearson class of diffusions is closed under location and scale-transformations. To be specific, if X is an ergodic Pearson diffusion, then so is $\tilde{X}$ where $\tilde{X}_t = \gamma X_t + \delta$. The parameters of the stochastic differential equation (1.81) for $\tilde{X}$ are $\tilde{a} = a$, $\tilde{b} = b\gamma - 2a\delta$, $\tilde{c} = c\gamma^2 - b\gamma\delta + a\delta^2$, $\tilde{\beta} = \beta$, and $\tilde{\alpha} = \gamma\alpha + \delta$. Hence, up to transformations of location and scale, the ergodic Pearson diffusions can take the following forms. Note that we consider scale transformations in a general sense where multiplication by a negative real number is allowed, so to each case of a diffusion with state space $(0, \infty)$ there corresponds a diffusion with state space $(-\infty, 0)$.

Case 1: $\sigma^2(x) = 2\beta$**.** The solution to (1.81) is an Ornstein-Uhlenbeck process. The state space is $\mathbb{R}$, and the invariant distribution is the *normal distribution* with mean α and variance 1. The eigenfunctions are the Hermite polynomials.

Case 2: $\sigma^2(x) = 2\beta x$**.** The solution to (1.81) is the square root process (CIR process) (1.37) with state space $(0, \infty)$. Condition 1.3.1 that ensures ergodicity is satisfied if and only if $\alpha > 1$. If $0 < \alpha \leq 1$, the boundary 0 can with positive probability be reached at a finite time point, but if the boundary is made instantaneously reflecting, we obtain a stationary process. The invariant distribution is the *gamma distribution* with scale parameter 1 and shape parameter α. The eigenfunctions are the Laguerre polynomials.

Case 3: $a > 0$ and $\sigma^2(x) = 2\beta a(x^2 + 1)$. The state space is the real line, and the scale density is given by $s(x) = (x^2 + 1)^{\frac{1}{2a}} \exp(-\frac{\alpha}{a} \tan^{-1} x)$. By Condition 1.3.1, the solution is ergodic for all $a > 0$ and all $\alpha \in \mathbb{R}$. The invariant density is given by $\mu_\theta(x) \propto (x^2 + 1)^{-\frac{1}{2a} - 1} \exp(\frac{\alpha}{a} \tan^{-1} x)$ If $\alpha = 0$ the invariant distribution is a scaled t*-distribution* with $\nu = 1 + a^{-1}$ degrees of freedom and scale parameter $\nu^{-\frac{1}{2}}$. If $\alpha \neq 0$ the invariant distribution is skew and has tails decaying at the same rate as the t-distribution with $1 + a^{-1}$ degrees of freedom. A fitting name for this distribution is the *skew* t*-distribution*. It is also known as *Pearson's type IV distribution*. In either case the mean is α and the invariant distribution has moments of order k for $k < 1 + a^{-1}$. Because of its skew and heavy tailed marginal distribution, the class of diffusions with $\alpha \neq 0$ is potentially very useful in many applications, e.g. finance. It was studied and fitted to financial data by Nagahara (1996) using the local linearization method of Ozaki (1985). We consider this process in more detail below.

Case 4: $a > 0$ and $\sigma^2(x) = 2\beta a x^2$. The state space is $(0, \infty)$ and the scale density is $s(x) = x^{\frac{1}{a}} \exp(\frac{\alpha}{ax})$. Condition 1.3.1 holds if and only if $\alpha > 0$. The invariant distribution is given by $\mu_\theta(x) \propto x^{-\frac{1}{a} - 2} \exp(-\frac{\alpha}{ax})$, and is thus an *inverse gamma distribution* with shape parameter $1 + \frac{1}{a}$ and scale parameter $\frac{a}{\alpha}$. The invariant distribution has moments of order k for $k < 1 + a^{-1}$. This process is sometimes referred to as the GARCH diffusion model. The polynomial eigenfunctions are known as the Bessel polynomials.

Case 5: $a > 0$ and $\sigma^2(x) = 2\beta a x(x + 1)$.The state space is $(0, \infty)$ and the scale density is $s(x) = (1 + x)^{\frac{\alpha + 1}{a}} x^{-\frac{\alpha}{a}}$. The ergodicity Condition 1.3.1 holds if and only if $\frac{\alpha}{a} \geq 1$. Hence for all $a > 0$ and all $\mu \geq a$, a unique ergodic solution to (1.81) exists. If $0 < \alpha < 1$, the boundary 0 can be reached at a finite time point with positive probability, but if the boundary is made instantaneously reflecting, a stationary process is obtained. The density of the invariant distribution is given by $\mu_\theta(x) \propto (1+x)^{-\frac{\alpha+1}{a} - 1} x^{\frac{\alpha}{a} - 1}$. This is a scaled

F-distribution with $\frac{2\alpha}{a}$ and $\frac{2}{a}+2$ degrees of freedom and scale parameter $\frac{\alpha}{1+a}$. The invariant distribution has moments of order k for $k < 1 + a^{-1}$.

Case 6: $a < 0$ and $\sigma^2(x) = 2\beta a x(x-1)$. The state space is $(0,\infty)$ and the scale density is $s(x) = (1-x)^{\frac{1-\alpha}{a}} x^{\frac{\alpha}{a}}$. Condition 1.3.1 holds if and only if $\frac{\alpha}{a} \leq -1$ and $\frac{1-\alpha}{a} \leq -1$. Hence for all $a < 0$ and all $\alpha > 0$ such that $\min(\alpha, 1-\alpha) \geq -a$, a unique ergodic solution to (1.81) exists. If $0 < \alpha < -a$, the boundary 0 can be reached at a finite time point with positive probability, but if the boundary is made instantaneously reflecting, a stationary process is obtained. Similar remarks apply to the boundary 1 when $0 < 1-\alpha < -a$. The invariant distribution is given by $\mu_\theta(x) \propto (1-x)^{-\frac{1-\alpha}{a}-1} x^{-\frac{\alpha}{a}-1}$ and is thus the *Beta distribution* with shape parameters $\frac{\alpha}{-a}$ and $\frac{1-\alpha}{-a}$. This class of diffusions will be discussed in more detail below. It is often referred to as the *Jacobi diffusions* because the related eigenfunctions are Jacobi polynomials. Multivariate Jacobi diffusions were considered by Gourieroux & Jasiak (2006).

Example 1.3.18 The *skew t-distribution* with mean zero, ν degrees of freedom, and skewness parameter ρ has (unnormalized) density

$$\begin{aligned} &f(z) \propto \\ &\qquad \left\{(z/\sqrt{\nu}+\rho)^2+1\right\}^{-(\nu+1)/2} \exp\left\{\rho(\nu-1)\tan^{-1}\left(z/\sqrt{\nu}+\rho\right)\right\}, \end{aligned}$$

which is the invariant density of the diffusion $Z_t = \sqrt{\nu}(X_t - \rho)$ with $\nu = 1 + a^{-1}$ and $\rho = \alpha$, where X is as in Case 3. An expression for the normalizing constant when ν is integer valued was derived in Nagahara (1996). By the transformation result above, the corresponding stochastic differential equation is

$$dZ_t = -\beta Z_t dt + \sqrt{2\beta(\nu-1)^{-1}\{Z_t^2 + 2\rho\nu^{\frac{1}{2}} Z_t + (1+\rho^2)\nu\}}dW_t. \quad (1.84)$$

For $\rho = 0$ the invariant distribution is the t-distribution with ν degrees of freedom.

The skew t-diffusion (1.84) has the eigenvalues $\lambda_n = n(\nu-n)(\nu-1)^{-1}\beta$ for $n < \nu$. The four first eigenfunctions are

$$\begin{aligned} p_1(z) &= z, \\ p_2(z) &= z^2 - \frac{4\rho\nu^{\frac{1}{2}}}{\nu-3}z - \frac{(1+\rho^2)\nu}{\nu-2}, \\ p_3(z) &= z^3 - \frac{12\rho\nu^{\frac{1}{2}}}{\nu-5}z^2 + \frac{24\rho^2\nu + 3(1+\rho^2)\nu(\nu-5)}{(\nu-5)(\nu-4)}z + \frac{8\rho(1+\rho^2)\nu^{\frac{3}{2}}}{(\nu-5)(\nu-3)}, \end{aligned}$$

and

$$p_4(z) = z^4 - \frac{24\rho\nu^{\frac{1}{2}}}{\nu-7}z^3 + \frac{144\rho^2\nu - 6(1+\rho^2)\nu(\nu-7)}{(\nu-7)(\nu-6)}z^2$$

$$+\frac{8\rho(1+\rho^2)\nu^{\frac{3}{2}}(\nu-7)+48\rho(1+\rho^2)\nu^{\frac{3}{2}}(\nu-6)-192\rho^3\nu^{\frac{3}{2}}}{(\nu-7)(\nu-6)(\nu-5)}z$$
$$+\frac{3(1+\rho^2)^2\nu(\nu-7)-72\rho^2(1+\rho^2)\nu^2}{(\nu-7)(\nu-6)(\nu-4)},$$

provided that $\nu > 4$. If $\nu > 2i$, the first i eigenfunctions are square integrable and thus satisfy (1.72). Hence (1.73) holds, and the eigenfunctions can be used to construct explicit martingale estimating functions. □

Example 1.3.19 The model

$$dX_t = -\beta[X_t - (m+\gamma z)]dt + \sigma\sqrt{z^2 - (X_t - m)^2}dW_t, \tag{1.85}$$

where $\beta > 0$ and $\gamma \in (-1,1)$, has been proposed as a model for the random variation of the logarithm of an exchange rate in a target zone between realignments by De Jong, Drost & Werker (2001) ($\gamma = 0$) and Larsen & Sørensen (2007). This is a diffusion on the interval $(m-z, m+z)$ with mean reversion around $m+\gamma z$. It is a *Jacobi diffusion* obtained by a location-scale transformation of the diffusion in Case 6 above. The parameter γ quantifies the asymmetry of the model. When $\beta(1-\gamma) \geq \sigma^2$ and $\beta(1+\gamma) \geq \sigma^2$, X is an ergodic diffusion, for which the stationary distribution is a Beta-distribution on $(m-z, m+z)$ with parameters $\kappa_1 = \beta(1-\gamma)\sigma^{-2}$ and $\kappa_2 = \beta(1+\gamma)\sigma^{-2}$. If the parameter restrictions are not satisfied, one or both of the boundaries can be hit in finite time, but if the boundaries are made instantaneously reflecting, a stationary process is obtained.

The eigenfunctions for the generator of the diffusion (1.85) are

$$\phi_i(x;\beta,\gamma,\sigma,m,z) = P_i^{(\kappa_1-1,\,\kappa_2-1)}((x-m)/z), \quad i = 1,2,\ldots$$

where $P_i^{(a,b)}(x)$ denotes the Jacobi polynomial of order i given by

$$P_i^{(a,b)}(x) = \sum_{j=0}^{i} 2^{-j}\binom{n+a}{n-j}\binom{a+b+n+j}{j}(x-1)^j, \quad -1 < x < 1.$$

The eigenvalue of ϕ_i is $i(\beta + \frac{1}{2}\sigma^2(i-1))$. Since (1.72) is obviously satisfied, (1.73) holds, so that the eigenfunctions can be used to construct explicit martingale estimating functions. □

Explicit formulae for the *conditional moments* of a Pearson diffusion can be obtained from the eigenfunctions by means of (1.73). Specifically,

$$\mathrm{E}(X_t^n \mid X_0 = x) = \sum_{k=0}^{n}\left(\sum_{\ell=0}^{n} q_{n,k,\ell}e^{-\lambda_\ell t}\right)x^k, \tag{1.86}$$

where $q_{n,k,n} = p_{n,k}$, $q_{n,n,\ell} = 0$ for $\ell \leq n-1$, and

$$q_{n,k,\ell} = - \sum_{j=k\vee\ell}^{n-1} p_{n,j} q_{j,k,\ell}$$

for $k, \ell = 0, \ldots, n-1$ with λ_ℓ and $p_{n,j}$ given by (1.82) and (1.83). For details see Forman & Sørensen (2008).

Also the *moments* of the Pearson diffusions can, when they exist, be found explicitly by using the fact that the integral of the eigenfunctions with respect to the invariant probability measure is zero.We have seen above that $E(|X_t|^\kappa) < \infty$ if and only if $a < (\kappa - 1)^{-1}$. Thus if $a \leq 0$ all moments exist, while for $a > 0$ only the moments satisfying that $\kappa < a^{-1} + 1$ exist. In particular, the expectation always exists. The moments of the invariant distribution can be found by the recursion

$$\mathrm{E}(X_t^n) = a_n^{-1}\{b_n \cdot \mathrm{E}(X_t^{n-1}) + c_n \cdot \mathrm{E}(X_t^{n-2})\}, \quad n = 2, 3, \ldots, \tag{1.87}$$

where $a_n = n\{1-(n-1)a\}\beta$, $b_n = n\{\alpha+(n-1)b\}\beta$, and $c_n = n(n-1)c\beta$. The initial conditions are given by $\mathrm{E}(X_t^0) = 1$, and $\mathrm{E}(X_t) = \alpha$. This can be found from the expressions for the eigenfunctions, but is more easily seen as follows. By Ito's formula

$$\begin{aligned} dX_t^n &= -\beta n X_t^{n-1}(X_t - \mu)dt + \beta n(n-1)X_t^{n-2}(aX_t^2 + bX_t + c)dt \\ &\qquad + nX_t^{n-1}\sigma(X_t)dW_t, \end{aligned}$$

and if $E(X_t^{2n})$ is finite, i.e. if $a < (2n-1)^{-1}$, the last term is a martingale with expectation zero.

Example 1.3.20 Equation (1.87) allows us to find the moments of the *skewed t-distribution*, in spite of the fact that the normalizing constant of the density is unknown. In particular, for the diffusion (1.84),

$$\begin{aligned} \mathrm{E}(Z_t) &= 0, \\ \mathrm{E}(Z_t^2) &= \frac{(1+\rho^2)\nu}{\nu-2}, \\ \mathrm{E}(Z_t^3) &= \frac{4\rho(1+\rho^2)\nu^{\frac{3}{2}}}{(\nu-3)(\nu-2)}, \\ \mathrm{E}(Z_t^4) &= \frac{24\rho^2(1+\rho^2)\nu^2 + 3(\nu-3)(1+\rho^2)^2\nu^2}{(\nu-4)(\nu-3)(\nu-2)}. \end{aligned}$$

□

For a diffusion $T(X)$ obtained from a solution X to (1.81) by a twice differentiable and invertible transformation T, the eigenfunctions of the generator are

$p_n\{T^{-1}(x)\}$, where p_n is an eigenfunction of the generator of X. The eigenvalues are the same as for the original eigenfunctions. Since the original eigenfunctions are polynomials, the eigenfunctions of $T(X)$ are of the form (1.77) with $\kappa = T^{-1}$. Hence *explicit optimal martingale estimating functions are also available for transformations of Pearson diffusions*, which is a very large and flexible class of diffusion processes. Their stochastic differential equations can, of course, be found by Ito's formula.

Example 1.3.21 For the Jacobi-diffusion (case 6) with $\mu = -a = \frac{1}{2}$, i.e.

$$dX_t = -\beta(X_t - \tfrac{1}{2})dt + \sqrt{\beta X_t(1 - X_t)}dW_t$$

the invariant distribution is the uniform distribution on $(0, 1)$ for all $\beta > 0$. For any strictly increasing and twice differentiable distribution function F, we therefore have a class of diffusions given by $Y_t = F^{-1}(X_t)$ or

$$\begin{aligned} dY_t &= -\beta \frac{(F(Y_t) - \frac{1}{2})f(Y_t)^2 + \frac{1}{2}F(Y_t)\{1 - F(Y_t)\}}{f(Y_t)^3} dt \\ &\qquad\qquad + \frac{\beta F(Y_t)\{1 - F(Y_t)\}}{f(Y_t)} dW_t, \end{aligned}$$

which has invariant distribution with density $f = F'$. A particular example is the logistic distribution

$$F(x) = \frac{e^x}{1 + e^x}, \quad x \in \mathbb{R},$$

for which

$$dY_t = -\beta \left\{\sinh(x) + 8\cosh^4(x/2)\right\} dt + 2\sqrt{\beta}\cosh(x/2)dW_t.$$

If the same transformation $F^{-1}(y) = \log(y/(1 - y))$ is applied to the general Jacoby diffusion (case 6), then we obtain

$$\begin{aligned} dY_t &= -\beta \left\{1 - 2\mu + (1 - \mu)e^{Y_t} - \mu e^{-1} - 8a\cosh^4(Y_t/2)\right\} dt \\ &\qquad\qquad + 2\sqrt{-a\beta}\cosh(Y_t/2)dW_t, \end{aligned}$$

a diffusion for which the invariant distribution is the generalized logistic distribution with density

$$f(x) = \frac{e^{\kappa_1 x}}{(1 + e^x)^{\kappa_1 + \kappa_2} B(\kappa_1, \kappa_2)}, \quad x \in \mathbb{R},$$

where $\kappa_1 = -(1 - \alpha)/a$, $\kappa_2 = \alpha/a$ and B denotes the Beta-function. This distribution was introduced and studied in Barndorff-Nielsen, Kent & Sørensen (1982).

□

Example 1.3.22 Let again X be a general Jacobi-diffusion (case 6). If we apply the transformation $T(x) = \sin^{-1}(2x-1)$ to X_t, we obtain the diffusion

$$dY_t = -\rho \frac{\sin(Y_t) - \varphi}{\cos(Y_t)} dt + \sqrt{-a\beta/2} dW_t,$$

where $\rho = \beta(1 + a/4)$ and $\varphi = (2\alpha - 1)/(1 + a/4)$. The state space is $(-\pi/2, \pi/2)$. Note that Y has dynamics that are very different from those of the Jacobi diffusion: the drift is highly non-linear and the diffusion coefficient is constant. This model was considered in Example 1.3.15.

□

1.3.8 Implementation of martingale estimating functions

An R-package, where a number of methods for calculating estimators for diffusion models are implemented, has been developed by Stefano Iacus and is described in the book Iacus (2008), which outlines the underlying theory too. The R-package also contains implementations of methods for simulating solutions to stochastic differential equations. It is, however, useful to notice that for many martingales estimating functions the estimators, or asymptotically equivalent estimators, can be calculated by means of standard statistical software packages. Specifically, they can be calculated as weighted least squares estimators for non-linear regression models.

To see this, consider the weighted least squares estimator obtained by minimizing

$$C_n(\theta) = \tag{1.88}$$
$$\sum_{i=1}^{n} \left[f(X_{t_i}) - \pi_\Delta^\theta(f)(X_{t_{i-1}}) \right]^T V_i^{-1} \left[f(X_{t_i}) - \pi_\Delta^\theta(f)(X_{t_{i-1}}) \right],$$

with $f(x) = (f_1(x), \ldots, f_N(x))$ and

$$V_i = V_h(X_{t_{i-1}}; \tilde{\theta}_n), \tag{1.89}$$

where V_h is the $N \times N$-matrix given by (1.44), and $\tilde{\theta}_n$ is a consistent estimator of θ. As usual, π_Δ^θ denotes the transition operator (1.42). The consistent estimator can, for instance, be the non-weighted least squares estimator obtained by minimizing (1.88) with $V_i = I_N$, where I_N is the $N \times N$ identity matrix. The weighted least squares estimator obtained from (1.88) with V_i given by (1.89) solves the estimating equation

$$\sum_{i=1}^{n} B_h(X_{t_{i-1}}; \theta) V_h(X_{t_{i-1}}; \tilde{\theta}_n)^{-1} \left[f(X_{t_i}) - \pi_\Delta^\theta(f)(X_{t_{i-1}}) \right] = 0 \tag{1.90}$$

with B_h given by (1.45). Therefore this estimator has the same asymptotic variance as the optimal G_n^*-estimator with h given by (1.40); see e.g. Jacod & Sørensen (2009). The estimating function (1.90) is similar in spirit to (1.39). The estimators obtained by minimizing (1.88) is the weighted least squares estimator for a regression model for the data $f(X_{t_i})$, $i = 1, \ldots, n$ with $X_{t_{i-1}}$, $i = 1, \ldots, n$, as explanatory regression variables, the non-linear regression function $\pi_\Delta^\theta(f)(X_{t_{i-1}})$, and the weight matrix V_i. In some particularly nice cases, the regression function is linear in the parameters, and the estimator is a linear regression estimator.

Example 1.3.23 Let X be the square root process (1.37), and suppose we have the observations $X_{i\Delta}$, $i = 0, \ldots, n$. Let us think of $(X_{i\Delta}, X_{i\Delta}^2)$, $i = 1, \ldots, n$, as data with explanatory regression variables $X_{(i-1)\Delta}$, $i = 1, \ldots, n$, and with the non-linear regression function

$$\pi_\Delta^\theta(f)(X_{(i-1)\Delta}) = \left(\begin{array}{l} F(X_{(i-1)\Delta};\theta) \\ \phi(X_{(i-1)\Delta};\theta) + F(X_{(i-1)\Delta};\theta)^2 \end{array} \right),$$

where F and ϕ are as in Example 1.3.6. Then we obtain a weighted least squares estimator for θ, by minimizing (1.88) with $f_1(x) = x$, $f_2(x) = x^2$, and

$V_h(x;\theta) =$

$$\begin{pmatrix} \phi(x;\theta) & \eta(x;\theta)+2F(x;\theta)\phi(x;\theta)^2 \\ \eta(x;\theta)+2F(x;\theta)\phi(x;\theta)^2 & \psi(x;\theta)+4F(x;\theta)^2\phi(x;\theta)+4F(x;\theta)\eta(x;\theta) \end{pmatrix},$$

where η and ψ are as in Example 1.3.6.

This estimator has the same efficiency as the estimator obtained from the optimal martingale estimating function of form (1.30) with $N = 2$ and

$$\begin{aligned} h_1(x,y;\theta) &= y - F(x;\theta) \\ h_2(x,y;\theta) &= y^2 - \phi(x;\theta) - F(x;\theta)^2. \end{aligned}$$

The optimal estimating function of this form is equivalent to the optimal estimating function in Example 1.3.6. For the square root process some simplification can be achieved by using the Gaussian approximation (1.38) in the definition of the matrix V_h.

□

Example 1.3.24 Consider the process (1.74), and suppose we have the observations $X_{i\Delta}$, $i = 0, \ldots, n$. Let us think of $\sin(X_{i\Delta})$, $i = 1, \ldots, n$, as data with explanatory regression variables $X_{(i-1)\Delta}$, $i = 1, \ldots, n$ and with the non-linear regression function $\pi_\Delta^\theta(\sin)(X_{(i-1)\Delta}) = e^{-(\theta+1/2)\Delta}\sin(X_{(i-1)\Delta})$. Again we can obtain a weighted least squares estimator for θ, by minimizing (1.88) with $f(x) = \sin(x)$ and

$$V_h(x;\theta) = \tfrac{1}{2}(e^{2(\tilde{\theta}_n+1)\Delta} - 1)/(\tilde{\theta}_n + 1) - (e^{\Delta} - 1)\sin^2(X_{(i-1)\Delta}),$$

where $\tilde{\theta}_n$ is a consistent estimator, for instance the simple estimator (1.76). Note that the non-linear regression is, in fact, a linear regression in the parameter $\xi = e^{-\theta\Delta}$. The regression estimator equals the estimator obtained from the estimating function

$$G_n^\bullet(\theta) = \sum_{i=1}^{n} \frac{\sin(X_{(i-1)\Delta})[\sin(X_{i\Delta}) - e^{-(\theta+\frac{1}{2})\Delta}\sin(X_{(i-1)\Delta})]}{\frac{1}{2}(e^{2(\tilde{\theta}_n+1)\Delta} - 1)/(\tilde{\theta}_n + 1) - (e^{\Delta} - 1)\sin^2(X_{(i-1)\Delta})},$$

which has the same efficiency as the optimal estimator obtained in Example 1.3.17. If instead we minimize (1.88) with the approximation $V_h(x;\theta) = \cos^2(x)$, then we obtain the estimator $\tilde{\theta}_n$ from Example 1.3.17, and if we minimize (1.88) with the more crude approximation $V_h(x;\theta) = 1$, then we obtain the simple estimator (1.76) from Example 1.3.15.

□

More generally, an estimator with the same efficiency as the optimal estimator from (1.30) with optimal weights (1.32) is obtained by minimizing the objective function

$$\sum_{i=1}^{n} h(X_{t_{i-1}}, X_{t_i};\theta)^T V_i^{-1} h(X_{t_{i-1}}, X_{t_i};\theta) \tag{1.91}$$

with V_i defined as in (1.89), but here with V_h given by (1.34). This estimator can be found by applying standard software for minimizing objective functions to (1.91).

