On asymptotics of estimating functions.

Michael Sørensen
Department of Theoretical Statistics, University of Copenhagen
and MaPhySto.*

Abstract

The asymptotic theory of estimators obtained from estimating functions is reviewed and some new results on the multivariate parameter case are presented. Specifically, results about existence of consistent estimators and about asymptotic normality of these are given. First a very general stochastic process setting is considered. Then it is demonstrated how more specific conditions for existence of $\sqrt{n}$-consistent and asymptotically normal estimators can be given for martingale estimating functions in the case of observations of a Markov process.

Key words: asymptotic normality, consistency, diffusion processes, estimating equations, likelihood inference, Markov processes, martingale estimating functions, misspecified models, statistical inference for stochastic processes, quasi likelihood.

*MaPhySto - Centre for Mathematical Physics and Stochastics, funded by a grant from The Danish National Research Foundation

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. An estimator is obtained by equating the estimating function to zero and solving the resulting equation with respect to the parameter. The estimating function approach has for example turned out to be very useful in obtaining estimators for discretely observed diffusion models, where the likelihood function is usually not explicitly known; see Bibby and Sørensen (1995, 1996, 1997 and 1998), Kessler and Sørensen (1998), and Sørensen (1997b). Also maximum likelihood estimators are obviously covered by the theory as they are obtained when the score function is used as estimating function.

In Section 2, we give results about existence of consistent estimators and about asymptotic normality of these in a very general stochastic process setting. In Section 3, we consider the special case of ergodic Markov processes and estimating functions that are martingales for a certain parameter value. Under these restrictions, simpler and more specific conditions can be given for existence of $\sqrt{n}$-consistent and asymptotically normal estimators.

Before the general theory is presented, let us give some examples of estimating functions.

Example 1.1 *Quasi likelihood for independent observations.* Consider n independent observations $X_1, \ldots, X_n$ with p-dimensional explanatory covariates $t_1, \ldots, t_n$. Suppose we only want to specify the variance function $V(\mu)$, and that the mean value of X_i, μ_i, is related to the covariates by an invertible link function g in the following way: $g(\mu_i) = \sum_{j=1}^p \beta_j t_{ij}$, where the β_j's are unknown parameters about which we want to draw inference, and where t_{ij} denotes the jth coordinate of t_i. In this situation it was proposed by Wedderburn (1974) to use the estimating function $G_n(\beta)$ with the kth coordinate given by

$$G_n(\beta)_k = \sum_{i=1}^n \frac{t_{ik}}{V(\mu_i(\beta))g'(\mu_i(\beta))}[X_i - \mu_i(\beta)]. \tag{1.1}$$

Here $\mu_i(\beta) = g^{-1}(\sum_{j=1}^p \beta_j t_{ij})$. An estimator of β is obtained by solving the equation $G_n(\beta) = 0$. This is, in fact, the maximum likelihood estimator for the corresponding generalized linear model. The estimating function (1.1) is optimal among all estimating functions of the form $\sum_{i=1}^n a_i(\beta)[X_i - \mu_i(\beta)]$ in the sense of Godambe and Heyde (1987). Obviously, $G_n(\beta)$ is a martingale with respect to the natural filtration provided that $\mu_i(\beta)$ is the true mean value of X_i for all $i = 1, \ldots$. The natural filtration $(\mathcal{F}_n)$ is given by $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Thus the estimating function (1.1) satisfies Condition 3.3 in Section 3.

□

Example 1.2 Next consider the *autoregression of order one* defined by

$$X_i = \theta X_{i-1} + \epsilon_i,\ \ X_0 = x_0,$$

where the ϵ_i's are independent, identically distributed random variables with $E(\epsilon_i) = 0$. We will not specify the model further, except that we will assume that the second moment of ϵ_i exists. Let us assume that we have the data $X_1, X_2, \ldots, X_n$. Since

$$E_\theta(X_i \,|\, X_{i-1} = x) = \theta x,$$

it seems natural to find an estimator for θ by minimizing

$$K_n(\theta) = \sum_{i=1}^{n} (X_i - \theta X_{i-1})^2.$$

This least squares estimator can be found by using the estimating function

$$G_n(\theta) = \sum_{i=1}^{n} X_{i-1}(X_i - \theta X_{i-1}).$$

The resulting estimator is

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

This is the maximum likelihood estimator when the ϵ_i's are Gaussian. Again it is not difficult to see that $G_n(\theta)$ is a martingale with respect to the natural filtration, provided that θ is the true parameter value.

□

Example 1.3 Consider a *Galton-Watson branching process,* where X_i denotes the size of the ith generation. The model is defined as follows. The initial population size $X_0 = x_0$ is assumed to be given. The size of the ith population is given by

$$X_i = \sum_{j=1}^{X_{i-1}} Y_{ij},$$

where the Y_{ij}'s are independent, identically distributed random variables. Thus, Y_{ij} is the number of offspring that the jth individual in the ith generation gets. We will not specify the distribution of Y_{ij} (the offspring distribution) completely, but only denote its mean value by θ, and suppose that it has finite variance.

Suppose we wish to draw inference about θ on the basis of the generation size data $X_1, \ldots, X_n$. Then a possible estimating function is

$$G_n(\theta) = \sum_{i=1}^{n} (X_i - \theta X_{i-1}),$$

which is a martingale if θ is the true mean value of the offspring distribution. The corresponding estimator

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

is the maximum likelihood estimator when the offspring distribution is assumed to belong to an exponential family, see e.g. Küchler and Sørensen (1997, p.23).

□

Example 1.4 *Discrete time observations from Markov processes.* Consider statistical inference for a class of Markov processes (possibly with continuous time). We denote the observed process by X, and suppose that the distribution of X depends on an unknown p-dimensional parameter θ that varies in the set $\Theta \subseteq \mathbb{R}^p$. For simplicity, we let the initial value be fixed: $X_0 = x_0$. It is, moreover, assumed that the distribution of X_t given $X_s = x,\ t > s$, has a strictly positive density with respect to a dominating measure on the state space E. We denote the density by

$$y \mapsto p(t-s, x, y; \theta) > 0,\ y \in E,\ \theta \in \Theta.$$

Suppose we have data of the form $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$, where $0 < t_1 < \ldots < t_n$.

