Bernoulli **8**(5), 2002, 643–668

Optimality and small Δ-optimality of martingale estimating functions

MARTIN JACOBSEN

Department of Statistics and Operations Research, University of Copenhagen, 5 Universitetsparken, DK-2100 Copenhagen Ø, Denmark. E-mail: martin@math.ku.dk

Martingale estimating functions determined from a given collection (the base) of conditional expectations are considered for estimating the parameters of a discretely observed diffusion. Small Δ-optimality of these functions (i.e. near-efficiency when the observations are close together) is discussed, and in particular it is shown that this can be achieved provided the base is large enough. It is also shown that the optimal martingale estimating function with a given base is automatically small Δ-optimal, provided only that the base is sufficiently large. In both cases the critical dimension of the base is the same and determined by the dimension of the diffusion, and by whether the squared diffusion matrix is parameter-dependent or not; the critical number does not depend, however, on the dimension of the parameter.

Keywords: asymptotic variance; diffusions; estimating functions; generalized Cox–Ingersoll–Ross process; martingales; optimality; small Δ-optimality

1. Introduction

Suppose a d-dimensional, ergodic, time-homogeneous diffusion X is observed at finitely many equidistant points $i\Delta$ in time, and $i = 0, \ldots, n$ with $\Delta > 0$ is the interval between neighbouring observations. In order to estimate the unknown parameter θ that determines the distribution of X, rather than using maximum likelihood, which may well prove unfeasible, one often resorts to unbiased estimating functions, some of the most successful of which are based on conditional expectations, resulting in estimating equations of the form

$$\sum_{i=1}^{n}\sum_{q=1}^{r} h^q(X_{(i-1)\Delta})(f^q(X_{i\Delta}) - \mathrm{E}_\theta(f^q(X_{i\Delta})|X_{(i-1)\Delta})) = 0, \tag{1.1}$$

where, for the moment, we consider θ to be one-dimensional (see (2.6) and (2.1) below for the general set-up). Equation (1.1) is the prime example of an estimating equation obtained from an unbiased *martingale* estimating function.

The study of estimating equations of the form (1.1) was initiated by Bibby and Sørensen (1995) who focused on the case where $r = 1$ and $f(x) = x$, for which they showed that under mild conditions (1.1) has a consistent and $\sqrt{n}$-asymptotically Gaussian solution in θ (as $n \to \infty$ with $\Delta > 0$ fixed). These asymptotic results readily generalize to other types of unbiased estimating equations – see, for example, Sørensen (1999).

In the present paper we consider estimating equations of the form (1.1) for d-dimensional diffusions X with a p-dimensional parameter θ. The main issue is the discussion of the choice of *base* $(f^q)_{1 \leqslant q \leqslant r}$, in particular the choice of r, the dimension or rank of the base, and the choice of *weights* h^q. For this purpose we think of $\Delta > 0$ as arbitrary and consider families (for Δ varying) of estimating equations of the form (1.1) where h^q, but not f^q (the base should be the same for all Δ), is allowed to depend on Δ, and such that, for any fixed $\Delta > 0$, (1.1) has a consistent and asymptotically Gaussian solution as described above.

Given the base (f^q), for any given $\Delta > 0$ there is an *optimal* choice of weights; see Proposition 2.1 below. The resulting estimator will typically not be efficient, while an estimator that is *small* Δ-*optimal* will be nearly efficient for small values of Δ – and, of course, still consistent and asymptotically Gaussian for all Δ, although not optimal. As we shall see, while it may be difficult to find the optimal estimator (to obtain and use the weights one needs the explicit form of the inverse of an $r \times r$ matrix with all elements conditional variances), it is quite easy to determine weights that lead to small Δ-optimal estimators.

The concept of small Δ-optimality was introduced by Jacobsen (2001a), and the main purpose of the present paper is to discuss conditions for small Δ-optimality for the type of martingale estimating functions underlying (1.1).

The first main result, Theorem 2.2, shows that given the base, provided only that the dimension r is large enough, there always exist weights such that small Δ-optimality is achieved. Furthermore, it is easy to find the weights – it is just a matter of solving, at any point x in the range for the diffusion, a set of linear equations, and in the statement of the theorem a concrete solution is exhibited.

The second main result, Theorem 2.3, shows that for any base, again provided r is large enough, for the optimal choice of weights from Proposition 2.1 below, small Δ-optimality is automatic.

In both theorems the same critical value r_0 for the dimension of the base appears: for $r \geqslant r_0$ small Δ-optimality can be achieved for any base, while for $r < r_0$ this may only be possible, if at all, for a special choice of base – for an important example of this, see Remark 2 below. The value of r_0 depends on the structure of the model but not on the dimension of the parameter, with $r_0 = d$, the dimension of the diffusion, if the squared diffusion matrix does not depend on the parameter, and $r_0 = d(d+3)/2$ otherwise. Thus one natural choice of base is $f^i(x_1, \ldots, x_d) = x_i$ if $r_0 = d$, and these f^i supplemented by $f^{ij}(x_1, \ldots, x_d) = x_i x_j$ for $1 \leqslant i \leqslant j \leqslant d$ if $r_0 = d(d+3)/2$.

The paper is concluded (Section 3) with two examples, one describing a generalized Cox–Ingersoll–Ross model, the second the finite-dimensional Ornstein–Uhlenbeck processes. For the latter it turns out that, using the base of first- and second-order moments (f^i, f^{ij}) described above, the concrete small Δ-optimal estimating function exhibited in Theorem 2.2 for any Δ yields the maximum likelihood estimator.

Although, both here and in Jacobsen (2001a), small Δ-optimality is discussed exclusively for diffusions, it should be pointed out that the concept makes perfect sense for any model involving discrete observations from an ergodic, time-homogeneous Markov process in continuous time.

2. Optimality and small Δ-optimality

Let $X = (X_t)_{t \geq 0}$ be a d-dimensional ergodic diffusion solving the stochastic differential equation

$$\mathrm{d}X_t = b_\theta(X_t)\mathrm{d}t + \sigma_\theta(X_t)\mathrm{d}B_t, \qquad X_0 = U,$$

where $b_\theta(x) \in \mathbb{R}^{d\times 1}$, $\sigma_\theta(x) \in \mathbb{R}^{d\times d}$, B is a standard d-dimensional Brownian motion and U is a d-dimensional random variable, independent of B. Both the drift b_θ and the diffusion coefficient σ_θ are allowed to depend on the p-dimensional parameter $\theta \in \Theta$, with the parameter set Θ an open subset of $\mathbb{R}^p$ (typically delimited by the stationarity requirement). The invariant distribution for X is denoted μ_θ, i.e. if U has distribution μ_θ, then X is a strictly stationary Markov process.

We shall assume that X takes its values within some open subset D of $\mathbb{R}^d$, which of course is not allowed to depend on θ. We assume that $C_\theta(x) := \sigma_\theta(x)\sigma_\theta^{\mathrm{T}}(x) \succ 0$ for all θ and all $x \in D$, i.e. the symmetric positive semidefinite matrix $C_\theta(x)$ is assumed to be strictly positive definite always ($^{\mathrm{T}}$ denotes matrix transposition).

We shall also write μ_θ for the density of μ_θ and assume that, for all θ, $\mu_\theta > 0$ everywhere on D. The transition density is denoted $p_{t,\theta}(x, y)$:

$$P_\theta(X_{s+t} \in \mathrm{d}y | X_s = x) = p_{t,\theta}(x, y)\mathrm{d}y.$$

The underlying probability P_θ depends not only on θ but also on the distribution of X_0. It is denoted P_θ^μ if X_0 has distribution μ_θ and P_θ^x if $X_0 \equiv x$. The corresponding expectations are written E_θ^μ and E_θ^x.

We shall denote by $Q_{t,\theta}$ the joint distribution of (X_s, X_{s+t}) under P_θ^μ (for any s), and by $q_{t,\theta}$ the density of $Q_{t,\theta}$:

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

Finally, the transition operators for X are denoted $\pi_{t,\theta}$,

$$\pi_{t,\theta}f(x) = \mathrm{E}_\theta^x f(X_t),$$

provided the integral makes sense (e.g. for f bounded or $f \in L^1(\mu_\theta)$), and the differential operator determining the infinitesimal generator is denoted A_θ,

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

for sufficiently smooth functions f.

Suppose now that X is observed at finitely many equidistant time-points, X_0, $X_\Delta, \ldots, X_{n\Delta}$. We shall discuss asymptotic optimality properties of estimators based on martingale estimating functions, i.e. the estimator $\hat{\theta}_n$ of θ is found by solving the estimating equation

$$G_{n,\Delta}(\theta) = \sum_{i=1}^n g_{\Delta,\theta}(X_{(i-1)\Delta}, X_{i\Delta}) = 0, \tag{2.1}$$

where $(t, \theta, x, y) \to g_{t,\theta}(x, y)$ is a p-variate function such that each coordinate $g^k_{t,\theta}$ satisfies the martingale condition

$$g^{k*}_{t,\theta}(x) = 0 \text{ for all } x, \qquad g^{k*}_{t,\theta}(x) := \mathrm{E}^x_\theta g^k_{t,\theta}(X_0, X_t), \tag{2.2}$$

ensuring that $(G_{n,\Delta}(\theta))_{n\geqslant 1}$ is a p-dimensional P_θ-martingale (whatever the initial distribution of X).

At this point we wish to emphasize that the only type of asymptotics considered in this paper is that consisting of fixing $\Delta > 0$ and letting n, the number of observations, tend to ∞ – in particular, we do not consider schemes where $\Delta = \Delta_n \to 0$ and $n \to \infty$ simultaneously. It is of critical importance, however, that we allow Δ to be arbitrary, with an estimating function available for each $t = \Delta$ (as is natural with the estimating functions (2.6) below, originating from a given base (f^q) that does not depend on Δ), and hence we obtain an asymptotic covariance matrix (see (2.4)) for each Δ. It is the properties of this function of covariances that are used to define the concept of small Δ-optimality (see the detailed discussion following the proof of Proposition 2.1 below).

Following the terminology in Jacobsen (2001a), we shall refer to $\mathcal{G} = (g_{t,\theta})_{t>0,\theta\in\Theta}$, where the $g_{t,\theta}$ satisfy (2.2), as a *well-behaved flow of martingale estimating functions*, $\mathcal{G} \subset \mathcal{M}$ (the space of flows of martingale estimating functions), if each $g^k_{t,\theta} \in L^2(Q_{t,\theta})$, with

$$\mathrm{E}^\mu_\theta g_{t,\theta'}(X_0, X_t) = 0 \text{ if and only if } \theta = \theta'; \tag{2.3}$$

if, furthermore, $\mathrm{E}^\mu_\theta(g_{t,\theta} g^{\mathrm{T}}_{t,\theta})(X_0, X_t) \succ 0$ for all t, θ; if $\partial_{\theta_\ell} g^k_{t,\theta} \in L^1(Q_{t,\theta})$ for all t, θ, k, ℓ; and finally, if, for every θ and every fixed $t = \Delta > 0$, there is with P^μ_θ-probability tending to 1 a consistent solution $\hat\theta_n = \hat\theta_n(\Delta)$ (henceforth, Δ is suppressed from the notation) to (2.1) such that $\sqrt{n}(\hat\theta_n - \theta)$ converges in distribution for $n \to \infty$ to the p-dimensional Gaussian distribution with mean vector 0 and covariance matrix $\mathrm{var}_{\Delta,\theta_0}(g, \hat\theta)$ (with $\hat\theta$ denoting the sequence $(\hat\theta_n)$ of estimators) given by

$$\mathrm{var}_{\Delta,\theta_0}(g, \hat\theta) = \Lambda^{-1}_{\Delta,\theta_0}(g)\mathrm{E}^\mu_{\theta_0}\left(g_{\Delta,\theta_0} g^{\mathrm{T}}_{\Delta,\theta_0}\right)(X_0, X_\Delta)\left(\Lambda^{-1}_{\Delta,\theta_0}(g)\right)^{\mathrm{T}}. \tag{2.4}$$

Here

$$\Lambda_{t,\theta} := \mathrm{E}^\mu_\theta \dot g_{t,\theta}(X_0, X_t), \tag{2.5}$$

the dot signifying differentiation with respect to θ so that $\dot g_{t,\theta}(x, y) \in \mathbb{R}^{p\times p}$ is given by

$$(\dot g_{t,\theta}(x, y))_{k\ell} = \partial_{\theta_\ell} g^k_{t,\theta}(x, y).$$

Condition (2.3) specifies that the estimation function is *unbiased*, as follows from (2.2), and that it *identifies* the parameters. In Section 3.2 an example is given where (2.3) does not hold.

The reader is reminded that asymptotic normality of $\hat\theta$, as specified above, holds under quite weak assumptions (see Sørensen 1999) and that certainly (2.4) is the natural expression for the asymptotic covariance. The most critical among the assumptions needed is that $\Lambda_{t,\theta} \in \mathbb{R}^{p\times p}$ must be non-singular for all t, θ.

