Estimating Functions for Discretely Sampled Diffusion-Type Models

Bo Martin Bibby
Institute of Mathematics and Physics
The Royal Veterinary and Agricultural University
Thorvaldsensvej 40
DK-1871 Frederiksberg C, Denmark

Martin Jacobsen
Department of Applied Mathematics and Statistics
University of Copenhagen
Universitetsparken 5
DK-2100 København Ø, Denmark

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

July 16, 2004

1 Introduction

Estimating functions provide a general framework for finding estimators and studying their properties in many different kinds of statistical models, including stochastic process models. An estimating function is a function of the data as well as of the parameter to be estimated. An estimator is obtained by equating the estimating function to zero and solving the resulting estimating equation with respect to the parameter. 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).

The estimating function approach has turned out to be very useful in obtaining, improving and studying estimators for discretely sampled parametric diffusion-type models, where the likelihood function is usually not explicitly known. Estimating functions are often 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. The

approach is thus closely related to the generalized method of moments (GMM) of Hansen (1982). By differentiation of the GMM criterion an estimating function is obtained for which the GMM estimator is a zero point. Thus GMM estimators are covered by the theory in this chapter provided the GMM criterion is differentiable. More generally, estimators obtained by maximization or minimization of a differentiable objective function are zero points for the estimating function obtained by differentiating the objective function. In particular, the estimating function corresponding to the likelihood function is the score function (the derivative of the log-likelihood function).

There are a number approaches that render likelihood inference and Bayesian inference feasible for ordinary diffusion models; for likelihood inference, see Pedersen (1995), Poulsen (1999), Aït-Sahalia (2002), Durham & Gallant (2002), and Aït-Sahalia, Hansen & Scheinkman (2003), and for Markov chain Monte Carlo methods, see Elerian, Chib & Shephard (2001), Eraker (2001), Roberts & Stramer (2001), and Johannes & Polson (2003). Markov chain Monte Carlo methods can also be used for more general diffusion-type models such as stochastic volatility models, while it is not clear that the approaches to likelihood inference mentioned here can be generalized to such more general models. Moreover, the usual asymptotic results for the maximum likelihood estimator (and Bayesian estimators) have not yet been established for stochastic volatility models. An approximate likelihood function for stochastic volatility models with tractable asymptotics was proposed by H. Sørensen (2003). Another useful approach to inference for general diffusion-type models is indirect inference, see Gallant & Tauchen (1996) and Gallant & Tauchen (2003).

In this chapter we shall only to a very limited degree be concerned with estimators obtained by maximizing or minimizing objective functions. The focus of the chapter is on estimating functions constructed directly by combining functions of observations at one or more time points. We shall demonstrate that such estimating functions can be found not only for ordinary diffusions, but also for stochastic volatility models and diffusions with jumps. For stochastic volatility models the estimating functions will be constructed in such a way that asymptotic properties of the estimator can easily be established. The main advantage of the estimating functions discussed in this chapter is that they usually require less computation than the alternative methods listed above and in several cases actually provide explicit estimators. It is therefore a particularly useful approach when quick estimators are needed. These simple estimators have a rather high efficiency when the estimating function is well-chosen. This is explained by the general theory presented in this chapter. The hall-mark of the estimating functions approach is the use of a given collection of relations between observations at different time points to construct the most efficient estimator possible on the basis of these, i.e. to combine the relations in the way that extracts as much information from the data as possible (with the given relations).

Let us give a few examples of estimating functions for a diffusion model given by the stochastic differential equation

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

where W is a Wiener process, and where θ is a parameter to be estimated. To simplify the exposition, let us assume here that X and θ are one-dimensional and that the data are observations of X at the time points $1, 2, \ldots, n$. Hansen & Scheinkman (1995) proposed the following simple and broadly applicable estimating function. For any twice continuously differentiable

function h an estimating function can be defined by

$$G_n(\theta) = \sum_{i=1}^{n} \left(b(X_i;\theta)h'(X_i) + \tfrac{1}{2}\sigma^2(X_i;\theta)h''(X_i) \right).$$

One advantage of this estimating function is that it is an explicit function of θ. The estimator obtained by solving the estimating equation $G_n(\theta) = 0$ is consistent under weak conditions. Hansen & Scheinkman (1995) also introduced an easily implementable estimating function where each term depends on a pair of consecutive observations that will be presented in Subsection 3.3. Another type of estimating function introduced by Bibby & Sørensen (1995) is

$$G_n(\theta) = \sum_{i=1}^{n} \frac{\partial_\theta b(X_{i-1};\theta)}{\sigma^2(X_{i-1};\theta)} \left[X_i - \mathrm{E}_\theta(X_i \,|\, X_{i-1}) \right],$$

which is an approximation to the optimal estimating function based on the relationship given by the function $h(x, y; \theta) = y - F(x;\theta)$ with $F(x;\theta) = \mathrm{E}_\theta(X_2 \,|\, X_1 = x)$. It can also be obtained by compensating a discretization of the continuous-time score function. This estimating function is a martingale, which simplifies the asymptotic theory. A disadvantage is that for most models there is not an explicit expression for the conditional expectation $F(x;\theta)$, which must in such cases be determined numerically. For models with mean reversion there is an explicit expression for $F(x;\theta)$. Let us finish this list of examples with an explicit martingale estimating function. For the diffusion on the interval $(-\pi/2, \pi/2)$ with $b(x;\theta) = -\theta\tan(x)$ and $\sigma(x;\theta) = 1$

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

(Kessler & Sørensen (1999)) is an approximation to the optimal estimating function based on the relationship given by the function $h(x, y; \theta) = \sin(y) - e^{-(\theta+\frac{1}{2})}\sin(x)$. The estimating equation $G_n(\theta) = 0$ has an explicit solution. The three examples given here will be treated more fully later in this chapter.

The asymptotic theory for estimating functions and the estimators obtained from them is presented in Section 2. Particular emphasis is given to martingale estimating functions for which the limit theory is relatively simple. A collection of useful asymptotic results for ergodic diffusion processes is the contents of Subsection 2.3. Various types of estimating functions for diffusion models are presented in Section 3. The maximum likelihood estimator is briefly considered. Then several types of martingale estimating functions are presented. A thorough discussion is given of how to construct explicit estimating functions, whether martingales or not. Computational aspects in cases where the estimating function is not explicit are briefly discussed. Finally, estimating functions that can be used for non-Markovian diffusion-type models are treated. Stochastic volatility models are discussed in particular. The general theory of optimal estimating functions is presented in Section 4. Again emphasis is given to the important case of martingale estimating functions. It is explained how to find the optimal linear combination of a given collection of relations between the observations, i.e. the combination that yields the estimator with the smallest asymptotic variance. In Section 5 the general theory is applied to the estimating functions introduced in Section 3. In practice, considerable computational simplifications can often be obtained by using a suitable approximation to the optimal estimating function. This aspect is discussed in Subsection 5.2. A new global optimality criterion for estimating functions for diffusion models is the subject of Section 5.3. This criterion is particularly suitable at high sampling frequencies.

2 Asymptotic Theory for Estimating Functions

Suppose as a model for the data $X_1, X_2, \ldots, X_n$ that they are observations from a stochastic process model indexed by a p-dimensional parameter $\theta \in \Theta$. The model could be a continuous time model observed at discrete time points that need not be equidistant. An *estimating function* is a p-dimensional function of the parameter θ and the data:

$$G_n(\theta; X_1, X_2, \ldots, X_n).$$

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

$$G_n(\theta) = 0.$$

It is possible that there are more than one solution or no solution at all. An estimating function is called *unbiased* if $E_\theta(G_n(\theta)) = 0$. This natural requirement is also called Fisher consistency. It ensures consistency of the estimator as $n \to \infty$ under weak regularity conditions like those imposed below.

2.1 Asymptotic Properties of Estimators

We first present a general result concerning asymptotic properties of estimators obtained from estimating functions. To simplify matters, we will only consider unbiased estimating functions of the form

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

where the function g is p-dimensional, and where r is a fixed integer smaller than n. This form is sufficiently general to cover all examples considered in this chapter. We will suppose that for all values of θ the process $\{X_i\}$ is stationary and that

$$Q_\theta(g(\theta)) = 0 \quad \text{and} \quad Q_\theta(g(\theta)^2) < \infty \tag{2.2}$$

for all $\theta \in \Theta$. Here Q_θ denotes the joint distribution of $(X_1, \ldots, X_r)$, and $Q_\theta(f)$ is the expectation of $f(X_1, \ldots, X_r)$ for a function $f : \mathbb{R}^r \mapsto \mathbb{R}$. We will further assume that $\{X_i\}$ is sufficiently mixing that as $n \to \infty$

$$\frac{1}{n} \sum_{i=r}^{n} f(X_{i-r+1}, \ldots, X_i) \xrightarrow{P_\theta} Q_\theta(f) \tag{2.3}$$

for any function $f : \mathbb{R}^r \mapsto \mathbb{R}$ such that $Q_\theta(|f|) < \infty$, and that

$$\frac{1}{\sqrt{n}} \sum_{i=r}^{n} g(X_{i-r+1}, \ldots, X_i; \theta) \xrightarrow{\mathcal{D}} N(0, V(\theta)) \tag{2.4}$$

for some $p \times p$-matrix $V(\theta)$. Note that (2.2) and (2.3) implies that

$$n^{-1} G_n(\theta) \xrightarrow{P_\theta} 0, \tag{2.5}$$

which is necessary for the consistency of the estimator obtained by solving $G_n(\theta) = 0$. Conditions ensuring that (2.3) and (2.4) hold for a diffusion model will be given in Subsection 2.3. For martingale estimating functions the situation is particularly simple and will be discussed in Subsection 2.2.

The conditions imposed on the process $\{X_i\}$ and the following condition on the function $g(x_1, \ldots, x_r; \theta)$ are sufficient to ensure the existence of a consistent and asymptotically normal estimator. From now on θ_0 will denote the true value of θ.

Condition 2.1
(1) The function g is twice continuously differentiable with respect to θ for all $x_1, \ldots, x_r$.

(2) The functions $(x_1, \ldots, x_r) \mapsto g_i(x_1, \ldots, x_r; \theta)$, $(x_1, \ldots, x_r) \mapsto \partial_{\theta_j} g_i(x_1, \ldots, x_r; \theta)$, and $(x_1, \ldots, x_r) \mapsto \partial_{\theta_i}\partial_{\theta_j} g_k(x_1, \ldots, x_r; \theta)$, $i, j, k = 1, \ldots, p$, are all locally dominated integrable w.r.t. Q_{θ_0}.

(3) The $p \times p$ matrix

$$S(\theta_0) = \left\{ Q_{\theta_0}^{\Delta} \left(\partial_{\theta_j} g_i(\theta_0) \right) \right\} \tag{2.6}$$

is invertible.

A function $f(x_1, \ldots, x_r; \theta)$ is called locally dominated integrable w.r.t. the measure Q on $\mathbb{R}^r$ if for each $\theta_* \in \Theta$ there exists a neighbourhood U_{θ_*} of θ_* and a non-negative Q-integrable function h_{θ_*} such that $| f(x_1, \ldots, x_r; \theta) | \leq h_{\theta_*}(x_1, \ldots, x_r)$ for all $x_1, \ldots, x_r$ in the support of Q and for all $\theta \in U_{\theta_*}$.

Theorem 2.2 *Suppose* (2.2), (2.3), (2.4), *and Condition 2.1 are satisfied. Then for every n an estimator $\hat{\theta}_n$ exists that solves the estimating equation $G_n(\hat{\theta}_n) = 0$ with a probability tending to one as $n \to \infty$. Moreover,*

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

as $n \to \infty$, and

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

Theorem 2.2 can be proved in complete analogy with the proof for the case $r = 2$ given in Sørensen (1999). Similar results can be found in Hansen (1982).

The asymptotic covariance matrix is the inverse of $S(\theta_0)V(\theta_0)^{-1}(S(\theta_0)^T$, which is an asymptotic version of the *Godambe information* that will be discussed further in Section 4.

2.2 Martingale Estimating Functions

Estimating functions that are martingales have particularly nice properties and a relatively simple asymptotic theory based on the well-developed martingale limit theory. In this subsection we shall give an asymptotic result for martingale estimating functions, i.e. estimating functions G_n satisfying that

$$\mathrm{E}_\theta(G_n(\theta)|\mathcal{F}_{n-1}) = G_{n-1}(\theta), \quad n = 1, 2, \ldots,$$

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. As will be discussed in Subsection 3.1, the score function is usually a martingale estimating function (for more details see e.g. Barndorff-Nielsen & Sørensen (1994)). When a more easily calculated alternative is needed, it is natural to approximate the score function by a simpler martingale estimating function. A simple typical example of a martingale estimating function is

$$G_n(\theta) = \sum_{i=1}^{n} a(X_{i-1}; \theta) \left[f(X_i) - \mathrm{E}_\theta(f(X_i) \,|\, \mathcal{F}_{i-1}) \right],$$

where f is some suitable (possibly multivariate) function of the data, while the function a (typically a matrix) can be chosen to ensure that the dimension of $G_n(\theta)$ equals the dimension of the parameter and to improve the estimator. We shall discuss how best to choose a in Section 4. The function a is usually called an instrument in the econometric literature. In the following, a version of the central limit theorem for martingales is given.

We write the martingale $G_n(\theta)$ in the form $G_n(\theta) = \sum_{i=1}^n H_i(\theta)$ with $H_i(\theta) = G_i(\theta) - G_{i-1}(\theta)$. From now on we assume, as we did in Subsection 2.1, that $G_n(\theta)$ has variance, and define the *quadratic characteristic* of $G_n(\theta)$ as the random positive semi-definite $p \times p$-matrix

$$\langle G(\theta) \rangle_n = \sum_{i=1}^{n} \mathrm{E}_\theta \left(H_i(\theta) H_i(\theta)^T | \mathcal{F}_{i-1} \right). \tag{2.7}$$

Theorem 2.3 *Suppose that as* $n \to \infty$

$$\frac{1}{n} \sum_{i=1}^{n} E_\theta \left(H_i(\theta) H_i(\theta)^T \right) \to \Sigma_\theta, \tag{2.8}$$

$$\langle G(\theta) \rangle_n / n \xrightarrow{P_\theta} \Sigma_\theta, \tag{2.9}$$

and

$$\frac{1}{\sqrt{n}} \sup_{i \le n} |H_i(\theta)| \xrightarrow{P_\theta} 0, \tag{2.10}$$

where Σ_θ *is a positive definite* $p \times p$*-matrix. Then*

$$\langle G(\theta) \rangle_n^{-\frac{1}{2}} G_n(\theta) \xrightarrow{\mathcal{D}} N(0, I_p). \tag{2.11}$$

Here I_p is the $p \times p$ identity matrix, and $N(0, I_p)$ denotes the standard normal distribution in $\mathbb{R}^p$. A proof of a one-dimensional version of the theorem (and other more general versions of the central limit theorem for martingales) can be found in Hall & Heyde (1980). The multivariate version follows by the Cramér-Wold device. More general versions of the multivariate central limit theorem for martingales can be found in Heyde (1997) and Küchler & Sørensen (1999). It is not difficult to see that under the conditions of Theorem 2.3 we have a weak law of large numbers:

$$n^{-1} G_n(\theta) \xrightarrow{P_\theta} 0, \tag{2.12}$$

which is a necessary condition for consistency, cf. (2.5).

2.3 Limit Results for Diffusion Processes

In this subsection we review asymptotic results for ergodic diffusion processes. We shall mainly consider one-dimensional diffusion models, i.e. solutions to stochastic differential equations of the form

$$dX_t = b(X_t;\theta)dt + \sigma(X_t;\theta)dW_t, \tag{2.13}$$

where W is a standard Wiener process. We assume that the drift b and the diffusion coefficient σ depend on a parameter θ which varies in a subset Θ of $\mathbb{R}^p$. Estimation of θ will not be discussed in this section, but the parameter is included for consistency with the rest of the section. The coefficients b and σ are assumed to be smooth enough functions of the state to ensure the existence of a unique weak solution for all θ in Θ. The *state space* of X, i.e. the set of possible values of the process, is an interval from ℓ to r, where ℓ could possibly be $-\infty$ and r might be ∞. The state space is assumed not to depend on θ.

First we give a condition ensuring that the solution X of (2.13) is ergodic. The *scale measure* of X is a measure on the state space of X with the density

$$s(x;\theta) = \exp\left(-2\int_{x^\#}^{x} \frac{b(y;\theta)}{v(y;\theta)}dy\right) \tag{2.14}$$

with respect to the Lebesgue measure. Here $x^\#$ is an arbitrary point in (ℓ, r). Since we shall often need the squared diffusion coefficient, we define

$$v(x;\theta) = \sigma^2(x;\theta). \tag{2.15}$$

Condition 2.4 *The following holds for all $\theta \in \Theta$:*

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

and

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

Under Condition 2.4 the process X is ergodic with an invariant probability measure that has density

$$\mu_\theta(x) = [A(\theta)s(x;\theta)v(x;\theta)]^{-1}, \quad x \in (\ell, r), \tag{2.16}$$

with respect to the Lebesgue measure on (ℓ, r). We will assume that $X_0 \sim \mu_\theta$, so that X is a stationary process with $X_t \sim \mu_\theta$ for all $t \geq 0$. The distribution of (X_t, X_{t+s}) $t > 0, s > 0$ has density

$$Q_\theta^s(x, y) = \mu_\theta(x)p(s, x, y;\theta), \tag{2.17}$$

where $y \mapsto p(s, x, y;\theta)$ is the transition density, i.e. the conditional density of X_{t+s} given that $X_t = x$. For a function $f : (\ell, r)^2 \mapsto \mathbb{R}$, we use the notation

$$Q_\theta^s(f) = \int_{(\ell,r)^2} f(x, y)p(s, x, y;\theta)\mu_\theta(x)dydx$$

(provided, of course, that the integral exists). Similarly we define

$$\mu_\theta(f) = \int_{\ell}^{r} f(x)\mu_\theta(x)dx$$

for a function $f : (\ell, r) \mapsto \mathbb{R}$.

Suppose Condition 2.4 holds, that $f : (\ell, r) \mapsto \mathbb{R}$ satisfies $\mu_\theta(|f|) < \infty$, and that $g : (\ell, r)^2 \mapsto \mathbb{R}$ satisfies $Q_\theta^\Delta(|g|) < \infty$ for a $\Delta > 0$. Then

$$\frac{1}{n}\sum_{i=1}^{n} f(X_{i\Delta}) \xrightarrow{a.s.} \mu_\theta(f) \tag{2.18}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} g(X_{(i-1)\Delta}, X_{i\Delta}) \xrightarrow{a.s.} Q_\theta^\Delta(g) \tag{2.19}$$

as $n \to \infty$, see Billingsley (1961b). The result (2.18) is obviously a particular case of (2.19). Note that these results require equidistant observations, i.e. $t_i = \Delta i$ to ensure that the sequences $f(X_{i\Delta})$ and $g(X_{(i-1)\Delta}, X_{i\Delta})$ are stationary.

If we assume that the sum $\sum_{i=1}^n g(X_{(i-1)\Delta}, X_{i\Delta})$ is a martingale with finite variance, i.e. that

$$\int_\ell^r g(x, y)p(\Delta, x, y; \theta)dy = 0 \quad \text{for all } x \in (\ell, r)$$

and that $Q_\theta^\Delta(g^2) < \infty$, then under Condition 2.4

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_{(i-1)\Delta}, X_{i\Delta}) \xrightarrow{\mathcal{D}} N\left(0, Q_\theta^\Delta(g^2)\right) \tag{2.20}$$

as $n \to \infty$, see Billingsley (1961a).

In cases where $\sum_{i=1}^n g(X_{(i-1)\Delta}, X_{i\Delta})$ is not a martingale, stronger conditions on the diffusion are needed to ensure a central limit result like (2.20). The following sufficient condition was given by Genon-Catalot, Jeantheau & Larédo (2000).

Condition 2.5
(i) The function b is continuously differentiable with respect to x and σ is twice continuously differentiable 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 $\gamma(x; \theta) = \partial_x\sigma(x; \theta) - 2b(x; \theta)/\sigma(x; \theta)$.

This condition implies that the process X is geometrically α-mixing, i.e. α-mixing with mixing coefficients that tend to zero geometrically fast. Other conditions for geometrical α-mixing were given by Veretennikov (1987), Doukhan (1994), Hansen & Scheinkman (1995), and Kusuoka & Yoshida (2000), see also Aït-Sahalia, Hansen & Scheinkman (2003).

In order to discuss the non-martingale case, we need the transition operator π_s^θ that maps a function f satisfying $\mu_\theta(|f|) < \infty$ into the function $\pi_s^\theta(f)$ given by

$$\pi_s^\theta(f)(x) = \int_\ell^r f(y)p(s, x, y; \theta)dy = \mathrm{E}_\theta(f(X_s) \mid X_0 = x). \tag{2.21}$$

Under Condition 2.5 the operator $I - \pi_s^\theta$ has a bounded inverse U_s^θ on the set of functions f satisfying that $\mu_\theta(f^2) < \infty$ and $\mu_\theta(f) = 0$. Here I denotes the identity operator defined by $I(f) = f$. The operator U_s^θ is called the potential of X and can be represented as

$$U_s^\theta(f)(x) = \sum_{k=0}^{\infty} \pi_{ks}^\theta(f)(x) = \sum_{k=0}^{\infty} \left(\pi_s^\theta\right)^k (f)(x) \tag{2.22}$$

where the convergence of the sum is in the $L_2(\mu_\theta)$-sense, i.e. with respect to the norm $\|f\| = \mu_\theta(f^2)^{\frac{1}{2}}$. The potential operator can also be applied to a function of two variables $g(x, y)$, provided that $Q_\theta^\Delta(g^2) < \infty$ and $Q_\theta^\Delta(g) = 0$. For such a function, $U_s^\theta(g) = U_s^\theta(\pi_s^\theta(g))$, where $\pi_s^\theta(g)(x) = \mathrm{E}_\theta(g(x, X_s) \,|\, X_0 = x)$.

Suppose Conditions 2.4 and 2.5 hold, and that the function $g : (\ell, r)^2 \mapsto \mathbb{R}$ satisfies $Q_\theta^\Delta(g^2) < \infty$ and $Q_\theta^\Delta(g) = 0$ for a $\Delta > 0$. Then as $n \to \infty$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_{(i-1)\Delta}, X_{i\Delta}) \xrightarrow{\mathcal{D}} N(0, V(\theta)) \tag{2.23}$$

where

$$V(\theta) = Q_\theta^\Delta(\tilde{g}_\theta^2), \tag{2.24}$$

with

$$\tilde{g}_\theta(x, y) = g(x, y) + U_\Delta^\theta(g)(y) - U_\Delta^\theta(g)(x). \tag{2.25}$$

It is not difficult to see that

$$\sum_{i=1}^{n} \tilde{g}_\theta(X_{(i-1)\Delta}, X_{i\Delta}) \tag{2.26}$$

is a martingale. This fact can be utilized to prove (2.23) via the central limit theorem for a martingale sum (2.20). For a detailed proof, see Florens-Zmirou (1984) and Florens-Zmirou (1989). See also the discussion in Hansen & Scheinkman (1995) and in Jacobsen (2001a).

Finally, we consider briefly the case where X is a multivariate diffusion, i.e. when X is the d-dimensional process that solves (2.13) with b now a d-dimensional vector, σ a $d \times d$-matrix and W a d-dimensional standard Wiener process. We assume that X moves freely on an open, connected set $D \subseteq \mathbb{R}^d$ (that does not depend on θ), that $C(x; \theta) = \sigma(x; \theta)\, \sigma(x; \theta)^T$ is strictly positive definite for all $x \in D$, $\theta \in \Theta$, and that X is ergodic for all θ with an invariant density $\mu_\theta(x)$. Under these assumptions the above results (2.18), (2.19) and (2.20) hold in the multivariate case too. The problem is, that there are no simple conditions ensuring ergodicity similar to those given for the one-dimensional case. Also (2.23) holds provided X is geometrically α-mixing.

