Annals of the Institute of Statistical Mathematics manuscript No.
(will be inserted by the editor)

Recursive Parameter Estimation: Asymptotic expansion

Teo Sharia

Received: date / Revised: date

Abstract This paper is concerned with the asymptotic behaviour of estimation procedures which are recursive in the sense that each successive estimator is obtained from the previous one by a simple adjustment. The results of the paper can be used to determine the form of the recursive procedure which is expected to have the same asymptotic properties as the corresponding non-recursive one defined as a solution of the corresponding estimating equation. Several examples are given to illustrate the theory, including an application to estimation of parameters in exponential families of Markov processes.

Keywords recursive estimation · estimating equations · stochastic approximation

1 Introduction

Let $X_1, \dots, X_n$ be random variables with a joint distribution depending on an unknown parameter θ. Then an M-estimator of θ is defined as a solution of the estimating equation

$$\sum_{i=1}^{n} \psi_i(v) = 0, \tag{1}$$

where $\psi_i(v) = \psi_i(X_1^i; v)$, $i = 1, 2, \dots, n$, are suitably chosen functions which may, in general, depend on the vector $X_1^i = (X_1, \dots, X_i)$ of all past and present observations. If $f_i(x, \theta) = f_i(x, \theta | X_1, \dots, X_{i-1})$ is the conditional probability density function or probability function of the observation X_i, given $X_1, \dots, X_{i-1}$, then one can obtain a MLE (maximum likelihood estimator) on choosing $\psi_i(v) = f_i'(X_i, v)/f_i(X_i, v)$. Besides MLEs, the class of M-estimators

T. Sharia
Department of Mathematics, Royal Holloway University of London,
Egham, Surrey TW20 0EX
E-mail: t.sharia@rhul.ac.uk

includes estimators with special properties such as robustness. Under certain regularity and ergodicity conditions, it can be proved that there exists a consistent sequence of solutions of (1) which has the property of local asymptotic linearity.

If ψ-functions are nonlinear, it is rather difficult to work with the corresponding estimating equations. In this paper, we consider estimation procedures which are recursive in the sense that each successive estimator is obtained from the previous one by a simple adjustment. In particular, we consider a class of estimators

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \Gamma_n^{-1}(\hat{\theta}_{n-1})\psi_n(\hat{\theta}_{n-1}), \qquad n \geq 1, \tag{2}$$

where ψ_n is a suitably chosen vector process, Γ_n is a normalizing matrix process, and $\hat{\theta}_0 \in \mathbb{R}^m$ is an initial value. A detailed discussion and a heuristic justification of this estimation procedure are given in Sharia (2007a).

In i.i.d. models, estimating procedures of type (2) have been studied by a number of authors using methods of stochastic approximation theory. Some work has been done for non i.i.d. models as well. A review of the existing literature can be found in Sharia (2007a).

We study multidimensional estimation procedures of type (2) for the general statistical model. This work can be regarded as the final part of a series of three papers: see Sharia (2007a) and Sharia (2007b). In Sharia (2007a), we study convergence of the recursive estimators for an arbitrary starting value $\hat{\theta}_0$. In Sharia (2007b), we present results on the rate of the convergence. In this paper, we are concerned with asymptotic behaviour of the estimators defined by (2). The main objective is to prove that $\hat{\theta}_n$ is locally asymptotically linear, that is, for each θ there exist a matrix process $G_n(\theta)$ such that

$$\hat{\theta}_n - \theta = G_n^{-1}(\theta)\sum_{i=1}^{n}\psi_i(\theta) + \varepsilon_n^{\theta},$$

where $G_n^{1/2}(\theta)\varepsilon_n^{\theta} \to 0$ in probability P^{θ} (see Section 2).

Since $\psi_t(\theta)$ is typically a martingale-difference, asymptotic distribution of an asymptotically linear estimator can be studied using a suitable form of the central limit theorem for martingales; see, e.g., Feigin (1985), Hutton and Nelson (1986), Jacod and Shiryayev (1987). Detailed discussion of the literature on this subject can be found in Barndorff-Nielsen and Sorensen (1994), Heyde (1997) and Prakasa-Rao (1999). For example, results in Shiryayev (1984) (see, e.g., Ch.VII, §8, Theorem 4) show that, under certain conditions, local asymptotic linearity implies asymptotic normality.

In the case of one dimensional parameter θ, an estimator is said to be *asymptotically efficient* if it is asymptotically linear with

$$\psi_n(\theta) = f_n'(\theta, X_n|X_1^{n-1})/f_n(\theta, X_n|X_1^{n-1}) \quad \text{and} \quad G_n(\theta) = I_n(\theta),$$

where $I_n(\theta)$ is the conditional Fisher information. This kind of efficiency is called asymptotic first order efficiency. The motivation behind this general

definition is the same as in the classical scheme of i.i.d. observations. For a detailed discussion of this notion see, e.g., Hall and Heyde (1980), Section 6.2. Under relatively mild conditions, asymptotically efficient estimators are asymptotically equivalent to the MLE T_n, i.e.

$$I_n^{1/2}(\theta)(\hat{\theta}_n - T_n) \to 0$$

in probability; see, e.g., Hall and Heyde (1980), Section 6.2, Theorem 6.2. A generalisation of these concepts can be found in Heyde (1997).

Note that the recursive procedure (2) is not a numerical solution of (1) (see the corresponding discussion in Sharia (2007a)). Nevertheless, as the results of the paper show, the recursive estimator and the corresponding M-estimator are expected to have the same or equivalent asymptotic linearity expansions under quite mild conditions. It therefore follows that they are asymptotically equivalent, in the sense that, depending on the regularity and ergodicity properties of the underlying model, they both have the same asymptotic distribution.

Note also that the global convergence results for (2) were obtained in Sharia (2007a) under conditions that allow Γ_n to belong to quite a wide class of processes which does not directly depend on the choice of ψ_n's. It turns out that to ensure local asymptotic linearity, one has to restrict this class to an explicit choice of Γ_n, depending on the choice of ψ_n. In other words, the results of the paper can be used to determine the form of a recursive procedure (see Remark 3 (iv)–(vi) below) which is expected to have the same asymptotic properties as the corresponding non-recursive one defined as a solution of the equation (1). The fact that one is restricted to this choice of Γ_t is probably not very surprising in retrospective, but this issue does not seem to have been discussed in the existing literature.