Throughout the paper, derivatives are understood as matrices by analogy with (2.5): if ϕ is a ρ-variate function of a υ-dimensional variable $z = (z_1, \ldots, z_\upsilon) \in \mathbb{R}^\upsilon$, $\partial_z \phi$ denotes the

$\rho \times v$ matrix of partial derivatives with mth row $(\partial_{z_1}\phi^m, \ldots, \partial_{z_v}\phi^m)$. The dot notation is used exclusively for differentiation with respect to θ, $\dot{\phi} = \partial_\theta \phi$.

In the remainder of the paper we shall focus on martingale estimating functions derived from conditional expectations of given functionals, i.e. we assume that

$$g^k_{t,\theta}(x, y) = \sum_{q=1}^{r} h^{qk}_{t,\theta}(x)(f^q_\theta(y) - (\pi_{t,\theta} f^q_\theta)(x)), \tag{2.6}$$

or, in matrix notation,

$$g_{t,\theta}(x, y) = h^{\mathrm{T}}_{t,\theta}(x)(f_\theta(y) - (\pi_{t,\theta} f_\theta)(x)),$$

with $h_{t,\theta}(x) \in \mathbb{R}^{r\times p}$, $f_\theta(x) \in \mathbb{R}^{r\times 1}$. (The integrability assumptions imposed on general g above make it natural to assume here that f^q_θ and $h^{qk}_{t,\theta} \in L^4(\mu_\theta)$, while θ-derivatives of f^q_θ and $h^{qk}_{t,\theta}$ must belong to $L^2(\mu_\theta)$. We shall not be too concerned about these conditions – it is tacitly assumed everywhere that the flow $\mathcal{G}$ given by (2.6) is well behaved.)

Estimating functions of the form (2.6) were first used by Bibby and Sørensen (1995); see also Jacobsen (2001a, Section 3) for an overview.

We shall refer to the functions $(f^1_\theta, \ldots, f^r_\theta)$ as the *base* for the flow of estimating functions given by (2.6). The problem studied in this paper is that of finding good choices for the dimension of the base and for the *weights* $h_{t,\theta}$ given the base.

Assumption A. *The functions $f^q_\theta(x)$ are supposed to be differentiable in θ and twice differentiable in x. Also, the base $(f^1_\theta, \ldots, f^r_\theta)$ is supposed to have full affine rank r on the domain D for all θ, i.e. for an arbitrary θ the identity*

$$\sum_{q=1}^{r} a^q_\theta f^q_\theta(x) + \alpha_\theta = 0, \qquad x \in D,$$

for constants a^q_θ, α_θ, implies $a^1_\theta = \ldots = a^r_\theta = \alpha_\theta = 0$.

The functions $h^{qk}_{t,\theta}$ are supposed to be such that, for any t, θ, the p r-variate functions $x \to (h^{1k}_{t,\theta}(x), \ldots, h^{rk}_{t,\theta}(x))$ forming the columns of $h_{t,\theta}$ are linearly independent on D.

Note that if the base $(f^1_\theta, \ldots, f^r_\theta)$ does not have full affine rank, then there is a representation (2.6) of the $g^k_{t,\theta}$ with r replaced by $r-1$. The condition that the base has full rank is equivalent to assuming that the r d-variate functions $\partial_x f^q_\theta$, for $1 \leqslant q \leqslant r$, are linearly independent. In the main results, Theorems 2.2 and 2.3 below, Assumption A is supplemented by conditions on the pointwise behaviour of $\partial_x f_\theta$ and $\partial^2_{xx} f_\theta$.

If, for some t, the columns of $h_{t,\theta}$ are not linearly independent for all θ, i.e. there exists $\beta_{t,\theta} \in \mathbb{R}^{p\times 1}\backslash 0$ such that $h_{t,\theta}(x)\beta_{t,\theta} = 0$ for all x, then $\beta^{\mathrm{T}}_{t,\theta} g_{t,\theta}(x, y) = 0$ for all x, y, so that one of the p estimating equations in (2.1) can be obtained from the others and it is impossible to estimate all p parameters θ_ℓ – formally, both matrices $\Lambda_{t,\theta}(g)$ and $\mathrm{E}^\mu_\theta(g_{t,\theta} g^{\mathrm{T}}_{t,\theta})(X_0, X_t)$ become singular and (2.4) does not make sense.

Note that we allow the base $(f^1_\theta, \ldots, f^r_\theta)$ to depend on θ, but *not* on t, i.e. the same base is used for all t.

For a given base, it is easy to determine the optimal choices for the $h_{t,\theta}^{qk}$, i.e. the choices minimizing $\mathrm{var}_{t,\theta}(g, \hat{\theta})$ given by (2.4). We use the notation $A \geqslant B$ between symmetric, positive semidefinite matrices to signify that $A - B$ is also positive semidefinite.

Proposition 2.1. *Assume that, for all x, t and θ, the symmetric $r \times r$ matrix*

$$\Pi_{t,\theta} f_\theta(x) := \pi_{t,\theta}\big(f_\theta f_\theta^{\mathrm{T}}\big)(x) - (\pi_{t,\theta} f_\theta)(x)\big(\pi_{t,\theta} f_\theta^{\mathrm{T}}\big)(x)$$

is non-singular, and define

$$h_{t,\theta}^{\mathrm{opt}} = (\Pi_{t,\theta} f_\theta)^{-1}\big(\partial_\theta(\pi_{t,\theta} f_\theta) - \pi_{t,\theta}\big(\dot{f}_\theta\big)\big). \tag{2.7}$$

Provided that differentiation and integration can be interchanged in

$$\partial_\theta \int p_{t,\theta}(x, y) f_\theta(y) \mathrm{d}y = \int \partial_\theta(p_{t,\theta}(x, y) f_\theta(y)) \mathrm{d}y$$

and the flow $\mathcal{G}^{\mathrm{opt}}$ given by

$$g_{t,\theta}^{\mathrm{opt}}(x, y) = \Big(h_{t,\theta}^{\mathrm{opt}}\Big)^{\mathrm{T}}(x)(f_\theta(y) - \pi_{t,\theta} f_\theta(x)) \tag{2.8}$$

is well behaved, then

$$\mathrm{var}_{t,\theta}(g^{\mathrm{opt}}, \hat{\theta}) \leqslant \mathrm{var}_{t,\theta}(g, \hat{\theta})$$

for any well-behaved flow $\mathcal{G} = (g_{t,\theta})$ of the form (2.6) *with base $(f_\theta^1, \ldots, f_\theta^r)$.*

Proof. Since f is allowed to depend on θ, this extends (2.10) in Bibby and Sørensen (1995) and Example 4 in Jacobsen (2001a), so we merely indicate the proof. By the projection theorem (Kessler 2000, Proposition 1; Jacobsen 2001a, Proposition 3), $g_{t,\theta}^{k,\mathrm{opt}}$ is found by projecting the kth coordinate of the score function, $\partial_{\theta_k} p_{t,\theta}(x, y)/p_{t,\theta}(x, y)$, onto the subspace of $L^2(Q_{t,\theta})$ spanned by functions of the form (2.6) with the f_θ^q fixed and arbitrary $h_{t,\theta}^{qk} \in L^2(\mu_\theta)$. Thus $h_{t,\theta}^{qk,\mathrm{opt}}$ satisfies, for all $1 \leqslant q_0 \leqslant r$, $1 \leqslant k \leqslant p$ and all $h \in L^2(\mu_\theta)$, the equality

$$\begin{aligned} 0 = \mathrm{E}_\theta^\mu \Bigg[& \frac{\partial_{\theta k} p_{t,\theta}}{p_{t,\theta}}(X_0, X_t) - \sum_{q=1}^{r} h_{t,\theta}^{qk,\mathrm{opt}}(X_0)(f_\theta^q(X_t) - \pi_{t,\theta} f_\theta^q(X_0)) \Bigg] \\ & \times h(X_0)(f_\theta^{q_0}(X_t) - \pi_{t,\theta} f_\theta^{q_0}(X_0)). \end{aligned} \tag{2.9}$$

Using the fact that $\dot{\boldsymbol{p}}_{t,\theta}/p_{t,\theta}$ is a martingale estimating function and that

$$\begin{aligned} \mathrm{E}_\theta^x \frac{\partial_{\theta k} p_{t,\theta}}{p_{t,\theta}}(X_0, X_t) f_\theta^{q_0}(X_t) &= \int \partial_{\theta_k} p_{t,\theta}(x, y) f_\theta^{q_0}(y) \mathrm{d}y \\ &= \partial_{\theta_k} \int p_{t,\theta}(x, y) f_\theta^{q_0}(y) \mathrm{d}y - \int p_{t,\theta}(x, y) \partial_{\theta_k} f_\theta^{q_0}(y) \mathrm{d}y, \end{aligned}$$

(2.9) may be written

$$0 = \mathrm{E}_\theta^\mu h(X_0)\Bigg\{ \partial_{\theta_k}(\pi_{t,\theta} f_\theta^{q_0})(X_0) - \pi_{t,\theta}(\partial_{\theta_k} f_\theta^{q_0})(X_0) \\ - \sum_{q=1}^{r} h_{t,\theta}^{qk,\mathrm{opt}}(X_0)(\pi_{t,\theta}(f_\theta^q f_\theta^{q_0})(X_0) - \pi_{t,\theta} f_\theta^q(X_0)\pi_{t,\theta} f_\theta^{q_0}(X_0)) \Bigg\}$$

for all $h \in L^2(\mu_\theta)$, i.e. the expression in braces must vanish P_θ^μ-almost surely and the result follows. □

Proposition 2.1 is a result on *local optimality*, i.e. it exhibits, for any given $t = \Delta > 0$, the best member of a given, restricted class of estimating functions – best from the point of view of minimizing the asymptotic covariance of the resulting estimator when $n \to \infty$; cf. the concept of *A*-optimality (Heyde 1988). But only in exceptional cases will this choice be *globally optimal*, i.e. the (locally) optimal estimator will be efficient with respect to the maximum likelihood estimator.

By contrast, the concept of small Δ-optimality introduced by Jacobsen (2001a, Section 6) gives conditions for global optimality not for any given $\Delta > 0$ but only for $\Delta \to 0$ in the following sense: for any given $\Delta > 0$, when $n \to \infty$ we still have the asymptotic covariance $\mathrm{var}_{\Delta,\theta_0}(g, \hat{\theta})$ from (2.4), but now consider it as a function of Δ and, rather than minimizing for $\Delta > 0$ fixed, use the fact that there is an expansion of the covariance in powers of Δ (see (2.10) below) and that there are universal (not depending on $\mathcal{G} = (g_{t,\theta})$) lower bounds on one or more of the leading coefficient matrices in this expansion; see Jacobsen (2001a, Section 6). Small Δ-optimality holds if these lower bounds are achieved by the flow $\mathcal{G}$, and sufficient conditions for this are contained in the main result, Theorem 1, in Jacobsen (2001a), that we now recapitulate.

With $\mathcal{G} \subset \mathcal{M}$ a well-behaved flow of estimating functions, it is first of all essential to assume that there is a smooth extension of $g_{t,\theta}(x, y)$ (which is defined only for $t > 0$) to allow $t = 0$, i.e. after a possible renormalization of $g_{t,\theta}$ by a factor (non-zero scalar or non-singular $p \times p$ matrix) depending on t, θ but not on x, y (so the solution of (2.1) is not affected), the limit

$$g_{0,\theta}(x, y) = \lim_{t \to 0} g_{t,\theta}(x, y)$$

must exist with $(x, y) \to g_{0,\theta}(x, y)$ not identically 0, of full rank p in a suitable sense (see Jacobsen 2001a, Theorem 1) and sufficiently smooth as required by the conditions below.

With this smooth extension of $g_{t,\theta}$ available, using Itô–Taylor expansions of the random variables appearing in (2.4), it is shown in Jacobsen (2001a) that, subject to important integrability conditions (see below), the asymptotic covariance for $\hat{\theta}$ is given, as $\Delta \to 0$, by

$$\mathrm{var}_{\Delta,\theta}(g, \hat{\theta}) = \frac{1}{\Delta} v_{-1,\theta}(g, \hat{\theta}) + v_{0,\theta}(g, \hat{\theta}) + o(1) \tag{2.10}$$

with (complicated) coefficient matrices, for instance for case (i) below; see Jacobsen (2001a, Corollary 1). For the discussion of small Δ-optimality, three cases for the structure of the

diffusion model now arise (to achieve the structure in (iii) it may be necessary first to reparametrize the model):

(i) $C_\theta = C$ *does not depend on* θ. Then the main term in (2.10) is always present and small Δ-optimality is achieved by globally (over all g) minimizing $v_{-1,\theta}(g, \hat{\theta})$. A sufficient condition for a given flow $(g_{t,\theta})$ to be small Δ-optimal is that

$$\partial_y g_{0,\theta}(x, x) = K_\theta \dot{b}_\theta^{\mathrm{T}}(x) C^{-1}(x) \tag{2.11}$$

for some non-singular $K_\theta \in \mathbb{R}^{p\times p}$. ($\partial_y g_{0,\theta}(x, x)$ evaluates $\partial_y g_{0,\theta}(x, y)$ along the diagonal $y = x$.)

