Estimating functions for diffusion-type processes

Michael Sørensen
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark

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 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 3.2. 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 when enough data are available.

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).

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 asymptotic scenario, where the time between observations, Δ, is fixed, and the number of observations, n, goes to infinity. Since Δ is fixed, we simplify the notation and denote the observations by $X_1, X_2, \ldots, X_n$. We will generally suppress Δ in the notation in this section. As usual, we assume that the statistical model is indexed by a p-dimensional parameter θ, 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.

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

$$G_n(\theta) = \frac{1}{n} \sum_{i=r}^{n} g(X_{i-r+1}, \ldots, X_i; \theta), \tag{2.1}$$

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. A priori there is no guarantee that a unique solution to (1.1) exists. 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. Also a uniqueness result will be given.

We assume that, under the true probability measure, $\{X_i\}$ is a stationary process with state space $D \subseteq \mathbb{R}^d$. We let Q denote the joint distribution of $(X_1, \ldots, X_r)$, and $Q(f)$ the expectation of $f(X_1, \ldots, X_r)$ for a function $f : D^r \mapsto \mathbb{R}$. 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 $n \to \infty$

$$\frac{1}{n} \sum_{i=r}^{n} f(X_{i-r+1}, \ldots, X_i) \xrightarrow{P} Q(f) \tag{2.2}$$

for any function $f : D^r \mapsto \mathbb{R}$ such that $Q(|f|) < \infty$, and that the estimating function (2.1) satisfies

$$\frac{1}{\sqrt{n}} \sum_{i=r}^{n} g(X_{i-r+1}, \ldots, X_i; \theta) \xrightarrow{\mathcal{L}} N(0, V(\theta)) \tag{2.3}$$

under P for any θ 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 $\xrightarrow{\mathcal{L}}$ 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 2.1 *There is a parameter value $\bar{\theta} \in \operatorname{int} \Theta$ and a neighbourhood N of $\bar{\theta}$ in Θ, 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

$$Q\left(g(\bar{\theta})\right) = 0. \tag{2.4}$$

(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

$$W = Q\left(\partial_{\theta^T} g(\bar{\theta})\right) \tag{2.5}$$

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 $f : D^r \times \Theta \mapsto \mathbb{R}^q$ 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 2.2 *Assume Condition 2.1 and (2.3). Then a $\bar{\theta}$-consistent G_n–estimator $\hat{\theta}_n$ exists, and*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N_p\left(0, W^{-1} V W^{T^{-1}}\right) \tag{2.6}$$

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

$$Q(g(\theta)) \neq 0 \quad \text{for all} \quad \theta \neq \bar{\theta}, \tag{2.7}$$

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 $\hat{\theta}_n$ converges in probability to $\bar{\theta}$ 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 $\hat{\theta}_n$ is most useful if Theorem 2.2 holds with $\bar{\theta} = \theta_0$. Note that because $\bar{\theta} \in \Theta$, 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 2.2, we need the following uniform law of large numbers.

Lemma 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 $\theta \mapsto Q(f(\theta))$ is continuous, and*

$$\sup_{\theta \in K} \| \frac{1}{n} \sum_{i=r}^{n} f(X_{i-r+1}, \ldots, X_i; \theta) - Q(f(\theta)) \| \xrightarrow{P} 0. \tag{2.8}$$

Proof: That $Q(f(\theta))$ is continuous follows from the dominated convergence theorem. To prove (2.8), 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 $\|\theta - \theta'\| < \eta$ implies that $\|Q(f(\theta)) - Q(f(\theta'))\| \leq \epsilon$. 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

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

we have

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

where

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

By (2.2), $P(Z_n > \epsilon) \to 0$ 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 2.2): The existence of a $\bar{\theta}$-consistent G_n–estimator $\hat{\theta}_n$ follows from Theorem 8.2. Condition (i) follows from (2.2) and (2.4). Define the function $W(\theta) = Q\left(\partial_{\theta^T} g(\theta)\right)$. Then condition (iii) in Theorem 8.2 is equal to Condition 2.1 (4). Finally, let M be a compact subset of N containing $\bar{\theta}$. Then the conditions of Lemma 2.3 are satisfied for $f = \partial_{\theta^T} g$, so (8.1) is satisfied. The asymptotic normality, (3.15), follows from Theorem 8.5 and (2.3).

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 2.3. Hence (8.2) 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 (2.7) 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 8.3 it follows that (8.4) 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 (8.4) the estimator $\hat{\theta}''_n$ is consistent, so by Theorem 8.2, $P(\hat{\theta}_n \neq \hat{\theta}''_n) \to 0$ as $n \to \infty$. Hence $\hat{\theta}_n$ is eventually the unique G_n–estimator on K.

□

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

$$dX_t = b(X_t; \theta)dt + \sigma(X_t; \theta)dW_t, \tag{3.1}$$

where σ is a $d \times d$-matrix and W a d-dimensional standard Wiener process. We denote the state space 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 (3.1) 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

$$G_n(\theta) = \sum_{i=1}^{n} g(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta). \tag{3.2}$$

where g is p-dimensional function that satisfies that

$$\int_D g(\Delta, x, y; \theta) p(\Delta, x, y; \theta) dy = 0 \tag{3.3}$$

for all $\Delta > 0$, $x \in D$ and all $\theta \in \Theta$. 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*.

3.1 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

$$L_n(\theta) = \prod_{i=1}^{n} p(t_i - t_{i-1}, X_{t_{i-1}}, X_{t_i}; \theta), \tag{3.4}$$

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 4.

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

$$U_n(\theta) = \partial_\theta \log L_n(\theta) = \sum_{i=1}^{n} \partial_\theta \log p(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta), \tag{3.5}$$

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 presented in the next subsection applies to the maximum likelihood estimator. Asymptotic results

for the maximum likelihood estimator in the fixed Δ (low frequency) asymptotic scenario considered in this section 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).

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

$$F(\Delta, x; \theta) = \mathrm{E}_\theta(X_\Delta | X_0 = x) = \int_\ell^r y p(\Delta, x, y; \theta) dy. \tag{3.6}$$

and

$$\begin{aligned} &\phi(\Delta, x; \theta) = \\ &\quad \mathrm{Var}_\theta(X_\Delta | X_0 = x) = \int_\ell^r [y - F(\Delta, x; \theta)]^2 p(\Delta, x, y; \theta) dy. \end{aligned} \tag{3.7}$$

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

$$\begin{aligned} \partial_\theta \log QL_n(\theta) = \sum_{i=1}^n \Bigg\{ &\frac{\partial_\theta F(\Delta_i, X_{t_{i-1}}; \theta)}{\phi(\Delta_i, X_{t_{i-1}}; \theta)} [X_{t_i} - F(\Delta_i, X_{t_{i-1}}; \theta)] \\ &+ \frac{\partial_\theta \phi(\Delta_i, X_{t_{i-1}}; \theta)}{2\phi(\Delta_i, X_{t_{i-1}}; \theta)^2} \left[(X_{t_i} - F(\Delta_i, X_{t_{i-1}}; \theta))^2 - \phi(\Delta_i, X_{t_{i-1}}; \theta)\right] \Bigg\}. \end{aligned} \tag{3.8}$$

It is clear from (3.6) and (3.7) 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).

3.2 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 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

$$s(x;\theta) = \exp\left(-2\int_{x^\#}^{x} \frac{b(y;\theta)}{\sigma^2(y;\theta)}dy\right), \quad x \in (\ell, r), \tag{3.9}$$

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

Condition 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 3.1 the process X is ergodic with an invariant probability measure with Lebesgue density

$$\mu_\theta(x) = [A(\theta)s(x;\theta)\sigma^2(x;\theta)]^{-1}, \quad x \in (\ell, r). \tag{3.10}$$

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

$$Q_\theta(dx, dy) = \mu_\theta(x)p(\Delta, x, y; \theta)dxdy. \tag{3.11}$$

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 (2.2) 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 (3.2)

$$\begin{aligned} &Q_\theta\left(g(\theta)^T g(\theta)\right) = \\ &\quad \int_{D^2} g(y, x; \theta)^T g(y, x; \theta)\mu_\theta(x)p(x, y; \theta)dydx < \infty, \end{aligned} \tag{3.12}$$

for all $\theta \in \Theta$. By (2.2),

$$\frac{1}{n}\sum_{i=1}^{n} g(X_{\Delta i}, X_{\Delta(i-1)}; \theta') \xrightarrow{P_\theta} Q_\theta(g(\theta')). \tag{3.13}$$

Since the estimating function $G_n(\theta)$ is a martingale under P_θ, the asymptotic normality in (2.3) follows without further conditions from the central limit theorem for martingales, see Hall & Heyde (1980). This result goes back to Billingsley (1961). In the martingale case the asymptotic covariance matrix $V(\theta)$ in (2.3) is given by

$$V(\theta) = Q_{\theta_0}\left(g(\theta)g(\theta)^T\right). \tag{3.14}$$

Thus we have the following particular case of Theorem 2.2.

Theorem 3.2 *Assume Condition 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 (2.3) holds for $\theta = \theta_0$ with $V(\theta)$ given by (3.14). Then a θ_0-consistent G_n–estimator $\hat{\theta}_n$ exists, and*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N_p\left(0, W^{-1}VW^{T^{-1}}\right) \tag{3.15}$$

under P_{θ_0}, where W is given by (2.5) with $\bar{\theta} = \theta_0$ and $V = V(\theta_0)$. If, moreover, the function $g(x, y; \theta)$ is locally dominated integrable with respect to Q_{θ_0} and

$$Q_{\theta_0}(g(\theta)) \neq 0 \quad \text{for all} \quad \theta \neq \theta_0, \tag{3.16}$$

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 3.2 hold for a neighbourhood of 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 3.3 *Under Condition 2.1 (2) – (4) (with $r = 2$, $\bar{\theta} = \theta_0$, and $Q = Q_{\theta_0}$),*

$$W_n = \frac{1}{n}\sum_{i=1}^{n} \partial_{\theta^T} g(X_{(i-1)\Delta}, X_{i\Delta}; \hat{\theta}_n) \xrightarrow{P_{\theta_0}} W, \tag{3.17}$$

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

$$V_n = \frac{1}{n}\sum_{i=1}^{n} g(X_{(i-1)\Delta}, X_{i\Delta}; \hat{\theta}_n) g(X_{(i-1)\Delta}, X_{i\Delta}; \hat{\theta}_n)^T \xrightarrow{P_{\theta_0}} V. \tag{3.18}$$

Proof: Let C be a compact subset of N such that $\theta_0 \in \operatorname{int} C$. By Lemma 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 (3.17) 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 2.3. □

In the case of likelihood inference, the function $Q_{\theta_0}(g(\theta))$ appearing in the identifiability condition (3.16) 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.