Ideally, we would base the statistical inference on the likelihood function. Since we have assumed that the observed process X is a Markov process, the likelihood function is

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

where $\Delta_i = t_i - t_{i-1}$ (with $t_0 = 0$). Unfortunately, the transition density p is not known explicitly for many Markovian continuous time models. This is an important motivation for the study of the theory of estimating function that provides an alternative inference method. In fact, inference based on optimal estimating function can be thought of as an approximation to likelihood inference.

The score function is

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

We use the notation $\partial_\theta f(\theta)$ for the column vector of partial derivatives of a function f. The estimating function $U_n(\theta)$ is of the form studied in Section 3, and standard arguments show, under regularity conditions allowing the interchange of differentiation and integration, that $U_n(\theta)$ is a martingale with respect to the natural filtration when θ is the true parameter value; see e.g. Barndorff-Nielsen and Sørensen (1994). Thus Condition 3.3 is usually satisfied for likelihood inference.

□

2 A general theory of asymptotics for estimating functions

In this section we consider a general set-up with a probability space on which a parametrized family of probability measures is given:

$$(\Omega, \mathcal{F}, \{P_\theta : \theta \in \Theta\}),\ \Theta \subseteq \mathbb{R}^p.$$

The data $X_1, X_2, \ldots, X_n$ are just assumed to be observations from some stochastic process defined on $(\Omega, \mathcal{F}, P_0)$. Note that the true probability measure P_0 under which the data have been generated might not be included in the parametric statistical model $\{P_\theta : \theta \in \Theta\}$.

An *estimating function* is a p-dimensional function of the parameter θ and the data:

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

Usually we suppress the dependence on the observations in the notation. We get an estimator by solving $G_n(\theta) = 0$. There might be more than one solution or no solution at all. We shall give conditions under which an estimator exists and is consistent and asymptotically normal.

We will use the following notation. For a $d \times d$-matrix $A = \{a_{ij}\}$, we use the norm

$$\|A\|^2 = \sum_{i,j=1}^{d} a_{i,j}^2 = \mathrm{tr}\,(AA^T).$$

If A is positive semi-definite, we denote by $A^{\frac{1}{2}}$ the positive semi-definite square root of A. For a vector a, the Euclidean norm of a is denoted by $\|a\|$. The reader is reminded that a class $\mathcal{R}$ of random variables is called *stochastically bounded* if for every $\epsilon > 0$ there exists a $K_\epsilon > 0$ such that

$$\sup_{X \in \mathcal{R}} P(\,\|X\| > K_\epsilon) < \epsilon.$$

If $\{X_n\}$ converges in distribution as $n \to \infty$, then it is stochastically bounded.

Condition 2.1 *$G_n(\theta)$ is continuously differentiable with respect to θ.*

Under this condition we can define the $p \times p$-matrix

$$J_n(\theta) = \partial_{\theta^T} G_n(\theta). \tag{2.1}$$

By this expression we mean that the ith row of the matrix consists of the partial derivatives with respect to θ of the ith coordinate of G_n, i.e. the ith row equals $(\partial_\theta G_n(\theta)_i)^T$. We denote transposition of a vector or a matrix by T. For $\theta^{(i)} \in \Theta,\ i = 1, \ldots, p$, we define a second $p \times p$-matrix by

$$J(\theta^{(1)}, \ldots, \theta^{(p)}) = \begin{pmatrix} \partial_{\theta^T} G(\theta^{(1)})_1 \\ \vdots \\ \partial_{\theta^T} G(\theta^{(p)})_p \end{pmatrix}. \tag{2.2}$$

Condition 2.2 *Suppose that there exists a parameter value $\bar\theta \in \mathrm{int}\,\Theta$ such that*

$$\sup_{\theta^{(i)} \in M_n^{(\alpha)}(\bar\theta)} \|K_n(\bar\theta)^T J_n(\theta^{(1)}, \ldots, \theta^{(p)}) K_n(\bar\theta) - W(\bar\theta)\| \to 0 \tag{2.3}$$

in P_0-probability on

$$C(\bar\theta) = \{\det\,(W(\bar\theta)) > 0\}$$

for all $\alpha > 0$ as $n \to \infty$. Here $\{K_n(\bar\theta) : n \in \mathbb{N}\}$ is a sequence of non-random invertible $p \times p$-matrices satisfying

$$K_n(\bar\theta) \to 0 \quad as \quad n \to \infty,$$

$W(\bar\theta)$ is a (possibly) random symmetric positive semi-definite matrix, and

$$M_n^{(\alpha)}(\bar\theta) = \{\theta \in \Theta : \|K_n(\bar\theta)^{-1}(\theta - \bar\theta)\| \le \alpha\}.$$

Assume, moreover, that the class of random variables

$$\{K_n(\bar\theta)^T G_n(\bar\theta) : n \in \mathbb{N}\}$$

is stochastically bounded.

Theorem 2.3 *Suppose the Conditions 2.1 and 2.2 hold. Then for every n, an estimator $\hat\theta_n$ exists on $C(\bar\theta)$ that solves the estimating equation $G_n(\hat\theta_n) = 0$ with a probability tending to $P_0(C(\bar\theta))$ as $n \to \infty$. Moreover,*

$$\hat\theta_n \xrightarrow{p} \bar\theta$$

on $C(\bar\theta)$ as $n \to \infty$, and

$$K_n(\bar\theta)^T J_n(a_n^{(1)}, \ldots, a_n^{(p)}) K_n(\bar\theta) \xrightarrow{p} W(\bar\theta) \tag{2.4}$$

on $C(\bar\theta)$, where for each n, $a_n^{(i)}$ is a convex combination of $\bar\theta$ and $\hat\theta_n$ $(i = 1, \ldots, p)$.

Remark: The last result in the theorem is a technicality needed later in order to prove asymptotic normality of the estimator. The following proof is inspired by a proof in Sweeting (1980), see also Barndorff-Nielsen and Sørensen (1994).