The paper is organized as follows. Section 2 introduces the main objects and definitions. The main results are obtained in Section 3 which also contains various comments and explanations of the conditions used there. In Section 4, we consider examples to illustrate the results of the paper.

2 Basic model

Let $X_t, \quad t = 1, 2, \ldots,$ be observations taking values in a measurable space $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ equipped with a σ-finite measure μ. Suppose that the distribution of the process X_t depends on an unknown parameter $\theta \in \Theta$, where Θ is an open subset of the m-dimensional Euclidean space $\mathbb{R}^m$. Suppose also that for each $t = 1, 2, \ldots,$ there exists a regular conditional probability density of X_t given values of past observations of $X_{t-1}, \ldots, X_2, X_1$, which will be denoted by

$$f_t(\theta, x_t \mid x_1^{t-1}) = f_t(\theta, x_t \mid x_{t-1}, \ldots, x_1),$$

where $f_1(\theta, x_1 \mid x_1^0) = f_1(\theta, x_1)$ is the probability density of the random variable X_1. Without loss of generality we assume that all random variables are

defined on a probability space $(\Omega, \mathcal{F})$ and denote by $\left\{P^\theta,\ \theta \in \Theta\right\}$ the family of the corresponding distributions on $(\Omega, \mathcal{F})$.

Let $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ be the σ-field generated by the random variables $X_1, \dots, X_t$. By $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ we denote the m-dimensional Euclidean space with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^m)$. Transposition of matrices and vectors is denoted by T. By (u, v) we denote the standard scalar product of $u, v \in \mathbb{R}^m$, that is, $(u, v) = u^T v$, and the corresponding norm is denoted by $\|u\|$.

Suppose that h is a real valued function defined on $\Theta \subset \mathbb{R}^m$. We denote by $\dot{h}(\theta)$ the row-vector of partial derivatives of $h(\theta)$ with respect to the components of θ, that is,

$$\dot{h}(\theta) = \left(\frac{\partial}{\partial \theta^1} h(\theta), \dots, \frac{\partial}{\partial \theta^m} h(\theta) \right).$$

The $m \times m$ identity matrix is denoted by $\mathbf{1}$.

If for each $t = 1, 2, \dots$, the derivative $\dot{f}_t(\theta, x_t \mid x_1^{t-1})$ w.r.t. θ exists, then we can define

$$l_t(\theta, x_t \mid x_1^{t-1}) = \frac{1}{f_t(\theta, x_t \mid x_1^{t-1})} \dot{f}_t^T(\theta, x_t \mid x_1^{t-1})$$

and the process

$$l_t(\theta) = l_t(\theta, X_t \mid X_1^{t-1})$$

with the convention $0/0 = 0$. Let us denote

$$i_t(\theta \mid x_1^{t-1}) = \int l_t(\theta, z \mid x_1^{t-1}) l_t^T(\theta, z \mid x_1^{t-1}) f_t(\theta, z \mid x_1^{t-1}) \mu(dz).$$

The *one step conditional Fisher information matrix* for $t = 1, 2, \dots$ is defined as

$$i_t(\theta) = i_t(\theta \mid X_1^{t-1}).$$

Note that the process $i_t(\theta)$ is *"predictable"*, that is, $i_t(\theta)$ is $\mathcal{F}_{t-1}$ measurable for each $t \geq 1$. Note also that by definition, $i_t(\theta)$ is a version of the conditional expectation w.r.t. $\mathcal{F}_{t-1}$, that is,

$$i_t(\theta) = E_\theta \left\{ l_t(\theta) l_t^T(\theta) \mid \mathcal{F}_{t-1} \right\}.$$

Everywhere in the present work conditional expectations are meant to be calculated as integrals w.r.t. the conditional probability densities.

The *conditional Fisher information* at time t is

$$I_t(\theta) = \sum_{s=1}^{t} i_s(\theta), \qquad t = 1, 2, \dots.$$

We say that $\psi = \{\psi_t(\theta, x_t, x_{t-1}, \dots, x_1)\}_{t \geq 1}$ is a sequence of estimating functions and write $\psi \in \mathbf{\Psi}$, if for each $t \geq 1$, $\psi_t(\theta, x_t, x_{t-1}, \dots, x_1) : \Theta \times \mathbf{X}^t \to \mathbb{R}^m$ is a Borel function.

Let $\psi \in \mathbf{\Psi}$ and denote $\psi_t(\theta) = \psi_t(\theta, X_t, X_{t-1}, \ldots, X_1)$. We write $\psi \in \mathbf{\Psi^M}$ if $\psi_t(\theta)$ is a martingale-difference process for each $\theta \in \Theta$, i.e., if $E_\theta \{\psi_t(\theta) \mid \mathcal{F}_{t-1}\} = 0$ for each $t = 1, 2, \ldots$. We assume that the conditional expectations above are well-defined and $\mathcal{F}_0$ is the trivial σ-algebra.

Note that if differentiation of the equation $1 = \int f_t(\theta, z \mid x_1^{t-1})\mu(dz)$ is allowed under the integral sign, then $\{l_t(\theta)\}_{t\geq 1} \in \mathbf{\Psi^M}$.

Suppose that $\psi \in \mathbf{\Psi}$ and $\mathit{\Gamma}_t(\theta)$ is a predictable $m \times m$ matrix process with $\det \mathit{\Gamma}_t(\theta) \neq 0$. We say that an estimator $\hat{\theta}_t$ is *locally asymptotically linear* if for each $\theta \in \Theta$,

$$\hat{\theta}_t = \theta + \mathit{\Gamma}_t^{-1}(\theta) \sum_{s=1}^{t} \psi_s(\theta) + \varepsilon_t^\theta, \tag{3}$$

and $A_t(\theta)\varepsilon_t^\theta \to 0$ in probability P_θ, where $A_t(\theta)$ is a sequence of invertible $m \times m$ matrices such that $A_t^{-1}(\theta) \to 0$ in probability P^θ and $A_t(\theta)\mathit{\Gamma}_t^{-1}(\theta)A_t(\theta) \to \eta(\theta)$ weakly w.r.t. P^θ for some random matrix $\eta(\theta)$. That is, $\hat{\theta}_t$ is locally asymptotically linear if

$$A_t(\theta)(\hat{\theta}_t^* - \hat{\theta}_t) \to 0 \tag{4}$$

in probability P^θ, where

$$\hat{\theta}_t^* = \theta + \mathit{\Gamma}_t^{-1}(\theta) \sum_{s=1}^{t} \psi_s(\theta), \tag{5}$$

is a linear statistic.

