Discretely observed diffusions: classes of estimating functions and small $\Delta-$optimality

Martin Jacobsen*
Department of Theoretical Statistics
University of Copenhagen
5 Universitetsparken
2100 Copenhagen Ø
Denmark

November 11, 1998

Abstract

Ergodic diffusions in several dimensions, depending on an unknown multivariate parameter are considered. For estimation, when the diffusion is observed only at finitely many equidistant timepoints, unbiased estimating functions leading to consistent and asymptotically Gaussian estimators are used. Different types of estimating functions are discussed and the concept of small $\Delta-$optimality is introduced to help select good estimating functions. Explicit criteria for small $\Delta-$optimality are given. Also some exact optimality conditions are presented as well as, for one-dimensional diffusions, methods for improving estimators using time reversibility.

KEYWORDS AND PHRASES: estimating functions of type: simple, martingale, transition dependent, reversible; asymptotic normality of estimators; minimizing asymptotic covariances.

* MaPhySto – Centre for Mathematical Physics and Stochastics, funded by a grant from the Danish National Research Foundation.

Contents

1. Introduction

The main purpose of this paper is to discuss criteria for choosing good estimating functions when estimating the parameters of an ergodic diffusion that is observed only at discrete points in time. While in principle one should of course use maximum-likelihood, in practice this is difficult because only in a few cases are the transition densities of the diffusion available in closed analytic form. Thus the existing methods for maximum-likelihood, see Pedersen [13] and Aït-Sahalia [1], rely on approximations of the transition densities and the methods prove quite computer intensive. (Pedersen uses numerical approximations based on iterations of the Gaussian transition densities emanating from the Euler scheme, while Aït-Sahalia, using a specific transformation of the diffusion, is able to obtain accurate theoretical approximations based on (transformed) Hermite function expansions).

Because of the difficulty in performing accurate maximum-likelihood, much research has focused on finding alternatives in the form of various types of unbiased estimating functions, but again here there may be practical problems because explicit analytic expressions are not always available.

The inspiration for the present paper came from following the work of Michael Sørensen and his colleagues, notably Mathieu Kessler, and was initiated by the desire for finding explicit estimating functions beyond the class of simple ones discussed by Kessler [9]. This led first to what proved merely a rediscovery of the class of explicit, transition dependent estimating functions presented by Hansen and Scheinkman [7], but eventually also to the study of optimality criteria presented below.

The basic model, involving only ergodic diffusions observed at equidistant timepoints distance Δ apart, and the basic assumptions together with the notation is presented in Section 2. In Section 3 we give an overview of the different types of estimating functions known from the literature, and in Section 4 survey the asymptotic theory (consistency, asymptotic normality) for the estimators obtained from the estimating functions. Section 6 treats the special results that arise from the time-reversibility of one-dimensional ergodic diffusions, while optimality is discussed in Sections 5 and 7. The main results are the conditions for *small* $\Delta-$*optimality* presented in Theorems 7.5 and 7.8: while exact optimality (i.e. minimizing the asymptotic covariance for an estimator within a given class) leads to results that are difficult to implement, see Section 5, it turns out that it is possible to give explicit, easily verifiable conditions on *flows* of estimating functions that ensure nearly efficient estimation if the discrete observations of the diffusion are close together. (A flow is simply a sufficiently smooth family of

estimating functions, one for each $\Delta > 0$). The conditions are flexible enough that there are many choices for small $\Delta-$optimal flows. A drawback is of course that a flow that performs well for small Δ, even though the estimator is always consistent and asymptotically Gaussian, may behave poorly for large Δ, and in the last part of Section 7 we make some preliminary suggestions on how to resolve this difficulty.

2. The model; setup and notation

Consider a statistical model for a $d-$dimensional diffusion process X, solving the stochastic differential equation

$$dX_t = b_\theta\left(X_t\right)\, dt + \sigma_\theta\left(X_t\right)\, dB_t \tag{2.1}$$

with some initial condition $X_0 = U$. Here B is a standard $d-$dimensional Brownian motion and b and σ are known functions, $b_\theta(x), \sigma_\theta(x)$ of an unknown $p-$dimensional parameter θ and $x \in \mathbb{R}^{d\times 1}$, b taking values in $\mathbb{R}^{d\times 1}$ and σ matrix-valued, $\sigma_\theta(x) \in \mathbb{R}^{d\times d}$.

Notation. Throughout the paper, column vectors are denoted x, θ etc., row vectors are written as transposed columns, x^T, θ^T etc. Indexes i, j are reserved for objects relating to the state space $D \subset \mathbb{R}^d$ of the diffusion, indexes k, l for objects relating to the parameter space $\Theta \subset \mathbb{R}^p$. We write x_i for the coordinates of x, θ_k for the coordinates of θ, but for processes or the functions b, σ, where time, respectively θ, appears as a subscript, use superscripts to designate the coordinates, $X^i, b^i_\theta, \sigma^{ij}_\theta$. Thus writing (2.1) coordinatewise,

$$dX^i_t = b^i_\theta\left(X_t\right)\, dt + \sum_{j=1}^{d} \sigma^{ij}_\theta\left(X_t\right)\, dB^j_t, \quad X^i_0 = U^i.$$

Formally, the diffusion X is defined on a filtered measurable space $(\Omega, \mathcal{F}, \mathcal{F}_t)$ with U $\mathcal{F}_0-$measurable. We shall assume that for any $\theta \in \Theta$, and any probability ν on $\mathbb{R}^d$, there is a probability measure P^ν_θ on $(\Omega, \mathcal{F})$ with respect to which the $\sigma-$algebra $\mathcal{F}_0$ and the Brownian motion B are independent, and such that for the prescribed $\theta-$value, (2.1) has a unique strong solution (in particular the law of the solution is unique) with ν the distribution of U. Subject to P^ν_θ, X is then a time-homogeneous diffusion with transition probabilities that do not depend on ν and depend on the diffusion coefficient σ_θ through

$$C_\theta := \sigma_\theta \sigma^T_\theta$$

only (with ^{T} denoting matrix transposition). Apart from these assumptions of a rather general nature, we shall need some more specific assumptions:

Assumption 1. The parameter θ belongs to an open subset $\Theta \subset \mathbb{R}^p$, and for each $\theta \in \Theta$, the diffusion X is ergodic, i.e. admits of a uniquely determined invariant distribution μ_θ: subject to $P_\theta^{\mu_\theta}$, X is strictly stationary. It is also assumed that the range of X does not depend on θ: there is an open connected subset $D \subset \mathbb{R}^d$ such that for all θ, ν,

$$P_\theta^\nu \bigcap_{t \geq 0} (X_t \in D) = 1,$$

and $\mu_\theta(D \cap O) > 0$ for all open O with the interior of $D \cap O$ non-empty.

Remark. The assumption about D not depending on θ is made of course for statistical reasons – the mere fact that an observation X_{t_0} belongs to some subset of D must not contain information about θ. We have assumed for convenience that D is open, but in principle one could allow for diffusions with accessible reflecting boundaries. In that case, typically (but not always), (2.1) is not sufficient to describe X and a complete description involves local time along the boundary. The assumption D connected is there simply to ensure that the invariant measure is uniquely determined. The reader is reminded that for $d = 1$, Assumption 1 amounts to the following when $\sigma_\theta(x) > 0$ for all x: $D =]l, r[$ for some $-\infty \leq l < r \leq \infty$ and if S_θ is a *scale function* with derivative (with respect to x)

$$S'_\theta(x) = \exp\left(-2 \int_{x_0}^x \frac{b_\theta(y)}{\sigma_\theta^2(y)}\, dy\right) \tag{2.2}$$

for some $x_0 \in D$, then $S_\theta(l) = -\infty, S_\theta(r) = \infty$; furthermore, $k_\theta := \int_l^r \kappa_\theta(x)\, dx < \infty$ where

$$\kappa_\theta = 2/\left(\sigma_\theta^2 S'_\theta\right) \tag{2.3}$$

is the density of the *speed measure* matching the scale S_θ and μ_θ has density $\kappa_\theta(x)/k_\theta$.

If $d \geq 2$ not too much is known about the range D of the solution to (2.1), although there are of course well known conditions ensuring that $D = \mathbb{R}^d$.

Notation. With ν a probability on D, we write $\nu(f)$ for $\int_D f\, d\nu$. The Lebesgue density of the invariant distribution of μ_θ will also be denoted μ_θ. If $\nu = \mu_\theta$ we write P_θ^μ instead of $P_\theta^{\mu_\theta}$ and if $\nu = \varepsilon_x$, simply P_θ^x. We shall write E_θ^ν for expectations with respect to P_θ^ν and $\pi_{t,\theta}$ for the transition operators for X when

θ is the parameter, $p_{t,\theta}(x,y)$ for the transition densities. Thus, for functions f integrable with respect to $p_{t,\theta}(x,y)\,dy$,

$$E_\theta^\nu\left(f(X_{s+t})\,|X_s = x\right) = \pi_{t,\theta}f(x) = \int_D dy\, p_{t,\theta}(x,y)f(y)$$

for all s, x, ν. Finally, $Q_{t,\theta}$ will denote the joint distribution of (X_s, X_{s+t}) under P_θ^μ (for any s) and $q_{t,\theta}$ its density,

$$q_{t,\theta}(x,y) = \mu_\theta(x)p_{t,\theta}(x,y).$$

Assumption 2. The following assumptions are made on the drift and diffusion coefficients b and σ from (2.1): for all $\theta \in \Theta$, $b_\theta(x), C_\theta(x)$ are continuous in x, and for each $x \in D$, $b_\theta(x), C_\theta(x)$ are continuously differentiable in θ. Also $C_\theta(x) > 0$ for all θ, x.

Notation. $C_\theta(x)$ is a symmetric, positive semidefinite $d \times d-$matrix. If A_1, A_2 are two symmetric matrices, we write $A_1 \geq A_2$ if $A_1 - A_2$ is positive semidefinite and $A_1 > 0$ if A_1 is symmetric and positive definite.

Throughout the paper the following notation is used for differentiation: if $a(r) = (a_1(r), \ldots, a_q(r)) \in \mathbb{R}^q$ is a differentiable function of a $q'-$dimensional variable $r = (r_1, \ldots, r_{q'})$, $\partial_r a(r)$ denotes the $q \times q'-$dimensional matrix of partial derivatives $(\partial_r a)_{ij} = \frac{\partial a_i}{\partial r_j}$ evaluated at r. Thus, if $q = 1$, $\partial_r a(r)$ is a $q'-$dimensional row vector, and for general q, the i'th row of $\partial_r a(r)$ is $\partial_r a_i(r)$. If differentiation is with respect to θ, we use a dot, $\dot{}$, instead of the symbol ∂. Thus (cf Assumption 2),

$$\dot{b}_\theta\,(x) \in \mathbb{R}^{d\times p}, \qquad \dot{b}_{\theta,ik}\,(x) = \frac{\partial b_\theta^i(x)}{\partial \theta_k}$$

$$\dot{C}_\theta\,(x) \in \mathbb{R}^{d^2\times p}, \qquad \dot{C}_{\theta,(ij)k}\,(x) = \frac{\partial C_\theta^{ij}(x)}{\partial \theta_k}.$$

Introduce the differential operator associated with the infinitesimal generator for the transition semigroup $(\pi_{t,\theta})_{t\geq 0}$ for the diffusion X,

$$A_\theta f(x) = \sum_{i=1}^d b_\theta^i(x)\frac{\partial}{\partial x_i}f(x) + \tfrac{1}{2}\sum_{i,j=1}^d C_\theta^{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j}f(x)$$

acting on the space $C^2(D)$ of functions $f : D \to \mathbb{R}$ that are twice continuously differentiable.

If $f \in C^2(D)$, by Itô's formula

$$df(X_t) = A_\theta f(X_t)\,dt + \partial_x f(X_t)\sigma_\theta(X_t)\,dB_t \tag{2.4}$$

(with the three 'factors' in the last term of dimensions $1 \times d, d \times d, d \times 1$ respectively). The last term defines a local martingale, which is a true martingale if

$$E_\theta^\mu \partial_x f(X_t) C_\theta(X_t) (\partial_x f)^T(X_t) = \int_D dx\, \mu_\theta(x) \partial_x f(x) C_\theta(x) (\partial_x f)^T(x) < \infty, \quad (2.5)$$

hence it follows that if $f \in L^1(\mu_\theta) \cap C^2(D), A_\theta f \in L^1(\mu_\theta)$ and (2.5) holds, then

$$\mu_\theta(A_\theta f) = 0, \quad (2.6)$$

which is the equation used to determine μ_θ.

Here we shall be concerned with a L^2-generator for X. Recall that $\pi_{t,\theta} : L^2(\mu_\theta) \to L^2(\mu_\theta)$, so we have a semigroup of operators on $L^2(\mu_\theta)$, and then define the domain $\mathcal{D}_\theta$ for the L^2-generator $\mathcal{A}_\theta$ as the subspace comprising all $f \in L^2(\mu_\theta) \cap C^2(D)$ for which $A_\theta f \in L^2(\mu_\theta)$ and (2.5) holds, and for such f, define $\mathcal{A}_\theta f = A_\theta f$. It is well known that $\mathcal{D}_\theta$ is dense in $L^2(\mu_\theta)$. Clearly for any θ, $\mathcal{D}_\theta$ contains the space $\mathcal{DC}$ consisting of the $f \in C^2(D)$ with compact support. More generally, $\mathcal{D}_\theta \supset \mathcal{DB}_\theta$, the space of $f \in C^2(D)$ such that f and $A_\theta f$ are bounded: to verify this, the problem is to show (2.5), i.e. that $E_\theta^\mu [f(X)]_t < \infty$, where $[f(X)]_t = \int_0^t ds\, \partial_x f(X_s) C_\theta(X_s) \partial_x^T f(X_s)$ is the quadratic variation of $f(X)$. Introducing the stopping times $\tau_K := \inf\{t : [f(X)]_t = K\}$, one finds that M_K and $\left(M_K^2 - [f(X)]_{\tau_K \wedge t}\right)_{t \geq 0}$ are martingales, where

$$M_{K,t} = f(X_{\tau_K \wedge t}) - f(X_0) - \int_0^{\tau_K \wedge t} ds\, A_\theta f(X_s).$$

Thus $E_\theta^\mu M_{K,t}^2 = E_\theta^\mu [f(X)]_{\tau_K \wedge t}$ and for $K \uparrow \infty$, by monotone convergence the expectation on the right increases to $E_\theta^\mu [f(X)]_t$ while, since $f, A_\theta f$ are bounded, dominated convergence applied to $E_\theta^\mu M_{K,t}^2$ yields a finite limit and (2.5) is proved.

Note that for $d = 1$, $\mathcal{DC}$ is dense in $L_0^2(\mu_\theta) = 1_\theta^\perp := \{f \in L^2(\mu_\theta) : \mu_\theta(f) = 0\}$, a fact that is definitely not true for $d > 1$: thus, to determine μ_θ if $d = 1$ one need only consider (2.6) for $f \in \mathcal{DC}$, but for $d > 1$ it is necessary to involve, say, all $f \in \mathcal{DB}_\theta$, which makes the determination of μ_θ more difficult. (A swift proof of the formula for $\mu_\theta(x)$ for $d = 1$ (see (2.3) and the next line) is the following: if $f \in \mathcal{DC}$, using partial integration on $\int A_\theta f\, d\mu_\theta = 0$ gives $\int \left(b_\theta \mu_\theta f' - \frac{1}{2}(\sigma_\theta^2 \mu_\theta)' f'\right) dx = 0$ and since f' satisfies $\int_D f'\, dx = 0$ and the collection of these derivatives is large enough, the density μ_θ must satisfy the differential equation $b_\theta \mu_\theta = \frac{1}{2}(\sigma_\theta^2 \mu_\theta)'$ from which the desired expression follows).

Notation. Write $\mathcal{D}_{0,\theta} = \mathcal{D}_\theta \cap 1_\theta^\perp$.

We shall need a final, more restrictive condition. Define

$$\mathrm{spec}(\mathcal{A}_\theta) = \{\lambda \in \mathbb{C} : \mathcal{A}_\theta f = \lambda f \text{ for some complex-valued } f \in \mathcal{D}_\theta\}.$$

Assumption 3. There exists $\lambda_0 < 0$ such that $\mathrm{Re}\,\lambda \leq \lambda_0$ for all $\lambda \in \mathrm{spec}(A_\theta)\backslash 0$.

If $\lambda \in \mathrm{spec}(A_\theta)$, then certainly $\mathrm{Re}\,\lambda \leq 0$ with $0 \in \mathrm{spec}(\mathcal{A}_\theta)$ always. With a *spectral gap* $\lambda_0 < 0$, we have for all $f \in 1_\theta^\perp$ that $\|\pi_{t,\theta} f\| \leq e^{\lambda_0 t}\|f\|$, where $\|\cdot\|$ is the norm on $L^2(\mu_\theta)$, and hence for *any* $f \in L^2_\theta(\mu_\theta)$, $\pi_{t,\theta} f$ converges in $L^2(\mu_\theta)$ at an exponential rate as $t \to \infty$ to the function which is constant and equal to $\mu_\theta(f)$. With Assumption 3 in force, it is therefore possible to define the *potential operator* $U_{t,\theta}$ for any $t > 0$ by

$$U_{t,\theta} f = \sum_{n=0}^{\infty} \pi_{nt,\theta} f = \sum_{n=0}^{\infty} (\pi_{t,\theta})^n f \tag{2.7}$$

for $f \in 1_\theta^\perp$ with the series converging in $L^2(\mu_\theta)$ to a limit in $1_\theta^\perp$.