(ii) C_θ *depends on all parameters* $\theta_1, \ldots, \theta_p$. Then the main term in (2.10) vanishes provided $\partial_y g_{0,\theta}(x, x) \equiv 0$ and small Δ-optimality is achieved by minimizing $v_{0,\theta}(g, \hat{\theta})$. A sufficient condition for $(g_{t,\theta})$ to be small Δ-optimal is that

$$\partial_y g_{0,\theta}(x, x) = 0, \qquad \partial^2_{yy} g_{0,\theta}(x, x) = K_\theta \dot{C}_\theta^{\mathrm{T}}(x) \big(C_\theta^{\otimes 2}(x)\big)^{-1} \tag{2.12}$$

for some non-singular $K_\theta \in \mathbb{R}^{p\times p}$. (Here $\dot{C}_\theta(x) \in \mathbb{R}^{d^2\times p}$ with $(\dot{C}_\theta(x))_{ij,k} = \partial_{\theta_k} C^{ij}_\theta(x)$.)

(iii) C_θ *depends on the parameters* $\theta_1, \ldots, \theta_{p'}$, *but not on* $\theta_{p'+1}, \ldots, \theta_p$, *for some* p' *with* $1 \leqslant p' < p$. Then parts of the main term in (2.10) can be made to disappear so that

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

Furthermore, the matrix $v_{22,-1,\theta}(g, \hat{\theta}) \in \mathbb{R}^{(p-p')\times(p-p')}$ can be minimized and small Δ-optimality is achieved by, in addition, minimizing the upper left block $v_{11,0,\theta}(g, \hat{\theta})$ of $v_{0,\theta}(g, \hat{\theta})$. A sufficient condition for small Δ-optimality is that

$$\partial_y g_{0,\theta}(x, x) = c_\theta \begin{pmatrix} 0_{p'\times d} \\ \dot{b}_{2,\theta}^{\mathrm{T}}(x) C_\theta^{-1}(x) \end{pmatrix},$$

$$\partial^2_{yy} g_{1,0,\theta}(x, x) = K'_\theta \dot{C}_{1,\theta}^{\mathrm{T}}(x) \big(C_\theta^{\otimes 2}(x)\big)^{-1} \tag{2.13}$$

for some constant $c_\theta \neq 0$ and some non-singular $K'_\theta \in \mathbb{R}^{p'\times p'}$. ($\dot{b}_{2,\theta} \in \mathbb{R}^{d\times(p-p')}$ comprises the last $p - p'$ columns of $\dot{b}_\theta$, $g_{1,0,\theta}$ the first p' coordinates of $g_{0,\theta}$, and $\dot{C}_{1,\theta} \in \mathbb{R}^{d^2\times p'}$ the first p' columns of $\dot{C}_\theta$.)

As mentioned above, to check for small Δ-optimality more is required than just checking (2.11), (2.12) or (2.13): it must be verified that various matrices involving expectations of quantities related to $\dot{b}$, $\dot{C}$, $\partial_y g_{0,\theta}$ and $\partial^2_{yy} g_{0,\theta}$ must be non-singular; see Theorem 1 in Jacobsen (2001a) and also (2.14) and (2.15) below.

The same theorem also gives the lower bounds for v_{-1} and v_0. In case (i), the leading coefficient matrix v_{-1} is present with lower bound

$$\big(\mathrm{E}^{\mu}_\theta \dot{b}_\theta^{\mathrm{T}}(X_0) C^{-1}(X_0) \dot{b}_\theta(X_0)\big)^{-1}, \tag{2.14}$$

while in case (ii) under small Δ-optimality $v_{-1} = 0$ and the lower bound for v_0 is

$$2\big(\mathrm{E}_\theta^\mu \dot{C}_\theta^{\mathrm{T}}(X_0)\big(C_\theta^{\otimes 2}(X_0)\big)^{-1}\dot{C}_\theta(X_0)\big)^{-1}, \tag{2.15}$$

with the lower bounds for case (iii) a suitable mixture of those for cases (i) and (ii) (Jacobsen (2001a, Theorem 1 (iii)).

Thus, for $\Delta \to 0$, in case (i) $\mathrm{var}_{\Delta,\theta}(g, \hat{\theta}) = O(\Delta^{-1})$, while in case (ii) and partly in case (iii) it is possible to obtain $\mathrm{var}_{\Delta,\theta}(g, \hat{\theta}) = O(1)$, i.e. for high-frequency data the parameters appearing in C_θ can be estimated much more precisely than those that appear only in the drift. An explanation for this is provided by observing that for $\Delta > 0$ small we are close to continuous-time observation of X, and that if $\mathbb{P}_{\theta,t}$ is the distribution of $(X_s)_{0\leqslant s\leqslant t}$ under P_θ^μ, in (i) it is typically the case that $\mathbb{P}_{\theta',t} \ll \mathbb{P}_{\theta,t}$ for $\theta' \neq \theta$ with the information about θ proportional to t, while in case (ii) it may well happen that $\mathbb{P}_{\theta',t} \perp \mathbb{P}_{\theta,t}$, i.e. the true value of θ can be read off from the observations $(X_s)_{0\leqslant s\leqslant t}$. Of course, in case (ii), if one is not using a small Δ-optimal estimating flow or at least one satisfying $\partial_y g_{0,\theta}(x, x) \equiv 0$, the leading $O(\Delta^{-1})$ term in (2.10) is present and the resulting estimator will have efficiency close to 0 against one that is small Δ-optimal. An example of this phenomenon is given in the simulation study by Jacobsen (2001b, Section 2.2) where, for a one-parameter model belonging to case (ii), the optimal martingale estimating flow using a base of dimension 1 yields an estimator that for small Δ is much worse than that derived from a small Δ-optimal flow with a base of dimension 2.

We shall now show that small Δ-optimality of martingale estimating functions is easy to achieve. The three cases refer to (i), (ii) and (iii) above.

Let $J := \{(i', j') : 1 \leqslant i' \leqslant j' \leqslant d\}$. Thus J has $|J| = d + d(d-1)/2$ elements and can be used as an index set for characterizing the elements of a symmetric $d \times d$ matrix. We write $R \in \mathbb{R}^{d^2\times J}$ for the reduction matrix with elements

$$R_{ij,i'j'} = \delta_{ii'}\delta_{jj'} \qquad (1 \leqslant i, j \leqslant d, (i', j') \in J).$$

Thus, if $M \in \mathbb{R}^{A\times d^2}$, then $MR \in \mathbb{R}^{A\times J}$ with $(MR)_{a,i'j'} = M_{a,i'j'}$, as is used frequently below.

As a counterpart to R, the expansion matrix $\tilde{R} \in \mathbb{R}^{J\times d^2}$ is defined by

$$\tilde{R}_{i'j',ij} = \begin{cases} \delta_{i'i}\delta_{j'j} & \text{if } i \leqslant j, \\ \delta_{i'j}\delta_{j'i} & \text{if } i > j. \end{cases}$$

Then

$$\tilde{R}R = I_{J\times J} \tag{2.16}$$

and, for any matrix $N \in \mathbb{R}^{S\times d^2}$, symmetric in the sense that $N_{s,ij} = N_{s,ji}$ for all s, i, j,

$$N(R\tilde{R}) = N. \tag{2.17}$$

Define

$$r_0(d) := d + |J| = \frac{d(d+3)}{2},$$

a number that plays a critical role below. We will usually write r_0 rather than $r_0(d)$.

Theorem 2.2. *Let $(f_\theta^1, \ldots, f_\theta^r)$ be a base for a martingale estimating function, of full affine rank r.*

(i) *Suppose that $r \geqslant d$, that for μ_θ-almost all x the matrix $\partial_x f_\theta(x) \in \mathbb{R}^{r\times d}$ is of full rank d, and that the p d-variate functions forming the columns of $\dot{b}_\theta$ are linearly independent. Then there exists $h_{t,\theta}(x) = h_\theta(x) \in \mathbb{R}^{r\times p}$, not depending on t, such that $g_{t,\theta}(x, y) := h_\theta^{\mathrm{T}}(x)(f_\theta(y) - \pi_{t,\theta} f(x))$ satisfies the small Δ-optimality condition* (2.11)*. In particular, for $r = d$, one may choose*

$$h_\theta^{\mathrm{T}}(x) = \dot{b}_\theta^{\mathrm{T}}(x) C^{-1}(x)(\partial_x f_\theta(x))^{-1}, \tag{2.18}$$

and this h_θ has linearly independent columns as required in Assumption A.

(*ii*) *Suppose that $r \geqslant r_0$, that for μ_θ-almost all x, the matrix*

$$\begin{pmatrix} \partial_x f_\theta(x) & \partial^2_{xx} f_\theta(x) \end{pmatrix} \in \mathbb{R}^{r\times(d+d^2)}$$

is of full rank r_0 and that the p d^2-variate functions forming the columns of $\dot{C}_\theta$ are linearly independent. Then there exists $h_{t,\theta} = h_\theta \in \mathbb{R}^{r\times p}$, not depending on t, such that $g_{t,\theta}(x, y) := h_\theta^{\mathrm{T}}(x)(f_\theta(y) - \pi_{t,\theta} f(x))$ satisfies the small Δ-optimality condition (2.12)*. In particular, for $r = r_0$, one may choose*

$$h_\theta^{\mathrm{T}}(x) = \begin{pmatrix} 0_{p\times d} & \dot{C}_\theta^{\mathrm{T}}(x)(C_\theta^{\otimes 2}(x))^{-1} R \end{pmatrix} (D_{1,2} f_\theta(x))^{-1}, \tag{2.19}$$

where

$$D_{1,2} f_\theta(x) = \begin{pmatrix} \partial_x f_\theta(x) & \partial^2_{xx} f_\theta(x) R \end{pmatrix}, \tag{2.20}$$

and this h_θ has linearly independent columns as required in Assumption A.

(*iii*) *Suppose that $r \geqslant r_0$, that for μ_θ-almost all x, the matrix*

$$\begin{pmatrix} \partial_x f_\theta(x) & \partial^2_{xx} f_\theta(x) \end{pmatrix} \in \mathbb{R}^{r\times(d+d^2)}$$

is of full rank r_0, that the $p - p'$ d-variate functions forming the columns of $\dot{b}_{2,\theta}$ are linearly independent, and that the p' d^2-variate functions forming the columns of $\dot{C}_{1,\theta}$ are linearly independent. Then there exists $h_{t,\theta} = h_\theta \in \mathbb{R}^{r\times p}$, not depending on t, such that $g_{t,\theta}(x, y) := h_\theta^{\mathrm{T}}(x)(f_\theta(y) - \pi_{t,\theta} f(x))$ satisfies the small Δ-optimality condition (2.13)*. In particular, for $r = r_0$ one may choose, with $D_{1,2} f_\theta$ as in* (2.20)*,*

$$h_\theta^{\mathrm{T}}(x) = \begin{pmatrix} 0_{p'\times d} & \dot{C}_{1,\theta}^{\mathrm{T}}(x)\big(C_\theta^{\otimes 2}(x)\big)^{-1} R \\ \dot{b}_{2,\theta}^{\mathrm{T}}(x) C_\theta^{-1}(x) & * \end{pmatrix} (D_{1,2} f_\theta(x))^{-1}. \tag{2.21}$$

with $*$ a $(p - p') \times J$ matrix depending arbitrarily on θ and x. If $*$ is chosen equal to 0, then this h_θ has linearly independent columns as required in Assumption A.

Proof. Since h_θ does not depend on t,

$$g_{0,\theta}(x, y) = h_\theta^{\mathrm{T}}(x)(f_\theta(y) - f_\theta(x)),$$

whence

$$\partial_y g_{0,\theta}(x, x) = h_\theta^{\mathrm{T}}(x)\partial_x f_\theta(x), \qquad \partial^2_{yy} g_{0,\theta}(x, x) = h_\theta^{\mathrm{T}}(x)\partial^2_{xx} f_\theta(x).$$

Thus, for each x, (2.11), (2.12) or (2.13) gives a system of linear equations for determining the elements of $h_\theta(x)$. The conditions of the theorem ensure that these equations have at least one solution, and exactly one in case (i) if $r = d$ and in case (ii) if $r = r_0$. (For case (ii), note that since $\partial^2_{x_i x_j} = \partial^2_{x_j x_i}$, the rank of $\partial^2_{xx} f_\theta$ is at most $|J|$. With h_θ given by (2.19), one now finds

$$\partial^2_{yy} g_{0,\theta}(x, x)R = \dot{C}^{\mathrm{T}}_\theta(x)\big(C^{\otimes 2}_\theta(x)\big)^{-1}R,$$

and, using (2.17), this implies the second identity in (2.12).)