Convention *Everywhere in the present work $\theta \in \mathbb{R}^m$ is an arbitrary but fixed value of the parameter. Convergence and all relations between random variables are meant with probability one w.r.t. the measure P^θ unless specified otherwise. A sequence of random variables $(\xi_t)_{t\geq 1}$ has some property eventually if for every ω in a set Ω^θ of P^θ probability 1, ξ_t has this property for all t greater than some $t_0(\omega) < \infty$.*

3 Main results

Suppose that $\psi \in \mathbf{\Psi}$ and $\mathit{\Gamma}_t(\theta)$, for each $\theta \in \mathbb{R}^m$, is a predictable $m \times m$ matrix process with $\det \mathit{\Gamma}_t(\theta) \neq 0$, $t \geq 1$. Consider the estimator $\hat{\theta}_t$ defined by

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \mathit{\Gamma}_t^{-1}(\hat{\theta}_{t-1})\psi_t(\hat{\theta}_{t-1}), \qquad t \geq 1, \tag{6}$$

where $\hat{\theta}_0 \in \mathbb{R}^m$ is an arbitrary initial point.

Let $\theta \in \mathbb{R}^m$ be an arbitrary but fixed value of the parameter and for any $u \in \mathbb{R}^m$ define

$$R_t(\theta, u) = \mathit{\Gamma}_t(\theta)\mathit{\Gamma}_t^{-1}(\theta + u)E_\theta \{\psi_t(\theta + u) \mid \mathcal{F}_{t-1}\}.$$

Denote $\mathit{\Delta}_t = \hat{\theta}_t - \theta$. Then (6) can be rewritten as

$$\mathit{\Delta}_t = \mathit{\Delta}_{t-1} + \mathit{\Gamma}_t^{-1}(\theta)R_t(\theta, \mathit{\Delta}_{t-1}) + \mathit{\Gamma}_t^{-1}(\theta)\varepsilon_{\theta t}, \tag{7}$$

where

$$\varepsilon_{\theta t} = \Gamma_t(\theta)\Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1}) - R_t(\theta, \Delta_{t-1})$$

is a P^θ-martingale difference.

Lemma 1 *Suppose that $\psi \in \mathbf{\Psi}$ and there exists a sequence of invertible random matrices $A_t(\theta)$ such that $A_t^{-1}(\theta) \to 0$ in probability P^θ and*

(E)

$$A_t(\theta)\Gamma_t^{-1}(\theta)A_t(\theta) \to \eta(\theta)$$

weakly w.r.t. P^θ, where $\eta(\theta)$ is a random matrix with $\eta(\theta) < \infty$ P^θ-a.s.;

(S1)

$$\lim_{t\to\infty} A_t^{-1}(\theta) \sum_{s=1}^{t} (\triangle\Gamma_s(\theta)\Delta_{s-1} + R_s(\theta, \Delta_{s-1})) = 0$$

in probability P^θ;

(S2)

$$\lim_{t\to\infty} A_t^{-1}(\theta) \sum_{s=1}^{t} \mathcal{E}_s(\theta) = 0$$

in probability P^θ, where

$$\mathcal{E}_s(\theta) = \Gamma_s(\theta)\Gamma_s^{-1}(\theta+\Delta_{s-1}) \left[\psi_s(\theta + \Delta_{s-1}) - E_\theta \left\{\psi_s(\theta + \Delta_{s-1}) \mid \mathcal{F}_{s-1}\right\}\right] - \psi_s(\theta).$$

Then $A_t(\theta)(\hat\theta_t^* - \hat\theta_t) \to 0$ in probability P^θ, i.e., $\hat\theta_t^*$ is locally asymptotically linear.

Proof. To simplify notation we drop the fixed argument or the index θ in some of the expressions below. Rewrite (7) as

$$\Delta_t = \left(\mathbf{1} - \Gamma_t^{-1}\triangle\Gamma_t\right)\Delta_{t-1} + \Gamma_t^{-1}(\triangle\Gamma_t\Delta_{t-1} + R_t(\theta, \Delta_{t-1})) + \Gamma_t^{-1}\varepsilon_t. \quad (8)$$

Denote $\mathcal{H}_t := \sum_{s=1}^t (\triangle\Gamma_s\Delta_{s-1} + R_s(\theta, \Delta_{s-1}))$ and $\bar M_t := \sum_{s=1}^t \varepsilon_s$. Then the expression

$$\Delta_t = \Gamma_t^{-1}\left\{\bar M_t + \mathcal{H}_t + \Delta_0\right\}, \qquad t \geq 1$$

can easily be obtained by inspecting the difference between t'th and $(t-1)$'th term of this sequence to check that (8) holds. Therefore, denoting

$$\delta_t := \hat\theta_t - \hat\theta_t^* = \Delta_t - (\hat\theta_t^* - \theta),$$

we obtain

$$\delta_t = \Gamma_t^{-1}\left\{M_t + \mathcal{H}_t + \Delta_0\right\}, \qquad t \geq 1,$$

where $M_t := \sum_{s=1}^t (\varepsilon_s - \psi_s)$. Now, (S1) implies that $A_t^{-1}\mathcal{H}_t \to 0$ in probability P^θ. Also, by (S2), $A_t^{-1}M_t = A_t^{-1}(\theta)\sum_{s=1}^t \mathcal{E}_s(\theta) \to 0$ in probability P^θ. So, using (E), it follows that $A_t\delta_t \to 0$ in probability P^θ. $\diamondsuit$

The next result gives sufficient conditions for (S1) and (S2).

Proposition 2 (a) *Suppose that $A_t(\theta)$ in Lemma 1 are diagonal matrices with non-decreasing (w.r.t. t) elements and*

(L1)

$$A_t^{-2}(\theta)\sum_{s=1}^{t} A_s(\theta)[\triangle\Gamma_s(\theta)\Delta_{s-1} + R_s(\theta,\Delta_{s-1})] \to 0$$

in probability P^θ;

Then (S1) holds.