Note also that with respect to P_θ^μ, for any $t > 0$ fixed, the strictly stationary sequence $(X_{mt})_{m \geq 0}$ is mixing: since the sequence is a Markov chain, to show this it suffices to show that for all bounded and measurable f, all $n_0 \in \mathbb{N}$ and all sets $A = ((X_0, X_t, \ldots, X_{n_0 t}) \in B)$ with B Borel,

$$\lim_{n \to \infty} E_\theta^\mu 1_A f(X_{(n_0+n)t}) = P_\theta^\mu(A)\, \mu_\theta(f). \tag{2.8}$$

But

$$E_\theta^\mu 1_A f(X_{(n_0+n)t}) = E_\theta^\mu \left(1_A \left(\pi_{nt,\theta}(f)\right)(X_{n_0 t})\right)$$

and since $\pi_{\theta,nt}(f) \to \mu_\theta(f)$ in $L^2(\mu_\theta)$ it follows that

$$E_\theta^\mu 1_A \left[\left(\pi_{nt,\theta}(f)\right)(X_{n_0 t}) - \mu_\theta(f)\right]^2 \leq \|\pi_{nt,\theta}(f) - \mu_\theta(f)\|_{\mu_\theta}^2 \to 0$$

with (2.8) now an easy consequence.

The mixing property ensures that the ergodic theorem holds, e.g.

$$\frac{1}{n} \sum_{m=1}^{n} \varphi\left(X_{(m-1)t}, X_{mt}\right) \underset{n \to \infty}{\to} E_\theta^\mu \varphi(X_0, X_t) \tag{2.9}$$

with convergence P_θ^μ−a.s. and in $L^1(P_\theta^\mu)$ if $\varphi \in L^1(Q_{t,\theta})$, and also in $L^2(P_\theta^\mu)$ if $\varphi \in L^2(Q_{t,\theta})$.

An important observation to be used later is

Proposition 2.1. *If $f \in \mathcal{D}_{0,\theta}$ and $\mathcal{A}_\theta f \equiv 0$, then $f \equiv 0$.*

Proof. The process $f(X)$ is a L^2-bounded $P_\theta^\mu-$martingale, hence $\lim_{t\to\infty} f(X_t)$ exists $P_\theta^\mu-$a.s. By ergodicity f visits any open subset of D infinitely often (cf. (2.9)) $P_\theta^\mu-$a.s, hence f must be constant, and since $\mu_\theta(f)=0$, $f\equiv 0$ follows. □

Remark. If $d=1$ the differential equation $A_\theta f\equiv 0$ is solved by the collection of scale functions S_θ given by (2.2). Hence $S_\theta\notin\mathcal{D}_\theta$. More generally, if d is arbitrary and $f\in L^1(\mu_\theta)\cap C^2(D)$ with $A_\theta f\equiv 0$ and (2.5) holds, then f is constant: $f(X)$ is a L^1-bounded $P_\theta^\mu-$martingale and one can copy the argument from the proof of Proposition 2.1.

3. Classes of estimation functions

With X given by (2.1), suppose that finitely many observations $X_0, X_{\Delta_1},\dots,X_{\Delta_n}$ are available, where $0<\Delta_1<\cdots<\Delta_n$. We wish to estimate θ, assuming that the distribution of X is determined by $P_{\theta_0}^\mu$ for some true parameter value θ_0. The estimation is performed using a *flow of estimating functions* $\mathcal{G}=(g_{t,\theta})_{t>0,\theta\in\Theta}$ with each $g_{t,\theta}: D^2\to\mathbb{R}^{p\times 1}$ belonging to $L^2(Q_{t,\theta})$, i.e.

$$E_\theta^\mu g_{t,\theta}^T g_{t,\theta}(X_0,X_t)<\infty\quad(t>0,\theta\in\Theta) \tag{3.1}$$

and satisfying the vital *unbiasedness condition*

$$E_\theta^\mu g_{t,\theta}(X_0,X_t)=0\quad(t>0,\theta\in\Theta). \tag{3.2}$$

Using g, an estimator $\hat\theta$ of θ is obtained by solving the equation

$$\sum_{m=1}^n g_{\Delta_m-\Delta_{m-1},\theta}\left(X_{\Delta_{m-1}},X_{\Delta_m}\right)=0 \tag{3.3}$$

in θ. (Notation: $\Delta_0=0$).

(3.3) is of course the type of estimating equation used in the literature on inference for discretely observed diffusions, with the idea of using g depending on neighbouring observations motivated by the Markovian nature of X, which in particular renders the estimating equation for the maximum likelihood estimator of the form (3.3), see (3.6). Below we discuss some of the different types of g's that have been suggested.

We shall from now on assume that the observations are equidistant, $\Delta_m=m\Delta$ for some $\Delta>0$. Then (3.3) becomes

$$\sum_{m=1}^n g_{\Delta,\theta}\left(X_{(m-1)\Delta},X_{m\Delta}\right)=0, \tag{3.4}$$

an equation that, as is well known, under mild regularity conditions leads to estimators that are consistent and asymptotically Gaussian as $n \to \infty$, (see Section 4). An important condition on g is of course that g can distinguish different parameter values, e.g.

$$E_\theta^\mu g_{\Delta,\theta'}(X_0, X_\Delta) = 0 \tag{3.5}$$

iff $\theta = \theta'$. Although this condition is not formally required below, we subsume it and will comment further on its importance in Section 4, see the paragraph below (4.7).

We write $g_{\Delta,\theta} \in \mathcal{E}$ if (3.1), (3.2) hold for $t = \Delta$ and $g_{\Delta,\theta}(x,y)$ is one time continuously differentiable in θ. In Section 7 we shall also need other partial derivatives. In accordance with the notation from the previous section, the derivatives are written

$$\begin{aligned}
\partial_y g_{t,\theta}(x,y) &= \left(\frac{\partial}{\partial y_i} g^k_{t,\theta}(x,y)\right)_{ki} \in \mathbb{R}^{p\times d},\\
\partial^2_{yy} g_{t,\theta}(x,y) &= \left(\frac{\partial^2}{\partial y_i \partial y_j} g^k_{t,\theta}(x,y)\right)_{k(ij)} \in \mathbb{R}^{p\times d^2},\\
\dot{g}_{t,\theta}(x,y) &= \left(\frac{\partial}{\partial \theta_l} g^k_{t,\theta}(x,y)\right)_{kl} \in \mathbb{R}^{p\times p},\\
\partial_t g_{t,\theta}(x,y) &= \left(\frac{\partial}{\partial t} g^k_{t,\theta}(x,y)\right)_{k} \in \mathbb{R}^{p\times 1}.
\end{aligned}$$

We write $\mathcal{G} \subset \mathcal{E}$ if $g_{t,\theta} \in \mathcal{E}$ for all $t > 0, \theta \in \Theta$.

From (3.4) we have p equations to solve for p unknowns $\theta_1, \ldots, \theta_p$. The equations may involve estimating functions g of different types and we now list some of the types studied in the literature. When we refer to these types later, it is always assumed that $\mathcal{G} \subset \mathcal{E}$ or, if a given Δ is considered, that $g_{\Delta,\theta} \in \mathcal{E}$.

MLE. The flow $(s_{t,\theta})_{t>0,\theta\in\Theta}$ used for obtaining the *maximum likelihood estimator* (conditionally on X_0),

$$s_{t,\theta}(x,y) = \frac{\dot{p}^T_{t,\theta}(x,y)}{p_{t,\theta}(x,y)}. \tag{3.6}$$

$\mathcal{M}$. The class of *martingale* estimating functions: for a given k and $\Delta > 0$, $g^k_{\Delta,\theta} \in \mathcal{M}$ if $\left(\sum_{m=1}^n g^k_{\Delta,\theta}\left(X_{(m-1)\Delta}, X_{m\Delta}\right), \mathcal{F}_{n\Delta}\right)_{n\geq 1}$ is a $P^\mu_\theta-$martingale.

$\mathcal{S}$. The class of *simple* estimating functions: for a given k and $\Delta > 0$, $g^k_{\Delta,\theta} \in \mathcal{S}$ if it is of the form $g^k_{\Delta,\theta}(x,y) = f_\theta(x)$ or $= h_\theta(y)$.

$\mathcal{T}$. The class of *explicit, transition dependent* estimating functions: for a given k and $\Delta > 0$, $g^k_{\Delta,\theta} \in \mathcal{T}$ if it is of the form

$$g^k_{\Delta,\theta}(x,y) = A_\theta f(x)\, h(y) - f(x) A_\theta h(y). \tag{3.7}$$

In Section 6 we shall introduce

$\mathcal{R}$. The class of *reversible* estimating functions.

Notation. If $\mathcal{K}$ is one of the classes of estimating functions, we write $g_{\Delta,\theta} \in \mathcal{K}$ if all $g^k_{\Delta,\theta} \in \mathcal{K}$. Also we write $T(f,h)(x,y)$ for the right hand side of (3.7).

As we shall see, the martingale estimating functions are particularly important, not only for applications but also for the theory to be developed presently. The classes $\mathcal{S}$ and $\mathcal{T}$ are important because they provide explicit estimating functions – but $\mathcal{T}$ (and also $\mathcal{R}$) are mostly relevant only if $d = 1$, see Section 6.

Of course (subject to mild regularity conditions) each coordinate of the MLE-estimating function (3.6) belongs to $\mathcal{M}$. Bibby and Sørensen [3] studied functions in $\mathcal{M}$ of the form

$$g^k_{\Delta,\theta}(x,y) = h_{\Delta,\theta}(x)\left(f(y) - \pi_{\Delta,\theta} f(x)\right), \tag{3.8}$$

in particular with $f(y) = y$ or $= y^2$ in the case $d = 1$. Other special cases of (3.8) are obtained if $f = f_\theta$ is an eigenfunction for $\mathcal{A}_\theta$ corresponding to an eigenvalue $\lambda_\theta < 0$, in which case (3.8) becomes

$$g^k_{\Delta,\theta}(x,y) = h_{\Delta,\theta}(x)\left(f_\theta(y) - e^{\Delta\lambda_\theta} f(x)\right),$$

see Kessler and Sørensen [11].

The bilinearity inherent in (3.8) immediately gives rise to the following large class in $\mathcal{M}$,

$$g^k_{\Delta,\theta}(x,y) = \sum_{q=1}^{r} h^q_{\Delta,\theta}(x)\left(f^q(y) - \pi_{\Delta,\theta} f^q(x)\right) \tag{3.9}$$

for different pairs $\left(h^q_{\Delta,\theta}, f^q_\theta\right)_{1\le q\le r}$, see e.g. Sørensen [15], Section 3.

The defining property of $\mathcal{M}$ requires that

$$E^\nu_\theta\left(g^k_{\Delta,\theta}\left(X_{(m-1)\Delta}, X_{m\Delta}\right)\middle| X_{(m-1)\Delta}\right) = 0 \quad P^\mu_\theta - \text{a.s.}$$

for all $m \in \mathbb{N}$. Thus $g^k_{\Delta,\theta} \in \mathcal{M}$ if (and essentially only if)

$$g^{k,*}_{\Delta,\theta} \equiv 0 \tag{3.10}$$

where in general $g^{k,*}_{\Delta,\theta} \in 1^{\perp}_{\theta}$ is given by

$$g^{k,*}_{\Delta,\theta}(x) := \int_D dy\, p_{\Delta,\theta}(x,y) g^k_{\Delta,\theta}(x,y),$$

see Kessler [8] and also [15]. Another useful characterization is that $g^k_{\Delta,\theta} \in \mathcal{M}$ iff $g^k_{\Delta,\theta} \perp \{\varphi \in L^2(Q_{\Delta,\theta}) : \varphi(x,y) = h(x) \text{ for some } h\}$.

Both the characterization (3.10) of $\mathcal{M}$ and the definition of $g^{k,*}_{\Delta,\theta}$ will be used frequently in the remainder of the paper.

The class $\mathcal{S}$ was studied in particular by Kessler [9]. Among his examples are

$$g_{\Delta,\theta}(x,y) = \frac{\dot{\mu}^T_\theta(x)}{\mu_\theta(x)}, \tag{3.11}$$

the collection of estimating functions one would use if the $X_{m\Delta}$ were independent and identically distributed with the correct marginal distribution μ_{θ_0}; also $g^k_{\Delta,\theta}(x,y) = f_\theta(x)$ with f_θ an eigenfunction for $\mathcal{A}_\theta$ with an eigenvalue < 0 (in which case (3.2) is automatic), and finally and most importantly, the class of *simple, explicit* estimating functions of the form

$$g^k_{\Delta,\theta}(x,y) = \mathcal{A}_\theta f(x) \tag{3.12}$$

(see also Baddeley [2] and Hansen and Scheinkman [7], C1, p. 774) which satisfy (3.2) because of (2.6).

Hansen and Scheinkman [7], C2, p. 775 also introduced the class $\mathcal{T}$ of estimating functions, which we shall discuss in Section 6 and which in the version presented here, apply only to reversible models, i.e. essentially only for $d = 1$: for $d > 1$ (3.2) will hold only in special models or for particular choices of the f, h appearing in (3.7).

In some of the cases discussed above, we stressed that the estimating functions were explicit, meaning essentially that they could be written in closed analytic form. Of course only in very few cases is MLE explicit and for (3.8) to be explicit one may have to use particular choices for f. Also, for the examples involving eigenfunctions, explicit expressions are only rarely available. Thus the main classes of explicit estimating functions are (3.12) and $\mathcal{T}$. Here, while the collection of functions obtained from (3.12) when f varies, forms a linear space, the bilinearity in (3.7) permits the definition of a large class of explicit functions, viz. finite sums

$$g^k_{\Delta,\theta}(x,y) = \sum_{q=1}^{r} \left(A_\theta f^q(x)\, h^q(y) - f^q(x) A_\theta h^q(y)\right). \tag{3.13}$$

A different possibility, not yet fully explored, for obtaining explicit estimating functions, is to look for real-valued $\tilde{f}_{t,\theta}(x)$ such that $\left(\tilde{f}_{t,\theta}(X_t), \mathcal{F}_t\right)_{t\geq 0}$ is a continuous-time $P_\theta^\mu-$martingale, and then use

$$g_\Delta^k(x,y) = h_{\Delta,\theta}(x)\left(\tilde{f}_{\Delta,\theta}(y) - \tilde{f}_{0,\theta}(x)\right),$$

which is in $\mathcal{M}$. Here, subject to differentiability and integrability requirements, $\tilde{f}$ must satisfy

$$\partial_t \tilde{f}_{t,\theta} + \mathcal{A}_\theta \tilde{f}_{t,\theta} \equiv 0,$$

as is seen from Itô's formula.

A last comment for now on estimating functions is that it is perfectly possible, but perhaps not very useful, to allow f in (3.8) and f and h in (3.7) to depend on both θ and Δ: the critical unbiasedness (3.2) still holds.

4. Asymptotics

We shall briefly review the asymptotic theory for the estimators obtained from the estimating equations (3.4) with g satisfying the unbiasedness condition (3.2) and the integrability condition (3.1). The asymptotics are done for equidistant observations $(X_{m\Delta})_{0\leq m\leq n}$ for $\Delta > 0$ fixed, $n \to \infty$ with the diffusion X satisfying Assumptions 1,2,3 from Section 2.

Recall the definition (2.7) of the potential operator $U_{\Delta,\theta}$ acting on functions in $\mathcal{D}_{0,\theta}$ with values in $1_\theta^\perp$. Of course

$$(I - \pi_{\Delta,\theta})\, U_{\Delta,\theta} f = f \quad (f \in \mathcal{D}_{0,\theta}). \tag{4.1}$$

Suppose now that $\mathcal{G} = (g_{t,\theta})$ is a flow of estimating functions in $\mathcal{E}$. Since each $g_{\Delta,\theta}^{k,*} \in 1_\theta^\perp$ (because of (3.2) and since $g_{\Delta,\theta}^k \in L^2(Q_{\Delta,\theta})$, see (3.1)), we find that

$$\breve{g}_{\Delta,\theta}(x,y) := g_{\Delta,\theta}(x,y) + U_{\Delta,\theta} g_{\Delta,\theta}^*(y) - U_{\Delta,\theta} g_{\Delta,\theta}^*(x) \tag{4.2}$$

(with U acting separately on each component $g^{k,*}$) defines a $p-$dimensional estimating function with all components in $\mathcal{M}$: (3.10) holds for $\breve{g}$ because using (4.1) it is seen that

$$\begin{aligned}
\breve{g}_{\Delta,\theta}^*(x) &= \int_D dy\, p_{\Delta,\theta}(x,y)\breve{g}_{\Delta,\theta}(x,y) \\
&= g_{\Delta,\theta}^*(x) + \pi_{\Delta,\theta}\left(U_{\Delta,\theta} g_{\Delta,\theta}^*\right)(x) - U_{\Delta,\theta} g_{\Delta,\theta}^*(x) \\
&= 0.
\end{aligned}$$

We shall call $\breve{g}_{\Delta,\theta}$ the *martingale estimating function associated with* $g_{\Delta,\theta}$. It will play a very important role in the remainder of this paper. The fact that *any*

estimating function is associated with an element of $\mathcal{M}$ means that for theoretical purposes, it suffices to consider martingale estimating functions. (For practical purposes $\breve{g}$ is of course useless – with the transition operators known explicitly only in exceptional cases, there are of course even fewer (if any) examples where the potential operators can be expressed explicitly).

The potential operators appear naturally in the statement of the central limit theorem for ergodic Markov chains. In our setup we obtain as a direct consequence of the martingale central limit theorem and the ergodicity property (2.8) that

$$\frac{1}{\sqrt{n}} \sum_{m=1}^{n} g_{\Delta,\theta}\left(X_{(m-1)\Delta}, X_{m\Delta}\right) \stackrel{d_\theta}{\to} N_p\left(0, v_{\Delta,\theta}(\breve{g})\right) \tag{4.3}$$

as $n \to \infty$, where the limiting covariance matrix is

$$v_{\Delta,\theta}(\breve{g}) = E_\theta^\mu \breve{g}_{\Delta,\theta} \breve{g}_{\Delta,\theta}^T\left(X_0, X_\Delta\right), \tag{4.4}$$

cf. Lemma 7 in Dacunha-Castelle and Florens-Zmirou [4].