Example 1.3.25 Let again X be the square root process (1.37), and consider the martingale estimating function of form (1.30) with $N = 2$ and h_1 and h_2 as in Example 1.3.4. In this case an optimal estimator is obtained by minimizing (1.91) with

$$V_h(x;\theta) = \begin{pmatrix} \phi(x;\theta) & \eta(x;\theta) \\ \eta(x;\theta) & \psi(x;\theta) \end{pmatrix},$$

where ϕ, η and ψ are as in Example 1.3.6. Here a considerable simplification can be obtained by the Gaussian approximation (1.38). With this approximation

$$V_h(x;\theta) = \begin{pmatrix} \phi(x;\theta) & 0 \\ 0 & 2\phi(x;\theta)^2 \end{pmatrix}.$$

□

1.4 The likelihood function

The likelihood function for a discretely observed diffusion model, (1.24) is a product of transitions densities. Unfortunately, the transition density of a diffusion process is only rarely explicitly known, but several numerical approaches make likelihood inference feasible for diffusion models.

Pedersen (1995) proposed a method for obtaining an approximation to the likelihood function by rather extensive simulation. Pedersen's method was very considerably improved by Durham & Gallant (2002), whose method is computationally much more efficient. Poulsen (1999) obtained an approximation to the transition density by numerically solving a partial differential equation, whereas Aït-Sahalia (2002) and Aït-Sahalia (2008) proposed to approximate the transition density by means of expansions. A Gaussian approximation to the likelihood function obtained by local linearization of (1.11) was proposed by Ozaki (1985), while Forman & Sørensen (2008) proposed to use an approximation in terms of eigenfunctions of the generator of the diffusion. Bayesian estimators with the same asymptotic properties as the maximum likelihood estimator can be obtained by Markov chain Monte Carlo methods, see Elerian, Chib & Shephard (2001), Eraker (2001), and Roberts & Stramer (2001). Finally, exact and computationally efficient likelihood-based estimation methods were presented by Beskos et al. (2006). The latter approach is presented in Chapter XXX. In the following we will outline the expansion approach of Aït-Sahalia (2002) for scalar diffusion models. The various other approaches to calculation of the likelihood function will not be considered further in this chapter.

Assume that the diffusion process (1.11) is one-dimensional and that the state space is either $(-\infty, \infty)$ or $(0, \infty)$, i.e. $r = \infty$ and ℓ is either $-\infty$ or 0. The coefficients b and σ are assumed to satisfy the following condition.

Condition 1.4.1
(i) The functions $b(x;\theta)$ and $\sigma(x;\theta)$ are infinitely often differentiable w.r.t. x and three times continuously differentiable w.r.t. θ for all $x \in (\ell, r)$ and $\theta \in \Theta$.

(ii-a) If $\ell = -\infty$, there exists a constant $c > 0$ such that $\sigma(x;\theta) > c$ for all $x \in (\ell, r)$ and all $\theta \in \Theta$.

(ii-b) If $\ell = 0$, then σ is non-degenerate on $(0, \infty)$ in the sense that for each $\xi > 0$ there exists a constant $c_\xi > 0$ such that $\sigma(x;\theta) \geq c_\xi$ for all $x \geq \xi$ and all $\theta \in \Theta$. Moreover, if $\lim_{x \to 0} \sigma(x;\theta) = 0$, then constants ξ_0, ω and ρ exist such that $\sigma(x;\theta) \geq \omega x^\rho$ for all $x \in (0, \xi_0)$ and all $\theta \in \Theta$.

The idea is to make an expansion of the transition density. However, the distribution of X_Δ given X_0 can be so far from a normal distribution that a convergent expansion with the normal density as leading term is not possible. This is possible for a diffusion with constant diffusion coefficient. Therefore the standard transformation

$$h(x;\theta) = \int_{x^*}^{x} \frac{1}{\sigma(u;\theta)} du,$$

where x^* is arbitrary, is applied to obtain the diffusion process

$$Y_t = h(X_t;\theta).$$

Since $\sigma > 0$, the transformation h is increasing, and by Ito's formula

$$dY_t = a(Y_t;\theta)dt + dW_t, \tag{1.92}$$

where

$$a(y;\theta) = \frac{b(h^{-1}(y;\theta);\theta)}{\sigma(h^{-1}(y;\theta);\theta)} - \tfrac{1}{2}\sigma'(h^{-1}(y;\theta);\theta)$$

with $\sigma'(x;\theta) = \partial_x \sigma(x;\theta)$. The state space of Y, (ℓ_Y, r_Y) could in principle depend on θ, but we assume that this is not the case. If only one of the boundaries ℓ_Y and r_Y is finite, it can always be arranged that the finite boundary equals zero by choosing x^* suitably. For instance if $r_Y = \infty$ and ℓ_Y is finite, then we can choose $x^* = \ell$ to obtain $\ell_Y = 0$. We will assume that ℓ_Y is either $-\infty$ or 0, and that r_Y is either 0 or ∞. It is further assumed that a satisfies the following condition (which can be translated into a condition on b and σ).

Condition 1.4.2

(i) For all $\theta \in \Theta$, the drift coefficient $a(y;\theta)$ and its derivatives w.r.t. y and θ have at most polynomial growth near the boundaries, and

$$\lim[a(y;\theta)^2 + \partial_y a(y;\theta)] > -\infty \quad \textit{as } y \downarrow \ell_Y \quad \textit{and} \quad y \uparrow r_Y.$$

(ii-a) If $\ell_Y = 0$, then there exist constants $\epsilon_0 > 0$, κ and α such that $a(y;\theta) \geq \kappa y^{-\alpha}$ for all $y \in (0, \epsilon_0)$ and all $\theta \in \Theta$, where either $\alpha > 1$ and $\kappa > 0$, or $\alpha = 1$ and $\kappa \geq 1$. If $\ell_Y = -\infty$, then there exists constants $E_0 > 0$ and $K > 0$ such that $a(y;\theta) \geq Ky$ for all $y \leq -E_0$ and all $\theta \in \Theta$.

(ii-b) If $r_Y = 0$, then there exist constants $\epsilon_0 > 0$, κ and α such that $a(y;\theta) \leq -\kappa|y|^{-\alpha}$ for all $y \in (-\epsilon_0, 0)$ and all $\theta \in \Theta$, where either $\alpha > 1$ and $\kappa > 0$, or $\alpha = 1$ and $\kappa \geq 1/2$. If $r_Y = \infty$, then there exist constants $E_0 > 0$ and $K > 0$ such that $a(y;\theta) \leq Ky$ for all $y \geq E_0$ and all $\theta \in \Theta$.

A real function f is said to be of polynomial growth near a boundary at ∞ or $-\infty$ if there exist constants $C > 0$, $K > 0$ and $p > 0$ such that $|f(x)| \leq C|x|^p$ for $x > K$ or $x < -K$. If the boundary is at zero, polynomial growth means that there exist constants $C > 0$, $\epsilon > 0$ and $p > 0$ such that $|f(x)| \leq C|x|^{-p}$ for $|x| \leq \epsilon$.

Under the assumptions imposed, a solution exists to (1.92) with a transition density that is sufficiently regular for likelihood inference. This is the contents of the following proposition from Aït-Sahalia (2002).

Proposition 1.4.3 *Under the Conditions 1.4.1 and 1.4.2, the stochastic differential equation (1.92) has a unique weak solution for every initial distribution. The boundaries are unattainable. The solution Y has a transition density $p_Y(\Delta, y_0, y; \theta)$ that is continuously differentiable w.r.t. Δ, infinitely often differentiable w.r.t. $y \in (\ell_Y, r_Y)$, and three times continuously differentiable w.r.t. $\theta \in \Theta$.*

This result implies that the original stochastic differential equation (1.11) has a unique weak solution, and by the transformation theorem for density functions, it has a similarly regular transition density given by

$$p(\Delta, x_0, x; \theta) = p_Y(\Delta, h(x_0; \theta), h(x; \theta); \theta)/\sigma(x; \theta). \tag{1.93}$$

Instead of expanding the transition density of Y, i.e. the conditional density function of Y_Δ given $Y_0 = y_0$, we expand the conditional density of the normalized increment

$$Z = \Delta^{-1}(Y_\Delta - y_0)$$

given $Y_0 = y_0$. This is because p_Y gets peaked around y_0 as Δ gets close to zero, whereas the distribution of Z is sufficiently close to the $N(0,1)$-distribution to make it the appropriate transformation of X_Δ to obtain a convergent expansion of the conditional density function with the standard normal density function as the leading term. Obviously,

$$p_Y(\Delta, y_0, y; \theta) = \Delta^{-1/2} p_Z(\Delta, \Delta^{-1/2}(y - y_0) \,|\, y_0; \theta), \tag{1.94}$$

where $p_Z(\Delta, z \,|\, y_0; \theta)$ is the conditional density of Z given that $Y_0 = y_0$.

We can now obtain an approximation to the transition density of X, and hence an approximation to the likelihood function, by expanding the conditional density, p_Z, of Z given $Y_0 = y_0$ in terms of Hermite polynomials up to order J:

$$p_Z^J(\Delta, z \,|\, y_0; \theta) = \varphi(z) \sum_{j=0}^{J} \eta_j(\Delta, y_0; \theta) H_j(z), \tag{1.95}$$

where φ denotes the density function of the standard normal distribution, and H_j is the jth Hermite polynomial, which is defined by

$$H_j(x) = (-1)^j e^{x^2/2} \frac{d^j}{dx^j} e^{-x^2/2}, \quad j = 0, 1, \ldots.$$

The Hermite polynomials up to order 4 are

$$\begin{aligned} H_0(x) &= 1 \\ H_1(x) &= x \\ H_2(x) &= x^2 - 1 \\ H_3(x) &= x^3 - 3x \\ H_4(x) &= x^4 - 6x^2 + 3. \end{aligned}$$

The coefficients $\eta_j(\Delta, y_0; \theta)$ can be found by using that the Hermite polynomials are orthogonal in the space $L^2(\varphi)$:

$$\int_{-\infty}^{\infty} H_i(x) H_j(x) \varphi(x) dx = \begin{cases} 0 & \text{if } i \neq j \\ i! & \text{if } i = j. \end{cases}$$

Hence if

$$p_Z(\Delta, z \,|\, y_0; \theta) = \varphi(z) \sum_{j=0}^{\infty} \eta_j(\Delta, y_0; \theta) H_j(z),$$

it follows that

$$\begin{aligned} \int_{-\infty}^{\infty} H_i(z) p_Z(\Delta, z \,|\, y_0; \theta) dz &= \sum_{j=0}^{\infty} \eta_j(\Delta, y_0; \theta) \int_{-\infty}^{\infty} H_i(z) H_j(z) \varphi(z) dz \\ &= i!\, \eta_i(\Delta, y_0; \theta). \end{aligned}$$

By inserting the expansion (1.95) in (1.94) and (1.93), we obtain the following approximations to the transitions densities p_Y and p

$$p_Y^J(\Delta, y_0, y; \theta) = \Delta^{-1/2} \varphi(\Delta^{-1/2}(y - y_0)) \sum_{j=0}^{J} \eta_j(\Delta, y_0; \theta) H_j(\Delta^{-1/2}(y - y_0)) \tag{1.96}$$

and

$$p^J(\Delta, x_0, x; \theta) = \tag{1.97}$$

$$\frac{\varphi\left(\frac{h(x;\theta) - h(x_0;\theta)}{\sqrt{\Delta}}\right)}{\sqrt{\Delta}\, \sigma(x; \theta)} \sum_{j=0}^{J} \eta_j(\Delta, h(x_0; \theta); \theta) H_j\left(\frac{h(x;\theta) - h(x_0;\theta)}{\sqrt{\Delta}}\right).$$

Aït-Sahalia (2002) gave the following theorem about the convergence of the approximation p^J to the exact transition density p.

Theorem 1.4.4 *Under the Conditions 1.4.1 and 1.4.2, there exists $\bar{\Delta} > 0$ such that*

$$\lim_{J \to \infty} p^J(\Delta, x_0, x; \theta) = p(\Delta, x_0, x; \theta)$$

for all $\Delta \in (0, \bar{\Delta})$, $\theta \in \Theta$ and $(x_0, x) \in (\ell, r)^2$.

If $r_Y = \infty$ and $a(y; \theta) \leq 0$ near r_Y, and if $a(y; \theta) \geq 0$ near ℓ_Y (which is either 0 or $-\infty$), then $\bar{\Delta} = \infty$, see Proposition 2 in Aït-Sahalia (2002).

In order to use the expansions of the transition densities to calculate likelihood functions in practice, it is necessary to determine the coefficients $\eta_j(\Delta, y_0; \theta)$.

Note that by inserting (1.94) in the expression above for $\eta_i(\Delta, y_0; \theta)$ we find that

$$\begin{aligned}\eta_i(\Delta, y_0; \theta) &= \frac{1}{i!}\int_{-\infty}^{\infty} H_i(z)\Delta^{1/2} p_Y(\Delta, y_0, \Delta^{1/2}z + y_0; \theta)dz \\ &= \frac{1}{i!}\int_{-\infty}^{\infty} H_i(\Delta^{-1/2}(y-y_0)) p_Y(\Delta, y_0, y; \theta)dy \\ &= \mathrm{E}_\theta\left(H_i(\Delta^{-1/2}(Y_\Delta - y_0)) \,|\, Y_0 = y_0\right).\end{aligned}$$

Thus the coefficients $\eta_i(\Delta, y_0; \theta)$, $i = 0, 1, \ldots$, are conditional moments of the process Y, and can therefore be found by simulation of Y or X. An approximation to $\eta_i(\Delta, y_0; \theta)$ can be obtained by applying the expansion (1.57) to the functions $(y-x)^i$, $i = 1, \ldots, J$. For instance, we find that

$$\begin{aligned}\eta_1(\Delta, y_0; \theta) &= \Delta^{1/2} a(y_0; \theta) + \tfrac{1}{2}\Delta^{3/2}\left(a(y_0; \theta)\partial_y a(y_0; \theta) + \tfrac{1}{2}\partial_y^2 a(y_0; \theta)\right) \\ &\qquad + O(\Delta^{5/2}) \\ \eta_2(\Delta, y_0; \theta) &= \Delta\left(a(y_0; \theta)^2 + \partial_y a(y_0; \theta)\right) + O(\Delta^2).\end{aligned}$$

By expanding the coefficients $\eta_i(\Delta, y_0; \theta)$ suitably and collecting terms of the same order in Δ, Aït-Sahalia (2002) found the following approximation to p_Y

$$\tilde{p}_Y^K(\Delta, y_0, y; \theta) = \\ \Delta^{-1/2}\varphi\left(\frac{y-y_0}{\sqrt{\Delta}}\right)\exp\left(\int_{y_0}^{y} a(w, \theta)dw\right)\sum_{k=0}^{K}\frac{\Delta^k}{k!}c_k(y_0, y; \theta),$$

where $c_0(y_0, y; \theta) = 1$, and

$$c_k(y_0, y; \theta) = \\ k(y-y_0)^{-k}\int_{y_0}^{y}(w-y_0)^{k-1}\left[\lambda(w; \theta)c_{k-1}(y_0, w; \theta) + \tfrac{1}{2}\partial_w^2 c_{k-1}(y_0, w; \theta)\right]dw,$$

for $k \geq 1$, where

$$\lambda(w; \theta) = -\tfrac{1}{2}\left(a(w; \theta)^2 + \partial_w a(w; \theta)\right).$$

1.5 Non-martingale estimating functions

1.5.1 Asymptotics

When the estimating function

$$G_n(\theta) = \sum_{i=r}^{n} g(X_{(i-r+1)\Delta}, \ldots, X_{i\Delta}; \theta)$$

is not a martingale under P_θ, further conditions on the diffusion process must be imposed to ensure the asymptotic normality in (1.5). A sufficient condition that (1.5) holds under P_{θ_0} with $V(\theta)$ given by (1.98) is that the diffusion process is stationary and geometrically α-mixing, that

$$\begin{aligned} V(\theta) \;=\;& Q_{\theta_0}\left(g(\theta)g(\theta)^T\right) \qquad (1.98)\\ &+\sum_{k=1}^{\infty}\Big[E_{\theta_0}\left(g(X_\Delta,\ldots,X_{r\Delta})g(X_{(k+1)\Delta},\ldots,X_{(k+r)\Delta})^T\right)\\ &\qquad + E_{\theta_0}\left(g(X_{(k+1)\Delta},\ldots,X_{(k+r)\Delta})g(X_\Delta,\ldots,X_{r\Delta})^T\right)\Big], \end{aligned}$$

converges and is strictly positive definite, and that $Q_{\theta_0}(g_i(\theta)^{2+\epsilon}) < \infty$, $i = 1,\ldots,p$ for some $\epsilon > 0$, see e.g. Doukhan (1994). Here g_i is the ith coordinate of g, and Q_θ is the joint distribution of $X_\Delta,\ldots,X_{r\Delta}$ under P_θ. To define the concept of α-mixing, let $\mathcal{F}_t$ denote the σ-field generated by $\{X_s \mid s \leq t\}$, and let $\mathcal{F}^t$ denote the σ-field generated by $\{X_s \mid s \geq t\}$. A stochastic process X is said to be α-*mixing* under P_{θ_0}, if

$$\sup_{A\in\mathcal{F}_t, B\in\mathcal{F}^{t+u}} |P_{\theta_0}(A)P_{\theta_0}(B) - P_{\theta_0}(A\cap B)| \leq \alpha(u)$$

for all $t > 0$ and $u > 0$, where $\alpha(u) \to 0$ as $u \to \infty$. This means that X_t and X_{t+u} are almost independent, when u is large. If positive constants c_1 and c_2 exist such that

$$\alpha(u) \leq c_1 e^{-c_2 u},$$

for all $u > 0$, then the process X is called geometrically α-mixing. For one-dimensional diffusions there are simple conditions for geometric α-mixing. If all non-zero eigenvalues of the generator (1.46) are larger than some $\lambda > 0$, then the diffusion is geometrically α-mixing with $c_2 = \lambda$. This is for instance the case if the spectrum of the generator is discrete. Ergodic diffusions with a linear drift $-\beta(x-\alpha)$, $\beta > 0$, for instance the Pearson diffusions, are geometrically α-mixing with $c_2 = \beta$; see Hansen, Scheinkman & Touzi (1998).

Genon-Catalot, Jeantheau & Larédo (2000) gave the following simple sufficient condition for the one-dimensional diffusion that solves (1.11) to be geometrically α-mixing, provided that it is ergodic with invariant probability density μ_θ.

Condition 1.5.1
(i) The function b is continuously differentiable with respect to x, and σ is twice continuously differentiable with respect to x, $\sigma(x;\theta) > 0$ for all $x \in (\ell, r)$, and a constant $K_\theta > 0$ exists such that $|b(x;\theta)| \leq K_\theta(1+|x|)$ and $\sigma^2(x;\theta) \leq K_\theta(1+x^2)$ for all $x \in (\ell, r)$.

(ii) $\sigma(x;\theta)\mu_\theta(x) \to 0$ as $x \downarrow \ell$ and $x \uparrow r$.

(iii) $1/\gamma(x;\theta)$ *has a finite limit as* $x \downarrow \ell$ *and* $x \uparrow r$*, where* $\gamma(x;\theta) = \partial_x \sigma(x;\theta) - 2b(x;\theta)/\sigma(x;\theta)$.

Other conditions for geometric α-mixing were given by Veretennikov (1987), Hansen & Scheinkman (1995), and Kusuoka & Yoshida (2000).

For geometrically α-mixing diffusions processes and estimating functions G_n satisfying Condition 1.2.1, the existence of a $\bar{\theta}$-consistent and asymptotically normal G_n-estimator follows from Theorem 1.2.2, which also contains a result about eventual uniqueness of the estimator.

1.5.2 Explicit non-martingale estimating functions

Explicit martingale estimating functions are only available for the relatively small, but versatile, class of diffusions for which explicit eigenfunctions of the generator are available; see the Subsections 1.3.6 and 1.3.7. Explicit non-martingale estimating functions can be found for all diffusions, but cannot be expected to approximate the score functions as well as martingale estimating functions, and therefore usually give less efficient estimators. As usual we consider ergodic diffusion processs with invariant probability density μ_θ.

First we consider estimating functions of the form

$$G_n(\theta) = \sum_{i=1}^{n} h(X_{\Delta i};\theta), \tag{1.99}$$

where h is a p-dimensional function. We assume that the diffusion is geometrically α-mixing, so that a central limit theorem holds (under an integrability condition), and that Condition 1.2.1 holds for $r = 1$ and $\bar{\theta} = \theta_0$. The latter condition simplifies considerably, because for estimating functions of the form (1.99), it does not involve the transition density, but only the invariant probability density μ_θ, which for one-dimensional ergodic diffusions is given explicitly by (1.15). In particular, (1.6) and (1.7) simplifies to

$$\mu_{\theta_0}(h(\theta_0)) = \int_\ell^r h(x;\theta_0)\mu_{\theta_0}(x)dx = 0 \tag{1.100}$$

and

$$W = \mu_{\theta_0}(\partial_{\theta^T} h(\theta_0)) = \int_\ell^r \partial_{\theta^T} h(x;\theta_0)\mu_{\theta_0}(x)dx.$$

The condition for eventual uniqueness of the G_n-estimator (1.9) is here that θ_0 is the only root of $\mu_{\theta_0}(h(\theta))$.

Kessler (2000) proposed

$$h(x;\theta) = \partial_\theta \log \mu_\theta(x), \tag{1.101}$$

which is the score function (the derivative of the log-likelihood function) if we pretend that the observations are an i.i.d. sample from the stationary distribution. If Δ is large, this might be a reasonable approximation. That (1.100) is satisfied for this specification of h follows under standard conditions that allow the interchange of differentiation and integration.

$$\int_\ell^r \left(\partial_\theta \log \mu_\theta(x)\right) \mu_\theta(x) dx = \int_\ell^r \partial_\theta \mu_\theta(x) dx = \partial_\theta \int_\ell^r \mu_\theta(x) dx = 0.$$

A modification of the simple estimating function (1.101) was shown by Kessler, Schick & Wefelmeyer (2001) to be efficient in the sense of semiparametric models. The modified version of the estimating function was derived by Kessler & Sørensen (2005) in a completely different way.

Hansen & Scheinkman (1995) and Kessler (2000) proposed and studied the generally applicable specification

$$h_j(x;\theta) = A_\theta f_j(x;\theta), \tag{1.102}$$

where A_θ is the generator (1.46), and f_j, $j = 1, \ldots, p$, are twice differentiable functions chosen such that Condition 1.2.1 holds. The estimating function with h given by (1.102) can easily be applied to multivariate diffusions, because an explicit expression for the invariant density μ_θ is not needed. The following lemma for one-dimensional diffusions shows that only weak conditions are needed to ensure that (1.100) holds for h_j given by (1.102).

Lemma 1.5.2 *Suppose $f \in C^2((\ell, r))$, $A_\theta f \in L^1(\mu_\theta)$ and*

$$\lim_{x \to r} f'(x)\sigma^2(x;\theta)\mu_\theta(x) = \lim_{x \to \ell} f'(x)\sigma^2(x;\theta)\mu_\theta(x). \tag{1.103}$$

Then

$$\int_\ell^r (A_\theta f)(x)\mu_\theta(x) dx = 0.$$

Proof: Note that by (1.15), the function $\nu(x;\theta) = \frac{1}{2}\sigma^2(x;\theta)\mu_\theta(x)$ satisfies that $\nu'(x;\theta) = b(x;\theta)\mu_\theta(x)$. In this proof all derivatives are with respect to x. It follows that

$$\begin{aligned} &\int_\ell^r (A_\theta f)(x)\mu_\theta(x) dx \\ &= \int_\ell^r \left(b(x;\theta)f'(x) + \tfrac{1}{2}\sigma^2(x;\theta)f''(x)\right) \mu_\theta(x) dx \\ &= \int_\ell^r \left(f'(x)\nu'(x;\theta) + f''(x)\nu(x;\theta)\right) dx = \int_\ell^r \left(f'(x)\nu(x;\theta)\right)' dx \\ &= \lim_{x \to r} f'(x)\sigma^2(x;\theta)\mu_\theta(x) - \lim_{x \to \ell} f'(x)\sigma^2(x;\theta)\mu_\theta(x) = 0. \end{aligned}$$

□

Example 1.5.3 Consider the square-root process (1.37) with $\sigma = 1$. For $f_1(x) = x$ and $f_2(x) = x^2$, we see that

$$A_\theta f(x) = \begin{pmatrix} -\beta(x-\alpha) \\ -2\beta(x-\alpha)x + x \end{pmatrix},$$

which gives the simple estimators

$$\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^{n} X_{i\Delta}, \qquad \hat{\beta}_n = \frac{\displaystyle\frac{1}{n}\sum_{i=1}^{n} X_{i\Delta}}{2\left(\displaystyle\frac{1}{n}\sum_{i=1}^{n} X_{i\Delta}^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_{i\Delta}\right)^2\right)}.$$

The condition (1.103) is obviously satisfied because the invariant distribution is a normal distribution.