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

$$\int_D h_j(x, y; \theta) p(x, y; \theta) dy = 0 \tag{3.19}$$

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 (2.1), 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

$$G_n(\theta) = \sum_{i=1}^{n} a(X_{(i-1)\Delta}, \theta) h(X_{(i-1)\Delta}, X_{i\Delta}; \theta), \tag{3.20}$$

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 (3.20) is P_θ-integrable. It follows from (3.19) 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 9. The *optimal weight matrix* a^* gives the estimating function of the form (3.20) that provides the best possible approximation to the score function (3.5) 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 9.

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 Sections 3.6 and 3.7, we shall present ways of choosing these functions that usually work well in practice.

Example 3.4 The martingale estimating function (3.8) is of the type (3.20) 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 (3.6) and (3.7). The weight matrix is

$$\left(\frac{\partial_\theta F(\Delta, x; \theta)}{\phi(\Delta, x; \theta)}, \quad \frac{\partial_\theta \phi(\Delta, x; \theta)}{2\phi^2(\Delta, x; \theta)\Delta} \right), \tag{3.21}$$

which we shall see is approximately optimal. □

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 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 (2.2)

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

Hence the estimating function $\partial_\theta H_n(\theta)$ is asymptotically equivalent to an estimating function of the form (3.20) with a constant weight matrix

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

and we see that GMM-estimators 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 (3.20) with the optimal weight matrix. We assume that the functions h_j satisfy the following condition.

Condition 3.5

(1) The functions h_j, $j = 1, \ldots N$, are linearly independent.
(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 9.3. By (9.16), the optimal choice of the weight matrix a is given by

$$a^*(x; \theta) = B_h(x; \theta)\, V_h(x; \theta)^{-1}, \tag{3.22}$$

where

$$B_h(x;\theta) = \int_D \partial_\theta h(x,y;\theta)^T p(x,y;\theta)dy \tag{3.23}$$

and

$$V_h(x;\theta) = \int_D h(x,y;\theta)h(x,y;\theta)^T p(x,y;\theta)dy. \tag{3.24}$$

The matrix $V_h(x;\theta)$ is invertible because the functions h_j, $j = 1,\ldots N$ are linearly independent.Compared to (9.16), 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 (3.15) because in this case the matrices W and V given by (2.5) and (3.14) are equal and given by (3.25). This is a general property of optimal estimating functions as discussed in Section 9. The result can easily be verified under the assumption that $a^*(x;\theta)$ is a differentiable function of θ: by (3.19)

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

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

$$V = \mu_{\theta_0}\left(B_h(\theta_0)V_h(\theta_0)^{-1}B_h(\theta_0)^T\right). \tag{3.25}$$

Thus we have as a corollary to Theorem 2.2 that is $g^*(x,y,\theta) = a^*(x;\theta)h(x,y;\theta)$ satisfies the conditions of Theorem 2.2, then a sequence $\hat{\theta}_n$ of G_n^*–estimators has the asymptotic distribution

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N_p\left(0, V^{-1}\right). \tag{3.26}$$

Example 3.6 Consider the martingale estimating function of form (3.20) with $N = 2$ and with h_1 and h_2 as in Example 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 CIR-model)

$$dX_t = -\beta(X_t - \alpha)dt + \tau\sqrt{X_t}dW_t, \tag{3.27}$$

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

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

We give a method to derive these expression in Section 3.6.

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

$$\eta(t, x; \theta) \approx 0 \quad \text{and} \quad \psi(t, x; \theta) \approx 2\phi(t, x; \theta)^2, \tag{3.28}$$

which would be exactly true if transition density were a normal distribution. If we insert the Gaussian approximations into the expressions for a_1^* and a_2^*, we obtain the weight functions in (3.8). When Δ is not large this can be justified, because the transition distribution is not far from Gaussian. □

In Subsections 3.6 and 3.7 we shall 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 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

$$G_n^\bullet(\theta) = \sum_{i=1}^n a^*(X_{(i-1)\Delta}; \tilde{\theta}_n) h(X_{(i-1)\Delta}, X_{i\Delta}; \theta), \tag{3.29}$$

where $\tilde{\theta}_n$ is a weakly 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. Under weak regularity conditions, the $G_n^\bullet$-estimator has the same efficiency as the optimal G_n^*-estimator; see e.g. Jacod & Sørensen (2008).

Most martingale estimating functions proposed in the literature are of the form (3.20) with

$$h_j(x, y; \theta) = f_j(y; \theta) - \pi_\Delta^\theta(f_j(\theta))(x), \tag{3.30}$$

or more specifically,

$$G_n(\theta) = \sum_{i=1}^n a(X_{(i-1)\Delta}, \theta)\left[f(X_{i\Delta}; \theta) - \pi_\Delta^\theta(f(\theta))(X_{(i-1)\Delta})\right]. \tag{3.31}$$

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

$$\pi_s^\theta(f)(x) = \int_D f(y)p(s, x, y; \theta)dy = \mathrm{E}_\theta(f(X_s) \mid X_0 = x), \tag{3.32}$$

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 (3.31), the expression for the optimal weight matrix simplifies a bit because

$$B_h(x;\theta)_{ij} = \pi^\theta_\Delta(\partial_{\theta_i} f_j(\theta))(x) - \partial_{\theta_i}\pi^\theta_\Delta(f_j(\theta))(x), \tag{3.33}$$

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

$$V_h(x;\theta)_{ij} = \pi^\theta_\Delta(f_i(\theta)f_j(\theta))(x) - \pi^\theta_\Delta(f_i(\theta))(x)\pi^\theta_\Delta(f_j(\theta))(x), \tag{3.34}$$

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

$$B_h(x;\theta)_{ij} = -\partial_{\theta_i}\pi^\theta_\Delta(f_j)(x). \tag{3.35}$$

A useful *approximations to the optimal weight matrix* can be obtained by applying the formula

$$\pi^\theta_s(f)(x) = \sum_{i=0}^{k} \frac{s^i}{i!} A^i_\theta f(x) + O(s^{k+1}), \tag{3.36}$$

where A_θ denotes the *generator* of the diffusion

$$A_\theta f(x) = \sum_{k=1}^{d} b_k(x;\theta)\partial_{x_k} f(x) + \tfrac{1}{2}\sum_{k,\ell=1}^{d} C_{k\ell}(x;\theta)\partial^2_{x_k x_\ell} f(x), \tag{3.37}$$

where $C = \sigma\sigma^T$. The formula (3.36) holds for $2(k+1)$ times continuously differentiable functions under weak conditions which ensure that the remainder term has the correct order, see Kessler (1997) and Subsection 3.4. 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 $d = 1$

$$B_h(x;\theta) \approx \Delta\left[\partial_\theta b(x;\theta) f'(x) + \tfrac{1}{2}\partial_\theta \sigma^2(x;\theta) f''(x)\right] \tag{3.38}$$

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

$$V_h(x;\theta) \approx \Delta\left[A_\theta(f^2)(x) - 2f(x)A_\theta f(x)\right] = \Delta\,\sigma^2(x;\theta) f'(x)^2. \tag{3.39}$$

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 (3.36).

Example 3.7 If we simplify the optimal weight matrix found in Example 3.6 by the expansion (3.36) and the Gaussian approximation (3.28), we obtain the approximately optimal quadratic martingale estimating function

$$\begin{aligned} G^\circ_n(\theta) = \sum_{i=1}^{n} \Bigg\{ & \frac{\partial_\theta b(X_{(i-1)\Delta};\theta)}{\sigma^2(X_{(i-1)\Delta};\theta)}[X_{i\Delta} - F(X_{(i-1)\Delta};\theta)] \\ & + \frac{\partial_\theta \sigma^2(X_{(i-1)\Delta};\theta)}{2\sigma^4(X_{(i-1)\Delta};\theta)\Delta}\left[(X_{i\Delta} - F(X_{(i-1)\Delta};\theta))^2 - \phi(X_{(i-1)\Delta};\theta)\right]\Bigg\}. \end{aligned} \tag{3.40}$$

As in Example 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 $f(t) = \mathrm{E}_\theta(X_t \,|\, X_0 = x)$ 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 (3.20) 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 let $\bar{X}^\kappa = \sum_{i=1}^n w_i^\kappa X_{i\Delta}$ and $\bar{X}_{-1}^\kappa = \sum_{i=1}^n w_i^\kappa X_{(i-1)\Delta}$ be 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 for ergodic processes, the probability that $r_n^\kappa > 0$ tends to one as n tends to infinity. Specifically, we obtain the explicit estimators

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

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 (3.27), 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).

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 3.6, $\phi(x;\theta)$ is explicitly known for the *square-root process (CIR-model)* given by (3.27). For this model the approximately optimal quadratic martingale estimating function is

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

This expression is obtained from (3.40) 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

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

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 for ergodic processes. □

When the optimal weight matrix is approximated by means of (3.36), there is a certain loss of efficiency, which as in the previous example is often quite small; see Bibby & Sørensen (1995) and Section 6 on high frequency asymptotics below. Therefore the relatively simple estimating function (3.40) is often a good choice in practice.

It is tempting to go on to approximate $\pi_\Delta^\theta(f_j(\theta))(x)$ in (3.31) by (3.36) in order to obtain an explicit estimating function, but as will be demonstrated in Subsection 3.6, 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 3.5. Moreover, as we shall also see in Subsection 3.6, $\pi_\Delta^\theta(f_j(\theta))(x)$ can be found explicitly for a quite flexible class of diffusions.

3.4 Small Δ-optimality

The Godambe-Heyde optimal estimating functions discussed above are optimal within a 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 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 (3.20) and (3.30) 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

$$dX_t = b(X_t;\alpha)dt + \sigma(X_t;\beta)dW_t, \tag{3.41}$$

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 (3.2) satisfies the conditions of Theorem 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 (2.5) with $\bar{\theta} = \theta_0$ and $V = V(\theta_0)$ with $V(\theta)$ given by (3.14).