The assertions about h_θ having linearly independent columns follow readily from the assumptions made on the columns of $\dot{b}_\theta$ (case (i)), $\dot{C}_\theta$ (case (ii)) and $\dot{b}_{2,\theta}$ and $\dot{C}_{1,\theta}$ (case (iii)). □

Theorem 2.2 only gives a solution for h_θ such that (2.11), (2.12) or (2.13) is satisfied. To check small Δ-optimality one further has to check the required integrability conditions (e.g. that all $h^{qk}_\theta \in L^4(\mu_\theta)$) as well as the conditions for the estimators to be well behaved asymptotically.

In Theorem 2.2 we have exhibited a concrete choice of small Δ-optimal estimating functions from a given base (f^q_θ). But it is then easy to define a host of others that are also small Δ-optimal, but may behave better for a given Δ, namely, flows $(g_{t,\theta})$ of the form

$$g^k_{t,\theta}(x, y) = \sum_{q=1}^{r} a^{qk}_\theta(t) h^{qk}_\theta(x)(f^q_\theta(y) - \pi_{t,\theta} f^q_\theta(x)), \tag{2.22}$$

with h^{T}_θ given by (2.18), (2.19) or (2.21) and each $a^{qk}_\theta(t)$ a non-random function of t, continuous with $a^{qk}_\theta(0) = 1$: for this flow, $g_{0,\theta}$ is the same as for the original flow, so small Δ-optimality still holds. However, there is no obvious optimal choice for the $a^{qk}_\theta(t)$, in particular the projection technique from the proof of Proposition 2.1 does not apply.

Remark 1. In Theorem 2.2, case (iii), the expression (2.21) for $h^{\mathrm{T}}_\theta(x)$ depends on the choice of $*$. This choice can be avoided by using a different procedure that is perhaps better suited to practical applications. By inspection of (2.13) it is seen that small Δ-optimality in case (iii) can be obtained as follows. First, fix $(\theta_{p'+1}, \ldots, \theta_p)$ and find a small Δ-optimal estimating flow for estimating $(\theta_1, \ldots, \theta_{p'})$ as in case (ii); see (2.12). Then, for $(\theta_1, \ldots, \theta_{p'})$ fixed, find a small Δ-optimal estimating flow for estimating $(\theta_{p'+1}, \ldots, \theta_p)$ as in case (i); see (2.11). Formally, this is done by combining the small Δ-optimal weights from Theorem 2.2, (i) and (ii), and for this purpose considering an r_0-dimensional base $f^{\mathrm{T}}_\theta = (f^{\circ\mathrm{T}}_\theta \quad \tilde{f}^{\mathrm{T}}_\theta)$ satisfying Assumption A whose two components have dimension d and $|J|$ respectively – typically one would use $\tilde{f}^{i'j'}_\theta = f^{\circ i'}_\theta f^{\circ j'}_\theta$ for $(i', j') \in J$ so that f°_θ determines the entire base, with of course $f^{\circ i}(x) = x_i$ the most natural example. The matrix h^{T}_θ of small Δ-optimal weights now takes the form

$$h^{\mathrm{T}}_\theta = \begin{pmatrix} & h^{\mathrm{T}}_{1,\theta} \\ h^{\mathrm{T}}_{2,\theta} & 0_{(p-p')\times J} \end{pmatrix}, \tag{2.23}$$

with $h_{1,\theta}^{\mathrm{T}} \in \mathbb{R}^{p' \times r_0}$ and $h_{2,\theta}^{\mathrm{T}} \in \mathbb{R}^{(p-p') \times d}$ given by

$$h_{1,\theta}^{\mathrm{T}}(x) = \Big(0_{p \times d} \quad \dot{C}_{1,\theta}^{\mathrm{T}}(x)\big(C_\theta^{\otimes 2}(x)\big)^{-1} R\Big)(D_{1,2} f_\theta(x))^{-1}, \tag{2.24}$$

$$h_{2,\theta}^{\mathrm{T}}(x) = \dot{b}_{2,\theta}^{\mathrm{T}}(x) C_\theta^{-1}(x)(\partial_x f_\theta^{\circ}(x))^{-1}$$

and, in block matrix notation,

$$D_{1,2} f_\theta(x) = \begin{pmatrix} \partial_x f_\theta^{\circ} & \partial_{xx}^2 f_\theta^{\circ} R \\ \partial_x \tilde{f}_\theta & \partial_{xx}^2 \tilde{f}_\theta R \end{pmatrix}.$$

With h_θ determined by (2.23) and (2.24), it is readily verified that the columns of h_θ are linearly independent as required by Assumption A, and also that h_θ is the same as was given by (2.21) when

$$* = \dot{b}_{2,\theta}^{\mathrm{T}} C_\theta^{-1} (\partial_x f_\theta^{\circ})^{-1} \partial_{xx}^2 f_\theta^{\circ} R. \tag{2.25}$$

In particular, if $f_\theta^{\circ i}(x) = x_i$ the two expressions agree for $* = 0$.

Remark 2. We mention one important special structure for the diffusion model that in cases (ii) and (iii) permits small Δ-optimal weights using a base of dimension $r < r_0$. Suppose that there is a decomposition $\{1, \ldots, d\} = \bigcup_{\nu=1}^{\kappa} I^{(\nu)}$ of the coordinates into disjoint non-empty sets $I^{(\nu)}$ with $\kappa \geqslant 2$, $|I^{(\nu)}| = d_\nu$ and $\sum d_\nu = d$ such that (assuming for convenience that $I^{(1)}$ comprises the first d_1 coordinates, $I^{(2)}$ the next d_2, etc.), for all θ and x, $C_\theta(x)$ can be written in block-diagonal form

$$C_\theta(x) = \begin{pmatrix} C_\theta^{(1)}(x) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & C_\theta^{(\kappa)}(x) \end{pmatrix}, \tag{2.26}$$

with each $C_\theta^{(\nu)}(x) \in \mathbb{R}^{d^{(\nu)} \times d^{(\nu)}}$.

The most important special case of this is of course when the components $X^{(\nu)} = (X_i)_{i \in I^{(\nu)}}$, $1 \leqslant \nu \leqslant \kappa$, are stochastically independent (assuming the initial values $X_0^{(\nu)}$ to be independent) so that in addition $C_\theta^{(\nu)}(x)$ and $(b_\theta^i(x))_{i \in I^{(\nu)}}$ depend on x through $x^{(\nu)} = (x_i)_{i \in I^{(\nu)}}$ only. However, independence is not required for the following discussion.

It is intuitively clear, at least under independence, that for case (ii) or (iii) models small Δ-optimal weights h_θ can be found from a base of dimension $r_1 = \sum_\nu r_0(d_\nu) < r_0$ (with a further reduction possible if one of the blocks $C_\theta^{(\nu)}$ does not depend on θ, or of course if a finer block structure than (2.26) is possible), and this we now verify.

Suppose the model belongs to case (ii) and that (2.26) holds. For each ν, determine a base $f_\theta^{(\nu)}$ of dimension $r_0(d_\nu)$ such that $f_\theta^{(\nu)}(x)$ as a function of x depends on $x^{(\nu)}$ only. If each $f_\theta^{(\nu)}$ satisfies Assumption A (as a function on $\mathbb{R}^{I^{(\nu)}}$), then so does the combined base f_θ given by $f_\theta^{\mathrm{T}} = (f_\theta^{(1)\mathrm{T}} \cdots f_\theta^{(\kappa)\mathrm{T}})$ of dimension r_1 (as a function on $\mathbb{R}^d$). Now define, by analogy with (2.19) (with $R^{(\nu)}$ the obvious reduction matrix for $I^{(\nu)}$),

$$h_\theta^{(\nu\mathrm{T})}(x) = \Big(0_{p \times d_\nu} \quad \dot{C}_\theta^{(\nu)\mathrm{T}}(x)\Big(C_\theta^{(\nu)\otimes 2}(x)\Big)^{-1} R^{(\nu)}\Big)\Big(D_{1,2} f_\theta^{(\nu)}\big(x^{(\nu)}\big)\Big)^{-1}, \tag{2.27}$$

where (cf. (2.20))

$$D_{1,2}f_\theta^{(\nu)}(x^{(\nu)}) = \Big(\partial_{x^{(\nu)}} f_\theta^{(\nu)}(x^{(\nu)}) \quad \partial^2_{x^{(\nu)}x^{(\nu)}} f_\theta^{(\nu)}(x^{(\nu)}) R^{(\nu)}\Big),$$

and combine the $h_\theta^{(\nu)\mathrm{T}}$ into

$$h_\theta^{(1)\mathrm{T}} = (h_\theta^{(1)\mathrm{T}} \cdots h_\theta^{(\kappa)\mathrm{T}}) \tag{2.28}$$

by juxtaposition. Then the weights h_θ are small Δ-optimal,

$$h_\theta^{\mathrm{T}}\partial_x f_\theta = 0, \qquad h_\theta^{\mathrm{T}}\partial^2_{xx} f_\theta = \dot{C}_\theta^{\mathrm{T}}\big(C_\theta^{\otimes 2}\big)^{-1},$$

as is easily verified using the special structure of f_θ and of course (2.26), which implies that

$$\big(\dot{C}_\theta^{\mathrm{T}}\big(C_\theta^{\otimes 2}\big)^{-1}\big)_{k,ij} = \Big(\dot{C}_\theta^{(\nu)\mathrm{T}}\Big(C_\theta^{(\nu)\otimes 2}\Big)^{-1}\Big)_{k,ij}$$

if $i, j \in I^{(\nu)}$, and taking the value 0 if i, j belong to two different $I^{(\nu)}$.

For case (iii) one may use the construction from Remark 1 and let each $f_\theta^{(\nu)}$ consist of $f_\theta^{\circ(\nu)}$ of dimension d_ν and $\tilde{f}_\theta^{(\nu)}$ of dimension $r_0(d_\nu) - d_\nu$. The small Δ-optimal weights h_θ then take the form (2.28) but with each $h_\theta^{(\nu)}$ now with the structure from (2.23),

$$h_\theta^{(\nu)\mathrm{T}} = \begin{pmatrix} & h_{1,\theta}^{(\nu)\mathrm{T}} \\ h_{2,\theta}^{(\nu)\mathrm{T}} & 0_{(p-p')\times J^{(\nu)}} \end{pmatrix},$$

where $h_{1,\theta}^{(\nu)\mathrm{T}}$ is given by (2.27) when replacing $\dot{C}_\theta^{(\nu)}$ by $\dot{C}_{1,\theta}^{(\nu)}$ and with $f_\theta^{(\nu)}$ comprising $f_\theta^{\circ(\nu)}$ and $\tilde{f}_\theta^{(\nu)}$ as just described, and where

$$h_{2,\theta}^{(\nu)\mathrm{T}}(x) = \dot{b}_{2,\theta}^{(\nu)\mathrm{T}}(x)\Big(C_\theta^{(\nu)}(x)\Big)^{-1}\big(\partial_{x^{(\nu)}} f^{\circ(\nu)}\big(x^{(\nu)}\big)\big)^{-1},$$

with $b_\theta^{(\nu)}$ of course collecting the coordinates $(b_\theta^i)_{i\in I^{(\nu)}}$ of b_θ and $\dot{b}_{2,\theta}^{(\nu)}$ signifying differentiation of $b_\theta^{(\nu)}$ with respect to the parameters $\theta_{p'+1}, \ldots, \theta_p$ not appearing in C_θ.

Remark 3. The results and remarks above have shown that there are lots of small Δ-optimal estimating functions. Yet another way of achieving small Δ-optimality may be by using generalized method of moments estimators (Hansen 1982). Start with $G_{n,\Delta}$ as in (2.1) with $g_{\Delta,\theta}$ of the form (2.6), but allow the dimension to be $s \geqslant p$. Then introduce a random weight matrix a_n not depending on θ, of dimension $s' \times s$ for some $s' \geqslant p$, and minimize the scalar $G_{n,\Delta}^{\mathrm{T}}(\theta)a_n^{\mathrm{T}}a_n G_{n,\Delta}(\theta)$ as a function of θ to obtain the estimator. Here we shall not investigate under what conditions this estimator is small Δ-optimal.

We return now to the discussion of the optimal martingale estimating function (2.8) determined by the base $(f_\theta^1, \ldots, f_\theta^r)$. Since, for any given $t = \Delta > 0$, $g_{\Delta,\theta}^{\mathrm{opt}}$ is better than a $g_{\Delta,\theta}$ where $h_{\Delta,\theta} = h_\theta$ is determined as in Theorem 2.2, the flow $\mathcal{G}^{\mathrm{opt}} = (g_{t,\theta}^{\mathrm{opt}})$ should be small Δ-optimal if $r \geqslant d$ in case (i) or $r \geqslant r_0$ in cases (ii) or (iii). What, however, does not follow from Theorem 2.2 is that $\mathcal{G}^{\mathrm{opt}}$ satisfies condition (2.11), (2.12) or (2.13). We shall now verify that this is the case (for $r = d$ and r_0, respectively). We also believe that in general (excepting for cases (ii) and (iii) the block structure from Remark 2 and possibly other special structures for the model) the lower bounds d and r_0 for r cannot be improved upon. In the statement of the result, the three cases are treated separately as usual.