Proof: First note that (2.3) implies that

$$\sup_{\theta^{(i)} \in S_n^{(\alpha)}(\bar\theta)} \|K_n(\bar\theta)^T J_n(\theta^{(1)}, \ldots, \theta^{(p)}) K_n(\bar\theta) - W(\bar\theta)\| \xrightarrow{p} 0 \tag{2.5}$$

on $C(\bar\theta)$ as $n \to \infty$ for all $\alpha > 0$ (under the true probability measure P_0, of course). Here

$$S_n^{(\alpha)} = \{\theta \in \Theta : \|W(\bar\theta)^{\frac{1}{2}} K_n(\bar\theta)^{-1}(\theta - \bar\theta)\| \le \alpha\}.$$

If no solution of $G_n(\theta) = 0$ exists, we set $\hat\theta_n = \infty$. Otherwise, choose a solution in the smallest $S_n^{(m)}$, $m = 1, \ldots$, in which there is a solution.

Fix $\epsilon > 0$. Since the class of random variables $\{K_n(\bar\theta)^T G_n(\bar\theta) : n \in \mathbb{N}\}$ is stochastically bounded, so is $\{W(\bar\theta)^{-\frac{1}{2}} K_n(\bar\theta)^T G_n(\bar\theta) 1_{C(\bar\theta)} : n \in \mathbb{N}\}$. Hence, we can find a constant $K > 0$ such that for all $n \in \mathbb{N}$, the event

$$A_n = \{\|W(\bar\theta)^{-\frac{1}{2}} K_n(\bar\theta)^T G_n(\bar\theta)\| \le K\} \cap C(\bar\theta)$$

has P_0-probability larger than $P_0(C(\bar\theta)) - \epsilon$.

Next, fix $\delta > 0$, and choose $c > 0$ large enough that $c^2 - Kc - \delta > 0$. Consider the Taylor expansion

$$G_n(\theta) = G_n(\bar\theta) + J_n(\theta^{(1)}, \ldots, \theta^{(p)})(\theta - \bar\theta),$$

where for each i, $\theta^{(i)}$ is a convex combination of θ and $\bar{\theta}$. We will consider this expansion only when $\theta \in \operatorname{bd} S_n^{(c)}$, i.e. when $\|W(\bar{\theta})^{\frac{1}{2}} K_n(\bar{\theta})^{-1}(\theta - \bar{\theta})\| = c$. This implies that $\theta^{(i)} \in S_n^{(c)}$, $i = 1, \ldots, p$. It follows from the expansion that for $\theta \in \operatorname{bd} S_n^{(c)}$ we have that on $C(\bar{\theta})$

$$\begin{aligned}
&(\theta - \bar{\theta})^T G_n(\theta) \\
&\quad = (\theta - \bar{\theta})^T (K_n(\bar{\theta})^{-1})^T W(\bar{\theta})^{\frac{1}{2}} W(\bar{\theta})^{-\frac{1}{2}} K_n(\bar{\theta})^T G_n(\bar{\theta}) \\
&\qquad + \underbrace{(\theta - \bar{\theta})^T (K_n(\bar{\theta})^{-1})^T W(\bar{\theta}) K_n(\bar{\theta})^{-1} (\theta - \bar{\theta})}_{=c^2} + V_n(\theta),
\end{aligned}$$

where

$$\begin{aligned}
&V_n(\theta) = \\
&\quad (\theta - \bar{\theta})^T (K_n(\bar{\theta})^{-1})^T [K_n(\bar{\theta})^T J_n(\theta^{(1)}, \ldots, \theta^{(p)}) K_n(\bar{\theta}) - W(\bar{\theta})] K_n(\bar{\theta})^{-1} (\theta - \bar{\theta}).
\end{aligned}$$

Now define

$$U_n = \max_{\theta \in \operatorname{bd} S_n^{(c)}} |V_n(\theta)|$$

and

$$B_n = \{U_n < \delta\} \cap C(\bar{\theta}),$$

where the value of δ was fixed earlier. By (2.5) we can find $N \in \mathbb{N}$ such that

$$P_0(B_n) \geq P_0(C(\bar{\theta})) - \epsilon$$

for $n \geq N$. On $A_n \cap B_n$,

$$(\theta - \bar{\theta})^T G_n(\theta) \geq c^2 - Kc - \delta > 0$$

for all $\theta \in \operatorname{bd} S_n^{(c)}$, and since $G_n(\theta)$ is a continuous function of θ, the equation $G_n(\theta) = 0$ has a least one solution in $S_n^{(c)}$. This follows from Brouwer's fixed point theorem (see e.g. Aitchison and Silvey, 1958, Lemma 2, where also a related application to a statistical problem is given). In conclusion,

$$A_n \cap B_n \subseteq \{\hat{\theta}_n \in S_n^{(c)}\},$$

and since

$$P_0(A_n \cap B_n) \geq P_0(C(\bar{\theta})) - 2\epsilon$$

for $n \geq N$, we have proved that

$$P_0(\{G_n(\hat{\theta}_n) = 0\} \cap C(\bar{\theta})) \to P_0(C(\bar{\theta}))$$

as $n \to \infty$. Since $\hat{\theta}_n \in S_n^{(c)}$ on $A_n \cap B_n$, and because $S_n^{(c)}$ shrinks towards $\bar{\theta}$ as $n \to \infty$, it follows that $\hat{\theta}_n \xrightarrow{p} \bar{\theta}$ on $C(\bar{\theta})$ under P_0 as $n \to \infty$. Moreover, the result (2.4) follows from (2.5) by the same arguments.

□

For a class of *ergodic Markov processes* we can usually choose $K_n(\bar{\theta}) = I_p/\sqrt{n}$, where I_p denotes the $p \times p$ identity matrix, and the matrix $W(\bar{\theta})$ is non-random. Therefore, the assumptions made in Theorem 2.3 can be considerably simplified as formulated in the following condition and corollary.

Condition 2.4 *Suppose that there exists a $\bar{\theta} \in \operatorname{int}\Theta$ and a non-random symmetric positive definite matrix $W(\bar{\theta})$ such that*

$$\sup_{\theta^{(i)} \in M_n^{(\alpha)}(\bar{\theta})} \left\| \frac{1}{n} J_n(\theta^{(1)}, \ldots, \theta^{(p)}) - W(\bar{\theta}) \right\| \to 0$$

in probability as $n \to \infty$ for all $\alpha > 0$, where

$$M_n^{(\alpha)}(\bar{\theta}) = \{\theta \in \Theta : \|\theta - \bar{\theta}\| \leq \alpha/\sqrt{n}\}. \tag{2.6}$$

Assume, moreover, that the class of random variables

$$\{G_n(\bar{\theta})/\sqrt{n} : n \in \mathbb{N}\}$$

is stochastically bounded.