□

Conley et al. (1997) proposed a model-based choice of the f_js in (1.102): $f_j = \partial_{\theta_j} \log \mu_\theta(x)$, i.e. the i.i.d. score function used in (1.101). Thus they obtained an estimating function of the form (1.99) with

$$h(x;\theta) = A_\theta \partial_\theta \log \mu_\theta(x). \tag{1.104}$$

Sørensen (2001) independently derived the same estimating function as an approximation to the score function for continuous-time observation of the diffusion process. Jacobsen (2001) showed that this estimating function is small Δ-optimal. This result was later rediscovered by Aït-Sahalia & Mykland (2008) who obtained a similar result for estimating functions given by (1.105).

An estimating function of the simple form (1.99) cannot be expected to yield as efficient estimators as an estimating function that depends on pairs of consecutive observations, and therefore can use the information contained in the transitions. Hansen & Scheinkman (1995) proposed non-martingale estimating functions of the form (1.12) with g given by

$$g_j(\Delta, x, y; \theta) = h_j(y) A_\theta f_j(x) - f_j(x) \hat{A}_\theta h_j(y), \tag{1.105}$$

where the functions f_j and h_j satisfy weak regularity conditions ensuring that (1.6) holds for $\bar{\theta} = \theta_0$. The differential operator $\hat{A}_\theta$ is the generator of the time reversal of the observed diffusion X. For a multivariate diffusion it is given by

$$\hat{A}_\theta f(x) = \sum_{k=1}^{d} \hat{b}_k(x;\theta)\partial_{x_k} f(x) + \tfrac{1}{2}\sum_{k,\ell=1}^{d} C_{k\ell}(x;\theta)\partial^2_{x_k x_\ell} f(x),$$

where $C = \sigma\sigma^T$ and

$$\hat{b}_k(x;\theta) = -b_k(x;\theta) + \frac{1}{\mu_\theta(x)} \sum_{\ell=1}^{d} \partial_{x_\ell} \left(\mu_\theta C_{kl}\right)(x;\theta).$$

For one-dimensional ergodic diffusions, $\hat{A}_\theta = A_\theta$. That $\hat{b} = b$ for a one-dimensional diffusion follows from (1.15). Obviously, the estimating function of the form (1.99) with $h_j(x;\theta) = A_\theta f_j(x)$ is a particular case of (1.105) with $h_j(y) = 1$.

1.5.3 Approximate martingale estimating functions

For martingale estimating functions of the form (1.30) and (1.40), we can always, as discussed in Subsection 1.3.3, obtain an explicit approximation to the optimal weight matrix by means of the expansion (1.47). For diffusion models where there is no explicit expression for the transition operator, it is tempting to go on and approximate the conditional moments $\pi_\Delta^\theta(f_j(\theta))(x)$ using (1.47), and thus, quite generally, obtain *explicit approximate martingale estimating function*. Such estimators were the first type of estimators for discretely observed diffusion processes to be studied in the literature. They have been considered by Dorogovcev (1976), Prakasa Rao (1988), Florens-Zmirou (1989), Yoshida (1992), Chan et al. (1992), Kloeden et al. (1996), Kessler (1997), Kelly, Platen & Sørensen (2004), and many othes.

It is, however, important to note that there is a dangerous pitfall when using these simple approximate martingale estimating functions. They do not satisfy the condition that $Q_{\theta_0}(g(\theta_0)) = 0$, and hence the estimators are inconsistent. To illustrate the problem, consider an estimating function of the form (1.12) with

$$g(x,y;\theta) = a(x,\theta)[f(y) - f(x) - \Delta A_\theta f(x)], \tag{1.106}$$

where A_θ is the generator (1.46), i.e., we have replaced $\pi_\Delta^\theta f(x)$ by a first order expansion. To simplify the exposition, we assume that θ, a and f are one-dimensional. We assume that the diffusion is geometrically α-mixing, that the other conditions mentioned above for the weak convergence result (1.5) hold, and that Condition 1.2.1 is satisfied. Then by Theorem 1.2.2, the estimator obtained using (1.106) converges to the solution, $\bar{\theta}$, of

$$Q_{\theta_0}(g(\bar{\theta})) = 0, \tag{1.107}$$

where, as usual, θ_0 is the true parameter value. We assume that the solution is unique. Using the expansion (1.47), we find that

$$\begin{aligned} Q_{\theta_0}(g(\theta)) &= \mu_{\theta_0}\Big(a(\theta)[\pi_\Delta^{\theta_0} f - f - \Delta A_\theta f]\Big) \\ &= \Delta\mu_{\theta_0}\left(a(\theta)[A_{\theta_0} f - A_\theta f + \tfrac{1}{2}\Delta A_{\theta_0}^2 f]\right) + O(\Delta^3) \end{aligned}$$

$$= \quad (\theta_0 - \theta)\Delta\mu_{\theta_0}\left(a(\theta_0)\partial_\theta A_{\theta_0} f\right) + \tfrac{1}{2}\Delta^2\mu_{\theta_0}\left(a(\theta_0)A^2_{\theta_0} f\right)$$
$$+O(\Delta|\theta - \theta_0|^2) + O(\Delta^2|\theta - \theta_0|) + O(\Delta^3).$$

If we neglect all O-terms, we obtain that

$$\bar{\theta} \doteq \theta_0 + \Delta\tfrac{1}{2}\mu_{\theta_0}\left(a(\theta_0)A^2_{\theta_0} f\right)/\mu_{\theta_0}\left(a(\theta_0)\partial_\theta A_{\theta_0} f\right),$$

which indicates that when Δ is small, the asymptotic bias is of order Δ. However, the bias can be huge when Δ is not sufficiently small as the following example shows.

Example 1.5.4 Consider again a diffusion with linear drift,

$$b(x;\theta) = -\beta(x - \alpha).$$

In this case (1.106) with $f(x) = x$ gives the estimating function

$$G_n(\theta) = \sum_{i=1}^{n} a(X_{\Delta(i-1)};\theta)[X_{\Delta i} - X_{\Delta(i-1)} + \beta\left(X_{\Delta(i-1)} - \alpha\right)\Delta],$$

where a is 2-dimensional. For a diffusion with linear drift, we found in Example 1.3.8 that

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

Using this, we obtain that

$$Q_{\theta_0}(g(\theta)) = c_1(e^{-\beta_0\Delta} - 1 + \beta\Delta) + c_2\beta(\alpha_0 - \alpha),$$

where

$$c_1 = \int_D a(x)x\mu_{\theta_0}(dx) - \mu_{\theta_0}(a)\alpha_0, \qquad c_2 = \mu_{\theta_0}(a)\Delta.$$

Thus

$$\bar{\alpha} = \alpha_0$$

and

$$\bar{\beta} = \frac{1 - e^{-\beta_0\Delta}}{\Delta} \leq \frac{1}{\Delta}.$$

We see that the estimator of α is consistent, while the estimator of β will tend to be small if Δ is large, whatever the true value β_0 is. We see that what determines how well $\hat{\beta}$ works is the magnitude of $\beta_0\Delta$, so it is not enough to know that Δ is small. Moreover, we cannot use $\hat{\beta}\Delta$ to evaluate whether there is a problem, because this quantity will always tend to be smaller than one. If $\beta_0\Delta$ actually is small, then the bias is proportional to Δ as expected

$$\bar{\beta} = \beta_0 - \tfrac{1}{2}\Delta\beta_0^2 + O(\Delta^2).$$

We get an impression of how terribly misled we can be when estimating the parameter β by means of the dangerous estimating function given by (1.106) from a simulation study in Bibby & Sørensen (1995) for the square root process (1.37). The result is given in Table 1.1. For the weight function a, the approximately optimal weight function was used, cf. Example 1.3.8. For different values of Δ and the sample size, 500 independent datasets were simulated, and the estimators were calculated for each dataset. The expectation of the estimator $\hat{\beta}$ was determined as the average of the simulated estimators. The true parameter values were $\alpha_0 = 10$, $\beta_0 = 1$ and $\tau_0 = 1$, and the initial value was $x_0 = 10$. When Δ is large, the behaviour of the estimator is bizarre. □

Δ	# obs.	mean	Δ	# obs.	mean
0.5	200	0.81	1.5	200	0.52
	500	0.80		500	0.52
	1000	0.79		1000	0.52
1.0	200	0.65	2.0	200	0.43
	500	0.64		500	0.43
	1000	0.63		1000	0.43

Table 1.1 *Empirical mean of 500 estimates of the parameter β in the CIR model. The true parameter values are $\alpha_0 = 10$, $\beta_0 = 1$, and $\tau_0 = 1$.*

The asymptotic bias given by (1.107) is small when Δ is sufficiently small, and the results in the following section on high frequency asymptotics show that in this asymptotic scenario the approximate martingale estimating functions work well. However, how small Δ needs to be depends on the parameter values, and without prior knowledge about the parameters, it is safer to use an exact martingale estimating function, which gives consistent estimators at all sampling frequencies.

1.6 High-frequency asymptotics

An expression for the asymptotic variance of estimators was obtained in Theorem 1.3.2 using a low frequency asymptotic scenario, where the time between observations is fixed. This expression is rather complicated and is not easy to use for comparing the efficiency of different estimators. Therefore the relative merits of estimators have often been investigated by simulation studies, and the general picture has been rather confusing. A much simpler and more manageable expression for the asymptotic variance of estimators can be obtained by considering the high frequency scenario,

$$n \to \infty, \qquad \Delta_n \to 0, \qquad n\Delta_n \to \infty. \tag{1.108}$$

The assumption that $n\Delta_n \to \infty$ is needed to ensure that parameters in the drift coefficient can be consistently estimated.

For this type of asymptotics Sørensen (2007) obtained simple conditions for rate optimality and efficiency for ergodic diffusions, which allow identification of estimators that work well when the time between observations, Δ_n, is not too large. How small Δ_n needs to be for the high frequency scenario to be relevant, depends on the speed with which the diffusion moves. For financial data the speed of reversion is usually slow enough that this type of asymptotics works for daily, sometimes even weekly observations. A main result of the theory in this section is that under weak conditions optimal martingale estimating functions give rate optimal and efficient estimators.

It is also interesting that the high frequency asymptotics provides a very clear statement of the important fact that parameters in the diffusion coefficient can be estimated more exactly than drift parameters when the time between observations is small. A final advantage of high frequency asymptotics is that it also gives useful results about the approximate martingale estimating functions discussed in Subsection 1.5.3, in situations where they work.

To simplify the exposition, we restrict attention to a one-dimensional diffusion given by

$$dX_t = b(X_t;\alpha)dt + \sigma(X_t;\beta)dW_t, \tag{1.109}$$

where $\theta = (\alpha, \beta) \in \Theta \subseteq \mathrm{I\!R}^2$. The results below can be generalized to multivariate diffusions and parameters of higher dimension. We consider estimating functions of the general form (1.3), where the two-dimensional function $g = (g_1, g_2)$ for some $\kappa \geq 2$ and for all $\theta \in \Theta$ satisfies

$$\begin{aligned} E_\theta(g(\Delta_n, X_{\Delta_n i}, X_{\Delta_n (i-1)};\theta) \mid X_{\Delta_n (i-1)}) & \\ = \Delta_n^\kappa R(\Delta_n, X_{\Delta_n (i-1)};\theta).& \end{aligned} \tag{1.110}$$

Martingale estimating functions obviously satisfy (1.110) with $R = 0$, but for instance the approximate martingale estimating functions discussed at the end of the previous section satisfy (1.110) too. Here and later $R(\Delta, y, x;\theta)$ denotes a function such that $|R(\Delta, y, x;\theta)| \leq F(y, x;\theta)$, where F is of polynomial growth in y and x uniformly for θ in compact sets. This means that for any compact subset $K \subseteq \Theta$, there exist constants $C_1, C_2, C_3 > 0$ such that $\sup_{\theta\in K} |F(y, x;\theta)| \leq C_1(1 + |x|^{C_2} + |y|^{C_3})$ for all x and y in the state space of the diffusion.

The main results in this section are simple conditions on the function $g(\Delta, y, x;\theta)$ that ensure rate optimality and efficiency of estimators. The condition for *rate optimality* is

Condition 1.6.1

$$\partial_y g_2(0, x, x;\theta) = 0 \tag{1.111}$$

for all $x \in (\ell, r)$ *and all* $\theta \in \Theta$.

By $\partial_y g_2(0, x, x; \theta)$ we mean $\partial_y g_2(0, y, x; \theta)$ evaluated at $y = x$. This condition is called *the Jacobsen condition*, because it was first found in the theory of small Δ-optimal estimation developed in Jacobsen (2001), cf. (1.62) in Subsection 1.3.4.

The condition for *efficiency* is

Condition 1.6.2

$$\partial_y g_1(0, x, x; \theta) = \partial_\alpha b(x; \alpha)/\sigma^2(x; \beta) \tag{1.112}$$

and

$$\partial_y^2 g_2(0, x, x; \theta) = \partial_\beta \sigma^2(x; \beta)/\sigma^4(x; \beta), \tag{1.113}$$

for all $x \in (\ell, r)$ *and all* $\theta \in \Theta$.

Also (1.112) and (1.113) were found as conditions for small Δ-optimality in Jacobsen (2002), cf. (1.61) and (1.63). This is not surprising. The following theorem provides an interpretation of small Δ-optimality in terms of the classical statistical concepts rate optimality and efficiency. As usual, $\theta_0 = (\alpha_0, \beta_0)$ denotes the true parameter value.

Theorem 1.6.3 *Assume that the diffusion is ergodic, that* $\theta_0 \in \text{int}\,\Theta$*, and that the technical regularity Condition 1.6.4 given below holds. Denote the density function of the invariant probability measure by* μ_θ*. Suppose that* $g(\Delta, y, x; \theta)$ *satisfies Condition 1.6.1. Assume, moreover, that the following identifiability condition is satisfied*

$$\begin{aligned} \int_\ell^r [b(x, \alpha_0) - b(x, \alpha)] \partial_y g_1(0, x, x; \theta) \mu_{\theta_0}(x) dx &\neq 0 \qquad \textit{when } \alpha \neq \alpha_0, \\ \int_\ell^r [\sigma^2(x, \beta_0) - \sigma^2(x, \beta)] \partial_y^2 g_2(0, x, x; \theta) \mu_{\theta_0}(x) dx &\neq 0 \qquad \textit{when } \beta \neq \beta_0, \end{aligned}$$

and that

$$\begin{aligned} S_1 &= \int_\ell^r \partial_\alpha b(x; \alpha_0) \partial_y g_1(0, x, x; \theta_0) \mu_{\theta_0}(x) dx \neq 0, \\ S_2 &= \tfrac{1}{2} \int_\ell^r \partial_\beta \sigma^2(x; \beta_0) \partial_y^2 g_2(0, x, x; \theta_0) \mu_{\theta_0}(x) dx \neq 0. \end{aligned}$$

Then a consistent G_n*–estimator* $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n)$ *exists and is unique in any compact subset of* Θ *containing* θ_0 *with probability approaching one as* $n \to \infty$*. If, moreover,*

$$\partial_\alpha \partial_y^2 g_2(0, x, x; \theta) = 0, \tag{1.114}$$

then for a martingale estimating function, and for more general estimating functions if $n\Delta^{2(\kappa-1)} \to 0$,

$$\begin{pmatrix} \sqrt{n\Delta_n}(\hat{\alpha}_n - \alpha_0) \\ \sqrt{n}(\hat{\beta}_n - \beta_0) \end{pmatrix} \xrightarrow{\mathcal{D}} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{W_1}{S_{11}^2} & 0 \\ 0 & \frac{W_2}{S_{22}^2} \end{pmatrix}\right) \tag{1.115}$$

where

$$\begin{aligned} W_1 &= \int_\ell^r \sigma^2(x;\beta_0)[\partial_y g_1(0,x,x;\theta_0)]^2 \mu_{\theta_0}(x)dx \\ W_2 &= \tfrac{1}{2}\int_\ell^r \sigma^4(x;\beta_0)[\partial_y^2 g_2(0,x,x;\theta_0)]^2 \mu_{\theta_0}(x)dx. \end{aligned}$$

Note that the estimator of the diffusion coefficient parameter, β, converges faster than the estimator of the drift parameter, α, and that the two estimators are asymptotically independent. Gobet (2002) showed, under regularity conditions, that a discretely sampled diffusion model is locally asymptotically normal under high frequency asymptotics, and that the optimal rate of convergence for a drift parameter is $1/\sqrt{n\Delta_n}$, while it is $1/\sqrt{n}$ for a parameter in the diffusion coefficient. Thus under the conditions of Theorem 1.6.3 the estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ are rate optimal. More precisely, Condition 1.6.1 implies rate optimality. If this condition is not satisfied, the estimator of the diffusion coefficient parameter, β, does not use the information about the diffusion coefficient contained in the quadratic variation and therefore converges at the same relatively slow rate $1/\sqrt{n\Delta_n}$ as estimators of α, see Sørensen (2007).

Gobet gave the following expression for the Fisher information matrix

$$\mathcal{I} = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}, \tag{1.116}$$

where

$$W_1 = \int_\ell^r \frac{(\partial_\alpha b(x;\alpha_0))^2}{\sigma^2(x;\beta_0)} \mu_{\theta_0}(x)dx, \tag{1.117}$$

$$W_2 = \int_\ell^r \left[\frac{\partial_\beta \sigma^2(x;\beta_0)}{\sigma^2(x;\beta_0)}\right]^2 \mu_{\theta_0}(x)dx. \tag{1.118}$$

By comparing the covariance matrix in (1.115) to (1.116), we see that Condition 1.6.2 implies that $S_1 = W_1$ and $S_2 = W_2$, with W_1 and W_2 given by (1.117) and (1.118), and that hence the asymptotic covariance matrix of $(\hat{\alpha}_n, \hat{\beta}_n)$ under Condition 1.6.2 equals the inverse of the Fisher information matrix (1.116). Thus Condition 1.6.2 ensures efficiency of $(\hat{\alpha}_n, \hat{\beta}_n)$. Under the conditions of Theorem 1.6.3 and Condition 1.6.2, we see that for a martingale estimating function, and more generally if $n\Delta^{2(\kappa-1)} \to 0$,

$$\begin{pmatrix} \sqrt{n\Delta_n}(\hat{\alpha}_n - \alpha_0) \\ \sqrt{n}(\hat{\beta}_n - \beta_0) \end{pmatrix} \xrightarrow{\mathcal{D}} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathcal{I}^{-1} \right). \tag{1.119}$$

Note that condition (1.114) is automatically satisfied under the efficiency Condition 1.6.2.

Proof of Theorem 1.6.3: Only a brief outline the proof is given; for details see Sørensen (2007). Consider the normalized estimating function

$$G_n(\theta) = \frac{1}{n\Delta_n} \sum_{i=1}^{n} g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta).$$

First the conditions of Theorem 1.10.2 must be checked. Using Lemma 9 in Genon-Catalot & Jacod (1993), it can be shown that $G_n(\theta_0) \to 0$ in P_{θ_0}-probability, and that $\partial_{\theta^T} G_n(\theta)$ under P_{θ_0} converges pointwise to a matrix, which for $\theta = \theta_0$ is upper triangular and has diagonal elements equal to S_1 and S_2, and thus is invertible. In order to prove that the convergence is uniform for θ in a compact set K, we show that the sequence

$$\zeta_n(\cdot) = \frac{1}{n\Delta_n} \sum_{i=1}^{n} g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}, \cdot)$$

converges weakly to the limit $\gamma(\cdot, \theta_0)$ in the space, $C(K)$, of continuous functions on K with the supremum norm. Since the limit is non-random, this implies uniform convergence in probability for $\theta \in K$. We have proved pointwise convergence, so the weak convergence result follows because the family of distributions of $\zeta_n(\cdot)$ is tight. The tightness is shown by checking the conditions in Corollary 14.9 in Kallenberg (1997). Thus the conditions of Theorem 1.10.2 are satisfied, and we conclude the existence of a consistent and eventually unique G_n-estimator. The uniqueness on compact subsets follows from Theorem 1.10.3 because the identifiability condition in Theorem 1.6.3 implies (1.160).

The asymptotic normality of the estimators follows from Theorem 1.10.4 with

$$A_n = \begin{pmatrix} \sqrt{\Delta_n n} & 0 \\ 0 & \sqrt{n} \end{pmatrix}.$$

The weak convergence of $A_n G_n(\theta_0)$ follows from a central limit theorem for martingales, e.g. Corollary 3.1 in Hall & Heyde (1980). The uniform convergence of $A_n \partial_{\theta^T} G_n(\theta) A_n^{-1}$ was proved for three of the entries when the conditions of Theorem 1.10.2 were checked. The result for the last entry is proved in a similar way using (1.114).

□

The reader is reminded of the trivial fact that for any non-singular 2×2 matrix, M_n, the estimating functions $M_n G_n(\theta)$ and $G_n(\theta)$ have exactly the same roots and hence give the same estimator(s). We call them *versions* of the same estimating function. The matrix M_n may depend on Δ_n. The point is that a version must exist which satisfies the conditions (1.111) – (1.113), but not all versions of an estimating function satisfy these conditions.

It follows from results in Jacobsen (2002) that to obtain a rate optimal and efficient estimator from an estimating function of the form (1.41), we need that $N \geq 2$ and that the matrix

$$D(x) = \begin{pmatrix} \partial_x f_1(x;\theta) & \partial_x^2 f_1(x;\theta) \\ \partial_x f_2(x;\theta) & \partial_x^2 f_2(x;\theta) \end{pmatrix}$$

is invertible for μ_θ-almost all x. Under these conditions, Sørensen (2007) showed that Godambe-Heyde optimal martingale estimating functions give rate optimal and efficient estimators. For a d-dimensional diffusion, Jacobsen (2002) gave the conditions $N \geq d(d+3)/2$, and that the $N \times (d+d^2)$-matrix $D(x) = \big(\partial_x f(x;\theta)\ \partial_x^2 f(x;\theta)\big)$ has full rank $d(d+3)/2$, which are needed to ensure the existence of a rate optimal and efficient estimator from an estimating function of the form (1.41).

We conclude this section by an example, but first we state technical conditions under which the results in this section hold. The assumptions about polynomial growth are far too strong, but simplify the proofs. These conditions can most likely be weakened considerably.

Condition 1.6.4 *The diffusion is ergodic with invariant probability density μ_θ, and the following conditions hold for all $\theta \in \Theta$:*

(1) $\int_\ell^r x^k \mu_\theta(x)dx < \infty$ *for all* $k \in \mathbb{N}$.

(2) $\sup_t E_\theta(|X_t|^k) < \infty$ *for all* $k \in \mathbb{N}$.

(3) $b, \sigma \in C_{p,4,1}((\ell, r) \times \Theta)$.

(4) There exists a constant C_θ such that for all $x, y \in (\ell, r)$

$$|b(x;\alpha) - b(y;\alpha)| + |\sigma(x;\beta) - \sigma(y;\beta)| \leq C_\theta |x-y|$$

(5) $g(\Delta, y, x; \theta) \in C_{p,2,6,2}(\mathbb{R}_+ \times (\ell, r)^2 \times \Theta)$ *and has an expansion in powers of Δ:*

$$g(\Delta, y, x; \theta) = \\ g(0, y, x; \theta) + \Delta g^{(1)}(y, x; \theta) + \tfrac{1}{2}\Delta^2 g^{(2)}(y, x; \theta) + \Delta^3 R(\Delta, y, x; \theta),$$

where

$$\begin{aligned} g(0, y, x; \theta) &\in C_{p,6,2}((\ell, r)^2 \times \Theta), \\ g^{(1)}(y, x; \theta) &\in C_{p,4,2}((\ell, r)^2 \times \Theta), \end{aligned}$$

$$g^{(2)}(y,x;\theta) \quad \in \quad C_{p,2,2}((\ell,r)^2 \times \Theta).$$

We define $C_{p,k_1,k_2,k_3}(\mathbb{R}_+ \times (\ell,r)^2 \times \Theta)$ as the class of real functions $f(t,y,x;\theta)$ satisfying that

(i) $f(t,y,x;\theta)$ is k_1 times continuously differentiable with respect t, k_2 times continuously differentiable with respect y, and k_3 times continuously differentiable with respect α and with respect to β

(ii) f and all partial derivatives $\partial_t^{i_1}\,\partial_y^{i_2}\,\partial_\alpha^{i_3}\,\partial_\beta^{i_4} f$, $i_j = 1,\ldots k_j$, $j = 1,2$, $i_3 + i_4 \leq k_3$, are of polynomial growth in x and y uniformly for θ in a compact set (for fixed t).

The classes $C_{p,k_1,k_2}((\ell,r)\times\Theta)$ and $C_{p,k_1,k_2}((\ell,r)^2\times\Theta)$ are defined similarly for functions $f(y;\theta)$ and $f(y,x;\theta)$, respectively.

Example 1.6.5 We can now interpret the findings in Example 1.3.11 as follows. The general quadratic martingale estimating function (1.64) gives rate optimal estimators in the high frequency asymptotics considered in this section. Moreover, the estimators are efficient in three particular cases: the optimal estimating function given in Example 1.3.6 and the approximations (1.28) and (1.51).