3 Estimating Functions for Diffusion-Type Processes

Suppose a d-dimensional continuous time process X has been observed at discrete time points: $X_{t_0}, X_{t_1}, \cdots, X_{t_n}$, $t_0 = 0 < t_1 < \cdots < t_n$. As a model for these data, we assume that X is a d-dimensional diffusion, i.e. that X solves the stochastic differential equation (2.13) with b a d-dimensional vector, σ a $d \times d$-matrix and W a d-dimensional standard Wiener process. We assume that the drift b and the diffusion coefficient σ are known apart from the parameter θ

which varies in a subset Θ of $\mathbb{R}^p$. These functions are assumed to be smooth enough to ensure the existence of a unique weak solution for all θ in Θ. The statistical problem considered here is to draw inference about the parameter θ based on the observations. We consider only the case where the sampling-times are not random. The effect of random sampling-times can be considerable, see Aït-Sahalia & Mykland (2003).

3.1 Maximum Likelihood Estimation

The diffusion process X is a Markov process, so the likelihood function (conditional on X_0) is

$$L_n(\theta) = \prod_{i=1}^{n} p(t_i - t_{i-1}, X_{t_{i-1}}, X_{t_i}; \theta), \tag{3.1}$$

where $y \mapsto p(s, x, y; \theta)$ is the transition density, i.e. the conditional density of X_{t+s} given that $X_t = x$ ($s > 0$). 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 there are a number of numerical approaches that render likelihood inference feasible for diffusion models. Pedersen (1995) proposed a method for obtaining an approximation to the likelihood function by rather extensive simulation, Poulsen (1999) obtained an approximation to the transition density by numerically solving a partial differential equation, while Aït-Sahalia (2002) and Aït-Sahalia (2003) proposed to approximate the transition density by means of a Hermite expansion, see also Aït-Sahalia, Hansen & Scheinkman (2003). 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), Roberts & Stramer (2001), and Johannes & Polson (2003). These various approaches to maximum likelihood estimation will not be considered further in this chapter.

The vector $U_n(\theta)$ of partial derivatives of the log-likelihood function $\log L_n(\theta)$ with respect to the coordinates of θ is called the score function (or score vector). The maximum likelihood estimator solves the estimating equation $U_n(\theta) = 0$. The score function based on the observations $X_{t_0}, X_{t_1}, \cdots, X_{t_n}$ is

$$U_n(\theta) = \sum_{i=1}^{n} \partial_\theta \log p(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta), \tag{3.2}$$

where $\Delta_i = t_i - t_{i-1}$. The score function is a martingale estimating function, which is easily seen provided that the following interchange of differentiation and integration is allowed:

$$E_\theta\left(\partial_\theta \log p(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta)\Big| X_{t_1}, \ldots, X_{t_{i-1}}\right) = E_\theta\left(\frac{\partial_\theta p(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta)}{p(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta)}\Big| X_{t_{i-1}}\right)$$

$$= \int_\ell^r \frac{\partial_\theta p(\Delta_i, X_{t_{i-1}}, y; \theta)}{p(\Delta_i, X_{t_{i-1}}, y; \theta)} p(\Delta_i, X_{t_{i-1}}, y, \theta) dy = \partial_\theta \underbrace{\int_\ell^r p(\Delta_i, X_{t_{i-1}}, y; \theta) dy}_{=1} = 0.$$

A wide spectre of estimators based on estimating functions other than the score function have been proposed and are useful alternatives to the maximum likelihood estimator in situation where simpler estimators that require less computation are needed. Some of these alternatives

are not much less efficient than the maximum likelihood estimator, and in some cases they are even fully efficient. Another advantage of these alternative approaches is that the estimators are often more robust to model misspecification than the maximum likelihood estimator, because typically the estimating functions do not involve the full model specification. For instance the martingale estimating functions considered below depends only on the conditional moments of certain functions of the observations. In the following subsections some of these alternative estimating functions will be reviewed and discussed.

Estimators obtained by maximizing or minimizing other objective functions than the likelihood function can also be thought of as solutions to the estimating equation obtained by differentiating the objective function, provided of course that it is differentiable. An important example is the generalized method of moments of Hansen (1982). We will not explicitly treat such estimators in this chapter, rather we consider estimating functions that are obtained from suitable functions of the data or relationships between the observations at different time points.

3.2 Martingale Estimating Functions for Diffusion models

The score function is a martingale estimating function of the form

$$G_n(\theta) = \sum_{i=1}^{n} g(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta). \tag{3.3}$$

It is therefore natural to approximate the score function by martingale estimating functions of the general form (3.3) with

$$g(\Delta, x, y; \theta) = \sum_{j=1}^{N} a_j(\Delta, x; \theta) h_j(\Delta, x, y; \theta), \tag{3.4}$$

where $h_j(\Delta, x, y, ; \theta)$, $j = 1, \ldots, N$ are given real valued functions satisfying that

$$\int_{\ell}^{r} h_j(\Delta, x, y; \theta) p(\Delta, x, y; \theta) dy = 0$$

for all $\Delta > 0$, $x \in (\ell, r)$, and $\theta \in \Theta$. Each of the functions h_j could separately be used to define an estimating function of the form (3.3), but more efficient estimators are obtained by combining them in an optimal way. The p-dimensional functions a_j in (3.4) determine how much weight is given in the estimation procedure to each of the relationships defined by the h_js. These functions, which we will refer to as the *weights*, can be chosen in an optimal way using the theory of optimal estimating functions. This is quite straightforward and is the subject of Section 4. 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, on the other hand, also a source of weakness, since it is not always clear how best to choose the h_js. However, for diffusion models the global small Δ-optimality criterion presented in Section 5.3 gives some guidance to the choice of the functions h_j. In what follows and in Subsection 3.3, we shall present some standard ways of choosing these functions that usually work in practice. Note that the weights a_j are usually called *instruments* in the econometric literature.

Martingale estimating functions have turned out to be very useful in obtaining estimators for discretely sampled diffusion-type models; see for instance Bibby & Sørensen (1995), Bibby &

Sørensen (1996), Sørensen (1997), Kessler & Sørensen (1999), Kessler (2000), Bibby & Sørensen (2001), and Jacobsen (2001a). An application to financial data can be found in Bibby & Sørensen (1997), while Pedersen (2000) used the method to estimate the nitrous oxide emission rate from a soil surface. In Christensen, Poulsen & Sørensen (2001) the martingale estimating functions approach is compared to other methods commonly used in financial econometrics.

A simple type of estimating function is the *linear estimating function* obtained for $N = 1$ and

$$h_1(\Delta, x, y; \theta) = y - F(\Delta, x; \theta),$$

where

$$F(\Delta, x; \theta) = \mathrm{E}_\theta(X_\Delta | X_0 = x) = \int_\ell^r y p(\Delta, x, y; \theta) dy. \tag{3.5}$$

In some models the conditional expectation $F(\Delta, x; \theta)$ and the conditional variance $\phi(\Delta, x; \theta)$ are known, but in most cases they are not and must be determined by simulations which can usually be done easily. The simplest way is straightforward. Fix θ and simulate numerically M independent trajectories $\{X_{\delta i}^{(j)}, i = 1, \ldots, N\}$ with $X_0 = x$ $(j = 1, \ldots, M)$, where $\delta = \Delta/N$ for N sufficiently large. If M is sufficiently large, the approximation $F(\Delta, x; \theta) \doteq \frac{1}{M} \sum_{j=1}^M X_\Delta^{(j)}$ can be applied. Methods for simulating a diffusion process can be found in Kloeden & Platen (1999).

Linear martingale estimating functions for diffusion models were studied by Bibby & Sørensen (1995), where they were derived as an approximation to the continuous time likelihood function. An advantage of this type of estimating functions is that the estimators are very robust to model misspecification. If only the first moment F of the transition distribution is correctly specified, the estimator is consistent.

When the diffusion coefficient (the volatility) σ depends on a parameter, the linear estimating function are too simple to be useful, whereas the *quadratic estimating functions* are a natural, generally applicable choice. They are given by $N = 2$ and, when the diffusion is one-dimensional, by

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

where

$$\phi(\Delta, x; \theta) = \mathrm{Var}_\theta(X_\Delta | X_0 = x) = \int_\ell^r [y - F(\Delta, x; \theta)]^2 p(\Delta, x, y; \theta) dy. \tag{3.6}$$

The version for multivariate diffusions is defined in an analogous way. An argument for using this type of estimating function goes as follows. When Δ is small, the transition density $p(\Delta, x, y; \theta)$ is well approximated by a Gaussian density function with expectation $F(\Delta, x; \theta)$ and variance $\phi(\Delta, x; \theta)$. By inserting this Gaussian density in the expression for the likelihood function (3.1), an approximate likelihood function is obtained, and the corresponding approximate score function is

$$\begin{aligned} \sum_{i=1}^n \Bigg\{ & \frac{\partial_\theta F(\Delta_i, X_{t_{i-1}}; \theta)}{\phi(\Delta_i, X_{t_{i-1}}; \theta)} [X_{t_i} - F(\Delta_i, X_{t_{i-1}}; \theta)] \\ & + \frac{\partial_\theta \phi(\Delta_i, X_{t_{i-1}}; \theta)}{2\phi^2(\Delta_i, X_{t_{i-1}}; \theta)\Delta_i} [(X_{t_i} - F(\Delta_i, X_{t_{i-1}}; \theta))^2 - \phi(\Delta_i, X_{t_{i-1}}; \theta)] \Bigg\}. \end{aligned} \tag{3.7}$$

Quadratic martingale estimating functions for diffusion models were considered in Bibby & Sørensen (1996). Estimators obtained from this type of estimating functions are also rather robust to model misspecification. If the first and the second moments, F and ϕ, of the transition distribution are correctly specified, the estimator is consistent.

Example 3.1 For a *mean-reverting diffusion model* given by

$$dX_t = -\beta(X_t - \alpha)dt + \sigma(X_t)dW_t, \tag{3.8}$$

where $\beta > 0$,

$$F(t, x; \alpha, \beta) = xe^{-\beta t} + \alpha(1 - e^{-\beta t}) \tag{3.9}$$

under weak conditions on σ. This can be seen by noting that for fixed x, α and β the function $f(t) = F(t, x; \alpha, \beta)$ solves the ordinary differential equation $f' = -\beta(f - \alpha)$. Thus linear estimating functions can be easily calculated.

If we make the further assumption that $\sigma(x) = \tau\sqrt{x}$ $(\tau > 0)$, we obtain the model proposed by Cox, Ingersoll, Jr. & Ross (1985) for interest rates (the spot rate). In the rest of the chapter we will refer to this model as the CIR model or the CIR process. For the CIR model the function ϕ and hence quadratic estimating functions can be found explicitly:

$$\phi(x; \alpha, \beta, \tau) = \frac{\tau^2}{\beta}\left((\tfrac{1}{2}\alpha - x)e^{-2\beta} - (\alpha - x)e^{-\beta} + \tfrac{1}{2}\alpha\right).$$

Another model, where ϕ can be found explicitly is the mean-reverting model with $\sigma = \sqrt{\beta + x^2}$. For this model

$$\phi(x; \beta) = x^2 e^{-2\beta}(e - 1) + \frac{\beta}{2\beta - 1}\left(1 - e^{1-2\beta}\right)$$

when $\alpha = 0$. □

A natural generalization of the quadratic martingale estimating functions is obtained by choosing h_js of the form

$$h_j(\Delta, x, y; \theta) = f_j(y; \theta) - \pi_\Delta^\theta(f_j(\theta))(x) \tag{3.10}$$

for suitably chosen functions f_j and with the transition operator π_Δ^θ defined by (2.21). We will refer to the functions f_j, $j = 1, \ldots, N$ as the *base* of the estimating function. Almost all martingale estimating functions proposed in the literature are of this form. An example is the higher order *polynomial martingale estimating functions* considered by Pedersen (1994a) and Kessler (1996). These are obtained by choosing the base as $f_j(y) = y^j$, $j = 1, \ldots, N$. However, there is no reason to believe that polynomial estimating functions are in general the best possible way to approximate the true score function when the transition distribution is far from Gaussian, and it may be useful to choose a base that is tailored to a particular diffusion model. An example are the estimating functions based on eigenfunctions of the generator of the diffusion that were proposed by Kessler & Sørensen (1999). These estimating functions and other examples will be discussed in the next subsection.

In many cases the conditional expectations needed in a martingale estimating function are not explicitly available and must be calculated numerically, for instance by means of simulations. Let us briefly consider the *effect on the variance of the estimator caused by simulation.* Suppose

that we need the conditional expectation of $f(X_{t+\Delta})$ given that $X_t = x$ for a particular value θ of the parameter. Then we can use one of the approximation schemes in Kloeden & Platen (1999) with a step size δ much smaller than Δ to generate an approximation $Y(\delta, \theta, x)$ to X starting at x. A simple example is the Euler scheme

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

where the Z_is are independent and $Z_i \sim N(0, \delta)$. As usual it is assumed that X solves (2.13). By generating N independent simulations $Y^{(j)}(\delta, \theta, x)$, $j = 1, \ldots, N$, we can approximate the conditional expectation of $f(X_{t+\Delta})$ given that $X_t = x$ by

$$\frac{1}{N}\sum_{j=1}^{N} f(Y_\Delta^{(j)}(\delta, \theta, x)).$$

This procedure is closely related to the simulated method of moments, see Duffie & Singleton (1993) and Clement (1997). The asymptotic properties of the estimators obtained when the conditional moments are approximated by simulation were investigated by Kessler & Paredes (2002), who considered approximations to martingale estimating functions of the form

$$G_n(\theta) = \sum_{i=1}^{n} \left[f(X_{i\Delta}, X_{(i-1)\Delta}; \theta) - F(X_{(i-1)\Delta}; \theta) \right], \tag{3.11}$$

where $F(x; \theta)$ is the conditional expectation of $f(X_\Delta, x; \theta)$ given $X_0 = x$ when the parameter value is θ. Let $\hat{\theta}_n^{N,\delta}$ denote the estimator obtained from the approximate martingale estimating function

$$G_n^{N,\delta}(\theta) = \sum_{i=1}^{n} \left[f(X_{i\Delta}, X_{(i-1)\Delta}; \theta) - \frac{1}{N}\sum_{j=1}^{N} f(Y_\Delta^{(j)}(\delta, \theta, X_{(i-1)\Delta}), X_{(i-1)\Delta}; \theta) \right], \tag{3.12}$$

and suppose that $Y(\delta, \theta, x)$ satisfies that there exists a $\delta > 0$ such that

$$|E_\theta(g(X_\Delta(x), x; \theta)) - E(g(Y_\Delta(\delta, \theta, x), x; \theta))| \leq R(x; \theta)\delta^\beta \tag{3.13}$$

for all $x \in \mathbb{R}$ and $\theta \in \Theta$ and for δ sufficiently small. Here $g(y, x; \theta) = f(y, x; \theta) - F(x; \theta)$, $X_t(x)$ is a solution of (2.13) with $X_0(x) = x$, and $R(x; \theta)$ is of polynomial growth in x uniformly for θ in compact sets. Under this and a number of further regularity conditions Kessler & Paredes (2002) showed that if δ goes to zero sufficiently fast that $\sqrt{n}\delta^\beta \to 0$ as $n \to \infty$, then

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

where θ_0 is the true parameter value, and where Σ denotes the asymptotic covariance matrix for the estimator obtained from the estimating function (3.11). Thus for δ sufficiently small and N sufficiently large, it does not matter much that the conditional moment $F(x; \theta)$ has been determined by simulation in (3.12). However, when $0 < \lim_{n\to\infty} \sqrt{n}\delta^\beta < \infty$,

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

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

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

in probability. Here the p-dimensional vector $m(\theta_0)$ depends on f and is generally different from zero. According to Kessler & Paredes (2002) condition (3.13) is satisfied by the order β weak schemes based on Ito-Taylor expansions given in Chapter 14 of Kloeden & Platen (1999).

3.3 Explicit Estimating Functions

In this subsection we shall focus on estimating functions for which explicit analytic expressions are available. These are particularly useful, because the problem of finding the resulting estimators then amounts to solving p explicitly given equations, and although typically the solution must be obtained numerically, that will not create practical problems if the dimension of the parameter is not too large – in particular no simulations are required for the calculations.

We start the discussion of explicit estimating functions by considering first *martingale estimating functions* of the form (3.3), (3.4) and (3.10), i.e.

$$G_n(\theta) = \sum_{i=1}^{n} a(\Delta_i, X_{t_{i-1}}, \theta) \left(f(X_{t_i}; \theta) - \pi_\Delta^\theta (f(\theta))(X_{t_{i-1}}) \right) \tag{3.14}$$

with $f = (f_j)_{1 \le j \le N}$ a (column) vector of given functions, the *base*, and $a = (a_{kj})_{1 \le k \le p, 1 \le j \le N}$ a $p \times N$-matrix of given functions, the *weights*. The transition operator π_Δ^θ is defined by (2.21). We shall call $G_n(\theta)$ explicit if all the f_j and a_{kj} are given in explicit form *and* the conditional expectations $\pi_\Delta^\theta (f(\theta))(x)$ can be determined explicitly. In this section the weight matrix a can be chosen in any way we please, so we shall not be concerned with the explicit determination of a. In the next section we shall discuss how to determine a in an optimal or approximately optimal way. Then we shall also discuss when an explicit expression for the optimal a is available.

By far the simplest case in which $\pi_\Delta^\theta (f(\theta))(x)$ can be found explicitly, is when the base consists of eigenfunctions for the generator of the diffusion as proposed by Kessler & Sørensen (1999) for one-dimensional diffusions. The differential operator

$$L_\theta = b(x;\theta)\frac{d}{dx} + \tfrac{1}{2}\sigma^2(x;\theta)\frac{d^2}{dx^2} \tag{3.15}$$

is called the *generator* of the diffusion process given by (2.13). Generators of Markov processes are treated more fully in Aït-Sahalia, Hansen & Scheinkman (2003). A twice differentiable function $\phi(x;\theta)$ is called an *eigenfunction* for the generator L_θ if

$$L_\theta \phi(x;\theta) = -\lambda(\theta)\phi(x;\theta), \tag{3.16}$$

where the real number $\lambda(\theta) \ge 0$ is called the *eigenvalue* corresponding to $\phi(x;\theta)$. Under weak regularity conditions, see e.g. Kessler & Sørensen (1999),

$$\pi_\Delta^\theta(\phi(\theta))(x) = \mathrm{E}_\theta(\phi(X_\Delta;\theta)|X_0 = x) = e^{-\lambda(\theta)\Delta}\phi(x;\theta). \tag{3.17}$$

We can therefore define a martingale estimating function by (3.3) and (3.4) with

$$h_j(\Delta, x, y; \theta) = \phi_j(y;\theta) - e^{-\lambda_j(\theta)\Delta}\phi_j(x;\theta), \tag{3.18}$$

where $\phi_1(\cdot;\theta), \cdots, \phi_N(\cdot;\theta)$ are eigenfunctions for L_θ with eigenvalues $\lambda_1(\theta), \cdots, \lambda_N(\theta)$.

Example 3.2 For the Cox-Ingersoll-Ross model the eigenfunctions are the Laguerre polynomials, and we obtain polynomial estimating functions, some of which were discussed in Example 3.1.

□

Example 3.3 The class of diffusions which solve the equation

$$dX_t = -\theta \tan(X_t)dt + dW_t,\ X_0 = x_0$$

is more interesting, because here the eigenfunctions are not polynomials, and we get estimating functions that we have not seen before. For $\theta \geq \frac{1}{2}$ the process X is an ergodic diffusion on the interval $(-\pi/2, \pi/2)$, which can be thought of as an Ornstein-Uhlenbeck process on a finite interval. The eigenfunctions are $\phi_i(x;\theta) = C_i^\theta(\sin(x)),\ i = 0, 1, \cdots$, with eigenvalues $i(\theta + i/2),\ i = 0, 1, \cdots$, where C_i^θ is the Gegenbauer polynomial of order i. This model was studied in more detail in Kessler & Sørensen (1999). An asymmetric version was introduced in Larsen & Sørensen (2003).

□

Example 3.4 In Larsen & Sørensen (2003) the following model is proposed for the random variation of an *exchange rate in a target zone* between realignments. Let X denote the logarithm of the exchange rate. Then

$$dX_t = -\beta[X_t - (m + \gamma Z)]dt + \sigma\sqrt{Z^2 - (X_t - m)^2}dW_t, \tag{3.19}$$

where $\beta > 0$ and $\gamma \in (-1, 1)$. This is a diffusion on the interval $(m - Z, m + Z)$ with mean reversion around $m + \gamma Z$. Here m denotes the central parity and $Z = \log(1 + z)$ with z denoting the largest deviation from m that is allowed. The quantities m and z are known fixed quantities. When $\beta(1 - \gamma) \geq \sigma^2$ and $\beta(1 + \gamma) \geq \sigma^2$, X an ergodic diffusion, for which the stationary distribution is the Beta-distribution on $(m - Z, m + Z)$ with parameters $\beta(1-\gamma)\sigma^{-2}$ and $\beta(1+\gamma)\sigma^{-2}$. For $\gamma = 0$ the target zone model proposed by De Jong, Drost & Werker (2001) is obtained. The purpose of introducing the parameter γ is to allow an asymmetric stationary distribution, which is usually needed to fit observations of exchange rates in a target zone, see Larsen & Sørensen (2003). The eigenfunctions for the generator of the diffusion (3.19) are the Jacobi polynomials

$$\phi_i(x;\beta,\gamma,\sigma) = \sum_{j=1}^{i} 2^{-j} \binom{\beta(1-\gamma)\sigma^{-2} + i - 1}{i - j} \binom{2\beta\sigma^{-2} - 2 + i + j}{j} [(x - m)/Z - 1]^j$$

with eigenvalues $\lambda_i = i[\beta + \frac{1}{2}\sigma^2(i - 1)]$, $i = 1, 2, \ldots$.

□

While it is quite natural to search for eigenfunctions for the generator of a one-dimensional diffusion, it is less natural in higher dimensions (e.g. the eigenvalues need no longer be real). Instead one may use invariant subspaces and do the following. Let X be a general d-dimensional diffusion satisfying (2.13) with b a d-dimensional vector and σ a $d \times d$-matrix. For a d-dimensional diffusion, the generator is defined by

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

where f is a real twice differentiable function defined on the d-dimensional state space of X and $C = \sigma\sigma^T$ with σ^T denoting the transpose of σ. Suppose that for every θ, $\mathcal{L}_\theta$ is a finite-dimensional vector space of twice differentiable real-valued functions f^* such that $L_\theta f^* \in \mathcal{L}_\theta$

for all $f^* \in \mathcal{L}_\theta$ (the simplest case is of course when $\mathcal{L}_\theta$ is a one-dimensional eigen-space). If $(f_j)_{1\le j\le N}$ is a basis for $\mathcal{L}_\theta$, we may write

$$L_\theta f = \Psi_\theta f, \tag{3.20}$$

where Ψ_θ is an $N \times N$-matrix of constants, and f is the column vector $(f_j)_{1\le j\le N}$. The basis f will typically depend on θ, but that dependence is suppressed in the notation. By $L_\theta f$ we mean that L_θ is applied to each coordinate of f, i.e. $L_\theta f$ is the column vector $(L_\theta f_j)_{1\le j\le N}$. Then by Itô's formula

$$\pi_t^\theta f(x) = f(x) + \int_0^t \Psi_\theta \left(\pi_s^\theta f\right)(x)\, ds \quad (x \in D) \tag{3.21}$$

provided all $f_j(X_s)$ are integrable, $\mathrm{E}_\theta\left(|f_j(X_s)|\,|X_0 = x\right) < \infty$, for all x, and provided each of the local martingales

$$M_t^{f_j} = \sum_{k=1}^d \int_0^t \partial_{x_k} f_j(X_s) \sum_{\ell=1}^d \sigma_{k\ell}(X_s)\, dW_{\ell,s}$$

is a true martingale conditionally on $X_0 = x$. In that case, (3.21) gives $\partial_t \pi_t^\theta f = \Psi_\theta \pi_t^\theta f$ with the boundary condition $\pi_0^\theta f = f$ so that

$$\pi_t^\theta f(x) = e^{t\Psi_\theta} f(x) \quad (x \in D) \tag{3.22}$$

with the exponential matrix defined through its series expansion,

$$e^{t\Psi_\theta} = \sum_{m=0}^\infty \frac{t^m}{m!} \Psi_\theta^m.$$

It is perhaps debatable whether (3.22) is an explicit expression, but at least, if N is not too large, a more compact expression may be found.