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

$$M(g) = \frac{1}{\Delta} v_{-1}(g) + v_0(g) + o(1). \tag{3.42}$$

Small Δ-optimal estimating functions minimize the leading term in (3.42). Jacobsen (2001) obtained (3.42) 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 $(x,y) \in D^2$ that are continuous differentiable in s and twice continuously differentiable in y. The operator $\mathcal{A}_\theta$ is given by

$$\mathcal{A}_\theta \varphi(s,x,y) = \partial_s \varphi(s,x,y) + A_\theta \varphi(s,x,y), \tag{3.43}$$

where A_θ is the generator (3.37) , which for every 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 $(t, X_t)_{t<geq0}$. We also need the probability measure Q_θ^Δ given by (3.11). Note that in this section the dependence on Δ is explicit in the notation.

Condition 3.8 *The function φ belongs to Γ and satisfies that*

$$\begin{aligned}
&\int_{D^2} \varphi(s,x,y) Q_{\theta_0}^s(dx,dy) < \infty \\
&\int_{D^2} (\mathcal{A}_{\theta_0}\varphi(s,x,y))^2 Q_{\theta_0}^s(dx,dy) < \infty \\
&\int_{D^2} (\partial_y \varphi(s,x,y))^2 \sigma^2(y;\beta_0) Q_{\theta_0}^s(dx,dy) < \infty
\end{aligned}$$

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 3.8 is each component of the functions satisfies this condition.

Suppose φ satisfies Condition 3.8. Then by Ito's formula

$$\begin{aligned}
&\varphi(t,X_0,X_t) = \\
&\quad \varphi(0,X_0,X_0) + \int_0^t \mathcal{A}_{\theta_0}\varphi(s,X_0,X_s)ds + \int_0^t \partial_y \varphi(s,X_0,X_s) dW_s
\end{aligned} \tag{3.44}$$

under P_{θ_0}. A significant consequence of Condition 3.8 is that the Ito-integral in (3.44) 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 3.8, a similar result holds for this functions, which we can insert in the Lebesgue integral in (3.44). By doing so and then taking the conditional expectation given $X_0 = x$ on both sides of (3.44), we obtain

$$\pi_t^{\theta_0}(\varphi)(t,x) = \varphi(0,x,x) + t\mathcal{A}_{\theta_0}\varphi(0,x,x) + O(t^2), \tag{3.45}$$

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 3.8, where $\mathcal{A}_{\theta_0}^i$ denotes i-fold application of the operator $\mathcal{A}_{\theta_0}$, we obtain by similar arguments that

$$\pi_t^{\theta_0}(\varphi)(t,x) = \sum_{i=0}^k \frac{s^i}{i!}\mathcal{A}_{\theta_0}^i\varphi(0,x,x) + O(t^{k+1}). \tag{3.46}$$

Note that $\mathcal{A}_\theta^0$ is the identity operator. The previously used expansion (3.36) is a particular case of (3.46). In the case where φ does not depend on x (or y) the integrals in Condition 3.8 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 3.9 *Suppose that the function $g(\Delta, x, y; \theta_0)$ in (3.2) is such that g, $\partial_{\theta^T} g$, gg^T and $\mathcal{A}_{\theta_0} g$ satisfy Condition 3.8. 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

$$S = \int_\ell^r B_{\theta_0}(x) \mu_{\theta_0}(x) dx \tag{3.47}$$

is invertible, where

$$B_\theta(x) = \begin{pmatrix} \partial_\alpha b(x;\alpha) \partial_y g_1(0,x,x;\theta) & \frac{1}{2} \partial_\beta v(x;\beta) \partial_y^2 g_1(0,x,x;\theta) \\ \partial_\alpha b(x;\alpha) \partial_y g_2(0,x,x;\theta) & \frac{1}{2} \partial_\beta v(x;\beta) \partial_y^2 g_2(0,x,x;\theta) \end{pmatrix}, \tag{3.48}$$

then (3.42) holds with

$$v_{-1}(g) \geq \begin{pmatrix} \left(\int_\ell^r \left(\partial_\alpha b(x;\alpha_0) \right)^2 / \sigma^2(x;\beta_0) \mu_{\theta_0}(x) dx \right)^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.49}$$

These is equality in (3.49) if

$$\partial_y g_1(0,x,x;\theta_0) = \partial_\alpha b(x;\alpha_0) / \sigma^2(x;\beta_0), \tag{3.50}$$

$$\partial_y g_2(0,x,x;\theta_0) = 0 \tag{3.51}$$

for all $x \in (\ell, r)$. In this case, the second term in (3.42) 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

$$\partial_y^2 g_2(0,x,x;\theta_0) = \partial_\beta \sigma^2(x;\beta_0) / \sigma^2(x;\beta_0)^2, \tag{3.52}$$

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

Thus the conditions for small Δ-optimality are (3.50), (3.51) and (3.52). For a proof of Theorem 3.9, see Jacobsen (2001). The condition (3.51) 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 (3.42) 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 (3.51) 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 (3.50) – (3.52). The point is that a version must exist which satisfies these conditions.

Example 3.10 Consider a quadratic martingale estimating function of the form

$$g(\Delta, y, x; \theta) = \begin{pmatrix} a_1(x, \Delta; \theta)[y - F(\Delta, x; \theta)] \\ a_2(x, \Delta; \theta)\left[(y - F(\Delta, x; \theta))^2 - \phi(\Delta, x; \theta)\right] \end{pmatrix}, \tag{3.53}$$

where F and ϕ are given by (3.6) and (3.7). By (3.36), $F(\Delta, x; \theta) = x + O(\Delta)$ and $\phi(\Delta, x; \theta) = O(\Delta)$, so

$$g(0, y, x; \theta) = \begin{pmatrix} a_1(x, 0; \theta)(y - x) \\ a_2(x, 0; \theta)(y - x)^2 \end{pmatrix}. \tag{3.54}$$

Since $\partial_y g_2(0, y, x; \theta) = 2a_2(x, \Delta; \theta)(y - x)$, the Jacobsen condition (3.51) is satisfied for all quadratic martingale estimating functions. Using again (3.36), it is not difficult to see that the two other conditions (3.50) and (3.52) are satisfied in three particular cases: the optimal estimating function given in Example 3.6 and the approximations (3.8) and (3.40). □

The following theorem gives conditions ensuring, for given functions $f_1, \ldots, f_N$, that a small Δ-optimal estimating function of the form (3.20) and (3.30) 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 3.11 *Suppose that $N \geq 2$, that the functions f_j are twice continuously differentiable and satisfies that the matrix*

$$D(x) = \begin{pmatrix} \partial_x f_1(x; \theta) & \partial_x^2 f_1(x; \theta) \\ \partial_x f_2(x; \theta) & \partial_x^2 f_2(x; \theta) \end{pmatrix} \tag{3.55}$$

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 $a(x; \theta)$, independent of Δ, exists such that the estimating function (3.20) satisfies the conditions (3.51), (3.50) and (3.52). When $N = 2$, these conditions are satisfy for

$$a(x; \theta) = \begin{pmatrix} \partial_\alpha b(x; \alpha)/v(x; \beta) & c(x; \theta) \\ 0 & \partial_\beta v(x; \beta)/v(x; \beta)^2 \end{pmatrix} D(x)^{-1} \tag{3.56}$$

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

For a proof of Theorem 3.11, see Jacobsen (2002). In Section 6, we shall see that the Godambe-Heyde optimal choice (3.22) of the weight-matrix in (3.20) gives an estimating function which has a version that satisfies the conditions for small Δ-optimality, (3.50) – (3.52).

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

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

In the latter equation, $\mathrm{vec}(M)$ 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 $\mathrm{vec}(M)\, M^\bullet$ 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 3.11 ensuring the existence of a small Δ-optimal estimating function of the form (3.20) 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 $d(d+3)/2$.

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 3.6 and 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 $X_0 = x$. 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 (3.36). Methods for numerical simulation of diffusion models can be found in Kloeden & Platen (1999).

The approach just described is sufficient when calculating the conditional expectation appearing in (3.30), although it is important to use the same random numbers (seed) when calculating the estimating functions for different values of the parameter θ, for instance when using a search algorithm to find a solution to the estimating equation. More care is needed if the optimal weight functions are calculated numerically. The problem is that the

optimal weight matrix typically contain derivatives with respect to θ of functions that must be determined numerically, see e.g. Example 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 (3.36), possibly supplemented by Gaussian approximations, as explained in Subsection 3.3. This is not only much simpler, but also avoids potentially serious problems of numerical instability, and by results in Section 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

$$G_n(\theta) = \sum_{i=1}^{n} \left[f(X_{i\Delta}, X_{(i-1)\Delta}; \theta) - F(X_{(i-1)\Delta}; \theta) \right], \tag{3.57}$$

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 (3.1). 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 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

$$Y_{i\delta} = Y_{(i-1)\delta} + b(Y_{(i-1)\delta}; \theta)\delta + \sigma(Y_{(i-1)\delta}; \theta) Z_i, \qquad Y_0 = x, \tag{3.58}$$

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 (3.57)

$$G_n^{M,\delta}(\theta) = \sum_{i=1}^{n} \left[f(X_{i\Delta}, X_{(i-1)\Delta}; \theta) - \frac{1}{M} \sum_{j=1}^{M} f(Y_\Delta^{(j)}(\delta, \theta, X_{(i-1)\Delta}), X_{(i-1)\Delta}; \theta) \right], \tag{3.59}$$

where $Y^{(j)}(\delta, \theta, x)$, $j = 1, \ldots, N$ 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

$$|E_\theta(g(X_\Delta(x, \theta), x; \theta)) - E(g(Y_\Delta(\delta, \theta, x), x; \theta))| \leq R(x; \theta)\delta^\beta \tag{3.60}$$

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 $\sup_{\theta \in K} |R(x;\theta)| \leq C_1(1+|x|^{C_2})$ for all x in the state space of the diffusion. The inequality (3.60) 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 (3.60). In particular, the Euler scheme (3.58) is of weak order one if the coefficients of the stochastic differential equation (3.1) are smooth enough.

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 (3.57). 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 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 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 (3.59). 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.

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 (3.31) where the functions f_j, $i = 1, \ldots, N$ are *eigenfunctions* for the generator (3.37), 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, i.e.

$$\pi_t^\theta(f_j(\theta))(x) = e^{-\lambda_j(\theta)t} f_j(x;\theta). \tag{3.61}$$

for all $t > 0$. Thus the function h_j in (3.30) is explicit.