Notation. $\stackrel{d_\theta}{\to}$ means convergence in distribution under P_θ^μ. $N_p(\xi, \Gamma)$ denotes the $p-$dimensional Gaussian distribution with mean vector ξ, covariance matrix Γ.

A comment on the derivation of (4.3): a direct application of the martingale central limit theorem gives that (4.3) holds with g replaced by $\breve{g}$. But

$$\frac{1}{\sqrt{n}} \sum_{m=1}^{n} \left(\breve{g}_{\Delta,\theta} - g_{\Delta,\theta}\right)\left(X_{(m-1)\Delta}, X_{m\Delta}\right) = \frac{1}{\sqrt{n}}\left(U_{\Delta,\theta} g_{\Delta,\theta}^*(X_{n\Delta}) - U_{\Delta,\theta} g_{\Delta,\theta}^*(X_0)\right)$$

which converges to 0 in $P_\theta^\mu-$probability and (4.3) itself follows.

For $g_{\Delta,\theta}$ a simple estimating function, $\breve{g}_{\Delta,\theta}$ was used by Florens-Zmirou [5] and also by Kessler [9]. Sørensen (the review [16] for instance), used (4.3) together with a Taylor expansion to obtain asymptotic normality for the estimator $\hat{\theta}_n$ solving the estimating equation (3.4), viz.

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \stackrel{d_\theta}{\to} N\left(0, \Lambda_{\Delta,\theta}^{-1}(g)\, v_{\Delta,\theta}(\breve{g})\, \left(\Lambda_{\Delta,\theta}^{-1}(g)\right)^T\right) \tag{4.5}$$

where

$$\Lambda_{\Delta,\theta}(g) := E_\theta^\mu\, \dot{g}_{\Delta,\theta}\left(X_0, X_\Delta\right), \tag{4.6}$$

i.e. the (k,l)'th element of the $p \times p-$matrix $\Lambda_{\Delta,\theta}(g)$ is $E_\theta^\mu \partial_{\theta_l} g_{\Delta,\theta}^k\left(X_0, X_\Delta\right)$.

The asymptotic variance matrix in (4.5) we denote by $\mathrm{var}_{\Delta,\theta}\left(g, \hat{\theta}\right)$,

$$\mathrm{var}_{\Delta,\theta}\left(g, \hat{\theta}\right) = \Lambda_{\Delta,\theta}^{-1}(g)\, v_{\Delta,\theta}(\breve{g})\, \left(\Lambda_{\Delta,\theta}^{-1}(g)\right)^T. \tag{4.7}$$

We shall not here discuss the precise conditions under which (4.5) holds, but refer the reader to the literature, e.g. [16]. Typically the assertion of the result yielding (4.5) is that with a $P_{\theta_0}^{\mu}-$probability tending to 1, (θ_0 denoting the true parameter value), (3.4) has a consistent solution which is asymptotically Gaussian as stated in (4.5). For this it is important that $g_{\Delta,\theta}$ *distinguishes between parameter values*, e.g. that (3.5) holds. Note that obviously $\Lambda_{\Delta,\theta}(\breve{g}) = \Lambda_{\Delta,\theta}(g)$.

Most of what follows relies on (4.5) being true. We shall make some formal assumptions so that the statement at least makes sense and otherwise refer to $g_{\Delta,\theta} \in \mathcal{E}$ as *well behaved* if it satisfies Assumption 4 below and if, for all θ_0, there is with $P_{\theta_0}^{\mu}-$probability tending to 1, a consistent solution to (3.4), which is asymptotically Gaussian according to (4.5) with $\theta = \theta_0$.

Assumption 4. Assume that $g_{\Delta,\theta} \in \mathcal{E}$ that $v_{\Delta,\theta}(\breve{g}) > 0$, that $\partial_{\theta_l} g_{\Delta,\theta}^k \in L^1(Q_{\Delta,\theta})$ for all k, l and that $\Lambda_{\Delta,\theta}(g)$ is non-singular.

The assumption that $v_{\Delta,\theta}(\breve{g}) > 0$ means essentially that the estimating equation enables us to estimate *all* the parameters θ_k. The assumption that Λ be non-singular, which is obviously necessary for (4.5) to make sense, is more critical and may fail in innocuous looking situations, as we shall now see. Here and below, by a *reparametrization* we mean a differentiable homeomorphism $\iota : \Theta \to \widetilde{\Theta}$, $\widetilde{\Theta}$ an open subset of $\mathbb{R}^p$, with $\widetilde{\theta} = \iota(\theta)$ the parameter vector corresponding to θ after the reparametrization.

Example 4.1. *Suppose that μ_θ does not depend on all the parameters, i.e. that, possibly after a reparametrization, μ_θ depends on $(\theta_1, \ldots, \theta_{p'})$ only where $p' < p$. Then θ cannot be estimated according to (4.5)–asymptotics using only simple estimating functions, more precisely $\Lambda_{\Delta,\theta}(g)$ is singular for all well behaved $g_{\Delta,\theta} \in \mathcal{S}$ for which the differentiation under the integral sign below is valid: suppose $g_{\Delta,\theta}(x, y) = f_{\Delta,\theta}(x)$ so that*

$$\Lambda_{\Delta,\theta}(g) = \int_D dx\, \mu_\theta(x) \dot{f}_{\Delta,\theta}(x).$$

Then for $l > p'$, using (3.2),

$$\begin{aligned} 0 &= \partial_{\theta_l} \left(\int_D dx\, \mu_\theta(x) f_{\Delta,\theta}(x) \right) \\ &= \int_D dx\, \mu_\theta(x) \partial_{\theta_l} f_{\Delta,\theta}(x), \end{aligned}$$

hence the last $p - p'$ columns of $\Lambda_{\Delta,\theta}(g)$ vanish.

Example 4.2. *Suppose that, possibly after a reparametrization, the differential operator A_θ is linear in $\theta' = (\theta_1, \ldots, \theta_{p'})$ for some $p' \geq 1$, in the sense that*

$$b_\theta(x) = \widetilde{b}_{\theta_-}(x)\theta', \quad C_\theta(x) = \widetilde{C}_{\theta_-}(x)\theta'$$

for some $\widetilde{b}_{\theta_-}(x) \in \mathbb{R}^{d \times p'}, \widetilde{C}_{\theta_-}(x) \in \mathbb{R}^{d^2 \times p'}$, writing $\theta_- = (\theta_{p'+1}, \ldots, \theta_p)$. (For $p' = p$, $\tilde{b}, \tilde{C}$ must not depend on θ). Then, subject to mild analytic conditions, $\Lambda_{\Delta,\theta}(g)$ is singular for all well behaved g, which are either in $\mathcal{S}$ of the form (3.12),

$$g^k_{\Delta,\theta}(x,y) = A_\theta h^k_{\Delta,\theta}(x) \tag{4.8}$$

for all k, or, for the reversible models where the members of $\mathcal{T}$ can be used, of the form

$$g^k_{\Delta,\theta}(x,y) = T\left(f^k_{\Delta,\theta}, h^k_{\Delta,\theta}\right)(x,y), \tag{4.9}$$

see (3.7) and the notation introduced just below (3.7).

To see that $\Lambda_{\Delta,\theta}(g)$ is singular, consider (4.9) for a fixed k. Then by calculation, omitting subscripts from $f^k_{\Delta,\theta}, h^k_{\Delta,\theta}$,

$$\sum_{l=1}^{p'} \theta_l \partial_{\theta_l} g^k_{\Delta,\theta}(x,y) = g^k_{\Delta,\theta}(x,y) + \sum_{l=1}^{p'} \theta_l \left(T(\partial_{\theta_l} f^k, h^k) + T(f^k, \partial_{\theta_l} h^k)\right)(x,y),$$

and taking expectations with respect to $Q_{\Delta,\theta}$ it follows that the linear combination with coefficients θ_l of the first p' columns of $\Lambda_{\Delta,\theta}(g)$ is equal to 0. (Note that for any $\theta = (0, \ldots 0, \theta_{p'+1}, \ldots \theta_p)$, $A_\theta \equiv 0$ which is not a diffusion generator, and hence $\theta' \neq 0$ for all $\theta \in \Theta$).

Simple examples of models that have the linearity property described here include the one-dimensional Ornstein-Uhlenbeck models

$$\begin{array}{ll} dX_t = -\theta X_t\, dt + \sqrt{2\theta}\, dB_t & (\theta > 0) \\ dX_t = -\theta X_t\, dt + \sigma\, dB_t & (\theta, \sigma > 0) \\ dX_t = (\alpha - \theta X_t)\, dt + \sigma\, dB_t & (\theta, \sigma > 0, \alpha \in \mathbb{R}). \end{array}$$

Note that for the first case, $\mu_\theta = N(0,1)$ does not depend on θ, so that no simple estimating functions can be used. Note also that in the two last cases, if one parameter is assumed known it is perfectly possible to find a combination of g^k of the form (4.8) or (4.9) for each k, such that Λ is non-singular.

5. Optimality

Let $\Delta > 0$ be fixed, and consider a given class $(g_{\Delta,\theta})_{\theta \in \Theta}$ of well behaved estimation functions. To find the best element from the class one would try to minimize

$\mathrm{var}_{\Delta,\theta}\left(g,\hat{\theta}\right)$ from (4.7), (which for $p \geq 2$ means minimizing in the partial ordering of symmetric positive semidefinite matrices, cf. the notation introduced after Assumption 2 above). Of course there may not be a variance minimizing g (in particular not if $p \geq 2$), and even if there is, it may be difficult to find. However, with a suitable linear structure imposed on the class of estimating functions, it is possible theoretically to describe the optimal g, but actually finding it may still be difficult, and even then the expression may be too complicated for practical purposes. In Section 7 we shall return to the optimality problem, proposing a solution that yields explicit results and which we believe can be useful in practice.

The proofs of the optimality inequalities rely on the following result, where, if $M = (M_{ij})$ is a matrix-valued random variable, EM is the matrix (EM_{ij}) of expectations.

Lemma 5.1. *Suppose that Y, Z, S are matrix-valued random variables of dimensions $a \times b, a \times b, b \times b$ respectively with S symmetric and strictly positive definite with probability 1. Assuming that all entries in the matrices $YZ^T, YSY^T, ZS^{-1}Z^T$ are integrable, it then holds that:*

(i) *if $E\left(YSY^T\right)$ is non-singular, then*

$$E\left(ZY^T\right)\left(E\left(YSY^T\right)\right)^{-1}E\left(YZ^T\right) \leq E\left(ZS^{-1}Z^T\right); \tag{5.1}$$

(ii) *if $E\left(YSY^T\right)$ and $E(YZ^T)$ are non-singular, then*

$$\left(E\left(ZY^T\right)\right)^{-1}\left(E\left(ZS^{-1}Z^T\right)\right)\left(E\left(YZ^T\right)\right)^{-1} \geq \left(E\left(YSY^T\right)\right)^{-1}; \tag{5.2}$$

(iii) *in (5.1), (5.2) there is equality if for some non-random, non-singular $K \in \mathbb{R}^{a\times a}$, $Y = KZS^{-1}$, equivalently if $Z = K^{-1}YS$.*

Proof. (ii) follows directly from (i) and (iii) is verified immediately. To show (5.1), define

$$\Sigma_{11} = E\left(ZS^{-1}Z^T\right), \quad \Sigma_{12} = E\left(ZY^T\right), \quad \Sigma_{22} = E\left(YSY^T\right),$$

and consider the symmetric $2a \times 2a$ matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}.$$

If $\Sigma \geq 0$ also $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T \geq 0$, which is precisely (5.1). So we complete the proof by showing that

$$s(u,v) := \begin{pmatrix} u^T & v^T \end{pmatrix}\Sigma\begin{pmatrix} u \\ v \end{pmatrix} \geq 0$$

for all $u, v \in \mathbb{R}^{a\times 1}$. But

$$s(u,v) = E\left(u^T Z S^{-1} Z^T u + 2u^T Z Y^T v + v^T Y S Y^T v\right),$$

and defining $\widetilde{u} = Z^T u$, $\widetilde{v} = SY^T v$, the random variable under the expectation becomes

$$\left(\widetilde{u}^T + \widetilde{v}^T\right) S^{-1} \left(\widetilde{u} + \widetilde{v}\right) \geq 0.$$

□

In the scalar case $a = b = 1$, (5.1), (5.2) follow from the Cauchy-Scwartz inequality and equality holds iff $Y = cZ/S$ for some constant $c \neq 0$. If $b = 1$, $S \equiv 1$, (5.2) gives for Y, Z $a-$dimensional random vectors

$$\left(E\left(ZY^T\right)\right)^{-1} E(ZZ^T) \left(E\left(YZ^T\right)\right)^{-1} \geq \left(E\left(YY^T\right)\right)^{-1} \tag{5.3}$$

with equality if $Z = K^{-1}Y$ for some non-random matrix K. Inequalities of this type were used by Godambe and Heyde [6].

For the first optimality result, that maximum-likelihood is best, we consider the estimation function

$$s_{\Delta,\theta} = \frac{\dot{p}^T_{\Delta,\theta}}{p_{\Delta,\theta}},$$

see (3.6).

Notation. To lighten the notation, if $\Delta > 0$ is given and $\psi(x,y)$ is a scalar, vector- or matrix-valued function on $D \times D$, we also write ψ for the random variable $\psi(X_0, X_\Delta)$.

Assuming $s_{\Delta,\theta}$ to be well behaved, the asymptotic variance for the MLE is of course

$$\mathrm{var}_{\Delta,\theta}\left(s, \hat{\theta}\right) = \left(E^\mu_\theta s_{\Delta,\theta} s^T_{\Delta,\theta}\right)^{-1} \tag{5.4}$$

as is seen from (4.7) with $g = s$, provided differentiation and integration can be interchanged as is done in

$$0 = \partial_\theta E^x_\theta s_{\Delta,\theta} = \int_D dy\ \left(s_{\Delta,\theta}(x,y)\dot{p}_{\Delta,\theta}(x,y) + p_{\Delta,\theta}(x,y)\dot{s}_{\Delta,\theta}(x,y)\right)$$

for all x, since then (see (4.6))

$$\Lambda_{\Delta,\theta}(s) = -E^\mu_\theta s_{\Delta,\theta} s^T_{\Delta,\theta}.$$

Proposition 5.2. *If $s_{\Delta,\theta}$ is well behaved with the asymptotic variance for the **MLE** given by (5.4), then*

$$\mathrm{var}_{\Delta,\theta}\left(g,\hat{\theta}\right) \geq \mathrm{var}_{\Delta,\theta}\left(s,\hat{\theta}\right)$$

for all Δ, θ and any well behaved g, provided $E^{\mu}_{\theta}\breve{g}_{\Delta,\theta}s^{T}_{\Delta,\theta}$ is non-singular and, for all x,

$$\partial_\theta \int_D dy\ \left(p_{\Delta,\theta}\breve{g}_{\Delta,\theta}\right)(x,y) = \int_D dy\, \partial_\theta \left(p_{\Delta,\theta}\breve{g}_{\Delta,\theta}\right)(x,y). \tag{5.5}$$

Proof. Because $E^{x}_{\theta}\breve{g}_{\Delta,\theta} = 0$ (which is just (3.10) applied to $\breve{g}$), the left hand side of (5.5) is 0. Thus $E^{x}_{\theta}\ \dot{\breve{g}}_{\Delta,\theta} = -E^{x}_{\theta}\breve{g}_{\Delta,\theta}s^{T}_{\Delta,\theta}$, and then also

$$E^{\mu}_{\theta}\ \dot{\breve{g}}_{\Delta,\theta} = -E^{\mu}_{\theta}\breve{g}_{\Delta,\theta}s^{T}_{\Delta,\theta}.$$

Now apply (5.3) with $Z = \breve{g}_{\Delta,\theta}$, $Y = s_{\Delta,\theta}$. □

The next result is a simple multivariate generalization of Kessler's [9] projection result Lemma 3.3.

Let $\Delta > 0$ be given and let for each θ, $\mathcal{L}_\theta$ be a closed linear subspace of the Hilbert space $L^2(Q_{\Delta,\theta})$ such that $\int_{D^2} dxdy\, q_{\Delta,\theta}(x,y)\kappa(x,y) = 0$ for all $\kappa \in \mathcal{L}_\theta$. Let $\breve{\mathcal{L}}_\theta$ denote the closure in $L^2(Q_{\Delta,\theta})$ of the space of functions $\breve{\kappa}$ of the form $\breve{\kappa}(x,y) = \kappa(x,y) + U_{\Delta,\theta}\kappa^*(y) - U_{\Delta,\theta}\kappa^*(y)$ for some $\kappa \in \mathcal{L}_\theta$, cf. (4.2). Also define $\mathcal{L}_{\theta,p}$ as the linear space of all $p-$variate functions $g_{\Delta,\theta} = \left(g^{k}_{\Delta,\theta}\right)_{1\leq k\leq p}$ with each $g^{k}_{\Delta,\theta} \in \mathcal{L}_\theta$ and finally define

$$\gamma_{\Delta,\theta} = \mathrm{proj}_{\breve{\mathcal{L}}_\theta}\frac{\dot{p}^{T}_{\Delta,\theta}}{p_{\Delta,\theta}}$$

where the projection is performed componentwise, within $L^2(Q_{\Delta,\theta})$.