Note that (3.9) in Example 3.1 (where the diffusion is one-dimensional) may be deduced as a special case of (3.22) with $\mathcal{L}_\theta = \mathcal{L}$ equal to the space of polynomials of degree less than or equal to one. We have $N = 2$ and can use $f_1(x) = 1$ and $f_2(x) = x$ as basis for $\mathcal{L}$. Then $L_\theta f_1 = 0$, $L_\theta f_2 = \alpha\beta f_1 - \beta f_2$ so that

$$\Psi_\theta = \begin{pmatrix} 0 & 0 \\ \alpha\beta & -\beta \end{pmatrix}.$$

A straightforward calculation gives

$$e^{t\Psi_\theta} = \begin{pmatrix} 1 & 0 \\ \alpha\left(1 - e^{-t\beta}\right) & e^{-t\beta} \end{pmatrix},$$

and by multiplying from the right with the vector $(1, x)^T$, formula (3.9) is recovered.

The integrability conditions from above may be verified as follows. If X has an invariant density μ_θ, and *all* f_j *are* μ_θ*-integrable*, then since

$$\int_D \mu_\theta(dx)\ \mathrm{E}_\theta\left(|f_j(X_s)|\,|X_0 = x\right) = \mu_\theta\left(|f_j|\right) < \infty$$

it follows (at least for μ_θ-almost all x) that $f_j(X_s)$ is integrable. Similarly, if *all functions*

$$\eta_\ell(x) = \left(\sum_k \partial_{x_k} f_j(x) \sigma_{k\ell}(x)\right)^2 \quad (1 \le \ell \le d)$$

are μ_θ-integrable, it can be verified that for μ_θ-almost all x, M^{f_j} is a true martingale when conditioning on $X_0 = x$.

A particularly nice case of the setup above arises when $\mathcal{L}_\theta = \mathcal{L}$ is the space of polynomials of degree less than or equal to r for some $r \in \mathbb{N}$. Then the invariance, $L_\theta \mathcal{L} \subseteq \mathcal{L}$, holds for *all* r provided each $b_k(x;\theta)$ is a polynomial in x of degree less than or equal to one and each $C_{k\ell}(x;\theta)$ is a polynomial in x of degree less than or equal to two. These conditions are for instance satisfied by the *affine term structure models,* see Duffie & Kan (1996), where the $C_{k\ell}$ are of degree ≤ 1. Thus, with these conditions on b and C satisfied, the conditional moments

$$\pi_t\left(\prod_{k=1}^d x_k^{p_k}\right) = \mathrm{E}\left(\prod_{k=1}^d X_t^{p_k} \middle| X_0 = x\right)$$

with all $p_k \in \mathbb{N}_0$ may be found from (3.22) provided they exist and the relevant local martingales are true martingales.

Note that since $L_\theta \mathbf{1} = 0$, where $\mathbf{1} = (1, \ldots, 1)^T$ the constant functions may always be included in $\mathcal{L}$, and it is not really required that the basis f satisfy the linear relationship (3.20) – it is sufficient that there is a vector $\mathbf{c}$ of constant functions such that $L_\theta f = \mathbf{c} + \Psi_\theta f$.

We now turn to some estimating functions of the form (3.3) that are *not martingale estimating functions*, but can be found in explicit form. Consider first *simple* estimating functions, where the function g appearing in (3.3) is of the form

$$g(\Delta, x, y; \theta) = h(x;\theta)$$

(or $= h(y;\theta)$). We assume in the following that the diffusion X is ergodic with invariant density μ_θ. The unbiasedness property, $\mathrm{E}_\theta(G_n(\theta)) = 0$, is satisfied if $\mu_\theta(h(\theta)) = 0$. Note that here it holds only when X_0 has distribution μ_θ, in contrast to (3.14) where it holds regardless of the distribution of X_0.

A simple example is

$$h(x;\theta) = \partial_\theta \log \mu_\theta(x),$$

which was proposed by Kessler (2000). This estimating function corresponds to assuming that all observations are independent with density μ_θ. The unbiasedness condition is satisfied under usual conditions allowing the interchange of differentiation and integration. An somewhat complex modification of this simple estimating function was shown by Kessler, Schick & Wefelmeyer (2001) to be efficient in the sense of semiparametric models. The semiparametric model for X considered in that paper was that the process is Markovian with only the invariant measures $\{\mu_\theta | \, \theta \in \Theta\}$ specified parametrically. The modified version of the estimating function was derived by Kessler & Sørensen (2002) in a completely different way.

The unbiasedness property holds for all h of the form

$$h(x;\theta) = L_\theta f(x) \tag{3.23}$$

provided each coordinate f_j and $L_\theta f_j$ belong to $L^2(\mu_\theta)$. This is the basic property of the invariant measure expressed in terms of the generator, a fact noted and used to construct estimating functions by Hansen & Scheinkman (1995), see also Kessler (2000), Baddeley (2000), and Aït-Sahalia, Hansen & Scheinkman (2003).

One might consider it a major weakness that (3.23) depends on the argument x only. In fact, Hansen & Scheinkman (1995) proved that only parameters on which the invariant density μ_θ depends can be estimated by (3.23). Hence the importance of the class of *explicit, transition dependent estimating functions* introduced and studied thoroughly by Hansen & Scheinkman (1995), viz. each coordinate g_j is of the form

$$g_{j,\Delta}(x, y; \theta) = h_j(y) L_\theta f_j(x) - f_j(x) \hat{L}_\theta h_j(y). \tag{3.24}$$

Both here and in (3.23) the functions f and h are allowed to depend on θ and Δ – mostly however we think of cases where they do not. The general form of (3.24) requires an explanation: When X_0 has distribution μ_θ, the process X is stationary (for that value of θ), and for any finite $T > 0$, the fragment $(X_{T-t})_{0 \le t \le T}$, has the same distribution as $(\hat{X}_t)_{0 \le t \le T}$ where $\hat{X}$ is another diffusion, stationary with $\hat{X}_0$ having distribution μ_θ. This new diffusion, the *time reversal of* X, has generator

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

where

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

see e.g. Hansen & Scheinkman (1995). It is the dual generator $\hat{L}_\theta$ that appears in (3.24).

We call X *reversible* if $\hat{X}$ and X are the same diffusion, i.e. X is reversible if and only if $\hat{b}(x; \theta) = b(x; \theta)$ for all x. For $d = 1$, X is always reversible – the equation $\hat{b} \equiv b$ when solved for μ_θ immediately gives the expression (2.16). For $d \ge 2$ it is the exception rather than the rule that X be reversible, and that makes (3.24) an explicit estimating function only if μ_θ is known explicitly which, again in contrast to the one-dimensional case, hardly ever happens. Thus in practice, the class (3.24) of estimating functions will be relevant mostly for reversible diffusion models, in particular always when $d = 1$. For reversible models it seems natural to enlarge the class (3.24) by considering g of the form

$$g_{i,\Delta}(x, y; \theta) = \sum_{q=1}^{r} \left[h_{iq}(y) L_\theta f_{iq}(x) - f_{iq}(x) L_\theta h_{iq}(y) \right]. \tag{3.25}$$

Example 3.5 One of the best examples known of the successful use of a simple estimating function is Kessler's estimator of the drift parameter in the one-dimensional Ornstein-Uhlenbeck model

$$dX_t = -\theta X_t \, dt + dW_t$$

where $\theta > 0$ in order to make X ergodic (Kessler (2000)). Kessler uses (3.23) with $f(x) = x^2$, so that $h(x; \theta) = -2\theta x^2 + 1$ resulting in the estimator

$$\hat{\theta} = \frac{n}{2 \sum_{i=1}^{n} X^2_{(i-1)\Delta}},$$

which he shows is the most efficient of all estimators that can be obtained using estimating functions of the form (3.23), and which performs remarkably well with an asymptotic efficiency

relative to the (complicated) maximum-likelihood estimator that is always greater than or equal to 95.6%, no matter what Δ is.

That simple estimating functions can also be very bad, is illustrated by Kessler (2000) using the example

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

where the estimator based on (3.23) with $f(x) = x^2$ behaves terribly for all values of Δ.

□

While (3.23) has often been used with f a polynomial, and thus with the choice of f not related to the model at hand in any particular way, H. Sørensen (2001), using approximations to the score function for continuous time observation of the diffusion X, argued that one should use the model based choice $f = \partial_\theta \log \mu_\theta$. The resulting estimator is *small Δ-optimal,* a concept that is the subject of the Subsection 5.3.

3.4 Non-Markovian Models

An important type of a non-Markovian model that is widely used in finance is the *stochastic volatility model*

$$\begin{aligned} dY_t &= \sqrt{v_t}dW_t \\ dv_t &= b(v_t;\theta)dt + c(v_t;\theta)dB_t, \end{aligned} \tag{3.26}$$

where W and B are independent standard Wiener processes. We assume that v is an ergodic, positive diffusion with invariant probability measure μ_θ, and that $v_0 \sim \mu_\theta$ and is independent of B and W. The process Y is for instance used as a model for the logarithm of the price of a stock. The returns $X_i = Y_{\Delta i} - Y_{\Delta(i-1)}$ are observations from a stationary non-Markovian process. There are more complex stochastic volatility models, but for simplicity we will here only consider the most basic type.

A number of approaches are available for inference about the parameters in stochastic volatility models considered below. One is indirect inference or the efficient method of moments, see Gourieroux, Monfort & Renault (1993), Gallant & Tauchen (1996), and Gallant & Tauchen (2003). Likelihood based methods for stochastic volatility models have been proposed by Kim, Shephard & Chib (1998) and H. Sørensen (2003), and simulation based Bayesian methods using Markov chain Monte Carlo have been developed by Elerian, Chib & Shephard (2001) and Eraker (2001), see also Johannes & Polson (2003). Estimating functions for stochastic volatility models were proposed by Genon-Catalot, Jeantheau & Larédo (1999) and Genon-Catalot, Jeantheau & Larédo (2000). Here we will concentrate on the prediction-based estimating functions introduced by Sørensen (2000) that are widely applicable to non-Markovian diffusion models. An example is the application to observations of integrals of diffusions in disjoint intervals in Ditlevsen & Sørensen (2002).

We will consider the general situation when the model for the data $X_1, \ldots, X_n$ is a stationary non-Markovian process. Here it is in many cases not possible to find a martingale estimating function that can be easily calculated. Obviously,

$$\sum_{i=1}^{n} a(X_1, \ldots, X_{i-1};\theta)\left[f(X_i) - \mathrm{E}_\theta\left(f(X_i)|X_1, \ldots, X_{i-1}\right)\right] \tag{3.27}$$

is a martingale estimating function, but it can usually not be used in practice because of the problems involved in calculating the conditional expectation, analytically as well as numerically. Also for non-Markovian models the score function is usually a martingale, so it must be expected that it is best approximated by a martingale estimating function. When the conditional expectations appearing in the martingales are too complicated for practical use, it therefore seems desirable to approximate them as well as possible by other predictors. This is the idea behind the prediction-based estimating functions.

Assume that f_j, $j = 1, \ldots, N$, are one-dimensional functions, defined on the state space of X, such that $\mathrm{E}_\theta(f_j(X_1)^2) < \infty$ for all $\theta \in \Theta$. We shall now introduce an estimating function with a structure similar to (3.27) where the intractable conditional expectation is replaced by a simpler expression that can be interpreted as an approximation to the conditional expectation. For each j we will predict $f_j(X_i)$ by predictors of the form

$$\pi_j^{(i-1)} = \alpha_{j0} + \alpha_{j1} h_{j1}(X_{i-1}, \ldots, X_{i-s}) + \cdots + \alpha_{jq} h_{jq}(X_{i-1}, \ldots, X_{i-s}), \tag{3.28}$$

where h_{jk}, $k = 1, \ldots q$ are given functions from $\mathbb{R}^s$ into $\mathbb{R}$ satisfying that $\mathrm{E}_\theta(h_{jk}(X_1, \ldots, X_s)^2) < \infty$. Note that the predictor depends only on observations s time steps back in time. This is essential and simplifies the asymptotic theory for the estimators enormously. The minimum mean square error unbiased predictor of $f_j(X_i)$ of the form (3.28) is given by

$$\breve{\pi}_j^{(i-1)}(\theta) = \breve{\alpha}_{j0}(\theta) + \breve{\alpha}_j(\theta)^T Z_j^{(i-1)}, \tag{3.29}$$

where $Z_j^{(i-1)} = (Z_{j1}^{(i-1)}, \ldots, Z_{jq}^{(i-1)})^T$ with $Z_{jk}^{(i-1)} = h_{jk}(X_{i-1}, \ldots, X_{i-s})$, $k = 1, \ldots q$, where $\breve{\alpha}_j(\theta) = (\breve{\alpha}_{j1}(\theta), \ldots, \breve{\alpha}_{jq}(\theta))^T$ equals

$$\breve{\alpha}_j(\theta) = C_j(\theta)^{-1} b_j(\theta), \tag{3.30}$$

and where

$$\breve{\alpha}_{j0}(\theta) = E_\theta\left(f_j(X_1)\right) - \breve{\alpha}_j(\theta)^T E_\theta\left(Z_j^{(s)}\right). \tag{3.31}$$

As earlier, T denotes transposition of vectors and matrices. In formula (3.30), $C_j(\theta)$ denotes the covariance matrix of $Z_j^{(s)}$, while

$$b_j(\theta) = \left(\mathrm{Cov}_\theta\left(Z_{j1}^{(s)}, f_j(X_{s+1})\right), \ldots, \mathrm{Cov}_\theta\left(Z_{jq}^{(s)}, f_j(X_{s+1})\right)\right)^T. \tag{3.32}$$

Thus a prediction-based estimating function can be calculated provided only that we can find the covariances in $C_j(\theta)$ and $b_j(\theta)$. When these moments cannot be determined explicitly, they are usually easy to obtain numerically. A simple and natural choice of $Z_j^{(i-1)}$ is $Z_j^{(i-1)} = (f_j(X_{i-1}), \ldots, f_j(X_{i-q}))^T$. In this case, the coefficients $\breve{\alpha}_{j0}, \ldots, \breve{\alpha}_{jq}$ can easily be found by means of the Durbin-Levinson algorithm, see Brockwell & Davis (1991).

The minimum mean square error unbiased predictor of $f_j(X_i)$ is the projection in the L_2-space of functions of $X_i, X_{i-1}, \ldots, X_{i-s}$ with finite variance onto the linear subspace of functions on the form (3.28). Therefore $\breve{\pi}_j^{(i-1)}(\theta)$ satisfies the normal equation

$$\mathrm{E}_\theta\left(\pi_j^{(i-1)}\left\{f_j(X_i) - \breve{\pi}_j^{(i-1)}(\theta)\right\}\right) = 0 \tag{3.33}$$

for all $\pi_j^{(i-1)}$ of the form (3.28). This implies (3.30). The fact that $\breve{\pi}_j^{(i-1)}(\theta)$ is a projection also shows that it can be interpreted as an approximation to the conditional expectation of $f_j(X_i)$

given $X_1, \ldots, X_{i-1}$, because this conditional expectation is the projection of $f_j(X_i)$ onto the linear space of all functions of $X_1, \ldots, X_{i-1}$ with finite variance.

It follows from (3.31) that the estimating function

$$\sum_{i=s+1}^{n} \sum_{j=1}^{N} \left\{ f_j(X_i) - \breve{\pi}_j^{(i-1)}(\theta) \right\}$$

is unbiased, so that we can expect it to produce consistent estimators, cf. Theorem 2.2. The normal equations (3.33) indicate, how we can choose weights that can improve the efficiency of the estimators. In fact,

$$G_n(\theta) = \sum_{i=s+1}^{n} \sum_{j=1}^{N} \Pi_j^{(i-1)}(\theta) \left\{ f_j(X_i) - \breve{\pi}_j^{(i-1)}(\theta) \right\}, \tag{3.34}$$

where $\Pi_j^{(i-1)}(\theta)$ is a p-dimensional vector, is an unbiased estimating function whenever $\Pi_j^{(i-1)}(\theta) = (\pi_{1,j}^{(i-1)}(\theta), \ldots, \pi_{p,j}^{(i-1)}(\theta))^T$, where the coordinates are of the form (3.28). We shall find the optimal choice for $\Pi_j^{(i-1)}(\theta)$ in Subsection 5.4. An estimating function of the type (3.34) is called a *prediction-based estimating function.*

Note the computational advantage of prediction-based estimating functions in comparison to martingale estimating functions when these are not explicit. Here we need only unconditional moments that are relatively easy to compute by simulation, whereas for martingale estimating functions moments conditional on all data points are needed which involve much more computation. The conventional wisdom is that estimators based on conditional moments are more efficient than estimators based on unconditional moments. The argument is that with conditional moments one can construct a martingale estimating function that is a close approximation to the score function, which as we saw in Subsection 3.1 is itself a martingale, and thus obtain a highly efficient estimator. However, by choosing the functions h_{jk} suitably, a good approximation can be obtained to the conditional expectation, so there is reason to believe that the estimators presented in this section can have a high efficiency too.

Note also that since $\breve{\pi}_j^{(i-1)}(\theta)$ depends exclusively on the first and second order moments of the random vector $\left(f_j(X_i), Z_{j1}^{(i-1)}, \ldots, Z_{jq}^{(i-1)}\right)$, only parameters appearing in these moments for at least one j can be estimated using (3.34). This is intuitively obvious and indeed follows from Condition 3.7 given later in this section. Of course, one would usually choose the functions f_j and h_{jk} in such a way that it is possible to estimate all parameters of interest.

A non-optimal prediction-based estimating function is obtained by differentiating the logarithm of the pseudo-likelihood function obtained by pretending that the process $\{f(X_i)\}$ is Gaussian with the correct first and second order moments and multiplying the Gaussian conditional densities of $f(X_i)$ given $(f(X_{i-1}), \ldots, f(X_{i-q}))$ for $i = q+1, \ldots, n$. In this particular case, $N = 1$, $q = s$, and $h_k(x_q, \ldots, x_1) = f(x_{q+1-k})$, $k = 1, \ldots, q$.

Example 3.6 For the stochastic volatility model (3.26), the returns $X_i = Y_{\Delta i} - Y_{\Delta(i-1)}$ satisfy that

$$X_i = \int_{(i-1)\Delta}^{i\Delta} \sqrt{v_t} dW_t, \tag{3.35}$$

so that

$$X_i = \sqrt{S_i} Z_i, \tag{3.36}$$

where

$$S_i = \int_{(i-1)\Delta}^{i\Delta} v_t dt \tag{3.37}$$

and where the Z_is are independent, identically standard normal distributed random variables, that are independent of $\{S_i\}$. It is easy to see from (3.36) that the returns are uncorrelated in accordance with empirical findings. Therefore the function $f(x) = x$ cannot be used to define a prediction based estimating function. A simple alternative is to base the estimating function on the squared returns X_i^2. This can be done by using the estimating function with $N = 1$, $f(x) = x^2$, and $h_k(x_q, \ldots, x_1) = x_{q-k+1}^2$, $k = 1, \ldots, q$. Here $s = q$. In this way we obtain an estimating function of the form

$$G_n(\theta) = \sum_{i=q+1}^{n} \Pi^{(i-1)}(\theta) \left\{ X_i^2 - \breve{\alpha}_0(\theta) - \breve{\alpha}_1(\theta) X_{i-1}^2 - \cdots - \breve{\alpha}_q(\theta) X_{i-q}^2 \right\}, \tag{3.38}$$

where the quantities $\breve{\alpha}_k(\theta)$, $k = 0, \ldots, q$, are given by

$$\breve{\alpha}_0(\theta) = \mathrm{E}_\theta\left(X_1^2\right) \{1 - (\breve{\alpha}_1(\theta) + \cdots + \breve{\alpha}_q(\theta))\}$$

and (3.30) with $C(\theta)$ equal to the covariance matrix of the stochastic vector $\left(X_q^2, \ldots, X_1^2\right)$, and with $b(\theta) = \left(\mathrm{Cov}_\theta\left(X_{q+1}^2, X_q^2\right), \ldots, \mathrm{Cov}_\theta\left(X_{q+1}^2, X_1^2\right)\right)^T$.

In order to ensure that the quantities $C(\theta)$ and $b(\theta)$ are well-defined, we must assume that $\mathrm{E}_\theta(X_1^4) < \infty$. This is the case provided that the second moment of the volatility process v exists. Let us briefly discuss how to calculate the covariances. It follows from (3.36) that

$$\begin{aligned} \mathrm{E}_\theta(X_i^2) &= \mathrm{E}_\theta(S_1) \\ \mathrm{Var}_\theta\left(X_i^2\right) &= 3\mathrm{Var}_\theta\left(S_1\right) + 2\mathrm{E}_\theta(S_1)^2 \\ \mathrm{Cov}_\theta\left(X_i^2, X_{i+j}^2\right) &= \mathrm{Cov}_\theta\left(S_1, S_{1+j}\right). \end{aligned}$$

Define

$$\begin{aligned} \xi(\theta) &= E_\theta(v_t) \\ \omega(\theta) &= \mathrm{Var}_\theta(v_t) \\ r(u;\theta) &= \mathrm{Cov}_\theta(v_t, v_{t+u})/\omega(\theta). \end{aligned}$$

Using (3.37), it is not difficult to see that

$$\mathrm{E}_\theta\left(X_n^2\right) = \Delta\xi(\theta) \tag{3.39}$$

$$\mathrm{Var}_\theta\left(X_n^2\right) = 6\omega(\theta)R^*(\Delta;\theta) + 2\Delta^2\xi(\theta)^2 \tag{3.40}$$

$$\begin{aligned} \mathrm{Cov}_\theta\left(X_n^2, X_{n+i}^2\right) &= \omega(\theta)\left[R^*(\Delta(i+1);\theta)\right. \\ &\quad \left. - 2R^*(\Delta i;\theta) + R^*(\Delta(i-1);\theta)\right], \end{aligned} \tag{3.41}$$

where

$$R^*(t;\theta) = \int_0^t \int_0^s r(u;\theta) du ds;$$

see Barndorff-Nielsen & Shephard (2001). For numerical calculations it is perhaps more useful that

$$\mathrm{Cov}_\theta\left(X_n^2, X_{n+i}^2\right) = \omega(\theta) \int_{(i-1)\Delta}^{i\Delta} \int_s^{s+\Delta} r(u;\theta) du ds, \tag{3.42}$$

which follows by easy calculations. Thus all that is needed to compute the minimal mean squared error predictor in (3.38) are the first and second order moments of the volatility process. Examples of models where these moments can be found explicitly are given in Example 5.11.

□

We finish by discussing the asymptotic properties of an estimator obtained from an estimating function of the general type (3.34), which we will first give a more compact form. Write the ℓth coordinate of the vector $\Pi_j^{(i-1)}(\theta)$ as

$$\pi_{\ell,j}^{(i-1)}(\theta) = \sum_{k=0}^{q} \alpha_{\ell jk}(\theta) Z_{jk}^{(i-1)},$$

with $Z_{j0}^{(i-1)} = 1$. Then (3.34) can be written in the form

$$G_n(\theta) = A(\theta) \sum_{i=s+1}^{n} Z^{(i-1)} \left(F(X_i) - \breve{\pi}^{(i-1)}(\theta) \right), \tag{3.43}$$

where

$$A(\theta) = \begin{pmatrix} \alpha_{110}(\theta) & \cdots & \alpha_{11q}(\theta) & \cdots & \cdots & \alpha_{1N0}(\theta) & \cdots & \alpha_{1Nq}(\theta) \\ \vdots & & \vdots & & & \vdots & & \vdots \\ \alpha_{p10}(\theta) & \cdots & \alpha_{p1q}(\theta) & \cdots & \cdots & \alpha_{pN0}(\theta) & \cdots & \alpha_{pNq}(\theta) \end{pmatrix}, \tag{3.44}$$

where $F(x) = (f_1(x), \ldots, f_N(x))^T$, $\breve{\pi}^{(i-1)}(\theta) = (\breve{\pi}_1^{(i-1)}(\theta), \ldots, \breve{\pi}_N^{(i-1)}(\theta))^T$, and where $Z^{(i-1)}$ is the $N(q+1) \times N$-matrix given by

$$Z^{(i-1)} = \begin{pmatrix} Z_{10}^{(i-1)} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ Z_{1q}^{(i-1)} & 0 & \cdots & 0 \\ 0 & Z_{20}^{(i-1)} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & Z_{2q}^{(i-1)} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & Z_{N0}^{(i-1)} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & Z_{Nq}^{(i-1)} \end{pmatrix}. \tag{3.45}$$

The following condition will imply that the conditions of Theorem 2.2 are satisfied and will thus ensure the existence of a $\sqrt{n}$-consistent and asymptotically normal estimator. In this condition we need two further definitions:

$$\breve{\alpha}(\theta) = (\breve{\alpha}_{10}(\theta), \breve{\alpha}_{11}(\theta), \ldots, \breve{\alpha}_{1q}(\theta), \ldots, \breve{\alpha}_{N0}(\theta), \ldots \breve{\alpha}_{Nq}(\theta))^T, \tag{3.46}$$

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

$$D(\theta) = \mathrm{E}_\theta \left(Z^{(i-1)} (Z^{(i-1)})^T \right). \tag{3.47}$$

Finally let θ_0 denote the true parameter value.