Kessler (1997) considered an approximation to the Gaussian quasi-likelihood presented in Subsection 1.3.2, where the conditional mean F and the conditional variance Φ are approximated as follows. The conditional mean is replaced by the expansion

$$r_k(\Delta,x;\theta) = \sum_{i=0}^{k} \frac{\Delta^i}{i!} A_\theta^i f(x) = x + \Delta \sum_{i=0}^{k-1} \frac{\Delta^i}{(i+1)!} A_\theta^i b(x;\alpha),$$

where $f(x) = x$, cf. (1.47). For fixed x, y and θ the function $(y - r_k(\Delta,x;\theta))^2$ is a polynomial in Δ of order $2k$. Define $g_{x,\theta}^j(y)$, $j = 0,1,\cdots,k$ by

$$(y - r_k(\Delta,x;\theta))^2 = \sum_{j=0}^{k} \Delta^j g_{x,\theta}^j(y) + O(\Delta^{k+1}).$$

For instance, for $k = 2$

$$(y - r_2(\Delta,x;\theta))^2 =$$
$$(y-x)^2 - 2(y-x)b(x;\alpha)\Delta + \left[(y-x)A_\theta b(x;\alpha) + b(x;\alpha)^2\right]\Delta^2 + O(\Delta^3),$$

from which we can see the expressions for $g_{x,\theta}^j(y)$, $j = 0,1,2$. The conditional variance can be approximated by

$$\Gamma_k(\Delta,x;\theta) = \sum_{j=0}^{k} \Delta^j \sum_{r=0}^{k-j} \frac{\Delta^r}{r!} A_\theta^r g_{x,\theta}^j(x).$$

In particular,

$$\Gamma_2(\Delta, x; \theta) = \Delta\sigma^2(x;\beta) + \tfrac{1}{2}\Delta^2 \left[A_\theta \sigma^2(x;\beta) - \sigma^2(x;\beta)\partial_x b(x;\alpha)\right].$$

By inserting these approximations in (1.28), we obtain the approximate martingale estimating function

$$\begin{aligned} H_n^{(k)}(\theta) \quad &= \quad \sum_{i=1}^{n} \frac{\partial_\theta r_k(\Delta_i, X_{t_{i-1}};\theta)}{\Gamma_{k+1}(\Delta_i, X_{t_{i-1}};\theta)}[X_{t_i} - r_k(\Delta_i, X_{t_{i-1}};\theta)] \qquad (1.120)\\ &+ \sum_{i=1}^{n} \frac{\partial_\theta \Gamma_{k+1}(\Delta_i, X_{t_{i-1}};\theta)}{2\Gamma_{k+1}(\Delta_i, X_{t_{i-1}};\theta)^2}[(X_{t_i} - r_k(\Delta_i, X_{t_{i-1}};\theta))^2 - \Gamma_{k+1}(\Delta_i, X_{t_{i-1}};\theta)]. \end{aligned}$$

Kessler (1997) (essentially) showed that for ergodic diffusions satisfying Condition 1.6.4 (1) – (4), the estimator obtained from $H_n^{(k)}(\theta)$ satisfies (1.119) provided that $n\Delta^{2k+1} \to 0$.

□

1.7 High-frequency asymptotics in a fixed time-interval

We will now briefly consider a more extreme type of high-frequence asymptotics, where the observation times are restricted to a bounded interval, which, without loss of generality, we can take to be $[0, 1]$. Suppose that the d-dimensional diffusion X which solves (1.109) has been observed at the time points $t_i = i/n$, $i = 0, \ldots, n$. Note that in this section W in equation (1.109) is a d-dimensional standard Wiener process, and σ is a $d \times d$-matrix. We assume that the matrix $C(x;\beta) = \sigma(x;\beta)\sigma(x;\beta)^T$ is invertible for all x in the state space, D, of X. Because the observation times are bounded, the drift parameter, α, cannot be consistently estimated as $n \to \infty$, so in the following we consider estimation of β only, and concentrate on the following Gaussian quasi-likelihood function:

$$Q_n(\beta) = \qquad (1.121)$$
$$\sum_{i=1}^{n} \left[\log\det C(X_{t_{i-1}};\beta) + n(X_{t_i} - X_{t_{i-1}})^T C(X_{t_{i-1}};\beta)^{-1}(X_{t_i} - X_{t_{i-1}})\right].$$

This is an approximation to a multivariate version of the Gaussian quasi-likelihood in Subsection 1.3.2 with $b = 0$, where the conditional mean $F(x;\theta)$ is approximated by x, and the conditional covariance matrix Φ is approximated by $n^{-1}C$. An estimator is obtained by minimizing $Q_n(\beta)$. This estimator can also be obtained from the approximate martingale estimating function which we get by differentiating $Q_n(\beta)$ with respect to β. The drift may be known, but in general we allow it to depend on an unknown parameter α. We assume that $\theta = (\alpha, \beta) \in A \times B = \Theta$, and we denote the true parameter value by $\theta_0 = (\alpha_0, \beta_0)$.

Genon-Catalot & Jacod (1993) showed the following theorem under the assumption that $\beta \in B$, where B is a compact subset of $\mathbb{R}^q$, which ensures that a $\hat{\beta}_n \in B$ that minimizes $Q_n(\beta)$ always exists. The results in the theorem hold for any $\hat{\beta}_n$ that minimizes $Q_n(\beta)$.

Theorem 1.7.1 *Assume that Condition 1.7.2 given below holds. Then the estimator $\hat{\beta}_n$ is consistent, and provided that $\beta_0 \in \operatorname{int} B$,*

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} Z,$$

where the distribution of Z is a normal variance mixture with characteristic function

$$s \mapsto E_{\theta_0}\left(\exp\left(-\tfrac{1}{2}s^T W(\beta_0)^{-1}s\right)\right)$$

with $W(\beta)$ given by (1.122). Conditional on $W(\beta_0)$, the asymptotic distribution of $\sqrt{n}(\hat{\beta}_n - \beta_0)$ is a centered q-dimensional normal distribution with covariate matrix $W(\beta_0)^{-1}$.

We will not prove Theorem 1.7.1 here. Note, however, that to do so we need the full generality of the Theorems 1.10.2, 1.10.3 and 1.10.4, where the matrix $W(\theta)$ (equal to $W_0(\theta)$ in Theorem 1.10.4) is random. Only if the matrix $B(x;\beta)$ defined below does not depend on x, is $W(\beta)$ non-random, in which case the limit distribution is simply the centered q-dimensional normal distribution with covariate matrix $W(\beta_0)^{-1}$. A simple example of a non-random $W(\beta)$ is when β is one-dimensional and a $q \times q$-matrix $F(x)$ exists such that $C(x;\beta) = \beta F(x)$. So for instance for the Ornstein-Uhlenbeck process and the square-root diffusion (1.37), $W(\beta)$ is non-random, and the limit distribution is normal.

Condition 1.7.2 *The stochastic differential equation (1.109) has a non-exploding, unique strong solution for $t \in [0,1]$, and the following conditions hold for all $\theta = (\alpha, \beta) \in \Theta$:*

(1) $b(x;\alpha)$ is a continuous function of x, and the partial derivatives $\partial_x^2 \sigma(x;\beta)$, $\partial_x \partial_\beta \sigma(x;\beta)$, $\partial_\beta^2 \sigma(x;\beta)$ exist and are continuous functions of $(x,\beta) \in D \times B$.

(2) With P_θ-probability one it holds that for all $\beta_1 \neq \beta$, the functions $t \mapsto C(X_t;\beta_1)$ and $t \mapsto C(X_t;\beta)$ are not equal.

(3) The random $q \times q$- matrix

$$W(\beta) = \int_0^1 B(X_t;\beta)dt, \tag{1.122}$$

where the ijth entry of $B(x;\beta)$ is given by

$$B(x;\beta)_{ij} = 2\,\mathrm{tr}\left(\partial_{\beta_i} C(x;\beta)C(x;\beta)^{-1}\partial_{\beta_j} C(x;\beta)C(x;\beta)^{-1}\right),$$

is invertible P_θ-almost surely.

The Condition 1.7.2 (2) can be difficult to check because it depends on the path of the process X. It is implied by the stronger condition that for all $\beta_1 \neq \beta$, $C(x;\beta_1) \neq C(x;\beta)$ for almost all $x \in D$.

Gobet (2001) showed, under regularity conditions, that for the high-frequency asymptotics in a fixed time-interval considered in this section, the diffusion model is locally asymptotically mixed normal (LAMN) with rate $\sqrt{n}$ and conditional variance given by $W(\beta)$; see e.g. Le Cam & Yang (2000) for the definition of LAMN. Therefore the estimator discussed above is efficient in the sense of Jeganathan (1982) and Jeganathan (1983).

Example 1.7.3 Consider the one-dimensional model given by

$$dX_t = -(X_t - \alpha)dt + \sqrt{\beta + X_t^2}dW_t,$$

where $\alpha > 0$ and $\beta > 0$. In this case $c(x;\beta) = \beta + x^2$, so

$$W(\beta) = \int_0^1 \frac{2X_t^4}{(\beta + X_t^2)^2}dt,$$

which is random.

□

1.8 Small-diffusion asymptotics

Under the high-frequency asymptotics with bounded observation times considered in the previous section, drift parameters could not be consistently estimated. Here we combine the high-frequency asymptotics with small-diffusion asymptotics to show that if the diffusion coefficient is small, we can find accurate estimators of drift parameters even when we have only observations in a bounded time-interval, which we again take to be $[0,1]$.

We consider observations that the time points $t_i = i/n$, $i = 1,\ldots,n$, of a d-dimensional diffusion process that solves the stochastic differential equation

$$dX_t = b(X_t,\alpha)dt + \varepsilon\sigma(X_t,\beta)dW_t, \quad X_0 = x_0, \tag{1.123}$$

with $\varepsilon > 0$ and $(\alpha,\beta) \in A \times B$, where $A \subseteq \mathbb{R}^{q_1}$ and $B \subseteq \mathbb{R}^{q_2}$ are convex, compact subsets. It is assumed that ϵ is known, while the parameter $\theta = (\alpha,\beta) \in \Theta = A \times B$ must be estimated. In (1.123) W is a d-dimensional standard Wiener process, and σ is a $d \times d$-matrix. We assume that the matrix $C(x;\beta) = \sigma(x;\beta)\sigma(x;\beta)^T$ is invertible for all x in the state space, D, of X.

In this section the asymptotic scenario is that $n \to \infty$ and $\varepsilon \to 0$ with a suitable balance between the rate of convergence of the two. Small diffusion asymptotics, where $\varepsilon \to 0$, has been widely studied and has proved fruitful in applied problems, see e.g. Freidlin & Wentzell (1998). Applications to contingent claim pricing and other financial problems can be found in Takahashi & Yoshida (2004) and Uchida & Yoshida (2004a), and applications to filtering problems in Picard (1986) and Picard (1991). The estimation problem outlined above was studied by Genon-Catalot (1990), Sørensen & Uchida (2003), and Gloter & Sørensen (2009). Here we follow Gloter & Sørensen (2009), which generalize results in the other papers, and consider the asymptotic scenario:

$$n \to \infty \qquad \varepsilon_n \to 0 \qquad \liminf_{n\to\infty} \varepsilon_n n^\rho > 0 \tag{1.124}$$

for some $\rho > 0$. When ρ is large, ϵ can go faster to zero than when ρ is relatively small. The value of ρ depends on the quasi-likelihood, as we shall see below.

The solution to (1.123) for $\epsilon = 0$ plays a crucial role in the theory. It is obviously non-random. More generally, we define the flow $\xi_t(x,\alpha)$ as the solution to the equation

$$\partial_t \xi_t(x,\alpha) = b(\xi_t(x,\alpha),\alpha), \qquad \xi_0(x,\alpha) = x, \tag{1.125}$$

for all $x \in D$. The solution to (1.123) for $\epsilon = 0$ is given by $\xi_t(x_0,\alpha)$. A related function of central importance is

$$\tilde{\delta}_n(x,\alpha) = \xi_{1/n}(x,\alpha) - x. \tag{1.126}$$

When ε is small, $\tilde{\delta}_n(X_{t_{i-1}},\alpha) + X_{t_{i-1}}$ approximates the conditional expectation of X_{t_i} given $X_{t_{i-1}}$, and can be used to define a Gaussian quasi-likelihood. However, equation (1.125) does not generally have en explicit solution, so $\xi_t(x,\alpha)$ is usually not explicitly available. Therefore we replace it by an approximation $\delta(x,\alpha)$ that satisfies Condition 1.8.2 (5) given below. Using this approximation, we define a Gaussian quasi-log-likelihood by

$$U_{\varepsilon,n}(\theta) = \sum_{k=1}^{n} \left\{ \log\det C_{k-1}(\beta) + \varepsilon^{-2} n P_k(\alpha)^T C_{k-1}(\beta)^{-1} P_k(\alpha) \right\}, \tag{1.127}$$

where

$$\begin{aligned} P_k(\alpha) &= X_{k/n} - X_{(k-1)/n} - \delta_n(X_{(k-1)/n},\alpha) \\ C_k(\beta) &= \sigma(X_{k/n},\beta)\sigma(X_{k/n},\beta)^T. \end{aligned}$$

This is the log-likelihood function that would have been obtained if the conditional distribution of X_{t_i} given $X_{t_{i-1}}$ were a normal distribution with mean $\delta_n(X_{t_{i-1}},\alpha) + X_{t_{i-1}}$ and covariance matrix $(t_i - t_{i-1})\varepsilon^2 C_{k-1}(\beta)$.

When ξ is explicitly available, a natural choice is $\delta_n(x,\alpha) = \tilde{\delta}_n(x,\alpha)$. Otherwise, simple useful approximations to $\tilde{\delta}_n(x,\alpha)$ are given by

$$\delta_n^k(x,\alpha) = \sum_{j=1}^{k} \frac{n^{-j}}{j!} (\mathcal{L}_\alpha)^{j-1} (b(\cdot,\alpha))(x),$$

$k = 1, 2 \ldots$, where the operator $\mathcal{L}_\alpha$ is defined by

$$\mathcal{L}_\alpha(f)(x) = \sum_{i=1}^{d} b_i(x,\alpha)\partial_{x_i} f(x).$$

By $(\mathcal{L}_\alpha)^j$ we denote j-fold application of the operator $\mathcal{L}_\alpha$. The approximation δ_n^k satisfies Conditions 1.8.2 (5)-(6), when $k - 1/2 \geq \rho$. The first two approximations are

$$\delta_n^1(x,\alpha) = n^{-1} b(x,\alpha),$$

for which the quasi-likelihood studied in Sørensen & Uchida (2003) is obtained, and

$$\delta_n^2(x,\alpha) = n^{-1} b(x,\alpha) + \tfrac{1}{2} n^{-2} \sum_{i=1}^{d} b_i(x,\alpha)\partial_{x_i} b(x,\alpha).$$

Since the parameter space Θ is compact, a $\hat{\theta}_{\varepsilon,n} = (\hat{\alpha}_{\varepsilon,n}, \hat{\beta}_{\varepsilon,n})$ that minimizes the Gaussian quasi- log-likelihood $U_{\varepsilon,n}(\theta)$ always exists. The results in the following theorem hold for any $\hat{\theta}_{\varepsilon,n}$ that minimizes $U_{\varepsilon,n}(\theta)$. As usual, $\theta_0 = (\alpha_0, \beta_0)$ denotes the true parameter value.

Theorem 1.8.1 *Assume that Condition 1.8.2 given below holds, that* $\theta_0 \in \operatorname{int}\Theta$*, and that the matrix*

$$I(\theta_0) = \begin{pmatrix} I_1(\theta_0) & 0 \\ 0 & I_2(\theta_0) \end{pmatrix}$$

is invertible, where the ij*th entries of the* $q_1 \times q_1$ *matrix* I_1 *and of the* $q_2 \times q_2$ *matrix* I_2 *are given by*

$$I_1^{i,j}(\theta_0) = \int_0^1 \partial_{\alpha_i} b(\xi_s(x_0,\alpha_0),\alpha_0)^T C^{-1}(\xi_s(x_0,\alpha_0),\beta_0)\partial_{\alpha_j} b(\xi_s(x_0,\alpha_0),\alpha_0) ds$$

and

$$I_\sigma^{i,j}(\theta_0) = \tfrac{1}{2} \int_0^1 \operatorname{tr}\left[(\partial_{\beta_i} C)C^{-1}(\partial_{\beta_j} C)C^{-1}(\xi_s(x_0,\alpha_0),\beta_0)\right] ds.$$

Then, under the asymptotic scenario (1.124), the estimator $\hat{\theta}_{\varepsilon,n}$ *is consistent, and*

$$\begin{pmatrix} \varepsilon^{-1}(\hat{\alpha}_{\varepsilon,n} - \alpha_0) \\ \sqrt{n}(\hat{\beta}_{\varepsilon,n} - \beta_0) \end{pmatrix} \xrightarrow{\mathcal{D}} N\left(0, I(\theta_0)^{-1}\right).$$

We do not prove the theorem here, but a similar result for the estimating function obtained by differentiation of $U_{\varepsilon,n}(\theta)$ with respect to θ can be proved using the asymptotic results in Section 1.10. Note that the estimators of the drift and diffusion coefficient parameters are asymptotically independent. The two parameters are not estimated at the same rate. For the approximation δ_n^1 the conditions below are satisfied if ϵ^{-1} converges at a rate smaller than or equal to $\sqrt{n}$, so in this case the rate of convergence of $\hat{\alpha}_{\varepsilon,n}$ is slower than or equal to that of $\hat{\beta}_{\varepsilon,n}$. For the approximations δ_n^k, $k \geq 2$, the rate of convergence of $\hat{\alpha}_{\varepsilon,n}$ can be slower than or faster than that of $\hat{\beta}_{\varepsilon,n}$, dependent on how fast ϵ goes to zero.

The matrix I_1 equals the Fisher information matrix when the data is a continuous sample path in $[0,1]$ and $\varepsilon \to 0$, cf. Kutoyants (1994), so $\hat{\alpha}_{\varepsilon,n}$ is efficient. Probably $\hat{\beta}_{\varepsilon,n}$ is efficient too, but this cannot be seen in this simple way and has not yet been proved.

We now give the technical conditions that imply Theorem 1.8.1.

Condition 1.8.2 *The following holds for all $\varepsilon > 0$:*

(1) The stochastic differential equation (1.123) has a unique strong solution for $t \in [0,1]$ for all $\theta = (\alpha, \beta) \in \Theta$.

(2) $b(x;\alpha)$ is a smooth (i.e. $\mathcal{C}^\infty$) function of (x,α), and a constant c exists such that for all $x, y \in D$ and all $\alpha_1, \alpha_2 \in A$:

$$|b(x;\alpha_1) - b(y;\alpha_2)| \leq c(|x-y| + |\alpha_1 - \alpha_2|).$$

(3) $\sigma(x;\beta)$ is continuous, and there exists an open convex subset $\mathcal{U} \subseteq D$ such that $\xi_t(x_0,\alpha_0) \in \mathcal{U}$ for all $t \in [0,1]$, and $\sigma(x;\beta)$ is smooth on $\mathcal{U} \times B$.

(4) If $\alpha \neq \alpha_0$, then the two functions $t \mapsto b(\xi_t(x_0,\alpha_0);\alpha)$ and $t \mapsto b(\xi_t(x_0,\alpha_0);\alpha_0)$ are not equal. If $\beta \neq \beta_0$, then the two functions $t \mapsto C(\xi_t(x_0,\alpha_0);\beta)$ and $t \mapsto C(\xi_t(x_0,\alpha_0);\beta_0)$ are not equal.

(5) The function $\delta_n(x;\alpha)$ is smooth, and for any compact subset $K \subseteq D$, a constant $c(K)$ exists such that

$$\sup_{x\in K,\alpha\in A} \left|\delta_n(x;\alpha) - \tilde{\delta}_n(x;\alpha)\right| \leq c(K)\varepsilon n^{-3/2}.$$

Similar bounds hold for the first two derivatives of δ_n w.r.t. α.

(6) For any compact subset $K \subseteq D \times A$, there exists a constant $c(K)$, independent of n, such that

$$|n\delta_n(x;\alpha_1) - n\delta_n(x;\alpha_2)| \leq c(K)|\alpha_1 - \alpha_2|$$

for all $(x,\alpha_1),(x,\alpha_2) \in K$ and for all $n \in \mathbb{N}$. The same holds for derivatives of any order w.r.t. α of $n\delta_n$.

It can be shown that $\delta_n(x,\alpha) = \tilde{\delta}_n(x,\alpha)$ satisfies Condition 1.8.2 (6), under Condition 1.8.2 (2). This choice of δ_n trivially satisfies Condition 1.8.2 (5).

Example 1.8.3 Consider the two dimensional diffusion $X = (Y, R)$ given by

$$\begin{aligned} dY_t &= (R_t + \mu_1)dt + \varepsilon\kappa_1 dW_t^1 \\ dR_t &= -\mu_2(R_t - m)dt + \varepsilon\kappa_2\sqrt{R_t}\left(\rho dW_t^1 + \sqrt{1-\rho^2}dW_t^2\right), \end{aligned}$$

where $(Y_0, R_0) = (y_0, r_0)$ with $r_0 > 0$. This model was used in finance by Longstaff & Schwartz (1995). In their mode, the second component represents the short term interest rate, while Y is the logarithm of the price of some asset. The second component is the square-root diffusion. The parameters are $\theta = (\alpha, \beta)$, where $\alpha = (\mu_1, \mu_2, m)$ and $\beta = (\kappa_1^2, \kappa_2^2, \rho)$. The parameter ρ allows correlation between the innovation terms of the two coordinates. The diffusion process (Y, R) satisfies Condition 1.8.2 (1) – (3), and (4) is holds if $r_0 \neq m_0$. The equation (1.125) is linear and has the solution

$$\xi_t(y, r, \mu_1, \mu_2, m) = \begin{pmatrix} y + (\mu_1 + m)t + \mu_2^{-1}(r - m)(1 - e^{-\mu_2 t}) \\ m + (r - m)e^{-\mu_2 t} \end{pmatrix}.$$

Therefore we can choose $\delta_n(x,\alpha) = \tilde{\delta}_n(x,\alpha)$, which satisfies Condition 1.8.2 (5) – (6). The matric $I(\theta_0)$ is invertible when $r_0 \neq m_0$ and is given by

$$I_1(\theta) = (1-\rho^2)^{-1} \begin{pmatrix} \kappa_1^{-2} & 0 & \\ 0 & \frac{-m(\mu_2 + \log(q)) + (m - r_0)(e^{-\mu_2} - 1))}{\kappa_2^2 \mu_2} & \frac{-\mu_1 + \log(q)}{\kappa_2^2} \\ 0 & \frac{-\mu_1 + \log(q)}{\kappa_2^2} & -\frac{\mu_2 \log(q)}{m\kappa_2^2}, \end{pmatrix},$$

where $q = r_0/(r_0 + m(e^{\mu_2} - 1))$, and

$$I_2(\theta) = \begin{pmatrix} 2\kappa_1^4 & 2\rho^2\kappa_1^2\kappa_2^2 & \rho(1-\rho^2)\kappa_1^2 \\ \rho^2\kappa_1^2\kappa_2^2 & 2\kappa_2^4 & \rho(1-\rho^2)\kappa_2^2 \\ \rho(1-\rho^2)\kappa_1^2 & \rho(1-\rho^2)\kappa_2^2 & (1-\rho^2)^2 \end{pmatrix}.$$

Note that the asymptotic variance of the estimators of the drift parameter goes to zero, as the correlation parameter ρ goes to one.

□

Several papers have studied other aspects of small diffusion asymptotics for estimators of parameters in diffusion models. First estimation of the parameter α based on a continuously observed sample path of the diffusion process was considered by Kutoyants (1994). Semiparametric estimation for the same type of data was studied later by Kutoyants (1998) and Iacus & Kutoyants (2001). Information criteria were investigated by Uchida & Yoshida (2004b). Uchida

(2004) and Uchida (2008) studied approximations to martingale estimating functions for discretely sampled diffusions under small diffusion asymptotics. Martingale estimating functions were studied by Sørensen (2000b) under an extreme type of small diffusion asymptotics where n is fixed.

1.9 Non-Markovian models

In this section we consider estimating functions that can be used when the observed process is not a Markov process. In this situation, it is usually not easy to find a tractable martingale estimating function. For instance a simple estimating function of the form (1.41) is not a martingale. To obtain a martingale, the conditional expectation given $X_{(i-1)\Delta}$ in (1.41) must be replaced by the conditional expectation given all previous observations, which can only very rarely be found explicitly, and which it is rather hopeless to find by simulation. Instead we will consider a generalization of the martingale estimating functions, called the prediction-based estimating functions, which can be interpreted as approximations to martingale estimating functions.