(b) *Suppose that $A_t(\theta)$ in Lemma 1 are diagonal non-random matrices, $\psi \in \mathbf{\Psi^M}$ and*

(L2) for each $j = 1,\dots,m$,

$$\lim_{t\to\infty} \frac{1}{(A_t^{(jj)}(\theta))^2}\sum_{s=1}^{t} E_\theta\left\{\left(\mathcal{E}_s^{(j)}(\theta)\right)^2 \mid \mathcal{F}_{s-1}\right\} = 0$$

in probability P^θ, where $A_t^{(jj)}(\theta)$ is the j-th diagonal element of the matrix $A_t(\theta)$ and $\mathcal{E}_s^{(j)}(\theta)$ is the j-th component of $\mathcal{E}_s(\theta)$ which is defined in (S2).

Then (S2) holds.

(c) *Suppose that $A_t(\theta)$ in Lemma 1 are diagonal with non-decreasing elements $A_t^{(jj)}(\theta) \to \infty$, $\psi \in \mathbf{\Psi^M}$ and*

(LL2) for each $j = 1,\dots,m$,

$$\sum_{s=1}^{\infty} \frac{E_\theta\left\{(\mathcal{E}_s^{(j)}(\theta))^2 \mid \mathcal{F}_{s-1}\right\}}{(A_s^{(jj)}(\theta))^2} < \infty$$

P^θ-a.s., where $\mathcal{E}_s^{(j)}(\theta)$ is the j-th component of $\mathcal{E}_s(\theta)$ which is defined in (S2).

Then (S2) holds.

Proof. See Section 5.

Remark 3 (i) As was mentioned above, strong consistency of the recursive estimator $\hat\theta_t$, that is the convergence $\Delta_t = \hat\theta_t - \theta \to 0$ (P^θ-a.s.) is established in Sharia (2007a). Here we are interested in the asymptotic behaviour of the recursive estimator given that it is consistent. Note that although consistency is not formally required in Lemma 1, it is easy to see that if $\hat\theta_t$ is not consistent, conditions (S1) and (S2) will be satisfied for very special cases only. Note also that given that $\Delta_t = \hat\theta_t - \theta \to 0$, conditions (S1) and (S2) are local in the sense that they are determined by the local behaviour of the corresponding functions w.r.t. the parameter.

(ii) Condition (E) is an ergodicity type assumption on the statistical model. If $\Gamma_t(\theta) = I_t(\theta)$ and $A_t(\theta)$ and $\eta(\theta)$ are non-random, then the model is called

ergodic. Further discussion of this concept and related work appears in Basawa and Scott (1983), Hall and Heyde (1980) § 6.2, and Barndorff-Nielsen and Sorensen (1994).

(iii) Let us examine condition (S2) in Lemma 1. Given that $\Delta_t = \hat{\theta}_t - \theta \to 0$, if the functions $\psi_t(\theta)$ and $\Gamma_t(\theta)$ are continuous w.r.t. θ with certain uniformity w.r.t. t, we expect $\mathcal{E}_t(\theta) \to 0$. Parts (b) and (c) in Proposition 2 give sufficient conditions for (S2). If there exists a non-random sequence $A_t(\theta)$, then obviously (L2) is less restrictive then (LL2). But unfortunately, (L2) can only be used for non-random $A_t(\theta)$. In the case of random $A_t(\theta)$, when (LL2) may be used, just the convergence $E_\theta\left\{(\mathcal{E}_t(\theta))^2 \mid \mathcal{F}_{t-1}\right\} \to 0$ may not be enough since in many models the components of $A_t(\theta)$ have the rate $\sqrt{t}$. In such cases one may also use the result on the rate of convergence of $\hat{\theta}_t$ presented in Sharia (2007b); see examples 4.1 and 4.3 in the next section.

(iv) Condition (S1) gives an important clue for an optimal choice of the normalizing sequence $\Gamma_t(\theta)$. To see this, let us assume that $\psi \in \mathbf{\Psi^M}$ so that $R_t(\theta, 0) = 0$ and consider (S1) and (L1) in the case of the one dimensional parameter $\theta \in \mathbb{R}$. Now we can write

$$\triangle\Gamma_t(\theta)\Delta_{t-1} + R_t(\theta, \Delta_{t-1}) = \left(\triangle\Gamma_t(\theta) + \frac{R_t(\theta, \Delta_{t-1})) - R_t(\theta, 0)}{\Delta_{t-1}}\right)\Delta_{t-1}.$$

In most applications, the rate of A_t is $\sqrt{t}$ and the best one can hope for is that $\sqrt{t}\Delta_t$ is stochastically bounded. Therefore we must at least have the convergence $\triangle\Gamma_t(\theta) + (R_t(\theta, \Delta_{t-1})) - R_t(\theta, 0))/\Delta_{t-1} \to 0$. Given that $\Delta_{t-1} \to 0$ we expect $\triangle\Gamma_t(\theta) \approx -\partial/\partial u\ R_t(\theta, u) \mid_{u=0}$ for large t's. Also, since $R_t(\theta, 0) = E_\theta\{\psi_t(\theta) \mid \mathcal{F}_{t-1}\} = 0$, if $\Gamma_t(\theta)/\Gamma_t(\theta + u)$ is smooth in $u = 0$, we can write that $\partial/\partial u\ R_t(\theta, u) \mid_{u=0} = \partial/\partial u\ E_\theta\{\psi_t(\theta + u) \mid \mathcal{F}_{t-1}\} \mid_{u=0}$. So,

$$\triangle\Gamma_t(\theta) \approx -b'_t(\theta, 0), \tag{9}$$

$$\text{where} \quad b_t(\theta, u) = E_\theta\{\psi_t(\theta + u) \mid \mathcal{F}_{t-1}\} \quad \text{and} \quad b'_t(\theta, 0) = \frac{\partial}{\partial u} b_t(\theta, u) \mid_{u=0} .$$