Proposition 5.3. *Suppose that $\gamma_{\Delta,\theta}$ is well behaved. Then for any well behaved $g_{\Delta,\theta} \in \mathcal{L}_{p,\theta}$,*

$$\mathrm{var}_{\Delta,\theta}(g,\hat{\theta}) \geq \mathrm{var}_{\Delta,\theta}(\gamma,\hat{\theta}) = \left(E^{\mu}_{\theta}\gamma_{\Delta,\theta}\gamma^{T}_{\Delta,\theta}\right)^{-1}. \tag{5.6}$$

Proof. γ is characterized by the properties $\gamma^{k}_{\Delta,\theta} \in \breve{\mathcal{L}}_\theta$ for each k and

$$E^{\mu}_{\theta}\breve{\kappa}\left(r^{k}_{\Delta,\theta} - \gamma^{k}_{\Delta,\theta}\right) = 0 \tag{5.7}$$

for all k and all $\kappa \in \mathcal{L}_\theta$, where $r_{\Delta,\theta} = \dot{p}^T_{\Delta,\theta}/p_{\Delta,\theta}$. But if differentiation and integration can be interchanged so that

$$0 = \partial_\theta \int_D dxdy\, (q_{\Delta,\theta}\breve{g}_{\Delta,\theta})(x,y) = \int_D dxdy\, \partial_\theta (q_{\Delta,\theta}\breve{g}_{\Delta,\theta})(x,y),$$

using (5.7) with κ an arbitrary component of $g_{\Delta,\theta}$, we obtain

$$\Lambda_{\Delta,\theta}(g) = -E^\mu_\theta \breve{g}_{\Delta,\theta} r^T_{\Delta,\theta} = -E^\mu_\theta \breve{g}_{\Delta,\theta}\gamma^T_{\Delta,\theta}. \tag{5.8}$$

Now use (5.3) with $Z = \breve{g}_{\Delta,\theta}$, $Y = \gamma_{\Delta,\theta}$ to obtain the lower bound on the right of (5.6) for $\mathrm{var}_{\Delta,\theta}(g,\hat{\theta})$. Since $\gamma_{\Delta,\theta} \in \mathcal{M}$, it follows finally from (5.8) with $g = \breve{g} = \gamma$ that $\mathrm{var}_{\Delta,\theta}(\gamma,\hat{\theta})$ has the form stated in (5.6). □

Remark. With $p \geq 2$ it is only possible to formulate this result if all components $g^k_{\Delta,\theta}$ are allowed to vary in the *same* space $\mathcal{L}_\theta$. But for estimation purposes it is also quite natural that it should be so: introducing a subspace $\mathcal{L}^k_\theta$ for each k would imply that the p equations used for estimating θ had been numbered in some fashion, something which would be quite artificial since for instance an arbitrary permutation of the p equations does not change the estimator $\hat{\theta}$.

Example 5.4. *Let $d = 1$. Kessler [9] studied the case where $\mathcal{L}_\theta = \mathcal{S}_\theta$, the space of simple unbiased estimating functions $g_{\Delta,\theta}(x,y) = f_{\Delta,\theta}(x)$, and found the projection using eigenfunction expansions. We give here an expression valid in general, that does not involve eigenvalues or eigenfunctions.*

We have

$$\breve{\mathcal{L}}_\theta = \breve{\mathcal{S}}_\theta = \left\{ h(x) + U_{\Delta,\theta}h(y) - U_{\Delta,\theta}h(x) : h \in 1^\perp_\theta \right\}$$

and write

$$\psi^k(x) + U_{\Delta,\theta}\psi^k(y) - U_{\Delta,\theta}\psi^k(x)$$

for the projection onto $\breve{\mathcal{S}}_\theta$ of $r^k := \partial_{\theta_k} p_{\Delta,\theta}/p_{\Delta,\theta}$. Thus

$$E^\mu_\theta Z\, (h(X_0) + U_{\Delta,\theta}h(X_\Delta) - U_{\Delta,\theta}h(X_0)) = 0 \tag{5.9}$$

for all $h \in 1^\perp_\theta$, where

$$Z = r^k\,(X_0, X_\Delta) - \psi^k(X_0) - U_{\Delta,\theta}\psi^k(X_\Delta) + U_{\Delta,\theta}\psi^k(X_0).$$

The idea is now to rewrite the expectation in (5.9) as an inner product in $L^2\,(\mu_\theta)$ between h and some function $\tilde{\psi} \in 1^\perp_\theta$. It then follows that $\tilde{\psi} \equiv 0$, and from this equation ψ^k is identified. But using stationarity, the Markov property and time

reversal (e.g. that $q_{\Delta,\theta}(x,y)$ is symmetric in x,y and that $U_{\Delta,\theta}$ is selfadjoint, see Section 6), writing $\pi = \pi_{\Delta,\theta}$, $U = U_{\Delta,\theta}$, one finds

$$\begin{aligned}
E_\theta^\mu Z h(X_0) &= E_\theta^\mu r^{k*}(X_0)h(X_0), \\
E_\theta^\mu Z U h(X_\Delta) &= E_\theta^\mu \left(r^{k*} + \frac{\partial_{\theta_k}\mu_\theta}{\mu_\theta} - \left(\pi \frac{\partial_{\theta_k}\mu_\theta}{\mu_\theta}\right) - (I+\pi)\psi^k \right)(X_0) U h(X_0) \\
&= E_\theta^\mu \left(U r^{k*} + \frac{\partial_{\theta_k}\mu_\theta}{\mu_\theta} - (2U - I)\,\psi^k \right)(X_0) h(X_0), \\
E_\theta^\mu Z (Uh)(X_0) &= E_\theta^\mu U r^{k*}(X_0) h(X_0).
\end{aligned}$$

Since $r^{k*} \equiv 0$,

$$\tilde{\psi}(x) = \frac{\partial_{\theta_k}\mu_\theta}{\mu_\theta}(x) - (2U - I)\,\psi^k(x)$$

so the equation $\tilde{\psi} \equiv 0$ translates into

$$(I - \pi)\frac{\partial_{\theta_k}\mu_\theta}{\mu_\theta} = (I + \pi)\psi^k$$

or, equivalently,

$$\psi^k = U_{2\Delta,\theta}\,(I - \pi_{\Delta,\theta})^2\,\frac{\partial_{\theta_k}\mu_\theta}{\mu_\theta}.$$

We have shown that

$$\mathrm{proj}_{\breve{\mathcal{S}}_\theta}\,\frac{\dot{p}^T_{\Delta,\theta}}{p_{\Delta,\theta}}(x,y) = \psi(x) + U_{\Delta,\theta}\psi(y) - U_{\Delta,\theta}\psi(x),$$

where ψ is the unique solution with all $\psi^k \in 1_\theta^\perp$ to the equation

$$(I - \pi_{\Delta,\theta})\,\frac{\dot{\mu}^T_\theta}{\mu_\theta} = (I + \pi_{\Delta,\theta})\,\psi.$$

Example 5.5. *For $p = 1$, Bibby and Sørensen [3] studied optimality of estimating functions of the form (3.8). In our formulation this would correspond to taking, for a given $f \in L^2(\mu_\theta)$ <u>not</u> depending on θ,*

$$\mathcal{L}_\theta = \left\{ h(x)\,(f(y) - \pi_{\Delta,\theta} f(x)) : h \in L^2\,(\mu_\theta) \right\}.$$

In this case $\breve{\mathcal{L}}_\theta = \mathcal{L}_\theta$ and the projection takes the form

$$\gamma_{\Delta,\theta}(x,y) = h^{opt}_{\Delta,\theta}(x)\,(f(y) - \pi_{\Delta,\theta} f(x))$$

and using (5.7) it is not difficult to find that the optimal h is given by

$$h^{opt}_{\Delta,\theta} = \frac{\partial_\theta \left(\pi_{\Delta,\theta} f\right)}{\pi_{\Delta,\theta} f^2 - \left(\pi_{\Delta,\theta} f\right)^2},$$

cf. [3], provided

$$\partial_\theta \int dy\, \kappa(y) p_{\Delta,\theta}(x,y) = \int dy\, \kappa(y) \dot{p}_{\Delta,\theta}(x,y)$$

for $\kappa \equiv f$, $\kappa \equiv 1$.

For $p \geq 2$ the natural approach is to fix p functions $f^k \in L^2(\mu_\theta)$ not depending on θ, and then use

$$\mathcal{L}_\theta = \breve{\mathcal{L}}_\theta = \left\{ \sum_{k=1}^{p} h^k(x) \left(f^k(y) - \pi_{\Delta,\theta} f^k(x) \right) : h^1, \ldots, h^p \in L^2(\mu_\theta) \right\}, \tag{5.10}$$

cf. (3.9). The result is

$$h^{opt}_{\Delta,\theta} = \left[\pi_{\Delta,\theta} \left(f f^T \right) - \left(\pi_{\Delta,\theta} f\right) \left(\pi_{\Delta,\theta} f\right)^T \right]^{-1} \partial_\theta \left(\pi_{\Delta,\theta} f\right)$$

with the k'th column of $h^{opt}_{\Delta,\theta} \in \mathbb{R}^{p \times p}$ giving the choices for the $h^1, \ldots, h^p$ in (5.10) when projecting $\partial_{\theta_k} p_{\Delta,\theta} / p_{\Delta,\theta}$ onto $\mathcal{L}_\theta$. For discussions of optimality for this class, see Kessler [8] (quoted in Sørensen [15], Theorem 3.1), and Pedersen [12].

6. Time reversal

Since the diffusion X has an invariant measure, it is possible to define X as a strictly stationary process $X = (X_t)_{-\infty<t<\infty}$ in doubly infinite time, each X_t having density μ_θ. But then also the *reversed* process $\overleftarrow{X} = \left(\overleftarrow{X}_t\right)_{-\infty<t<\infty}$, where $\overleftarrow{X}_t = X_{-t}$, is a diffusion with invariant density μ_θ and transition densities

$$\overleftarrow{p}_{t,\theta}(x,y) = \mu_\theta(y) p_{t,\theta}(y,x) \frac{1}{\mu_\theta(x)}. \tag{6.1}$$

If $d = 1$, for the diffusions we are considering, $\overleftarrow{p}_{t,\theta} = p_{t,\theta}$, i.e. the *one-dimensional diffusions are reversible.* Put differently, for $d = 1$, the transition operators $\pi_{t,\theta}$ (acting on $L^2(\mu_\theta)$) are selfadjoint, and the generator $\mathcal{A}_\theta$ (acting on $\mathcal{D}_\theta$) is selfadjoint. (A quick proof of the reversibility property is the following: as noted p. 7 above, for $d = 1$, the space $\mathcal{DC}$ of functions with compact support is dense in $1_\theta^\perp$, and it therefore suffices to show that for any $f, h \in \mathcal{DC}$,

$$\int_D dx\, \mu_\theta(x) A_\theta f(x) h(x) = \int_D dx\, \mu_\theta(x) f(x) A_\theta h(x).$$

But $A_\theta f = k_\theta \left(f'/S'_\theta\right)'$ (see (2.2) for the definition of S'_θ), for some consatnt k_θ and by partial integration the expression on the left then becomes $-k_\theta \int dx \ \left(f'h'\right)/S'_\theta$, which is symmetric in f, h).

For $d \geq 1$, it is known (see e.g. the overview in [7]) that $\overleftarrow{X}$ satisfies the stochastic differential equation

$$d\overleftarrow{X}_t = \overleftarrow{b}_\theta(\overleftarrow{X}_t)\, dt + \overleftarrow{\sigma}_\theta(\overleftarrow{X}_t)\, d\overleftarrow{B}_t$$

where $\overleftarrow{B}$ is a $d-$dimensional Brownian motion, and

$$\overleftarrow{b}^i_\theta(x) = -b^i(x) + \frac{1}{\mu_\theta(x)} \sum_{j=1}^{d} \partial_{x_j} \left(\mu_\theta C^{ij}_\theta\right)(x), \quad (1 \leq i \leq d), \tag{6.2}$$

$$\overleftarrow{C}_\theta(x) := \left(\overleftarrow{\sigma}_\theta \overleftarrow{\sigma}^T_\theta\right)(x) = C_\theta(x),$$

so that typically $\overleftarrow{b}_\theta \neq b_\theta$, while always $\overleftarrow{C}_\theta = C_\theta$. Thus, only in special cases, is a multidimensional diffusion model reversible.

In the remainder of this section, unless explicitly stated otherwise, we assume that X is reversible for all θ (with $d = 1$ by far the most important case). Let $\mathcal{G} = (g_{t,\theta})_{t\geq 0, \theta\in\Theta}$ be a flow of well behaved estimating functions, and define for each t, θ,

$$\overleftarrow{g}_{t,\theta}(x, y) = g_{t,\theta}(y, x).$$

Because of the reversibility, $\overleftarrow{g}$ satisfies (3.1), (3.2). If $\overleftarrow{g}_{\Delta,\theta}$ is also well behaved for a given Δ, the estimating function $\overleftarrow{g}_{\Delta,\theta}$ may be used as an alternative to $g_{\Delta,\theta}$, and as we shall now see, by combining the two by simply averaging, one may obtain an estimating function which is better than either.

In general models (reversible or not), we shall call an estimating function $g_{\Delta,\theta}$ such that $E^\mu_\theta \overleftarrow{g}_{\Delta,\theta}(X_0, X_\Delta) = 0$ for all θ, *reversible* if (cf. (4.2))

$$g_{\Delta,\theta}(x, y) + U_{\Delta,\theta} g^*_{\Delta,\theta}(y) - U_{\Delta,\theta} g^*_{\Delta,\theta}(x) = g_{\Delta,\theta}(y, x) + U_{\Delta,\theta}\, {}^*g_{\Delta,\theta}(y) - U_{\Delta,\theta}\, {}^*g_{\Delta,\theta}(x) \tag{6.3}$$

for $Q_{\Delta,\theta}-$almost all (x, y). Here

$${}^*g_{\Delta,\theta}(x) = \int_d dy\, p_{\Delta,\theta}(x, y) g_{\Delta,\theta}(y, x),$$

so (6.3) merely states that the martingale estimating functions associated with $g_{\Delta,\theta}$ and $\overleftarrow{g}_{\Delta,\theta}$ are the same.

We denote the class of reversible estimating functions by $\mathcal{R}$ (cf. p. 11).

Again, from now on X is assumed to be reversible.

Proposition 6.1. *If $g, \overleftarrow{g}$ are both well behaved, then*

$$\mathrm{var}_{\Delta,\theta}(\overleftarrow{g}, \hat{\theta}) = \mathrm{var}_{\Delta,\theta}(g, \hat{\theta}). \tag{6.4}$$

Furthermore, if also $\overline{g} := \frac{1}{2}(\overleftarrow{g} + g)$ is well behaved, then

$$\mathrm{var}_{\Delta,\theta}(\overline{g}, \hat{\theta}) \leq \mathrm{var}_{\Delta,\theta}\left(g, \hat{\theta}\right) \tag{6.5}$$

with equality if and only if $g \in \mathcal{R}$.

Proof. (6.4) is intuitively obvious and corresponds to comparing the estimates of θ obtained using g as estimating function on the observations $(X_{m\Delta})_{0\leq m\leq n}$ and $\left(\overleftarrow{X}_{m\Delta}\right)_{0\leq m\leq n} = (X_{-m\Delta})_{0\leq m\leq n}$ respectively. Note however the asymmetry as $n \to \infty$ with the most recent observations, $X_{n\Delta}$ and $X_{-n\Delta}$ respectively, being added after, respectively before, the earlier observations. We shall therefore give a formal proof of (6.4) which, as it turns out, is not an entirely trivial matter.

Clearly $\Lambda_{\Delta,\theta}(\overleftarrow{g}) = \Lambda_{\Delta,\theta}(g)$ (see (4.6)) by reversibility, so we need only show that

$$E_\theta^\mu \overleftarrow{\breve{g}} \overleftarrow{\breve{g}}^T = E_\theta^\mu \breve{g}\breve{g}^T, \tag{6.6}$$

where

$$\overleftarrow{\breve{g}}(x, y) = g(y, x) + U^* g(y) - U^* g(x).$$

(To ease the notation, the subscripts $_{\Delta,\theta}$ have been omitted from $\overleftarrow{\breve{g}}, g, U$). But writing out

$$\breve{g}\breve{g}^T = (g + (Ug^*)(X_\Delta) - (Ug^*)(X_0))\left(g^T + (Ug^*)^T(X_\Delta) - (Ug^*)^T(X_0)\right),$$

into 9 terms, taking expectations, conditioning on X_0 or X_Δ as appropriate and using reversibility and stationarity, one arrives at

$$\begin{aligned} E_\theta^\mu \breve{g}\breve{g}^T \;=\; & E_\theta^\mu g g^T \\ & + E_\theta^\mu({}^*g\,(Ug^*)^T - g^*\,(Ug^*)^T + (Ug^*)\,{}^*g^T + (Ug^*)\,(Ug^*)^T \\ & - (\pi U g^*)\,(Ug^*)^T - (Ug^*)\,g^{*T} - (Ug^*)\,(\pi U g^*)^T + (Ug^*)\,(Ug^*)^T), \end{aligned}$$

writing $\pi = \pi_{\Delta,\theta}$, and where all terms in the last expectation are evaluated at X_0. Using $\pi U = U - I$ this collapses to

$$E_\theta^\mu \breve{g}\breve{g}^T = E_\theta^\mu g g^T + E_\theta^\mu\left({}^*g\,(Ug^*)^T + (Ug^*)\,{}^*g^T\right).$$

The same calculation applied to $\overleftarrow{g}$ gives

$$E_\theta^\mu \overleftarrow{\breve{g}}\,\overleftarrow{\breve{g}}^T = E_\theta^\mu \overleftarrow{g}\,\overleftarrow{g}^T + E_\theta^\mu\left({}^*\overleftarrow{g}\,(U\overleftarrow{g}^*)^T + (U\overleftarrow{g}^*)^*\,\overleftarrow{g}^T\right)$$

and (6.6) follows since $E_\theta^\mu g g^T = E_\theta^\mu \overleftarrow{g}\,\overleftarrow{g}^T$ (reversibility), ${}^*\overleftarrow{g} = g^*$, $\overleftarrow{g}^* = {}^*g$, and the fact that U is selfadjoint.