To clarify our thoughts, we will consider a concrete model type. Let the D-dimensional process X be the stationary solution to the stochastic differential equation

$$dX_t = b(X_t;\theta)dt + \sigma(X_t;\theta)dW_t, \tag{1.128}$$

where b is D-dimensional, σ is a $D \times D$-matrix, and W a D-dimensional standard Wiener process. As usual the parameter θ varies in a subset Θ of IR^p. However, we do not observed X directly. What we observe is

$$Y_i = k(X_{t_i}) + Z_i, \quad i = 1, \ldots, n, \tag{1.129}$$

where k maps IR^D into IR^d ($d < D$), and $\{Z_i\}$ is a sequence of independent identically distributed measurement errors with mean zero. We assume that the measurement errors are independent of the process X. Obviously, the discrete time process $\{Y_i\}$ is not a Markov-process.

1.9.1 Prediction-based estimating functions

In the following we will outline the method of prediction-based estimating functions introduced in Sørensen (2000a). Assume that $f_j, j = 1, \ldots, N$, are functions that map $\mathrm{IR}^{s+1} \times \Theta$ into IR such that $E_\theta(f_j(Y_{s+1}, \ldots, Y_1;\theta)^2) < \infty$ for all $\theta \in \Theta$. Let $\mathcal{P}^\theta_{i-1,j}$ be a closed linear subset of the L_2-space, L^θ_{i-1}, of all functions of $Y_1, \ldots, Y_{i-1}$ with finite variance under P_θ. The set $\mathcal{P}^\theta_{i-1,j}$ can be interpreted as a set of predictors of $f_j(Y_i, \ldots, Y_{i-s};\theta)$ based on $Y_1, \ldots, Y_{i-1}$.

A prediction-based estimating function has the form

$$G_n(\theta) = \sum_{i=s+1}^{n} \sum_{j=1}^{N} \Pi_j^{(i-1)}(\theta) \left[f_j(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}_j^{(i-1)}(\theta) \right],$$

where $\Pi_j^{(i-1)}(\theta)$ is a p-dimensional vector, the coordinates of which belong to $\mathcal{P}_{i-1,j}^{\theta}$, and $\breve{\pi}_j^{(i-1)}(\theta)$ is the minimum mean square error predictor in $\mathcal{P}_{i-1,j}^{\theta}$ of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ under P_θ. When $s = 0$ and $\mathcal{P}_{i-1,j}^{\theta}$ is the set of all functions of $Y_1, \ldots, Y_{i-1}$ with finite variance, then $\breve{\pi}_j^{(i-1)}(\theta)$ is the conditional expectation under P_θ of $f_j(Y_i; \theta)$ given $Y_1, \ldots, Y_{i-1}$, so in this case we obtain a martingale estimating function. Thus for a Markov process, a martingale estimating function of the form (1.41) is a particular case of a prediction-based estimating function.

The minimum mean square error predictor in $\mathcal{P}_{i-1,j}^{\theta}$ of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ is the projection in L_{i-1}^{θ} of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ onto the subspace $\mathcal{P}_{i-1,j}$. Therefore $\breve{\pi}_j^{(i-1)}(\theta)$ satisfies the normal equation

$$\mathrm{E}_\theta \left(\pi_j^{(i-1)} \left[f_j(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}_j^{(i-1)}(\theta) \right] \right) = 0 \qquad (1.130)$$

for all $\pi_j^{(i-1)} \in \mathcal{P}_{i-1,j}^{\theta}$. This implies that a prediction-based estimating function satisfies that

$$\mathrm{E}_\theta \left(G_n(\theta) \right) = 0. \qquad (1.131)$$

We can interpret the minimum mean square error predictor as an approximation to the conditional expectation of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ given $X_1, \ldots, X_{i-1}$, which is the projection of $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ onto the subspace of all functions of $X_1, \ldots, X_{i-1}$ with finite variance.

To obtain estimators that can relatively easily be calculated in practice, we will from now on restrict attention to predictor sets, $\mathcal{P}_{i-1,j}^{\theta}$, that are finite dimensional. Let $h_{jk}, j = 1, \ldots, N,\ k = 0, \ldots, q_j$ be functions from $\mathbb{R}^r$ into $\mathbb{R}$ ($r \geq s$), and define (for $i \geq r+1$) random variables by

$$Z_{jk}^{(i-1)} = h_{jk}(Y_{i-1}, Y_{i-2}, \ldots, Y_{i-r}).$$

We assume that $E_\theta((Z_{jk}^{(i-1)})^2) < \infty$ for all $\theta \in \Theta$, and let $\mathcal{P}_{i-1,j}$ denote the subspace spanned by $Z_{j0}^{(i-1)}, \ldots, Z_{jq_j}^{(i-1)}$. We set $h_{j0} = 1$ and make the natural assumption that the functions $h_{j0}, \ldots, h_{jq_j}$ are linearly independent. We write the elements of $\mathcal{P}_{i-1,j}$ in the form $a^T Z_j^{(i-1)}$, where $a^T = (a_0, \ldots, a_{q_j})$ and

$$Z_j^{(i-1)} = \left(Z_{j0}^{(i-1)}, \ldots, Z_{jq_j}^{(i-1)} \right)^T$$

are $(q_j + 1)$-dimensional vectors. With this specification of the predictors, the estimating function can only include terms with $i \geq r+1$:

$$G_n(\theta) = \sum_{i=r+1}^{n} \sum_{j=1}^{N} \Pi_j^{(i-1)}(\theta) \left[f_j(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}_j^{(i-1)}(\theta) \right]. \quad (1.132)$$

It is well-known that the minimum mean square error predictor, $\breve{\pi}_j^{(i-1)}(\theta)$, is found by solving the normal equations (1.130). Define $C_j(\theta)$ as the covariance matrix of $(Z_{j1}^{(r)}, \ldots, Z_{jq_j}^{(r)})^T$ under P_θ, and $b_j(\theta)$ as the vector for which the ith coordinate is

$$b_j(\theta)_i = \mathrm{Cov}_\theta(Z_{ji}^{(r)}, f_j(Y_{r+1}, \ldots, Y_{r+1-s}; \theta)), \quad (1.133)$$

$i = 1, \ldots, q_j$. Then we have

$$\breve{\pi}_j^{(i-1)}(\theta) = \breve{a}_j(\theta)^T Z_j^{(i-1)}, \quad (1.134)$$

where $\breve{a}_j(\theta)^T = (\breve{a}_{j0}(\theta), \breve{a}_{j*}(\theta)^T)$ with

$$\breve{a}_{j*}(\theta) = C_j(\theta)^{-1} b_j(\theta) \quad (1.135)$$

and

$$\breve{a}_{j0}(\theta) = E_\theta(f_j(Y_{s+1}, \ldots, Y_1; \theta)) - \sum_{k=1}^{q_j} \breve{a}_{jk}(\theta) E_\theta(Z_{jk}^{(r)}). \quad (1.136)$$

That $C_j(\theta)$ is invertible follows from the assumption that the functions h_{jk} are linearly independent. If $f_j(Y_i, \ldots, Y_{i-s}; \theta)$ has mean zero under P_θ for all $\theta \in \Theta$, we need not include a constant in the space of predictors, i.e. we need only the space spanned by $Z_{j1}^{(i-1)}, \ldots, Z_{jq_j}^{(i-1)}$.

Example 1.9.1 An important particular case when $d = 1$ is $f_j(y) = y^j$, $j = 1, \ldots, N$. For each $i = r+1, \ldots, n$ and $j = 1, \ldots, N$, we let $\{Z_{jk}^{(i-1)} \mid k = 0, \ldots, q_j\}$ be a subset of $\{Y_{i-\ell}^\kappa \mid \ell = 1, \ldots, r, \kappa = 0, \ldots, j\}$, where $Z_{j0}^{(i-1)}$ is always equal to 1. Here we need to assume that $\mathrm{E}_\theta(Y_i^{2N}) < \infty$ for all $\theta \in \Theta$. To find $\breve{\pi}_j^{(i-1)}(\theta)$, $j = 1, \ldots, N$, by means of (1.135) and (1.136), we must calculate moments of the form

$$\mathrm{E}_\theta(Y_1^\kappa Y_k^j), \quad 0 \leq \kappa \leq j \leq N, \quad k = 1, \ldots, r. \quad (1.137)$$

To avoid the matrix inversion in (1.135), the vector of coefficients $\breve{a}_j$ can be found by means of the N-dimensional Durbin-Levinson algorithm applied to the process $\{(Y_i, Y_i^2, \ldots, Y_i^N)\}_{i \in \mathbb{N}}$, see Brockwell & Davis (1991). Suppose the diffusion process X is exponentially ρ-mixing, see Doukhan (1994) for a definition. This is for instance the case for a Pearson diffusion (see Subsection 1.3.7) or for a one-dimensional diffusion that satisfies Condition 1.5.1. Then the observed process Y inherits this property, which implies that constants

$K > 0$ and $\lambda > 0$ exist such that $|\mathrm{Cov}_\theta(Y_1^j, Y_k^j)| \leq Ke^{-\lambda k}$. Therefore a small value of r can usually be used.

In many situations it is reasonable to choose $N = 2$ with the following simple predictor sets where $q_1 = r$ and $q_2 = 2r$. The predictor sets are generated by $Z_{j0}^{(i-1)} = 1$, $Z_{jk}^{(i-1)} = Y_{i-k}$, $k = 1, \ldots, r, j = 1, 2$ and $Z_{2k}^{(i-1)} = Y_{i+r-k}^2$, $k = r+1, \ldots, 2r$. In this case the minimum mean square error predictor of Y_i can be found using the Durbin-Levinson algorithm for real processes, while the predictor of Y_i^2 can be found by applying the two-dimensional Durbin-Levinson algorithm to the process (Y_i, Y_i^2). Including predictors in the form of lagged terms $Y_{i-k}Y_{i-k-l}$ for a number of lags l's might also be of relevance.

We illustrate the use of the Durbin-Levinson algorithm in the simplest possible case, where $N = 1$, $f(x) = x$, $Z_0^{(i-1)} = 1$, $Z_k^{(i-1)} = Y_{i-k}$, $k = 1, \ldots, r$. We suppress the superfluous j in the notation. Let $K_\ell(\theta)$ denote the covariance between Y_1 and $Y_{\ell+1}$ under P_θ, and define $\phi_{1,1}(\theta)) = K_1(\theta)/K_0(\theta)$ and $v_0(\theta) = K_0(\theta)$. Then the Durbin-Levinson algorithm works as follows

$$\phi_{\ell,\ell}(\theta) = \left(K_\ell(\theta) - \sum_{k=1}^{\ell-1} \phi_{\ell-1,k}(\theta) K_{\ell-k}(\theta) \right) v_{\ell-1}(\theta)^{-1},$$

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

and

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

The algorithm is run for $\ell = 2, \ldots, r$. Then

$$\breve{a}_*(\theta) = (\phi_{r,1}(\theta), \ldots, \phi_{r,r}(\theta)),$$

while $\breve{a}_0$ can be found from (1.136), which here simplifies to

$$\breve{a}_0(\theta) = E_\theta(Y_1) \left(1 - \sum_{k=1}^{r} \phi_{r,k}(\theta) \right).$$

The quantity $v_r(\theta)$ is the prediction error $\mathrm{E}_\theta\left((Y_i - \breve{\pi}^{(i-1)})^2\right)$. Note that if we want to include a further lagged value of Y in the predictor, we just iterate the algorithm once more.

□

We will now find the optimal prediction-based estimating function of the form (1.132) in the sense explained in Section 1.11. First we express the estimating function in a more compact way. The ℓth coordinate of the vector $\Pi_j^{(i-1)}(\theta)$ can be written as

$$\pi_{\ell,j}^{(i-1)}(\theta) = \sum_{k=0}^{q_j} a_{\ell jk}(\theta) Z_{jk}^{(i-1)}, \quad \ell = 1, \ldots, p.$$

With this notation, (1.132) can be written in the form

$$G_n(\theta) = A(\theta) \sum_{i=r+1}^{n} H^{(i)}(\theta), \tag{1.138}$$

where

$$A(\theta) = \begin{pmatrix} a_{110}(\theta) & \cdots & a_{11q_1}(\theta) & \cdots & \cdots & a_{1N0}(\theta) & \cdots & a_{1Nq_N}(\theta) \\ \vdots & & \vdots & & & \vdots & & \vdots \\ a_{p10}(\theta) & \cdots & a_{p1q_1}(\theta) & \cdots & \cdots & a_{pN0}(\theta) & \cdots & a_{pNq_N}(\theta) \end{pmatrix},$$

and

$$H^{(i)}(\theta) = Z^{(i-1)} \left(F(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}^{(i-1)}(\theta) \right), \tag{1.139}$$

with $F = (f_1, \ldots, f_N)^T$, $\breve{\pi}^{(i-1)}(\theta) = (\breve{\pi}_1^{(i-1)}(\theta), \ldots, \breve{\pi}_N^{(i-1)}(\theta))^T$, and

$$Z^{(i-1)} = \begin{pmatrix} Z_1^{(i-1)} & 0_{q_1} & \cdots & 0_{q_1} \\ 0_{q_2} & Z_2^{(i-1)} & \cdots & 0_{q_2} \\ \vdots & \vdots & & \vdots \\ 0_{q_N} & 0_{q_N} & \cdots & Z_N^{(i-1)} \end{pmatrix}. \tag{1.140}$$

Here 0_{q_j} denotes the q_j-dimensional zero-vector. When we have chosen the functions f_j and the predictor spaces, the quantities $H^{(i)}(\theta)$ are completely determined, whereas we are free to choose the matrix $A(\theta)$ in an optimal way, i.e. such that the asymptotic variance of the estimators is minimized.

We will find en explicit expression for the optimal weight matrix, $A^*(\theta)$, under the following condition, in which we need one further definition:

$$\breve{a}(\theta) = (\breve{a}_{10}(\theta), \ldots, \breve{a}_{1q_1}(\theta), \ldots, \breve{a}_{N0}(\theta), \ldots \breve{a}_{Nq_N}(\theta))^T, \tag{1.141}$$

where the quantities $\breve{a}_{jk}$s define the minimum mean square error predictors, cf. (1.134).

Condition 1.9.2
(1) *The function $F(y_1, \ldots, y_{s+1}; \theta)$ and the coordinates of $\breve{a}(\theta)$ are continuously differentiable functions of θ.*

(2) $p \leq \bar{p} = N + q_1 + \cdots + q_N$.

(3) *The* $\bar{p} \times p$*-matrix* $\partial_{\theta^T} \breve{\alpha}(\theta)$ *has rank* p.

(4) *The functions* $1, f_1, \ldots, f_N$ *are linearly independent (for fixed* θ*) on the support of the conditional distribution of* $(Y_i, \ldots, Y_{i-s})$ *given* $(X_{i-1}, \ldots, X_{i-r})$.

(5) *The* $p \times p$*-matrix*

$$U(\theta)^T = E_\theta \left(Z^{(i-1)} \partial_{\theta^T} F(Y_i, \ldots, Y_{i-s}; \theta) \right) \tag{1.142}$$

exists.

If we denote the optimal prediction-based estimating function by $G_n^*(\theta)$, then

$$\mathrm{E}_\theta \left(G_n(\theta) G_n^*(\theta)^T \right) = (n-r) A(\theta) \bar{M}_n(\theta) A_n^*(\theta)^T,$$

where

$$\begin{aligned} \bar{M}_n(\theta) = {} & \mathrm{E}_\theta \left(H^{(r+1)}(\theta) H^{(r+1)}(\theta)^T \right) \\ & + \sum_{k=1}^{n-r-1} \frac{(n-r-k)}{(n-r)} \left\{ \mathrm{E}_\theta \left(H^{(r+1)}(\theta) H^{(r+1+k)}(\theta)^T \right) \right. \\ & \left. + \mathrm{E}_\theta \left(H^{(r+1+k)}(\theta) H^{(r+1)}(\theta)^T \right) \right\}, \end{aligned} \tag{1.143}$$

which is the covariance matrix of $\sum_{i=r+1}^n H^{(i)}(\theta)/\sqrt{n-r}$. The sensitivity function (1.166) is given by

$$S_{G_n}(\theta) = (n-r) A(\theta) \left[U(\theta)^T - D(\theta) \partial_{\theta^T} \breve{a}(\theta) \right],$$

where $D(\theta)$ is the $\bar{p} \times \bar{p}$-matrix

$$D(\theta) = \mathrm{E}_\theta \left(Z^{(i-1)} (Z^{(i-1)})^T \right). \tag{1.144}$$

It follows from Theorem 1.11.1 that $A_n^*(\theta)$ is optimal if $\mathrm{E}_\theta \left(G_n(\theta) G_n^*(\theta)^T \right) = S_{G_n}(\theta)$. Under Condition 1.9.2 (4) the matrix $\bar{M}_n(\theta)$ is invertible, see Sørensen (2000a), so it follows that

$$A_n^*(\theta) = (U(\theta) - \partial_\theta \breve{a}(\theta)^T D(\theta)) \bar{M}_n(\theta)^{-1}, \tag{1.145}$$

so that the estimating function

$$G_n^*(\theta) = A_n^*(\theta) \sum_{i=s+1}^n Z^{(i-1)} \left(F(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}^{(i-1)}(\theta) \right), \tag{1.146}$$

is Godambe optimal. When the function F does not depend on θ, the expression for $A_n^*(\theta)$ simplifies slightly as in this case $U(\theta) = 0$.

Example 1.9.3 Consider again the type of prediction-based estimating function discussed in Example 1.9.1. In order to calculate (1.143), we need mixed moments of the form

$$\mathrm{E}_\theta[Y_{t_1}^{k_1} Y_{t_2}^{k_2} Y_{t_3}^{k_3} Y_{t_4}^{k_4}], \tag{1.147}$$

for $t_1 \leq t_2 \leq t_3 \leq t_4$ and $k_1 + k_2 + k_3 + k_4 \leq 4N$, where $k_i,\ i = 1, \ldots, 4$ are non-negative integers.

□

1.9.2 Asymptotics

A prediction-based estimating function of the form (1.138) gives consistent and asymptotically normal estimators under the following condition, where θ_0 is the true parameter value.

Condition 1.9.4
(1) *The diffusion process X is stationary and geometrically α-mixing.*

(2) *There exists a $\delta > 0$ such that*

$$E_{\theta_0}\left(\left|Z_{jk}^{(r)} f_j(X_{r+1}, \ldots, X_{r+1-s}; \theta_0)\right|^{2+\delta}\right) < \infty$$

and

$$E_{\theta_0}\left(\left|Z_{jk}^{(r)} Z_{j\ell}^{(r)}\right|^{2+\delta}\right) < \infty,$$

for $j = 1, \ldots, N,\ k, \ell = 0, \ldots q_j$.

(3) *The function $F(y_1, \ldots, y_{s+1}; \theta)$ and the components of $A(\theta)$ and $\breve{a}(\theta)$, given by (1.141) are continuously differentiable functions of θ.*

(4) *The matrix $W = A(\theta_0)(U(\theta_0) - D(\theta_0)\partial_{\theta^T}\breve{a}(\theta_0))$ has full rank p. The matrices $U(\theta)$ and $D(\theta)$ are given by* (1.142) *and* (1.144).

(5)

$$A(\theta)\left(E_{\theta_0}\left(Z^{(i-1)} F(Y_i, \ldots, Y_{i-s}; \theta)\right) - D(\theta_0)\partial_{\theta^T}\breve{a}(\theta))\right) \neq 0$$

for all $\theta \neq \theta_0$.

Condition 1.9.4 (1) and (2) ensures that the central limit theorem (1.5) holds and that $\bar{M}_n(\theta_0) \rightarrow M(\theta_0)$, where

$$M(\theta) = \mathrm{E}_\theta\left(H^{(r+1)}(\theta) H^{(r+1)}(\theta)^T\right)$$

$$+\sum_{k=1}^{\infty}\left\{\mathrm{E}_{\theta}\left(H^{(r+1)}(\theta)H^{(r+1+k)}(\theta)^T\right)\right.$$
$$\left.+\,\mathrm{E}_{\theta}\left(H^{(r+1+k)}(\theta)H^{(r+1)}(\theta)^T\right)\right\}.$$

The asymptotic covariance matrix in (1.5) is $V(\theta) = A(\theta)M(\theta)A(\theta)^T$. The concept of geometric α-mixing was explained in Subsection 1.5.1, where also conditions for geometric α-mixing were discussed. It is not difficult to see that if the basic diffusion process X is geometrically α-mixing, then the observed process Y inherits this property. We only need to check Condition 1.2.1 with $\bar{\theta} = \theta_0$ to obtain asymptotic results for prediction-based estimators. The condition (1.6) is satisfied because of (1.131). It is easy to see that Condition 1.9.4 (3) and (4) implies that $\theta \mapsto g(y_1, \ldots, y_{r+1})$ is continuously differentiable and that g as well as $\partial_{\theta^T} g$ are locally dominated integrable under P_{θ_0}. Finally, for a prediction-based estimating function, the condition (1.9) is identical to Condition 1.9.4 (5). Therefore it follows from Theorem 1.2.2 that a consistent G_n–estimator $\hat{\theta}_n$ exists and is the unique G_n–estimator on any bounded subset of Θ containing θ_0 with probability approaching one as $n \to \infty$. The estimator satisfies that

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

as $n \to \infty$.

1.9.3 Measurement errors

Suppose a one-dimensional diffusion has been observed with measurement errors so that the data are

$$Y_i = X_{t_i} + Z_i, \quad i = 1, \ldots, n,$$

where X solves (1.11), and the measurement errors Z_i are independent and identically distributed and independent of X. Since the observed process (Y_i) is not a Markov process, it is usually not possible to find a feasible martingale estimating function. Instead we can use a prediction-based estimating function of the type considered Example 1.9.1. To find the minimum mean square error predictor, we must find mixed moments of the form (1.137). By the binomial formula,

$$\begin{aligned}\mathrm{E}_{\theta}(Y_1^{k_1}Y_2^{k_2}) &= \mathrm{E}_{\theta}\left((X_{t_1}+Z_1)^{k_1}(X_{t_2}+Z_2)^{k_2}\right)\\ &= \sum_{i_1=0}^{k_1}\sum_{i_2=0}^{k_2}\binom{k_1}{i_1}\binom{k_2}{i_2}\mathrm{E}_{\theta}(X_{t_1}^{i_1}X_{t_2}^{i_2})\mathrm{E}_{\theta}(Z_1^{k_1-i_1})\mathrm{E}_{\theta}(Z_2^{k_2-i_2}).\end{aligned}$$

Note that the distribution of the measurement error Z_i can depend on components of the unknown parameter θ. We need to find the mixed moments

$\mathrm{E}_\theta(X_{t_1}^{i_1} X_{t_2}^{i_2})$, ($t_1 < t_2$). If expressions for the moments and conditional moments of X_t are available, these mixed moments can be found explicitly. As an example, consider the Pearson diffusions discussed in Subsection 1.3.7, for which the conditional moments are given by (1.86). Thus

$$\begin{aligned} \mathrm{E}_\theta(X_{t_1}^{i_1} X_{t_2}^{i_2}) &= \mathrm{E}_\theta(X_{t_1}^{i_1} \mathrm{E}_\theta(X_{t_2}^{i_2} \,|X_{t_1})) \qquad (1.148) \\ &= \sum_{k=0}^{i_2} \left(\sum_{\ell=0}^{i_2} q_{i_2,k,\ell} e^{-\lambda_\ell (t_2 - t_1)} \right) \mathrm{E}_\theta(X_{t_1}^{i_1+k}), \end{aligned}$$

where $\mathrm{E}_\theta(X_{t_1}^{i_1+k})$ can be found by (1.87), provided, of course, that it exists. For stationary, ergodic one-dimensional diffusions, the polynomial moments can usually be found because we have an explicit expression for the marginal density functions, at least up to a multiplicative constant, cf. (1.15). In order to find the optimal prediction-based estimating functions of the form considered in Example 1.9.1, we must find the mixed moments of the form (1.147), which can be calculated in a similar way.

1.9.4 Integrated diffusions and hypoelliptic stochastic differential equations

Sometimes a diffusion process, X, cannot be observed directly, but data of the form

$$Y_i = \frac{1}{\Delta} \int_{(i-1)\Delta}^{i\Delta} X_s \, ds, \quad i = 1, \ldots, n, \qquad (1.149)$$

are available for some fixed Δ. Such observations might be obtained when the process X is observed after passage through an electronic filter. Another example is provided by ice-core records. The isotope ratio $^{18}O/^{16}O$ in the ice is a proxy for paleo-temperatures. The average isotope ratio is measured in pieces of ice, each of which represent a time interval. The variation of the paleo-temperature can be modelled by a stochastic differential equation, and hence the ice-core data can be modelled as an integrated diffusion process, see Ditlevsen, Ditlevsen & Andersen (2002). Estimation based on this type of data was considered by Gloter (2000), Bollerslev & Wooldridge (1992), Ditlevsen & Sørensen (2004), Gloter (2006), and Baltazar-Larios & Sørensen (2009). Non-parametric inference was studied in Comte, Genon-Catalot & Rozenholc (2009).