Using the similar arguments, for the multidimensional case, we expect (9) to hold for large t's, where $b'_t(\theta, 0)$ is the total differential of $b_t(\theta, u)$ in $u = 0$. Therefore,

$$\Gamma_t(\theta) = -\sum_{s=1}^{t} b'_s(\theta, 0) \tag{10}$$

is an obvious candidate for the normalizing sequence. If $\psi_t(\theta)$ is differentiable in θ and differentiation of $b_t(\theta, u) = E_\theta\{\psi_t(\theta + u) \mid \mathcal{F}_{t-1}\}$ is allowed under the integral sign, then $b'_t(\theta, 0) = E_\theta\{\dot{\psi}_t(\theta) \mid \mathcal{F}_{t-1}\}$. This implies that, for a given sequence of estimating functions $\psi_t(\theta)$, another possible choice of the normalizing sequence is

$$\Gamma_t(\theta) = -\sum_{s=1}^{t} E_\theta\{\dot{\psi}_s(\theta) \mid \mathcal{F}_{s-1}\}, \tag{11}$$

or any sequence with the increments $\Delta\Gamma_t(\theta) = -E_\theta\{\dot{\psi}_t(\theta) \mid \mathcal{F}_{t-1}\}$. Also, if the differentiation w.r.t. θ of $0 = \int \psi_t(\theta, z \mid X_1^{t-1}) f_t(\theta, z \mid X_1^{t-1})\mu(dz)$ is allowed under the integral sign, using the product rule it is easy to obtain that $E_\theta\{\dot{\psi}_t(\theta) \mid \mathcal{F}_{t-1}\} = -E_\theta\{\psi_t(\theta) l_t^T(\theta) \mid \mathcal{F}_{t-1}\}$. Therefore, another possible choice of the normalizing sequence is

$$\Gamma_t(\theta) = \sum_{s=1}^{t} E_\theta\{\psi_s(\theta) l_s^T(\theta) \mid \mathcal{F}_{s-1}\} = \langle M^\theta, U^\theta \rangle_t \tag{12}$$

where $\langle M^\theta, U^\theta \rangle_t$ is the mutual quadratic characteristic of the martingales

$$M_t^\theta = \sum_{s=1}^{t} \psi_s(\theta) \quad \text{and} \quad U_t^\theta = \sum_{s=1}^{t} l_s(\theta).$$

(v) Let us consider a likelihood case, that is $\psi_t(\theta) = l_t(\theta)$. Then the process (12) in this case is the conditional Fisher information $I_t(\theta) = \sum_{s=1}^{t} i_s(\theta)$. So, the corresponding recursive procedure is

$$\hat{\theta}_t = \hat{\theta}_{t-1} + I_t^{-1}(\hat{\theta}_{t-1}) l_t(\hat{\theta}_{t-1}), \qquad t \geq 1, \tag{13}$$

Also, given that the model possesses certain ergodicity properties, asymptotic linearity of (13) implies asymptotic efficiency. In particular, in the case of i.i.d. observations, it follows that the above recursive procedure is asymptotically normal with parameters $(0,\ i^{-1}(\theta))$; see Corollary 4 in Section 4.

(vi) Normalizing sequences suggested in (iv) have been derived from the asymptotic considerations. In practice however, behaviour of Γ sequence for the first several steps might also be important. This can happen when the number of observations is small or even moderately large. According to (iv), to achieve asymptotic linearity, one has to choose a normalizing sequence Γ with the property that $\triangle\Gamma_t(\theta) \approx -b_t'(\theta, 0)$ for large t's. So, we can consider any sequence of the form $C + c_t\Gamma_t$, where Γ_t is one of the sequences introduced above (by (10), (11), or (12)), c_t is a sequence of non-negative r.v.'s such that $c_t = 1$ eventually, and C is a suitably chosen constant. In practice, c_t and C can be treated as tuning constants to control behaviour of the procedure for the first several steps; see Remark 4.4 in Sharia (2007a). Under certain assumptions, at each step, the recursive procedure (6) on average moves towards the direction of the unknown parameter; see Remark 3.2 in Sharia (2007a) for details. Nevertheless, if the values of the normalizing sequence are too small for the first several steps, then the procedure will oscillate excessively around the true value of the parameter. On the other hand, too large values of the normalizing sequence will result in slower convergence of the procedure. A good balance can be achieved by using the tuning constants. The detailed discussion of these and related topics will appear elsewhere, but as a rough guide, the graph of $\hat{\theta}_t$ against t should ideally have a shape of those in Figure 1 in Sharia (2007a); that is, a reasonable oscillation at the beginning of the procedure before settling down at a particular level.

4 SPECIAL MODELS AND EXAMPLES

4.1 **The i.i.d. scheme**

Consider the classical scheme of i.i.d. observations $X_1, X_2, \dots$, with a common probability density/mass function $f(\theta, x)$, $\theta \in \mathbb{R}^m$. Suppose that $\psi(\theta, x)$ is an estimating function with

$$E_\theta(\psi(\theta, X_1)) = \int \psi(\theta, z) f(\theta, z) \mu(dz) = 0.$$

Let us define the recursive estimator $\hat{\theta}_t$ by

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{1}{t}\gamma^{-1}(\hat{\theta}_{t-1})\psi(\hat{\theta}_{t-1}, X_t), \qquad t \geq 1, \tag{14}$$

where $\hat{\theta}_0 \in \mathbb{R}^m$ is any initial value. According to Remark 3 (iv) and the condition (V) below, an optimal choice of $\gamma(\theta)$ would be either

$$\gamma(\theta) = E_\theta(\dot{\psi}(\theta, X_1))$$

or

$$\gamma(\theta) = E_\theta(\psi(\theta, X_1) l^T(\theta, X_1)) \qquad \text{where} \qquad l(\theta, x) = \frac{\dot{f}^T(\theta, x)}{f(\theta, x)},$$

or any non-random invertible matrix function that satisfies conditions listed below.

Suppose that

$$j_\psi(\theta) = \int \psi(\theta, z)\psi^T(\theta, z) f(\theta, z) \mu(\, dz) < \infty$$

and consider the following conditions.

(I) For any $0 < \varepsilon < 1$,

$$\sup_{\varepsilon \leq \|u\| \leq \frac{1}{\varepsilon}} u^T \ \gamma^{-1}(\theta + u) \int \psi(\theta + u, x) f(\theta, x) \mu(\, dx) < 0.$$

(II) For each $u \in \mathbb{R}^m$,

$$\int \left\|\gamma^{-1}(\theta + u)\psi(\theta + u, x)\right\|^2 f(\theta, x) \mu(\, dx) \leq K_\theta(1 + \|u\|^2)$$

for some constant K_θ.