Condition 3.7
(1) *The process X is stationary and geometrically α-mixing.*

(2) $E_{\theta_0}\left(\left|Z_{jk}^{(s)} f_j(X_{s+1})\right|^{2+\delta}\right) < \infty$ *and* $E_{\theta_0}\left(\left|Z_{jk}^{(s)} Z_{j\ell}^{(s)}\right|^{2+\delta}\right) < \infty$, $j = 1, \ldots, N$, $k, \ell = 0, \ldots q$.

(3) *The vector $\breve{a}(\theta)$ given by* (3.30), (3.31) *and* (3.46) *and the matrix $A(\theta)$ are twice continuously differentiable with respect to θ.*

(4) *The matrix $A(\theta_0)D(\theta_0)\partial_{\theta^T}\breve{a}(\theta_0)$ has rank p. The matrix $D(\theta_0)$ is given by* (3.47).

Under Condition 3.7 (1)–(2) the process $Z^{(i-1)}\left(F(X_i) - \breve{\pi}^{(i-1)}(\theta)\right)$, $i = s+1, s+2, \ldots$ is geometrically α-mixing, and satisfies the moment condition of the central limit theorem for such sequences, see e.g. Doukhan (1994), Theorem 1 in his Section 1.5. Therefore, the law of large numbers (2.3) and the following hold. The matrix

$$\bar{M}_n(\theta) = \mathrm{E}_\theta\left(H^{(s+1)}(\theta)H^{(s+1)}(\theta)^T\right) + \sum_{k=1}^{n-s-1} \frac{(n-s-k)}{(n-s)} \left\{\mathrm{E}_\theta\left(H^{(s+1)}(\theta)H^{(s+1+k)}(\theta)^T\right) + \mathrm{E}_\theta\left(H^{(s+1+k)}(\theta)H^{(s+1)}(\theta)^T\right)\right\} \tag{3.48}$$

with

$$H^{(i)}(\theta) = Z^{(i-1)}\left(F(X_i) - \breve{\pi}^{(i-1)}(\theta)\right).$$

satisfies that

$$\bar{M}_n(\theta_0) \to M(\theta_0),$$

and the central limit theorem

$$\frac{1}{\sqrt{n}} G_n(\theta_0) \xrightarrow{\mathcal{D}} N\left(0, A(\theta_0)M(\theta_0)A(\theta_0)^T\right)$$

holds as $n \to \infty$, provided that $M(\theta_0)$ is strictly positive definite. The matrix $\bar{M}_n(\theta)$ is the covariance matrix of $\sum_{i=s+1}^n H^{(i)}(\theta)/\sqrt{n-s}$. The rest of Condition 3.7 implies Condition 2.1. Note in particular, that

$$S(\theta_0) = -\mathrm{E}_{\theta_0}(A(\theta_0)Z^{(i-1)}(Z^{(i-1)})^T\partial_{\theta^T}\breve{\alpha}(\theta_0)) = -A(\theta_0)D(\theta_0)\partial_{\theta^T}\breve{\alpha}(\theta_0).$$

Under Condition 3.7 and the condition on $M(\theta_0)$ we therefore have the conclusions of Theorem 2.2, i.e. with a probability tending to one as $n \to \infty$ the estimating equation $G_n(\theta) = 0$ defines a $\sqrt{n}$-consistent estimator $\hat{\theta}_n$ satisfying that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N\left(0, S(\theta_0)^{-1}A(\theta_0)M(\theta_0)A(\theta_0)^T(S(\theta_0)^{-1})^T\right). \tag{3.49}$$

Example 3.6 (continued). For the stochastic volatility model (3.26), to ensure (1) in Condition 3.7 it is enough that the volatility process v is geometrically α-mixing, see Lemma 6.3 in Sørensen (2000) or Genon-Catalot, Jeantheau & Larédo (2000). Condition 2.5 together with the condition for ergodicity discussed in Section 2.3 are sufficient to ensure that a diffusion process like the volatility process is geometrically α-mixing.

□

4 The General Theory of Optimal Estimating Functions

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 closely related generalized estimating equations developed by Liang & Zeger (1986) to deal with problems of longitudinal data analysis, see also Prentice (1988) and Li (1997).

It is interesting that a related parallel development in the theory of the generalized method of moments has taken place independently in the econometric literature, see e.g. Hansen (1982) and Hansen (1985). There the emphasis has been on bounds on the asymptotic covariance matrix of estimators, rather than on criteria for optimality of estimating functions and methods to construct optimal estimating functions.

4.1 General Estimating Functions

In this subsection we first review the general theory of estimating functions for stochastic process models. A modern review of the theory is given in Heyde (1997).

Here we consider again a general model. The data $X_1, X_2, \ldots, X_n$ are assumed to be observations from a stochastic process model indexed by a p-dimensional parameter $\theta \in \Theta$. 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

$$S_{G_n}(\theta) = E_\theta(\partial_\theta G_n(\theta)) \tag{4.1}$$

is called the *sensitivity* function for G_n. Here ∂_θ denotes the partial derivative with respect to θ. 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 $\mathrm{E}_\theta(G_n(\theta))$ 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 as our estimating function the score function $U_n(\theta) = \partial_\theta \log L_n(\theta)$. 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 variance. Then it can be proven under usual regularity conditions allowing the interchange of integration and differentiation 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

$$K_{G_n}(\theta) = S_{G_n}(\theta)^2 / \mathrm{Var}_\theta(G_n(\theta)) = S_{G_n}(\theta)^2 / \mathrm{E}_\theta(G_n(\theta)^2), \tag{4.2}$$

which is known as the *Godambe information*. This makes intuitive sense: $K_{G_n}(\theta)$ 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 *F-optimal* or *Godambe-optimal* in $\mathcal{G}_n$ if

$$K_{G_n^*}(\theta) \geq K_{G_n}(\theta) \tag{4.3}$$

for all $\theta \in \Theta$ and for all $G_n \in \mathcal{G}_n$. The optimal estimating function G_n^* is sometimes called a *quasi score function*, while an estimator θ_n^* obtained by solving the equation $G_n^*(\theta) = 0$ is called a *quasi likelihood estimator*.

When the parameter θ is multivariate ($p > 1$), the sensitivity function is the $p \times p$-matrix

$$S_{G_n}(\theta) = E_\theta(\partial_{\theta^T} G_n(\theta)) = \begin{pmatrix} \partial_{\theta_1} G_n(\theta)_1 & \cdots & \partial_{\theta_p} G_n(\theta)_1 \\ \vdots & & \vdots \\ \partial_{\theta_1} G_n(\theta)_p & \cdots & \partial_{\theta_p} G_n(\theta)_p \end{pmatrix}. \tag{4.4}$$

We denote the transpose of a vector or a matrix a by a^T. Vectors are column vectors. For a multivariate parameter, the Godambe information is the $p \times p$-matrix

$$K_{G_n}(\theta) = S_{G_n}(\theta)^T \mathrm{E}_\theta \left(G_n(\theta) G_n(\theta)^T \right)^{-1} S_{G_n}(\theta), \tag{4.5}$$

and an optimal estimating function (or a quasi score function) G_n^* can be defined by (4.3) with the inequality referring to the partial ordering of the set of positive semi-definite $p \times p$-matrices. Whether an F-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 (4.3), 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). For theoretical purposes a standardized version of the estimating functions is useful. The standardized version of $G_n(\theta)$ is given by

$$G_n^{(s)}(\theta) = -S_{G_n}(\theta)^T \mathrm{E}_\theta \left(G_n(\theta) G_n(\theta)^T \right)^{-1} G_n(\theta).$$

The rationale behind this standardization is that $G_n^{(s)}(\theta)$ satisfies the *second Bartlett-identity*

$$\mathrm{E}_\theta \left(G_n^{(s)}(\theta) G_n^{(s)}(\theta)^T \right) = -\mathrm{E}_\theta(\partial_{\theta^T} G_n^{(s)}(\theta)), \tag{4.6}$$

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 (4.4).

An F-optimal estimating function is close to the score function U_n in an L_2-sense. Suppose G_n^* is F-optimal in $\mathcal{G}_n$. Then the standardized version $G_n^{*(s)}(\theta)$ satisfies the inequality

$$E_\theta \left((G_n^{(s)}(\theta) - U_n(\theta))^T (G_n^{(s)}(\theta) - U_n(\theta)) \right) \geq E_\theta \left((G_n^{*(s)}(\theta) - U_n(\theta))^T (G_n^{*(s)}(\theta) - U_n(\theta)) \right)$$

for all $\theta \in \Theta$ and for all $G_n \in \mathcal{G}_n$, see Heyde (1997). In fact, if $\mathcal{G}_n$ is a closed subspace of the L_2-space of all square integrable random vectors, 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 can often be used to find the optimal estimating function.

Theorem 4.1 (Heyde (1988)) *If $G_n^* \in \mathcal{G}_n$ satisfies*

$$S_{G_n}(\theta)^{-1} E_\theta \left(G_n(\theta) G_n^*(\theta)^T \right) = S_{G_n^*}(\theta)^{-1} E_\theta \left(G_n^*(\theta) G_n^*(\theta)^T \right) \tag{4.7}$$

for all $\theta \in \Theta$ and for all $G_n \in \mathcal{G}_n$, then it is F-optimal in $\mathcal{G}_n$. When $\mathcal{G}_n$ is closed under addition, any F-optimal estimating function G_n^* satisfies (4.7).

In many situations the condition (4.7) can 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*, (4.6), so that

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

Example 4.2 Almost all estimating functions considered in this chapter have the following form. Suppose that one has a number of functions $h_{ij}(x_1, \ldots, x_i; \theta)$, $j = 1, \ldots, N$, $i = 1, \ldots n$ satisfying that

$$\mathrm{E}_\theta(h_{ij}(X_1, \ldots, X_i; \theta)) = 0.$$

Such functions define relationships (dependent on θ) between an observation X_i and the previous observations $X_1, \ldots, X_{i-1}$ (or some of them) that are on average equal to zero. It is natural to use such relationships to estimate θ by solving the equations $\sum_{i=1}^n h_{ij}(X_1, \ldots, X_i; \theta) = 0$. In order to estimate θ it is necessary that $N \geq p$, but if $N > p$ we have too many equations. The theory of optimal estimating functions tells us how to combine the N relations in an optimal way.

Let h_i denote the N-dimensional vector $(h_{i1}, \ldots, h_{iN})^T$, and define an N-dimensional estimating function by $H_n(\theta) = \sum_{i=1}^n h_i(X_1, \ldots, X_i; \theta)$. First we consider the class of p-dimensional estimating functions of the form

$$G_n(\theta) = A_n(\theta) H_n(\theta),$$

where $A_n(\theta)$ is a non-random $p \times N$-matrix that is differentiable with respect to θ. By $A_n^*(\theta)$ we denote the optimal choice of $A_n(\theta)$. It is not difficult to see that

$$S_{G_n}(\theta) = A_n(\theta) S_{H_n}(\theta)$$

and

$$\mathrm{E}_\theta \left(G_n(\theta) G_n^*(\theta)^T \right) = A_n(\theta) \mathrm{E}_\theta \left(H_n(\theta) H_n(\theta)^T \right) A_n^*(\theta)^T,$$

where $S_{H_n}(\theta) = E_\theta(\partial_{\theta^T} H_n(\theta))$. If we choose

$$A_n^*(\theta) = -S_{H_n}(\theta)^T \mathrm{E}_\theta \left(H_n(\theta) H_n(\theta)^T \right)^{-1},$$

then (4.7) is satisfied for all $G_n \in \mathcal{G}_n$, so that $G_n^*(\theta) = A_n^*(\theta)H_n(\theta)$ is F-optimal.

Sometimes there are good reasons to use functions h_{ij} satisfying that

$$\mathrm{E}_\theta(h_{ij}(X_1, \ldots, X_i; \theta)h_{i'j'}(X_1, \ldots, X_{i'}; \theta)) = 0 \tag{4.8}$$

for all $j, j' = 1, \ldots, N$ when $i \neq i'$. For such functions the random variables $h_{ij}(X_1, \ldots, X_i; \theta)$, $i = 1, 2, \ldots$ are uncorrelated, and in this sense the "new" random variation of $h_{ij}(X_1, \ldots, X_i; \theta)$ depends only on the innovation in the ith observation. This is for instance the case for martingale estimating functions, see (4.14). In this situation it is natural to consider the larger class of estimating functions given by

$$G_n(\theta) = \sum_{i=1}^{n} a_i(\theta)h_i(X_1, \ldots, X_i; \theta),$$

where $a_i(\theta)$, $i = 1, \ldots n$, are $p \times N$ matrices that do not depend on the data and are differentiable with respect to θ. Here

$$\begin{aligned} S_{G_n}(\theta) &= \sum_{i=1}^{n} a_i(\theta)\mathrm{E}_\theta(\partial_{\theta^T} h_i(X_1, \ldots, X_i; \theta)) \\ \mathrm{E}_\theta\left(G_n(\theta)G_n^*(\theta)^T\right) &= \sum_{i=1}^{n} a_i(\theta)\mathrm{E}_\theta\left(h_i(X_1, \ldots, X_i; \theta)h_i(X_1, \ldots, X_i; \theta)^T\right) a_i^*(\theta)^T, \end{aligned}$$

where $a_i^*(\theta)$ denotes the optimal choice of $a_i(\theta)$. We see that with

$$a_i^*(\theta) = -\mathrm{E}_\theta(\partial_{\theta^T} h_i(X_1, \ldots, X_i; \theta))^T \left(\mathrm{E}_\theta\left(h_i(X_1, \ldots, X_i; \theta)h_i(X_1, \ldots, X_i; \theta)^T\right)\right)^{-1}$$

the condition (4.7) is satisfied. □

4.2 Martingale Estimating Functions

In Subsections 2.2 and 2.3 we saw that martingale estimating functions have a particularly simple asymptotic theory. The martingale limit theory also allows a lucid theory asymptotic optimality. The optimality criterion discussed in the following is particular to martingale estimating functions.

Suppose the estimating function $G_n(\theta)$ satisfies the conditions of the central limit theorem for martingales (Theorem 2.3), and let $\hat{\theta}_n$ be a solution of the equation $G_n(\theta) = 0$. Under usual regularity conditions and using standard manipulations (including a Taylor expansion around the true parameter value θ_0), it can be proved that

$$\langle G(\theta)\rangle_n^{-\frac{1}{2}} \bar{G}_n(\theta)(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, I_p). \tag{4.9}$$

Here $\partial_\theta G_n(\theta)$ has been replaced by a so-called predictable version (called its compensator)

$$\bar{G}_n(\theta) = \sum_{i=1}^{n} \mathrm{E}_\theta\left(\partial_\theta H_i(\theta)|\mathcal{F}_{i-1}\right).$$

For details see Heyde (1997). We see that the inverse of the random matrix

$$I_{G_n}(\theta) = \bar{G}_n(\theta)^T \langle G(\theta)\rangle_n^{-1} \bar{G}_n(\theta) \tag{4.10}$$

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, the *Heyde-information*. It is also called the martingale information, and 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,

$$\mathrm{E}_\theta\left(\bar{G}_n(\theta)\right) = S_{G_n}(\theta) \quad \text{and} \quad \mathrm{E}_\theta\left(\langle G(\theta)\rangle_n\right) = \mathrm{E}_\theta\left(G_n(\theta)G_n(\theta)^T\right).$$

We can thus think of the Heyde-information as a stochastic or estimated version of the Godambe information.

Let $\mathcal{G}_n$ be a class of martingale estimating functions with variance. We say that a martingale estimating function G_n^* is *A-optimal* or *Heyde-optimal* in $\mathcal{G}_n$ if

$$I_{G_n^*}(\theta) \geq I_{G_n}(\theta) \tag{4.11}$$

P_θ-almost surely for all $\theta \in \Theta$, for all $G_n \in \mathcal{G}_n$, and for all $n \in \mathbb{N}$. As was the case for F-optimality, an A-optimal estimating function G_n^* is sometimes called a *quasi score function*, while an estimator θ_n^* obtained by solving the equation $G_n^*(\theta) = 0$ is called a *quasi likelihood estimator*.

The following useful result is similar to Theorem 4.1. 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:

$$\langle G, \tilde{G}\rangle_n = \sum_{i=1}^{n} \mathrm{E}\left(H_i \tilde{H}_i^T | \mathcal{F}_{i-1}\right), \tag{4.12}$$

where $H_i = G_i - G_{i-1}$ and $\tilde{H}_i = \tilde{G}_i - \tilde{G}_{i-1}$.

Theorem 4.3 (Heyde (1988)). *If $G_n^* \in \mathcal{G}_n$ satisfies*

$$\bar{G}_n(\theta)^{-1}\langle G(\theta), G^*(\theta)\rangle_n = \bar{G}_n^*(\theta)^{-1}\langle G^*(\theta)\rangle_n \tag{4.13}$$

for all $\theta \in \Theta$, $G_n \in \mathcal{G}_n$, and $n \in \mathbb{N}$, then it is is A-optimal in $\mathcal{G}_n$. When $\mathcal{G}_n$ is closed under addition, any A-optimal estimating function G_n^* satisfies (4.13). *Moreover, if $\bar{G}_n^*(\theta)^{-1}\langle G^*(\theta)\rangle_n$ is non-random, then G_n^* is also F-optimal in $\mathcal{G}_n$.*

Since in many situations condition (4.13) 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 A-optimality implies F-optimality.

Example 4.4 Let us again consider the situation in Example 4.2, where a number of functions $h_{ij}(x_1, \ldots, x_i; \theta)$, $j = 1, \ldots, N$, $i = 1, \ldots n$ define relationships (dependent on θ) between an observation X_i and the previous observations $X_1, \ldots, X_{i-1}$ (or a subset of them) that can be used to estimate θ. If the functions h_{ij} satisfy that

$$\mathrm{E}_\theta(h_{ij}(X_1, \ldots, X_i; \theta)|\mathcal{F}_{i-1}) = 0$$

for $j = 1, \ldots, N$, $i = 1, \ldots n$, then

$$G_n(\theta) = \sum_{i=1}^{n} a_i(X_1, \ldots, X_{i-1}; \theta) h_i(X_1, \ldots, X_i; \theta), \tag{4.14}$$

is a p-dimensional unbiased martingale estimating function. Here h_i denotes the N-dimensional vector $(h_{i1}, \ldots, h_{iN})^T$, and $a_i(x_1, \ldots, x_{i-1}; \theta)$ is a function from $\mathbb{R}^{i-1} \times \Theta$ into the set of $p \times N$-matrices that is differentiable with respect to θ. We will now find the matrices a_i that combine the N relations in an optimal way. Let $\mathcal{G}_n$ be the class of martingale estimating functions of the form (4.14) that have finite variance. Note that the functions h_{ij} satisfy condition (4.8) in Example 4.2. However, here we consider more general weights a_i that can depend on past observations, and thus we obtain a much larger class of estimating functions than the one considered in Example 4.2.

Since

$$\bar{G}_n(\theta) = \sum_{i=1}^{n} a_i(X_1, \ldots, X_{i-1}; \theta) \mathrm{E}_\theta(\partial_{\theta^T} h_i(X_1, \ldots, X_i; \theta) | \mathcal{F}_{i-1})$$

and

$$\langle G(\theta), G^*(\theta) \rangle_n = \sum_{i=1}^{n} a_i(X_1, \ldots, X_{i-1}; \theta) V_{h_i}(X_1, \ldots, X_{i-1}; \theta) a_i^*(X_1, \ldots, X_{i-1}; \theta)^T,$$

where

$$G_n^*(\theta) = \sum_{i=1}^{n} a_i^*(X_1, \ldots, X_{i-1}; \theta) h_i(X_1, \ldots, X_i; \theta), \tag{4.15}$$

and where

$$V_{h_i}(X_1, \ldots, X_{i-1}; \theta) = \mathrm{E}_\theta \left(h_i(X_1, \ldots, X_i; \theta) h_i(X_1, \ldots, X_i; \theta)^T | \mathcal{F}_{i-1} \right)$$

is the conditional covariance matrix of the random vector $h_i(X_1, \ldots, X_i; \theta)$ given $(X_1, \ldots, X_{i-1})$, we see that with

$$a_i^*(X_1, \ldots, X_{i-1}; \theta) = -\mathrm{E}_\theta(\partial_{\theta^T} h_i(X_1, \ldots, X_i; \theta) | \mathcal{F}_{i-1})^T V_{h_i}(X_1, \ldots, X_{i-1}; \theta)^{-1}, \tag{4.16}$$

the condition (4.13) is satisfied. Hence by Theorem 4.3 the estimating function $G_n^*(\theta)$ is Heyde-optimal. Since $\bar{G}_n^*(\theta)^{-1} \langle G^*(\theta) \rangle_n = -I_p$, the estimating function $G_n^*(\theta)$ is also Godambe-optimal.

Let $p_i(x; \theta | x_1, \ldots, x_{i-1})$ denote the conditional density of X_i given that $(X_1, \ldots, X_{i-1}) = (x_1, \ldots, x_{i-1})$. Then the likelihood function for θ based on the data $(X_1, \ldots, X_n)$ is

$$L_n(\theta) = \prod_{i=1}^{n} p_i(X_i; \theta | X_1, \ldots, X_{i-1})$$

(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

$$U_n(\theta) = \sum_{i=1}^{n} \partial_\theta \log p_i(X_i; \theta | X_1, \ldots, X_{i-1}). \tag{4.17}$$

We shall now see, in exactly what sense the optimal estimating function (4.15) approximates the score function. Let us fix i, $x_1, \ldots, x_{i-1}$ and θ. We let $\mathbf{x}_{i-1}$ denote the vector $(x_1, \ldots, x_{i-1})$ and consider the L_2-space $\mathcal{K}_i(\mathbf{x}_{i-1}, \theta)$ of functions $f : \mathbb{R} \mapsto \mathbb{R}$ for which

$$\int f(x)^2 p_i(x; \theta|\mathbf{x}_{i-1}) dx < \infty.$$

We equip $\mathcal{K}_i(\mathbf{x}_{i-1}, \theta)$ with the usual inner product

$$\langle f, g \rangle = \int f(x) g(x) p_i(x; \theta|\mathbf{x}_{i-1}) dx,$$

and let $\mathcal{H}_i(\mathbf{x}_{i-1}, \theta)$ denote the N-dimensional subspace of $\mathcal{K}_i(\mathbf{x}_{i-1}, \theta)$ spanned by the functions $x \mapsto h_{ij}(\mathbf{x}_{i-1}, x; \theta)$, $j = 1, \ldots, N$. That the functions are linearly independent in $\mathcal{K}_i(\mathbf{x}_{i-1}, \theta)$ follows from the earlier assumption that the covariance matrix $V_{h_i}(\mathbf{x}_{i-1}; \theta)$ is regular.