The model for data of the type (1.149) is a particular case of (1.128) with

$$d \begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} = \begin{pmatrix} b(X_{1,t};\theta) \\ X_{1,t} \end{pmatrix} dt + \begin{pmatrix} \sigma(X_{1,t};\theta) \\ 0 \end{pmatrix} dW_t,$$

with $X_{2,0} = 0$, where W and the two components are one-dimensional, and only the second coordinate, $X_{2,t}$, is observed. The second coordinate is not stationary, but if the first coordinate is a stationary process, then the observed increments $Y_i = (X_{2,i\Delta} - X_{2,(i-1)\Delta})/\Delta$ form a stationary sequence. A stochastic differential equation of the form (1.150) is called *hypoelliptic*. Hypoelliptic stochastic differential equations are, for instance, used to model molecular dynamics, see e.g. Pokern, Stuart & Wiberg (2009). The unobserved component, $X_{1,t}$, can more generally be multivariate and have coefficients that depend on the observed component $X_{2,t}$ too. The observed smooth component can also be multivariate. The drift is typically minus the derivative of a potential. A simple example is the stochastic harmonic oscillator

$$\begin{aligned} dX_{1,t} &= -(\beta_1 X_{1,t} + \beta_2 X_{2,t})\,dt + \gamma dW_t \\ dX_{2,t} &= X_{1,t}\,dt, \end{aligned}$$

$\beta_1, \beta_2, \gamma > 0$. Observations of the form (1.149), or more generally discrete time observations of the smooth components of a hypoelliptic stochastic differential equation, do not form a Markov process, so usually a feasible martingale estimating function is not available, but prediction-based estimating functions can be used instead. For instance, the stochastic harmonic oscillator above is a Gaussian process. Therefore all the mixed moments needed in the optimal prediction-based estimating function of the form considered in Example 1.9.1 can be found explicitly.

In the following we will again denote the basic diffusion by X (rather than X_1), and assume that the data are given by (1.149). Suppose that $4N$'th moment of X_t is finite. The moments (1.137) and (1.147) can be calculated by

$$E\left[Y_1^{k_1} Y_{t_1}^{k_2} Y_{t_2}^{k_3} Y_{t_3}^{k_4}\right] = \\ \frac{\int_A E[X_{v_1} \cdots X_{v_{k_1}} X_{u_1} \cdots X_{u_{k_2}} X_{s_1} \cdots X_{s_{k_3}} X_{r_1} \cdots X_{r_{k_4}}]\,d\mathbf{t}}{\Delta^{k_1+k_2+k_3+k_4}}$$

where $1 \leq t_1 \leq t_2 \leq t_3$, $A = [0\,,\,\Delta]^{k_1} \times [(t_1 - 1)\Delta\,,\,t_1\Delta]^{k_2} \times [(t_2 - 1)\Delta\,,\,t_2\Delta]^{k_3} \times [(t_3 - 1)\Delta\,,\,t_3\Delta]^{k_4}$, and $d\mathbf{t} = dr_{k_4} \cdots dr_1\,ds_{k_3} \cdots ds_1\,du_{k_2} \cdots du_1\,dv_{k_1} \cdots dv_1$. The domain of integration can be reduced considerably by symmetry arguments, but the point is that we need to calculate mixed moments of the type $E(X_{t_1}^{\kappa_1} \cdots X_{t_k}^{\kappa_k})$, where $t_1 < \cdots < t_k$. For the Pearson diffusions discussed in Subsection 1.3.7, these mixed moments can be calculated by a simple iterative formula obtained from (1.86) and (1.87), as explained in the previous subsection. Moreover, for the Pearson diffusions, $E(X_{t_1}^{\kappa_1} \cdots X_{t_k}^{\kappa_k})$ depends on $t_1, \ldots, t_k$ through sums and products of exponential functions, cf. (1.86) and (1.148). Therefore the integral above can be explicitly calculated, and thus explicit optimal estimating functions of the type considered in Example 1.9.1 are available for observations of integrated Pearson diffusions.

Example 1.9.5 Consider observation of an integrated square root process (1.37) and a prediction-based estimating function with $f_1(x) = x$ and $f_2(x) = x^2$ with predictors given by $\pi_1^{(i-1)} = a_{1,0} + a_{1,1}Y_{i-1}$ and $\pi_2^{(i-1)} = a_{2,0}$. Then the minimum mean square error predictors are

$$\begin{aligned} \breve{\pi}_1^{(i-1)}(Y_{i-1};\theta) &= \mu\,(1-\breve{a}(\beta)) + \breve{a}(\beta)Y_{i-1}, \\ \breve{\pi}_2^{(i-1)}(\theta) &= \alpha^2 + \alpha\tau^2\beta^{-3}\Delta^{-2}(e^{-\beta\Delta} - 1 + \beta\Delta) \end{aligned}$$

with

$$\breve{a}(\beta) = \frac{(1-e^{-\beta\Delta})^2}{2(\beta\Delta - 1 + e^{-\beta\Delta})}.$$

The optimal prediction-based estimating function is

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ Y_{i-1} \\ 0 \end{pmatrix} [Y_i - \bar{\breve{\pi}}_1^{(i-1)}(Y_{i-1};\theta)] + \sum_{i=1}^{n} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} [Y_i^2 - \bar{\breve{\pi}}_2^{(i-1)}(\theta)],$$

from which we obtain the estimators

$$\begin{aligned} \hat{\alpha} &= \frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{a(\hat{\beta})Y_n - Y_1}{(n-1)(1-a(\hat{\beta}))} \\ \sum_{i=2}^{n} Y_{i-1}Y_i &= \hat{\alpha}(1-a(\hat{\beta}))\sum_{i=2}^{n} Y_{i-1} + a(\hat{\beta})\sum_{i=2}^{n} Y_{i-1}^2 \\ \hat{\sigma}^2 &= \frac{\hat{\beta}^3\Delta^2\sum_{i=2}^{n}\left(Y_i^2 - \hat{\alpha}^2\right)}{(n-1)\hat{\alpha}(e^{-\hat{\beta}\Delta} - 1 + \hat{\beta}\Delta)}. \end{aligned}$$

The estimators are explicit apart from $\hat{\beta}$, which can easily be found numerically by solving a non-linear equation in one variable. For details, see Ditlevsen & Sørensen (2004).

□

1.9.5 Sums of diffusions

An autocorrelation function of the form

$$\rho(t) = \phi_1 \exp(-\beta_1 t) + \ldots + \phi_D \exp(-\beta_D t), \qquad (1.150)$$

where $\sum_{i=1}^{D} \phi_i = 1$ and $\phi_i, \beta_i > 0$, is found in many observed time series. Examples are financial time series, see Barndorff-Nielsen & Shephard (2001), and turbulence, see Barndorff-Nielsen, Jensen & Sørensen (1990) and Bibby, Skovgaard & Sørensen (2005).

A simple model with autocorrelation function of the form (1.150) is the sum of diffusions

$$Y_t = X_{1,t} + \ldots + X_{D,t}, \qquad (1.151)$$

where the D diffusions

$$dX_{i,t} = -\beta_i(X_{i,t} - \alpha_i) + \sigma_i(X_{i,t})dW_{i,t}, \quad i = 1, \ldots, D,$$

are independent. In this case

$$\phi_i = \frac{\text{Var}(X_{i,t})}{\text{Var}(X_{1,t}) + \cdots + \text{Var}(X_{D,t})}.$$

Sums of diffusions of this type with a pre-specified marginal distribution of Y were considered by Bibby & Sørensen (2003) and Bibby, Skovgaard & Sørensen (2005), while Forman & Sørensen (2008) studied sums of Pearson diffusions. The same type of autocorrelation function is obtained for sums of independent Ornstein-Uhlenbeck processes driven by Lévy processes. This class of models was introduced and studied in Barndorff-Nielsen, Jensen & Sørensen (1998).

Example 1.9.6 *Sum of square root processes.* If $\sigma_i^2(x) = 2\beta_i bx$ and $\alpha_i = \kappa_i b$ for some $b > 0$, then the stationary distribution of Y_t is a gamma-distribution with shape parameter $\kappa_1 + \cdots + \kappa_D$ and scale parameter b. The weights in the autocorrelation function are $\phi_i = \kappa_i/(\kappa_1 + \cdots + \kappa_D)$.

□

For sums of the Pearson diffusions presented in Subsection 1.3.7, we have explicit formulae that allow calculation of (1.137) and (1.147), provided these mixed moments exists. Thus for sums of Pearson diffusions we have explicit optimal prediction-based estimating functions of the type considered in Example 1.9.1. By the multinomial formula,

$$E(Y_{t_1}^{\kappa} Y_{t_2}^{\nu}) = \\ \sum\sum \binom{\kappa}{\kappa_1, \ldots, \kappa_D} \binom{\nu}{\nu_1, \ldots, \nu_D} E(X_{1,t_1}^{\kappa_1} X_{1,t_2}^{\nu_1}) \ldots E(X_{D,t_1}^{\kappa_D} X_{D,t_2}^{\nu_D})$$

where

$$\binom{\kappa}{\kappa_1, \ldots, \kappa_D} = \frac{\kappa!}{\kappa_1! \cdots \kappa_D!}$$

is the multinomial coefficient, and where the first sum is over $0 \leq \kappa_1, \ldots, \kappa_D$ such that $\kappa_1 + \ldots \kappa_D = \kappa$, and the second sum is analogous for the ν_is. Higher order mixed moments of the form (1.147) can be found by a similar formula with four sums and four multinomial coefficients. Such formulae may appear daunting, but are easy to program. For a Pearson diffusion, mixed moments of the form $E(X_{t_1}^{\kappa_1} \cdots X_{t_k}^{\kappa_k})$ can be calculated by a simple iterative formula obtained from (1.86) and (1.87), as explained in Subsection 1.9.3.

Example 1.9.7 *Sum of two skew t-diffusions.* If

$$\sigma_i^2(x) = 2\beta_i(\nu_i - 1)^{-1}\{x^2 + 2\rho\sqrt{\nu_i}x + (1 + \rho^2)\nu\}, \quad i = 1, 2,$$

the stationary distribution of $X_{i,t}$ is a skew t-distribution,

$$\text{Var}(Y) = (1+\rho^2)\left(\frac{\nu_1}{\nu_1-2} + \frac{\nu_2}{\nu_2-2}\right),$$

and in (1.150) the weights are given by $\phi_i = \nu_i(\nu_i-2)^{-1}/\{\nu_1(\nu_1-2)^{-1} + \nu_2(\nu_2-2)^{-1}\}$. To simplify the exposition we assume that the correlation parameters β_1, β_2, ϕ_1, and ϕ_2 are known or have been estimated in advance, for instance by fitting (1.150) with $D = 2$ to the empirical autocorrelation function. We will find the optimal estimating function in the simple case where predictions of Y_i^2 are made based on predictors of the form $\pi^{(i-1)} = a_0 + a_1 Y_{i-1}$. The estimating equations take the form

$$\sum_{i=2}^{n} \begin{bmatrix} Y_i^2 - \sigma^2 - \zeta Y_{i-1} \\ Y_{i-1}Y_i^2 - \sigma^2 Y_{i-1} - \zeta Y_{i-1}^2 \end{bmatrix} = 0, \tag{1.152}$$

with

$$\begin{aligned} \sigma^2 &= \text{Var}(Y_i) = (1+\rho^2)\left\{\frac{\nu_1}{\nu_1-2} + \frac{\nu_2}{\nu_2-2}\right\}, \\ \zeta &= \frac{\text{Cov}(Y_{i-1}, Y_i^2)}{\text{Var}(Y_i)} = 4\rho\left\{\frac{\sqrt{\nu_1}}{\nu_1-3}\phi_1 e^{-\beta_1\Delta} + \frac{\sqrt{\nu_2}}{\nu_2-3}\phi_2 e^{-\beta_2\Delta}\right\}. \end{aligned}$$

Solving equation (1.152) for ζ and σ^2 we get

$$\begin{aligned} \hat{\zeta} &= \frac{\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}Y_i^2 - (\frac{1}{n-1}\sum_{i=2}^n Y_{i-1})(\frac{1}{n-1}\sum_{i=2}^n Y_i^2)}{\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}^2 - (\frac{1}{n-1}\sum_{i=2}^n Y_{i-1})^2}, \\ \hat{\sigma}^2 &= \tfrac{1}{n-1}\textstyle\sum_{i=2}^n Y_i^2 + \hat{\zeta}\tfrac{1}{n-1}\sum_{i=2}^n Y_{i-1}. \end{aligned}$$

In order to estimate ρ we restate ζ as

$$\begin{aligned} \zeta = \sqrt{32(1+\rho^2)} \cdot \rho \cdot \Bigg\{ & \frac{\sqrt{9(1+\rho^2) - \phi_1\sigma^2}}{3(1+\rho^2) - \phi_1\sigma^2}\phi_1 e^{-\beta_1\Delta} \\ & + \frac{\sqrt{9(1+\rho^2) - \phi_2\sigma^2}}{3(1+\rho^2) - \phi_2\sigma^2}\phi_2 e^{-\beta_2\Delta}\Bigg\} \end{aligned}$$

and insert $\hat{\sigma}^2$ for σ^2. Thus, we get a one-dimensional estimating equation, $\zeta(\beta, \phi, \hat{\sigma}^2, \rho) = \hat{\zeta}$, which can be solved numerically. Finally by inverting $\phi_i = \frac{1+\rho^2}{\sigma^2}\frac{\nu_i}{\nu_i-2}$, we find the estimates $\hat{\nu_i} = \frac{2\phi_i\hat{\sigma}^2}{\phi_i\hat{\sigma}^2 - (1+\hat{\rho}^2)}$, $i = 1, 2$.

□

A more complex model is obtained if the observations are integrals of the process Y given by (1.151). In this case the data are

$$Z_i = \frac{1}{\Delta}\int_{(i-1)\Delta}^{i\Delta} Y_s\, ds = \frac{1}{\Delta}\left(\int_{(i-1)\Delta}^{i\Delta} X_{1,s} ds + \cdots + \int_{(i-1)\Delta}^{i\Delta} X_{D,s} ds\right), \tag{1.153}$$

$i = 1, \ldots, n$. Also here the moments of form (1.137) and (1.147), and hence optimal prediction-based estimating functions, can be found explicitly for Pearson diffusions. This is because each of the observations Z_i is a sum of processes of the type considered in Subsection 1.9.4. To calculate $E(Z_{t_1}^{k_1} Z_{t_2}^{k_2} Z_{t_3}^{k_3} Z_{t_4}^{k_4})$, first apply the multinomial formula as above to express this quantity in terms of moments of the form $E(Y_{j,t_1}^{\ell_1} Y_{j,t_2}^{\ell_2} Y_{j,t_3}^{\ell_3} Y_{j,t_4}^{\ell_4})$, where

$$Y_{j,i} = \frac{1}{\Delta} \int_{(i-1)\Delta}^{i\Delta} X_{j,s}\, ds.$$

Now proceed as in Subsection 1.9.4.

1.9.6 Stochastic volatility models

A stochastic volatility model is a generalization of the Black-Scholes model for the logarithm of an asset price $dX_t = (\kappa + \beta\sigma^2)dt + \sigma dW_t$, that takes into account the empirical finding that the volatility σ^2 varies randomly over time:

$$dX_t = (\kappa + \beta v_t)dt + \sqrt{v_t} dW_t. \tag{1.154}$$

Here the volatility v_t is a stochastic process that cannot be observed directly. If the data are observations at the time points Δi, $i = 0, 1, 2, \ldots, n$, then the returns $Y_i = X_{i\Delta} - X_{(i-1)\Delta}$ can be written in the form

$$Y_i = \kappa\Delta + \beta S_i + \sqrt{S_i} A_i, \tag{1.155}$$

where

$$S_i = \int_{(i-1)\Delta}^{i\Delta} v_t dt, \tag{1.156}$$

and where the A_i's are independent, standard normal distributed random variables. Prediction-based estimating functions for stochastic volatility models were considered in detail in Sørensen (2000a).

Here we consider the case where the volatility process v is a sum of independent Pearson diffusions with state-space $(0, \infty)$ (the cases 2, 4 and 5). Barndorff-Nielsen & Shephard (2001) demonstrated that an autocorrelation function of the type (1.150) fits empirical autocorrelation functions of volatility well, while an autocorrelation function like that of a single Pearson diffusion is too simple to obtain a good fit. Stochastic volatility models where the volatility process is a sum of independent square root processes were considered by Bollerslev & Zhou (2002) and Bibby & Sørensen (2003). We assume that v and W are independent, so that the sequences $\{A_i\}$ and $\{S_i\}$ are independent. By the multinomial formula we find that

$$\mathrm{E}\left(Y_1^{k_1} Y_{t_1}^{k_2} Y_{t_2}^{k_3} Y_{t_3}^{k_4}\right) =$$

$$\sum K_{k_{11},\ldots,k_{43}} \mathrm{E}\left(S_1^{k_{12}+k_{13}/2} S_{t_1}^{k_{22}+k_{23}/2} S_{t_2}^{k_{32}+k_{33}/2} S_{t_3}^{k_{42}+k_{43}/2}\right)$$
$$\cdot\, \mathrm{E}\left(A_1^{k_{13}}\right) \mathrm{E}\left(A_{t_1}^{k_{23}}\right) \mathrm{E}\left(A_{t_2}^{k_{33}}\right) \mathrm{E}\left(A_{t_3}^{k_{43}}\right),$$

where the sum is over all non-negative integers k_{ij}, $i = 1,2,3,4$, $j = 1,2,3$ such that $k_{i1} + k_{i2} + k_{i3} = k_i$ ($i = 1,2,3,4$), and where

$$K_{k_{11},\ldots,k_{43}} =$$
$$\binom{k_1}{k_{11},k_{12},k_{13}}\binom{k_2}{k_{21},k_{22},k_{23}}\binom{k_3}{k_{31},k_{32},k_{33}}\binom{k_4}{k_{41},k_{42},k_{43}}(\kappa\Delta)^{k_{\cdot 1}}\beta^{k_{\cdot 2}}$$

with $k_{\cdot j} = k_{1j} + k_{2j} + k_{3j} + k_{4j}$. The moments $E(A_i^{k_{i3}})$ are the well-known moments of the standard normal distribution. When k_{i3} is odd, these moments are zero. Thus we only need to calculate the mixed moments of the form $E(S_1^{\ell_1} S_{t_1}^{\ell_2} S_{t_2}^{\ell_3} S_{t_3}^{\ell_4})$, where $\ell_1, \ldots, \ell_4$ are integers. When the volatility process is a sum of independent Pearson diffusions, S_i of the same form as Z_i in (1.153) (apart from the factor $1/\Delta$), so we can proceed as in the previous subsection to calculate the necessary mixed moments. Thus also for the stochastic volatility models defined in terms of Pearson diffusions, we can explicitly find optimal estimating functions based on prediction of powers of returns, cf. Example 1.9.1.

1.9.7 Compartment models

Diffusion compartment models are D-dimensional diffusions with linear drift,

$$dX_t = [B(\theta)X_t - b(\theta)]\,dt + \sigma(X_t;\theta)dW_t, \tag{1.157}$$

where only a subset of the coordinates are observed. Here $B(\theta)$ is a $D \times D$-matrix, $b(\theta)$ is a D-dimensional vector, $\sigma(x;\theta)$ is a $D \times D$-matrix, and W a D-dimensional standard Wiener process. Compartment models are used to model the dynamics of the flow of a substance between different parts (compartments) of, for instance, an ecosystem or the body of a human being or an animal. The process X_t is the concentration in the compartments, and flow from a given compartment into other compartments is proportional to the concentration in the given compartment, but modified by the random perturbation given by the diffusion term. The vector $b(\theta)$ represents input to or output from the system, for instance infusion or degradation of the substance. The complication is that only a subset of the compartments can be observed, for instance the first compartment, in which case the data are $Y_i = X_{1,t_i}$.

Example 1.9.8 The two-compartment model given by

$$B = \begin{pmatrix} -\beta_1 & \beta_2 \\ \beta_1 & -(\beta_1+\beta_2) \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \sigma = \begin{pmatrix} \tau_1 & 0 \\ 0 & \tau_2 \end{pmatrix},$$

where all parameters are positive, was used by Bibby (1995) to model how a radioactive tracer moved between the water and the biosphere in a certain ecosystem. Samples could only be taken from the water, the first compartment, so $Y_i = X_{1,t_i}$. Likelihood inference, which is feasible because the model is Gaussian, was studied by Bibby (1995). All mixed moments of the form (1.137) and (1.147) can be calculated explicitly, again because the model is Gaussian. Therefore also explicit optimal prediction-based estimating functions of the type considered in Example 1.9.1 are available to estimate the parameters and were studied by Düring (2002).

□

Example 1.9.9 A non-Gaussian diffusion compartment model is obtained by the specification $\sigma(x,\theta) = \text{diag}(\tau_1\sqrt{x_1}, \ldots, \tau_D\sqrt{x_D})$. This multivariate version of the square root process was studied by Düring (2002), who used methods in Down, Meyn & Tweedie (1995) to show that the D-dimensional process is geometrically α-mixing and established the asymptotic normality of prediction-based estimators of the type considered in Example 1.9.1. As in the previous example, only the first compartment is observed, i.e. the data are $Y_i = X_{1,t_i}$. For the multivariate square root model, the mixed moments (1.137) and (1.147) must be calculated numerically.

□

1.10 General asymptotics results for estimating functions

In this section we review some general asymptotic results for estimators obtained from estimating functions for stochastic process models. Proofs can be found in Jacod & Sørensen (2009).

Suppose as a statistical model for the data $X_1, X_2, \ldots, X_n$ that they are observations from a stochastic process. The corresponding probability measures (P_θ) are indexed by a p-dimensional parameter $\theta \in \Theta$. An estimating function is a function of the parameter and the observations, $G_n(\theta; X_1, X_2, \ldots, X_n)$, with values in $\mathbb{R}^p$. Usually we suppress the dependence on the observations in the notation and write $G_n(\theta)$. We get an estimator by solving the equation (1.1) and call such an estimator a G_n*-estimator*. It should be noted that n might indicate more than just the sample size: the distribution of the data $X_1, X_2, \ldots, X_n$ might depend on n. For instance, the data might be observations of a diffusion process at time points $i\Delta_n$, $i = 1, \ldots, n$, where Δ_n decreases as n increases; see Sections 1.6 and 1.7. Another example is that the diffusion coefficient might depend on n; see Section 1.8.

We will not necessarily assume that the data are observations from one of the probability measures $(P_\theta)_{\theta\in\Theta}$. We will more generally denote the *true probability measure* by P. If the statistical model contains the true model, in the sense that there exists a $\theta_0 \in \Theta$ such that $P = P_{\theta_0}$, then we call θ_0 the *true parameter value*.

A priory, there might be more than one solution or no solution at all to the estimating equation (1.1), so conditions are needed to ensure that a unique solution exists when n is sufficiently large. Moreover, we need to be careful when formally defining our estimator. In the following definition, δ denotes a "special" point, which we take to be outside Θ and $\Theta_\delta = \Theta \cup \{\delta\}$.

Definition 1.10.1 *a) The domain of G_n-estimators (for a given n) is the set A_n of all observations $x = (x_1, \ldots, x_n)$ for which $G_n(\theta) = 0$ for at least one value $\theta \in \Theta$.*

b) A G_n-estimator, $\hat\theta_n(x)$, is any function of the data with values in Θ_δ, such that for P–almost all observations we have either $\hat\theta_n(x) \in \Theta$ and $G_n(\hat\theta_n(x), x) = 0$ if $x \in A_n$, or $\hat\theta_n(x) = \delta$ if $x \notin A_n$.

We usually suppress the dependence on the observations in the notation and write $\hat\theta_n$.

The following theorem gives conditions which ensure that, for n large enough, the estimating equation (1.1) has a solution that converges to a particular parameter value $\bar\theta$. When the statistical model contains the true model, the estimating function should preferably be chosen such that $\bar\theta = \theta_0$. To facilitate the following discussion, we will refer to an estimator that converges to $\bar\theta$ in probability as a $\bar\theta$–consistent estimator, meaning that it is a (weakly) consistent estimator of $\bar\theta$. We assume that $G_n(\theta)$ is differentiable with respect to θ and denote by $\partial_{\theta^T} G_n(\theta)$ the $p \times p$-matrix, where the ijth entry is $\partial_{\theta_j} G_n(\theta)_i$.