Corollary 2.5 *Suppose the Conditions 2.1 and 2.4 hold. 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,*

$$\hat{\theta}_n \to \bar{\theta}$$

in probability as $n \to \infty$, and

$$\frac{1}{n} J_n(a_n^{(1)}, \ldots, a_n^{(p)}) \to W(\bar{\theta}) \tag{2.7}$$

in probability as $n \to \infty$, where for each n, $a_n^{(i)}$ is a convex combination of $\bar{\theta}$ and $\hat{\theta}_n$ $(i = 1, \ldots, p)$.

For a one-dimensional parameter the condition that the matrix $W(\bar{\theta})$ is positive semi-definite just means that it is non-negative, which it can always be arranged to be. In the case of a multi-dimensional parameter, the condition is sometimes too strong. For estimating functions obtained from pseudo-likelihood functions or contrast functions by differentiation, the matrix $W(\bar{\theta})$ is positive semi-definite, and for optimal estimating functions this is often the case too; but otherwise $W(\bar{\theta})$ is rarely positive semi-definite. There are, however, ways around this problem.

Suppose the estimating function G_n satisfies Conditions 2.1 and 2.4 except that the matrix $W(\bar{\theta})$ is only assumed to be invertible (and not necessarily positive definite), and assume moreover that for every $\theta \in \Theta$ there exists a non-random, invertible matrix $W(\theta)$ such that $\partial_{\theta^T} G_n(\theta)/n \xrightarrow{p} W(\theta)$. Then it is easy to see that the estimating function $\widetilde{G}_n(\theta) = W(\theta)^{-1} G_n(\theta)$ has the same roots as $G_n(\theta)$ and that $\partial_{\theta^T} \widetilde{G}_n(\bar{\theta})/n \xrightarrow{p} I_p$, where I_p denotes the $p \times p$ identity matrix. However, extra conditions are needed on the original estimating function G_n to ensure that $\widetilde{G}_n$ satisfies Conditions 2.1 and 2.4. These extra conditions must, for instance, ensure that $W(\theta)$ is continuously differentiable and that $G_n(\theta)/n$ converges uniformly to zero on the shrinking balls $M_n^{(\alpha)}(\bar{\theta})$. We will go a slightly different way, which has the advantage that it can be generalized to the case of a random matrix $W(\theta)$. To simplify matters, we will only give a result that is relevant to ergodic processes. We impose the following condition.

Condition 2.6
(i) The mapping $\theta \mapsto G_n(\theta)$ *is twice continuously differentiable.*

(ii) There exist a $\bar{\theta} \in \operatorname{int}\Theta$ *and an invertible non-random* $p \times p$*-matrix* $A(\bar{\theta})$ *such that*

$$\sup_{\theta^{(i)} \in M_n^{(\alpha)}(\bar{\theta})} \left\| \frac{1}{n} J_n(\theta^{(1)}, \ldots, \theta^{(p)}) - A(\bar{\theta}) \right\| \to 0$$

in probability as $n \to \infty$ *for all* $\alpha > 0$*, where* $M_n^{(\alpha)}(\bar{\theta})$ *is given by (2.6).*

(iii) There exist p *non-random* $p \times p$*-matrices* $B_i(\bar{\theta})$*,* $i = 1, \ldots, p$*, such that*

$$\sup_{\theta^{(i)} \in M_n^{(\alpha)}(\bar{\theta})} \left\| \frac{1}{n} Q_n^{(i)}(\theta^{(1)}, \ldots, \theta^{(p)}) - B_i(\bar{\theta}) \right\| \to 0$$

in probability as $n \to \infty$ *for all* $\alpha > 0$ *and all* $i = 1, \ldots, p$*, where*

$$Q_n^{(i)}(\theta) = \partial_\theta^2 G_n(\theta)_i.$$

(iv) $\{G_n(\bar{\theta})/\sqrt{n} : n \in \mathbb{N}\}$ *is stochastically bounded.*

(v) $\sup_{\theta \in M_n^{(\alpha)}(\bar{\theta})} \|G_n(\theta)/n\| \to 0$ *in probability as* $n \to \infty$ *for all* $\alpha > 0$*.*

Remark: By $\partial_\theta^2 G_n(\theta)_i$ is meant the $p \times p$-matrix of second order derivatives of the function $G_n(\theta)_i$.

Corollary 2.7 *Suppose Condition 2.6 is satisfied. Then the conclusions of Corollary 2.5 hold with* $W(\bar{\theta})$ *replaced by* $A(\bar{\theta})$ *in (2.7).*

Proof: The contrast function

$$H_n(\theta) = G_n(\theta)^T G_n(\theta)$$

takes its minimal value if $G_n(\theta) = 0$. Consider the estimating function

$$R_n(\theta) = \frac{1}{2n} \partial_\theta H_n(\theta) = \frac{1}{n} \partial_\theta G_n(\theta)^T G_n(\theta).$$

We will check that $R_n(\theta)$ satisfies Condition 2.4 with $W(\bar{\theta}) = A(\bar{\theta})^T A(\bar{\theta})$, which is positive definite. Then we can apply Corollary 2.5 to $R_n(\theta)$.

First note that $\{R_n(\bar{\theta})/\sqrt{n} : n \in \mathbb{N}\}$ is stochastically bounded because $\{G_n(\bar{\theta})/\sqrt{n} : n \in \mathbb{N}\}$ is so and $\frac{1}{n}\partial_\theta G_n(\bar{\theta})^T \to A^T$. Then consider the equation

$$\frac{1}{n} \partial_{\theta^T} R_n(\theta) - A^T A = \frac{1}{n} \partial_\theta G_n(\theta)^T \frac{1}{n} \partial_{\theta^T} G_n(\theta) - A^T A + \sum_{k=1}^{p} \frac{1}{n} Q_n^{(k)}(\theta) \frac{1}{n} G_n(\theta)_k.$$