Since $\Lambda_{\Delta,\theta}(\overline{g}) = \Lambda_{\Delta,\theta}(g)$, to show (6.5) it suffices to show that $E_\theta^\mu \breve{\overline{g}}\breve{\overline{g}}^T \le E_\theta^\mu \breve{g}\breve{g}^T$ with equality iff $g \in \mathcal{R}$. Let $u \in \mathbb{R}^p \backslash 0$. Of course

$$E_\theta^\mu \left(u^T g\right)^2 = E_\theta^\mu \left(u^T \overleftarrow{g}\right)^2, \tag{6.7}$$

hence

$$u^T \left(E_\theta^\mu \breve{\overline{g}}\breve{\overline{g}}^T\right) u = E_\theta^\mu \left(u^T \breve{\overline{g}}\right)^2 = \tfrac{1}{2} E_\theta^\mu \left(u^T \breve{g}\right)^2 + \tfrac{1}{2} E_\theta^\mu \left(u^T \breve{g}\right)\left(u^T \breve{\overleftarrow{g}}\right),$$

and here by Cauchy-Schwarz and (6.7),

$$\left| E_\theta^\mu \left(u^T \breve{g}\right)\left(u^T \breve{\overleftarrow{g}}\right)\right| \le E_\theta^\mu \left(u^T \breve{g}\right)^2 \tag{6.8}$$

proving (6.5). Furthermore, equality holds in (6.8) iff $u^T \breve{g} = K(u) u^T \breve{\overleftarrow{g}}$ $Q_{\Delta,\theta}-$a.s. for some constant $K(u)$. But (6.7) forces $K(u) = \pm 1$ and by inspection it is then seen that $E_\theta^\mu \left(u^T \breve{\overline{g}}\right)^2 = E_\theta^\mu \left(u^T \breve{g}\right)^2$ iff $K(u) = 1$. Letting u run through a countable dense subset of $\mathbb{R}^p$, it finally follows that equality in (6.5) holds iff $\breve{g} = \breve{\overleftarrow{g}}$ $Q_{\Delta,\theta}-$a.s. □

Proposition 6.1 shows that in reversible models, if $g \notin \mathcal{R}$, one can always improve an estimating function g by symmetrizing: using $\overline{g}$ instead of g. Of course $\overline{g} \in \mathcal{R}$.

We give two examples of estimating functions of the types dicussed earlier, where $g \in \mathcal{R}$ and symmetrizing therefore does not help.

Example 6.2. *By Propositions 5.2, 6.1, for a reversible model* $\boldsymbol{MLE} \in \mathcal{R}$. *A direct proof is the following: since* $s = s_{\Delta,\theta} \in \mathcal{M}$, $s^* \equiv 0$. *Further, from* $s(x,y) = \dot{q}_{\Delta,\theta}^T / q_{\Delta,\theta}(x,y) - \dot{\mu}_\theta^T / \mu_\theta(x)$ *and the symmetry of* $q_{\Delta,\theta}$ *follows that*

$$\overleftarrow{s}(x,y) = s(x,y) + \frac{\dot{\mu}_\theta^T}{\mu_\theta}(x) - \frac{\dot{\mu}_\theta^T}{\mu_\theta}(y), \tag{6.9}$$

and hence, since $s^* \equiv 0$, *that*

$$\overleftarrow{s}^* = {}^*s = (I - \pi_{\Delta,\theta}) \frac{\dot{\mu}_\theta^T}{\mu_\theta}.$$

Thus, $U_{\Delta,\theta}\, {}^*s = \dot{\mu}_\theta^T / \mu_\theta^T$, *and using (6.9),* $\breve{\overleftarrow{s}}\,(x,y) = s(x,y)$, *i.e.* $s \in \mathcal{R}$, *follows.*

Example 6.3. *If all $g_{t,\theta}$ are simple, also in non-reversible models (3.2) is automatic for $\overleftarrow{g}_{t,\theta}$, and as we shall see, $\mathcal{S} \subset \mathcal{R}$. If e.g. $g_{\Delta,\theta}(x,y) = f_{\Delta,\theta}(x)$, (each component of g has this form) we find (omitting subscripts $_{\Delta,\theta}$) that $g^* = f$, ${}^*g = \pi f$ which quickly gives $\breve{g}(x,y) = Uf(y) - \pi Uf(x) = \widetilde{\overleftarrow{g}}\,(x,y)$. The intuitive content behind this fact that simple estimating functions are reversible is this: the estimating equations based on g and $\overleftarrow{g}$ respectively are $\sum_0^{n-1} f_{\Delta,\theta}(X_{m\Delta}) = 0$ and $\sum_1^n f_{\Delta,\theta}(X_{m\Delta}) = 0$, and clearly they are asymptotically equivalent.*

We saw that $s_{\Delta,\theta} \in \mathcal{M}$, but other martingale estimating functions are typically not reversible: for g^k given by (3.8) to be reversible it is necessary and sufficient that either h is constant or that f and h are proportional. (The only term in $\breve{g}^k$, which is not just a function of x or a function of y, is the product $h(x)f(y)$. For g^k to be reversible this term must cancel against $h(y)f(x)$, which is possible only if f or h is constant or if f and h are proportional. Since f constant results in $g^k \equiv 0$ this possibility can be ruled out, and it is then easily checked directly that with h constant or $f \propto h$, g^k is indeed reversible).

For reversible models one can use the *explicit, transition dependent* estimating functions from the class $\mathcal{T}$, introduced by Hansen and Scheinkman [7], see (3.7),

$$g^k_{\Delta,\theta}(x,y) = \left(A_\theta f^k\right)(x)h^k(y) - f^k(x)\left(A_\theta h^k\right)(y), \tag{6.10}$$

(with (3.12) the special case $h^k \equiv 1$) where f^k, h^k are allowed to depend on Δ, θ. That (3.2) holds is an easy consequence of reversibility and the fact that $\pi_{\Delta,\theta}$ and A_θ are selfadjoint. Instead of (6.10) one could also use

$$g^k_{\Delta,\theta}(x,y) = \left(A_\theta f^k\right)(x)h^k(y) - f^k(y)\left(A_\theta h^k\right)(x). \tag{6.11}$$

Because of Proposition 6.1, if g^k is of the form (6.10) or (6.11) one should symmetrize. The two $\overline{g}^k-$functions obtained are identical, viz.

$$\tfrac{1}{2}\left(\left(A_\theta f^k\right)(x)h^k(y) - f^k(x)\left(A_\theta h^k\right)(y) + \left(A_\theta f^k\right)(y)h^k(x) - f^k(y)\left(A_\theta h^k\right)(x)\right),$$

which is therefore the preferred choice. Even better estimating functions are obtained by symmetrizing in (3.13).

In models that are not reversible, Hansen and Scheinkman [7] propose that (6.10) be replaced by

$$g^k_{\Delta,\theta}(x,y) = \left(\overleftarrow{A}_\theta f^k\right)(x)h^k(y) - f^k(x)\left(A_\theta h^k\right)(y),$$

with $\overleftarrow{A}_\theta$ the differential operator that defines the generator for $\overleftarrow{X}$. This gives an explicit, transiton dependent estimating function if μ_θ is known explicitly, cf. (6.2), but of course, for $d \geq 2$ it may be difficult to determine μ_θ.

We conclude this section with a brief comment on the class $\mathcal{U}$ of *useless estimating functions:* For a reversible model of course

$$g_{\Delta,\theta}(x,y) = \varsigma_{\Delta,\theta}(x,y) - \varsigma_{\Delta,\theta}(y,x)$$

satisfies the unbiasedness condition (3.2) for any $\varsigma_{\Delta,\theta} \in L^2(Q_{\Delta,\theta})$. But (3.5) is not satisfied and for any $g_{\Delta,\theta} \in \mathcal{U}$, $\Lambda_{\Delta,\theta}(g) = 0$ so there is no hope of asymptotics as in (4.5). It may however be noted that it is sometimes possible to obtain a variance reduction by considering the sum of a well behaved estimating function and a useless one. Also, by using functions from $\mathcal{U}$, it may be shown that an arbitrarily large amount in efficiency may be gained by symmetrizing as proposed in Proposition 6.1: suppose that $p = 1$ and let $g^\circ_{\Delta,\theta}$ be well-behaved for a given Δ. Now destroy the good properties of $g^\circ_{\Delta,\theta}$ by adding a member from $\mathcal{U}$ thus,

$$g_{\Delta,\theta}(x,y) = g^\circ_{\Delta,\theta}(x,y) + a\left(\varsigma(x,y) - \varsigma(y,x)\right)$$

with $|a|$ large. Then $\Delta_{\Delta,\theta}(g)$ does not depend on θ and it is an easy matter to arrange things so that $E^\mu_\theta \breve{g}^2_{\Delta,\theta}(X_0, X_\Delta) = O(a^2)$ as $|a| \to \infty$. But $\bar{g}_{\Delta,\theta} = g^\circ_{\Delta,\theta}$, and so

$$\frac{\mathrm{var}_{\Delta,\theta}(\bar{g})}{\mathrm{var}_{\Delta,\theta}(g)}$$

can be made arbitrarily small!

7. Small $\Delta-$optimality

Proposition 5.3 showed how the best estimating function from a linear space may be found using projections in $L^2(Q_{\Delta,\theta})$. But as already noted, this rarely leads to explicit formulas, and the method is therefore difficult to apply in practice. Here we shall propose the concept of *small* $\Delta-$*optimality* and discuss how it may be applied to give estimating functions with good properties.

For Δ large, the observations $(X_{m\Delta})_{0\le m\le n}$ are almost i.i.d. μ_θ, hence optimal inference as $\Delta \to \infty$ should be performed using the simple estimating function

$$g^{\mathrm{opt}}_{\infty,\theta}(x,y) := \frac{\dot{\mu}^T_\theta}{\mu_\theta}(x),$$

see (3.11), corresponding to finding the MLE if the $X_{m\Delta}$ were truly i.i.d. (Recall from Example 4.1 that only parameters appearing in μ_θ can be estimated using (3.11)). Therefore, if optimal estimating functions g^{opt}_0 can be found for $\Delta \to 0$, by a suitable interpolation between $g^{\mathrm{opt}}_{0,\theta}$ and $g^{\mathrm{opt}}_{\infty,\theta}$, one may find flows $\left(g^{\mathrm{opt}}_{t,\theta}\right)$ that perform well for *all* Δ. (And such a flow will give useful estimators even if Δ is

large and μ_θ does not depend on all the parameters, only in that case estimators for some parameters will have large variances: the model does not allow for precise estimation of all the parameters).

With the optimality as $\Delta \to \infty$ in place, we shall focus on the problem of minimizing for a given flow, $\mathrm{var}_{\Delta,\theta}(g, \hat{\theta})$ as $\Delta \to 0$. The conditions we arrive at (for multivariate, multiparameter diffusion models) are similar to the ones found by Kessler [10], (who studied models with $p = 2$ of the form $dX_t = b_{\theta_1}(X_t)\, dt + \sigma_{\theta_2}(X_t)\, dB_t$, $\theta = (\theta_1, \theta_2) \in \Theta \subset \mathbb{R}^2$, with n discrete observations Δ_n apart, and used moment estimating functions to obtain efficiency for $\Delta_n \to 0$ at a suitable rate), and Woerner [17] (who for $d = 1$ gave conditions for LAN in one-parameter models ($p = 1$) in the two cases where (i) the parameter enters in the drift b_θ only, (ii) σ_θ depends on θ and b_θ possibly also).

Our idea is to derive a power expansion in Δ of $\mathrm{var}_{\Delta,\theta}(g, \hat{\theta})$, which, as we shall see, takes the form

$$\mathrm{var}_{\Delta,\theta}(g, \hat{\theta}) = \frac{1}{\Delta} v_{-1} + v_0 + O(\Delta), \tag{7.1}$$

and then call $g \in \mathcal{G}$ *small* $\Delta-$*optimal* if the leading term(s) in this expansion, v_{-1} or v_0 or a suitable combination of the two, is minimized. Depending on the structure of the model, we shall distinguish between three cases:

(i) Minimizing v_{-1}: $\sigma_\theta = \sigma$ does not depend on θ;

(ii) Minimizing v_0 with $v_{-1} \equiv 0$: σ_θ depends on all parameters $\theta_1, \ldots, \theta_p$;

(iii) Minimizing a combination of v_{-1}, v_0: σ_θ depends on some, but not all the parameters.

As $\Delta \to 0$, we are approaching a limiting continuous time observation scheme, $(X_t)_{0 \le t \le n\Delta}$. For that it is essentially true that if $\mathbb{P}_{\theta,t}$ is the distribution of $(X_s)_{0 \le s \le t}$ under $\bar{P}_\theta^\mu$, then $\mathbb{P}_{\theta',t} \ll \mathbb{P}_{\theta,t}$ for all t, θ, θ' in case (i), while in case (ii), $\mathbb{P}_{\theta',t} \perp \mathbb{P}_{\theta,t}$ if $\theta' \neq \theta$, and in case (iii) absolute continuity holds iff the parameters on which σ_θ depend are fixed. For n fixed, $\Delta \to 0$, we would in case (i) expect the information about θ to be approximately proportional to Δ, explaining the leading term $\Delta^{-1} v_{-1}$ in (7.1). In case (ii), due to the continuous time singularity between measures, we gain an order of magnitude in estimation precision: it is possible to obtain $v_0 \equiv 0$.

In case (i) under local absolute continuity, the log-likelihood for observation on $[0, t]$ given X_0 has the form

$$\int_0^t b_\theta^T(X_s) C^{-1}(X_s)\, dX_s - \tfrac{1}{2} \int_0^t b_\theta^T(X_s) C^{-1}(X_s) b_\theta(X_s)\, ds,$$

corresponding to the score function

$$\begin{aligned} s^c_{t,\theta} &= \int_0^t \dot{b}_\theta^T(X_s)C^{-1}(X_s)\,dX_s - \int_0^t \dot{b}_\theta^T(X_s)C^{-1}(X_s)b_\theta(X_s)\,ds \\ &\approx \dot{b}_\theta^T(X_0)C^{-1}(X_0)\,(X_t - X_0) - t\dot{b}_\theta^T(X_0)C^{-1}(X_0)b_\theta(X_0). \end{aligned}$$

The latter approximation suggests the approximate estimating function (which does not in general satisfy (3.2)),

$$\begin{aligned} \tilde{g}_{t,\theta}(x,y) &= \dot{b}_\theta^T(x)C^{-1}(x)\,(y-x) - t\dot{b}_\theta^T(x)C^{-1}(x)b_\theta(x), \\ \tilde{g}_{0,\theta}(x,y) &= \dot{b}_\theta^T(x)C^{-1}(x)\,(y-x)\,. \end{aligned}$$

It should be noted that $\partial_y \tilde{g}_{0,\theta}(x,x) = \dot{b}_\theta^T(x)C^{-1}(x)$, which is precisely the optimality criterion for martingale estimating functions presented in Theorem 7.5 (i) below!

To obtain (7.1) we shall simply use Itô-Taylor expansions of the random matrices $\breve{g}_{\Delta,\theta}\breve{g}_{\Delta,\theta}^T$, $\dot{g}_{\Delta,\theta}$ that appear in the expression for $\mathrm{var}_{\Delta,\theta}(g,\hat{\theta})$, see (4.7), (4.6).

Suppose first that $(s,x,y) \to \phi_s(x,y)$ is a continuous real-valued function of $s \geq 0$ and $(x,y) \in D^2$, continuously differentiable in s, twice continuously differentiable in y. Let $\mathcal{H}_\theta$ denote the differential operator given by

$$\mathcal{H}_\theta\phi_s(x,y) = \partial_s\phi_s(x,y) + A_{\theta,y}\phi_s(x,y),$$

where $A_{\theta,y}\phi_s(x,y)$ denotes the function $A_\theta\phi_s(x,\cdot)$ for s,x fixed evaluated at y. By Itô's formula, under P_θ^ν for any ν,

$$\begin{aligned} \phi_t(X_0,X_t) = \phi_0(X_0,X_0) &+ \int_0^t ds\,\mathcal{H}_\theta\phi_s(X_0,X_s) \\ &+ \int_0^t \sum_{i,j=1}^d \partial_{y_i}\phi_s(X_0,X_s)\sigma_\theta^{ij}(X_s)\,dB_s^j, \end{aligned} \tag{7.2}$$

with $\partial_{y_i}\phi_s(x,z)$ denoting the function $\partial_{y_i}\phi_s(x,\cdot)$ for s,x fixed evaluated at z.

We shall denote by Φ_θ the class of functions ϕ satisfying the continuity and differentiability requirements above, together with

$$\begin{aligned} &E_\theta^\mu \phi_s^2(X_0,X_s) < \infty, \\ &E_\theta^\mu \left(\mathcal{H}_\theta\phi_s(X_0,X_s)\right)^2 < \infty, \\ &E_\theta^\mu \partial_y\phi_s(X_0,X_s)C_\theta(X_s)\partial_y^T\phi_s(X_0,X_s) < \infty, \end{aligned} \tag{7.3}$$

for all s, (so if $\phi_s(x,y) = f(y)$ is a function of y only, this is simply the conditions for $f \in \mathcal{D}_\theta$, see p. 7). The operator $\mathcal{H}_\theta$ acting on functions $\phi_s(x,z)$ in Φ_θ that do not depend on x, is the generator for the *space-time* process $(t,X_t)_{t\geq 0}$.

If $\phi \in \Phi_\theta$, the local martingale in (7.4) is a true $P_\theta^\mu-$martingale and

$$E_\theta^\mu \phi_t(X_0, X_t) = E_\theta^\mu \phi_0(X_0, X_0) + \int_0^t E_\theta^\mu \mathcal{H}_\theta \phi_s(X_0, X_s)\, ds. \tag{7.4}$$

Also, since $P_\theta^\mu = \int_D \mu_\theta(dx)\, P_\theta^x$, the three conditions in (7.3) are satisfied with E_θ^μ replaced by E_θ^x, at least for $\mu_\theta-$almost all x, and then (7.4) also holds with E_θ^x instead of E_θ^μ.

In some cases we shall need to expand the integrand on the right of (7.4). This will be done for $\phi \in \Phi_\theta$, with the function $(s, x, y) \to \mathcal{H}_\theta \phi_s(x, y)$ also in Φ_θ, which ensures that all local martingales are true martingales.