Theorem 3.12 *Let $\phi(x;\theta)$ be an eigenfunction for the generator (3.37) with eigenvalue $\lambda(\theta)$. Suppose*

$$\int_\ell^r [\partial_x \phi(x;\theta)\sigma(x;\theta)]^2 \mu_\theta(dx) < \infty \tag{3.62}$$

for all $t > 0$. Then

$$\pi_t^\theta(\phi(\theta))(x) = e^{-\lambda(\theta)t}\phi(x;\theta). \tag{3.63}$$

for all $t > 0$.

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

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

so by (3.62), Y is a true martingale, which implies (3.63). □

Note that if $\sigma(x;\theta)$ and $\partial_x\phi(x;\theta)$ are bounded functions of $x \in (\ell, r)$, then (3.62) holds. If ϕ is a polynomial of order k and $\sigma(x) \le C(1+x^m)$, then (3.62) holds if the $2(k+m-1)$'th moment of the invariant distribution μ_θ is finite.

Example 3.13 For the square-root model (CIR-model) defined by (3.27) 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

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

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 3.12, (3.63) holds, so we can calculate all conditional polynomial moments, of which the first four were given in Example 3.6. Thus all polynomial martingale estimating functions are explicit.

□

Example 3.14 The diffusion given as the solution of

$$dX_t = -\theta \tan(X_t)dt + dW_t, \tag{3.64}$$

is an ergodic diffusion on the interval $(-\pi/2, \pi/2)$ provided that $\theta \ge 1/2$, so that Condition 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 a 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 (3.62) in Theorem 3.12 is obviously satisfied because the state space is bounded, so (3.63) holds.

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

$$\check{G}_n(\theta) \sum_{i=1}^{n} \sin(X_{(i-1)\Delta})[\sin(X_{i\Delta})) - e^{-(\theta+1/2)\Delta} \sin(X_{(i-1)\Delta}))], \tag{3.65}$$

we obtain the very simple estimator for θ

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

which is defined when the numerator is positive.

An asymmetric generalization of (3.64) 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 (3.64) is obtained is when $\varphi = 0$, $m = 0$, and $z = \pi/2$. 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

$$\phi_i(y;\theta) = \sum_{j=0}^{i} a_{i,j}(\theta)\,\kappa(y)^j \tag{3.66}$$

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 (3.22) can be found explicitly too.

Theorem 3.15 *Suppose $2N$ eigenfunctions are of the form (3.66) 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 (3.30) using the first N eigenfunctions, then*

$$B_h(x,\theta)_{ij} = \sum_{k=0}^{j} \left(\partial_{\theta_i} a_{j,k}(\theta)\nu_k(x;\theta) - \partial_{\theta_i}[e^{-\lambda_j(\theta)\Delta}\phi_j(x;\theta)]\right) \tag{3.67}$$

and

$$V_h(x,\theta)_{i,j} = \sum_{r=0}^{i}\sum_{s=0}^{j} \left(a_{i,r}(\theta)a_{j,s}(\theta)\nu_{r+s}(x;\theta) - e^{-[\lambda_i(\theta)+\lambda_j(\theta)]\Delta}\,\phi_i(x;\theta)\phi_j(x;\theta)\right), \tag{3.68}$$

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

$$e^{-\lambda_i(\theta)\Delta}\phi_i(x;\theta) = \sum_{j=0}^{i} a_{i,j}(\theta)\nu_j(x;\theta) \quad i = 1, \ldots, 2N, \tag{3.69}$$

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

Proof: The expressions for B_h and V_h follow from (3.33) and (3.34) when the eigenfunctions are of the form (3.66), and (3.69) follows by applying π_Δ^θ to both sides of (3.66).

□

Example 3.16 Consider again the diffusion (3.64) in Example 3.14. We will find the optimal martingale estimating function based on the first non-trivial eigenfunction, $\sin(x)$ (where we have neglected a non-essential multiplicative function of θ) with eigenvalue $\theta + 1/2$. It follows from (3.33) 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 3.15. 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

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

where a constant has been omitted. When Δ is small it is a good idea to multiply $G_n^\circ(\theta)$ 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 $G_n^\circ(\theta) = 0$ numerically. Note however that for $\tilde{G}_n$ the square integrability under Q_{θ_0} (3.12) required in Theorem 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, which it is close to when the process is not near the boundaries. In that way we arrive at the simple estimating function (3.65), which is thus also approximately optimal.

□

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

$$dX_t = -\beta(X_t - \alpha)dt + \sqrt{2\beta(aX_t^2 + bX_t + c)}dW_t, \tag{3.70}$$

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 (3.37) maps polynomials into polynomials. It is therefore easy to find eigenfunctions among the polynomials

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

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, equating the coefficients we find that the eigenvalue is given by

$$\lambda_n = a_n = n\{1-(n-1)a\}\beta. \tag{3.71}$$

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

$$(a_j - a_n)p_{n,j} = b_{j+1}p_{n,j+1} + c_{j+2}p_{n,j+2}. \tag{3.72}$$

Equation (3.72) 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 (3.62) if $p_n(x)$ is square integrable with respect to the invariant distribution, which is the case if $a < (2n-1)^{-1}$. By Theorem 3.12 this implies that the transition operator satisfies (3.63), so that $p_n(x)$ can be used to construct *explicit optimal martingale estimating functions* as explained in Subsection 3.6. For Pearson diffusions with $a \leq 0$, $a < (2n-1)^{-1}$ is automatically satisfied, and 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. 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. (3.9). Since the density of the invariant probability measure is given by

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

cf. (3.10), it follows that

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

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 (3.70) 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 that 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 (3.70) 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 (3.70) is the square root process (CIR process) (3.27) with state space $(0, \infty)$. Condition 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 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 $k < 1 + a^{-1}$. With 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 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 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 $k < 1 + \frac{1}{a}$. 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 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 (3.70) exists. If $0 < \alpha < 1$, the boundary 0

can with positive probability be reached at a finite time point, 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 $k < 1 + \frac{1}{a}$.

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 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 (3.70) exists. If $0 < \alpha < -a$, the boundary 0 can with positive probability be reached at a finite time point, 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 $\frac{\alpha}{-a}$, $\frac{1-\alpha}{-a}$. 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 3.17 The *skew t-distribution* with mean zero, ν degrees of freedom, and skewness parameter ρ has (unnormalized) density

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

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

$$dZ_t = -\beta Z_t dt + \sqrt{2\beta(\nu-1)^{-1}\{Z_t^2 + 2\rho\nu^{\frac{1}{2}} Z_t + (1+\rho^2)\nu\}}\, dW_t. \tag{3.73}$$

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

The skew t-diffusion (3.73) 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

$$\begin{aligned}
p_4(z) &= z^4 - \frac{24\rho\nu^{\frac{1}{2}}}{\nu-7} z^3 + \frac{144\rho^2\nu - 6(1+\rho^2)\nu(\nu-7)}{(\nu-7)(\nu-6)} z^2 \\
&\quad + \frac{8\rho(1+\rho^2)\nu^{\frac{3}{2}}(\nu-7) + 48\rho(1+\rho^2)\nu^{\frac{3}{2}}(\nu-6) - 192\rho^3\nu^{\frac{3}{2}}}{(\nu-7)(\nu-6)(\nu-5)} z \\
&\quad + \frac{3(1+\rho^2)^2\nu(\nu-7) - 72\rho^2(1+\rho^2)\nu^2}{(\nu-7)(\nu-6)(\nu-4)},
\end{aligned}$$

provided that $\nu > 4$ If $\nu > 2i$ the first i eigenfunctions are square integrable and thus satisfy (3.62). Hence (3.63) holds, and the eigenfunctions can be used to construct explicit martingale estimating functions. □

Example 3.18 The model

$$dX_t = -\beta[X_t - (m + \gamma z)]dt + \sigma\sqrt{z^2 - (X_t - m)^2}dW_t \tag{3.74}$$

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 (3.74) 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 (3.62) is obviously satisfied, (3.63) 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 on (3.61). Specifically,

$$E(X_t^n \mid X_0 = x) = \sum_{k=0}^{n} \sum_{\ell=0}^{n} q_{n,k,\ell} \cdot e^{-\lambda_\ell t} \cdot x^k, \tag{3.75}$$

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 (3.71) and (3.72). 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

$$E(X_t^n) = a_n^{-1}\{b_n \cdot E(X_t^{n-1}) + c_n \cdot E(X_t^{n-2})\} \tag{3.76}$$

where $a_n = n\{1-(n-1)a\}\beta$, $b_n = n\{\alpha+(n-1)b\}\beta$, and $c_n = n(n-1)c\beta$ for $n = 0, 1, 2, \ldots$. The initial conditions are given by $E(X_t^0) = 1$, and $E(X_t) = \alpha$. This can be found from the expressions for the eigenfunctions, but is more easily seen as follows. By Ito's formula

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

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

Example 3.19 Equation (3.76) 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 (3.73),

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

□

For a diffusion $T(X)$ obtained from a solution X to (3.70) 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 (3.66) 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 3.20 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

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

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

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

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 3.21 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 non-linear and the diffusion coefficient is constant. This model was considered in Example 3.14.

□

4 The likelihood function

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 (3.1) 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).

5 Non-martingale estimating functions

5.1 Asymptotics

When the estimating function $G_n(\theta)$ is not a martingale under P_θ, further conditions on the diffusion process must be imposed to ensure the asymptotic normality in (2.3). A sufficient condition that (2.3) holds under P_{θ_0} with $V(\theta)$ given by (5.1) is that the diffusion is stationary and geometrically α-mixing, that

$$\begin{aligned} V(\theta) &= Q_{\theta_0}\left(g(\theta)g(\theta)^T\right) \\ &\quad + \sum_{k=1}^{\infty}\Big[E_{\theta_0}\left(g(X_\Delta,\ldots,X_{r\Delta})g(X_{(k+1)\Delta},\ldots,X_{(k+r)\Delta})^T\right) \\ &\qquad + E_{\theta_0}\left(g(X_{(k+1)\Delta},X_{(k+r)\Delta})g(X_\Delta,X_{r\Delta})^T\right)\Big], \end{aligned} \tag{5.1}$$

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). To define the concept of α-mixing, let $\mathcal{F}_t$ denote the σ-field generated by $\{X_s \,|\, s \le t\}$ and let $\mathcal{F}^t$ denote the σ-field generated by $\{X_s \,|\, s \ge t\}$. A stochastic process X is said to be *α-mixing*, if

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

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 there exit positive constants c_1 and c_2 such that

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

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 (3.37) are larger than $\lambda > 0$, 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 $-\beta(x-\alpha)$, $\beta > 0$, as 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 (3.1) to be geometrically α-mixing.