Simple evaluations show that

$$\sup_{\theta^{(i)} \in M_n^{(\alpha)}(\bar{\theta})} \left\| \frac{1}{n} \partial_{\theta^T} R_n(\theta^{(1)}, \ldots, \theta^{(p)}) - A^T A \right\| \leq [Y_n^{(1)} + 2\|A\|] Y_n^{(1)} + Y_n^{(3)} \left[Y_n^{(2)} + \sum_{k=1}^{p} \|B_k\| \right],$$

where

$$\begin{aligned}
Y_n^{(1)} &= \sup_{\theta^{(i)} \in M_n^{(\alpha)}(\bar{\theta})} \left\| \frac{1}{n} J_n(\theta^{(1)}, \ldots, \theta^{(p)}) - A(\bar{\theta}) \right\| \\
Y_n^{(2)} &= \max_i \sup_{\theta^{(i)} \in M_n^{(\alpha)}(\bar{\theta})} \left\| \frac{1}{n} Q_n^{(i)}(\theta^{(1)}, \ldots, \theta^{(p)}) - B_i(\bar{\theta}) \right\| \\
Y_n^{(3)} &= \sup_{\theta \in M_n^{(\alpha)}(\bar{\theta})} \| G_n(\theta)/n \| .
\end{aligned}$$

Hence, Corollary 2.5 ensures the existence of a sequence $\{\hat{\theta}_n\}$ such that $\hat{\theta}_n \to \bar{\theta}$ in probability as $n \to \infty$, for which

$$P_0(A_n) \to 1 \quad \text{as} \quad n \to \infty,$$

where

$$A_n = \{R_n(\hat{\theta}_n) = 0\}.$$

Finally, define the set

$$B_n = \{\partial_{\theta^T} G_n(\hat{\theta}_n) \ \text{ is invertible}\}.$$

Since $\partial_{\theta^T} G_n(\hat{\theta}_n)/n \to A$ in P_0-probability as $n \to \infty$, and A is invertible, it follows that

$$P_0(B_n) \to 1 \quad \text{as} \quad n \to \infty.$$

Remember that $R_n(\theta) = \frac{1}{n} \partial_\theta G_n(\theta)^T G_n(\theta)$. Therefore, $G_n(\hat{\theta}_n) = 0$ on $A_n \cap B_n$, and Corollary 2.7 follows because

$$P_0(A_n \cap B_n) \to 1 \quad \text{as} \quad n \to \infty.$$

□

We complete this section by giving a result on asymptotic normality of estimators obtained from estimating functions.

Theorem 2.8 *Suppose either that Conditions 2.1 and 2.4 hold or that Condition 2.6 holds. Assume, moreover, that*

$$\frac{1}{\sqrt{n}} G_n(\bar{\theta}) \xrightarrow{\mathcal{D}} N(0, V_0) \tag{2.8}$$

as $n \to \infty$. Then

$$\sqrt{n}(\hat{\theta}_n - \bar{\theta}) \xrightarrow{\mathcal{D}} N(0, A(\bar{\theta})^{-1} V_0 (A(\bar{\theta})^{-1})^T)$$

as $n \to \infty$. Under Condition 2.4, $A(\bar{\theta})$ should be replaced by $W(\bar{\theta})$.

Proof: Consider again

$$G_n(\hat{\theta}_n) = G_n(\bar{\theta}) + J_n(\alpha_n^{(1)}, \ldots, \alpha_n^{(p)})(\hat{\theta}_n - \bar{\theta}),$$

where each $\alpha_n^{(i)}$ is a convex combination of $\hat{\theta}_n$ and $\bar{\theta}$. By rearranging the terms, we get

$$[J_n(\alpha_n^{(1)}, \ldots, \alpha_n^{(p)})/n] \ \sqrt{n}(\hat{\theta}_n - \bar{\theta}) = -G_n(\bar{\theta})/\sqrt{n} + G_n(\hat{\theta}_n)/\sqrt{n},$$

and the theorem follows because of (2.7), (2.8), and

$$G_n(\hat{\theta}_n)/\sqrt{n} \xrightarrow{p} 0.$$

□

3 Discretely observed Markov processes

In this section, we use the general results proved in the previous section to give results for the case of observations from a Markov process. The basic setup is as in Section 2, but here we assume that we observe a stochastic process X which is Markovian with state space E under every P_θ, $\theta \in \Theta$, as well as under the true probability measure P_0. It does not matter whether the process has discrete time or continuous time, but the continuous time case is the more interesting, as there the likelihood functions is often not explicitly known so that it is difficult or impossible to apply inferential methods based on the likelihood function. In the case of a continuous time process, we assume that the observations are made at equidistant time points, i.e. that they are of the form $X_\Delta, X_{2\Delta}, \ldots, X_{n\Delta}$. We can denote the observations in the same way in the discrete time case, since we can take the time to be $\mathbb{N}$ and put $\Delta = 1$. For simplicity we suppose that $X_0 = x_0$ is fixed.

Let $y \mapsto p_0(\Delta, x, y)$ be the true transition density of the observed process X, i.e. the conditional density under P_0 of X_Δ given $X_0 = x$ with respect to a dominating measure ν on the state space E. We impose the following condition on the process X.

Condition 3.1 *Under P_0, the process X has a unique invariant measure which has the density $\mu_0(x)$ with respect to the measure ν on E. The transition distribution (i.e. the conditional distribution of X_Δ given $X_0 = x$) is, under P_0, absolutely continuous with respect to the invariant measure for all values of x.*

If $X_0 \sim \mu_0$, then X is stationary, and $X_t \sim \mu_0$ for all $t > 0$. Moreover,

$$(X_t, X_{t+\Delta}) \sim Q_0^\Delta$$

for all $t > 0$ and all $\Delta > 0$, where Q_0^Δ is the measure on E^2 with density

$$Q_0^\Delta(x, y) = \mu_0(x) p_0(\Delta, x, y)$$

with respect to ν^2. For a function $f : E^2 \mapsto \mathbb{R}$ we use the notation

$$Q_0^\Delta(f) = \int_{E^2} f(x, y) p_0(\Delta, x, y) \mu_0(x) \nu(dy) \nu(dx)$$

(provided, of course, that this makes sense). We shall need the following ergodic theorem (law of large numbers) and central limit theorem (for martingales).