Consider now a given flow $\mathcal{G} = (g_{t,\theta})_{t>0, \theta \in \Theta}$ of well behaved estimating functions. In the remainder of this section we shall assume that for all θ, all components $g_{t,\theta}^k(x, y)$ are one time continuously differentiable in t, two times continuously differentiable in y. We just write $\mathcal{H}_\theta g_{t,\theta}(x, y) \in \mathbb{R}^{p\times 1}$ for $\left(\mathcal{H}_\theta g_t^k(x, y)\right)_k$. One more crucial assumption will now be made: with $\mathcal{G}$ given, a renormalization, replacing $g_{t,\theta}$ by $K_{t,\theta} g_{t,\theta}$, where $K_{t,\theta} \in \mathbb{R}^{p\times p}$ is non-singular and does not depend on x, y, does not change the solutions of the estimating equation (3.4) for any $t = \Delta$, and in particular, does not change $\mathrm{var}_{\Delta,\theta}(g, \hat{\theta})$. We shall now assume that, possibly after renormalization with some $K_{t,\theta}$, the *flow* $(g_{t,\theta})$ *extends continuously to include* $t = 0$ with $g_{t,\theta}(x, y)$ for $t \geq 0$ continuously differentiable in t and twice differentiable in y. In particular we then have a power expansion,

$$g_{t,\theta}(x, y) = g_{0,\theta}(x, y) + t\partial_s g_{0,\theta}(x, y) + o_{\theta,x,y}(t) \tag{7.5}$$

and assume that $g_{0,\theta}$ *is non-vanishing* in the sense that $P_\theta^\mu\left(g_{0,\theta}^k(X_0, X_t) \neq 0\right) > 0$ for all $t > 0$, all θ, all components k.

Example 7.1. *Suppose X is a one-dimensional Ornstein-Uhlenbeck process (so $d = 1$),*

$$dX_t = -\theta X_t\, dt + \sigma\, dB_t, \tag{7.6}$$

cf. Example 4.2. Then $p_{t,\theta,\sigma}(x, \cdot) \sim N\left(e^{-\theta t}x, \sigma^2\left(1 - e^{-2\theta t}\right)/2\theta\right)$ and we shall see how the MLE estimating function $s_t = \dot{p}_t/p_t$ may be renormalized to give a non-trivial s_0 in three different one-parameter models ($p = 1$), as well as in the full model (7.6) ($p = 2$, $\theta > 0, \sigma > 0$).

(i) *$\theta > 0$, $\sigma = 1$. By computation,*

$$\lim_{t\to 0} s_{t,\theta}(x, y) = -\tfrac{1}{2}\left(y^2 - x^2\right),$$

so $s_{0,\theta}(x, y) = -\frac{1}{2}(y^2 - x^2)$, not depending on θ.

(ii) $\theta = 1$, $\sigma > 0$. *With* σ^2 *the parameter*

$$\lim_{t\to 0} t s_{t,\sigma^2}(x,y) = \frac{1}{2\sigma^4}(y-x)^2,$$

so $s_{0,\sigma^2}(x,y) = (y-x)^2/2\sigma^4$.

(iii) $\theta > 0$, $\sigma = \sqrt{2\theta}$. *Then*

$$\lim_{t\to 0} t s_{t,\theta}(x,y) = \frac{1}{4\theta^2}(y-x)^2$$

so $s_{0,\theta}(x,y) = (y-x)^2/4\theta^2$.

(iv) $\theta > 0$, $\sigma > 0$. *Then with* $(\theta_1,\theta_2) = (\sigma^2,\theta)$ *the labeling of the parameters,*

$$\lim_{t\to 0} \begin{pmatrix} t & 0 \\ 0 & 1 \end{pmatrix} s_{t,\sigma^2,\theta}(x,y) = \begin{pmatrix} \frac{1}{2\sigma^4}(y-x)^2 \\ -\frac{1}{2\sigma^2}(y^2-x^2) \end{pmatrix}$$

with $s_{0,\sigma^2,\theta}(x,y)$ *given by the expression on the right.*

Returning to the flow $\mathcal{G}$ satisfying (7.5), we also from now on assume that $\mathcal{G} \subset \Phi$, i.e. for every θ, each component $g^k_{t,\theta}(x,y)$ of g, viewed as a function of (t,x,y), belongs to Φ_θ.

The results about small $\Delta-$optimality will be established first if all $g_{t,\theta} \in \mathcal{M}$, and then in the general case by using the associated martingale estimating functions (4.2). Naturally, studying the behaviour of $g_{t,\theta}$ for $t \to 0$ implies that everything is expressed in terms of $g_{0,\theta}(x,y)$ and its partial derivatives evaluated along the diagonal $y = x$.

If $g_{\Delta,\theta} \in \mathcal{M}$, for all x (with $\mathcal{H}_\theta$ acting on g componentwise),

$$\begin{aligned} 0 &= E^x_\theta g_{\Delta,\theta}(X_0, X_\Delta) \\ &= g_{0,\theta}(x,x) + E^x_\theta \int_0^\Delta ds\, \mathcal{H}_\theta g_{s,\theta}(x, X_s) \\ &= g_{0,\theta}(x,x) + \Delta \mathcal{H}_\theta g_{0,\theta}(x,x) + o(\Delta), \end{aligned} \tag{7.7}$$

and thus

$$g_{0,\theta}(x,x) = 0, \tag{7.8}$$

$$\mathcal{H}_\theta g_{s,\theta}(x,x) = 0. \tag{7.9}$$

(In (7.7) the appearance of a remainder term $o(\Delta)$ is justified provided $s \to E^x_\theta \mathcal{H}_\theta g_{s,\theta}(x, X_s)$ is continuous. Conditions of this nature are deliberately ignored here and below!)

We begin by finding the expansions of $\Lambda_{\Delta,\theta}(g)$ and $E_\theta^\mu\left(g_{\Delta,\theta}g_{\Delta,\theta}^T\right)(X_0,X_\Delta)$.

Notation. In the remainder of this section, if ψ is a function of x,y (or x only) and possibly t,θ,the symbol ψ, when appropriate, will also denote the random variable $\psi(X_0,X_0)$ ($\psi(X_0)$ respectively), conforming with the notation from p. 18 when taking $\Delta=0$.

Proposition 7.2. *Suppose $\mathcal{G}\subset\Phi$ with expansion (7.5) and all $g_{\Delta,\theta}\in\mathcal{M}$. Suppose also that $\dot{g}_{\Delta,\theta}\in\Phi_\theta$, $g_{\Delta,\theta}g_{\Delta,\theta}^T\in\Phi_\theta$. Then*

$$E_\theta^\mu\dot{g}_{\Delta,\theta}(X_0,X_\Delta)=-\Delta E_\theta^\mu\left[\left(\partial_y g_{0,\theta}\right)\dot{b}_\theta+\tfrac{1}{2}\left(\partial_{yy}^2 g_{0,\theta}\right)\dot{C}_\theta\right]+o(\Delta), \tag{7.10}$$

$$E_\theta^\mu\left(g_{\Delta,\theta}g_{\Delta,\theta}^T\right)(X_0,X_\Delta)=\Delta E_\theta^\mu\left(\partial_y g_{0,\theta}\right)C_\theta\left(\partial_y g_{0,\theta}\right)^T+o(\Delta). \tag{7.11}$$

Proof. We have

$$E_\theta^\mu\dot{g}_{\Delta,\theta}(X_0,X_\Delta)=E_\theta^\mu\left[\partial_\theta g_{0,\theta}+\Delta\left(\partial_s\partial_\theta g_{0,\theta}+A_{\theta,y}\left(\partial_\theta g_{0,\theta}\right)\right)\right]+o(\Delta).$$

By (7.8) $\partial_\theta g_{0,\theta}\equiv 0$ while (7.9) implies

$$\begin{aligned}0&=\partial_\theta\left(\partial_s g_{0,\theta}+A_{\theta,y}g_{0,\theta}\right)(x,x)\\&=\left(\partial_s\partial_\theta g_{0,\theta}+A_{\theta,y}\left(\partial_\theta g_{0,\theta}\right)\right)(x,x)+\left(\left(\partial_y g_{0,\theta}\right)\dot{b}_\theta+\tfrac{1}{2}\left(\partial_{yy}^2 g_{0,\theta}\right)\dot{C}_\theta\right)(x,x),\end{aligned}$$

and (7.10) follows. Similarly, again using (7.8),

$$E_\theta^\mu\left(g_{\Delta,\theta}g_{\Delta,\theta}^T\right)(X_0,X_\Delta)=\Delta E_\theta^\mu\left(\partial_s\left(g_{0,\theta}g_{0,\theta}^T\right)+A_{\theta,y}\left(g_{0,\theta}g_{0,\theta}^T\right)\right)+o(\Delta).$$

But (7.8) also gives $\partial_s\left(g_{0,\theta}g_{0,\theta}^T\right)(x,x)\equiv 0$ and that

$$\partial_{y_iy_j}^2\left(g_{0,\theta}^k g_{0,\theta}^l\right)(x,x)=\left(\partial_{y_i}g_{0,\theta}^k\partial_{y_j}g_{0,\theta}^l+\partial_{y_j}g_{0,\theta}^k\partial_{y_i}g_{0,\theta}^l\right)(x,x),$$

whence

$$A_{\theta,y}\left(g_{0,\theta}g_{0,\theta}^T\right)(x,x)=\left(\partial_y g_{0,\theta}\right)C_\theta\left(\partial_y g_{0,\theta}\right)^T(x,x)$$

and (7.11) follows. □

As a direct consequence we obtain

Corollary 7.3. *If $\mathcal{G}$ is as in Proposition 7.2 and*

$$\Lambda_{0,\theta}(g):=E_\theta^\mu\left[\left(\partial_y g_{0,\theta}\right)\dot{b}_\theta+\tfrac{1}{2}\left(\partial_{yy}^2 g_{0,\theta}\right)\dot{C}_\theta\right] \tag{7.12}$$

is non-singular, then

$$\mathrm{var}_{\Delta,\theta}(g,\hat{\theta})=\frac{1}{\Delta}\Lambda_{0,\theta}^{-1}(g)\left(E_\theta^\mu\left(\partial_y g_{0,\theta}\right)C_\theta\left(\partial_y g_{0,\theta}\right)^T\right)\left(\Lambda_{0,\theta}^{-1}(g)\right)^T+o(\frac{1}{\Delta}). \tag{7.13}$$

The corollary will be used to obtain small $\Delta-$optimality in case (i), p. 28 above. For case (ii), if $\partial_y g_{0,\theta}(x,x)$ vanishes, the main term in (7.13) is 0, and a further expansion in (7.11) is required.

Proposition 7.4. *Suppose that* $g_{\Delta,\theta}g_{\Delta,\theta}^T \in \Phi_\theta$, $\mathcal{H}_\theta\left(g_{\Delta,\theta}g_{\Delta,\theta}^T\right) \in \Phi_\theta$ *and that* $\partial_y g_{0,\theta}(x,x) \equiv 0$. *Then*

$$E_\theta^\mu\left(g_{\Delta,\theta}g_{\Delta,\theta}^T\right)(X_0,X_\Delta) = \tfrac{1}{2}\Delta^2 E_\theta^\mu\left[\left(\partial_{yy}^2 g_{0,\theta}\right)C_\theta^{\otimes 2}\left(\partial_{yy}^2 g_{0,\theta}\right)^T\right] + o(\Delta^2). \tag{7.14}$$

If it is only known that $\partial_y g_{0,\theta}^k(x,x) \equiv \partial_y g_{0,\theta}^l(x,x) \equiv 0$ *for some* k,l, *the identity (7.14) still applies to the* kl*'th elements of the matrices involved.*

Note. Recall that $\partial_{yy}^2 g_{0,\theta} \in \mathbb{R}^{p\times d^2}$ and that $C_\theta^{\otimes 2} \in \mathbb{R}^{d^2\times d^2}$ with $(ij),(i'j')$'th element $C_\theta^{ii'}C_\theta^{jj'}$.

Proof. If $\rho_t(x,y) \in \Phi_\theta$ is a real-valued function of (t,x,y) with $\mathcal{H}_\theta\rho \in \Phi_\theta$, by Itô's formula,

$$E_\theta^\mu \rho_\Delta(X_0,X_\Delta) = E_\theta^\mu\left[\rho_0 + \Delta\mathcal{H}_\theta\rho_0 + \tfrac{1}{2}\Delta^2\mathcal{H}_\theta^2\rho_0\right](X_0,X_0) + o\left(\Delta^2\right). \tag{7.15}$$

We are going to apply this expansion to $\rho_t(x,y) = (g_{t,\theta}^k g_{t,\theta}^l)(x,y)$ and first note that (omitting the subscript θ in the remainder of the proof) $\mathcal{H}$ acts on products as follows,

$$\mathcal{H}\left(\varphi_t\psi_t\right)(x,y) = \left(\left(\mathcal{H}\varphi_t\right)\psi_t + \varphi_t\left(\mathcal{H}\psi_t\right)\right)(x,y) + \sum_{i,j=1}^d C^{ij}(y)\left(\partial_{y_i}\varphi_t\partial_{y_j}\psi\right)(x,y). \tag{7.16}$$

Now, with $\varphi = g^k, \psi = g^l$, because of (7.8) and the assumption $\partial_y g_{0,\theta}(x,x) \equiv 0$, applying (7.15) to $\rho_t = g_{t,\theta}^k g_{t,\theta}^l$ gives

$$E_\theta^\mu\left(g_\Delta^k g_\Delta^l\right)(X_0,X_\Delta) = \tfrac{1}{2}\Delta^2 E_\theta^\mu\mathcal{H}^2\left(g_0^k g_0^l\right)(X_0,X_0) + o(\Delta^2). \tag{7.17}$$

But again using (7.16), (7.8), (7.9) and the assumption on $\partial_y g_{0,\theta}$,

$$\mathcal{H}^2\left(g_0^k g_0^l\right)(x,x) = \sum_{i,j=1}^d \mathcal{H}\left(C_y^{ij}\partial_{y_i}g_0^k\partial_{y_j}g_0^l\right)(x,x) \tag{7.18}$$

where the notation C_y^{ij} signifies that the function of (t,x,y) on which $\mathcal{H}$ is acting, involves the factor $C^{ij}(y)$. By (7.16),

$$\mathcal{H}\left(C_y^{ij}\partial_{y_i}g_t^k\partial_{y_j}g_t^l\right) = \mathcal{H}(C_y^{ij})\,\partial_{y_i}g_t^k\partial_{y_j}g_t^l$$

$$+C_y^{ij}\left[\mathcal{H}\left(\partial_{y_i}g_t^k\right)\,\partial_{y_j}g_t^l+\partial_{y_i}g_t^k\,\mathcal{H}\left(\partial_{y_j}g_t^l\right)+\sum_{i',j'}C_y^{i'j'}\partial^2_{y_iy_{i'}}g_t^k\partial^2_{y_jy_{j'}}g_t^l\right]$$
$$+\sum_{i',j'}C_y^{i'j'}\left(\partial_{y_{i'}}C_y^{ij}\right)\partial_{y_{j'}}\left(\partial_{y_i}g_t^k\partial_{y_j}g_t^l\right),$$

and evaluating this for $t=0, y=x$, using the assumption $\partial_y g_0(x,x)=0$ all terms except those involving second derivatives $\partial^2_{yy}g$ vanish, and we are left with

$$\mathcal{H}\left(C_y^{ij}\partial_{y_i}g_0^k\partial_{y_j}g_0^l\right)(x,x)=\sum_{i,j,i',j'}\left(\partial^2_{y_iy_{i'}}g_0^kC_y^{ij}C_y^{i'j'}\partial^2_{y_jy_{j'}}g_0^l\right)(x,x),$$

which inserted into (7.18) and (7.17) gives (7.14). □

We can now state and prove the main result on small $\Delta-$optimality for martingale estimating functions. The three cases corresponds to the listing on page 28.

Theorem 7.5. *Suppose that $\mathcal{G}=(g_{t,\theta})\subset\Phi\cap\mathcal{M}$ is a well behaved flow of martingale estimating functions satisfying the expansion (7.5) with a non-vanishing $g_{0,\theta}$, such that for every $\Delta>0$, $\theta\in\Theta$, $\dot{g}_{\Delta,\theta}\in\Phi_\theta$, $g_{\Delta,\theta}g^T_{\Delta,\theta}\in\Phi_\theta$, and for (ii), (iii) also such that $\mathcal{H}_\theta g_{\Delta,\theta}g^T_{\Delta,\theta}\in\Phi_\theta$.*

(i) *If $C=\sigma\sigma^T$ does not depend on θ and if the $p\times p-$matrices*

$$E_\theta^\mu\left(\partial_y g_{0,\theta}\right)(X_0,X_0)\dot{b}_\theta(X_0),$$
$$E_\theta^\mu\dot{b}_\theta^T(X_0)C^{-1}(X_0)\dot{b}_\theta(X_0)$$

are non-singular, then

$$\mathrm{var}_{\Delta,\theta}(g,\hat{\theta})=\frac{1}{\Delta}v_{-1,\theta}(g,\hat{\theta})+o(\frac{1}{\Delta}),$$

where

$$v_{-1,\theta}(g,\hat{\theta})\geq\left(E_\theta^\mu\dot{b}_\theta^T(X_0)C^{-1}(X_0)\dot{b}_\theta(X_0)\right)^{-1}.$$

Here equality holds, and g is small $\Delta-$optimal, provided

$$\partial_y g_{0,\theta}(x,x)=K_\theta\dot{b}_\theta^T(x)C^{-1}(x)$$

for some constant, non-singular $p\times p-$matrix K_θ.