(III) $\gamma(\theta)$ is continuous in θ.

(IV)

$$\lim_{u \to 0} \int \|\psi(\theta + u, x) - \psi(\theta, x)\|^2 f(\theta, x) \mu(\, dx) = 0.$$

(V)
$$\int \psi(\theta+u,x)f(\theta,x)\mu(\,dx) = -\gamma(\theta+u)u + \alpha^{\theta}(u),$$

where $\alpha^{\theta}(u) = o(\|u\|^{1+\varepsilon})$ as $u \to 0$ for some $\varepsilon > 0$.

Corollary 4 *Suppose that for any $\theta \in \mathbb{R}^m$ conditions* **(I)** - **(V)** *are satisfied. Then the estimator $\hat{\theta}_t$ is strongly consistent and $t^{\delta}(\hat{\theta}_t - \theta) \to 0$ (P^{θ}-a.s.) for any $0 < \delta < 1/2$ and any initial value $\hat{\theta}_0$. Furthermore, $\hat{\theta}_t$ is asymptotically normal with parameters $(0,\ \gamma^{-1}(\theta)j(\theta,0)\gamma^{-1}(\theta))$, that is,*

$$\mathcal{L}\left(t^{1/2}(\hat{\theta}_t - \theta) \mid P^{\theta}\right) \xrightarrow{w} \mathcal{N}\left(0,\ \gamma^{-1}(\theta)j_{\psi}(\theta)\gamma^{-1}(\theta)\right).$$

In particular, in the case of the maximum likelihood type recursive procedure with $\psi(\theta,x) = \dot{f}^T(\theta,x)/f(\theta,z)$ and $\gamma(\theta) = i(\theta) = j_l(\theta)$, the estimator $\hat{\theta}_t$ is asymptotically efficient, i.e., asymptotically normal with parameters $\left(0,\ i^{-1}(\theta)\right)$.

Proof See Section 5.

The results in Corollary 4 are similar to those obtained in the classical works by Khas'minskii and Nevelson (1972; see Ch.8, §4) and Fabian (1978).

4.2 **Linear procedures**

Consider the recursive procedure

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \Gamma_t^{-1}\left(h_t - \gamma_t\hat{\theta}_{t-1}\right), \quad t \geq 1, \tag{15}$$

where the Γ_t and γ_t are predictable matrix processes, h_t is an adapted process, i.e., h_t is $\mathcal{F}_t$-measurable for $t \geq 1$. Assume also that all three processes are independent of θ. The following result gives a set of sufficient conditions for the asymptotic linearity of the estimator defined by (15) in the case when the linear $\psi_t(\theta) = h_t - \gamma_t\theta$ is a martingale-difference, i.e., $E_{\theta}\{h_t \mid \mathcal{F}_{t-1}\} = \gamma_t\theta$, for $t \geq 1$.

Corollary 5 *Suppose that $\Gamma_t \to \infty$ and*

$$\Gamma_t^{-1/2}\sum_{s=1}^{t}(\triangle\Gamma_s - \gamma_s)\Delta_{s-1} \to 0 \tag{16}$$

in probability P^{θ}, where $\Delta_{s-1} = \hat{\theta}_{s-1} - \theta$. Then the recursive estimator defined by (15) *is asymptotically linear with*

$$\Gamma_t^{1/2}(\hat{\theta}_t - \theta) = \Gamma_t^{-1/2}\sum_{s=1}^{t}\psi_s(\theta) + o_{P^{\theta}}(1), \tag{17}$$

where $o_{P^{\theta}}(1) \to 0$ in probability P_{θ}.

Proof Let us check the conditions of Lemma 1 for $A_t(\theta) = \Gamma_t^{1/2}$. Condition (E) trivially holds. Then, since $\psi_t(\theta) = h_t - \gamma_t\theta$ and

$$b_t(\theta, u) = E_\theta\left\{(\psi_t(\theta + u)) \mid \mathcal{F}_{t-1}\right\} = E_\theta\left\{(h_t - \gamma_t(\theta + u)) \mid \mathcal{F}_{t-1}\right\} = -\gamma_t u,$$

we have

$$R_t(\theta, u) = \Gamma_t(\theta)\Gamma_t^{-1}(\theta + u)b_t(\theta, u) = -\gamma_t u.$$

Therefore, (S1) is equivalent to (16). Then, it is easy to see that for $\mathcal{E}_s(\theta)$ defined in (S2) we have

$$\mathcal{E}_s(\theta) = \psi_s(\theta + \Delta_{s-1}) - b_s(\theta, \Delta_{s-1}) - \psi_s(\theta) = 0$$

implying that (S2) holds which completes the proof. $\diamondsuit$

Remark 6 Condition (16) trivially holds if $\Delta\Gamma_t = \gamma_t$, that is if $\Gamma_t = \sum_{s=1}^t \gamma_s$. In this case, the solution of (15) is

$$\hat{\theta}_t = \Gamma_t^{-1}\left(\hat{\theta}_0 + \sum_{s=1}^t h_s(X_s)\right). \tag{18}$$

This can be easily seen by inspecting the difference $\hat{\theta}_t - \hat{\theta}_{t-1}$ for the sequence (18) to check that (15) holds. Also, since (18) can obviously be rewritten as

$$\hat{\theta}_t = \Gamma_t^{-1}\hat{\theta}_0 + \Gamma_t^{-1}\sum_{s=1}^t (h_s(X_s) - \gamma_s\theta) + \theta,$$

it follows that in this case, $\Gamma_t \to \infty$ is indeed an obvious necessary and sufficient condition for $\hat{\theta}_t$ to be asymptotically linear for arbitrary starting value $\hat{\theta}_0$.