Now, assume that $\partial_{\theta_j} \log p_i(x|\mathbf{x}_{i-1}; \theta) \in \mathcal{K}_i(\mathbf{x}_{i-1}, \theta)$ for $j = 1, \ldots, p$, let g^*_{ij} denote the orthogonal projection with respect to $\langle \cdot, \cdot \rangle$ of $\partial_{\theta_j} \log p_i$ onto $\mathcal{H}_i(\mathbf{x}_{i-1}, \theta)$, and define a p-dimensional function by $g^*_i = (g_{i1}, \ldots, g_{ip})^T$. Then

$$g^*_i(\mathbf{x}_{i-1}, x; \theta) = a^*_i(\mathbf{x}_{i-1}; \theta) h_i(\mathbf{x}_{i-1}, x; \theta), \tag{4.18}$$

where a^*_i is the matrix defined by (4.16). To see this, note that g^* must have the form (4.18) 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 also 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 $\langle \partial_{\theta_j} \log p_i, \, h_{ik} \rangle$. If we can interchange differentiation and integration so that

$$\int \partial_{\theta_j} \left[h_{ik}(\mathbf{x}_{i-1}, x; \theta) p(\mathbf{x}_{i-1}, x; \theta) \right] dx = \partial_{\theta_j} \int h_{ik}(\mathbf{x}_{i-1}, x; \theta) p(\mathbf{x}_{i-1}, x; \theta) dx = 0,$$

it follows that

$$B_i = - \int \partial_{\theta^T} h_i(\mathbf{x}_{i-1}, x; \theta) p(\mathbf{x}_{i-1}, x; \theta) dx,$$

which proves (4.18).

The result (4.18) was first shown by Kessler (1996) in the case of a Markov process. The proof in the general case is essentially the same as that for a Markov process. It is important to note that if for all i the functions h_{ij} are chosen such that as $N \to \infty$ the subspace $\mathcal{H}_i(\mathbf{x}_{i-1}, \theta)$ converges to a subspace of $\mathcal{K}_i(\mathbf{x}_{i-1}, \theta)$ containing the functions $\partial_{\theta_j} \log p_i$, $j = 1, \ldots, p$, then the optimal estimating function will approach the score function, and it is possible to obtain a sequence of quasi-likelihood estimators that is asymptotically fully efficient.

□

Example 4.5 We now consider a simple example where the stochastic process $\{X_t\}$ is a Markov process, but not a standard diffusion process. The process is, in fact, a diffusion

with jumps. It is well known that if the price, P_t, of a stock is described by the Black-Scholes model (geometric Brownian motion), that is,

$$dP_t = \alpha^\diamond P_t dt + \sigma P_t dW_t,$$

then the logarithm of the price is a Brownian motion with drift, more precisely $X_t = \log P_t$ solves the stochastic differential equation

$$dX_t = \alpha dt + \sigma dW_t, \tag{4.19}$$

where $\alpha = \alpha^\diamond + \frac{1}{2}\sigma^2$. This follows from Itô's formula.

Suppose now that we want to allow jumps in the price process (and therefore also in the log-price process X). One of the simplest ways to achieve this is by adding a compound Poisson process term to the log-price process, that is to modify (4.19) in the following way,

$$dX_t = \alpha dt + \sigma dW_t + dZ_t, \tag{4.20}$$

where

$$Z_t = \sum_{j=0}^{N_t} Y_j,$$

and $\{N_t\}$ is a Poisson process with intensity λ. The stochastic process $\{N_t\}$ is thus a counting process with independent increments and N_t, the number of jumps in the time interval $[0, t]$, is Poisson distributed with parameter λt. The jump sizes Y_j, $j = 1, 2, \ldots$, are assumed to be i.i.d. normal with mean μ and variance τ^2. Furthermore, we assume that $\{W_t\}$, $\{N_t\}$ and $\{Y_j\}$ are independent and that $N_0 = Y_0 = 0$ so that $Z_0 = 0$. This is a simplified version of the kind of jump-diffusion models studied in Andersen, Benzoni & Lund (2002). The solution to (4.20) is given by

$$X_t = \alpha t + \sigma W_t + Z_t, \quad t \geq 0.$$

In figure 4.1 a simulated trajectory corresponding to the model in (4.20) is shown for the parameter values $\alpha = 0.0001$, $\sigma = 0.1$, $\lambda = 0.01$, $\mu = 1$, and $\tau = 0.1$.

For simplicity we consider observations $X_1, X_2, \ldots, X_n$. The parameter vector is in this case 5-dimensional, $\theta = (\alpha, \sigma^2, \lambda, \mu, \tau^2)^T$. We will derive an optimal martingale estimating function based on the functions

$$h(x, y; \theta) = \begin{pmatrix} y - F(x;\theta) \\ (y - F(x;\theta))^2 - \phi(x;\theta) \\ e^y - \kappa(x;\theta) \end{pmatrix},$$

where

$$\begin{aligned} F(x;\theta) &= \mathrm{E}_\theta(X_i | X_{i-1} = x) = x + \alpha + \lambda\mu, \\ \phi(x;\theta) &= \mathrm{Var}_\theta(X_i | X_{i-1} = x) = \sigma^2 + \lambda(\mu^2 + \tau^2), \\ \kappa(x;\theta) &= \mathrm{E}_\theta(e^{X_i} | X_{i-1} = x) = \exp\left(x + \alpha + \tfrac{1}{2}\sigma^2 + \lambda(e^{\mu + \frac{1}{2}\tau^2} - 1)\right). \end{aligned}$$

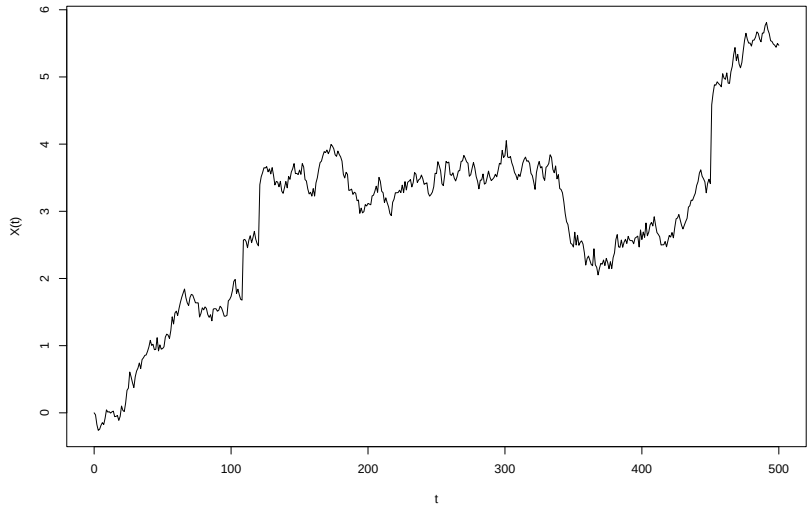

Figure 4.1: A simulated trajectory for the model given by (4.20) with parameter values $\alpha = 0.0001$, $\sigma = 0.1$, $\lambda = 0.01$, $\mu = 1$, and $\tau = 0.1$.

In order to find an expression for the optimal martingale estimating function based on h, we need the following quantities, see (4.16) and (4.15). The conditional covariance matrix for h is given by

$$\begin{aligned} V_h(x;\theta) &= \mathrm{E}_\theta\left(h(X_{i-1},X_i;\theta)h(X_{i-1},X_i;\theta)^T\Big|\, X_{i-1}=x\right) \\ &= \begin{pmatrix} \phi(x;\theta) & \eta(x;\theta) & \nu(x;\theta) \\ \eta(x;\theta) & \psi(x;\theta) & \rho(x;\theta) \\ \nu(x;\theta) & \rho(x;\theta) & \zeta(x;\theta) \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} \eta(x;\theta) &= \mathrm{E}_\theta((X_i - F(X_{i-1};\theta))^3|X_{i-1}=x) = \lambda\mu(\mu^2+3\tau^2), \\ \psi(x;\theta) &= \mathrm{E}_\theta((X_i - F(X_{i-1};\theta))^4|X_{i-1}=x) - \phi(x;\theta)^2 \\ &= 2\sigma^4 + \lambda[4\sigma^2(\mu^2+\tau^2) + (2\lambda+1)\mu^4 + (2\lambda+3)\tau^2(\tau^2+2\mu^2)], \\ \nu(x;\theta) &= \mathrm{E}_\theta((X_i - F(X_{i-1};\theta))(e^{X_i} - \kappa(X_{i-1};\theta))|X_{i-1}=x) \\ &= (\sigma^2 - \lambda\mu + \lambda(\mu+\tau^2)e^{\mu+\frac{1}{2}\tau^2})\kappa(x;\theta), \\ \rho(x;\theta) &= \mathrm{E}_\theta((X_i - F(X_{i-1};\theta))^2 - \phi(X_{i-1};\theta))(e^{X_i} - \kappa(X_{i-1};\theta))|X_{i-1}=x) \\ &= (\sigma^2 + \lambda(\tau^2 + (\mu+\tau^2)^2)e^{\mu+\frac{1}{2}\tau^2})\kappa(x;\theta) + \frac{(\nu(x;\theta) + F(x;\theta)\kappa(x;\theta))^2}{\kappa(x;\theta)} \end{aligned}$$

$$-2F(x;\theta)v(x;\theta) - F(x;\theta)^2\kappa(x;\theta) - \phi(x;\theta)\kappa(x;\theta),$$

$$\begin{aligned}\zeta(x;\theta) &= \mathrm{Var}_\theta(e^{X_i}|X_{i-1}=x)\\ &= e^{2x+2\alpha+\sigma^2-\lambda}\left(\exp\left(\sigma^2+\lambda e^{2\mu+2\tau^2}\right)-\exp\left(2\lambda e^{\mu+\frac{1}{2}\tau^2}-\lambda\right)\right).\end{aligned}$$

Furthermore we have that

$$-\mathrm{E}_\theta(\partial_\theta h(X_{i-1},X_i;\theta)|X_{i-1}=x) = \begin{pmatrix} 1 & 0 & \kappa(x;\theta)\\ 0 & 1 & \frac{1}{2}\kappa(x;\theta)\\ \mu & \mu^2+\tau^2 & (e^{\mu+\frac{1}{2}\tau^2}-1)\kappa(x;\theta)\\ \lambda & 2\lambda\mu & \lambda e^{\mu+\frac{1}{2}\tau^2}\kappa(x;\theta)\\ 0 & \lambda & \frac{1}{2}\lambda e^{\mu+\frac{1}{2}\tau^2}\kappa(x;\theta)\end{pmatrix}.$$

Hence an explicit expression for the optimal martingale estimating function is obtained, though the corresponding estimating equations have to be solved numerically. It should be noted that all subsets of the three functions defining h result in fewer than the required five estimating equations.

In Table 4.1 the empirical mean and standard error of 500 independent estimates of the five parameters are given. Each estimate is obtained from 500 simulated observations ($n = 500$ with $X_0 = 0$). The true parameter values are $\alpha = 0.0001$, $\sigma = 0.1$, $\lambda = 0.01$, $\mu = 1$, and $\tau = 0.1$ as in Figure 4.1.

Parameter	Mean	Standard error
α	-0.0009	0.0070
σ	0.0945	0.0180
λ	0.0155	0.0209
μ	0.9604	0.5126
τ	0.2217	0.3156

Table 4.1: Empirical mean and standard error of 500 estimates of the parameters in (4.20). The true parameter values are $\alpha = 0.0001$, $\sigma = 0.1$, $\lambda = 0.01$, $\mu = 1$, and $\tau = 0.1$.

From Table 4.1 we see that the mean of the parameter estimates in all cases are quite close to the true values. It is also clear, however, that for this particular choice of parameter values the estimates associated with the jumps of the process (μ and τ) are harder to estimate than the remaining parameters. This is not surprising as there are rather few jumps.

□

5 Optimal Estimating Functions for Diffusion Models

5.1 Optimal Linear Combinations of Relationships between Consecutive Observations

We will now again focus on diffusion models where X is supposed to be the solution to a stochastic differential equation (2.13). To simplify matters we will assume that X is one-dimensional.

Consider a class of estimating functions of the form (3.3) and (3.4), i.e.

$$G_n(\theta) = \sum_{i=1}^{n} a(\Delta_i, X_{t_{i-1}}, \theta) h(\Delta, X_{t_{i-1}}, X_{t_i}; \theta), \tag{5.1}$$

where $h = (h_1, \ldots, h_N)^T$ is a column vector of N given functions satisfying that

$$\int_\ell^r h_j(\Delta, x, y; \theta) p(\Delta, x, y; \theta) dy = 0$$

for all $\Delta > 0$, $x \in (\ell, r)$, and $\theta \in \Theta$, while the weight matrix a, a $p \times N$-matrix, can vary freely. The functions h_j define relationships (dependent on θ) between an observation X_i and the previous observation X_{i-1} that can be used to estimate θ. We shall now find the weight matrix a^* for which we obtain the Godambe and Heyde optimal combination of these relationships.

The class of estimating functions considered here is a particular case of the general type studied in Example 4.4, so by (4.16) the optimal estimating function is

$$G_n^*(\theta) = \sum_{i=1}^{n} a^*(\Delta_i, X_{t_{i-1}}; \theta) h(\Delta_i, X_{t_{i-1}}, X_{t_i}; \theta), \tag{5.2}$$

where

$$a^*(\Delta, x; \theta) = -\int_\ell^r \partial_\theta h(\Delta, x, y; \theta)^T p(\Delta, x, y; \theta) dy V_h(\Delta, x; \theta)^{-1}, \tag{5.3}$$

with

$$V_h(\Delta, x; \theta) = \int_\ell^r h(\Delta, x, y; \theta) h(\Delta, x, y; \theta)^T p(\Delta, x, y; \theta) dy. \tag{5.4}$$

Here it is assumed that $V_h(\Delta, x; \theta)$ is invertible, or equivalently that the functions h_j, $j = 1, \ldots N$ are linearly independent.

When the functions h are of the form (3.10) with π_Δ^θ defined by (2.21), the optimal estimating function is given by (5.2) with

$$a^*(\Delta, x; \theta) = B(\Delta, x; \theta) V(\Delta, x; \theta)^{-1}, \tag{5.5}$$

where

$$B(\Delta, x; \theta)_{ij} = -\int_\ell^r \partial_{\theta_i} f_j(y; \theta) p(\Delta, x, y; \theta) dy + \partial_{\theta_i} \pi_\Delta^\theta(f_j(\theta))(x), \tag{5.6}$$

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

$$V(\Delta, x; \theta)_{ij} = \int_\ell^r f_i(y; \theta) f_j(y; \theta) p(\Delta, x, y; \theta) dy - \pi_\Delta^\theta(f_i(\theta))(x) \pi_\Delta^\theta(f_j(\theta))(x), \tag{5.7}$$

$i, j = 1, \ldots, N$.

Particularly important examples are the linear and quadratic estimating functions. The optimal linear estimating function is

$$\sum_{i=1}^{n} \frac{\partial_\theta F(\Delta_i, X_{t_{i-1}}; \theta)}{\phi(\Delta_i, X_{t_{i-1}}; \theta)} [X_{t_i} - F(\Delta_i, X_{t_{i-1}}; \theta)], \tag{5.8}$$

where F and ϕ are given by (3.5) and (3.6). In the expression for the optimal linear estimating function the derivative of F appears. If F is determined by simulation, it is necessary to be careful to ensure that the derivative is correctly calculated. Pedersen (1994a) proposed a procedure for determining $\partial_\theta F(\Delta, x; \theta)$ by simulation based on results in Friedman (1975). However, it is often easier to use an approximation to the optimal estimating function, see the following subsection.

If the first and second moment of the transition distribution are both correctly specified, the estimator obtained from (5.8) is efficient in the non-parametric model that assumes X is a Markov process, but specifies only the first two moments, see Wefelmeyer (1996) and Wefelmeyer (1997).

The optimal quadratic estimating function depends on the third and fourth moments of the transition distribution. It is given by

$$\begin{aligned} \sum_{i=1}^{n} \Big\{ & \alpha^*(\Delta_i, X_{t_{i-1}}; \theta)[X_{t_i} - F(\Delta_i, X_{t_{i-1}}; \theta)] \\ & + \beta^*(\Delta_i, X_{t_{i-1}}; \theta)[(X_{t_i} - F(\Delta_i, X_{t_{i-1}}; \theta))^2 - \phi(\Delta_i, X_{t_{i-1}}; \theta)] \Big\}. \end{aligned} \tag{5.9}$$

with

$$\alpha^*(x; \theta) = \frac{\partial_\theta \phi(x; \theta)\eta(x; \theta) - \partial_\theta F(x; \theta)\psi(x; \theta)}{\phi(x; \theta)\psi(x; \theta) - \eta(x; \theta)^2}$$

and

$$\beta^*(x; \theta) = \frac{\partial_\theta F(x; \theta)\eta(x; \theta) - \partial_\theta \phi(x; \theta)\phi(x; \theta)}{\phi(x; \theta)\psi(x; \theta) - \eta(x; \theta)^2},$$

where the Δ's have been omitted,

$$\eta(x; \theta) = E_\theta([X_\Delta - F(x; \theta)]^3 | X_0 = x)$$

and

$$\psi(x; \theta) = E_\theta([X_\Delta - F(x; \theta)]^4 | X_0 = x) - \phi(x; \theta)^2.$$

If the first four moments of the transition distribution are correctly specified, the estimator is efficient in the non-parametric model that assumes a Markov process, but specifies only the first four moments, see Wefelmeyer (1996) and Wefelmeyer (1997).

Example 5.1 For a *mean-reverting diffusion model* given by (3.8) with $\beta > 0$, the first conditional moment F is given by (3.9). Hence the optimal linear estimating function

$$G_n^*(\alpha, \beta) = \begin{pmatrix} \displaystyle\sum_{i=1}^{n} \frac{1 - e^{-\beta}}{\phi(X_{i-1}; \alpha, \beta)} \left[X_i - X_{i-1}e^{-\beta} - \alpha(1 - e^{-\beta}) \right] \\ \\ \displaystyle\sum_{i=1}^{n} \frac{e^{-\beta}(\alpha - X_{i-1})}{\phi(X_{i-1}; \alpha, \beta)} \left[X_i - X_{i-1}e^{-\beta} - \alpha(1 - e^{-\beta}) \right] \end{pmatrix},$$

where for simplicity of exposition we have taken the observation times $t_i = i$. Here there is in general no explicit expression for the function ϕ that must be found by simulation or be approximated, see Subsection 5.2. The following simpler estimating function gives us exactly the same estimators:

$$\tilde{G}_n^*(\alpha,\beta) = \begin{pmatrix} \sum_{i=1}^n \frac{1}{\phi(X_{i-1};\alpha,\beta)} \left[X_i - X_{i-1}e^{-\beta} - \alpha(1-e^{-\beta})\right] \\ \sum_{i=1}^n \frac{X_{i-1}}{\phi(X_{i-1};\alpha,\beta)} \left[X_i - X_{i-1}e^{-\beta} - \alpha(1-e^{-\beta})\right] \end{pmatrix}.$$

This is because $\tilde{G}_n^*(\alpha,\beta) = M(\alpha,\beta)\, G_n^*(\alpha,\beta)$, where the matrix

$$M(\alpha,\beta) = \begin{pmatrix} \frac{1}{1-e^{-\beta}} & 0 \\ \frac{\alpha}{1-e^{-\beta}} & -e^{-\beta} \end{pmatrix}$$

is invertible. Quite generally, if $G(\theta)$ is an estimating function and if $M(\theta)$ is a deterministic invertible matrix, then the estimating function $M(\theta)G(\theta)$ defines the same estimators as $G(\theta)$. Moreover, if $G(\theta)$ is optimal, then so is $M(\theta)G(\theta)$. We say that the two estimating functions are *equivalent*. Usually we will use the simplest possible version of the estimating function. See also the discussion of this problem after formula (4.5).

For the CIR model where $\sigma(x) = \tau\sqrt{x}$ the functions ϕ, η and ψ and hence the optimal quadratic estimating function can be found explicitly:

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

In view of Example 3.2 and results given in the following, it is not surprising that the conditional moments can be found explicitly for the CIR model. □

The optimal estimating function takes a particularly simple form in the case where the base $f_1, \ldots, f_N$ of the class of estimating functions consists of eigenfunctions of the generator, see (3.15) and the discussion below that formula. For such a base, the optimal estimating function is given by (5.5) with

$$B(\Delta,x;\theta)_{ij} = -\int \partial_{\theta_i}\varphi_j(y;\theta)p(\Delta,x,y;\theta)dy + \partial_{\theta_i}[e^{-\lambda_j(\theta)\Delta}\varphi_j(x;\theta)],$$

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

$$C(\Delta,x;\theta)_{ij} = \int \varphi_i(y;\theta)\varphi_j(y;\theta)p(\Delta,x,y;\theta)dy - e^{-[\lambda_i(\theta)+\lambda_j(\theta)]\Delta}\varphi_i(x;\theta)\varphi_j(x;\theta),$$

$i, j = 1, \ldots, N$. These expressions are relatively easy to determine by simulation because the differentiation is inside the integral. As mentioned earlier, numerical determination of quantities like $\partial_\theta F$ in (5.8) requires some care, but this problem disappears in the case of an eigenfunction base.

For many models where eigenfunctions can be found, they are of the form

$$\varphi_i(y;\theta) = \sum_{j=0}^{i} a_{i,j}(\theta)\,\kappa(y)^j \tag{5.10}$$

where κ is a real function defined on the state space and independent of θ. In this situation the optimal estimating function is explicit. To see this, note that

$$C_{i,j}(x,\theta) = \sum_{r=0}^{i}\sum_{s=0}^{j} a_{i,r}(\theta)a_{j,s}(\theta)\int \kappa(y)^{r+s}p(\Delta,x,y;\theta)dy - e^{-[\lambda_i(\theta)+\lambda_j(\theta)]\Delta}\,\varphi_i(x;\theta)\varphi_j(x;\theta)$$

and

$$B_i(x,\theta) = -\sum_{j=0}^{i}\partial_\theta a_{i,j}(\theta)\int \kappa(y)^i p(\Delta,x,y;\theta)dy + \partial_\theta(e^{-\lambda_j(\theta)\Delta}\varphi_i)(x;\theta).$$

Hence if we can find the moments $\int \kappa(y)^i p(\Delta,x,y;\theta)dy$ for $1 \le i \le 2N$, we have found the optimal estimating function based on the first N eigenfunctions. But this is easy since by integrating both sides of (5.10) with respect to $p(\Delta,x,y;\theta)$ for $i = 1,\ldots,2N$, we obtain the following system of linear equations

$$e^{-\lambda_i(\theta)}\varphi_i(x;\theta) = \sum_{j=0}^{i} a_{i,j}(\theta)\int \kappa(y)^j p(\Delta,x,y;\theta)dy \tag{5.11}$$

for $i = 1,\ldots,2N$.

Example 5.2 For the model considered in Example 3.3 the eigenfunctions are $\phi_i(x;\theta) = C_i^\theta(\sin(x)),\ i = 0,1,\cdots$, with eigenvalues $i(\theta + i/2),\ i = 0,1,\cdots$, where C_i^θ is the Gegenbauer polynomial of order i. The optimal estimating function based on any set of eigenfunctions can thus be found explicitly using (5.11). The optimal estimating function based on the first non-trivial eigenfunction, $\sin(x)$, is

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

When Δ is small the optimal estimating function can be approximated by

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

which yields the explicit estimator

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

provided the numerator is positive. In a simulation study with $\Delta \le 0.5$ this estimator was almost as efficient as the optimal estimator based on G^*, see Kessler & Sørensen (1999).

□

Statistical inference based on an optimal estimating function with an eigenfunction base is *invariant* under twice continuously differentiable transformations of data, see Kessler & Sørensen (1999). After such a transformation the data are, by Itô's formula, still observations from a certain diffusion process, and the eigenfunctions transform in exactly the way needed to keep the optimal estimating function invariant. Inference based on polynomial estimating functions is not invariant under transformations of the data. As mentioned above, the optimal estimating functions with eigenfunction base have clear numerical advantages over other estimating functions. The disadvantage of these estimating functions, on the other hand, is that it is not always possible to find eigenfunction for the generator of a given diffusion model. If eigenfunctions cannot be found, the polynomial estimating functions, in particular the quadratic, provide a very useful alternative.