(ii) *If $C_\theta = \sigma_\theta \sigma_\theta^T$ depends on all parameters $\theta_1, \ldots, \theta_p$ and if the $p \times p-$matrices*

$$E_\theta^\mu \left(\partial_{yy}^2 g_{0,\theta} \right) (X_0, X_0) \dot{C}_\theta(X_0),$$
$$E_\theta^\mu \dot{C}_\theta^T(X_0) \left(C^{\otimes 2}(X_0) \right)^{-1} \dot{C}_\theta(X_0)$$

are non-singular, then

$$\mathrm{var}_{\Delta,\theta}(g, \hat{\theta}) = \frac{1}{\Delta} v_{-1,\theta}(g, \hat{\theta}) + v_{0,\theta}(g, \hat{\theta}) + o(\Delta),$$

with $v_{-1,\theta}(g, \hat{\theta}) = 0$ if $\partial_y g_0(x, x) = 0$ for all x, and

$$v_{0,\theta}\left(g, \hat{\theta}\right) \geq 2 \left(E_\theta^\mu \dot{C}_\theta^T(X_0) \left(C^{\otimes 2}(X_0) \right)^{-1} \dot{C}_\theta(X_0) \right)^{-1}.$$

Here equality holds, and g is small $\Delta-$optimal, if $\partial_y g_{0,\theta}(x, x) \equiv 0$ and

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

for some constant, non-singular $p \times p-$matrix K_θ.

(iii) *If for some p', $1 \leq p' < p$, $C_\theta = \sigma_\theta \sigma_\theta^T$ depends on $\theta_1, \ldots, \theta_{p'}$ but not on $\theta_{p'+1}, \ldots, \theta_p$, and if the matrices*

$$E_\theta^\mu \left[\left(\partial_y g_{0,\theta} \right)(X_0, X_0) \dot{b}_\theta(X_0) + \tfrac{1}{2} \left(\partial_{yy}^2 g_{0,\theta} \right)(X_0, X_0) \dot{C}_\theta(X_0) \right],$$
$$E_\theta^\mu \dot{b}_{2,\theta}^T(X_0) C_\theta^{-1}(X_0) \dot{b}_{2,\theta}(X_0),$$
$$E_\theta^\mu \left(\dot{C}_{1,\theta}^T(X_0) \left(C_\theta^{\otimes 2}(X_0) \right)^{-1} \dot{C}_{1,\theta}(X_0) \right)$$

are non-singular, then

$$\mathrm{var}_{\Delta,\theta}(g, \hat{\theta}) = \frac{1}{\Delta} v_{-1,\theta}(g, \hat{\theta}) + v_{0,\theta}(g, \hat{\theta}) + o(\Delta),$$

where

$$v_{-1,\theta}(g, \hat{\theta}) \geq \begin{pmatrix} 0 & 0 \\ 0 & \left(E_\theta^\mu \dot{b}_{2,\theta}^T(X_0) C_\theta^{-1}(X_0) \dot{b}_{2,\theta}(X_0) \right)^{-1} \end{pmatrix} \tag{7.19}$$

with equality if for some constant $c_\theta \neq 0$

$$\partial_y g_{0,\theta}(x, x) = c_\theta \begin{pmatrix} 0 \\ \dot{b}_{2,\theta}^T(x) C_\theta^{-1}(x) \end{pmatrix}. \tag{7.20}$$

In that case the upper left $p' \times p'$-block $v_{11,0,\theta}$ of $v_{0,\theta}$ satisfies

$$v_{11,0,\theta}\left(g,\hat{\theta}\right) \geq 2\left(E_\theta^\mu\left(\dot{C}_{1,\theta}^T(X_0)\left(C_\theta^{\otimes 2}(X_0)\right)^{-1}\dot{C}_{1,\theta}(X_0)\right)\right)^{-1} \tag{7.21}$$

and here equality holds, and g is small Δ-optimal, if $\partial_y g_{0,\theta}$ satisfies (7.20) and

$$\partial_{yy}^2 g_{1,0,\theta} = \tilde{K}_\theta \dot{C}_{1,\theta}^T(x)\left(C_\theta^{\otimes 2}(x)\right)^{-1}$$

for some constant, non-singular $p' \times p'$-matrix $\tilde{K}_\theta$.

Notation. In (iii) block-matrix notation is used. Thus in (7.19) the interesting part on the right refers to the last $p-p'$ rows and columns of $v_{-1,\theta}$ with $\dot{b}_{2,\theta} \in \mathbb{R}^{d\times(p-p')}$ comprising the last $p-p'$ columns of $\dot{b}_\theta$. Similarly, $\dot{C}_{1,\theta} \in \mathbb{R}^{d^2\times p'}$ consists of the first p' columns of $\dot{C}_\theta$ and $g_{1,0,\theta}$ contains the first p' components of $g_{0,\theta}$.

Proof.

(i) Since C does not depend on θ,

$$\Lambda_{0,\theta}(g) = E_\theta^\mu\left(\partial_y g_{0,\theta}\right)\dot{b}_\theta,$$

cf. (7.12). Now just apply Corollary 7.3 and Lemma 5.1 with $Z = \partial_y g_{0,\theta}$, $Y = \dot{b}_\theta^T$ and $S = C^{-1}$.

(ii) We have that if $\partial_y g_{0,\theta}(x,x) \equiv 0$, then (see (7.12))

$$\Lambda_{0,\theta}(g) = \tfrac{1}{2}E_\theta^\mu\left(\partial_{yy}^2 g_{0,\theta}\right)\dot{C}_\theta.$$

Now use Proposition 7.4 and Lemma 5.1 with $Z = \partial_{yy}^2 g_{0,\theta}$, $Y = \dot{C}_\theta^T$ and $S = \left(C_\theta^{\otimes 2}\right)^{-1}$.

(iii) (7.19) follows from Corollary 7.3 using Lemma 7.6 below with $Z = \partial_y g_{0,\theta}$, $Y = \dot{b}_\theta^T$, $U = \dot{C}_\theta^T$, $V = \frac{1}{2}\partial_{yy}^2 g_{0,\theta}$, and $S = C_\theta^{-1}$. Next, taking $\partial_y g_{0,\theta}$ as in (7.20), in block matrix notation $\Lambda_{0,\theta}(g)$ has the form

$$\begin{pmatrix} 0 & 0 \\ \alpha_1 & \alpha_2 \end{pmatrix} + \begin{pmatrix} \beta_1 & 0 \\ \beta_2 & 0 \end{pmatrix}$$

and with an expansion

$$E_\theta^\mu\left(g_{\Delta,\theta}g_{\Delta,\theta}^T\right)(X_0, X_\Delta) = \Delta\gamma + \Delta^2\begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix} + o(\Delta^2)$$

available, it is an easy matter to see that the upper left corner of $v_{0,\theta}(g,\hat{\theta})$ equals $\beta_1^{-1}\delta_{11}\left(\beta_1^T\right)^{-1}$, which must then be minimized. But $\beta_1 = \frac{1}{2}\partial^2_{yy}g_{1,0,\theta}\dot{C}_\theta$, while by Proposition 7.4, since $\partial_y g_{1,0,\theta} \equiv 0$,

$$\delta_{11} = \tfrac{1}{2}E^\mu_\theta\left[\left(\partial^2_{yy}g_{1,0,\theta}\right)C^{\otimes 2}_\theta\left(\partial^2_{yy}g_{1,0,\theta}\right)^T\right],$$

and the proof is completed using Lemma 5.1 with $Z = \partial^2_{yy}g_{1,0,\theta}$, $Y = \dot{C}_\theta$ and $S = \left(C^{\otimes 2}_\theta\right)^{-1}$. □

Remarks. Part (iii) of Theorem 7.5 is formulated as a mixture of (i) and (ii). Note that we have not minimized $v_{0,\theta}(g)$, which appears hopeless, only the upper left block. Suppose now that g is small $\Delta-$optimal and that

$$W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \sim N_p\left(0, \frac{1}{\Delta}v^{\text{opt}}_{-1,\theta}(g) + v^{\text{opt}}_{0,\theta}(g)\right),$$

with $v^{\text{opt}}_{-1,\theta}(g)$ the lower bound in (7.19) and the upper left block of $v^{\text{opt}}_{0,\theta}(g)$ the lower bound in (7.21), is a Gaussian random vector, which according to the theorem, for Δ small has a distribution close to the asymptotic distribution of $\sqrt{n}\left(\hat{\theta}-\theta\right)$ when the esimating function g is used on observations Δ apart. Thus the variance of W has the form

$$\frac{1}{\Delta}\begin{pmatrix} 0 & 0 \\ 0 & E^\mu_\theta \dot{b}_2^T C_\theta^{-1}\dot{b}_2 \end{pmatrix} + \begin{pmatrix} 2E^\mu_\theta \dot{C}^T_{1,\theta}\left(C^{\otimes 2}_\theta\right)^{-1}\dot{C}_{1,\theta} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix},$$

where the δ-matrices are unknown and depend on unspecified characteristics of g. However, regardless of the values of the δ's,

$$\begin{pmatrix} W_1 \\ \sqrt{\Delta}W_2 \end{pmatrix} \xrightarrow{\mathcal{D}_\theta} \begin{pmatrix} \tilde{W}_1 \\ \tilde{W}_2 \end{pmatrix}$$

as $\Delta \to 0$, where $\tilde{W}_1, \tilde{W}_2$ are independent mean zero Gaussian with covariances $2E^\mu_\theta \dot{C}^T_{1,\theta}\left(C^{\otimes 2}_\theta\right)^{-1}\dot{C}_{1,\theta}$ and $E^\mu_\theta \dot{b}_2^T C_\theta^{-1}\dot{b}_2$. Thus the result is sharp enough to yield the joint distribution of $\left(\hat{\theta}_1,\ldots,\hat{\theta}_{p'}\right)$ and $\left(\hat{\theta}_{p'+1},\ldots,\hat{\theta}_p\right)$ when the two vectors are properly scaled.

Theorem 7.5 (iii) should be compared to Kessler's [10] Theorem 1 and may perhaps be viewed as a generalization to a multidimensional, multiparameter setting of his result.

It was assumed in part (iii) that C_θ depend on $(\theta_1,\ldots,\theta_{p'})$ only. For applications it may be necessary to reparametrize before this is fulfilled. Of course also

any small $\Delta-$optimal g as described in (iii) may be relaced by $K_\theta g$ with K_θ a constant, non-singular $p \times p-$matrix.

If $d = 1$, $p = 1$ the lower bounds in parts (i), (ii) are exactly the variances for the limiting Gaussian experiment found by Woerner [17] in her discussion of local asymptotic normality (LAN).

In the proof of part (iii), Theorem 7.5, the following result was used:

Lemma 7.6. *Suppose that Y, Z, U, V, S are matrix-valued random variables of dimensions $a \times b, a \times b, a \times b', a \times b', b \times b$ respectively with S symmetric and strictly positive definite with probability 1. Suppose further that $U = \binom{U_1}{0}$, where $U_1 \in \mathbb{R}^{a' \times b'}$ comprises the first a' rows of U, and where $1 \leq a' < a$. Assuming that all entries in the matrices $YZ^T, UV^T, ZS^{-1}Z^T, Y_2SY_2^T$ are integrable, writing Y_2 for the last $a - a'$ rows of Y, if $E\left(YZ^T + UV^T\right)$, $EU_1U_1^T$ and $EY_2SY_2^T$ are non-singular, it holds that*

$$\left(E\left(ZY^T + VU^T\right)\right)^{-1} EZS^{-1}Z^T \left(E\left(YZ^T + UV^T\right)\right)^{-1} \geq \begin{pmatrix} 0 & 0 \\ 0 & \left(EY_2SY_2^T\right)^{-1} \end{pmatrix} \tag{7.22}$$

with equality if $EU_1V_1^T$ is non-singular and $Z = \kappa\binom{0}{Y_2S}$ for some constant $\kappa \neq 0$.

Proof. The left hand side M_0 of (7.22) equals $\lim_{h \to 0} M_h$, where M_h is given by the expression for M_0 when replacing $EZS^{-1}Z^T$ by $E\left(ZS^{-1}Z^T + hVV^T\right)$, $h > 0$. But by Lemma 5.1,

$$M_h \geq \left[E\begin{pmatrix} Y & U \end{pmatrix}\begin{pmatrix} S & 0 \\ 0 & \frac{1}{h}I \end{pmatrix}\begin{pmatrix} Y^T \\ U^T \end{pmatrix}\right]^{-1}, \tag{7.23}$$

(the assumptions $EU_1U_1^T > 0$ and $EY_2SY_2^T > 0$ ensuring that the inverse exists for h small enough), and by inspection the determinant of the matrix

$$N_h = \left(E\left(YSY^T + \tfrac{1}{h}UU^T\right)\right)^{-1}$$

on the right of (7.23) is of order $O(h^{-a'})$, while the subdeterminant obtained by deleting the k'th row and l'th column, because the last $a - a'$ rows of U vanish, is of the same order only if $k, l > a'$. Thus we may write

$$N_h = \begin{pmatrix} 0 & 0 \\ 0 & Q_0 \end{pmatrix} + hR + o(h),$$

and it is then an easy matter to verify that $Q_0 = \left(EY_2SY_2^T\right)^{-1}$. (7.22) now follows taking limits in (7.23).

Suppose now that $Z = \kappa\binom{0}{Y_2 S}$. Then $E\left(ZY^T + VU^T\right)$ is non-singular because $EU_1V_1^T$ is, and that equality holds in (7.22) amounts to showing

$$EZS^{-1}Z^T = \left(EZY^T\right)\begin{pmatrix} 0 & 0 \\ 0 & Q_0 \end{pmatrix}\left(EYZ^T\right), \tag{7.24}$$

for $Z = \binom{0}{Y_2 S}$, where we have used that

$$\begin{pmatrix} 0 & 0 \\ 0 & Q_0 \end{pmatrix} U = 0.$$

(7.24) is checked directly. □

Note. The assumption $EU_1U_1^T > 0$ may be replaced by $EU_1\bar{S}U_1^T > 0$ for any random $b' \times b'-$matrix $\bar{S} > 0$ almost surely. This is actually used in Theorem 7.5 (iii).

Let now $\mathcal{G} \subset \Phi$ be a well behaved flow of estimating functions, not necessarily in $\mathcal{M}$. Small $\Delta-$optimality is then discussed applying Theorem 7.5 to the flow of associated martingale estimating functions $\breve{g}_{t,\theta}$, cf. (4.2).

We assume that after a suitable renormalization $\mathcal{G}$ allows the expansion (7.5). To find the corresponding expansion of $(\breve{g}_{t,\theta})$ we use the following result

Lemma 7.7. *Consider a component $g_{t,\theta}^k$ of $g_{t,\theta}$. Then as $t \to 0$,*

$$U_{t,\theta}g_{t,\theta}^{k*}(x) = \frac{1}{t}a_{-1}^k(x) + a_0^k(x) + o(1), \tag{7.25}$$

where

(i) *if $g_{0,\theta}^k(x,x)$ is not identically 0, $a_{-1} \in \mathcal{D}_{0,\theta}$ is the unique solution to*

$$\mathcal{A}_\theta a_{-1}^k(x) = -g_{0,\theta}^k(x,x); \tag{7.26}$$

(ii) *if $g_{0,\theta}^k(x,x) \equiv 0$, then $a_{-1}^k \equiv 0$ and $a_0 \in \mathcal{D}_{0,\theta}$ is the unique solution to*

$$\mathcal{A}_\theta a_0^k(x) = -\left(\partial_s g_{0,\theta}^k + A_{\theta,y}g_{0,\theta}^k\right)(x,x). \tag{7.27}$$

Proof. Applying $I - \pi_{t,\theta}$ to both sides of (7.25) gives

$$g_{t,\theta}^{k*} = \frac{1}{t}\left(I - \pi_{t,\theta}\right)a_{-1}^k + \left(I - \pi_{t,\theta}\right)a_0^k + o(1).$$

But, see (7.5),

$$g^{k*}_{t,\theta}(x) = g^k_0(x,x) + t\left(\partial_s g^k_{0,\theta} + A_{\theta,y} g^k_{0,\theta}\right)(x,x) + o(t)$$

and comparing this with the expansion

$$\pi_{t,\theta} h = h + tA_\theta h + o(t)$$

applied to $h = a^k_{-1}, a^k_0$, the result follows, recalling that the uniqueness of the solutions to the equations determining a^k_{-1}, a^k_0 follows from Proposition 2.1. □

With this result in mind, we define the renormalized martingale flow $(\check{g}_{t,\theta})$ associated with $\mathcal{G}$ as follows (with $g^k_{0,\theta}(x,x) \neq 0$ meaning that $g^k_{0,\theta}(x,x)$ is not identically 0),

$$g^k_{t,\theta}(x,y) + U_{t,\theta} g^{k*}_{t,\theta}(y) - U_{t,\theta} g^{k*}_{t,\theta}(x) = \begin{cases} t^{-1}\check{g}^k_{t,\theta}(x,y) & \text{if } g^k_{0,\theta}(x,x) \neq 0 \\ \check{g}^k_{t,\theta}(x,y) & \text{if } g^k_{0,\theta}(x,x) \equiv 0 \end{cases}$$

so that

$$\check{g}^k_{0,\theta}(x,y) = \begin{cases} a^k_{-1}(y) - a^k_{-1}(x) & \text{if } g^k_{0,\theta}(x,x) \neq 0 \\ g^k_0(x,y) + a^k_0(y) - a^k_0(x) & \text{if } g^k_{0,\theta}(x,x) \equiv 0, \end{cases} \tag{7.28}$$

with a^k_{-1}, a^k_0 determined by (7.26), (7.27) respectively. Applying Theorem 7.5 with $\check{g}$ instead of g now gives the following general result, where e.g. in part (ii) the condition $g_{0,\theta}(x,x) \equiv 0$ is argued as follows: suppose that for some k, $g^k_{0,\theta}(x,x)$ is not identically 0; by Theorem 7.5 (ii), we then need $\partial_y \check{g}^k_{0,\theta}(x,x) \equiv 0$ for all k, and by (7.28) this forces $a^k_{-1} \equiv 0$ and therefore also $g^k_{0,\theta} \equiv 0$.