Theorem 2.3 *For the optimal flow $\mathcal{G}^{\mathrm{opt}} = (g^{\mathrm{opt}}_{t,\theta})$ of martingale estimating functions with base $(f^1_\theta, \ldots, f^r_\theta)$ the following results hold:*

(i) *If $r = d$ and the matrix $\partial_x f_\theta(x)$ is non-singular for μ_θ-almost all x, then*

$$g^{\mathrm{opt}}_{0,\theta}(x, y) = \lim_{t\to 0} g^{\mathrm{opt}}_{t,\theta}(x, y)$$

and

$$\partial_y g^{\mathrm{opt}}_{0,\theta}(x, x) = \dot{b}^{\mathrm{T}}_\theta(x) C^{-1}(x).$$

(*ii*) *If $r = r_0$ and the matrix*

$$(\partial_x f_\theta(x) \quad \partial^2_{xx} f_\theta(x)) \in \mathbb{R}^{r\times(d+d^2)}$$

is of full rank r_0 for μ_θ-almost all x, then

$$g^{\mathrm{opt}}_{0,\theta}(x, y) = \lim_{t\to 0} t g^{\mathrm{opt}}_{t,\theta}(x, y)$$

and

$$\partial_y g^{\mathrm{opt}}_{0,\theta}(x, x) = 0, \qquad \partial^2_{yy} g^{\mathrm{opt}}_{0,\theta}(x, x) = \dot{C}^{\mathrm{T}}_\theta(x)\big(C^{\otimes 2}_\theta(x)\big)^{-1}.$$

(*iii*) *If $r = r_0$ and the matrix*

$$\big(\partial_x f_\theta(x) \quad \partial^2_{xx} f_\theta(x)\big) \in \mathbb{R}^{r\times(d+d^2)}$$

is of full rank r_0 for μ_θ-almost all x, then

$$g^{\mathrm{opt}}_{0,\theta}(x, y) = \lim_{t\to 0} \begin{pmatrix} t g^{\mathrm{opt}}_{1,t,\theta} \\ g^{\mathrm{opt}}_{2,t,\theta} \end{pmatrix}$$

and

$$\partial_y g^{\mathrm{opt}}_{0,\theta}(x, x) = \begin{pmatrix} 0_{p'\times d} \\ \dot{b}^{\mathrm{T}}_{2,\theta}(x) C^{-1}_\theta(x) \end{pmatrix}, \qquad \partial^2_{yy} g^{\mathrm{opt}}_{1,0,\theta}(x, x) = \dot{C}^{\mathrm{T}}_{1,\theta}(x)\big(C^{\otimes 2}_\theta(x)\big)^{-1}. \tag{2.29}$$

Proof. The main difficulty lies in finding $g^{\mathrm{opt}}_{0,\theta}$ from (2.8) and (2.7). The reader is initially reminded that if $\rho(x)$ is twice differentiable in x with $\rho \in L^1(\mu_\theta)$, then, provided $A_\theta\rho \in L^1(\mu_\theta)$ and $(\partial_x\rho) C_\theta (\partial_x\rho)^{\mathrm{T}} \in L^1(\mu_\theta)$, the expansion

$$\pi_{t,\theta}\rho(x) = \rho(x) + t A_\theta \rho(x) + o(t, x) \tag{2.30}$$

holds with $o(t, x)/t \to 0$ as $t \to 0$ for each x. Such expansions are used repeatedly below, not only for some function ρ, but later in the proof also when ρ is replaced by $A_\theta\rho$, and it should be kept in mind that this presupposes that, for example, the sufficient (but far from necessary) conditions given above for (2.30) to hold are satisfied.

Note at this point that for sufficiently nice functions ϕ, ψ, since

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

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

(with $o(t)/t = o(t, x)/t \to 0$ for each x) and

$$A_\theta(\phi\psi) = (A_\theta\phi)\psi + \phi(A_\theta\psi) + (\partial_x\phi)C_\theta(\partial_x\psi)^{\mathrm{T}}, \tag{2.31}$$

it follows that

$$\pi_{t,\theta}(\phi\psi) - (\pi_{t,\theta}\phi)(\pi_{t,\theta}\psi) = t(\partial_x\phi)C_\theta(\partial_x\psi)^{\mathrm{T}} + o(t). \tag{2.32}$$

Now use (2.32) with $\phi = f^q$, $\psi = f^{q'}$, $1 \leqslant q, q' \leqslant r$ (with r arbitrary at the moment), to obtain

$$\pi_{t,\theta}(f_\theta f_\theta^{\mathrm{T}}) - (\pi_{t,\theta}f_\theta)(\pi_{t,\theta}f_\theta)^{\mathrm{T}} = t(\partial_x f_\theta)C_\theta(\partial_x f_\theta)^{\mathrm{T}} + o(t), \tag{2.33}$$

and it is seen that the main term on the right evaluated at x is a non-singular $r \times r$ matrix only if $r \leqslant d$ and $\partial_x f(x) \in \mathbb{R}^{r\times d}$ is of full rank r.

To find $g^{\mathrm{opt}}_{0,\theta}$ we also need to approximate the factor $\partial_\theta\pi_{t,\theta}f_\theta - \pi_{t,\theta}\dot{f}$ from (2.7). Assuming that $\partial_\theta o(t) = o(t)$ and that $\dot{f}_\theta$ is smooth enough,

$$\begin{aligned}\partial_\theta\pi_{t,\theta}f_\theta &= \partial_\theta(f_\theta + tA_\theta f_\theta + o(t))\\ &= \dot{f}_\theta + t\big(A_\theta\dot{f}_\theta + (\partial_x f_\theta)\dot{b}_\theta + \tfrac{1}{2}\big(\partial^2_{xx}f_\theta\big)\dot{C}_\theta\big) + o(t),\\ \pi_{t,\theta}\dot{f}_\theta &= \dot{f}_\theta + tA_\theta\dot{f}_\theta + o(t)\end{aligned}$$

and thus

$$\partial_\theta\pi_{t,\theta}f_\theta - \pi_{t,\theta}\dot{f}_\theta = t\big((\partial_x f_\theta)\dot{b}_\theta + \tfrac{1}{2}\big(\partial^2_{xx}f_\theta\big)\dot{C}_\theta\big) + o(t). \tag{2.34}$$

Case (i). Since $\dot{C}_\theta \equiv 0$, (2.34) reduces to

$$\partial_\theta\pi_{t,\theta}f_\theta - \pi_{t,\theta}\dot{f}_\theta = t(\partial_x f_\theta)\dot{b}_\theta + o(t),$$

and therefore, using (2.33), it follows that if $r \leqslant d$, then

$$\begin{aligned}g^{\mathrm{opt}}_{0,\theta}(x, y) &= \lim_{t\to 0} g^{\mathrm{opt}}_{t,\theta}(x, y)\\ &= \dot{b}_\theta^{\mathrm{T}}(x)(\partial_x f_\theta)^{\mathrm{T}}(x)\big[\partial_x f_\theta(x)C(x)(\partial_x f_\theta)^{\mathrm{T}}(x)\big]^{-1}(f_\theta(y) - f_\theta(x))\end{aligned}$$

so that

$$\partial_y g^{\mathrm{opt}}_{0,\theta}(x, x) = \dot{b}_\theta^{\mathrm{T}}(x)(\partial_x f_\theta)^{\mathrm{T}}(x)\big[\partial_x f_\theta(x)C(x)(\partial_x f_\theta)^{\mathrm{T}}(x)\big]^{-1}\partial_x f_\theta(x). \tag{2.35}$$

For $r = d$ this reduces to $\dot{b}_\theta^{\mathrm{T}}(x)C^{-1}(x)$ as required. (For $r < d$ the $d \times d$ matrix appearing as a factor to the right of $\dot{b}_\theta^{\mathrm{T}}(x)$ has rank r, hence can never equal the non-singular $d \times d$ matrix $C^{-1}(x)$. However, in some special cases it may still be possible to obtain $\partial_y g^{\mathrm{opt}}_{0,\theta}(x, x) = \dot{b}_\theta^{\mathrm{T}}(x)C^{-1}(x)$).

Case (ii). Assume that $r > d$. Then the main term on the right of (2.33) becomes singular and it is therefore necessary to expand further. But from the basic expansion

$$\pi_{t,\theta}\varphi = \varphi + tA_\theta\varphi + \tfrac{1}{2}t^2A_\theta^2\varphi + o(t^2),$$

using (2.31) repeatedly, it eventually follows that

$$\pi_{t,\theta}(f_\theta f_\theta^{\mathrm{T}}) - (\pi_{t,\theta}f_\theta)(\pi_{t,\theta}f_\theta)^{\mathrm{T}} = t(\partial_x f_\theta)C_\theta(\partial_x f_\theta)^{\mathrm{T}} + \tfrac{1}{2}t^2Q + o(t^2), \tag{2.36}$$

with Q of the form

$$Q = (\partial_{xx}^2 f_\theta)C_\theta^{\otimes 2}(\partial_{xx}^2 f_\theta)^{\mathrm{T}} + (\partial_x f_\theta)S + S^{\mathrm{T}}(\partial_x f_\theta)^{\mathrm{T}} \tag{2.37}$$

for some $S(x) \in \mathbb{R}^{d\times r}$. By Lemma A.1 in the Appendix, therefore (with $A = (\partial_x f_\theta)C_\theta(\partial_x f_\theta)^{\mathrm{T}}$, $B = \frac{1}{2}Q$),

$$\lim_{t\to 0} t^2\left[\pi_{t,\theta}(f_\theta f_\theta^{\mathrm{T}}) - (\pi_{t,\theta}f_\theta)(\pi_{t,\theta}f_\theta)^{\mathrm{T}}\right]^{-1} = \mathcal{O}_2^{\mathrm{T}}\left(\mathcal{O}_2\tfrac{1}{2}Q\mathcal{O}_2^{\mathrm{T}}\right)^{-1}\mathcal{O}_2, \tag{2.38}$$

where $\mathcal{O}^{\mathrm{T}}(x) = (\mathcal{O}_1^{\mathrm{T}}(\mathrm{x}) \quad \mathcal{O}_2^{\mathrm{T}}(x)) \in \mathbb{R}^{r\times r}$ is orthogonal for each x, $\mathcal{O}_1(x)$ comprising the first d and $\mathcal{O}_2(x)$ the last $r-d$ rows of $\mathcal{O}(x)$, and satisfies

$$\mathcal{O}(x)(\partial_x f_\theta(x))C_\theta(x)(\partial_x f_\theta)^{\mathrm{T}}(x)\mathcal{O}^{\mathrm{T}}(x) = \mathrm{diag}(\lambda_1(x), \ldots, \lambda_d(x), 0, \ldots, 0) \tag{2.39}$$

with $\lambda_1(x), \ldots, \lambda_d(x) > 0$ the non-zero eigenvalues for $(\partial_x f_\theta)C_\theta(\partial_x f_\theta)^{\mathrm{T}}$ evaluated at x.