Theorem 1.10.2 *Suppose the existence of a parameter value $\bar\theta \in \operatorname{int}\Theta$ (the interior of Θ), a connected neighbourhood M of $\bar\theta$, and a (possibly random) function W on M taking its values in the set of $p \times p$ matrices, such that the following holds:*

(i) $G_n(\bar\theta) \overset{P}{\rightarrow} 0$ *(convergence in probability, w.r.t. the true measure P) as $n \to \infty$.*

(ii) $G_n(\theta)$ *is continuously differentiable on M for all n, and*

$$\sup_{\theta\in M} \| \partial_{\theta^T} G_n(\theta) - W(\theta) \| \overset{P}{\rightarrow} 0. \tag{1.158}$$

(iii) *The matrix $W(\bar\theta)$ is non-singular with P–probability one.*

Then a sequence $(\hat\theta_n)$ of G_n-estimators exists which is $\bar\theta$-consistent. Moreover *this sequence is* eventually unique*, that is if* $(\hat{\theta}'_n)$ *is any other* $\bar{\theta}$*–consistent sequence of* G_n*–estimators, then* $P(\hat{\theta}_n \neq \hat{\theta}'_n) \to 0$ *as* $n \to \infty$*.*

Note that the condition (1.158) implies the existence of a subsequence $\{n_k\}$ such that $\partial_{\theta^T} G_{n_k}(\theta)$ converges uniformly to $W(\theta)$ on M with probability one. Hence W is a continuous function of θ (up to a null set), and it follows from elementary calculus that outside some P–null set there exists a unique continuously differentiable function G satisfying $\partial_{\theta^T} G(\theta) = W(\theta)$ for all $\theta \in M$ and $G(\bar{\theta}) = 0$. When M is a bounded set, (1.158) implies that

$$\sup_{\theta \in M} |G_n(\theta) - G(\theta)| \xrightarrow{P} 0. \qquad (1.159)$$

This observation casts light on the result of Theorem 1.10.2. Since $G_n(\theta)$ can be made arbitrarily close to $G(\theta)$ by choosing n large enough, and since $G(\theta)$ has a root at $\bar{\theta}$, it is intuitively clear that $G_n(\theta)$ must have a root near $\bar{\theta}$ when n is sufficiently large.

If we impose an identifiability condition, we can give a stronger result on any sequence of G_n–estimators. By $\bar{B}_\epsilon(\theta)$ we denote the closed ball with radius ϵ centered at θ.

Theorem 1.10.3 *Assume (1.159) for some subset* M *of* θ *containing* $\bar{\theta}$*, and that*

$$P\left(\inf_{M \backslash \bar{B}_\epsilon(\bar{\theta})} |G(\theta)| > 0 \right) = 1 \qquad (1.160)$$

for all $\epsilon > 0$*. Then for any sequence* $(\hat{\theta}_n)$ *of* G_n*–estimators*

$$P(\hat{\theta}_n \in M \backslash \bar{B}_\epsilon(\bar{\theta})) \to 0 \qquad (1.161)$$

as $n \to \infty$ *for every* $\epsilon > 0$

If $M = \Theta$, we see that any sequence $(\hat{\theta}_n)$ of G_n–estimators is $\bar{\theta}$–consistent. If the conditions of Theorem 1.10.3 hold for any compact subset M of Θ, then a sequence $(\hat{\theta}_n)$ of G_n–estimators is $\bar{\theta}$–consistent or converges to the boundary of Θ.

Finally, we give a result on the asymptotic distribution of a sequence $(\hat{\theta}_n)$ of $\bar{\theta}$–consistent G_n–estimators.

Theorem 1.10.4 *Assume the estimating function* G_n *satisfies the conditions of Theorem 1.10.2 and that there is a sequence of invertible matrices* A_n *such that each entry of* A_n^{-1} *tends to zero,*

$$\begin{pmatrix} A_n G_n(\bar{\theta}) \\ A_n \partial_{\theta^T} G_n(\bar{\theta}) A_n^{-1} \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} Z \\ W_0(\bar{\theta}) \end{pmatrix}, \qquad (1.162)$$

and there exists a connected neighbourhood M of $\bar{\theta}$ such that

$$\sup_{\theta \in M} \| A_n \partial_{\theta^T} G_n(\theta) A_n^{-1} - W_0(\theta) \| \overset{P}{\to} 0. \tag{1.163}$$

Here Z is a non-degenerate random variable, and W_0 is a random function taking values in the set of $p \times p$-matrices satisfying that $W_0(\bar{\theta})$ is invertible. Under these conditions, we have for any $\bar{\theta}$–consistent sequence $(\hat{\theta}_n)$ of G_n–estimators that

$$A_n(\hat{\theta}_n - \bar{\theta}) \xrightarrow{\mathcal{D}} -W_0(\bar{\theta})^{-1} Z. \tag{1.164}$$

When Z is normal distributed with expectation zero and covariance matrix V, and when Z is independent of $W_0(\bar{\theta})$, then the limit distribution is the normal variance-mixture with characteristic function

$$s \mapsto E\left(\exp\left(-\tfrac{1}{2} s^T W_0(\bar{\theta})^{-1} V W_0(\bar{\theta})^{T^{-1}} s\right)\right). \tag{1.165}$$

If, moreover, $W_0(\bar{\theta})$ is non-random, then the limit distribution is a normal distribution with expectation zero and covariance matrix $W_0(\bar{\theta})^{-1} V W_0(\bar{\theta})^{T^{-1}}$.

In the often occurring situation, where $W_0(\bar{\theta})$ is non-random, joint convergence of $A_n \partial_{\theta^T} G_n(\bar{\theta}) A_n^{-1}$ and $A_n G_n(\bar{\theta})$ is not necessary – marginal convergence of $A_n G_n(\bar{\theta})$ is enough.

1.11 Optimal estimating functions: general theory

The modern theory of optimal estimating functions dates back to the papers by Godambe (1960) and Durbin (1960), however the basic idea was in a sense already used in Fisher (1935). The theory was extended to stochastic processes by Godambe (1985), Godambe & Heyde (1987), Heyde (1988), and several others; see the references in Heyde (1997). Important particular instances are likelihood inference, the quasi-likelihood of Wedderburn (1974) and the generalized estimating equations developed by Liang & Zeger (1986) to deal with problems of longitudinal data analysis, see also Prentice (1988) and Li (1997). A modern review of the theory of optimal estimating functions can be found in Heyde (1997). The theory is very closely related to the theory of the generalized method of moments developed independently in parallel in the econometrics literature, where the foundation was laid by Hansen (1982), who followed Sagan (1958) by using selection matrices. Important extensions to the theory were made by Hansen (1985), Chamberlain (1987), Newey & West (1987), and Newey (1990); see also the discussion and references in Hall (2005). Particular attention is given to the time series setting in Hansen (1985), Hansen (1993), West (2001), and Kuersteiner (2002). A discussion of links between the econometrics and statistics literature can be found in Hansen (2001). In the following we present the theory as it was developed in the statistics literature by Godambe and Heyde.

The general setup is as in the previous section. We will only consider *unbiased* estimating functions, i.e., estimating functions satisfying that $E_\theta(G_n(\theta)) = 0$ for all $\theta \in \Theta$. This natural requirement is also called Fisher consistency. It often implies condition (i) of Theorem 1.10.2 for $\bar{\theta} = \theta_0$, which is an essential part of the condition for existence of a consistent estimator. Suppose we have a class $\mathcal{G}_n$ of unbiased estimating functions. How do we choose the best member in $\mathcal{G}_n$? And in what sense are some estimating functions better than others? These are the main problems in the theory of estimating functions.

To simplify the discussion, let us first assume that $p = 1$. The quantity

$$S_{G_n}(\theta) = E_\theta(\partial_{\theta^T} G_n(\theta)) \tag{1.166}$$

is called the *sensitivity* function for G_n. As in the previous section, it is assumed that $G_n(\theta)$ is differentiable with respect to θ. A large absolute value of the sensitivity implies that the equation $G_n(\theta) = 0$ tends to have a solution near the true parameter value, where the expectation of $G_n(\theta)$ is equal to zero. Thus a good estimating function is one with a large absolute value of the sensitivity.

Ideally, we would base the statistical inference on the likelihood function $L_n(\theta)$, and hence use the score function $U_n(\theta) = \partial_\theta \log L_n(\theta)$ as our estimating function. This usually yields an efficient estimator. However, when $L_n(\theta)$ is not available or is difficult to calculate, we might prefer to use an estimating function that is easier to obtain and is in some sense close to the score function. Suppose that both $U_n(\theta)$ and $G_n(\theta)$ have finite variance. Then it can be proven under usual regularity conditions that

$$S_{G_n}(\theta) = -\text{Cov}_\theta(G_n(\theta), U_n(\theta)).$$

Thus we can find an estimating function $G_n(\theta)$ that maximizes the absolute value of the correlation between $G_n(\theta)$ and $U_n(\theta)$ by finding one that maximizes the quantity

$$K_{G_n}(\theta) = S_{G_n}(\theta)^2/\text{Var}_\theta(G_n(\theta)) = S_{G_n}(\theta)^2/E_\theta(G_n(\theta)^2), \tag{1.167}$$

which is known as the *Godambe information*. This makes intuitive sense: the ratio $K_{G_n}(\theta)$ is large when the sensitivity is large and when the variance of $G_n(\theta)$ is small. The Godambe information is a natural generalization of the Fisher information. Indeed, $K_{U_n}(\theta)$ is the Fisher information. For a discussion of information quantities in a stochastic process setting, see Barndorff-Nielsen & Sørensen (1991) and Barndorff-Nielsen & Sørensen (1994). In a short while, we shall see that the Godambe information has a large sample interpretation too. An estimating function $G_n^* \in \mathcal{G}_n$ is called *Godambe-optimal* in $\mathcal{G}_n$ if

$$K_{G_n^*}(\theta) \geq K_{G_n}(\theta) \tag{1.168}$$

for all $\theta \in \Theta$ and for all $G_n \in \mathcal{G}_n$.

When the parameter θ is multivariate ($p > 1$), the sensitivity function is the $p \times p$-matrix

$$S_{G_n}(\theta) = E_\theta(\partial_{\theta^T} G_n(\theta)). \tag{1.169}$$

For a multivariate parameter, the Godambe information is the $p \times p$-matrix

$$K_{G_n}(\theta) = S_{G_n}(\theta)^T E_\theta \left(G_n(\theta) G_n(\theta)^T \right)^{-1} S_{G_n}(\theta), \tag{1.170}$$

and an optimal estimating function G_n^* can be defined by (1.168) with the inequality referring to the partial ordering of the set of positive semi-definite $p \times p$-matrices. Whether an Godambe-optimal estimating function exists and whether it is unique depends on the class $\mathcal{G}_n$. In any case, it is only unique up to multiplication by a regular matrix that might depend on θ. Specifically, if $G_n^*(\theta)$ satisfies (1.168), then so does $M_\theta G_n^*(\theta)$ where M_θ is an invertible deterministic $p \times p$-matrix. Fortunately, the two estimating functions give rise to the same estimator(s), and we refer to them as *versions* of the same estimating function. For theoretical purposes a standardized version of the estimating functions is useful. The *standardized version* of $G_n(\theta)$ is given by

$$G_n^{(s)}(\theta) = -S_{G_n}(\theta)^T E_\theta \left(G_n(\theta) G_n(\theta)^T \right)^{-1} G_n(\theta).$$

The rationale behind this standardization is that $G_n^{(s)}(\theta)$ satisfies the *second Bartlett-identity*

$$E_\theta \left(G_n^{(s)}(\theta) G_n^{(s)}(\theta)^T \right) = -E_\theta(\partial_{\theta^T} G_n^{(s)}(\theta)), \tag{1.171}$$

an identity usually satisfied by the score function. The standardized estimating function $G_n^{(s)}(\theta)$ is therefore more directly comparable to the score function. Note that when the second Bartlett identity is satisfied, the Godambe information equals minus the sensitivity matrix.

An Godambe-optimal estimating function is close to the score function U_n in an L_2-sense. Suppose G_n^* is Godambe-optimal in $\mathcal{G}_n$. Then the standardized version $G_n^{*(s)}(\theta)$ satisfies the inequality

$$\begin{aligned} & E_\theta \left((G_n^{(s)}(\theta) - U_n(\theta))^T (G_n^{(s)}(\theta) - U_n(\theta)) \right) \\ \geq \quad & E_\theta \left((G_n^{*(s)}(\theta) - U_n(\theta))^T (G_n^{*(s)}(\theta) - U_n(\theta)) \right) \end{aligned}$$

for all $\theta \in \Theta$ and for all $G_n \in \mathcal{G}_n$, see Heyde (1988). In fact, if $\mathcal{G}_n$ is a closed subspace of the L_2-space of all square integrable functions of the data, then the optimal estimating function is the orthogonal projection of the score function onto $\mathcal{G}_n$. For further discussion of this Hilbert space approach to estimating functions, see McLeish & Small (1988). The interpretation of an optimal estimating function as an approximation to the score function is important. By choosing a sequence of classes $\mathcal{G}_n$ that, as $n \to \infty$, converges to a subspace containing the score function U_n, a sequence of estimators that is asymptotically fully efficient can be constructed.

The following result by Heyde (1988) can often be used to find the optimal estimating function.

Theorem 1.11.1 *If $G_n^* \in \mathcal{G}_n$ satisfies the equation*

$$S_{G_n}(\theta)^{-1} E_\theta \left(G_n(\theta) G_n^*(\theta)^T \right) = S_{G_n^*}(\theta)^{-1} E_\theta \left(G_n^*(\theta) G_n^*(\theta)^T \right) \tag{1.172}$$

for all $\theta \in \Theta$ and for all $G_n \in \mathcal{G}_n$, then it is Godambe-optimal in $\mathcal{G}_n$. When $\mathcal{G}_n$ is closed under addition, any Godambe-optimal estimating function G_n^* satisfies (1.172) .

The condition (1.172) can often be verified by showing that $E_\theta(G_n(\theta)G_n^*(\theta)^T) = -E_\theta(\partial_{\theta^T} G_n(\theta))$ for all $\theta \in \Theta$ and for all $G_n \in \mathcal{G}_n$. In such situations, G_n^* satisfies the *second Bartlett-identity*, (1.171), so that

$$K_{G_n^*}(\theta) = E_\theta \left(G_n^*(\theta) G_n^*(\theta)^T \right).$$

Example 1.11.2 Suppose we have a number of functions $h_{ij}(x_1, \ldots, x_i; \theta)$, $j = 1, \ldots, N$, $i = 1, \ldots n$ satisfying that

$$\mathrm{E}_\theta(h_{ij}(X_1, \ldots, X_i; \theta)) = 0.$$

Such functions define relationships (dependent on θ) between an observation X_i and the previous observations $X_1, \ldots, X_{i-1}$ (or some of them) that are on average equal to zero. It is natural to use such relationships to estimate θ by solving the equations $\sum_{i=1}^n h_{ij}(X_1, \ldots, X_i; \theta) = 0$. In order to estimate θ it is necessary that $N \geq p$, but if $N > p$ we have too many equations. The theory of optimal estimating functions tells us how to combine the N relations in an optimal way.

Let h_i denote the N-dimensional vector $(h_{i1}, \ldots, h_{iN})^T$, and define an N-dimensional estimating function by $H_n(\theta) = \sum_{i=1}^n h_i(X_1, \ldots, X_i; \theta)$. First we consider the class of p-dimensional estimating functions of the form

$$G_n(\theta) = A_n(\theta) H_n(\theta),$$

where $A_n(\theta)$ is a non-random $p \times N$-matrix that is differentiable with respect to θ. By $A_n^*(\theta)$ we denote the optimal choice of $A_n(\theta)$. It is not difficult to see that

$$S_{G_n}(\theta) = A_n(\theta) S_{H_n}(\theta)$$

and

$$\mathrm{E}_\theta \left(G_n(\theta) G_n^*(\theta)^T \right) = A_n(\theta) \mathrm{E}_\theta \left(H_n(\theta) H_n(\theta)^T \right) A_n^*(\theta)^T,$$

where $S_{H_n}(\theta) = E_\theta(\partial_{\theta^T} H_n(\theta))$. If we choose

$$A_n^*(\theta) = -S_{H_n}(\theta)^T \mathrm{E}_\theta \left(H_n(\theta) H_n(\theta)^T \right)^{-1},$$

then (1.172) is satisfied for all $G_n \in \mathcal{G}_n$, so that $G_n^*(\theta) = A_n^*(\theta)H_n(\theta)$ is Godambe optimal.

Sometimes there are good reasons to use functions h_{ij} satisfying that

$$\mathrm{E}_\theta(h_{ij}(X_1, \ldots, X_i; \theta)h_{i'j'}(X_1, \ldots, X_{i'}; \theta)) = 0 \qquad (1.173)$$

for all $j, j' = 1, \ldots, N$ when $i \neq i'$. For such functions the random variables $h_{ij}(X_1, \ldots, X_i; \theta)$, $i = 1, 2, \ldots$ are uncorrelated, and in this sense the "new" random variation of $h_{ij}(X_1, \ldots, X_i; \theta)$ depends only on the innovation in the ith observation. This is for instance the case for martingale estimating functions, see (1.180). In this situation it is natural to consider the larger class of estimating functions given by

$$G_n(\theta) = \sum_{i=1}^{n} a_i(\theta)h_i(X_1, \ldots, X_i; \theta),$$

where $a_i(\theta)$, $i = 1, \ldots n$, are $p \times N$ matrices that do not depend on the data and are differentiable with respect to θ. Here

$$S_{G_n}(\theta) = \sum_{i=1}^{n} a_i(\theta)\mathrm{E}_\theta(\partial_{\theta^T} h_i(X_1, \ldots, X_i; \theta))$$

and

$$\mathrm{E}_\theta\left(G_n(\theta)G_n^*(\theta)^T\right) = \\ \sum_{i=1}^{n} a_i(\theta)\mathrm{E}_\theta\left(h_i(X_1, \ldots, X_i; \theta)h_i(X_1, \ldots, X_i; \theta)^T\right) a_i^*(\theta)^T,$$

where $a_i^*(\theta)$ denotes the optimal choice of $a_i(\theta)$. We see that with

$$a_i^*(\theta) = \\ -\mathrm{E}_\theta(\partial_{\theta^T} h_i(X_1, \ldots, X_i; \theta))^T \big(\mathrm{E}_\theta\left(h_i(X_1, \ldots, X_i; \theta)h_i(X_1, \ldots, X_i; \theta)^T\right)\big)^{-1}$$

the condition (1.172) is satisfied. □

1.11.1 Martingale estimating functions

More can be said about martingale estimating functions, i.e. estimating functions G_n satisfying that

$$E_\theta(G_n(\theta)|\mathcal{F}_{n-1}) = G_{n-1}(\theta), \quad n = 1, 2, \ldots,$$

where $\mathcal{F}_{n-1}$ is the σ-field generated by the observations $X_1, \ldots, X_{n-1}$ ($G_0 = 0$ and $\mathcal{F}_0$ is the trivial σ-field). In other words, the stochastic process $\{G_n(\theta) : n = 1, 2, \ldots\}$ is a martingale under the model given by the parameter value θ. Since the score function is usually a martingale (see e.g. Barndorff-Nielsen

& Sørensen (1994)), it is natural to approximate it by families of martingale estimating functions.

The well-developed martingale limit theory allows a straightforward discussion of the asymptotic theory, and motivates an optimality criterion that is particular to martingale estimating functions. Suppose the estimating function $G_n(\theta)$ satisfies the conditions of the central limit theorem for martingales and let $\hat{\theta}_n$ be a solution of the equation $G_n(\theta) = 0$. Under the regularity conditions of the previous section, it can be proved that

$$\langle G(\theta)\rangle_n^{-\frac{1}{2}} \bar{G}_n(\theta)(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, I_p). \tag{1.174}$$

Here $\langle G(\theta)\rangle_n$ is the *quadratic characteristic* of $G_n(\theta)$ defined by

$$\langle G(\theta)\rangle_n = \sum_{i=1}^{n} E_\theta \left((G_i(\theta) - G_{i-1}(\theta))(G_i(\theta) - G_{i-1}(\theta))^T | \mathcal{F}_{i-1} \right),$$

and $\partial_{\theta^T} G_n(\theta)$ has been replaced by its compensator

$$\bar{G}_n(\theta) = \sum_{i=1}^{n} E_\theta \left(\partial_{\theta^T} G_i(\theta) - \partial_{\theta^T} G_{i-1}(\theta) | \mathcal{F}_{i-1} \right),$$

using the extra assumption that $\bar{G}_n(\theta)^{-1} \partial_{\theta^T} G_n(\theta) \xrightarrow{P_\theta} I_p$. Details can be found in Heyde (1988). We see that the inverse of the data-dependent matrix

$$I_{G_n}(\theta) = \bar{G}_n(\theta)^T \langle G(\theta)\rangle_n^{-1} \bar{G}_n(\theta) \tag{1.175}$$

estimates the co-variance matrix of the asymptotic distribution of the estimator $\hat{\theta}_n$. Therefore $I_{G_n}(\theta)$ can be interpreted as an information matrix, called the *Heyde-information*. It generalizes the incremental expected information of the likelihood theory for stochastic processes; see Barndorff-Nielsen & Sørensen (1994). Since $\bar{G}_n(\theta)$ estimates the sensitivity function, and $\langle G(\theta)\rangle_n$ estimates the variance of the asymptotic distribution of $G_n(\theta)$, the Heyde-information has a heuristic interpretation similar to that of the Godambe-information. In fact,

$$E_\theta \left(\bar{G}_n(\theta) \right) = S_{G_n}(\theta) \quad \text{and} \quad E_\theta \left(\langle G(\theta)\rangle_n \right) = E_\theta \left(G_n(\theta) G_n(\theta)^T \right).$$

We can thus think of the Heyde-information as an estimated version of the Godambe information.

Let $\mathcal{G}_n$ be a class of martingale estimating functions with finite variance. We say that a martingale estimating function G_n^* is *Heyde-optimal* in $\mathcal{G}_n$ if

$$I_{G_n^*}(\theta) \geq I_{G_n}(\theta) \tag{1.176}$$

P_θ-almost surely for all $\theta \in \Theta$ and for all $G_n \in \mathcal{G}_n$.

The following useful result from Heyde (1988) is similar to Theorem 1.11.1. In order to formulate it, we need the concept of the *quadratic co-characteristic* of two martingales, G and $\tilde{G}$, both of which are assumed to have finite variance:

$$\langle G, \tilde{G} \rangle_n = \sum_{i=1}^{n} E\left((G_i - G_{i-1})(\tilde{G}_i - \tilde{G}_{i-1})^T | \mathcal{F}_{i-1} \right). \tag{1.177}$$

Theorem 1.11.3 *If $G_n^* \in \mathcal{G}_n$ satisfies that*

$$\bar{G}_n(\theta)^{-1} \langle G(\theta), G^*(\theta) \rangle_n = \bar{G}_n^*(\theta)^{-1} \langle G^*(\theta) \rangle_n \tag{1.178}$$

for all $\theta \in \Theta$ and all $G_n \in \mathcal{G}_n$, then it is is Heyde-optimal in $\mathcal{G}_n$. When $\mathcal{G}_n$ is closed under addition, any Heyde-optimal estimating function G_n^* satisfies (1.178)*. Moreover, if $\bar{G}_n^*(\theta)^{-1} \langle G^*(\theta) \rangle_n$ is non-random, then G_n^* is also Godambe-optimal in $\mathcal{G}_n$.*

Since in many situations condition (1.178) can be verified by showing that $\langle G(\theta), G^*(\theta) \rangle_n = -\bar{G}_n(\theta)$ for all $G_n \in \mathcal{G}_n$, it is in practice often the case that Heyde-optimality implies Godambe-optimality.

Example 1.11.4 Let us discuss a often occurring type of martingale estimating functions. To simplify the exposition we assume that the observed process is Markovian. For Markov processes it is natural to base martingale estimating functions on functions $h_{ij}(y, x; \theta)$, $j = 1, \ldots, N$, $i = 1, \ldots, n$ satisfying that

$$E_\theta(h_{ij}(X_i, X_{i-1}; \theta) | \mathcal{F}_{i-1}) = 0. \tag{1.179}$$

As in Example 1.11.2, such functions define relationships (dependent on θ) between consecutive observation X_i and X_{i-1} that are, on average, equal to zero and can be used to estimate θ. We consider the class of p-dimensional estimating functions of the form

$$G_n(\theta) = \sum_{i=1}^{n} a_i(X_{i-1}; \theta) h_i(X_i, X_{i-1}; \theta), \tag{1.180}$$

where h_i denotes the N-dimensional vector $(h_{i1}, \ldots, h_{iN})^T$, and $a_i(x; \theta)$ is a function from $\mathbb{R} \times \Theta$ into the set of $p \times N$-matrices that are differentiable with respect to θ. It follows from (1.179) that $G_n(\theta)$ is a p-dimensional unbiased martingale estimating function.