Theorem 7.8. *Suppose that $\mathcal{G} = (g_{t,\theta}) \subset \Phi$ is a well behaved flow of estimatimg functions satisfying the expansion (7.5) with a non-vanishing $g_{0,\theta}$ and let for every k, a^k_{-1}, a^k_0 denote the solutions to (7.26), (7.27) respectively. Further assume that for every $\Delta > 0$, $\theta \in \Theta$, $\dot{\check{g}}_{\Delta,\theta} \in \Phi_\theta$, $\check{g}_{\Delta,\theta}\check{g}^T_{\Delta,\theta} \in \Phi_\theta$, and for (ii), (iii) also that $\mathcal{H}_\theta \check{g}_{\Delta,\theta}\check{g}^T_{\Delta,\theta} \in \Phi_\theta$.*

(i) *If $C = \sigma\sigma^T$ does not depend on θ and if the $p \times p-$matrices*

$$E^\mu_\theta\left(\partial_y \check{g}_{0,\theta}\right)(X_0, X_0)\dot{b}_\theta(X_0),$$
$$E^\mu_\theta \dot{b}^T_\theta(X_0) C^{-1}(X_0)\dot{b}_\theta(X_0)$$

are non-singular, then $\mathrm{var}_{\Delta,\theta}(g, \hat{\theta})$ has the same expansion with the same lower bound as in Theorem 7.5 (i), and g is small $\Delta-$optimal, provided

$$\partial_y \check{g}_{0,\theta}(x,x) = K_\theta \dot{b}^T_\theta(x) C^{-1}(x)$$

for some constant, non-singular $p \times p-$matrix K_θ. Here, for every k

$$\partial_y \check{g}^k_{0,\theta}(x,x) = \begin{cases} \partial_x a^k_{-1}(x) & \text{if } g^k_{0,\theta}(x,x) \neq 0 \\ \partial_y g^k_0(x,x) + \partial_x a^k_0(x) & \text{if } g^k_{0,\theta}(x,x) \equiv 0. \end{cases} \tag{7.29}$$

(ii) *If $C_\theta = \sigma_\theta \sigma_\theta^T$ depends on all parameters $\theta_1, \ldots, \theta_p$ and if the $p \times p-$matrices*

$$E^\mu_\theta \left(\partial^2_{yy} \check{g}_{0,\theta}\right)(X_0, X_0) \dot{C}_\theta(X_0),$$
$$E^\mu_\theta \dot{C}^T_\theta(X_0) \left(C^{\otimes 2}(X_0)\right)^{-1} \dot{C}_\theta(X_0)$$

are non-singular, then $\mathrm{var}_{\Delta,\theta}(g, \hat{\theta})$ has the same expansion with the same lower bound as in Theorem 7.5 (ii), and g is small $\Delta-$optimal, provided $g_{0,\theta}(x,x) \equiv 0$, $\partial_y g_{0,\theta}(x,x) + \partial_x a_0(x) \equiv 0$ and

$$\partial^2_{yy} \check{g}_{0,\theta}(x,x) = K_\theta \dot{C}^T_\theta(x) \left(C^{\otimes 2}(x)\right)^{-1}$$

for some constant, non-singular $p \times p-$matrix K_θ. Here, for every k

$$\partial^2_{yy} \check{g}^k_{0,\theta}(x,x) = \partial^2_{yy} g^k_0(x,x) + \partial^2_{xx} a^k_0(x).$$

(iii) *If for some p', $1 \leq p' < p$, $C_\theta = \sigma_\theta \sigma_\theta^T$ depends on $\theta_1, \ldots, \theta_{p'}$ but not on $\theta_{p'+1}, \ldots, \theta_p$, and if the matrices*

$$E^\mu_\theta \left[(\partial_y \check{g}_{0,\theta})(X_0, X_0) \dot{b}_\theta(X_0) + \tfrac{1}{2} \left(\partial^2_{yy} \check{g}_{0,\theta}\right)(X_0, X_0) \dot{C}_\theta(X_0)\right],$$
$$E^\mu_\theta \dot{b}^T_{2,\theta}(X_0) C^{-1}_\theta(X_0) \dot{b}_{2,\theta}(X_0),$$
$$E^\mu_\theta \left(\dot{C}^T_{1,\theta}(X_0) \left(C^{\otimes 2}_\theta(X_0)\right)^{-1} \dot{C}_{1,\theta}(X_0)\right)$$

are non-singular, then $\mathrm{var}_{\Delta,\theta}(g, \hat{\theta})$ has the same expansion with the same lower bounds as in Theorem 7.5 (iii), and g is small $\Delta-$optimal, provided $g_{1,0,\theta}(x,x) \equiv 0$ and

$$\begin{aligned} \partial_y \check{g}_{0,\theta}(x,x) &= c_\theta \begin{pmatrix} 0 \\ \dot{b}^T_{2,\theta}(x) C^{-1}_\theta(x) \end{pmatrix}, \\ \partial^2_{yy} \check{g}_{1,0,\theta}(x,x) &= \tilde{K}_\theta \dot{C}^T_{1,\theta}(x) \left(C^{\otimes 2}_\theta(x)\right)^{-1} \end{aligned}$$

for some constant $c_\theta \neq 0$ and some constant, non-singular $\tilde{K}_\theta \in \mathbb{R}^{p' \times p'}$, $\check{g}_{1,0,\theta}$ comprising the first p' components of $\check{g}_{0,\theta}$ and $\dot{b}_{2,\theta}$ comprising the last $p - p'$ columns of $\dot{b}_\theta$. Here, for $k > p'$

$$\partial_y \check{g}^k_{0,\theta}(x,x) = \begin{cases} \partial_x a^k_{-1}(x) & \text{if } g^k_{0,\theta}(x,x) \neq 0 \\ \partial_y g^k_0(x,x) + \partial_x a^k_0(x) & \text{if } g^k_{0,\theta}(x,x) \equiv 0. \end{cases}$$

and for $k \leq p'$

$$\partial^2_{yy}\check{g}^k_{0,\theta}(x,x) = \partial^2_{yy}g^k_0(x,x) + \partial^2_{xx}a^k_0(x).$$

As an illustration of Theorem 7.8 we shall discuss when simple estimating functions (p. 10) are small Δ-optimal.

Suppose that for some k, $g^k_{t,\theta}(x,y) = f_{t,\theta}(x)$ has been properly normalized (see (7.5)) with $g^k_{0,\theta}(x,y) = f_{0,\theta}(x) \neq 0$. Then $g^k_{0,\theta}(x,x) \neq 0$ and $\check{g}^k_{0,\theta}(x,y) = a^k_{-1}(y) - a^k_{-1}(x)$, where $\mathcal{A}_\theta a^k_{-1} = -f_{0,\theta}$. Thus Theorem 7.8 (ii) does not apply and we see that simple estimating functions cannot be small $\Delta-$optimal if C_θ depends on all parameters. If however $C = C_\theta$ does not depend on θ, by Theorem 7.8 (i), $f_{t,\theta}$ satisfies the optimality criterion if

$$\partial_y a^k_{-1} = k\text{'th row of } K_\theta \dot{b}^T_\theta C^{-1}_\theta. \tag{7.30}$$

For $d = 1$, this is easy: just integrate the k'th element of $K_\theta \dot{b}^T_\theta(x) C^{-1}_\theta(x)$. But if $d > 1$, a miracle is required since (7.30) implies that

$$\partial_{y_j}\left(K_\theta \dot{b}^T_\theta C^{-1}_\theta\right)_{ki} = \partial^2_{y_i y_j} a^k_{-1}$$

must be symmetric in i, j. Thus the message is that simple estimating functions can be small Δ-optimal only if $d = 1$ and C_θ does not depend on all the parameters – but perhaps the truly surprising fact is that they can be optimal at all, since for Δ small one might consider transition dependence essential for effective estimation.

If $d = 1$ and the simple estimating function is of the form (3.12), i.e. $f^k_{t,\theta}(x) = \mathcal{A}_\theta h^k(x)$ $(1 \leq k \leq p)$, not depending on t, we find from (7.26) that $g^k_{0,\theta}(x,x) = \mathcal{A}_\theta h^k(x) = -\mathcal{A}_\theta a^k_{-1}(x)$ so $a^k_{-1} = -h^k$, and thus small $\Delta-$optimality is achieved using the flow

$$g_{t,\theta}(x,y) = \mathcal{A}_\theta h(x), \quad \partial_x h(x) = \frac{1}{\sigma^2(x)} \dot{b}^T_\theta(x), \tag{7.31}$$

where $h = \left(h^k\right)_k$ and $C(x) = \sigma^2(x)$ is scalar-valued.

The simple estimating function (7.31) was derived by Helle Sørensen [14] via a quite different approach.

Example 7.9. *As illustration of how small* $\Delta-$*optimality may be achieved, using different types of estimating functions, we consider the Ornstein-Uhlenbeck models from Example 7.1. Thus* $d = 1$.

(i) We have $p = 1$ *and can use (7.31). The result is that* $g_{t,\theta}(x,y) = \mathcal{A}_\theta h(x) = -\theta x h'(x) + \frac{1}{2}h''(x)$ *is small* $\Delta-$*optimal if* $h'(x) = -x$, *i.e.* $\theta x^2 - \frac{1}{2}$ *is small* $\Delta-$*optimal. This is the simple estimating function found by Kessler [9], Section 5.1, which he shows is optimal in* $\mathcal{S}$ *for any* Δ *(cf. Proposition 5.7) with a very high efficiency.*

(ii) *Here we cannot use simple estimating functions to obtain small $\Delta-$optimality, and instead consider martingale estimating functions of the form (3.9), viz.*

$$g_{t,\theta}(x,y) = h_t^{(1)}(x)\left(y - e^{-t}x\right) + h_t^{(2)}(x)\left(y^2 - e^{-2t}x^2 - \frac{\sigma^2}{2}\left(1 - e^{-2t}\right)\right).$$

Assuming that $h_t^{(q)} \to h_0^{(q)}$ for $q = 1, 2$ as $t \to 0$, we find

$$g_{0,\theta}(x,y) = h_0^{(1)}(x)(y-x) + h_0^{(2)}(x)\left(y^2 - x^2\right).$$

According to Theorem 7.5 (ii) we need $\partial_y g_{0,\theta}(x,x) \equiv 0$, i.e.

$$h_0^{(1)}(x) + 2xh_0^{(2)}(x) = 0$$

for all x, and also that $\partial^2_{yy} g_{0,\theta}(x,x)$ is constant, e.g. $2h_0^{(2)} \equiv 1$. Thus

$$g_{t,\theta}(x,y) = -x\left(y - e^{-t}x\right) + \frac{1}{2}\left(y^2 - e^{-2t}x^2 - \frac{\sigma^2}{2}\left(1 - e^{-2t}\right)\right)$$

is small $\Delta-$optimal. Also small $\Delta-$optimal is

$$\begin{aligned} g_{t,\theta}(x,y) &= -e^{-t}x\left(y - e^{-t}x\right) + \frac{1}{2}\left(y^2 - e^{-2t}x^2 - \frac{\sigma^2}{2}\left(1 - e^{-2t}\right)\right) \\ &= \frac{1}{2}\left(y - e^{-t}x\right)^2 - \frac{\sigma^2}{4}\left(1 - e^{-2t}\right) \end{aligned}$$

which, as is easily checked, for all $t = \Delta$ yields the maximum-likelihood estimator!

Of course the two estimating functions behave quite differently for large $t = \Delta$. The example shows that there may be many g that are small $\Delta-$optimal, but how to decide which one to use is for now an open problem.

(iii) *Here also $p = 1$, and for studying small $\Delta-$optimality, this model is equvalent to (ii) with $2\theta = \sigma^2$. For Δ large, inference in the two models differs radically, with no problems estimating σ^2 in (ii), while since μ_θ in (iii) does not depend on θ, estimation of θ for large Δ is virtually impossible.*

(iv) *Now $p = 2$, $p' = 1$ and $g_{t,\sigma^2,\theta}$ is small $\Delta-$optimal if e.g.*

$$\partial_y \check{g}_{0,\sigma^2,\theta}(x,x) = \kappa_{\sigma^2,\theta}\begin{pmatrix} 0 \\ -x/\sigma^2 \end{pmatrix}, \quad \partial^2_{yy}\check{g}_{1,0,\sigma^2,\theta}(x,x) = 1.$$

One finds therefore that

$$g_{t,\theta}(x,y) = \begin{pmatrix} \frac{1}{2}\left(y - e^{-\theta t}x\right)^2 - \frac{\sigma^2}{4}\left(1 - e^{-2\theta t}\right) \\ \theta x^2 - \frac{\sigma^2}{2} \end{pmatrix}$$

is small $\Delta-$optimal.

Our final result is obvious from Theorem 7.8 and Proposition 6.1. It is however quite interesting to verify directly that $\overleftarrow{g}_{t,\theta}$ satisfies the optimality criteria if $g_{t,\theta}$ does!

Proposition 7.10. *In a reversible model, if $(g_{t,\theta})$ is small $\Delta-$optimal, so is the reversed flow $(\overleftarrow{g}_{t,\theta})$ (provided it satisfies the relevant conditions listed in the opening paragraph of Theorem 7.8).* □

Even if a flow $\mathcal{G}$ is small $\Delta-$optimal, there is no guarantee that it will perform decently for Δ large. As mentioned earlier in this section, one may try to overcome this difficulty by combining small $\Delta-$optimal flows with the optimal simple estimating function as $\Delta \to \infty$, e.g. through convex combinations of the form

$$e^{-\lambda\Delta} g_{\Delta,\theta} + \left(1 - e^{-\lambda\Delta}\right) \frac{\dot{\mu}_\theta^T}{\mu_\theta}$$

for some $\lambda > 0$ (perhaps with a separate λ for each component). Here of course g must be chosen so that the convex combination is small $\Delta-$optimal.

More generally, if for a given $\Delta > 0$, $g_{(q),\theta}$ are well-behaved estimating functions, $1 \leq q \leq r$, one may look for constant $p \times p-$matrices $A_{(q)}$ such that the asymptotic covariance for the estimator determined by the estimating function

$$\sum_{q=1}^{r} A_{(q)} g_{(q),\theta}$$

is minimized. The result (as is seen using e.g. Lemma 5.1) is that, in block matrix notation, the $p \times pr-$matrix $\left(A_{(1)} \cdots A_{(r)}\right)$ should equal

$$K_\theta \left(\begin{array}{ccc} \Lambda_{\Delta,\theta}^T(g_{(1),\theta}) & \cdots & \Lambda_{\Delta,\theta}^T(g_{(r),\theta}) \end{array} \right) \left(E_\theta^\mu \left(\begin{array}{ccc} \breve{g}_{(1),\theta}\breve{g}_{(1),\theta}^T & \cdots & \breve{g}_{(1),\theta}\breve{g}_{(r),\theta}^T \\ \vdots & & \vdots \\ \breve{g}_{(r),\theta}\breve{g}_{(1),\theta}^T & \cdots & \breve{g}_{(r),\theta}\breve{g}_{(r),\theta}^T \end{array} \right) \right)^{-1}$$

for some constant, non-singular $K_\theta \in \mathbb{R}^{p\times p}$. This may appear useless for practical purposes, but through expansions for small Δ might perhaps be used to pinpoint better the best among several small $\Delta-$optimal flows.

ACKNOWLEDGEMENT. I would like to thank Michael Sørensen and Mathieu Kessler, not only for providing the inspiration, but also for many fruitful discussions.

References

[1] Aït-Sahalia, Ya. (1997, revised 1998). Maximum likelihood estimation of discretely sampled diffusions: a closed-form approach. Working paper 467, Graduate School of Business, University of Chicago.

[2] Baddeley, A.J. (1995). Time-invariance estimating equations. Research report 22, Department of Mathematics, University of Western Australia.

[3] Bibby, B.M. and Sørensen, M. (1995). Martingale estimating functions for discretely observed diffusion processes. *Bernoulli* **1**, 17-39.

[4] Dacunha-Castelle, D. and Florens-Zmirou, D. (1986). Estimation of the coefficients of a diffusion from discrete observations. *Stochastics* **19**, 263-284.

[5] Florens-Zmirou, D. (1987). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics* **20**, 547-557.

[6] Godambe, V.P. and Heyde, C.C. (1987). Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.* **55**, 231-244.

[7] Hansen, L.P. and Scheinkman, J.A. (1995). Back to the future: generating moment implications for continuous-time Markov processes. *Econometrica* **63**, 767-804.

[8] Kessler, M. (1995). Martingale estimating functions for a Markov chain. Preprint, Laboratoire de Probabilités, Université de Paris VI.

[9] Kessler, M. (1996). Simple and explicit estimating functions for a discretely observed diffusion process. Research report 336, Department of Theoretical Statistics, University of Aarhus. (To appear in *Scand. J. Statist*).

[10] Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.* **24**, 211-229.

[11] Kessler, M. and Sørensen, M. (1995). Estimating equations based on eigenfuctions for a discretely observed diffusion process. Research report 332, Department of Theoretical Statistics, University of Aarhus. (To appear in *Bernoulli*).

[12] Pedersen, A.R. (1994). Quasi-likelihood inference for discretely observed diffusion processes. Research report 295, Department of Theoretical Statistics, University of Aarhus.

[13] Pedersen, A.R. (1995). A new approach to maximum likelihood estimation for stocahstic differential equations based on discrete observations. *Scand. J. Statist.* **22**, 55-71.

[14] Sørensen, H. (1998). Approximation of the score function for diffusion processes. Preprint 8, Department of Theoretical Statistics, University of Copenhagen.

[15] Sørensen, M. (1997). 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, pp. 305-325.

[16] Sørensen, M. (1998). On asymptotics of estimating functions. Preprint 6, Department of Theoretical Statistics, University of Copenhagen.

[17] Woerner J.H.C. (1998). LAN for discretely observed diffusion processes in the ergodic case and applications to martingale estimating functions. Preprint Nr. 1, Albert-Ludwigs-Universität Freiburg, Mathematische Fakultät.

Martin Jacobsen
Department of Theoretical Statistics
University of Copenhagen
5 Universitetsparken
2100 Copenhagen Ø
Denmark

martin@math.ku.dk