A further justification for estimating functions based on eigenfunctions is that the eigenvalue problem (3.16) is a Sturm-Liouville problem. By a classical result of this theory, we have, for an ergodic diffusion with invariant probability μ_θ, a series expansion in terms of the eigenfunctions $(\phi_i)_{i\geq 0}$ of any function f satisfying that $\mu_\theta(f^2) < \infty$ (see Coddington & Levinson (1955)), i.e.

$$f(y) = \sum_{i=0}^{\infty} c_i \phi_i(y), \tag{5.12}$$

where (c_i) is a sequence of real numbers, and where the series converges with respect to the norm given by $\|f\|_\theta = \mu_\theta(f^2)^{\frac{1}{2}}$. Thus for a fixed x, $\sum_{j=0}^{k} \alpha_j(x;\theta)\,\phi_j(y;\theta)$ can be seen as a truncated series of the form (5.12). The estimating function given by (3.18) is obtained when one compensates the sum to obtain a martingale. The transition density can usually be expanded in the form (5.12), which mainly depends on the eigenfunctions with the smallest eigenvalues. In fact, the weights c_i decrease exponentially with the eigenvalues. If the score function can be expanded similarly, there is reason to expect rather efficient estimators. Suppose that the union $\cup_{k=1}^{\infty} V_k$, where V_k is the space spanned by $\{\phi_1(\cdot\,;\theta), \cdots, \phi_k(\cdot\,;\theta)\}$, is dense in the space $L_2(p(\Delta, x, y;\theta)dy)$ for every x. Then there exists a sequence N_n such that the estimator $\hat{\theta}_{n,N_n}$ is efficient. Here $\hat{\theta}_{n,N}$ is the optimal estimator based on N eigenfunctions and n observations. For details, see Kessler (1996). In particular in the case of a bounded state interval, where it is well known that the sequence $\phi_1(\cdot\,;\theta), \phi_2(\cdot\,;\theta), \cdots$ is complete in $L_2(\mu_\theta)$, the union $\cup_{k=1}^{\infty} V_k$ is dense in $L_2(p(\Delta, x, y;\theta)dy)$, so in this case there generally exists a sequence N_n such that the estimator $\hat{\theta}_{n,N_n}$ is efficient. In the case of an unbounded state interval, the sequence $\phi_1(.;\theta), \phi_2(.;\theta), \cdots$ is also complete in $L_2(\mu_\theta)$ when the set of eigenfunctions is discrete, but in order to deduce denseness of $\cup_{k=1}^{\infty} V_k$ in $L_2(p(\Delta, x, y;\theta)dy)$, additional conditions are needed. The efficiency of $\hat{\theta}_{n,N}$ obviously increases with increasing N, but so does the computational complexity. It is conjectured that for many models $\cup_{k=1}^{\infty} V_k$ is dense in $L_2(p(\Delta, x, y;\theta)dy)$ so that the efficiency is high, provided that N is sufficiently large, and a compromise between efficiency and computational feasibility must be found.

Example 5.3 For the target zone model in Example 3.4, the eigenfunctions are the Jacobi polynomials with eigenvalues $\lambda_i = i[\beta + \frac{1}{2}\sigma^2(i-1)]$, $i = 1, 2, \ldots$. Therefore it is easy to apply (5.11) to obtain explicit expressions for the optimal estimating function based on any fixed number of eigenfunctions. In Larsen & Sørensen (2003) the asymptotic variances of the estimators obtained from several combinations of eigenfunctions were calculated for certain parameter values. It turned out that in no case was the efficiency much above that obtained when using the optimal estimating function based on the first two eigenfunctions ϕ_1 and ϕ_2.

In view of the result on efficiency for diffusions with a bounded state space mentioned above, it is reasonable to assume that for the parameter values considered this estimating function is close to fully efficient.

□

5.2 Approximately Optimal Estimating Functions

For models where the optimal weight matrix $a^*(\Delta, x; \theta)$ is not explicit and must be calculated by means of simulations, it is often preferable to use a good approximation to $a^*(\Delta, x; \theta)$ instead. This will usually save a lot of computer time and make the estimation procedure more numerically robust. To make such an approximation, the following result is useful. As in the previous subsection we focus here on one-dimensional diffusions.

Suppose f is a $2(k+1)$ times continuously differentiable function. Then under weak conditions on f and the diffusion model

$$\mathrm{E}_\theta(f(X_{t+s}) \,|\, X_t) = \sum_{i=0}^{k} \frac{s^i}{i!} L_\theta^i f(X_t) + O(s^{k+1}), \tag{5.13}$$

where L_θ denotes the generator (3.15), see e.g. Kessler (1997). A sufficient conditions on f is that it is of polynomial growth. By applying this formula to $f(x) = x$ and $f(x) = x^2$ it follows that

$$\begin{aligned} E_\theta(X_\Delta | X_0 = x) = x \;&+\; \Delta b(x;\theta) + \tfrac{1}{2}\Delta^2\{b(x;\theta)\partial_x b(x;\theta) \\ &+\; \tfrac{1}{2} v(x;\theta)\partial_x^2 b(x;\theta)\} + O(\Delta^3) \end{aligned} \tag{5.14}$$

and

$$\begin{aligned} \mathrm{Var}_\theta(X_\Delta | X_0 = x) \;&=\; \Delta v(x;\theta) + \Delta^2[\tfrac{1}{2} b(x;\theta)\partial_x v(x;\theta) \\ &+\; v(x;\theta)\{\partial_x b(x;\theta) + \tfrac{1}{4}\partial_x^2 v(x;\theta)\}] + O(\Delta^3), \end{aligned} \tag{5.15}$$

where $v(x;\theta) = \sigma^2(x;\theta)$.

If we insert the approximations

$$\partial_\theta F(t, x; \theta) \doteq t\partial_\theta b(x;\theta) \quad \text{and} \quad \phi(t, x; \theta) \doteq t v(x;\theta) \tag{5.16}$$

in the expression for the optimal linear estimating function (5.8) we obtain the approximately optimal estimating function

$$\sum_{i=1}^{n} \frac{\partial_\theta b(X_{t_{i-1}};\theta)}{v(X_{t_{i-1}};\theta)} [X_{t_i} - F(\Delta_i, X_{t_{i-1}};\theta)], \tag{5.17}$$

which is usually considerably easier to calculate than (5.8). When t is small, the approximation (5.16) is good, but the approximately optimal estimating function (5.17) works surprisingly well for large values of Δ_i too. By means of the formulae (5.14) and (5.15) Bibby & Sørensen (1995) showed that in the case of equidistant sampling times (i.e. for $\Delta_i = \Delta$) the asymptotic variance of the estimators based on the optimal estimating function (5.8) and the approximation (5.17) coincide up to and including terms of order $O(\Delta^2)$. The term of order $O(\Delta)$ is equal to

the similar term for the maximum likelihood estimator found by Dacunha-Castelle & Florens-Zmirou (1986). Numerical calculations in Bibby & Sørensen (1995) indicate that for the CIR model the efficiencies of the two estimators are similar even for large values of Δ.

To simplify the optimal quadratic estimating function, we supplement (5.16) by the Gaussian approximations

$$\eta(t,x;\theta) \doteq 0 \quad \text{and} \quad \psi(t,x;\theta) \doteq 2\phi(t,x;\theta)^2 \tag{5.18}$$

that are also good for small Δ-values. By inserting these approximations into (5.9) we obtain the approximately optimal quadratic estimating function

$$\begin{aligned} &\sum_{i=1}^{n} \left\{ \frac{\partial_\theta b(X_{t_{i-1}};\theta)}{v(X_{t_{i-1}};\theta)} [X_{t_i} - F(\Delta_i, X_{t_{i-1}};\theta)] \right. \\ &\quad + \left. \frac{\partial_\theta v(X_{t_{i-1}};\theta)}{2v^2(X_{t_{i-1}};\theta)\Delta_i} \left[(X_{t_i} - F(\Delta_i, X_{t_{i-1}};\theta))^2 - \phi(\Delta_i, X_{t_{i-1}};\theta) \right] \right\}, \end{aligned} \tag{5.19}$$

which is a very considerable computational improvement over (5.9). This is not least because in (5.19) there are only derivatives of known functions, while (5.9) contains derivatives of functions that must often be determined by simulation. The approximately optimal quadratic estimating function (5.19) should be compared to the score corresponding to the Gaussian pseudo-likelihood (3.7).

Example 5.4 For the CIR model given by (3.8) with $\sigma(x) = \tau\sqrt{x}$ we obtain the approximately optimal quadratic estimating function

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

As earlier we have assumed that $t_i = i$ and given the simplest possible version of the estimating function, which is obtained by multiplying the estimating function obtained from (5.19) by the matrix

$$\begin{Bmatrix} \tau^2/\beta & 0 & 0 \\ \alpha\tau^2/\beta & -\tau^2 & 0 \\ 0 & 0 & \tau^3 \end{Bmatrix}.$$

We find the following explicit estimators of the parameters

$$\begin{aligned} \tilde{\alpha}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{e^{-\tilde{\beta}_n}}{n\left(1-e^{-\tilde{\beta}_n}\right)}(X_n - X_0) \\ e^{-\tilde{\beta}_n} &= \frac{n\sum_{i=1}^{n} X_i/X_{i-1} - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} X_{i-1}^{-1})}{n^2 - (\sum_{i=1}^{n} X_{i-1})(\sum_{i=1}^{n} X_{i-1}^{-1})} \\ \tilde{\tau}_n^2 &= \frac{\sum_{i=1}^{n} X_{i-1}^{-1} \left(X_i - X_{i-1}e^{-\tilde{\beta}_n} - \tilde{\alpha}_n(1-e^{-\tilde{\beta}_n}) \right)^2}{\sum_{i=1}^{n} X_{i-1}^{-1} \left((\frac{1}{2}\tilde{\alpha}_n - X_{i-1})e^{-2\tilde{\beta}_n} - (\tilde{\alpha}_n - X_{i-1})e^{-\tilde{\beta}_n} + \frac{1}{2}\tilde{\alpha}_n \right) / \tilde{\beta}_n}, \end{aligned}$$

which exist provided that the expression for $e^{-\tilde{\beta}_n}$ is strictly positive, an event that happens with a probability tending to one as $n \to \infty$. A simulation study and an investigation of the asymptotic variance of the estimators $\tilde{\alpha}_n$ and $\tilde{\beta}_n$ in Bibby & Sørensen (1995) indicate that these estimators are quite efficient; see also the simulation study in Overbeck & Rydén (1997). Note that the level α is essentially estimated by the average of the observations. In practice it is recommended to simply use the average as this is easier and causes no lost of asymptotic efficiency.

□

The expansion (5.13) can be used to simplify the expressions for the optimal weights in many other estimating functions. This 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_i is not too large, it will just lead to a minor loss of efficiency. The magnitude of this loss of efficiency can be calculated by means of (5.13), or in the case of the quadratic estimating function by means of (5.14) and (5.15).

Example 5.5 In this example we consider monthly observations of US one-month treasury bill yields from June 1964 to December 1989. These data were also analysed by Chan et al. (1992). The rates have been annualized and converted into continuously compounded yields. In Figure 5.1 the yields are plotted against time.

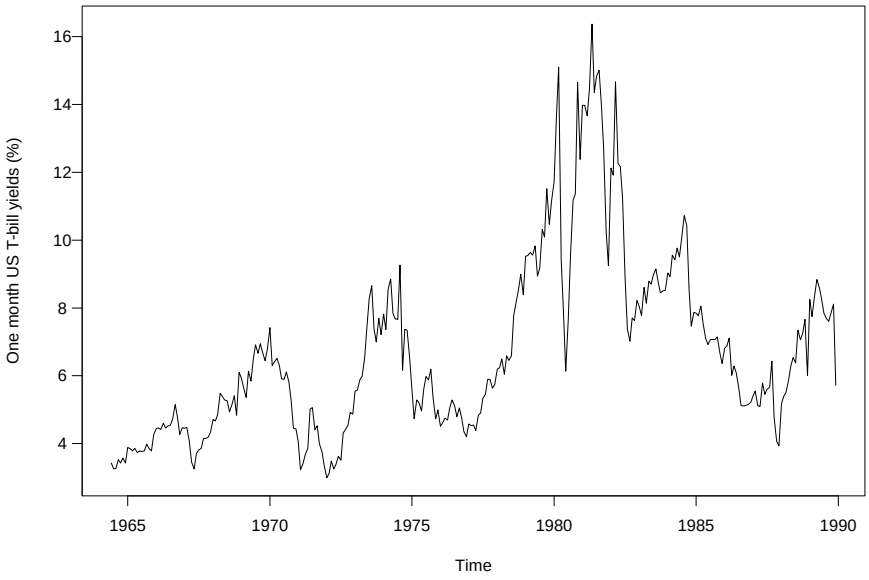

Figure 5.1: The one-month treasury bill yields plotted against time.

We use two different diffusion process models to describe the data, namely the model introduced in Chan et al. (1992), which we will refer to as the CKLS-model, and the generalized CIR-model (GCIR-model) introduced in Jacobsen (2002) and considered more closely in Example 5.9 below. If X_t denotes the yield at time t then the CKLS-model is given by the stochastic differential equation,

$$dX_t = \kappa(\theta - X_t)dt + \sigma X_t^{\gamma} dW_t.$$

The stochastic differential equation defining the GCIR-model is given in (5.36).

The observations are denoted $X_\Delta, X_{2\Delta}, \ldots, X_{n\Delta}$ where n is 307 and $\Delta = 1/12$. For both models the parameters are estimated using the approximation to the optimal quadratic martingale estimating function given by (5.19). For the CKLS-model the conditional expectation can be found explicitly, while the conditional variance is found using simulations. In case of the GCIR-model both the conditional mean and the conditional variance are determined by simulations. In Table 5.1 and Table 5.2 the estimates for the parameter in the two models are given based on both the whole time-series and for the period June 1964 to September 1979 ($n = 184$). The reason for considering the latter period separately is that between October 1979 and October 1982 the U.S. Federal Bank employed a monetary rather than an interest rate targeting policy resulting in a quite different stochastic regime.

	1964–1989	1964–1979
θ	0.0735	0.0676
κ	0.3309	0.3376
σ	1.0119	0.6311
γ	1.3833	1.2755

Table 5.1: Estimates for the parameters in the CKLS-model based on two periods.

	1964–1989	1964–1979
α	1.4093	0.7571
β	-1.2110	-0.5491
σ	0.3905	0.2987
γ	0.9997	0.9997

Table 5.2: Estimates for the parameters in the GCIR-model based on two periods.

For a more detailed analysis of these data based on the CKLS-model, see Christensen, Poulsen & Sørensen (2001). Note from table 5.2 that the estimate of the parameter γ in the GCIR-model is quite close to 1.

In Figure 5.2 uniform residuals corresponding to both models and both time periods are given. We see that the two models give almost equally good descriptions of the data. We may also note that the models clearly fit the data from June 1964 to September 1979 better than the whole data set. Model diagnostics based on uniform residuals was introduced and discussed by Pedersen (1994b).

□

It is tempting to go on and approximate the functions F and ϕ still appearing in (5.17) and (5.19) by $F(t, x; \theta) \doteq x + tb(x; \theta)$ and $\phi(t, x; \theta) \doteq tv(x; \theta)$. This certainly leads to a very simple estimation procedure that has often been applied, but it is important to note that there is a dangerous pitfall here. First, if F and ϕ are replaced by approximations, the martingale property is destroyed, so that stronger conditions on the process are needed to ensure asymptotic normality, see the discussion in Subsection 2.3. This is usually a minor problem. What is much worse is that the estimating function becomes biased, which implies that the estimator becomes inconsistent, at least under the kind of asymptotics considered so far. For the consistency result in Theorem 2.2 to hold for an estimating function of the form (3.3) it is important that the estimating function is unbiased, i.e. that $Q_\theta^\Delta(g(\Delta, \theta)) = 0$, so that as $n \to \infty$

$$\frac{1}{n} \sum_{i=1}^{n} g(\Delta, X_{\Delta(i-1)}, X_{\Delta i}; \theta) \to 0.$$

There is a version of Theorem 2.2 for biased estimating functions. Suppose for the true param-

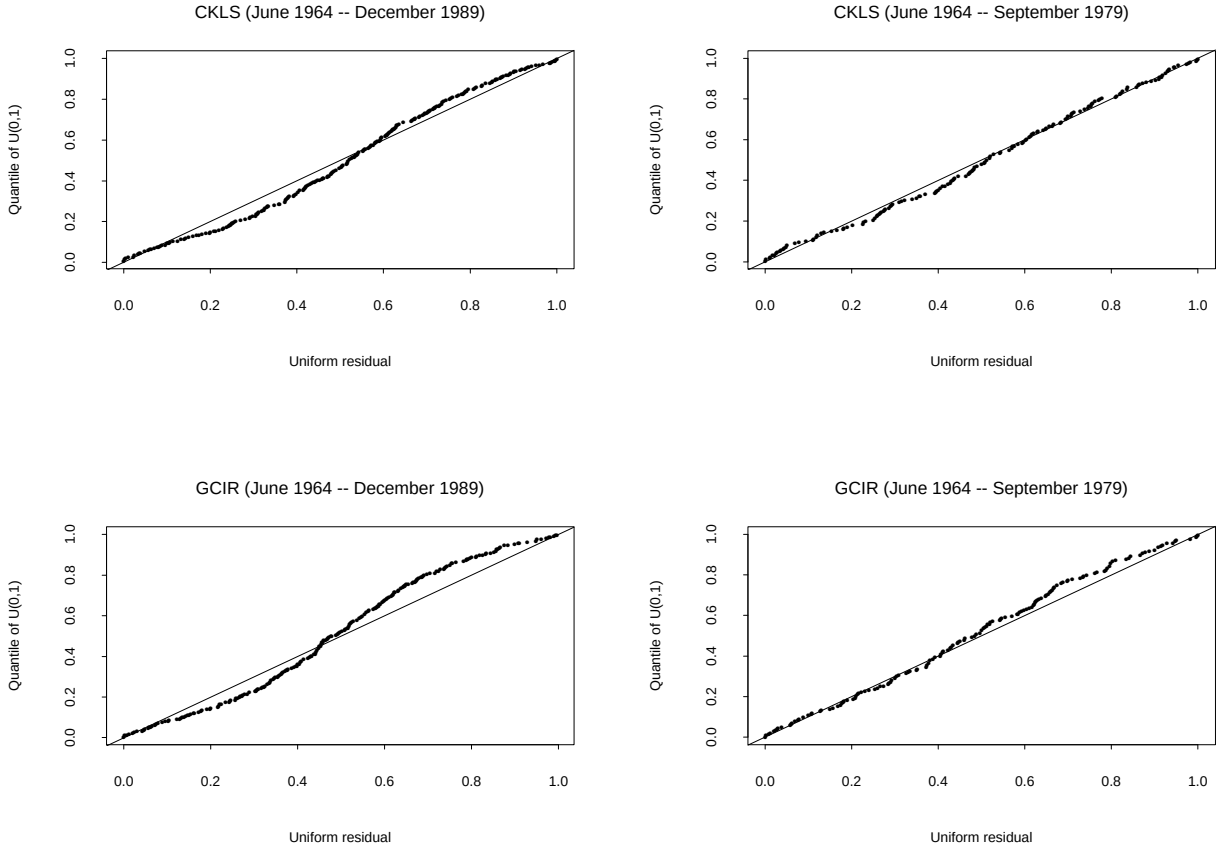

Figure 5.2: Uniform residuals corresponding to the CKLS-model and the GCIR-model based on observations in two periods.

eter value θ_0 the equation

$$Q_{\theta_0}^{\Delta}(g(\Delta, \bar{\theta})) = 0 \tag{5.20}$$

has a unique solution $\bar{\theta}$. Then according to the more general version of Theorem 2.2

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

as $n \to \infty$.

Example 5.6 For a general mean-reverting process (3.8) the approximate linear estimating function where the conditional expectation is replaced by the first order expansion is (for equidistant observation, $t_i = \Delta i$)

$$\begin{pmatrix} \sum_{i=1}^{n} \frac{1}{v(X_{\Delta(i-1)})} \left[X_{\Delta i} - X_{\Delta(i-1)} + \Delta\beta(X_{\Delta(i-1)} - \alpha) \right] \\ \sum_{i=1}^{n} \frac{X_{\Delta(i-1)}}{v(X_{\Delta(i-1)})} \left[X_{\Delta i} - X_{\Delta(i-1)} + \Delta\beta(X_{\Delta(i-1)} - \alpha) \right] \end{pmatrix}. \tag{5.21}$$

For the CIR process the weights are $X_{\Delta(i-1)}^{-1}$ and 1, and it is not difficult to find the explicit estimators obtained from (5.21) for this model:

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

$$\hat{\beta}_n = \frac{\frac{1}{n}(X_{\Delta n} - X_0)\sum_{i=1}^n X_{\Delta(i-1)}^{-1} - \sum_{i=1}^n X_{\Delta(i-1)}^{-1}(X_{\Delta i} - X_{\Delta(i-1)})}{\Delta[n - (\sum_{i=1}^n X_{\Delta(i-1)})(\sum_{i=1}^n X_{\Delta(i-1)}^{-1})/n]}.$$

The asymptotic bias of these estimators as $n \to \infty$ can easily be found using the ergodic theorem and the fact that the invariant probability measure for the CIR model is a gamma distribution. However, a result for general mean-reverting processes can be obtained by solving the equation (5.20) for the estimating function (5.21). The solutions are

$$\bar{\alpha} = \alpha_0 \qquad \text{and} \qquad \bar{\beta}\Delta = 1 - e^{-\beta_0 \Delta} \leq 1.$$

Thus the estimator of α is in fact consistent. Contrary to this, the estimator of the reversion parameter β is reasonable only when $\beta_0\Delta$ is considerably smaller than one. Note that $\bar{\beta} \leq \Delta^{-1}$, so the estimator will always converge to a limit smaller than the sampling frequency. When $\beta_0\Delta$ is large, the behaviour of the estimator is bizarre. Without prior knowledge of the value of β_0 it is thus a very dangerous estimator, which has unfortunately frequently been applied in the econometric literature, for instance in Chan et al. (1992).

□

Using (5.14) it is easy to see that in general the bias of

$$\sum_{i=1}^n \frac{\partial_\theta b(X_{\Delta(i-1)};\theta)}{v(X_{\Delta(i-1)};\theta)} \left[X_{\Delta i} - X_{\Delta(i-1)} - \Delta b(X_{\Delta(i-1)};\theta)\right]$$

is of order Δ^2 when the observation time points are equidistant. One would therefore expect that in an asymptotic scenario, where Δ goes to zero as $n \to \infty$ the estimator is consistent. This is in fact true. Dorogovcev (1976), Prakasa Rao (1983), Prakasa Rao (1988), and Florens-Zmirou (1989) proved that the estimator is consistent provided that $\Delta \to 0$ and $n\Delta \to \infty$ as $n \to \infty$. Moreover, the estimator is asymptotically normal if it is further assumed that $n\Delta^2 \to 0$. Prakasa Rao (1988) called this a rapidly increasing experimental design. A general result comprising also more accurate approximations of F and ϕ based on (5.13) was given by Kessler (1997). By choosing the approximations in a suitable way, Kessler obtained estimators that are asymptotically normal provided just that $n\Delta^k \to 0$ for a $k \in \mathbb{N}$ that depends on the order of the approximation.

5.3 Small Δ-optimality

We shall here discuss a new optimality criterion for unbiased estimating functions that was introduced by Jacobsen (2001a) and explored further in the case of martingale estimating functions of the form (3.14) in Jacobsen (2001b) and Jacobsen (2002). Throughout this subsection, we shall assume that the observation times are equidistant, i.e. $t_i = i\Delta$, $0 \leq i \leq n$, where Δ is fixed. That an estimating function is small Δ-optimal implies that for $\Delta > 0$ small, the resulting estimator is nearly efficient. Furthermore, as will be demonstrated, it is easy to find explicitly given estimating functions that are small Δ-optimal.