Theorem 3.2 *Suppose Condition 3.1 holds, and that $f : E^2 \mapsto \mathbb{R}$ satisfies that $Q_0^\Delta(|f|) < \infty$. Then*

$$\frac{1}{n} \sum_{i=1}^{n} f(X_{(i-1)\Delta}, X_{i\Delta}) \xrightarrow{a.s.} Q_0^\Delta(f)$$

as $n \to \infty$ under P_0. Suppose further that $Q_0^\Delta(f^2) < \infty$, and that

$$\int_E f(x, y) p_0(\Delta, x, y) \nu(dy) = 0 \quad \textit{for all } x \in E.$$

Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(X_{(i-1)\Delta}, X_{i\Delta}) \xrightarrow{\mathcal{D}} N(0, Q_0^\Delta(f^2))$$

as $n \to \infty$ under P_0.

The first part of the theorem can be found in Billingsley (1961a), while the second part was proved in Billingsley (1961b). Note that the condition in the second part of the theorem implies that the sum $\sum_{i=1}^{n} f(X_{(i-1)_\Delta}, X_{i\Delta})$ is a square integrable martingale under P_0.

In the rest of this section, we will consider the asymptotic properties of estimators based on estimating functions of the form

$$G_n(\theta) = \sum_{i=1}^{n} g(\Delta, X_{(i-1)\Delta}, X_{i\Delta}; \theta), \tag{3.1}$$

where g is p-dimensional. We will not assume that G_n is a martingale estimating function, but rather the following weaker condition, which is useful when considering misspecified models.

Condition 3.3 *There exists a parameter value* $\bar{\theta} \in \operatorname{int}\Theta$ *such that*

$$\int_E g(\Delta, x, y, \bar{\theta}) p_0(\Delta, x, y) \nu(dy) = 0$$

for all $x \in E$.

This condition states that there exists a parameter value $\bar{\theta}$ such that $G_n(\bar{\theta})$ is a martingale under the true probability measure P_0. In the case of a model that is not misspecified, i.e. if there exists a $\theta_0 \in \Theta$ such that $P_{\theta_0} = P_0$, Theorem 3.6 below ensures the existence of a $\sqrt{n}$-consistent estimator provided that Condition 3.3 is satisfied with $\bar{\theta} = \theta_0$. This simply amounts to assuming that G_n is a martingale estimating function.

We will formulate two sets of further conditions. The main difference is that the first set includes conditions on the second order derivatives with respect to θ, while the second set involves only the first order derivatives, but includes the stronger assumption that the mean value of the matrix of these derivatives is strictly positive definite. This stronger condition is usually only satisfied for optimal estimating functions or for estimating functions obtained by maximizing a contrast function.

Condition 3.4
(1) The function g is twice continuously differentiable with respect to θ for all x, y.

(2) The functions

$$(x, y) \mapsto g_i(\Delta, x, y; \theta), \quad i = 1, \ldots, p,$$
$$(x, y) \mapsto \partial_{\theta_j} g_i(\Delta, x, y; \theta), \quad i, j = 1, \ldots, p,$$

and

$$(x, y) \mapsto \partial_{\theta_i} \partial_{\theta_j} g_k(\Delta, x, y; \theta), \quad i, j, k = 1, \ldots, p,$$

are all locally dominated integrable w.r.t. Q_0^Δ. Moreover, the functions $(x, y) \mapsto g_i(\Delta, x, y; \theta)$, $i = 1, \ldots, p$, are in $L_2(Q_0^\Delta)$ for all $\theta \in \Theta$.

(3) The $p \times p$ matrix

$$A(\bar{\theta}) = \left\{ Q_0^\Delta \left(\partial_{\theta_j} g_i(\Delta; \bar{\theta}) \right) \right\} \tag{3.2}$$

is invertible.

Condition 3.5
(1) The function g is continuously differentiable with respect to θ for all x, y.

(2) The functions $(x,y) \mapsto g_i(\Delta, x, y; \theta),\ i = 1, \ldots, p$, are in $L_2(Q_0^\Delta)$ for all $\theta \in \Theta$.

(3) The functions
$$(x,y) \mapsto \partial_{\theta_j} g_i(\Delta, x, y; \theta), \quad i, j = 1, \ldots, p,$$
are all locally dominated integrable with respect to Q_0^Δ.

(4) The $p \times p$ matrix $A(\bar{\theta})$ given by (3.2) is positive definite.

Theorem 3.6 *Suppose Conditions 3.1 and 3.3 are satisfied. Assume further that either Condition 3.4 or Condition 3.5 holds. 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,*
$$\hat{\theta}_n \xrightarrow{p} \bar{\theta}$$
as $n \to \infty$, and
$$\sqrt{n}(\hat{\theta}_n - \bar{\theta}) \xrightarrow{\mathcal{D}} N(0, A(\bar{\theta})^{-1} V_0(\bar{\theta}) (A(\bar{\theta})^{-1})^T),$$
where
$$V_0(\bar{\theta}) = Q_0^\Delta(g(\Delta, \bar{\theta}) g(\Delta, \bar{\theta})^T).$$

Remark: If $P_{\bar{\theta}} = P_0$, then $G_n(\theta)$ is an unbiased martingale estimating function, and $\hat{\theta}_n$ converges in probability to the true parameter value as $n \to \infty$.

Proof: Under Condition 3.3, it follows from Theorem 3.2 that
$$\frac{1}{\sqrt{n}} x^T G_n(\bar{\theta}) \xrightarrow{\mathcal{D}} N\left(0, x^T V_0(\bar{\theta}) x\right)$$
as $n \to \infty$ for every $x \in \mathbb{R}^p \backslash \{0\}$. Hence
$$\frac{1}{\sqrt{n}} G_n(\bar{\theta}) \xrightarrow{\mathcal{D}} N(0, V_0(\bar{\theta}))$$
as $n \to \infty$. Here we use the so-called Cramér-Wold device (consider characteristic functions).