4.3 **Exponential family of Markov processes**

Consider a conditional exponential family of Markov processes in the sense of Feigin (1981); see also Barndorf-Nielson (1988). This is a time homogeneous Markov chain with the one-step transition density

$$f(y; \theta, x) = h(x, y)\exp\left(\theta^T m(y, x) - \beta(\theta; x)\right),$$

where $m(y, x)$ is a m-dimensional vector and $\beta(\theta; x)$ is one dimensional. Then in our notation $f_t(\theta) = f(X_t; \theta, X_{t-1})$ and

$$l_t(\theta) = \left(\frac{d}{d\theta}\log f_t(\theta)\right)^T = m(X_t, X_{t-1}) - \dot{\beta}^T(\theta; X_{t-1}).$$

It follows from standard exponential family theory (see, e.g., Feigin (1981)) that $l_t(\theta)$ is a martingale-difference and the conditional Fisher information is

$$I_t(\theta) = \sum_{s=1}^t \ddot{\beta}(\theta; X_{s-1}).$$

A maximum likelihood type recursive procedure can be defined as

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \left(\sum_{s=1}^{t} \ddot{\beta}(\hat{\theta}_{t-1}; X_{s-1}) \right)^{-1} \left(m(X_t, X_{t-1}) - \dot{\beta}^T(\hat{\theta}_{t-1}; X_{t-1}) \right), \quad t \geq 1.$$

Now suppose that θ is one dimensional and the process belongs to the conditionally additive exponential family, that is,

$$f(y; \theta, x) = h(x, y) \exp\left(\theta m(y, x) - \beta(\theta; x)\right),$$

with

$$\beta(\theta; x) = \gamma(\theta) h(x) \tag{19}$$

where $h(\cdot) \geq 0$ and $\ddot{\gamma}(\cdot) \geq 0$ (see Feigin (1981)). Then,

$$I_t(\theta) = \ddot{\gamma}(\theta) H_t \quad \text{where} \quad H_t = \sum_{s=1}^{t} h(X_{s-1}).$$

Assuming that $\ddot{\gamma}(\theta) \neq 0$, the likelihood recursive procedure is

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{1}{\ddot{\gamma}(\hat{\theta}_{t-1}) H_t} \left(m(X_t, X_{t-1}) - \dot{\gamma}(\hat{\theta}_{t-1}) h(X_{t-1}) \right). \tag{20}$$

Remark 7 Consistency and rate of convergence of the estimator derived by (20) are studied in Sharia (2007b). To ensure that (20) has the same asymptotic properties as the maximum likelihood estimator, one has to impose certain restrictions on the $\gamma(\theta)$ and H_t. In Corollary 9 in Section 5, the conditions of Section 3 written in terms of this model are presented. These conditions will be satisfied if there is a certain balance between requirements of smoothness on $\gamma(\cdot)$, the rate at which $H_t \to \infty$, and ergodicity of the model. For instance, suppose that the model is ergodic, that is, there exists a non-random sequence $\tilde{H}_t$ such that

$$H_t / \tilde{H}_t \to \eta < \infty$$

weakly. This will often follow from an ergodic theorem with $\tilde{H}_t = t$. Then

$$\frac{1}{I_t^{1/2}(\theta)} \sum_{s=1}^{t} \mathcal{E}_s(\theta) \to 0,$$

will hold if the process

$$\frac{1}{I_t(\theta)} \sum_{s=1}^{t} E_\theta \left\{ \mathcal{E}_s^2(\theta) \mid \mathcal{F}_{s-1} \right\} = \frac{1}{I_t(\theta)} \sum_{s=1}^{t} \triangle I_s(\theta) \left(\frac{\ddot{\gamma}(\theta + \Delta_{s-1}) - \ddot{\gamma}(\theta)}{\ddot{\gamma}(\theta + \Delta_{s-1})} \right)^2$$

converges to zero; criterion based on the Lenglart-Rebolledo inequality, see (L2) and formula (26) in Section 5. So, assuming that the estimator is consistent, that is $\Delta_t \to 0$, by the Toeplits lemma the above will be guaranteed by the continuity of $\ddot{\gamma}(\cdot)$. On the other hand, if the model is non-ergodic, then one

may need to impose smoothness of higher order on $\gamma(\cdot)$ function (see condition (iii) below) and restrictions on the growth of the sequence H_t (see condition (i) below). The following result gives one possible set of sufficient conditions for the recursive estimator to be consistent and to have the same asymptotic properties as the maximum likelihood estimator.

Proposition 8 *Suppose that* $H_t \to \infty$ *and*

(i)

$$\frac{h(X_t)}{H_t} \to 0;$$

(ii) there exists a constant B *such that*

$$\frac{1+\dot{\gamma}^2(u)}{\ddot{\gamma}^2(u)} \leq B(1+u^2)$$

for each $u \in \mathbb{R}$.

(iii) The function $\ddot{\gamma}(\cdot)$ *is locally Lipschitz , that is, for any* θ *there exists a constant* K_θ *and* $0 < \varepsilon_\theta \leq 1/2$ *such that*

$$|\ddot{\gamma}(\theta+u) - \ddot{\gamma}(\theta)| \leq K_\theta |u|^{\varepsilon_\theta}$$

for small u*'s.*

Then $\hat{\theta}_t$ *defined by* (20) *is strongly consistent, i.e.,* $\hat{\theta}_t \to \theta \;\; P^\theta$*-a.s. for any initial value* $\hat{\theta}_0$*. Furthermore,* $H_t^\delta(\hat{\theta}_t - \theta) \to 0 \;\; P^\theta$*-a.s. for any* $\delta \in]0, 1/2[$*, and* $\hat{\theta}_t$ *is asymptotically linear with*

$$H_t^{1/2}(\hat{\theta}_t - \theta) = H_t^{-1/2} \sum_{s=1}^{t} (m(X_s, X_{s-1}) - \dot{\gamma}(\theta) h(X_{s-1})) + o_{P^\theta}(1), \qquad (21)$$

where $o_{P^\theta}(1) \to 0$ *in probability* P_θ*.*

Proof See Section 5.

5 Appendix

Proof of Proposition 2 To simplify notation we drop the fixed argument or the index θ in some of the expressions below.