But from (2.39) it follows that

$$\mathcal{O}_2(x)(\partial_x f_\theta(x))C_\theta(x)(\partial_x f_\theta)^{\mathrm{T}}(x)\mathcal{O}_2^{\mathrm{T}}(x) = 0$$

or, since $C_\theta(x) \succ 0$, that

$$\mathcal{O}_2\partial_x f_\theta = 0. \tag{2.40}$$

Combining (2.38) with (2.34) and using (2.40), it follows that

$$\begin{aligned} g_{0,\theta}^{\mathrm{opt}}(x, y) &= \lim_{t\to 0} t g_{t,\theta}^{\mathrm{opt}}(x, y) \\ &= \dot{C}_\theta^{\mathrm{T}}(\partial_{xx}^2 f_\theta)^{\mathrm{T}}\mathcal{O}_2^{\mathrm{T}}[\mathcal{O}_2(\partial_{xx}^2 f_\theta)C_\theta^{\otimes 2}(\partial_{xx}^2 f_\theta)^{\mathrm{T}}\mathcal{O}_2^{\mathrm{T}}]^{-1}\mathcal{O}_2(f_\theta(y) - f_\theta(x)), \end{aligned}$$

with all factors to the left of $f_\theta(y)$ evaluated at x. Using (2.40), it is clear that $\partial_y g_{0,\theta}^{\mathrm{opt}}(x, x) = 0$ always, and hence, to obtain (2.12), it remains to check whether (omitting the argument x with $\partial_{xx}^2 g_{0,\theta}^{\mathrm{opt}}$ short for $\partial_{yy}^2 g_{0,\theta}^{\mathrm{opt}}(x, x)$)

$$\partial_{xx}^2 g_{0,\theta}^{\mathrm{opt}} = \dot{C}_\theta^{\mathrm{T}}(\partial_{xx}^2 f_\theta)^{\mathrm{T}}\mathcal{O}_2^{\mathrm{T}}\left[\mathcal{O}_2(\partial_{xx}^2 f_\theta)C_\theta^{\otimes 2}(\partial_{xx}^2 f_\theta)^{\mathrm{T}}\mathcal{O}_2^{\mathrm{T}}\right]^{-1}\mathcal{O}_2\partial_{xx}^2 f_\theta = \dot{C}_\theta^{\mathrm{T}}(C_\theta^{\otimes 2})^{-1}. \tag{2.41}$$

To achieve this we now assume that $r = r_0$, so that $r - d = |J|$, and use the assumption from the theorem that $\partial_{xx}^2 f_\theta(x)$ has full rank $|J|$ for all x. Then $\Gamma := (\partial_{xx}^2 f_\theta)R$ also has rank $|J|$ and $\mathcal{O}_2\Gamma \in \mathbb{R}^{J\times J}$ is non-singular, and, using $\partial_{xx}^2 f_\theta = \partial_{xx}^2 f_\theta(R\tilde{R})$ (cf. (2.17)), (2.41) therefore gives

$$\begin{aligned} (\partial_{xx}^2 g_{0,\theta}^{\mathrm{opt}})R &= \dot{C}_\theta^{\mathrm{T}}\tilde{R}^{\mathrm{T}}\Gamma^{\mathrm{T}}\mathcal{O}_2^{\mathrm{T}}\left[\mathcal{O}_2\Gamma\tilde{R}C_\theta^{\otimes 2}\tilde{R}^{\mathrm{T}}\Gamma^{\mathrm{T}}\mathcal{O}_2^{\mathrm{T}}\right]^{-1}\mathcal{O}_2\Gamma \\ &= \dot{C}_\theta^{\mathrm{T}}\tilde{R}^{\mathrm{T}}\left(\tilde{R}C_\theta^{\otimes 2}\tilde{R}^{\mathrm{T}}\right)^{-1}. \end{aligned} \tag{2.42}$$

That $\partial^2_{xx} g^{\mathrm{opt}}_{0,\theta} = \dot{C}^{\mathrm{T}}_\theta (C^{\otimes 2}_\theta)^{-1}$ will follow from (cf. (2.17))

$$(\partial^2_{xx} g^{\mathrm{opt}}_{0,\theta}) R = \dot{C}^{\mathrm{T}}_\theta (C^{\otimes 2}_\theta)^{-1} R,$$

and that the right-hand side here indeed equals that of (2.42) is verified by multiplying by $\tilde{R} C^{\otimes 2}_\theta \tilde{R}^{\mathrm{T}}$ from the right, again appealing to (2.17).

Case (iii). Here we initially proceed as in case (ii), arriving at (cf. (2.36))

$$g^{\mathrm{opt}}_{t,\theta}(x, y) = \big(\dot{b}^{\mathrm{T}}_\theta (\partial_x f_\theta)^{\mathrm{T}} + \tfrac{1}{2}\dot{C}^{\mathrm{T}}_\theta \big(\partial^2_{xx} f_\theta\big)^{\mathrm{T}} + o(1)\big) \times \big((\partial_x f_\theta) C_\theta (\partial_x f_\theta)^{\mathrm{T}} + \tfrac{1}{2} t Q + o(t)\big)^{-1} (f_\theta(y) - f_\theta(x) + o(1)) \tag{2.43}$$

with Q as in (2.37).

Considering first the last $p - p'$ components of $g^{\mathrm{opt}}_{t,\theta}$, since by assumption $\dot{C}_{2,\theta} \equiv 0$, it follows from Lemma A.1 that

$$g^{\mathrm{opt}}_{2,t,\theta}(x, y) = \dot{b}^{\mathrm{T}}_{2,\theta} (\partial_x f_\theta)^{\mathrm{T}} \big(\tfrac{1}{t} \mathcal{O}_2^{\mathrm{T}} \big(\mathcal{O}_2 \tfrac{1}{2} Q \mathcal{O}_2^{\mathrm{T}}\big)^{-1} \mathcal{O}_2 + N\big)(f_\theta(y) - f_\theta(x)) + o(1),$$

which, because of (2.40), in the limit reduces to

$$g^{\mathrm{opt}}_{2,0,\theta}(x, y) = \lim_{t \to 0} g^{\mathrm{opt}}_{2,t,\theta}(x, y) = \dot{b}^{\mathrm{T}}_{2,\theta} (\partial_x f_\theta)^{\mathrm{T}} N (f_\theta(y) - f_\theta(x))$$

with N of the form

$$N = \mathcal{O}_1^{\mathrm{T}} (\mathcal{O}_1 (\partial_x f_\theta) C_\theta (\partial_x f_\theta)^{\mathrm{T}} \mathcal{O}_1^{\mathrm{T}})^{-1} \mathcal{O}_1 + \mathcal{O}_2^{\mathrm{T}} \tilde{S} + \tilde{S}^{\mathrm{T}} \mathcal{O}_2.$$

But then, again using (2.40) and since $\mathcal{O}_1(\partial_x f_\theta) \in \mathbb{R}^{d \times d}$ is non-singular,

$$\begin{aligned} \partial_y g^{\mathrm{opt}}_{2,0,\theta}(x, x) &= \dot{b}^{\mathrm{T}}_{2,\theta} (\partial_x f_\theta)^{\mathrm{T}} \mathcal{O}_1^{\mathrm{T}} (\mathcal{O}_1 (\partial_x f_\theta) C_\theta (\partial_x f_\theta)^{\mathrm{T}} \mathcal{O}_1^{\mathrm{T}})^{-1} \mathcal{O}_1 (\partial_x f_\theta) \\ &= \dot{b}^{\mathrm{T}}_{2,\theta} C_\theta^{-1}, \end{aligned}$$

as required in the first part of (2.29).

As for the first p' components of $g^{\mathrm{opt}}_{t,\theta}$, obtain from (2.43) that

$$t g^{\mathrm{opt}}_{1,t,\theta}(x, y) = (\dot{b}^{\mathrm{T}}_{1,\theta} (\partial_x f_\theta)^{\mathrm{T}} + \tfrac{1}{2} \dot{C}^{\mathrm{T}}_{1,\theta} (\partial^2_{xx} f_\theta)^{\mathrm{T}}) \mathcal{O}_2^{\mathrm{T}} (\mathcal{O}_2 \tfrac{1}{2} Q \mathcal{O}_2^{\mathrm{T}})^{-1} \mathcal{O}_2 (f_\theta(y) - f_\theta(x)) + o(1),$$

whence

$$\begin{aligned} g^{\mathrm{opt}}_{1,0,\theta}(x, y) &= \lim_{t \to 0} t g^{\mathrm{opt}}_{1,t,\theta}(x, y) \\ &= \tfrac{1}{2} \dot{C}^{\mathrm{T}}_{1,\theta} (\partial^2_{xx} f_\theta)^{\mathrm{T}} \mathcal{O}_2^{\mathrm{T}} \big(\mathcal{O}_2 \tfrac{1}{2} \big(\partial^2_{xx} f_\theta\big) C^{\otimes 2} \big(\partial^2_{xx} f_\theta\big)^{\mathrm{T}} \mathcal{O}_2^{\mathrm{T}}\big)^{-1} \mathcal{O}_2 (f_\theta(y) - f_\theta(x)), \end{aligned}$$

once again using (2.40). But then (2.40) also gives

$$\partial_y g^{\mathrm{opt}}_{1,0,\theta}(x, x) = 0,$$

and arguing exactly as in the last part of case (ii), one finally finds that

$$\partial^2_{yy} g^{\mathrm{opt}}_{1,0,\theta}(x, x) = \dot{C}^{\mathrm{T}}_{1,\theta} \big(C^{\otimes 2}_\theta\big)^{-1},$$

and we have completed the proof of (2.29). □

We have not shown that (2.11) (or (2.12) or (2.13)) is satisfied for the optimal martingale estimating function when $r > d$ (or $r > r_0$). For case (i) with $r > d$ one may copy the argument involving $g^{\mathrm{opt}}_{2,\theta}$ given in case (iii) above. For cases (ii) and (iii), if $r > r_0$ a further expansion of (2.36) together with a refinement of Lemma A.1 is required, since, for example, the columns of $A = (\partial_x f_\theta) C_\theta (\partial_x f_\theta)^{\mathrm{T}}$ and $B = \frac{1}{2}Q$ cannot span a subspace of dimension r. We believe, however, that (2.12) (case (ii)) or (2.13) (case (iii)) is still valid for the optimal martingale estimating function, even if $r > r_0$.

Remark 4. For $d = 1$, Bibby and Sørensen (1995) studied martingale estimating functions with the one-dimensional ($r = 1$) base $f(x) = x$, and, apart from deriving the optimal estimating function G^*_n (their (2.15)), also suggested the use of an approximately optimal $\tilde{G}_n$ (their (2.14)). In general, the weights for $\tilde{G}_n$ are arrived at by replacing the true transition probabilities as they appear in our (2.7) by the Gaussian approximations corresponding to the Euler scheme, i.e. the conditional distribution of X_t given $X_0 = x$ is approximated by the normal distribution $n_{t,\theta}(x, \cdot)$ with mean $x + tb_\theta(x)$ and variance $tC_\theta(x)$. It may be shown, at least for $d = 1$ and $r = r_0$, that the estimator resulting from $\tilde{G}_n$ is small Δ-optimal.

3. Examples

We shall illustrate the foregoing results with two examples.

3.1. A generalized Cox–Ingersoll–Ross process

Consider the one-dimensional ($d = 1$) equation

$$\mathrm{d}X_t = \left(aX_t^{2\gamma-1} + bX_t\right)\mathrm{d}t + \sigma X_t^{\gamma}\mathrm{d}B_t, \tag{3.1}$$

where a, $b \in \mathbb{R}$, $\gamma \neq 1$ and $\sigma > 0$. For $\gamma = \frac{1}{2}$ this is the stochastic differential equation for the Cox–Ingersoll–Ross (CIR) process (see (3.2) below). The generalization (3.1) is arrived at by considering all powers $\tilde{X}^\rho$ of a CIR process with $\rho \neq 0$; more precisely, if X solves (3.1), then the associated CIR process is $\tilde{X} = X^{2-2\gamma}$ solving

$$\mathrm{d}\tilde{X}_t = \left(\tilde{a} + \tilde{b}\tilde{X}_t\right)\mathrm{d}t + \tilde{\sigma}\sqrt{\tilde{X}_t}\mathrm{d}B_t, \tag{3.2}$$

where

$$\tilde{b} = (2 - 2\gamma)b, \qquad \tilde{\sigma}^2 = (2 - 2\gamma)^2\sigma^2, \qquad \tilde{a} - \tfrac{1}{2}\tilde{\sigma}^2 = (2 - 2\gamma)\left(a - \tfrac{1}{2}\sigma^2\right) \tag{3.3}$$

(which also explains why $\gamma = 1$ is not allowed in (3.1)).

Because of the connection to the CIR process, the model described by (3.1) is much simpler to handle than the more standard Chan–Karolyi–Longstaff–Sanders model,

$$\mathrm{d}X_t = (a + bX_t)\mathrm{d}t + \sigma X_t^{\gamma}\mathrm{d}B_t;$$

in particular, for (3.1) it is easy to find martingale estimating functions of the type considered in the preceding sections.

In (3.1) the parameter space has dimension $p = 4$. We shall want X to be strictly positive and ergodic, which happens if and only if the associated CIR process $\tilde{X}$ is strictly positive and ergodic, i.e. $\tilde{b} < 0$ and $2\tilde{a} \geqslant \tilde{\sigma}^2$, or equivalently, either $\gamma < 1,\ b < 0,\ 2a \geqslant \sigma^2$ or $\gamma > 1,\ b > 0,\ 2a \leqslant \sigma^2$. As our *open* parameter set we shall therefore use

$$\Theta = \{(a, b, \gamma, \sigma^2) : \sigma^2 > 0 \text{ and either } \gamma < 1,\ b < 0,\ 2a > \sigma^2 \text{ or } \gamma > 1,\ b > 0,\ 2a < \sigma^2\}.$$

Note that if $\theta = (a, b, \gamma, \sigma^2) \in \Theta$ and $\rho \neq 0$, then X^ρ solves (3.1) with parameters $\theta^* = (a^*, b^*, \gamma^*, \sigma^{*2})$ given by

$$b^* = \rho b, \qquad \sigma^{*2} = \rho^2\sigma^2, \qquad 2 - 2\gamma^* = \frac{1}{\rho}(2 - 2\gamma), \qquad a^* - \tfrac{1}{2}\sigma^{*2} = \rho\left(a - \tfrac{1}{2}\sigma^2\right).$$

In particular, taking $\rho < 0$ corresponds to a switch from $\gamma < 1$ to $\gamma^* > 1$ (or from $\gamma > 1$ to $\gamma^* < 1$).