Condition 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 there exists a constant $K_\theta > 0$ such that $|b(x;\theta)| \le K_\theta(1+|x|)$ and $\sigma^2(x;\theta) \le K_\theta(1+x^2)$ 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 2.1 the existence of a $\bar{\theta}$-consistent and asymptotically normal G_n-estimator follows from Theorem 2.2, which also contains a result about eventual uniqueness of the estimator.

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 for the generator are available; see the Subsections 3.6 and 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 will therefore usually give less efficient estimators.

First we consider estimating function of the form

$$G_n(\theta) = \sum_{i=1}^{n} h(X_{\Delta i};\theta), \tag{5.2}$$

where h is a p-dimensional function. We assume that the diffusion is geometrically α-mixing, so that a central limit theorem holds, and that Condition 2.1 holds for $r = 1$ and $\bar{\theta} = \theta_0$. The latter condition simplifies considerably, because 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 (3.10). In particular, (2.4) and (2.5) simplifies to

$$\mu_{\theta_0}(h(\theta_0)) = \int_\ell^r h(x;\theta_0)\mu_{\theta_0}(x)dx = 0 \tag{5.3}$$

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 (2.7) is here that θ_0 is the only root of $\mu_{\theta_0}(h(\theta))$.

Kessler (2000) proposed

$$h(x;\theta) = \partial_\theta \log \mu_\theta(x), \tag{5.4}$$

which is the score functions (the derivative of the log-likelihood functions) 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 (5.3) is satisfied for this specification of h follows under standard conditions that allow the interchange of differentiation and integration.

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

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

$$h_j(x;\theta) = A_\theta f_j(x;\theta), \tag{5.5}$$

where A_θ is the generator (3.37), and f_j, $j = 1, \ldots, d$ are twice differentiable functions chosen such that Condition 2.1 holds. The estimating function with h given by (5.5) 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 (5.3).

Lemma 5.2 *Suppose* $f \in C^2((\ell, r))$, $A_\theta f \in L^1(\mu_\theta)$ *and*

$$\lim_{x \to r} f'(x)\sigma^2(x;\theta)\mu_\theta(x) = \lim_{x \to \ell} f'(x)\sigma^2(x;\theta)\mu_\theta(x). \tag{5.6}$$

Then

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

Proof: Note that by (3.10), 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 5.3 Consider the square-root process (3.27) 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

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

The condition (5.6) is obviously satisfied because the invariant distribution is a normal distribution.

□

Sørensen (2001) derived the estimating function of the form (5.2) with

$$h(x;\theta) = A_\theta \partial_\theta \log \mu_\theta(x) \tag{5.7}$$

as an approximation to the score function for continuous-time observation of the diffusion process. This is clearly a particular case of (5.5) with $f(x;\theta) = \partial_\theta \log \mu_\theta(x)$, which is the i.i.d. score function used in (5.4).

As mentioned above, an estimating function of the form (5.2) cannot be expected to yield as efficient estimators as an estimating function that depends on pairs of consecutive observations, and thus can use the information contained in the transitions. Hansen & Scheinkman (1995) proposed non-martingale estimating functions of the form (3.2) with g given by

$$g_j(\Delta, x, y; \theta) = h_j(y) A_\theta f_j(x) - f_j(x) \hat{A}_\theta h_j(y), \tag{5.8}$$

where the functions f_j and h_j satisfy weak regularity conditions ensuring that (2.4) 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, $\hat{A}_\theta = A_\theta$. Obviously, the estimating function of the form (5.2) with $h_j(x;\theta) = A_\theta f_j(x)$ is a particular case of (5.8) with $h_j(y) = 1$.

5.3 Approximate martingale estimating functions

For martingale estimating functions of the form (3.20) and (3.30), we can always, as discussed in Subsection 3.3, obtain an explicit approximation to the optimal weight matrix by means of the expansion (3.36). For diffusion models where there is no explicit expression for the transition operator, it is tempting to go on and approximate $\pi^\theta_\Delta(f_j(\theta))(x)$ using (3.36), and thus, quite generally, obtain explicit *approximate martingale estimating function*. Estimators of this type 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), Kessler (1997), and Kelly, Platen & Sørensen (2004).

It is, however, important to note that there is a dangerous pitfall when using these simple approximate martingale estimating functions. They do not satisfy 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 (3.2) with

$$g(x, y; \theta) = a(x, \theta)[f(y) - f(x) - \Delta A_\theta f(x)], \tag{5.9}$$

where A_θ is the generator (3.37), i.e., we have used a first order expansion of $\pi^\theta_\Delta f(x)$. 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 (2.3) hold, and that Condition 2.1 is satisfied. Then by Theorem 2.2, the estimator obtained using (5.9) converges to the solution, $\bar{\theta}$, of

$$Q_{\theta_0}(g(\bar{\theta})) = 0. \tag{5.10}$$

We assume that the solution is unique. Using the expansion (3.36), we find that

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

If we neglect all O-terms, we obtain that

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

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 5.4 Consider again a diffusion with linear drift, $b(x;\theta) = -\beta(x-\alpha)$. In this case (5.9) 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 3.7 that $F(x;\alpha,\beta) = -\beta(x-\alpha)\Delta$. 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, irrespective of the value of β_0. 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 can get an impression of what can happen when estimating the parameter β by means of the dangerous estimating function given by (5.9) from the simulation study in Bibby & Sørensen (1995) for the square root process (3.27). The result is given in Table 5.1. For the function a the approximately optimal weight function was used, cf. Example 3.7. 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 parameter values were $\alpha = 10$, $\beta = 1$ and $\tau = 1$, 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 5.1: Empirical mean of 500 estimates of the parameter β in the CIR model. The true parameter values are $\alpha = 10$, $\beta = 1$, and $\tau = 1$.

The asymptotic bias given by (5.10) is small when Δ is sufficiently small, and the results in the following section on high frequency asymptotics show that in this case 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.

6 High-frequency asymptotics

A large number of estimating functions have been proposed for diffusion models, and a large number of simulation studies have been performed to compare their relative merits, but the general picture has been rather confusing. By considering the high frequency scenario,

$$n \to \infty, \qquad \Delta_n \to 0, \qquad n\Delta_n \to \infty, \tag{6.1}$$

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. 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 this theory is that under weak conditions optimal martingale estimating functions give rate optimal and efficient estimators.

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

$$dX_t = b(X_t; \alpha)dt + \sigma(X_t; \beta)dW_t, \tag{6.2}$$

where $\theta = (\alpha, \beta) \in \Theta \subseteq \mathbb{R}^2$. The results below can be generalized to multivariate diffusions and parameters of higher dimension. We consider estimating functions of the general form (2.1), where the two-dimensional function $g = (g_1, g_2)$ for some $\kappa \geq 2$ and for all $\theta \in \Theta$ satisfies

$$E_\theta(g(\Delta_n, X_{\Delta_n i}, X_{\Delta_n(i-1)}; \theta) \,|\, X_{\Delta_n(i-1)}) = \Delta_n^\kappa R(\Delta_n, X_{\Delta_n(i-1)}; \theta). \tag{6.3}$$

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 a compact set[1]. We assume that the

[1]For any compact subset $K \subseteq \Theta$, there exist constants $C_1, C_2, C_3 > 0$ such that $\sup_{\theta \in K} |F(y, x; \theta)| \leq C_1(1 + |x|_2^C + |y|_3^C)$ for all x and y in the state space of the diffusion.

diffusion and the estimating functions satisfy the technical regularity Condition 6.2 given below.

Martingale estimating functions obviously satisfy (6.3) with $R = 0$, but for instance the approximate martingale estimating functions discussed at the end of the previous section satisfy (6.3) too.

Theorem 6.1 *Suppose that*

$$\begin{aligned} \partial_y g_2(0, x, x; \theta) &= 0, && (6.4)\\ \partial_y g_1(0, x, x; \theta) &= \partial_\alpha b(x; \alpha)/\sigma^2(x; \beta), && (6.5)\\ \partial_y^2 g_2(0, x, x; \theta) &= \partial_\beta \sigma^2(x; \beta)/\sigma^2(x; \beta)^2, && (6.6) \end{aligned}$$

for all $x \in (\ell, r)$ and $\theta \in \Theta$. Assume, moreover, that the following identifiability condition is satisfied

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

and that

$$\begin{aligned} W_1 &= \int_\ell^r \frac{(\partial_\alpha b(x; \alpha_0))^2}{\sigma^2(x; \beta_0)}\mu_{\theta_0}(x)dx \neq 0,\\ W_2 &= \int_\ell^r \left[\frac{\partial_\beta \sigma^2(x; \beta_0)}{\sigma^2(x; \beta_0)}\right]^2 \mu_{\theta_0}(x)dx \neq 0. \end{aligned}$$

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 $n \to \infty$. For a martingale estimating function, or more generally if $n\Delta^{2(\kappa-1)} \to 0$,

$$\begin{pmatrix} \sqrt{n\Delta_n}(\hat{\alpha}_n - \alpha_0) \\ \sqrt{n}(\hat{\beta}_n - \beta_0) \end{pmatrix} \xrightarrow{\mathcal{D}} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} W_1^{-1} & 0 \\ 0 & W_2^{-1} \end{pmatrix}\right). \quad (6.7)$$

An estimator satisfying (6.7) is rate optimal and efficient, cf. Gobet (2002), who showed that the model considered here is locally asymptotically normal. Note that the estimator of the diffusion coefficient parameter, β, converges faster than the estimator of the drift parameter, α. Condition (6.4) implies rate optimality. If this condition is not satisfied, the estimator of the diffusion coefficient parameter converges at the slower rate $\sqrt{n\Delta_n}$. This condition is called *the Jacobsen condition*, because it appears in the theory of *small Δ-optimal estimation* developed in Jacobsen (2001) and Jacobsen (2002). In this theory the asymptotic covariance matrix in (3.15) is expanded in powers of Δ, the time between observations. The leading term is minimal when (6.5) and (6.6) are satisfied. The same expansion of (3.15) was used by Aït-Sahalia & Mykland (2004).

The assumption $n\Delta_n \to \infty$ in (6.1) is needed to ensure that the drift parameter, α, can be consistently estimated. If the drift is known and only the diffusion coefficient parameter, β, needs to be estimated, this condition can be omitted, see Genon-Catalot & Jacod (1993). Another situation where the infinite observation horizon, $n\Delta_n \to \infty$, is not needed for

consistent estimation of α is when the high frequency asymptotic scenario is combined with the small diffusion scenario, where $\sigma(x;\beta) = \epsilon_n \zeta(x;\beta)$ and $\epsilon_n \to 0$, see Genon-Catalot (1990), Sørensen & Uchida (2003) and Gloter & Sørensen (2008).