To illustrate the main idea, consider a martingale estimating function as in (3.3). The covariance matrix of the asymptotic distribution of $\hat{\theta}_n$ is (with θ denoting the true parameter value)

$$\mathrm{Var}_{\Delta,\theta}\left(g, \hat{\theta}\right) = S(\theta)^{-1} V(\theta) \left(S^{-1}(\theta)\right)^T, \tag{5.22}$$

where the matrices $S(\theta) = (S_{ij}(\theta))_{i\leq i,j\leq p}$ and $V(\theta) = (V_{ij}(\theta))_{i\leq i,j\leq p}$ are given by

$$S_{ij}(\theta) = \mathrm{E}_\theta\left(\partial_{\theta_j} g_i(\Delta, X_0, X_\Delta; \theta)\right), \quad V_{ij}(\theta) = \mathrm{E}_\theta\left(g_i(\Delta, X_0, X_\Delta; \theta) g_j(\Delta, X_0, X_\Delta; \theta)\right); \quad (5.23)$$

see Theorem 2.2. Now allow $\Delta > 0$ to vary freely and consider the covariance matrix $\mathrm{Var}_{\Delta,\theta}\left(g, \hat{\theta}\right)$ as a function of Δ. The optimal martingale estimating function with base f (cf. (3.14)) comes about by minimizing $\mathrm{Var}_{\Delta,\theta}\left(g, \hat{\theta}\right)$ for a *given* $\Delta > 0$ when the weights vary (minimizing in the partial order on the space of covariance matrices). Different choices of f lead to different optimal martingale estimating functions of different quality. Each of them is *locally optimal* in the sense that the resulting estimator is the best within the subclass of estimators given by the chosen base f, but estimators from subclasses given by other choices of f may do better.

By contrast, for the discussion of small Δ-optimality, we consider $\mathrm{Var}_{\Delta,\theta}\left(g, \hat{\theta}\right)$ given by (5.22) for $\Delta \to 0$ and show that in the limit a universal lower bound for the asymptotic covariance can be obtained. This implies that for small values of Δ (high frequency data), the estimator obtained from a small Δ-optimal estimating function is in practice (almost) as good as the maximum-likelihood estimator. Thus small Δ-optimality is a *global optimality* criterion. Although small Δ-optimality refers explicitly to the limit $\Delta \to 0$, for any given fixed $\Delta > 0$ the estimator obtained is still $\sqrt{n}$-consistent and asymptotically Gaussian as the sample size goes to infinity. There is no guarantee that it is Godambe or Heyde optimal (relative to the base f), but for Δ not too large, it should still behave well, as has been verified in several examples.

The martingale estimating functions we shall use for the discussion here are of the form (3.3) with the ith coordinate of g given by

$$g_i(\Delta, x, y; \theta) = \sum_{j=1}^{N} a_{ij}(x; \theta)\left(f_j(y) - \pi_\Delta^\theta(f_j)(x)\right) \quad (1 \leq i \leq p). \quad (5.24)$$

It is assumed that neither the base functions f_j nor the weights a_{ij} depend on Δ (cf. (3.14)). The f_j may depend on θ, but for the time we ignore such a dependence. We also make the following vital assumption.

Condition 5.7 *The functions $f_j(x)$ are supposed to be twice differentiable in x. Also, the base f is supposed to have full affine rank N on the domain D, i.e. the identity*

$$\sum_{j=1}^{N} c_j f_j(x) + \gamma = 0 \quad (x \in D)$$

for some constants c_j, γ implies that $c_1 = \cdots = c_N = \gamma = 0$.

The functions $a_{ij}(x; \theta)$ are supposed to satisfy that for any θ, the p N-dimensional functions $x \to (a_{i1}(x; \theta), \ldots, a_{iN}(x; \theta))$ forming the rows of $a(x; \theta)$ are linearly independent on D.

As $\Delta \to 0$, neighbouring observations $\left(X_{(i-1)\Delta}, X_{i\Delta}\right)$ will, since X is continuous, get very close together. It is therefore not surprising that it is the limit

$$\begin{aligned} g_{i,0}(x, y; \theta) &= \lim_{\Delta \to 0} g_i(\Delta, x, y; \theta) \quad (5.25) \\ &= \sum_{j=1}^{N} a_{ij}(x, \theta)\left(f_j(y) - f_j(x)\right) \end{aligned}$$

and its behaviour close to the diagonal $y = x$ that determines the structure of $\mathrm{Var}_{\Delta,\theta}\left(g, \hat{\theta}\right)$ as $\Delta \to 0$. More specifically, using Itô-Taylor expansions of the random variables that determine the matrices $V(\theta)$ and $S(\theta)$ in the expression for $\mathrm{Var}_{\Delta,\theta}\left(g, \hat{\theta}\right)$, see (5.22) and (5.23), subject to integrability conditions, we obtain an expansion of the form

$$\mathrm{Var}_{\Delta,\theta}\left(g, \hat{\theta}\right) = \frac{1}{\Delta} v_{-1,\theta}\left(g, \hat{\theta}\right) + v_{0,\theta}\left(g, \hat{\theta}\right) + o(1) \tag{5.26}$$

as $\Delta \to 0$ (Jacobsen (2001a), Section 6). The expressions for the coefficient matrices $v_{-1,\theta}$ and $v_{0,\theta}$ depend in an essential way on the structure of the model, and we shall distinguish between three cases (i), (ii) and (iii) (where to achieve the structure in (iii), it may be necessary first to reparametrize the model). For each case we list conditions under which the relevant coefficients are minimized, i.e. conditions under which small Δ-optimality is achieved. For the cases (i) and (ii) we also give the universal lower bounds on $v_{-1,\theta}$ (case (i)) and $v_{0,\theta}$ (case(ii)).

(i) *$C(x;\theta) = C(x)$ does not depend on θ.* In this case the main term in (5.26) is always present and small Δ-optimality is achieved by minimizing globally (over all g) the quantity $v_{-1,\theta}\left(g, \hat{\theta}\right)$. A sufficient condition for a given g to be small $\Delta-$optimal is that

$$\partial_y g_0(x, x; \theta) = \dot{b}^T(x;\theta) C^{-1}(x). \tag{5.27}$$

Here $\partial_y g_0(x, x; \theta)$ evaluates $\partial_y g_0(x, y; \theta) = (\partial_{y_k} g_{i,0}(x, y; \theta)) \in \mathbb{R}^{p \times d}$ along the diagonal $y = x$, and $\dot{b}(x;\theta) \in \mathbb{R}^{d \times p}$ with $\left(\dot{b}(x;\theta)\right)_{ki} = \partial_{\theta_i} b_k(x;\theta)$. If (5.27) holds, $v_{-1,\theta}\left(g, \hat{\theta}\right)$ attains its lower bound

$$\left[E_\theta\left(\dot{b}_\theta^T\left(X_0\right) C^{-1}\left(X_0\right) \dot{b}_\theta\left(X_0\right)\right)\right]^{-1}.$$

(ii) *$C(x;\theta)$ depends on all parameters $\theta_1, \ldots, \theta_p$.* Then the main term in (5.26) vanishes provided $\partial_y g_0\left(x, x; \theta\right) \equiv 0$, and small Δ-optimality is achieved by minimizing $v_{0,\theta}\left(g, \hat{\theta}\right)$. A sufficient condition for g to be small Δ-optimal is that

$$\partial_y g_0(x, x; \theta) = 0, \quad \partial^2_{yy} g_0(x, x; \theta) = \dot{C}^T(x;\theta)\left(C^{\otimes 2}(x;\theta)\right)^{-1}, \tag{5.28}$$

where $\partial^2_{yy} g_0\left(x, x; \theta\right) \in \mathbb{R}^{p \times d^2}$ evaluates the second derivatives $\partial^2_{y_k y_\ell} g_{i,0}\left(x, y; \theta\right)$ along the diagonal $y = x$, $\dot{C}(x;\theta) \in \mathbb{R}^{d^2 \times p}$ with $\left(\dot{C}(x;\theta)\right)_{k\ell,i} = \partial_{\theta_i} C_{k\ell}(x;\theta)$, and $C^{\otimes 2} \in \mathbb{R}^{d^2 \times d^2}$ is given by $(C^{\otimes 2})_{k\ell,k'\ell'} = C_{kk'} C_{\ell\ell'}$. If (5.28) holds, $v_{0,\theta}\left(g, \hat{\theta}\right)$ attains its lower bound

$$2\left[E_\theta\left(\dot{C}_\theta^T\left(X_0\right)\left(C^{\otimes 2}\left(X_0\right)\right)^{-1} \dot{C}_\theta\left(X_0\right)\right)\right]^{-1}.$$

(iii) *C_θ depends on the parameters $\theta_1, \ldots, \theta_{p'}$, but not on $\theta_{p'+1}, \ldots, \theta_p$ for some p' with $1 \leq p' < p$.* Here parts of the main term in (5.26) can be made to disappear so that

$$v_{-1,\theta}\left(g, \hat{\theta}\right) = \begin{pmatrix} \mathbf{0}_{p' \times p'} & \mathbf{0}_{p' \times (p-p')} \\ \mathbf{0}_{(p-p') \times p'} & v_{22,-1,\theta}\left(g, \hat{\theta}\right) \end{pmatrix}.$$

Here $\mathbf{0}_{r \times s}$ denotes the $r \times s$-matrix with all entries equal to zero. Furthermore, the matrix $v_{22,-1,\theta}\left(g, \hat{\theta}\right) \in \mathbb{R}^{(p-p') \times (p-p')}$ can be minimized, and small Δ-optimality is achieved by

in addition minimizing the upper left $p' \times p'$-block $v_{11,0,\theta}\left(g, \hat{\theta}\right)$ of $v_{0,\theta}\left(g, \hat{\theta}\right)$. A sufficient condition for small Δ-optimality is that

$$\partial_y g_0(x, x; \theta) = \begin{pmatrix} \mathbf{0}_{p' \times d} \\ \dot{b}_2^T(x;\theta) C^{-1}(x;\theta) \end{pmatrix}, \tag{5.29}$$

$$\partial_{yy}^2 g_{1,0}(x, x; \theta) = \dot{C}_1^T(x;\theta) \left(C^{\otimes 2}(x;\theta)\right)^{-1}, \tag{5.30}$$

where $\dot{b}_2 \in \mathbb{R}^{d \times (p-p')}$ comprises the last $p - p'$ columns of $\dot{b}$, $g_{1,0}$ the first p' coordinates of g_0, and $\dot{C}_1 \in \mathbb{R}^{d^2 \times p'}$ the first p' columns of $\dot{C}$.

The complicated case (iii) may best be understood as follows: For $\theta_1, \ldots, \theta_{p'}$ fixed, (5.29) requires in particular that the last $p - p'$ coordinates of g be small Δ-optimal for estimating $\theta_{p'+1}, \ldots, \theta_p$, see case (i). And for $\theta_{p'+1}, \ldots, \theta_p$ fixed, (5.29) and (5.30) require that the first p' coordinates of g be small Δ-optimal for estimating $\theta_1, \ldots, \theta_{p'}$, see case (ii).

As mentioned above, to check for small Δ-optimality more is required than just checking (5.27), (5.28) or (5.29), (5.30), viz. it must be verified that various matrices involving expectations of quantities related to $\dot{b}$, $\dot{C}$, $\partial_y g_0$ and $\partial_{yy}^2 g_0$ are non-singular, see Theorem 2 in Jacobsen (2001a).

We used the special structure (5.24) above to get directly an expression for the limit $g_{i,0}(x, y; \theta)$ in (5.25). For a general martingale estimating function, the existence of a non-trivial (in particular non-zero) limit must be assumed, and to find it in concrete cases, it may be necessary to renormalize g, i.e. replace $g(\Delta, x, y; \theta)$ by $K_\Delta(\theta) g(\Delta, x, y; \theta)$ for some non-singular $p \times p$-matrix $K_\Delta(\theta)$ not depending on x or y. As discussed earlier, such a renormalization does not affect the solutions to the estimating equations. Small Δ-optimality can be discussed also for any family of unbiased estimating functions defined by a class of functions $(g_\Delta)_{\Delta > 0}$. Essentially, each g_Δ must then be replaced by the $\widetilde{g}_\Delta$ defined by (2.25), which yields the martingale estimating function (2.26). For details, see Jacobsen (2001a), Section 6.

It is important to comment further on the qualitatively different forms that the expansion (5.26) takes under small Δ-optimality in the three cases (i), (ii) and (iii). Obviously, a major gain in estimation accuracy is obtained for Δ small, if the leading term v_{-1} can be dispensed with, and the reason why this is possible in case (ii), partly in case (iii) and never in case (i) is best understood by considering complete observation of X in continuous time on a finite time interval – as $\Delta \to 0$ we are getting close to continuous time observation. So let $T > 0$ be fixed and denote by $P_{\theta,T}$ the distribution of $(X_t)_{0 \le t \le T}$ when X is stationary and the true parameter value is θ. In case (i), when θ varies, only the drift $b(x;\theta)$ changes and for $\theta \neq \theta'$ the measures $P_{\theta,T}$ and $P_{\theta',T}$ will typically be equivalent with a Radon-Nikodym derivative given by Girsanov's theorem. By contrast, in case (ii) where also $C(x;\theta)$ changes with θ, it may well happen that $P_{\theta,T}$ and $P_{\theta',T}$ are singular for $\theta \neq \theta'$, i.e. it is (in principle) possible to read off the exact value of θ from the observed sample path of X. Of course, for the discrete time observations $(X_{i\Delta})_{0 \le i \le n}$ perfect information about θ is not available, but through small Δ-optimality it is possible to increase the information about θ per observation $X_{i\Delta}$ from $O(\Delta)$ in case (i) to $O(1)$ in case (ii). Note that for the general martingale estimating functions, even in case (ii) the leading term v_{-1} will be present unless one is careful, and the result will then be an estimator that as $\Delta \to 0$ is infinitely worse than a small Δ-optimal estimator.

We shall now again return to the specific martingale estimating functions emanating from (5.24) and discuss when and how, for a given base $f = (f_j)_{1 \le j \le N}$ satisfying Condition 5.7, the

weights a may be chosen so as to achieve small Δ-optimality. In particular this will reveal a critical value

$$\dim(d) := d + \left(d + \binom{d}{2}\right) = d(d+3)/2$$

for the dimension N of the base. The value $\dim(d)$ comes about naturally by fixing a base f of dimension d and then supplementing this with the functions $f_j f_{j'}$ for $1 \leq j \leq j' \leq d$. The discussion splits into the same three cases as before, but for illustration we just consider case (i). From (5.25),

$$\partial_y g_0(x, x; \theta) = a(x; \theta)\, \partial_x f(x)$$

which is required to equal $\dot{b}^T(x; \theta)\, C^{-1}(x)$, see (5.27). Solving for $a(x; \theta)$ is clearly possible if $N = d$ provided the $d \times d$-matrix $\partial_x f(x)$ with jk'th element $\partial_{x_k} f_j(x)$ is non-singular, and possible also if $N > d$ provided $\partial_x f(x)$ has full rank d. In cases (ii) and (iii) similar linear equation systems are obtained (but now involving d first derivatives of g_0 and all the different second derivatives, i.e. $\dim(d)$ derivatives in all), resulting in the following shortened version of Theorem 2 of Jacobsen (2002). In the theorem we use the following notation: If $M \in \mathbb{R}^{r \times d^2}$ is a matrix with entries $M_{q,k\ell}$ for $1 \leq q \leq r$ and $1 \leq k, \ell \leq d$ that are symmetric in k and ℓ, we write $MR \in \mathbb{R}^{r \times \rho(d)}$ for the matrix with entries $M_{q,k\ell}$ for $1 \leq q \leq r$ and $1 \leq k \leq \ell \leq d$ obtained by multiplying M by the reduction matrix $R \in \mathbb{R}^{d^2 \times \rho(d)}$ with entries $R_{k'\ell',k\ell} = 1$ if $k' = k$ and $\ell' = \ell$ and $R_{k'\ell',k\ell} = 0$ otherwise ($1 \leq k', \ell' \leq d$ and $1 \leq k \leq \ell \leq d$). Here $\rho(d)$ is the number of choices for (k, ℓ) such that $1 \leq k \leq \ell \leq d$, i.e. $\rho(d) = d + \binom{d}{2} = \dim(d) - d$.

Theorem 5.8 *Consider martingale estimating functions of the form*

$$G_n(\theta) = \sum_{i=1}^{n} a^*(X_{(i-1)\Delta}, \theta) \left(f(X_{i\Delta}; \theta) - \pi_\Delta^\theta(f(\theta))(X_{(i-1)\Delta}) \right), \tag{5.31}$$

where the base $f = (f_j)_{1 \leq j \leq N}$ is of full affine rank N, and where the matrix-valued function $a^*(x, \theta)$ is chosen differently in the following three cases.

(i) Suppose that $N = d$, that for μ_θ-a.a. x the matrix $\partial_x f(x) \in \mathbb{R}^{d \times d}$ is non-singular, and that the p d-variate functions of x forming the columns of $\dot{b}(x; \theta)$ are linearly independent. Then the rows of

$$a^*(x; \theta) = \dot{b}^T(x; \theta) C^{-1}(x) \left(\partial_x f(x)\right)^{-1}, \tag{5.32}$$

are linearly independent as required by Condition 5.7, and the estimating function (5.31) satisfies the small Δ-optimality condition (5.27).

(ii) Suppose that $N = \dim(d)$, that for μ_θ-a.a. x, the matrix

$$Q(x) = \begin{pmatrix} \partial_x f(x) & \partial_{xx}^2 f(x) R \end{pmatrix} \in \mathbb{R}^{\dim(d) \times \dim(d)} \tag{5.33}$$

is non-singular and that the p d^2-variate functions of x forming the columns of $\dot{C}(x; \theta)$ are linearly independent. Then the rows of

$$a^*(x; \theta) = \begin{pmatrix} \mathbf{0}_{p \times d} & \dot{C}^T(x; \theta) \left(C^{\otimes 2}(x; \theta)\right)^{-1} R \end{pmatrix} (Q(x))^{-1}, \tag{5.34}$$

are linearly independent, and the estimating function (5.31) satisfies the small Δ-optimality condition (5.28).

(iii) Suppose that $N = \dim(d)$, that for μ_θ-a.a. x the matrix $Q(x)$ given by (5.33) is non-singular, that the $p - p'$ d-variate functions forming the columns of $\dot{b}_{2,\theta}$ are linearly independent, and that the p' d^2-variate functions forming the columns of $\dot{C}_{1,\theta}$ are linearly independent. Then the rows of

$$a^*(x;\theta) = \begin{pmatrix} \mathbf{0}_{p' \times d} & \dot{C}_1^T(x;\theta) \left(C^{\otimes 2}(x;\theta)\right)^{-1} R \\ \dot{b}_2^T(x;\theta) C^{-1}(x;\theta) & \mathbf{0}_{(p-p') \times \rho(d)} \end{pmatrix} (Q(x))^{-1}, \tag{5.35}$$

are linearly independent, and the estimating function (5.31) satisfies the small Δ-optimality conditions (5.29), (5.30).

For models with a special structure, the critical value $\dim(d)$ for the dimension of the base f may be lowered. This is, for instance, the case when $X = (X_1, \ldots, X_c)$ with $X_1, \ldots, X_c$ independent diffusions of dimensions $d_1, \ldots, d_c$ where $\sum_{m=1}^c d_m = d$. In this situation small Δ-optimality can be achieved using a base of dimension $\sum_{m=1}^c \dim(d_m)$. In general however, $\dim(d)$ is the critical dimension, even for the optimal martingale estimating function determined by a given base for any given $\Delta > 0$ to be small Δ-optimal (Jacobsen (2002), Theorem 2.3). Thus it may well happen if $d = 1$ for a model belonging to case (ii), that the optimal martingale estimating function determined by a base of dimension 1, will result in an estimator that behaves disastrously for high frequency data.

In case (iii) of Theorem 5.8 one may find a host of small Δ-optimal martingale estimating functions other than that specified by (5.35), in fact the entry $\mathbf{0}_{(p-p') \times \rho(d)}$ may be replaced by an arbitrary matrix depending on x and θ (subject to Condition 5.7 and smoothness and integrability requirements). Another useful recipe (adopted in Example 5.9 below) for finding small Δ-optimal estimating functions in case (iii), is to fix a base f° of dimension d, augment it to a base f of dimension $\dim(d)$ by adding the products $f_j^\circ f_{j'}^\circ$ for $1 \leq j \leq j' \leq d$, and then defining the first p' rows of $a^*(x;\theta)$ by

$$\begin{pmatrix} \mathbf{0}_{p' \times d} & \dot{C}_1^T(x;\theta) \left(C^{\otimes 2}(x;\theta)\right)^{-1} R \end{pmatrix} (Q(x))^{-1}$$

and the last $p - p'$ rows by

$$\dot{b}_2^T(x;\theta) C^{-1}(x;\theta) \left(\partial_x f(x)\right)^{-1}.$$

While it is easy to obtain small Δ-optimality for martingale estimating functions, it is not known what happens in general with the classes of simple and explicit, transition dependent estimating functions also discussed above, see (3.23) and (3.25). It is known (Jacobsen (2001a)) that for $d = 1$ and if $C = \sigma^2$ does not depend on θ, then the simple estimating function with h given by (3.23) is small Δ-optimal provided that f satisfies that $\partial_x f(x) = K_\theta \dot{b}_\theta^T(x)/\sigma^2(x)$ for some non-singular matrix K_θ not depending on x This is the case for Kessler's estimating function in the Ornstein-Uhlenbeck model, see Example 3.5, and, as follows easily using (2.16), also for H. Sørensen's estimating function mentioned at the end of Subsection 3.3 and obtained for $f = \partial_\theta \log \mu_\theta$. However, if either $d \geq 2$ or the model is of type (ii) or (iii), it seems virtually impossible to achieve small Δ-optimality. For the much wider class (3.25) nothing much is known, but it does appear difficult to obtain small Δ-optimality for models belonging to case (ii).

We shall conclude this section by showing how small Δ-optimality works for a one-dimensional diffusion model with four parameters. The model was introduced by Jacobsen (2002) and the simulation study below is from Jacobsen (2001b).

Example 5.9 Consider the one-dimensional ($d = 1$) stochastic differential equation

$$dX_t = \left(aX_t^{2\gamma-1} + bX_t\right)\,dt + \sigma X_t^{\gamma}\,dW_t \tag{5.36}$$

where $a, b \in \mathbb{R}$, $\gamma \neq 1$ and $\sigma > 0$ and where, as usual, W denotes a standard Wiener process. For $\gamma = \frac{1}{2}$ this is the stochastic differential equation for the CIR-process, see Example 3.1. The generalization (5.36) is arrived at by considering all powers $\widetilde{X}^{\rho}$ of a CIR-process with $\rho \neq 0$. More precisely if X solves (5.36), then the associated CIR–process is $\widetilde{X} = X^{2-2\gamma}$ solving

$$d\widetilde{X}_t = \left(\widetilde{a} + \widetilde{b}\widetilde{X}_t\right)\,dt + \widetilde{\sigma}\sqrt{\widetilde{X}_t}\,dW_t \tag{5.37}$$

where

$$\widetilde{b} = (2-2\gamma)\,b, \quad \widetilde{\sigma}^2 = (2-2\gamma)^2\,\sigma^2, \quad \widetilde{a} - \tfrac{1}{2}\widetilde{\sigma}^2 = (2-2\gamma)\left(a - \tfrac{1}{2}\sigma^2\right). \tag{5.38}$$

This also explains why $\gamma = 1$ is not allowed in (5.36).

Because of the connection to the CIR–process, the *generalized Cox-Ingersoll-Ross model* (or generalized CIR-process) described by (5.36) is much simpler to handle mathematically than the more standard CKLS–model (5.5) introduced in Chan et al. (1992) as a model for the spot rate. In particular, for (5.36) it is easy to find martingale estimating functions of the form (5.24) (although the base will now depend on the parameter γ).