Under Condition 3.4, the theorem follows from Corollary 2.7 and Theorem 2.8 if we can prove for all $\alpha > 0$ that as $n \to \infty$
$$\sup_{\theta \in M_n^{(\alpha)}(\bar{\theta})} |G_n(\theta)_i / n| \xrightarrow{p} 0$$
for $i = 1, \ldots, p$,
$$\sup_{\theta \in M_n^{(\alpha)}(\bar{\theta})} |n^{-1} \partial_{\theta_j} G_n(\theta)_i - A(\bar{\theta})_{ij}| \xrightarrow{p} 0$$

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

$$\sup_{\theta \in M_n^{(\alpha)}(\bar{\theta})} |n^{-1} \partial_{\theta_i} \partial_{\theta_j} G_n(\theta)_k - B^{(k)}(\bar{\theta})_{ij}| \xrightarrow{p} 0$$

for $i, j, k = 1, \ldots, p$, where $M_n^{(\alpha)}(\bar{\theta})$ is given by (2.6), and

$$B^{(k)}(\bar{\theta})_{ij} = Q_0^\Delta \left(\partial_{\theta_i} \partial_{\theta_j} g_k(\Delta; \bar{\theta}) \right).$$

Under Condition 3.5, we need only prove the second of these convergence result in order to deduce the theorem from Corollary 2.5 and Theorem 2.8.

The three convergence results are proved in exactly the same way, so we prove only the second. We will, in fact, prove almost sure convergence. Note that

$$\begin{aligned} & \sup_{\theta \in M_n^{(\alpha)}(\bar{\theta})} |n^{-1} \partial_{\theta_j} G_n(\theta)_i - A(\bar{\theta})_{ij}| \\ & \leq \sup_{\theta \in M_1^{(\alpha)}(\bar{\theta})} |n^{-1} \partial_{\theta_j} G_n(\theta)_i - A(\theta)_{ij}| + \sup_{\theta \in M_n^{(\alpha)}(\bar{\theta})} |A(\theta)_{ij} - A(\bar{\theta})_{ij}|, \end{aligned}$$

where

$$A(\theta)_{ij} = Q_0^\Delta \left(\partial_{\theta_j} g_i(\Delta; \theta) \right).$$

That each of the two terms on the right hand side tends to zero as $n \to \infty$ follows from the next lemma.

Lemma 3.7
a) $A(\theta)$ is continuous,

b) For every compact subset $K \subseteq \Theta$

$$\sup_{\theta \in K} |n^{-1} \partial_{\theta_j} G_n(\theta)_i - A(\theta)_{ij}| \xrightarrow{a.s.} 0$$

under P_0 as $n \to \infty$.

Proof: Define

$$k(\theta, \delta; x, y) = \sup_{\|\tilde{\theta} - \theta\| \leq \delta} \sum_{i,j} |\partial_{\theta_j} g_i(\Delta, x, y; \tilde{\theta}) - \partial_{\theta_j} g_i(\Delta, x, y; \theta)|.$$

By the dominated convergence theorem (using the local integrability of $\partial_{\theta_j} g_i$ with respect to Q_0^Δ)

$$\lim_{\delta \to 0} Q_0^\Delta(k(\theta, \delta)) = Q_0^\Delta(\lim_{\delta \to 0} k(\theta, \delta)) = 0.$$

Suppose $\theta_n \to \theta$. Then

$$\begin{aligned} \|A(\theta_n) - A(\theta)\| &= \left\| \{Q_0^\Delta(\partial_{\theta_j} g_i(\Delta; \theta_n)) - Q_0^\Delta(\partial_{\theta_j} g_i(\Delta; \theta))\} \right\| \\ &\leq \text{const} \sum_{ij} Q_0^\Delta(|\partial_{\theta_j} g_i(\Delta; \theta_n) - \partial_{\theta_j} g_i(\Delta; \theta)|) \\ &\leq \text{const } Q_0^\Delta(k(\theta, \delta_n)) \to 0, \end{aligned}$$

where $\delta_n = \|\theta_n - \theta\|$. Thus $A(\theta)$ is continuous.

Since $\partial_{\theta_j} g_i(\Delta, x, y; \theta)$ is locally dominated integrable with respect to Q_0^Δ, we can for every $\theta \in \Theta$ find a $\delta_\theta > 0$ such that

$$k(\theta, \delta; x, y) \in L_1(Q_0^\Delta) \quad \text{for} \quad 0 < \delta < \delta_\theta.$$

Fix $\epsilon > 0$. The function $A(\theta)$ is continuous, so for every $\theta \in \Theta$, we can find a $\lambda_\theta \in (0, \delta_\theta]$ such that

$$\|\tilde{\theta} - \theta\| < \lambda_\theta \Rightarrow |A(\tilde{\theta})_{ij} - A(\theta)_{ij}| < \tfrac{1}{2}\epsilon$$

and

$$Q_0^\Delta(k(\theta, \lambda_\theta)) < \tfrac{1}{2}\epsilon.$$

Let K be a compact subset of Θ. Then there exists a finite covering

$$K \subseteq \bigcup_{j=1}^{r} B(\theta_j, \lambda_{\theta_j}),$$

where $B(\theta, \lambda)$ is the open ball $B(\theta, \lambda) = \{\tilde{\theta} : \|\theta - \tilde{\theta}\| < \lambda\}$, and where $\theta_1, \ldots, \theta_r \in K$. For every $\theta \in K$, we can therefore choose θ_ℓ ($\ell \in \{1, \ldots, r\}$) such that

$$\|\theta - \theta_\ell\| < \lambda_{\theta_\ell}.$$

Then for $\theta \in K$,

$$\begin{aligned}
&|n^{-1}\partial_{\theta_j} G_n(\theta)_i - A(\theta)_{ij}| \\
&\leq \; |n^{-1}\partial_{\theta_j} G_n(\theta)_i - n^{-1}\partial_{\theta_j} G_n(\theta_\ell)_i| \\
&\quad + |n^{-1}\partial_{\theta_j} G_n(\theta_\ell)_i - A(\theta_\ell)_{ij}| + \underbrace{|A(\theta_\ell)_{ij} - A(\theta)_{ij}|}_{\leq \frac{1}{2}\epsilon} \\
&\leq \; \frac{1}{n}\sum_{\nu=1}^{n} |\partial_{\theta_j} g_i(\Delta, X_{\Delta(\nu-1)}, X_{\Delta\nu}; \theta) - \partial_{\theta_j} g_i(\Delta, X_{\Delta(\nu-1)}, X_{\Delta\nu}; \theta_\ell)| \\
&\quad + |n^{-1}\partial_{\theta_j} G_n(\theta_\ell)_i - A(\theta_\ell)_{ij}| + \epsilon/2 \\
&\leq \; \frac{1}{n}\sum_{\nu=1}^{n} k(\theta_\ell, \lambda_{\theta_\ell}; X_{\Delta(\nu-1)}, X_{\Delta\nu}) + |n^{-1}\partial_{\theta_j} G_n(\theta_\ell)_i - A(\theta_\ell)_{ij}| + \epsilon/2 \\
&\leq \; \left|\frac{1}{n}\sum_{\nu=1}^{n} k(\theta_\ell, \lambda_{\theta_\ell}; X_{\Delta(\nu-1)}, X_{\Delta\nu}) - Q_0^\Delta(k(\theta_\ell, \lambda_{\theta_\ell}))\right| \\
&\quad + \underbrace{|Q_0^\Delta(k(\theta_\ell, \lambda_{\theta_\ell}))|}_{\leq \epsilon/2} + |n^{-1}\partial_{\theta_j} G_n(\theta_\ell)_i - A(\theta_\ell)_{ij}| + \epsilon/2.
\end{aligned}$$