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 (6.4) – (6.6). The point is that a version must exist which satisfies 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 (3.31), 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 $D(x) = (\partial_x f(x;\theta) \ \partial_x^2 f(x;\theta))$ has full rank $d(d+3)/2$.

We conclude this section by stating 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 very considerably in a way similar to the proofs in Gloter & Sørensen (2008).

Condition 6.2 *The diffusion is ergodic 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) $g(\Delta, y, x; \theta) \in C_{p,2,6,2}(\mathbb{R}_+ \times (\ell, r)^2 \times \Theta)$ *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

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

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 $\partial_t^{i_1}\partial_y^{i_2}\partial_\alpha^{i_3}\partial_\beta^{i_4} f$, $i_j = 1, \dots k_j$, $j = 1, 2$, $i_3 + i_4 \leq k_3$, 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.

7 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 (3.31) is not a martingale. To obtain a martingale, the conditional expectation given $X_{(i-1)\Delta}$ in (3.31) 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

$$dX_t = b(X_t;\theta)dt + \sigma(X_t;\theta)dW_t, \tag{7.1}$$

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 $\mathbb{R}^p$. However, we do not observed X directly. What we observe is

$$Y_i = k(X_{t_i}) + Z_i, \tag{7.2}$$

where k maps $\mathbb{R}^D$ into $\mathbb{R}^d$ ($d < D$), 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.

7.1 Prediction-based estimating functions

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

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

where $\Pi_j^{(i-1)}(\theta)$ is a p-dimensional vector, the coordinates of which belong to $\mathcal{P}_{i-1,j}$, and $\breve{\pi}_j^{(i-1)}(\theta)$ is the minimum mean square error predictor in $\mathcal{P}_{i-1,j}$ of $f_j(Y_i, \dots, Y_{i-s};\theta)$ under P_θ. When $s = 0$ and $\mathcal{P}_{i-1,j}$ is the set of all functions of $Y_1, \dots, Y_{i-1}$ with finite variance,

$\breve{\pi}_j^{(i-1)}(\theta)$ 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 (3.31) is a particular case of a prediction-based estimating function.

The minimum mean square error predictor of $f_j(Y_i,\ldots,Y_{i-s};\theta)$ is the projection of $f_j(Y_i,\ldots,Y_{i-s};\theta)$ onto the subspace $\mathcal{P}_{i-1,j}$ of the L_2-space of all functions of $Y_1,\ldots,Y_i$ with finite variance under P_θ. Therefore $\breve{\pi}_j^{(i-1)}(\theta)$ satisfies the normal equation

$$\mathrm{E}_\theta\left(\pi_j^{(i-1)}\left\{f_j(Y_i,\ldots,Y_{i-s};\theta)-\breve{\pi}_j^{(i-1)}(\theta)\right\}\right)=0 \tag{7.3}$$

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

$$\mathrm{E}_\theta\left(G_n(\theta)\right)=0. \tag{7.4}$$

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, $\mathcal{P}_{i-1,j}$, 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 $a^TZ_j^{(i-1)}$, 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 $i\geq r+1$:

$$G_n(\theta)=\sum_{i=r+1}^{n}\sum_{j=1}^{N}\Pi_j^{(i-1)}(\theta)\left[f_j(Y_i,\ldots,Y_{i-s};\theta)-\breve{\pi}_j^{(i-1)}(\theta)\right] \tag{7.5}$$

It is well-known that the minimum mean square error predictor, $\breve{\pi}_j^{(i-1)}(\theta)$, is found by solving the normal equations (7.3). We 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

$$b_j(\theta)_i=\mathrm{Cov}_\theta(Z_{ji}^{(r)},f_j(Y_{r+1},\ldots,Y_{r+1-s};\theta)), \tag{7.6}$$

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

$$\breve{\pi}_j^{(i-1)}(\theta)=\breve{a}_j(\theta)^TZ_j^{(i-1)}$$

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

$$\breve{a}_{j*}(\theta)=C_j(\theta)^{-1}b_j(\theta) \tag{7.7}$$

and

$$\breve{a}_{j0}(\theta) = E_\theta(f_j(Y_{s+1}, \ldots, Y_1; \theta)) - \sum_{k=1}^{q_j} \breve{a}_{jk}(\theta) E_\theta(Z_{jk}^{(r)}). \tag{7.8}$$

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 7.1 An important particular case when $d = 1$ is $f_j(y) = y^j$, $j = 1, \ldots, m$. For each $i = r+1, \ldots, n$ and $j = 1, \ldots, m$, we let $\{Z_{jk}^{(i-1)} \mid k = 0, \ldots, q_j\}$ be a subset of $\{Y_{i-\ell}^\kappa \,|\, \ell = 1, \ldots, r, \kappa = 0, \ldots, j\}$, where $Z_{j0}^{(i-1)}$ is always equal to 1. Here we need to assume that $\mathrm{E}_\theta(Y_i^{2m}) < \infty$ for all $\theta \in \Theta$. To find $\breve{\pi}_j^{(i-1)}(\theta)$, $j = 1, \ldots, m$, by means of (7.7) and (7.8), we must calculate moments of the form

$$\mathrm{E}_\theta(Y_1^\kappa Y_k^j), \quad 0 \le \kappa \le j \le m, \quad k = 1, \ldots, r. \tag{7.9}$$

To avoid the matrix inversion in (7.7), the vector of coefficients $\breve{a}_j$ can be found by means of the m-dimensional Durbin-Levinson algorithm applied to the process $\{(Y_i, Y_i^2, \ldots, Y_i^m)\}_{i \in \mathbb{N}}$, see Brockwell & Davis (1991). Suppose the diffusion process X is exponentially ρ-mixing, see Doukhan (1994). This is for instance the case for a Pearson diffusion or for a one-dimensional diffusion that satisfies Condition 5.1. Then the observed process Y inherits this property, which implies that constants $K > 0$ and $\lambda > 0$ exist such that $|\mathrm{Cov}_\theta(Y_1^j, Y_k^j)| \le K e^{-\lambda k}$. Therefore r will usually not need to be chosen particularly large.

In many situations it is reasonable take $m = 2$ 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 will illustrate the use of the Durbin-Levinson algorithm in the simplest possible case, where $m = 1$, $f(x) = x$, $Z_0^{(i-1)} = 1$, $Z_k^{(i-1)} = Y_{i-k}$, $k = 1, \ldots, r$. 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 goes 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 (7.8), which 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 $\mathrm{E}_\theta(Y_i - \breve{\pi}^{(i-1)})$. 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 (7.5) in the sense explained in Section 9. 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, (7.5) can be written in the form

$$G_n(\theta) = A(\theta) \sum_{i=r+1}^{n} H^{(i)}(\theta), \tag{7.10}$$

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

$$H^{(i)}(\theta) = Z^{(i-1)}\left(F(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}^{(i-1)}(\theta)\right), \tag{7.11}$$

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

$$Z^{(i-1)} = \begin{pmatrix} Z_1^{(i-1)} & 0_{q_1} & \cdots & 0_{q_1} \\ 0_{q_2} & Z_2^{(i-1)} & \cdots & 0_{q_2} \\ \vdots & \vdots & & \vdots \\ 0_{q_N} & 0_{q_N} & \cdots & Z_N^{(i-1)} \end{pmatrix}. \tag{7.12}$$

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:

$$\breve{a}(\theta) = (\breve{a}_{10}(\theta), \ldots, \breve{a}_{1q_1}(\theta), \ldots, \breve{a}_{N0}(\theta), \ldots \breve{a}_{Nq_N}(\theta))^T, \tag{7.13}$$

where the $\breve{a}_{jk}$s define the minimum mean square error predictor. Specifically, $\breve{\pi}^{(i-1)}(\theta) = (Z^{(i-1)})^T \breve{a}(\theta)$.

Condition 7.2
(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*

$$U(\theta)^T = E_\theta \left(Z^{(i-1)} \partial_{\theta^T} F(Y_i, \ldots, Y_{i-s}; \theta) \right) \tag{7.14}$$

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

$$\begin{aligned} \bar{M}_n(\theta) &= \mathrm{E}_\theta \left(H^{(r+1)}(\theta) H^{(r+1)}(\theta)^T \right) \\ &+ \sum_{k=1}^{n-r-1} \frac{(n-r-k)}{(n-r)} \left\{ \mathrm{E}_\theta \left(H^{(r+1)}(\theta) H^{(r+1+k)}(\theta)^T \right) \right. \\ &\qquad \left. + \mathrm{E}_\theta \left(H^{(r+1+k)}(\theta) H^{(r+1)}(\theta)^T \right) \right\}, \end{aligned} \tag{7.15}$$

which is the covariance matrix of $\sum_{i=r+1}^n H^{(i)}(\theta)/\sqrt{n-r}$. The sensitivity function (9.1) is given by

$$S_{G_n}(\theta) = (n-r) A(\theta) (U(\theta)^T - D(\theta) \partial_{\theta^T} \breve{a}(\theta))$$

where the $\bar{p} \times \bar{p}$-matrix $D(\theta)$ is given by

$$D(\theta) = \mathrm{E}_\theta \left(Z^{(i-1)} (Z^{(i-1)})^T \right) \tag{7.16}$$

It follows from Theorem 9.1 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 7.2 (4) the matrix $\bar{V}_n(\theta)$ is invertible, see Sørensen (2000), so it follows that

$$A_n^*(\theta) = (U(\theta) - \partial_\theta \breve{a}(\theta)^T D(\theta)) \bar{M}_n(\theta)^{-1}, \tag{7.17}$$

so that the estimating function

$$G_n^*(\theta) = A_n^*(\theta) \sum_{i=s+1}^n Z^{(i-1)} \left(F(Y_i, \ldots, Y_{i-s}; \theta) - \breve{\pi}^{(i-1)}(\theta) \right), \tag{7.18}$$

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 7.3 Consider again the type of prediction-based estimating function discussed in Example 7.1. In order to calculate (7.15), we need mixed moments of the form

$$E_\psi[Y_1^{k_1} Y_{t_1}^{k_2} Y_{t_2}^{k_3} Y_{t_3}^{k_4}], \tag{7.19}$$

for $1 \leq t_1 \leq t_2 \leq t_3$ and $k_1 + k_2 + k_3 + k_4 \leq 4N$, where k_i, $i = 1, \ldots, 4$ are non-negative integers.