We will now find the matrices a_i that combine the N functions h_{ij} in an optimal way. Let $\mathcal{G}_n$ be the class of martingale estimating functions of the form (1.180) that have finite variance. Then

$$\bar{G}_n(\theta) = \sum_{i=1}^{n} a_i(X_{i-1}; \theta) E_\theta(\partial_{\theta^T} h_i(X_i, X_{i-1}; \theta) | \mathcal{F}_{i-1})$$

and

$$\langle G(\theta), G^*(\theta)\rangle_n = \sum_{i=1}^n a_i(X_{i-1};\theta) V_{h_i}(X_{i-1};\theta) a_i^*(X_{i-1};\theta)^T,$$

where

$$G_n^*(\theta) = \sum_{i=1}^n a_i^*(X_{i-1};\theta) h_i(X_i, X_{i-1};\theta), \tag{1.181}$$

and

$$V_{h_i}(X_{i-1};\theta) = E_\theta \left(h_i(X_i, X_{i-1};\theta) h_i(X_i, X_{i-1};\theta)^T | \mathcal{F}_{i-1} \right)$$

is the conditional covariance matrix of the random vector $h_i(X_i, X_{i-1};\theta)$ given $\mathcal{F}_{i-1}$. If we assume that $V_{h_i}(X_{i-1};\theta)$ is invertible and define

$$a_i^*(X_{i-1};\theta) = -E_\theta(\partial_{\theta^T} h_i(X_i, X_{i-1};\theta)|\mathcal{F}_{i-1})^T V_{h_i}(X_{i-1};\theta)^{-1}, \tag{1.182}$$

then the condition (1.178) is satisfied. Hence by Theorem 1.11.3 the estimating function $G_n^*(\theta)$ with a_i^* given by (1.182) is Heyde-optimal - provided, of course, that it has finite variance. Since $\bar{G}_n^*(\theta)^{-1}\langle G^*(\theta)\rangle_n = -I_p$ is non-random, the estimating function $G_n^*(\theta)$ is also Godambe-optimal. If a_i^* were defined without the minus, $G_n^*(\theta)$ would obviously also be optimal. The reason for the minus will be clear in the following.

We shall now see, in exactly what sense the optimal estimating function (1.181) approximates the score function. The following result was first given by Kessler (1996). Let $p_i(y;\theta|x)$ denote the conditional density of X_i given that $X_{i-1} = x$. Then the likelihood function for θ based on the data $(X_1, \ldots, X_n)$ is

$$L_n(\theta) = \prod_{i=1}^n p_i(X_i;\theta|X_{i-1})$$

(with p_1 denoting the unconditional density of X_1). If we assume that all p_is are differentiable with respect to θ, the score function is

$$U_n(\theta) = \sum_{i=1}^n \partial_\theta \log p_i(X_i;\theta|X_{i-1}). \tag{1.183}$$

Let us fix i, x_{i-1} and θ and consider the L_2-space $\mathcal{K}_i(x_{i-1},\theta)$ of functions $f : \mathbb{R} \mapsto \mathbb{R}$ for which $\int f(y)^2 p_i(y;\theta|x_{i-1})dy < \infty$. We equip $\mathcal{K}_i(x_{i-1},\theta)$ with the usual inner product

$$\langle f, g\rangle = \int f(y)g(y)p_i(y;\theta|x_{i-1})dy,$$

and let $\mathcal{H}_i(x_{i-1},\theta)$ denote the N-dimensional subspace of $\mathcal{K}_i(x_{i-1},\theta)$ spanned by the functions $y \mapsto h_{ij}(y, x_{i-1};\theta)$, $j = 1, \ldots, N$. That the functions are linearly independent in $\mathcal{K}_i(x_{i-1},\theta)$ follows from the earlier assumption that the covariance matrix $V_{h_i}(x_{i-1};\theta)$ is regular.

Now, assume that $\partial_{\theta_j} \log p_i(y|x_{i-1};\theta) \in \mathcal{K}_i(x_{i-1},\theta)$ for $j = 1,\ldots,p$, denote by g^*_{ij} the orthogonal projection with respect to $\langle \cdot,\cdot \rangle$ of $\partial_{\theta_j} \log p_i$ onto $\mathcal{H}_i(x_{i-1},\theta)$, and define a p-dimensional function by $g^*_i = (g^*_{i1},\ldots,g^*_{ip})^T$. Then (under weak regularity conditions)

$$g^*_i(x_{i-1},x;\theta) = a^*_i(x_{i-1};\theta)h_i(x_{i-1},x;\theta), \tag{1.184}$$

where a^*_i is the matrix defined by (1.182). To see this, note that g^* must have the form (1.184) with a^*_i satisfying the normal equations

$$\langle \partial_{\theta_j} \log p_i - g^*_j,\, h_{ik} \rangle = 0,$$

$j = 1,\ldots,p$ and $k = 1,\ldots,N$. These equations can be expressed in the form $B_i = a^*_i V_{h_i}$, where B_i is the $p \times p$-matrix whose (j,k)th element is $\langle \partial_{\theta_j} \log p_i,\, h_{ik} \rangle$. The main regularity condition needed to prove (1.184) is that we can interchange differentiation and integration so that

$$\int \partial_{\theta_j} \left[h_{ik}(y,x_{i-1};\theta)p(y,x_{i-1};\theta) \right] dy =$$

$$\partial_{\theta_j} \int h_{ik}(y,x_{i-1};\theta)p(x_{i-1},y;\theta)dy = 0,$$

from which it follows that

$$B_i = -\int \partial_{\theta^T} h_i(y,x_{i-1};\theta)p(x_{i-1},y;\theta)dy.$$

Thus a^*_i is given by (1.182).

□

Acknowledgements

The research was supported by the Danish Center for Accounting and Finance funded by the Danish Social Science Research Council and by the Center for Research in Econometric Analysis of Time Series funded by the Danish National Research Foundation.

Bibliography

Aït-Sahalia, Y. (2002). "Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach". *Econometrica*, 70:223–262.

Aït-Sahalia, Y. (2008). "Closed-form likelihood expansions for multivariate diffusions". *Ann. Statist.*, 36:906–937.

Aït-Sahalia, Y. & Mykland, P. (2003). "The effects of random and discrete sampling when estimating continuous-time diffusions". *Econometrica*, 71:483–549.

Aït-Sahalia, Y. & Mykland, P. A. (2004). "Estimators of diffusions with randomly spaced discrete observations: a general theory". *Ann. Statist.*, 32:2186–2222.

Aït-Sahalia, Y. & Mykland, P. A. (2008). "An analysis of Hansen-Scheinkman moment estimators for discretely and randomly sampled diffusions". *Journal of Econometrics*, 144:1–26.

Baltazar-Larios, F. & Sørensen, M. (2009). "Maximum likelihood estimation for integrated diffusion processes". Preprint, Department of Mathematical Sciences, University of Copenhagen.

Barndorff-Nielsen, O. E.; Jensen, J. L. & Sørensen, M. (1990). "Parametric Modelling of Turbulence". *Phil. Trans. R. Soc. Lond. A*, 332:439–455.

Barndorff-Nielsen, O. E.; Jensen, J. L. & Sørensen, M. (1998). "Some Stationary Processes in Discrete and Continuous Time". *Advances in Applied Probability*, 30:989–1007.

Barndorff-Nielsen, O. E.; Kent, J. & Sørensen, M. (1982). "Normal variance-mean mixtures and z-distributions". *International Statistical Review*, 50:145–159.

Barndorff-Nielsen, O. E. & Shephard, N. (2001). "Non-Gaussian Ornstein-Uhlenbeck-Based Models and some of their Uses in Financial Econometrics (with discussion)". *Journal of the Royal Statistical Society* **B**, 63:167–241.

Barndorff-Nielsen, O. E. & Sørensen, M. (1991). "Information quantities in non-classical settings". *Computational Statistics and Data Analysis*, 12:143–158.

Barndorff-Nielsen, O. E. & Sørensen, M. (1994). “A review of some aspects of asymptotic likelihood theory for stochastic processes”. *International Statistical Review*, 62:133–165.

Beskos, A.; Papaspiliopoulos, O. & Roberts, G. O. (2006). “Retrospective exact simulation of diffusion sample paths with applications”. *Bernoulli*, 12:1077–1098.

Beskos, A.; Papaspiliopoulos, O.; Roberts, G. O. & Fearnhead, P. (2006). “Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes”. *J. Roy. Statist. Soc. B*, 68:333–382.

Beskos, A. & Roberts, G. O. (2005). “Exact simulation of diffusions”. *Ann. Appl. Prob.*, 15:2422–2444.

Bibby, B. M. (1995). *Inference for diffusion processes with particular emphasis on compartmental diffusion processes*. PhD thesis, University of Aarhus.

Bibby, B. M.; Skovgaard, I. M. & Sørensen, M. (2005). “Diffusion-type models with given marginals and autocorrelation function”. *Bernoulli*, 11:191–220.

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

Bibby, B. M. & Sørensen, M. (1996). “On estimation for discretely observed diffusions: a review”. *Theory of Stochastic Processes*, 2:49–56.

Bibby, B. M. & Sørensen, M. (2003). “Hyperbolic processes in finance”. In Rachev, S., editor, *Handbook of Heavy Tailed Distributions in Finance*, pages 211–248. Elsevier Science.

Billingsley, P. (1961). “The Lindeberg-Lévy theorem for martingales”. *Proc. Amer. Math. Soc.*, 12:788–792.

Bollerslev, T. & Wooldridge, J. (1992). “Quasi-maximum likelihood estimators and inference in dynamic models with time-varying covariances”. *Econometric Review*, 11:143–172.

Bollerslev, T. & Zhou, H. (2002). “Estimating stochastic volatility diffusion using conditional moments of integrated volatility”. *Journal of Econometrics*, 109:33–65.

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

Campbell, J. Y.; Lo, A. W. & MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press, Princeton.

Chamberlain, G. (1987). “Asymptotic efficiency in estimation with conditional moment restrictions”. *Journal of Econometrics*, 34:305–34.

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

Christensen, B. J. & Sørensen, M. (2008). “Optimal inference in dynamic moels with conditional moment restrictions”. Working paper, CREATES, Center for Research in Econometric Analysis of Time Series.

Clement, E. (1997). “Estimation of diffusion processes by simulated moment methods”. *Scand. J. Statist.*, 24:353–369.

Comte, F.; Genon-Catalot, V. & Rozenholc, Y. (2009). “Nonparametric adaptive estimation for integrated diffusions”. *Stochastic processes and their applications*, 119:811–834.

Conley, T. G.; Hansen, L. P.; Luttmer, E. G. J. & Scheinkman, J. A. (1997). “Short-term interest rates as subordinated diffusions”. *Review of Financial Studies*, 10:525–577.

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

De Jong, F.; Drost, F. C. & Werker, B. J. M. (2001). “A jump-diffusion model for exchange rates in a target zone”. *Statistica Neerlandica*, 55:270–300.

Ditlevsen, P. D.; Ditlevsen, S. & Andersen, K. K. (2002). “The fast climate fluctuations during the stadial and interstadial climate states”. *Annals of Glaciology*, 35:457–462.

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

Dorogovcev, A. J. (1976). “The consistency of an estimate of a parameter of a stochastic differential equation”. *Theor. Probability and Math. Statist.*, 10:73–82.

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

Down, D.; Meyn, S. & Tweedie, R. (1995). “Exponential and uniform ergodicity of Markov processes”. *Annals of Probability*, 23:1671–1691.

Duffie, D. & Singleton, K. (1993). “Simulated moments estimation of Markov models of asset prices”. *Econometrica*, 61:929–952.

Durbin, J. (1960). “Estimation of parameters in time-series regression models”. *J. Roy. Statist. Soc. B*, 22:139–153.

Durham, G. B. & Gallant, A. R. (2002). “Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes”. *J. Business & Econom. Statist.*, 20:297–338.

Düring, M. (2002). “Den prediktions-baserede estimationsfunktion for diffusions puljemodeller". Master's thesis, University of Copenhagen. In Danish.

Elerian, O.; Chib, S. & Shephard, N. (2001). "Likelihood inference for discretely observed non-linear diffusions". *Econometrica*, 69:959–993.

Eraker, B. (2001). "MCMC Analysis of Diffusion Models with Application to Finance". *J. Business & Econom. Statist.*, 19:177–191.

Fisher, R. A. (1935). "The logic of inductive inference". *J. Roy. Statist. Soc.*, 98:39–54.

Florens-Zmirou, D. (1989). "Approximate discrete-time schemes for statistics of diffusion processes". *Statistics*, 20:547–557.

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

Freidlin, M. I. & Wentzell, A. D. (1998). *Random pertubations of dynamical systems.2nd Edition*. Springer, New York.

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

Genon-Catalot, V. (1990). "Maximum contrast estimation for diffusion processes from discrete observations". *Statistics*, 21:99–116.

Genon-Catalot, V. & Jacod, J. (1993). "On the estimation of the diffusion coefficient for multi-dimensional diffusion processes". *Ann. Inst. Henri Poincaré, Probabilités et Statistiques*, 29:119–151.

Genon-Catalot, V.; Jeantheau, T. & Larédo, C. (2000). "Stochastic volatility models as hidden Markov models and statistical applications". *Bernoulli*, 6:1051–1079.

Gloter, A. (2000). "Parameter estimation for a discrete sampling of an integrated Ornstein-Uhlenbeck process". *Statistics*, 35:225–243.

Gloter, A. (2006). "Parameter estimation for a discretely observed integrated diffusion process". *Scand. J. Statist.*, 33:83–104.

Gloter, A. & Sørensen, M. (2009). "Estimation for stochastic differential equations with a small diffusion coefficient". *Stoch. Proc. Appl.*, 119:679–699.

Gobet, E. (2001). "Local asymptotic mixed normality property for elliptic diffusion: a Malliavin calculus approach". *Bernoulli*, 7:899–912.

Gobet, E. (2002). "LAN property for ergodic diffusions with discrete observations". *Ann. Inst. Henri Poincaré, Probabilités et Statistiques*, 38:711–737.

Godambe, V. P. (1960). "An optimum property of regular maximum likelihood estimation". *Ann. Math. Stat.*, 31:1208–1212.

Godambe, V. P. (1985). "The foundations of finite sample estimation in stochastic processes". *Biometrika*, 72:419–428.

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

Gourieroux, C. & Jasiak, J. (2006). “Multivariate Jacobi process and with application to smooth transitions”. *Journal of Econometrics*, 131:475–505.

Gradshteyn, I. S. & Ryzhik, I. M. (1965). *Table of Integrals, Series, and Products, 4th Edition*. Academic Press, New-York.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press, New York.

Hall, P. & Heyde, C. C. (1980). *Martingale Limit Theory and Its Applications*. Academic Press, New York.

Hansen, L. P. (1982). “Large sample properties of generalized method of moments estimators”. *Econometrica*, 50:1029–1054.

Hansen, L. P. (1985). “A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators”. *Journal of Econometrics*, 30:203–238.

Hansen, L. P. (1993). “Semiparametric efficiency bounds for linear time-series models”. In Phillips, P. C. B., editor, *Models, Methods and Applications of Econometrics: Esseys in Honor of A. R. Bergstrom*, pages 253–271. Cambridge, MA: Blackwell.

Hansen, L. P. (2001). “Method of moments”. In *International Encyclopedia of the Social and Behavior Sciences*. New York: Elsevier.

Hansen, L. P. & Scheinkman, J. A. (1995). “Back to the future: generating moment implications for continuous-time Markov processes”. *Econometrica*, 63:767–804.

Hansen, L. P.; Scheinkman, J. A. & Touzi, N. (1998). “Spectral methods for identifying scalar diffusions”. *Journal of Econometrics*, 86:1–32.

Heyde, C. C. (1988). “Fixed sample and asymptotic optimality for classes of estimating functions”. *Contemporary Mathematics*, 80:241–247.

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

Hildebrandt, E. H. (1931). “Systems of polynomials connected with the Charlier expansions and the Pearson differential and difference equations”. *Ann. Math. Statist.*, 2:379–439.

Iacus, S. M. (2008). *Simulation and Inference for Stochastic Differential Equations*. Springer, New York.

Iacus, S. M. & Kutoyants, Y. (2001). “Semiparametric hypotheses testing for dynamical systems with small noise”. *Math. Methods Statist.*, 10:105–120.

Jacobsen, M. (2001). “Discretely observed diffusions; classes of estimating functions and small Δ-optimality”. *Scand. J. Statist.*, 28:123–150.

Jacobsen, M. (2002). “Optimality and small Δ-optimality of martingale estimating functions”. *Bernoulli*, 8:643–668.

Jacod, J. & Sørensen, M. (2009). “Aspects of asymptotic statistical theory for stochastic processes.”. Preprint, Department of Mathematical Sciences, University of Copenhagen. In preparation.

Jeganathan, P. (1982). “On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal”. *Sankhyā A*, 44:173–212.

Jeganathan, P. (1983). “Some asymptotic properties of risk functions when the limit of the experiment is mixed normal”. *Sankhyā A*, 45:66–87.

Kallenberg, O. (1997). *Foundations of Modern Probability*. Springer-Verlag, New York.

Kelly, L.; Platen, E. & Sørensen, M. (2004). “Estimation for discretely observed diffusions using transform functions”. *J. Appl. Prob.*, 41:99–118.

Kessler, M. (1996). *Estimation paramétrique des coefficients d’une diffusion ergodique à partir d’observations discrètes*. PhD thesis, Laboratoire de Probabilités, Université Paris VI.

Kessler, M. (1997). “Estimation of an ergodic diffusion from discrete observations”. *Scand. J. Statist.*, 24:211–229.

Kessler, M. (2000). “Simple and explicit estimating functions for a discretely observed diffusion process”. *Scand. J. Statist.*, 27:65–82.

Kessler, M. & Paredes, S. (2002). “Computational aspects related to martingale estimating functions for a discretely observed diffusion”. *Scand. J. Statist.*, 29:425–440.

Kessler, M.; Schick, A. & Wefelmeyer, W. (2001). “The information in the marginal law of a Markov chain”. *Bernoulli*, 7:243–266.

Kessler, M. & Sørensen, M. (1999). “Estimating equations based on eigenfunctions for a discretely observed diffusion process”. *Bernoulli*, 5:299–314.

Kessler, M. & Sørensen, M. (2005). “On time-reversibility and estimating functions for Markov processes”. *Statistical Inference for Stochastic Processes*, 8:95–107.

Kimball, B. F. (1946). “Sufficient statistical estimation functions for the parameters of the distribution of maximum values”. *Ann. Math. Statist.*, 17:299–309.

Kloeden, P. E. & Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations*. 3rd revised printing. Springer-Verlag, New York.

Kloeden, P. E.; Platen, E.; Schurz, H. & Sørensen, M. (1996). “On effects of discretization on estimates of drift parameters for diffusions processes”.

J. Appl. Prob., 33:1061–1076.

Kuersteiner, G. (2002). “Efficient instrumental variables estimation for autoregressive models with conditional heteroskedasticity”. *Econometric Theory*, 18:547–583.

Kusuoka, S. & Yoshida, N. (2000). “Malliavin calculus, geometric mixing, and expansion of diffusion functionals”. *Probability Theory and Related Fields*, 116:457–484.

Kutoyants, Y. (1994). *Identification of dynamical systems with small noise*. Kluwer.

Kutoyants, Y. (1998). “Semiparametric estimation for dynamical systems with small noise”. *Math. Methods Statist.*, 7:457–465.

Larsen, K. S. & Sørensen, M. (2007). “A diffusion model for exchange rates in a target zone”. *Mathematical Finance*, 17:285–306.

Le Cam, L. & Yang, G. L. (2000). *Asymptotics in Statistical. Some Basic Concepts. 2nd Edition*. Springer, New York.

Li, B. (1997). “On the consistency of generalized estimating equations”. In Basawa, I. V.; Godambe, V. P. & Taylor, R. L., editors, *Selected Proceedings of the Symposium on Estimating Functions*, pages 115–136. Hayward: Institute of Mathematical Statistics. IMS Lecture Notes – Monograph Series, Vol. 32.

Liang, K.-Y. & Zeger, S. L. (1986). “Longitudinal data analysis using generalized linear model”. *Biometrika*, 73:13–22.

Longstaff, F. & Schwartz, E. (1995). “A simple approach to valuing risky fixed and floating rate debt”. *J. Finance*, 1:789–819.

McLeish, D. L. & Small, C. G. (1988). *The Theory and Applications of Statistical Inference Functions*. Springer-Verlag, New York. Lecture Notes in Statistics 44.

Nagahara, Y. (1996). “Non-Gaussian distribution for stock returns and related stochastic differential equation”. *Financial Engineering and the Japanese Markets*, 3:121–149.

Newey, W. K. (1990). “Efficient instrumental variables estimation of nonlinear models”. *Econometrica*, 58:809–837.

Newey, W. K. & West, K. D. (1987). “Hypothesis testing with efficient method of moments estimation”. *International Economic Review*, 28:777–787.

Overbeck, L. & Rydén, T. (1997). “Estimation in the Cox-Ingersoll-Ross model”. *Econometric Theory*, 13:430–461.

Ozaki, T. (1985). “Non-linear time series models and dynamical systems”. In Hannan, E. J.; Krishnaiah, P. R. & Rao, M. M., editors, *Handbook of Statistics, Vol. 5*, pages 25–83. Elsevier Science Publishers.

Pearson, K. (1895). “Contributions to the Mathematical Theory of Evolution II. Skew Variation in Homogeneous Material”. *Philosophical Transactions of the Royal Society of London. A*, 186:343–414.

Pedersen, A. R. (1994). “Quasi-likelihood inference for discretely observed diffusion processes”. Research Report No. 295, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus.

Pedersen, A. R. (1995). “A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations”. *Scand. J. Statist.*, 22:55–71.

Picard, J. (1986). “Nonlinear filtering of one-dimensional diffusions in the case of a high signal-to-noise ratio”. *SIAM J. Appl. Math.*, 46:1098–1125.

Picard, J. (1991). “Efficiency of the extended Kalman filter for nonlinear system with small noise”. *SIAM J. Appl. Math.*, 51:843–885.

Pokern, Y.; Stuart, A. M. & Wiberg, P. (2009). “Parameter estimation for partially observed hypoelliptic diffusions”. *J. Roy. Statist. Soc. B*, 71:49–73.

Poulsen, R. (1999). “Approximate maximum likelihood estimation of discretely observed diffusion processes”. Working Paper 29, Centre for Analytical Finance, Aarhus.

Prakasa Rao, B. L. S. (1988). “Statistical inference from sampled data for stochastic processes”. *Contemporary Mathematics*, 80:249–284.

Prentice, R. L. (1988). “Correlated binary regression with covariates specific to each binary observation”. *Biometrics*, 44:1033–1048.

Roberts, G. O. & Stramer, O. (2001). “On inference for partially observed nonlinear diffusion models using Metropolis-Hastings algorithms”. *Biometrika*, 88:603–621.

Romanovsky, V. (1924). “Generalization of some types of the frequency curves of Professor Pearson”. *Biometrika*, 16:106–117.

Sagan, J. D. (1958). “The estimation of economic relationsships using instrumental variables”. *Econometrica*, 26:393–415.

Skorokhod, A. V. (1989). *Asymptotic Methods in the Theory of Stochastic Differential Equations*. American Mathematical Society, Providence, Rhode Island.

Sørensen, H. (2001). “Discretely observed diffusions: Approximation of the continuous-time score function”. *Scand. J. Statist.*, 28:113–121.

Sørensen, M. (2000a). “Prediction-Based Estimating Functions”. *Econometrics Journal*, 3:123–147.

Sørensen, M. (2000b). “Small dispersion asymptotics for diffusion martingale estimating functions”. Preprint 2, Department of Statistics and Operation Research, University of Copenhagen.

Sørensen, M. (2007). "Efficient estimation for ergodic diffusions sampled at high frequency". Preprint, Department of Mathematical Sciences, University of Copenhagen.

Sørensen, M. & Uchida, M. (2003). "Small-diffusion asymptotics for discretely sampled stochastic differential equations". *Bernoulli*, 9:1051–1069.

Takahashi, A. & Yoshida, N. (2004). "An asymptotic expansion scheme for optimal investment problems". *Statistical Inference for Stochastic Processes*, 7:153–188.

Uchida, M. (2004). "Estimation for discretely observed small diffusions based on approximate martingale estimating functions". *Scand. J. Statist.*, 31:553–566.

Uchida, M. (2008). "Approximate martingale estimating functions for stochastic differential equations with small noises". *Stoch. Proc. Appl.*, 118:1706–1721.

Uchida, M. & Yoshida, N. (2004a). "Asymptotic expansion for small diffusions applied to option pricing". *Stat. Inference Stoch. Process.*, 7:189–223.

Uchida, M. & Yoshida, N. (2004b). "Information criteria for small diffusions via the theory of Malliavin-Watanabe". *Stat. Inference Stoch. Process.*, 7:35–67.

Veretennikov, A. Y. (1987). "Bounds for the mixing rate in the theory of stochastic equations". *Theory of Probability and its Applications*, 32:273–281.

Wedderburn, R. W. M. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method". *Biometrika*, 61:439–447.

West, K. D. (2001). "On optimal instrumental variables estimation of stationary time series models". *International Economic Review*, 42:1043–1050.

Wong, E. (1964). "The construction of a class of stationary Markoff processes". In Bellman, R., editor, *Stochastic Processes in Mathematical Physics and Engineering*, pages 264–276. American Mathematical Society, Rhode Island.

Yoshida, N. (1992). "Estimation for diffusion processes from discrete observations". *Journal of Multivariate Analysis*, 41:220–242.