Since the invariant distribution for $\tilde{X}$ is a gamma distribution, the invariant distribution for X is that of a gamma-distributed random variable raised to the power $(2 - 2\gamma)^{-1}$. The density is

$$\mu_\theta(x) = \frac{|2 - 2\gamma|}{\Gamma(2\tilde{a}/\tilde{\sigma}^2)((2\gamma - 2)\sigma^2/2b)^{2\tilde{a}/\tilde{\sigma}^2}} x^{2a/\sigma^2 - 2\gamma} \exp\left(-\frac{2b}{(2\gamma - 2)\sigma^2} x^{2-2\gamma}\right) \tag{3.4}$$

for $x > 0$, where (cf. (3.3))

$$\frac{2\tilde{a}}{\tilde{\sigma}^2} = \frac{2a}{(2 - 2\gamma)\sigma^2} + \frac{1 - 2\gamma}{2 - 2\gamma}.$$

(For $\gamma = \frac{1}{2}$ the familiar invariant gamma density for the CIR process is obtained.)

Because a gamma distribution has finite moments of all orders $m \in \mathbb{N}$, we have $\mathrm{E}_\theta^\mu X_0^{(2-2\gamma)m} < \infty$ for all $m \in \mathbb{N}$, and, since

$$\mathrm{E}_\theta^\mu X_t^{(2-2\gamma)m} = \int_0^\infty \mathrm{d}x \mu_\theta(x) \pi_{t,\theta} x^{(2-2\gamma)m}$$

(where $\pi_{t,\theta} x^\beta$ is short for $\pi_{t,\theta} f(x)$ for $f(y) = y^\beta$), also

$$\pi_{t,\theta} x^{(2-2\gamma)m} < \infty$$

for all $t > 0$, $m \in \mathbb{N}$ and (Lebesgue almost all) $x > 0$.

The conditional moments for a CIR process are known and in any case easy to find using polynomial martingales: for $m \in \mathbb{N}$, let $\tilde{\xi}_m$ be the mth moment in the invariant distribution for $\tilde{X}$,

$$\tilde{\xi}_m = \left(-\frac{\tilde{\sigma}^2}{2\tilde{b}}\right)^m \frac{\Gamma(2\tilde{a}/\tilde{\sigma}^2 + m)}{\Gamma(2\tilde{a}/\tilde{\sigma}^2)},$$

and verify, for instance using induction on m and Itô's formula, that $M^{(m)}$ is a mean-zero martingale (see the note below) under each P_θ^x, where

$$M_t^{(m)} = \mathrm{e}^{-\tilde{b}mt} \sum_{i=1}^{m} \beta_i^{(m)} \left(X_t^{(2-2\gamma)i} - \tilde{\xi}_i \right) \tag{3.5}$$

with

$$\beta_i^{(m)} = \frac{1}{\tilde{\xi}_i} (-1)^{i-1} \binom{m}{i}.$$

(Equivalently, for each m, the polynomial $\sum_i \beta_i^{(m)} (x^i - \tilde{\xi}_i)$ of degree m is an eigenfunction for the generator for the CIR process (3.2) corresponding to the eigenvalue $\tilde{b}m$; see Kessler and Sørensen (1999) for estimating functions built from eigenfunctions, and their Example 2.1 for the CIR process).

Note that, because all conditional moments for the ergodic CIR process are finite, one verifies directly that the local martingale $M^{(m)}$ satisfies $\mathrm{E}_\theta^x [M^{(m)}]_t < \infty$ for all x and t, in particular $M^{(m)}$ is therefore a true martingale under P_θ^x (L^2-bounded on $[0, t]$ for all t).

Turning now to the problem of estimating θ from discrete observations of X, it is clear that (3.1) belongs to case (iii) with $p = 4$, $p' = 2$, so we shall apply Theorems 2.2 and 2.3 for that case with $r = r_0(1) = 2$. In view of the above, a natural candidate for the base (f^1, f^2) is

$$f^1(x) = x^{2-2\gamma}, \qquad f^2(x) = x^{4-4\gamma}, \tag{3.6}$$

which trivially satisfies Assumption A. Note that f^1, f^2 both depend on θ; cf. the comment immediately preceding Proposition 2.1.

In order to find an example of small Δ-optimal weights we use the recipe from Remark 1 (corresponding to the special choice (2.25) of $*$ in Theorem 2.2, case (iii)) and, listing the parameters in the order γ, σ^2, a, b, we find that

$$\dot{b}_{2,\theta}^{\mathrm{T}}(x) = \begin{pmatrix} x^{2\gamma-1} \\ x \end{pmatrix}, \qquad \dot{C}_{1,\theta}^{\mathrm{T}}(x) = \begin{pmatrix} 2\sigma^2 x^{2\gamma} \log x \\ x^{2\gamma} \end{pmatrix}$$

and eventually arrive at the estimating function

$$g_{t,\theta}(x, y) = \begin{pmatrix} -2\log x & x^{2\gamma-2} \log x \\ -2 & x^{2\gamma-2} \\ x^{2\gamma-2} & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y^{2-2\gamma} - \pi_{t,\theta} x^{2-2\gamma} \\ y^{4-4\gamma} - \pi_{t,\theta} x^{4-4\gamma} \end{pmatrix}, \tag{3.7}$$

which requires the use of (3.5) for $m = 1, 2$ in order to find the conditional expectations.

That $g_{t,\theta}$ given by (3.7) indeed satisfies the conditions (2.13) for small Δ-optimality is most easily verified directly. Note that the linear independence asserted in Theorem 2.2 between the columns in h_θ, i.e. the functions comprising the rows in the 4×2 matrix in (3.7), holds precisely because $\gamma \neq 1$.

For the flow $(g_{t,\theta})$ given by (3.7) one still needs to check the integrability assumptions from Jacobsen (2001a, Theorem 1) and the conditions on estimating flows made prior to

that theorem. In the case at hand the conditions in particular amount to requiring that $\mathrm{E}_\theta^\mu |g_{t,\theta}^k(X_0, X_t)|^K < \infty$ for all components k and moderate values of $K \in \mathbb{N}$. The problem is the appearance of the power $x^{2\gamma-2}$ in the expression for $g_{t,\theta}$, which translates into negative powers $\tilde{X}_0^{-1}$ and $\tilde{X}_0^{-2}$ of the CIR process $\tilde{X}$, and of course, for example, $\mathrm{E}_\theta^\mu \tilde{X}_0^{-K} = \mathrm{E}_\theta^\mu X_0^{(2\gamma-2)K} < \infty$ if and only if $2\tilde{a}/\tilde{\sigma}^2 > K$. Thus some care should be taken before applying (3.7): it must at least be assumed that $2\tilde{a}/\tilde{\sigma}^2$ is suitably large.

The estimation function (3.7) was used in a simulation study in Jacobsen (2001b, Section 2.1) with good results not only for small values of Δ.

To find the optimal martingale estimating function with base (f^1, f^2) given by (3.6), one needs (3.5) also for $m = 3, 4$ *and* conditional moments involving logarithms; see (2.7), in which the term $\pi_{t,\theta}\dot{f}_\theta$ appears. The latter moments are easy to find in terms of the conditional expectation $\mathrm{E}_\theta(\log \tilde{X}_t | \tilde{X}_0 = \tilde{x})$ for the CIR process starting at an arbitrary level $\tilde{x}$, but the explicit form for this is unpleasant to work with.

Whether one uses the small Δ-optimal flow (3.7) or the optimal flow, since $d = 1$ a slight improvement in efficiency may be gained by symmetrizing, using for example $\frac{1}{2}(g_{t,\theta}(x, y) + g_{t,\theta}(y, x))$ instead of (3.7); see Jacobsen (2001a, Proposition 4) and the discussion there about time reversal.

3.2. The finite-dimensional Gaussian diffusions

We consider now the d-dimensional diffusion

$$\mathrm{d}X_t = (A + BX_t)\mathrm{d}t + D\mathrm{d}W_t, \tag{3.8}$$

where the unknown parameters are $A \in \mathbb{R}^{d\times 1}$, $B \in \mathbb{R}^{d\times d}$ and $C := DD^{\mathrm{T}} \in \mathbb{R}^{d\times d}$, with the symmetric matrix C assumed strictly positive definite. (In this subsection the symbol B is used to denote the matrix of linear drift parameters, and the driving d-dimensional Brownian motion is denoted W instead of B.) Thus

$$p' = |J|, \qquad p = |J| + d + d^2.$$

The diffusion (3.8) has Gaussian transitions (for the expectation and second-order moments, see (3.9) and (3.10) below) and is ergodic if and only if $\mathrm{spec}(B) \subset \{\lambda \in \mathbb{C}: \mathrm{Re}(\lambda) < 0\}$.

For this model there is a genuine identification problem when considering equidistant observations $X_{i\Delta}$ for an arbitrary given $\Delta > 0$: it is possible to find $\theta \neq \theta'$ such that $\pi_{\Delta,\theta}(x, \cdot) = \pi_{\Delta,\theta'}(x, \cdot)$ for all $x \in \mathbb{R}^d$. For example, take $d = 2$ and define θ and θ' by $A_\theta = A_{\theta'} = 0_{2\times 1}$, $C_\theta = C_{\theta'} = I_2$ and $B_\theta = bI_2$,

$$B_{\theta'} = bI_2 + \begin{pmatrix} 0 & 2\pi k/\Delta \\ -2\pi k/\Delta & 0 \end{pmatrix}$$

for some $b < 0$ (to obtain ergodicity) and some $k \in \mathbb{Z}\backslash 0$. The θ'-process is a two-dimensional *rotating Ornstein–Uhlenbeck* process, while the θ-process starting, say, at a given $x \in \mathbb{R}^2$ is composed of two independent one-dimensional Ornstein–Uhlenbeck processes. Because, for all ϑ,

$$\exp\begin{pmatrix} 0 & \vartheta \\ -\vartheta & 0 \end{pmatrix} = \begin{pmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{pmatrix}$$

is orthogonal, it follows from (3.9) and (3.10) below that $\pi_{\Delta,\theta}(x,\cdot) = \pi_{\Delta,\theta'}(x,\cdot)$ is the Gaussian distribution with mean vector $\mathrm{e}^{b\Delta}x$ and covariance matrix $(-2b)^{-1}(1-\mathrm{e}^{2b\Delta})I_2$. Thus, for this example, the identification equation (2.3) can never hold for any $t > 0$.

As our base (of dimension r_0) for the martingale estimating functions we shall use (f^q) for $q \in \{1,\ldots,d\} \cup J$, where

$$f^i(x) = x_i \quad (1 \leqslant i \leqslant d), \qquad f^{i'j'}(x) = x_{i'}x_{j'} \quad ((i',j') \in J),$$

writing $x = (x_1,\ldots,x_d)$ for a generic point in $\mathbb{R}^d$. Clearly (f^q) satisfies the conditions from Assumption A and also, as a little work shows, the conditions on $\partial_x f$, $\partial^2_{xx} f$ from Theorem 2.2. To proceed we need the conditional moments $\pi_{t,\theta}f^q$, conveniently collected in the vector $\pi_{t,\theta}x = (\pi_{t,\theta}x_i)_{1\leqslant i\leqslant d}$ and the matrix $\pi_{t,\theta}xx^{\mathrm{T}}$ and known to be given by the expressions

$$\pi_{t,\theta}x = (\mathrm{e}^{tB} - I_d)B^{-1}A + \mathrm{e}^{tB}x, \tag{3.9}$$

$$\pi_{t,\theta}xx^{\mathrm{T}} = (\pi_{t,\theta}x)(\pi_{t,\theta}x)^{\mathrm{T}} + \int_0^t \mathrm{e}^{sB}C\mathrm{e}^{sB^{\mathrm{T}}}\mathrm{d}s. \tag{3.10}$$

(As in the previous example, notation like $\pi_{t,\theta}xx^{\mathrm{T}}$ is short for $\pi_{t,\theta}f(x)$, where $f(y) = yy^{\mathrm{T}}$.)