In (5.36) the parameter space has dimension $p = 4$. We shall want X to be strictly positive and ergodic, which happens if and only if the associated CIR-process $\widetilde{X}$ is strictly positive and ergodic, i.e. when $\widetilde{b} < 0$ and $2\widetilde{a} \geq \widetilde{\sigma}^2$, or equivalently, when either $\gamma < 1$, $b < 0$, $2a \geq \sigma^2$ or $\gamma > 1$, $b > 0$, $2a \leq \sigma^2$. Since the invariant distribution for $\widetilde{X}$ is a $\Gamma-$distribution, the invariant distribution for X is that of a $\Gamma-$distributed random variable raised to the power $(2-2\gamma)^{-1}$.

Because a $\Gamma-$distribution has finite moments of all orders $m \in \mathbb{N}$, we have $\mathrm{E}_\theta(X_0^{(2\gamma-2)m}) < \infty$ for all $m \in \mathbb{N}$ when $X_0 \sim \mu_\theta$, and $\pi_\Delta^\theta x^{(2\gamma-2)m} < \infty$ for all $\Delta > 0$, $m \in \mathbb{N}$, and all $x > 0$. Furthermore, since the conditional moments for a CIR-process are known, for the generalized CIR-process all $\pi_\Delta^\theta x^{(2\gamma-2)m}$ are known explicitly.

Turning now to the problem of estimating θ from discrete observations of X, it is clear that the model (5.36) belongs to case (iii) with $p = 4$, $p' = 2$. We need a base of dimension $\dim(1) = 2$ and shall simply use $f = (f_j)_{1\leq j\leq 2}$ given by

$$f_1(x) = x^{2-2\gamma} \quad \text{and} \quad f_2(x) = x^{4-4\gamma}, \tag{5.39}$$

which trivially satisfies Condition 5.7. This corresponds to choosing $f^\circ(x) = x^{2-2\gamma}$ (see page 51 for the general use of f°). By the methods described above one may then show that the martingale estimating function given by

$$g\,(\Delta, x, y; \theta) = \begin{pmatrix} -2\log x & x^{2\gamma-2}\log x \\ -2 & x^{2\gamma-2} \\ x^{2\gamma-2} & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y^{2-2\gamma} - \pi_\Delta^\theta x^{2-2\gamma} \\ y^{4-4\gamma} - \pi_\Delta^\theta x^{4-4\gamma} \end{pmatrix} \tag{5.40}$$

is small Δ-optimal. Here the conditional expectations are given by the expressions

$$\begin{aligned} \pi_\Delta^\theta x^{2-2\gamma} &= e^{\tilde{b}\Delta}\left(x^{2-2\gamma} - \tilde{\xi}_1\right) + \tilde{\xi}_1, \\ \pi_\Delta^\theta x^{4-4\gamma} &= e^{2\tilde{b}\Delta}\left(x^{4-4\gamma} - \tilde{\xi}_2 - 2(\tilde{\xi}_2/\tilde{\xi}_1)\left(x^{2-2\gamma} - \tilde{\xi}_1\right)\right) + 2(\tilde{\xi}_2/\tilde{\xi}_1)e^{\tilde{b}\Delta}\left(x^{2-2\gamma} - \tilde{\xi}_1\right) + \tilde{\xi}_2 \end{aligned}$$

where $\tilde{\xi}_1 = E_\theta\left(X_0^{2-2\gamma}\right) = -\tilde{a}/\tilde{b}$ and $\tilde{\xi}_2 = E_\theta\left(X_0^{4-4\gamma}\right) = \tilde{a}/(2\tilde{b}^2)\,(2\tilde{a}+\tilde{\sigma}^2)$ with $\tilde{a}, \tilde{b}, \tilde{\sigma}^2$ as in (5.38).

The results of a simulation study using this small Δ-optimal estimating function are given in Table 5.3. Note that in agreement with the theory, the estimators of a and b deteriorate for Δ small, while the estimators of γ and σ^2 perform well throughout.

Δ	success		mean	std. dev.	smallest	largest
0.01	50/50	a	1.77	0.864	0.737	4.51
		b	-1.88	0.872	-4.84	-0.612
		γ	0.493	0.054	0.396	0.641
		σ^2	1.00	0.073	0.806	1.17
0.1	50/50	a	1.04	0.207	0.685	1.62
		b	-1.08	0.262	-2.01	-0.662
		γ	0.494	0.050	0.393	0.571
		σ^2	1.00	0.086	0.786	1.18
0.5	45/50	a	1.22	0.335	0.597	1.92
		b	-1.22	0.308	-1.93	-0.674
		γ	0.545	0.081	0.361	0.680
		σ^2	0.995	0.087	0.730	1.24

Table 5.3: The result of a simulation study using the estimating function given by (5.40) with the parameter values $a = 1$, $b = -1$, $\gamma = \frac{1}{2}$, and $\sigma^2 = 1$. Simulations were done for the indicated values of Δ based on $n + 1 = 501$ observations. For each value of Δ, 50 data sets were simulated. The column labeled "success" indicates the proportion of data sets for which estimates for all four parameters were obtained. The mean value and the standard deviation of these estimates are given in the table. The columns labeled "smallest" and "largest" indicate the range of the estimates obtained.

□

5.4 Optimal Prediction Based Estimating Functions

We shall now return to the prediction-based estimating functions that were introduced for inference about non-Markovian models. We shall find the Godambe optimal choice of the weights $\Pi_j^{(i-1)}(\theta)$, $j = 1, \ldots, N$ for a class of prediction-based estimating functions of the general type (3.34). Since these estimating functions are not martingales, Heyde optimality does not apply.

Again the compact representation

$$\begin{aligned} G_n(\theta) &= A(\theta) \sum_{i=s+1}^{n} Z^{(i-1)} \left(F(X_i) - \breve{\pi}^{(i-1)}(\theta)\right) \\ &= A(\theta) \sum_{i=s+1}^{n} Z^{(i-1)} \left(F(X_i) - (Z^{(i-1)})^T \breve{\alpha}(\theta)\right) \end{aligned} \tag{5.41}$$

is convenient. Here $F(x) = (f_1(x), \ldots, f_N(x))^T$, $\breve{\pi}^{(i-1)}(\theta) = (\breve{\pi}_1^{(i-1)}(\theta), \ldots, \breve{\pi}_N^{(i-1)}(\theta))^T$, while the $p \times N(q+1)$-matrix $A(\theta)$, the $N(q+1) \times N$-matrix $Z^{(i-1)}$, and the $N(q+1)$-dimensional vector $\breve{\alpha}(\theta)$ are given by (3.44), (3.45), and (3.46), respectively.

We are free to choose the matrix $A(\theta)$ in an optimal way, whereas the $N(q+1)$-dimensional vectors $Z^{(i-1)}(F(X_i) - \breve{\pi}^{(i-1)}(\theta))$ are fixed by our earlier choice of the functions f_j and h_{jk}. The matrix $A(\theta)$ must have rank p in order that we can estimate all p parameters.

To find the optimal prediction-based estimating function, we impose the following condition.

Condition 5.10
(1) *The quantities $C_j(\theta)$ and $b_j(\theta)$, $j = 1, \ldots, N$ are differentiable functions of θ. Here $C_j(\theta)$ denotes the covariance matrix of $Z_j^{(s)}$, while $b_j(\theta)$ is given by (3.32).*

(2) $p \leq N(q+1)$.

(3) *The $N(q+1) \times p$-matrix $\partial_{\theta^T} \breve{\alpha}(\theta)$ has rank p.*

(4) *The functions $1, f_1, \ldots, f_N$ are linearly independent on the support of the conditional distribution of X_n given $(X_1, \ldots, X_{n-1})$.*

The sensitivity function is given by

$$S_{G_n}(\theta) = -(n-q)A(\theta)D(\theta)\partial_{\theta^T}\breve{\alpha}(\theta)$$

where the $N(q+1) \times N(q+1)$-matrix $D(\theta)$ is given by

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

If we denote the optimal choice of the matrix $A(\theta)$ by $A_n^*(\theta)$, then

$$\mathrm{E}_\theta \left(G_n(\theta) G_n^*(\theta)^T \right) = (n-q) A(\theta) \bar{M}_n(\theta) A_n^*(\theta)^T,$$

where $\bar{M}_n(\theta)$ is given by (3.48). It is the covariance matrix of $\sum_{i=s+1}^n H^{(i)}(\theta)/\sqrt{n-s}$ with

$$H^{(i)}(\theta) = Z^{(i-1)} \left(F(X_i) - \breve{\pi}^{(i-1)}(\theta) \right).$$

Under Condition 5.10 the matrix $\bar{M}_n(\theta)$ is invertible, see Sørensen (2000), so it follows from Theorem 4.1 that for

$$A_n^*(\theta) = \partial_\theta \breve{\alpha}(\theta)^T D(\theta) \bar{M}_n(\theta)^{-1}, \tag{5.42}$$

the estimating function

$$G_n^*(\theta) = A_n^*(\theta) \sum_{i=s+1}^n Z^{(i-1)} \left(F(X_i) - \breve{\pi}^{(i-1)}(\theta) \right), \tag{5.43}$$

is Godambe optimal and satisfies the second Bartlett identity. For the optimal estimating function, the covariance matrix of the asymptotic distribution is the inverse of $S(\theta_0) = \partial_\theta \breve{\alpha}(\theta_0)^T D(\theta_0) M(\theta_0)^{-1} D(\theta_0) \partial_{\theta^T} \breve{\alpha}(\theta_0)$.

An estimator with the same asymptotic variance as the estimator obtained from (5.43) is obtained if $A_n^*(\theta)$ is replaced by $A_n^*(\tilde{\theta}_n)$, where $\tilde{\theta}_n$ is some consistent preliminary estimator.

This modification of (5.43) is highly recommended, because the calculation of $A_n^*(\theta)$ usually requires very considerable simulation, so that a dramatic reduction of computing time can be achieved by calculating it at only one parameter value. The estimator $\tilde{\theta}_n$ can, for instance, be obtained from (5.41) with $A(\theta)$ equal to some simple matrix that does not depend on θ.

Note that when $p = N(q+1)$ the matrix $A_n^*(\theta)$ is invertible. Thus it does not influence the estimator and can be omitted. If we know $C_j(\theta)$, $b_j(\theta)$, $\mathrm{E}_\theta(Z_j^{(q)})$, $\mathrm{E}_\theta(f_j(X_1))$, $j = 1, \ldots, N$, their derivatives with respect to θ, and the moments appearing in (3.48), we can calculate the optimal prediction-based estimating function. Note that only moments and derivatives of moments are needed. Note also that $C_j(\theta)$, $b_j(\theta)$, $E_\theta(Z_j^{(q)})$, and $E_\theta(f_j(X_1))$ were needed earlier to find the predictor $\breve{\pi}_j^{(i-1)}(\theta)$, so the only new requirements to compute the optimal estimating function are the derivatives and the moments in (3.48). Many models are sufficiently mixing that there exist $K > 0$ and $\lambda > 0$ such that the absolute values of all entries in the expectation matrices in the kth term in the sum in (3.48) are dominated by $K\mathrm{e}^{-\lambda(k-q)}$ when $k > q$. Therefore, the sum in (3.48) can in practice often be truncated so that only a few moments need to be calculated.

When $A(\theta_0) = \partial_\theta \breve{\alpha}(\theta_0)^T D(\theta_0) M(\theta_0)^{-1}$ (the optimal choice), the covariance matrix of the asymptotic distribution is the inverse of $S(\theta_0) = \partial_\theta \breve{\alpha}(\theta_0)^T D(\theta_0) M(\theta_0)^{-1} D(\theta_0) \partial_{\theta^T} \breve{\alpha}(\theta_0)$, see (3.49).

Example 5.11 If we want to find the optimal estimating function of the form (3.38) for the stochastic volatility model (3.26), we must assume that $\mathrm{E}_\theta(X_i^8) < \infty$, and apart from the quantities mentioned above, we need to find $\mathrm{E}_\theta\left(X_i^2 X_j^2 X_1^2\right)$ and $\mathrm{E}_\theta\left(X_i^2 X_j^2 X_k^2 X_1^2\right)$ for $i \geq j \geq k$. We can essentially find these moments by integrating the moments $\mathrm{E}_\theta\left(v_s v_t v_u\right)$ and $\mathrm{E}_\theta\left(v_s v_t v_u v_z\right)$ of the volatility process as functions of s, t, u, and z over suitable sets. For details see Sørensen (2000).

The moments of the volatility process must in general be found by simulation, but can in some cases be found explicitly. This is for instance the case when the volatility process is the CIR-process

$$dv_t = -\theta(v_t - \alpha)dt + \sigma\sqrt{v_t}dB_t, \tag{5.44}$$

because for the CIR-process there are explicit expressions for all conditional moments, see e.g. Sørensen (2000). This stochastic volatility model was proposed by Hull & White (1988) and Heston (1993).

Another example of a stochastic volatility model, for which the necessary moments can be found explicitly, is when $v_t = \exp(U_t)$, where U is a stationary Gaussian Ornstein-Uhlenbeck process,

$$dU_t = -\theta(U_t - \alpha)dt + \sigma dB_t$$

with $\theta > 0$ (Wiggins (1987); Chesney & Scott (1989); Melino & Turnbull (1990)). This model can be obtained as a limit of the EGARCH(1,1) model, see Nelson (1990). Here we have for instance that

$$\mathrm{E}_\theta\left(v_s v_t v_u v_z\right) = \mathrm{E}_{\theta,\alpha,\sigma}\left\{\exp\left(U_s + U_t + U_u + U_z\right)\right\},$$

which is the Laplace transform of a known Gaussian distribution, and hence is explicitly known.

If one is not prepared to assume that $\mathrm{E}_\theta(X_i^8) < \infty$, a possible alternative to the estimating

function (3.38) is

$$G_n(\theta) = \sum_{i=q+1}^{n} \Pi^{(i-1)}(\theta) \left\{ |X_i|^\gamma - \breve{a}_0(\theta) - \breve{a}_1(\theta)|X_{i-1}|^\gamma - \cdots - \breve{a}_q(\theta)|X_{i-q}|^\gamma \right\}$$

with $\Pi^{(i-1)}(\theta) = A(\theta)\tilde{Z}^{(i-1)}$ and $\tilde{Z}^{(i-1)} = (1, |X_{i-1}|^\gamma, \ldots, |X_{i-q}|^\gamma)^T$. Here $A(\theta)$ is a $p \times (q+1)$-matrix to be chosen in an optimal way, while γ is some suitably chosen positive real number. If, for instance, $\gamma = \frac{1}{2}$, we need to assume only that $\mathrm{E}_\theta(X_i^2) < \infty$ for the optimal $A(\theta)$ to exist. The price is that it is not as easy to calculate the moments needed as for (3.38).

□

Acknowledgements

The research of Martin Jacobsen and Michael Sørensen was supported by MaPhySto – a Network in Mathematical Physics and Stochastics funded by The Danish National Research Foundation, and by the European Commission through the Research Training Network DYNSTOCH under the Human Potential Programme. Michael Sørensen was moreover supported by the Centre for Analytical Finance and the Danish Mathematical Finance Network, both financed by the Danish Social Science Research Council. The data were put at our disposal by the Centre for Analytical Finance.

References

Aït-Sahalia, Y. (2002). "Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-form Approximation Approach". *Econometrica*, 70:223–262.

Aït-Sahalia, Y. (2003). "Closed-form likelihood expansions for multivariate diffusions". Working paper, Princeton University.

Aït-Sahalia, Y.; Hansen, L. P. & Scheinkman, J. (2003). "Operator methods for continuous-time Markov models". In Aït-Sahalia, Y. & Hansen, L. P., editors, *Handbook of Financial Econometrics*. Amsterdam: North-Holland. Forthcoming.

Aït-Sahalia, Y. & Mykland, P. A. (2003). "The effect of random and discrete sampling when estimating continuous-time diffusions". *Econometrica*, 71:483–549.

Andersen, T. G.; Benzoni, L. & Lund, J. (2002). "An empirical investigation of continuous-time models for equity returns". *Journal of Finance*, 57:1239–1284.

Baddeley, A. J. (2000). "Time-invariance estimating equations". *Bernoulli*, 6:783–808.

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.

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. (1997). "A Hyperbolic Diffusion Model for Stock Prices". *Finance and Stochastics*, 1:25–41.

Bibby, B. M. & Sørensen, M. (2001). "Simplified Estimating Functions for Diffusion Models with a High-dimensional Parameter". *Scand. J. Statist.*, 28(1):99–112.

Billingsley, P. (1961a). "The Lindeberg-Lévy theorem for martingales". *Proc. Amer. Math. Soc.*, 12:788–792.

Billingsley, P. (1961b). *Statistical Inference for Markov Processes*. The University of Chicago Press, Chicago.

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

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.

Chesney, M. & Scott, L. (1989). "Pricing European currency options: A comparison of the modified Black-Scholes model and a random variance model". *Journal of Financial Quantitative Analalysis*, 24:267–289.

Christensen, B. J.; Poulsen, R. & Sørensen, M. (2001). "Optimal inference for diffusion processes with applications to the short rate of interest". Working Paper No. 102, Centre for Analytical Finance, University of Aarhus.

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

Coddington, E. A. & Levinson, N. (1955). *Theory of Ordinary Differential Equations*. McGraw-Hill, New York.

Cox, J. C.; Ingersoll, Jr., J. E. & Ross, S. A. (1985). "A Theory of the Term Structure of Interest Rates". *Econometrica*, 53(2):385–407.

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, S. & Sørensen, M. (2002). "Inference for observations of integrated diffusion processes.". Preprint No. 2, Department of Theoretical Statistics, University of Copenhagen. To appear in *Scand. J. Statist.*

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.

Duffie, D. & Singleton, K. (1993). "Simulated moments estimation of Markov models of asset prices". *Econometrica*, 61:929–952.

Duffie, D. J. & Kan, R. (1996). "A yield factor model of interest rates". *Math. Finance*, 6:379–406.

Durbin, J. (1960). "Estimation of parameters in time-series regression models". *J. Roy. Statist. Soc. Ser. 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.

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". *Journal of Business and Economic Statistics*, 19:177–191.

Fisher, R. A. (1935). "The logic of inductive inference". *J. Roy. Statist. Soc.*, 98:39–54.

Florens-Zmirou, D. (1984). "Théorème de limite centrale pour une diffusion et pout sa discrétisée". *C. R. Acad. Sc. Paris, Série I*, 299:995–998.

Florens-Zmirou, D. (1989). "Approximate discrete-time schemes for statistics of diffusion processes". *Statistics*, 20:547–557.

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

Gallant, A. R. & Tauchen, G. (1996). "Which Moments to Match?". *Econometric Theory*, 12:657–681.

Gallant, A. R. & Tauchen, G. (2003). "Simulated score methods and indirect inference for continuous-time models". In Aït-Sahalia, Y. & Hansen, L. P., editors, *Handbook of Financial Econometrics*. Amsterdam: North-Holland. Forthcoming.

Genon-Catalot, V.; Jeantheau, T. & Larédo, C. (1999). "Parameter estimation for discretely observed stochastic volatility models". *Bernoulli*, 5:855–872.

Genon-Catalot, V.; Jeantheau, T. & Larédo, C. (2000). "Stochastic Volatility Models as Hidden Markov Models and Statistical Applications". *Bernoulli*, 6:1051–1079.

Godambe, V. P. (1960). "An optimum property of regular maximum likelihood estimation". *Annals of Mathematical Statistics*, 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.; Monfort, A. & Renault, E. (1993). "Indirect inference". *Journal of Applied Econometrics*, 8:S85–S118.

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. & Scheinkman, J. A. (1995). "Back to the Future: Generating Moment Implications for Continuous-time Markov Processes". *Econometrica*, 63:767–804.

Heston, S. L. (1993). "A Closed-form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options". *Review of Financial Studies*, 6:327–343.

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.

Hull, J. & White, A. (1988). "An Analysis of the Bias in Option Pricing Caused by a Stochastic Volatility". *Advances in Futures and Options Research*, 3:29–61.

Jacobsen, M. (2001a). "Discretely Observed Diffusions; Classes of Estimating Functions and Small Δ-optimality". *Scand. J. Statist.*, 28:123–150.

Jacobsen, M. (2001b). "Small Δ-optimal martingale estimating functions for discretely observed diffusions: A simulation study". Preprint No. 2, Department of Theoretical Statistics, University of Copenhagen.

Jacobsen, M. (2002). "Optimality and small Δ-optimality of martingale estimating functions". *Bernoulli*, 8:643–668.

Johannes, M. & Polson, N. (2003). "MCMC methods for financial econometrics". In Aït-Sahalia, Y. & Hansen, L. P., editors, *Handbook of Financial Econometrics*. Amsterdam: North-Holland. Forthcoming.

Kessler, M. (1996). *Estimation paramétrique des coefficients d'une diffusion ergodique à partir d'observations discrètes*. PhD thesis, Laboratoire de Probabilités, Université Paris VI.

Kessler, M. (1997). "Estimation of an Ergodic Diffusion from Discrete Observations". *Scand. J. Statist.*, 24:211–229.

Kessler, M. (2000). "Simple and Explicit Estimating Functions for a Discretely Observed Diffusion Process". *Scand. J. Statist.*, 27:65–82.

Kessler, M. & Paredes, S. (2002). "Computational aspects related to martingale estimating functions for a discretely observed diffusion". *Scand. J. Statist.*, 29:425–440.

Kessler, M.; Schick, A. & Wefelmeyer, W. (2001). "The information in the marginal law of a Markov chain". *Bernoulli*, 7:243–266.

Kessler, M. & Sørensen, M. (2002). "On time-reversibility and estimating functions for Markov processes". Preprint No. 7, Department of Theoretical Statistics, University of Copenhagen. To appear in *Statistical Inference for Stochastic Processes*.

Kessler, M. & Sørensen, M. (1999). "Estimating Equations Based on Eigenfunctions for a Discretely Observed Diffusion Process". *Bernoulli*, 5:299–314.

Kim, S.; Shephard, N. & Chib, S. (1998). "Stochastic volatility: Likelihood inference and comparison with ARCH models". *Review of Economic Studies*, 65:361–393.

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.

Küchler, U. & Sørensen, M. (1999). "A note on limit theorems for multivariate martingales". *Bernoulli*, 5:483–493.

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. (2003). "A diffusion model for exchange rates in a target zone". Preprint No. 6, Department of Applied Mathematics and Statistics, University of Copenhagen.

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.

Melino, A. & Turnbull, S. M. (1990). "Pricing foreign currency options with stochastic volatility". *Journal of Econometrics*, 45:239–265.

Nelson, D. B. (1990). "ARCH models as diffusion approximations". *Journal of Econometrics*, 45:7–38.

Overbeck, L. & Rydén, T. (1997). "Estimation in the Cox-Ingersoll-Ross model". *Econometric Theory*, 13:430–461.

Pedersen, A. R. (1994a). "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. (1994b). "Uniform residuals for discretely observed diffusion processes". Research Report No. 292, Department of Theoretical Statistics, 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.

Pedersen, A. R. (2000). "Estimating the nitrous oxide emission rate from the soil surface by means of a diffusion model". *Scand. J. Statist.*, 27:385–403.

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. (1983). "Asymptotic theory for non-linear least squares estimator for diffusion processes". *Math. Operationsforsch. u. Statist., Ser. Statist.*, 14:195–209.

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.

Sørensen, H. (2001). "Discretely observed diffusions: Approximation of the continuous-time score function". *Scand. J. Statist.*, 28:113–121.

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

Sørensen, M. (1997). "Estimating Functions for Discretely Observed Diffusions: A Review". In Basawa, I. V.; Godambe, V. P. & Taylor, R. L., editors, *Selected Proceedings of the Symposium on Estimating Functions*, pages 305–325. Hayward: Institute of Mathematical Statistics. IMS Lecture Notes – Monograph Series, Vol. 32.

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

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

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.

Wefelmeyer, W. (1996). "Quasi-likelihood models and optimal inference". *Ann. Statist.*, 24:405–422.

Wefelmeyer, W. (1997). "Quasi-likelihood regression models for Markov chains". In Basawa, I. V.; Godambe, V. P. & Taylor, R. L., editors, *Selected Proceedings of the Symposium on Estimating Functions*, pages 149–173. Hayward: Institute of Mathematical Statistics. IMS Lecture Notes – Monograph Series, Vol. 32.

Wiggins, J. B. (1987). "Option values under stochastic volatility: Theory and empirical estimates". *Journal of Financial Economics*, 19:351–372.