□

7.2 Asymptotics

A prediction-based estimating function of the form (7.10) gives consistent and asymptotically normal estimators under the following condition, where θ_0 as usual is the true parameter value.

Condition 7.4
(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 (7.13) 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* (7.14) *and* (7.16).
(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 7.4 (1) and (2) ensures that the central limit theorem (2.3) holds and that $\bar{M}_n(\theta_0) \to M(\theta_0)$, where

$$\begin{aligned} M(\theta) &= \mathrm{E}_\theta\left(H^{(r+1)}(\theta) H^{(r+1)}(\theta)^T\right) \\ &\quad + \sum_{k=1}^{\infty}\Big\{\mathrm{E}_\theta\left(H^{(r+1)}(\theta) H^{(r+1+k)}(\theta)^T\right) \\ &\qquad\qquad + \mathrm{E}_\theta\left(H^{(r+1+k)}(\theta) H^{(r+1)}(\theta)^T\right)\Big\}. \end{aligned}$$

The concept of geometric α-mixing was explained in Subsection 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. As explained in Subsection 5.1, we only need to check Condition 2.1 with $\bar{\theta} = \theta_0$ to obtain asymptotic results for prediction-based estimators. The condition (2.4) is satisfied because of (7.4). It is easy to see that Condition 7.4 (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, the condition (2.7) is identical to Condition 7.4 (5). Therefore it follows from Theorem 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

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} N_p\left(0, W^{-1} A(\theta_0) M(\theta_0) A(\theta_0)^T W^{T^{-1}}\right).$$

7.3 Integrated diffusions

Sometimes a diffusion process cannot be observed directly, but that data of the form

$$Y_i = \frac{1}{\Delta}\int_{(i-1)\Delta}^{i\Delta} X_s\,ds, \qquad i = 1,\ldots,n \tag{7.20}$$

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, measured as an average in pieces of ice, each piece representing a time interval with time increasing as a function of the depth, is a proxy for paleo-temperatures. The variation of the paleo-temperature can be modelled by a stochastic differential equation, and it is natural to model the ice-core data 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), and Gloter (2006).

The model for data of the type (7.20) is a particular case of (7.1) with

$$b(x;\theta) = \begin{pmatrix} b_1(x_1;\theta) \\ x_1 \end{pmatrix}, \qquad \sigma(x;\theta) = \begin{pmatrix} \sigma_1(x_1;\theta) & 0 \\ 0 & 0 \end{pmatrix}$$

with $X_{2,0} = 0$, where only the second coordinate is observed. A stochastic differential equation of this form is called hypoelliptic. Clearly 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. In the following we will again denote the basic diffusion by X (rather than X_1).

Suppose that $4N$'th moment of X_t is finite. The moments (7.9) and (7.19) 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 $1 \le t_1 \le t_2 \le t_3$, $A = [0\,,\,\Delta]^{k_1} \times [(t_1-1)\Delta\,,\,t_1\Delta]^{k_2} \times [(t_2-1)\Delta\,,\,t_2\Delta]^{k_3} \times [(t_3-1)\Delta\,,\,t_3\Delta]^{k_4}$, 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 here 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 3.7, these mixed moments can be calculated by a simple iterative formula obtained from (3.75) and (3.76). 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. (3.75). Therefore the integral above can be explicitly calculated, so that explicit optimal estimating functions of the type considered in Example 7.1 are available for observations of integrated Pearson diffusions.

Example 7.5 Consider observation of an integrated square root process (3.27) and a prediction-based estimating function with $f_1(x) = x$ and $f_2(x) = x^2$ with predictors given by $\pi_1^{(i-1)} = \alpha_{1,0} + \alpha_{1,1}Y_{i-1}$ and $\pi_2^{(i-1)} = \alpha_{2,0}$. Then the minimal 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

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ Y_{i-1} \\ 0 \end{pmatrix} [Y_i - \bar{\pi}_1^{(i-1)}(Y_{i-1};\theta)] + \sum_{i=1}^{n} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} [Y_i^2 - \bar{\pi}_2^{(i-1)}(\theta)],$$

from which we obtain the estimators

$$\begin{aligned} \hat{\alpha} &= \frac{1}{n}\sum_{i=1}^{n} Y_i + \frac{a(\hat{\beta})Y_n - Y_1}{(n-1)(1-a(\hat{\beta}))} \\ \sum_{i=2}^{n} Y_{i-1}Y_i &= \hat{\alpha}(1-a(\hat{\beta}))\sum_{i=2}^{n} Y_{i-1} + a(\hat{\beta})\sum_{i=2}^{n} Y_{i-1}^2 \\ \hat{\sigma}^2 &= \frac{\hat{\beta}^3\Delta^2\sum_{i=2}^{n}(Y_i^2 - \hat{\alpha}^2)}{(n-1)\hat{\alpha}(e^{-\hat{\beta}\Delta} - 1 + \hat{\beta}\Delta)}. \end{aligned}$$

The estimators are explicit apart from $\hat{\beta}$, which can be found by solving a non-linear equation in one variable. Details can be found in Ditlevsen & Sørensen (2004).

□

7.4 Sums of diffusions

An autocorrelation function of the form

$$\rho(t) = \phi_1 \exp(-\beta_1 t) + \ldots + \phi_D \exp(-\beta_D t), \tag{7.21}$$

$\beta_i > 0$, is found in many time series data. Examples are financial time series, Barndorff-Nielsen & Shephard (2001), and turbulence, Barndorff-Nielsen, Jensen & Sørensen (1990) and Bibby, Skovgaard & Sørensen (2005).

A simple model with autocorrelation function of the form (7.21) is the sum of diffusions

$$Y_t = X_{1,t} + \ldots + X_{D,t}$$

where

$$dX_{i,t} = -\beta_i(X_{i,t} - \alpha_i) + \sigma_i(X_{i,t})dW_{i,t}, \quad i = 1, \ldots, D,$$

are independent. In this case

$$\phi_i = \frac{\mathrm{Var}(X_{i,t})}{\mathrm{Var}(X_{1,t}) + \cdots + \mathrm{Var}(X_{D,t})}.$$

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). 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 7.6 *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 3.7, we have explicit formulae that allow calculation of (7.9) and (7.19), 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 7.1. 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 (7.19) 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 (3.75) and (3.76).

Example 7.7 *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-diffusion. The distribution of Y_t is a convolution of skew t-diffusions,

$$\mathrm{Var}(Y) = (1 + \rho^2)\left(\frac{\nu_1}{\nu_1 - 2} + \frac{\nu_2}{\nu_2 - 2}\right),$$

and $\phi_i = \nu_i(\nu_i - 2)^{-1}/\{\nu_1(\nu_1 - 2)^{-1} + \nu_2(\nu_2 - 2)^{-1}\}$. 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 (7.21) 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 $Z_{1,1}^{(i-1)} = 1$ and $Z_{1,2}^{(i-1)} = Y_{i-1}$. The estimating equations take the form

$$\sum_{i=2}^{n} \begin{bmatrix} Y_i^2 - \sigma^2 - \zeta_{21} Y_{i-1} \\ Y_{i-1} Y_i^2 - \sigma^2 Y_{i-1} - \zeta_{21} Y_{i-1}^2 \end{bmatrix} = 0, \tag{7.22}$$

with

$$\begin{aligned} \sigma^2 &= \mathrm{Var}(Y_i) = (1 + \rho^2)\left\{\frac{\nu_1}{\nu_1 - 2} + \frac{\nu_2}{\nu_2 - 2}\right\}, \\ \zeta_{21} &= \frac{\mathrm{Cov}(Y_{i-1}, Y_i^2)}{\mathrm{Var}(Y_i)} = 4\rho\left\{\frac{\sqrt{\nu_1}}{\nu_1 - 3}\phi_1 e^{-\beta_1 \Delta} + \frac{\sqrt{\nu_2}}{\nu_2 - 3}\phi_2 e^{-\beta_2 \Delta}\right\}. \end{aligned}$$

Solving equation (7.22) for ζ_{21} and σ^2 we get

$$\begin{aligned}\hat{\zeta}_{21} &= \frac{\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}Y_i^2 - (\frac{1}{n-1}\sum_{i=2}^n Y_{i-1})(\frac{1}{n-1}\sum_{i=2}^n Y_i^2)}{\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}^2 - (\frac{1}{n-1}\sum_{i=2}^n Y_{i-1})^2}, \\ \hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=2}^n Y_i^2 + \hat{\zeta}_{21}\frac{1}{n-1}\sum_{i=2}^n Y_{i-1}.\end{aligned}$$

In order to estimate ρ we restate ζ_{21} as

$$\begin{aligned}\zeta_{21} &= \sqrt{32(1+\rho^2)}\cdot\rho\cdot\left\{\frac{\sqrt{9(1+\rho^2)-\phi_1\sigma^2}}{3(1+\rho^2)-\phi_1\sigma^2}\phi_1 e^{-\beta_1\Delta}\right. \\ &\qquad \left. + \frac{\sqrt{9(1+\rho^2)-\phi_2\sigma^2}}{3(1+\rho^2)-\phi_2\sigma^2}\phi_2 e^{-\beta_2\Delta}\right\}\end{aligned}$$

and insert $\hat{\sigma}^2$ for σ^2. Thus, we get a one-dimensional estimating equation, $\zeta_{21}(\beta,\phi,\hat{\sigma}^2,\rho) = \hat{\zeta}_{21}$, 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 $\hat{\nu}_i = \frac{2\phi_i\hat{\sigma}^2}{\phi_i\hat{\sigma}^2-(1+\hat{\rho}^2)}$, $i = 1, 2$.

□

7.5 Compartment models

Diffusion compartment models are multivariate diffusion models with linear drift,

$$dX_t = (B(\theta)X_t - b(\theta))dt + \sigma(X_t;\theta)dW_t, \tag{7.23}$$

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 certain 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 modified by the random perturbation given by the diffusion term. The vector $b(\theta)$ represents input to of output from the system.

Example 7.8 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 $Y_i = X_{1,t_i}$. The model is Gaussian, so likelihood inference is feasible and was studied by Bibby (1995). All mixed moments (7.9) and (7.19) can be calculated explicitly, so also an explicit optimal prediction-based estimating function of the type considered in Example 7.1 is available to estimate the parameters and was studied by Düring (2002).