Invoking Theorem 2.2 with $*$ in (2.21) equal to 0 (which here gives the same result as the method described in Remark 1), one eventually arrives at the small Δ-optimal estimating function $g_{t,\theta}$, with $g_{1,t,\theta} = (g^{i'j'}_{t,\theta})_{(i',j')\in J}$ and $g_{2,t,\theta}$ split into the vector-valued component $g^A_{2,t,\theta} = (g^i_{t,\theta})_{1\leqslant i\leqslant d}$ and the matrix-valued component $g^B_{2,t,\theta} = (g^{ij}_{t,\theta})_{1\leqslant i,j\leqslant d}$, and $g_{1,t,\theta}$, $g^A_{2,t,\theta}$ and $g^B_{2,t,\theta}$ given by

$$\begin{aligned} g^{i'j'}_{1,t,\theta}(x,y) &= (C^{-1}[-x(y-\pi_{t,\theta}x)^{\mathrm{T}} - (y-\pi_{t,\theta}x)x^{\mathrm{T}} \\ &\qquad + yy^{\mathrm{T}} - \pi_{t,\theta}(xx^{\mathrm{T}})]C^{-1})_{i'j'}, \\ g^A_{2,t,\theta}(x,y) &= C^{-1}(y-\pi_{t,\theta}x), \\ g^B_{2,t,\theta}(x,y) &= C^{-1}(y-\pi_{t,\theta}x)x^{\mathrm{T}}. \end{aligned}$$

For the calculations one uses the fact that

$$\dot{C}^{\mathrm{T}}_{1,\theta}(x) \in \mathbb{R}^{J\times d^2}, \qquad \left(\dot{C}^{\mathrm{T}}_{1,\theta}(x)\right)_{i'j',ij} = \begin{cases} \delta_{i'i}\delta_{j'j} & \text{if } i \leqslant j, \\ \delta_{i'j}\delta_{j'i} & \text{if } i > j, \end{cases}$$

$$\left(\dot{b}^A_{2,\theta}\right)^{\mathrm{T}}(x) = I_d \in \mathbb{R}^{d\times d},$$

$$\left(\dot{b}^B_{2,\theta}\right)^{\mathrm{T}}(x) = I_d \otimes x \in \mathbb{R}^{d^2\times d}$$

with $b^A_{2,\theta}(x) := A$ differentiated with respect to A only and $b^B_{2,\theta}(x) := Bx$ differentiated with respect to B only. Also note that

$$(D_{1,2}f(x))^{-1} = (\partial_x f(x) \quad \partial^2_{xx} f(x)R)^{-1} = \begin{pmatrix} I_d & 0_{d\times J} \\ P(x) & D_0 \end{pmatrix},$$

where $D_0 = \mathrm{diag}(d_{i'j'}) \in \mathbb{R}^{J\times J}$ with

$$d_{i'j'} = \begin{cases} 1 & \text{if } i' < j', \\ \frac{1}{2} & \text{if } i' = j', \end{cases}$$

and $P(x) \in \mathbb{R}^{J\times d}$ with

$$P_{i'j',j}(x) = -d_{i'j'}(\delta_{i'j}x_{j'} + \delta_{j'j}x_{i'}).$$

As in the previous example, the simplest way to verify the small Δ-optimality is to verify directly from these expressions that the conditions (2.13) are satisfied.

The resulting estimating equations are not affected by multiplication from the left and/or right by C, and it is now an easy task to write down the estimators of the parameter functions

$$\mathcal{A} := (\mathrm{e}^{\Delta B} - I_d)B^{-1}A, \qquad \mathrm{e}^{\Delta B}, \qquad \mathcal{C} := \int_0^\Delta \mathrm{e}^{sB} C \mathrm{e}^{sB^{\mathrm{T}}} \mathrm{d}s$$

based on the observations $X_0, X_\Delta, \ldots, X_{n\Delta}$: defining

$$\overline{X}_* := \frac{1}{n}\sum_{i=1}^n X_{(i-1)\Delta}, \qquad \overline{X}^* := \frac{1}{n}\sum_{i=1}^n X_{i\Delta},$$

one sees using (3.9) and (3.10) that the estimating equations obtained from g_1, g_2^A, g_2^B are equivalent to the equations

$$\sum_{i=1}^n \left(X_{i\Delta} - \mathcal{A} - \mathrm{e}^{\Delta B}X_{(i-1)\Delta}\right) = 0, \tag{3.11}$$

$$\sum_{i=1}^n \left(X_{i\Delta} - \mathcal{A} - \mathrm{e}^{\Delta B}X_{(i-1)\Delta}\right)X^{\mathrm{T}}_{(i-1)\Delta} = 0 \tag{3.12}$$

$$\sum_{i=1}^n \left(X_{i\Delta}X^{\mathrm{T}}_{i\Delta} - \left(\mathcal{A} + \mathrm{e}^{\Delta B}X_{(i-1)\Delta}\right)\left(\mathcal{A} + \mathrm{e}^{\Delta B}X_{(i-1)\Delta}\right)^{\mathrm{T}} - \mathcal{C}\right) = 0, \tag{3.13}$$

and hence

$$\hat{\mathcal{A}} = \overline{X}^* - \mathrm{e}^{\Delta \hat{B}} \overline{X}_*, \tag{3.14}$$

$$\mathrm{e}^{\Delta \hat{B}} = \left(\sum_{i=1}^{n} (X_{i\Delta} - \overline{X}^*) X_{(i-1)\Delta}^{\mathrm{T}} \right) \left(\sum_{i=1}^{n} \left(X_{(i-1)\Delta} - \overline{X}_* \right) \left(X_{(i-1)\Delta} - \overline{X}_* \right)^{\mathrm{T}} \right)^{-1} \tag{3.15}$$

$$\hat{\mathcal{C}} = \frac{1}{n} \sum_{i=1}^{n} (X_{i\Delta} X_{i\Delta}^{\mathrm{T}} - Z_i Z_i^{\mathrm{T}}) \tag{3.16}$$

where, in the last line,

$$Z_i := \hat{\mathcal{A}} + \mathrm{e}^{\Delta \hat{B}} X_{(i-1)\Delta}.$$

Note that (3.16) may be written

$$\hat{\mathcal{C}} = \frac{1}{n} \sum_{i=1}^{n} \left(X_{i\Delta} - \hat{\mathcal{A}} - \mathrm{e}^{\Delta \hat{B}} X_{(i-1)\Delta} \right) \left(X_{i\Delta} - \hat{\mathcal{A}} - \mathrm{e}^{\Delta \hat{B}} X_{(i-1)\Delta} \right)^{\mathrm{T}}, \tag{3.17}$$

as is seen using the fact that it follows from (3.11) and (3.12) that

$$\sum_{i=1}^{n} (X_{i\Delta} - \hat{\mathcal{A}} - \mathrm{e}^{\Delta \hat{B}} X_{(i-1)\Delta}) Z_i^{\mathrm{T}} = 0.$$

The likelihood function for observing $X_0, X_\Delta, \ldots, X_{n\Delta}$ conditionally on X_0 is

$$\prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |\mathcal{C}|} \exp\left(-\tfrac{1}{2} (X_{i\Delta} - \xi_i)^{\mathrm{T}} \mathcal{C}^{-1} (X_{i\Delta} - \xi_i) \right),$$

where

$$\xi_i = \pi_{\Delta,\theta}(X_{(i-1)\Delta}) = \mathcal{A} + \mathrm{e}^{\Delta B} X_{(i-1)\Delta}.$$

Maximizing this over $\mathcal{A}$, $\mathrm{e}^{\Delta B}$ and $\mathcal{C}$ varying *freely* in $\mathbb{R}^{d \times 1}$, $\mathbb{R}^{d \times d}$ and the space of symmetric positive definite $d \times d$ matrices yields the estimators $\hat{\mathcal{A}}$, $\mathrm{e}^{\Delta \hat{B}}$ and $\hat{\mathcal{C}}$ from (3.14), (3.15) and (3.16). For the model with $A = 0$, Kessler and Rahbek (2001) study the corresponding maximum likelihood estimator and also tackle the non-trivial problem of converting their versions of (3.15) and (3.16) into estimators for B and C: they provide conditions for (3.15) to be the exponential of a square matrix (which in our model must also satisfy the condition that all eigenvalues have negative real parts), and also provide conditions for this square matrix and the estimator for C to be uniquely determined.

Appendix

The following result was used in the proof of Theorem 2.3:

Lemma A.1. *Let A, $B \in \mathbb{R}^{m \times m}$ be symmetric and positive semidefinite matrices such that*

$1 \leqslant \operatorname{rank}(A) = m' < m$ *and such that the columns* (*or rows*) *of* A *and* B *jointly span all of* $\mathbb{R}^m$. *Further, let* $\mathcal{O}^{\mathrm{T}} = (\mathcal{O}_1^{\mathrm{T}} \quad \mathcal{O}_2^{\mathrm{T}})$ *be an orthogonal* $m \times m$ *matrix with* $\mathcal{O}_1$ *comprising the first* m' *and* $\mathcal{O}_2$ *the last* $m - m'$ *rows of* $\mathcal{O}$ *such that*

$$\mathcal{O}A\mathcal{O}^{\mathrm{T}} = \operatorname{diag}(\lambda_1, \ldots, \lambda_{m'}, 0, \ldots, 0),$$

$\lambda_1, \ldots, \lambda_{m'} > 0$ *denoting the non-zero eigenvalues for* A. *Then as* $t \to 0$,

$$(A + tB)^{-1} = \frac{1}{t}\mathcal{O}_2^{\mathrm{T}}(\mathcal{O}_2 B \mathcal{O}_2^{\mathrm{T}})^{-1}\mathcal{O}_2 + N + O(t),$$

where N is of the form

$$\mathcal{O}_1^{\mathrm{T}}(\mathcal{O}_1 A \mathcal{O}_1^{\mathrm{T}})^{-1}\mathcal{O}_1 + \mathcal{O}_2^{\mathrm{T}} S + S^{\mathrm{T}}\mathcal{O}_2$$

for some $(m - m') \times m$ matrix S.

Proof. Assume first that $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_{m'}, 0, \ldots, 0)$ (with all $\lambda_\ell > 0$) and write

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

with, for example, B_{22} the lower right $(m - m') \times (m - m')$ submatrix of B. Then

$$|A + tB| = t^{m-m'}\left(\prod_{\ell=1}^{m'} \lambda_\ell\right)|B_{22}| + O\big(t^{m-m'+1}\big),$$

as is seen by computing the determinant directly as the sum of signed products $\Pi_{\ell=1}^{m}(A + tB)_{\ell\sigma(\ell)}$ with σ an arbitrary permutation of $1, \ldots, m$. Also for the subdeterminants obtained by deleting the ℓth row and ℓ'th column,

$$|A + tB|_{\ell\ell'} = \begin{cases} O(t^{m-m'-1}) & \text{if } \ell, \ell' > m', \\ O(t^{m-m'}) & \text{otherwise.} \end{cases}$$

It follows from this that $(A + tB)^{-1}$ is of the form

$$\frac{1}{t}\begin{pmatrix} 0 & 0 \\ 0 & M \end{pmatrix} + N + O(t)$$

and it is then easy to see that, writing

$$D = \operatorname{diag}(\lambda_1, \ldots, \lambda_{m'}) \in \mathbb{R}^{m' \times m'},$$

one has

$$(A + tB)^{-1} = \frac{1}{t}\begin{pmatrix} 0 & 0 \\ 0 & B_{22}^{-1} \end{pmatrix} + \begin{pmatrix} D^{-1} & -D^{-1}B_{12}B_{22}^{-1} \\ -B_{22}^{-1}B_{21}D^{-1} & B_{22}^{-1}B_{21}D^{-1}B_{12}B_{22}^{-1} \end{pmatrix} + O(t). \tag{A.1}$$

For the general case, just use the fact that

$$(A + tB)^{-1} = \mathcal{O}^{\mathrm{T}}\big(\mathcal{O}A\mathcal{O}^{\mathrm{T}} + t\mathcal{O}B\mathcal{O}^{\mathrm{T}}\big)^{-1}\mathcal{O},$$

with $(\mathcal{O}A\mathcal{O}^{\mathrm{T}} + t\mathcal{O}B\mathcal{O}^{\mathrm{T}})^{-1}$ of the form (A.1) and $D = \mathcal{O}_1 A \mathcal{O}_1^{\mathrm{T}}$ □

Acknowledgements

This research was supported by MaPhySto – Centre for Mathematical Physics and Stochastics, funded by a grant from the Danish National Research Foundation, and by Dynstoch, part of the Human Potential Programme funded by the European Commission. Thanks are also due to the referees whose comments led to major improvements in the presentation.

References

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

Hansen, L.P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.

Heyde, C.C. (1988) Fixed sample and asymptotic optimality for classes of estimating functions. *Contemp. Math.*, **80**, 241–247.

Jacobsen, M. (2001a) Discretely observed diffusions: classes of estimating functions and small Δ-optimality. *Scand. J. Statist.*, **28**, 123–149.

Jacobsen, M. (2001b) Small Δ-optimal martingale estimating functions: a simulation study. Preprint 2, Department of Theoretical Statistics, University of Copenhagen.

Kessler, M. (2000) Simple and explicit estimating functions for a discretely observed diffusion process. *Scand. J. Statist.*, **27**, 65–82.

Kessler, M. and Rahbek, A. (2001) Identification and inference for cointegrated and ergodic Gaussian diffusions. Preprint 3, Department of Theoretical Statistics, University of Copenhagen.

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

Sørensen, M. (1999) On asymptotics of estimating functions. *Braz. J. Probab. Statist.*, **13**, 111–136.

Received May 2000 and revised May 2002