Hence,

$$\sup_{\theta \in K} |n^{-1}\partial_{\theta_j} G_n(\theta)_i - A(\theta)_{ij}|$$
$$\leq \max_{1\leq \ell \leq r} \left| \frac{1}{n} \sum_{\nu=1}^{n} k(\theta_\ell, \lambda_{\theta_\ell}; X_{\Delta(\nu-1)}, X_{\Delta\nu}) - Q_0^\Delta(k(\theta_\ell, \lambda_{\theta_\ell})) \right|$$
$$+ \max_{1\leq \ell \leq r} |n^{-1}\partial_{\theta_j} G_n(\theta_\ell)_i - A(\theta_\ell)_{ij}| + \epsilon,$$

so by the ergodic theorem (Theorem 3.2),

$$\lim_{n\to\infty} \sup_{\theta \in K} |n^{-1}\partial_{\theta_j} G_n(\theta)_i - A(\theta)_{ij}| \leq \epsilon$$

almost surely for all $\epsilon > 0$.

□

Note the importance of Condition 3.3 (or rather of the following weaker condition which it implies). If

$$\tilde{g}_\theta = Q_0^\Delta(g(\Delta;\theta)) = \int_E g(\Delta, x, y; \theta) p_0(\Delta, x, y) \mu_0(x) \nu(dy) \nu(dx) \neq 0,$$

then $G_n(\theta)/\sqrt{n}$ cannot be stochastically bounded under P_0 because by the ergodic theorem

$$\frac{1}{n} G_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(\Delta; X_{(i-1)\Delta}, X_{i\Delta}; \theta) \xrightarrow{a.s.} \tilde{g}_\theta$$

under P_0. We can therefore only hope for convergence of $\hat{\theta}_n$ to a parameter value $\bar{\theta}$ satisfying that $\tilde{g}_{\bar{\theta}} = 0$.

When $G_n(\bar{\theta})$ is not a martingale under P_0 (that is when Condition 3.3 is not satisfied), it is in many cases still possible to prove a result like Theorem 3.6 under the weaker condition that there exists a parameter value $\bar{\theta}$ such that $\tilde{g}_{\bar{\theta}} = 0$, provided that stronger regularity conditions are imposed on the process X. What is needed are conditions ensuring that the so-called potential of the Markov process is well-defined. The potential can be used to construct a martingale from $G_n(\bar{\theta})$ to which a martingale central limit theorem can be applied, see Florens-Zmirou (1984), Kessler (1996), and Jacobsen (1998). Unfortunately, the expression for the asymptotic variance becomes very difficult to calculate, but it can in many cases be estimated from the data; see Sørensen (1997a).

References

Aitchison, J. and Silvey, S.D. (1958): Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29**, 813–828.

Barndorff-Nielsen, O.E. and Sørensen, M. (1994): A review of some aspects of asymptotic likelihood theory for stochastic processes. *Int. Statist. Rev.* **62**, 133–165.

Bibby, B.M. and Sørensen, M. (1995): Martingale estimating functions for discretely observed diffusion processes. *Bernoulli* **1**, 17–39.

Bibby, B.M. and Sørensen, M. (1996): On estimation for discretely observed diffusions: A review. *Theory of Stochastic Processes* **2 (18)**, 49–56.

Bibby, B.M. and Sørensen, M. (1997): A hyperbolic diffusion model for stock prices. *Finance and Stochastics* **1**, 25–41.

Bibby, B.M. and Sørensen, M. (1998): Simplified estimating functions for diffusion models with a high-dimensional parameter. In preparation.

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

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

Florens-Zmirou, D. (1984): Théorème de limite centrale pour une diffusion et pour sa discrétisée. *C.A. Acad. Sci. Paris, Ser I* **299**, 995–998.

Godambe, V.P. (1960): An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208–1212.

Godambe, V.P. (1985): The foundations of finite sample estimation in stochastic processes. *Biometrika* **72**, 419–428.

Godambe, V.P. and Heyde, C.C. (1987): Quasi likelihood and optimal estimation. *Int. Statist. Rev.* **55**, 231–244.

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, New York.

Jacobsen, M. (1998): Discretely observed diffusions: Optimality and time reversibility of estimating functions. In preparation.

Kessler, M. (1996): Simple and explicit estimating functions for a discretely observed diffusion process. Research Report No. 336, Department of Theoretical Statistics, University of Aarhus.

Kessler, M. and Sørensen, M. (1998): Estimating equations based on eigenfunctions for a discretely observed diffusion process. Research Report No. 332, Department of Theoretical Statistics, University of Aarhus. To appear in *Bernoulli.*

Küchler, U. and Sørensen, M. (1997): *Exponential Families of Stochastic Processes.* Springer, New York.

Sørensen, M. (1997a): Exponential family inference for diffusions models. Research Report 383, Department of Theoretical Statistics, University of Aarhus.

Sørensen, M. (1997b): Estimating functions for discretely observed diffusions: A review. In Basawa, I.V., Godambe, V.P. and Taylor, R.L. (eds.): *Selected Proceedings of the Symposium on Estimating Functions.* IMS Lecture Notes - Monograph Series, Vol. 32, 305–325.

Sweeting, T.J. (1980): Uniform asymptotic normality of the maximum likelihood estimator. *Ann. Statist.* **8**, 1375–1381.

Wedderburn, R.W.M. (1974): Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.