□

Example 7.9 A non-Gaussian diffusion compartment model is obtained by the specification $\sigma(x,\theta) = \mathrm{diag}(\tau_1\sqrt{x_1},\ldots\sqrt{x_D})$. 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 7.1 when the first compartment is observed, i.e. when $Y_i = X_{1,t_i}$. In this case, the mixed moments (7.9) and (7.19) must be calculated numerically.

□

8 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 (2008).

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, X_i might be the observation of a diffusion process at time points $i\Delta_n$. Another example is that the diffusion coefficient might depend on n.

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 $P = P_\theta$, 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 8.1 *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 that 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 8.2 *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) $G_n(\bar{\theta}) \xrightarrow{P} 0$ *(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*

$$\sup_{\theta \in M} \| \partial_{\theta^T} G_n(\theta) - W(\theta) \| \xrightarrow{P} 0. \tag{8.1}$$

(iii) *The matrix $W(\bar{\theta})$ is non-singular with P–probability one.*

Then a sequence $(\hat{\theta}_n)$ of G_n-estimators 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 (8.1) 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 continuous (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, (8.1) implies that

$$\sup_{\theta \in M} |G_n(\theta) - G(\theta)| \xrightarrow{P} 0. \tag{8.2}$$

This observation casts light on the result of Theorem 8.2. Since $G_n(\theta)$ can be made arbitrarily close to $G(\theta)$ by choosing n large enough, and since $G(\theta)$ has a zero at $\bar{\theta}$, it is intuitively clear that $G_n(\theta)$ must have a zero 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 8.3 *Assume (8.2) for some subset M of θ containing $\bar{\theta}$, and that*

$$P\left(\inf_{M \setminus \bar{B}_\epsilon(\bar{\theta})} |G(\theta)| > 0 \right) = 1 \tag{8.3}$$

for all $\epsilon > 0$. Then for any sequence $(\hat{\theta}_n)$ of G_n–estimators

$$P(\hat{\theta}_n \in M \setminus \bar{B}_\epsilon(\bar{\theta})) \to 0 \tag{8.4}$$

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 8.3 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 8.4 *Assume the estimating function G_n satisfies the conditions of Theorem 8.2 and that there is a sequence of real numbers $a_n > 0$ increasing to ∞ such that*

$$\begin{pmatrix} a_n G_n(\bar{\theta}) \\ \partial_{\theta^T} G_n(\bar{\theta}) \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} Z \\ W(\bar{\theta}) \end{pmatrix}, \tag{8.5}$$

where Z is a non-degenerate random variable. When $W(\bar{\theta})$ is non-random it is enough to assume that

$$a_n G_n(\bar{\theta}) \xrightarrow{\mathcal{L}} Z. \tag{8.6}$$

Then for any $\bar{\theta}$–consistent sequence $(\hat{\theta}_n)$ of G_n–estimators,

$$a_n(\hat{\theta}_n - \bar{\theta}) \xrightarrow{\mathcal{L}} -W(\bar{\theta})^{-1} Z. \tag{8.7}$$

When Z is normal distributed with expectation zero and covariance matrix V, and when Z is independent of $W(\bar{\theta})$, then the limit distribution is the normal variance-mixture with characteristic function

$$s \mapsto E\left(\exp\left(-\tfrac{1}{2} s^* W(\bar{\theta})^{-1} V W(\bar{\theta})^{*-1} s\right)\right). \tag{8.8}$$

If moreover $W(\bar{\theta})$ is non-random, then the limit distribution is a normal distribution with expectation zero and covariance matrix $W(\bar{\theta})^{-1} V W(\bar{\theta})^{*-1}$.

9 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). The theory is very closely related to the theory of the generalized method of moments developed independently in parallel in the econometrics literature, see e.g. Hansen (1982), Hansen (1985) and Hansen, Heaton & Ogaki (1988). A modern review of the theory of optimal estimating functions can be found in Heyde (1997).

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 8.2 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

$$S_{G_n}(\theta) = E_\theta(\partial_{\theta^T} G_n(\theta)) \tag{9.1}$$

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

$$S_{G_n}(\theta) = -\mathrm{Cov}_\theta(G_n(\theta), U_n(\theta)).$$

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

$$K_{G_n}(\theta) = S_{G_n}(\theta)^2 / \mathrm{Var}_\theta(G_n(\theta)) = S_{G_n}(\theta)^2 / E_\theta(G_n(\theta)^2), \tag{9.2}$$

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 (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

$$K_{G_n^*}(\theta) \geq K_{G_n}(\theta) \tag{9.3}$$

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

$$S_{G_n}(\theta) = E_\theta(\partial_{\theta^T} G_n(\theta)). \tag{9.4}$$

For a multivariate parameter, the Godambe information is the $p \times p$-matrix

$$K_{G_n}(\theta) = S_{G_n}(\theta)^T E_\theta \left(G_n(\theta) G_n(\theta)^T \right)^{-1} S_{G_n}(\theta), \tag{9.5}$$

and an optimal estimating function G_n^* can be defined by (9.3) 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 (9.3), 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*

$$E_\theta \left(G_n^{(s)}(\theta) G_n^{(s)}(\theta)^T \right) = -E_\theta(\partial_{\theta^T} G_n^{(s)}(\theta)), \tag{9.6}$$

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 quasi-score 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 9.1 *If $G_n^* \in \mathcal{G}_n$ satisfies the equation*

$$S_{G_n}(\theta)^{-1}E_\theta\left(G_n(\theta)G_n^*(\theta)^T\right) = S_{G_n^*}(\theta)^{-1}E_\theta\left(G_n^*(\theta)G_n^*(\theta)^T\right) \tag{9.7}$$

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 (9.7) .

The condition (9.7) 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*, (9.6), so that

$$K_{G_n^*}(\theta) = E_\theta\left(G_n^*(\theta)G_n^*(\theta)^T\right).$$

9.1 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

$$\langle G(\theta)\rangle_n^{-\frac{1}{2}} \bar{G}_n(\theta)(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, I_p). \tag{9.8}$$

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

$$I_{G_n}(\theta) = \bar{G}_n(\theta)^T \langle G(\theta)\rangle_n^{-1} \bar{G}_n(\theta) \tag{9.9}$$

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

$$I_{G_n^*}(\theta) \geq I_{G_n}(\theta) \tag{9.10}$$

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 9.1. 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:

$$\langle G, \tilde{G}\rangle_n = \sum_{i=1}^{n} E \left((G_i - G_{i-1})(\tilde{G}_i - \tilde{G}_{i-1})^T | \mathcal{F}_{i-1} \right). \tag{9.11}$$

Theorem 9.2 *If $G_n^* \in \mathcal{G}_n$ satisfies*

$$\bar{G}_n(\theta)^{-1} \langle G(\theta), G^*(\theta)\rangle_n = \bar{G}_n^*(\theta)^{-1} \langle G^*(\theta)\rangle_n \tag{9.12}$$

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 (9.12)*. 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 (9.12) 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 9.3 Let us consider a common type of estimating functions. To simplify the exposition we assume that the observed process is Markovian. For Markov processes it is natural to base estimating functions on functions $h_{ij}(y, x; \theta)$, $j = 1, \ldots, N$, $i = 1, \ldots, n$ satisfying that

$$E_\theta(h_{ij}(X_i, X_{i-1}; \theta)|\mathcal{F}_{i-1}) = 0. \tag{9.13}$$

Such functions define relationships (dependent on θ) between consecutive observation X_i and X_{i-1} that are, on average, equal to zero. It is natural to use such relationships to estimate θ by solving the equations $\sum_{i=1}^n h_{ij}(X_i, X_{i-1}; \theta) = 0$. 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 functions in an optimal way. We consider the class of p-dimensional estimating functions of the form

$$G_n(\theta) = \sum_{i=1}^n a_i(X_{i-1}; \theta) h_i(X_i, X_{i-1}; \theta), \tag{9.14}$$

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 is differentiable with respect to θ. It follows from (9.13) 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 (9.14) 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

$$G_n^*(\theta) = \sum_{i=1}^n a_i^*(X_{i-1}; \theta) h_i(X_i, X_{i-1}; \theta), \tag{9.15}$$

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

$$a_i^*(X_{i-1}; \theta) = -E_\theta(\partial_{\theta^T} h_i(X_i, X_{i-1}; \theta)|\mathcal{F}_{i-1})^T V_{h_i}(X_{i-1}; \theta)^{-1}, \tag{9.16}$$

then the condition (9.12) is satisfied. Hence by Theorem 9.2 the estimating function $G_n^*(\theta)$ with a_i^* given by (9.16) 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 (9.15) 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

$$U_n(\theta) = \sum_{i=1}^{n} \partial_\theta \log p_i(X_i;\theta|X_{i-1}). \tag{9.17}$$

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)

$$g^*_i(x_{i-1}, x;\theta) = a^*_i(x_{i-1};\theta)h_i(x_{i-1}, x;\theta), \tag{9.18}$$

where a^*_i is the matrix defined by (9.16). To see this, note that g^* must have the form (9.18) 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 (9.18) 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 (9.16).

□

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.

References

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.

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. (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. & Fearnhead, P. (2006). "Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes". *J. Roy. Statist. Soc. B*, 68:333–382.

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.

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.

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.

Clement, E. (1997). "Estimation of diffusion processes by simulated moment methods". *Scand. J. Statist.*, 24:353–369.

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.* To appear.

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. (2008). "Estimation for stochastic differential equations with a small diffusion coefficient". *Stoch. Proc. Appl.* To appear.

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, 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.; Heaton, J. C. & Ogaki, M. (1988). "Efficiency bounds implied by multiperiod conditional restrictions". *Journal of the American Statistical Association*, 83:863–871.

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.

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. (2008). "Aspects of asymptotic statistical theory for stochastic processes.". Preprint, Department of Mathematical Sciences, University of Copenhagen. In preparation.

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. & Sørensen, M. (1999). "Estimating equations based on eigenfunctions for a discretely observed diffusion process". *Bernoulli*, 5:299–314.

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.

Kusuoka, S. & Yoshida, N. (2000). "Malliavin calculus, geometric mixing, and expansion of diffusion functionals". *Probability Theory and Related Fields*, 116:457–484.

Larsen, K. S. & Sørensen, M. (2007). "A diffusion model for exchange rates in a target zone". *Mathematical Finance*, 17:285–306.

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.

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.

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.

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.

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. (2000). "Prediction-Based Estimating Functions". *Econometrics Journal*, 3:123–147.

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.

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.

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.