To prove **(a)**, denote

$$\chi_s = A_s[\triangle \Gamma_s(\theta) \Delta_{s-1} + R_s(\theta, \Delta_{s-1})]$$

and

$$\mathcal{G}_t = A_t^{-1} \sum_{s=1}^{t} \; [\triangle \Gamma_s(\theta) \Delta_{s-1} + R_s(\theta, \Delta_{s-1})] = A_t^{-1} \sum_{s=1}^{t} \; A_s^{-1} \chi_s.$$

Applying the formula

$$\sum_{s=1}^{t} D_s \Delta C_s = D_t C_t - \sum_{s=1}^{t} \Delta D_s C_{s-1}, \;\; C_0 = 0 = D_0,$$

with $C_s = \sum_{m=1}^{s} \chi_m$ and $D_s = A_s^{-1}$ we obtain

$$\mathcal{G}_t = A_t^{-2} \sum_{s=1}^{t} \chi_s - A_t^{-1} \sum_{s=1}^{t} \triangle A_s^{-1} \sum_{m=1}^{s-1} \chi_m.$$

Then, $\triangle A_s^{-1} = A_s^{-1} - A_{s-1}^{-1} = -A_s^{-1}(A_s - A_{s-1})A_{s-1}^{-1} = -\triangle A_s A_s^{-1} A_{s-1}^{-1}$, where the last equality follows since A_s is diagonal. Therefore,

$$\mathcal{G}_t = A_t^{-2} \sum_{s=1}^{t} \chi_s + A_t^{-1} \sum_{s=1}^{t} \triangle A_s \left\{ A_s^{-1} A_{s-1}^{-1} \sum_{m=1}^{s-1} \chi_m \right\}.$$

Finally, since A_t's are diagonal with non-decreasing elements, applying the Toeplits Lemma to the components of the right hand side of latter formula we obtain that $\mathcal{G}_t \to 0$.

To prove **(b)** and **(c)** denote $M_t := \sum_{s=1}^{t} \mathcal{E}_s$. Since $\psi \in \mathbf{\Psi^M}$, it follows that M_t is a martingale. Denote by $M_t^{(j)}$ the j-th component of M_t. Then the quadratic characteristic $\langle M^{(j)} \rangle_t$ of the martingale $M_t^{(j)}$ is

$$\langle M^{(j)} \rangle_t = \sum_{s=1}^{t} E_\theta \left\{ \left(\mathcal{E}_s^{(j)} \right)^2 \mid \mathcal{F}_{s-1} \right\}$$

and, by (LL2), $\sum_{s=1}^{\infty} \triangle \langle M^{(j)} \rangle_s / (A_s^{(jj)})^2 < \infty$. It therefore follows that $M_t^{(j)} / {A_t}^{(jj)} \to 0$ P^θ -a.s. (see e.g., Shiryayev (1984), Ch.VII, §5, Theorem 4). This proves (c). Now, use of the Lenglart-Rebolledo inequality (see, e.g., Liptser and Shiryayev (1989), Ch.1, §9) yields

$$P^\theta \left\{ (M_t^{(j)})^2 \geq K^2 \left({A_t}^{(jj)} \right)^2 \right\} \leq \frac{\varepsilon}{K} + P^\theta \left\{ \langle M^{(j)} \rangle_t \;\geq\; \varepsilon \left({A_t}^{(jj)} \right)^2 \right\}$$

for each $K > 0$ and $\varepsilon > 0$. Then, by (L2), $\langle M^{(j)} \rangle_t / ({A_t}^{(jj)})^2 \to 0$ in probability P^θ. This implies that $M_t^{(j)} / {A_t}^{(jj)} \to 0$ in probability P^θ and so, since A_t is diagonal, (S2) follows. $\diamondsuit$

Proof of Corollary 4 Using Corollary 4.1 in Sharia (2007a) it follows that (I) and (II) imply $(\hat{\theta}_t - \theta) \to 0$. We have $\mathit{\Gamma}_t(\theta) = t\gamma(\theta)$ and $b(\theta, u) = \int \psi(\theta + u, z) f(\theta, z) \mu(\, dz)$. It is easy to see that (II) implies (B2) from Corollary 4.1 in Sharia (2007b), and (V) implies that (B1) of the same corollary holds with $C_\theta = \mathbf{1}$. So, for any $0 < \delta < 1/2$,

$$t^\delta (\hat{\theta}_t - \theta) \to 0. \tag{22}$$

Let us check that conditions of Lemma 1 are also satisfied with $A_t = \sqrt{t}\mathbf{1}$. Condition (E) trivially holds. According to Proposition 2, condition (S1) follows from (L1). To check (L1), it is sufficient to show that

$$\frac{1}{t}\sum_{s=1}^{t}[\gamma(\theta)\Delta_{s-1} + R(\theta, \Delta_{s-1})]\sqrt{s} \to 0, \tag{23}$$

where

$$R(\theta, u) = R_t(\theta, u) = \gamma(\theta)\gamma^{-1}(\theta + u)\int \psi(\theta + u, z)f(\theta, z)\mu(\,dz).$$

By (V), $\quad R(\theta, u) = -\gamma(\theta)u + \gamma(\theta)\gamma^{-1}(\theta + u)\alpha^{\theta}(u) \quad$ and

$$[\gamma(\theta)\Delta_{s-1}+R(\theta, \Delta_{s-1})]\sqrt{s} = \sqrt{s}\gamma(\theta)\gamma^{-1}(\theta+\Delta_{s-1})\alpha^{\theta}(\Delta_{s-1}) = \sqrt{s}\|\Delta_{s-1}\|^{1+\varepsilon}\delta_s$$

where, by (III) and (V), $\delta_s = \gamma(\theta)\gamma^{-1}(\theta + \Delta_{s-1})\alpha^{\theta}(\Delta_{s-1})/\|\Delta_{s-1}\|^{1+\varepsilon} \to 0$. Then,

$$\sqrt{s}\|\Delta_{s-1}\|^{1+\varepsilon}\delta_s = \sqrt{\frac{s}{s-1}}\left((s-1)^{\frac{1}{2(1+\varepsilon)}}\|\Delta_{s-1}\|\right)^{1+\varepsilon}\delta_s$$

which, by (22) (since $1/(2(1 + \varepsilon)) < 1/2$) converges to zero. Therefore, (23) is now a consequence of the Toeplits Lemma.

For the process $\mathcal{E}_s(\theta)$ from (L2) (since $\|u - v\|^2 \le 2\|u\|^2 + 2\|v\|^2$), we have

$$\begin{aligned}\|\mathcal{E}_s(\theta)\|^2 &= \|\gamma(\theta)\gamma^{-1}(\theta + \Delta_{s-1})\left(\psi(\theta + \Delta_{s-1}, X_s) - b(\theta, \Delta_{s-1})\right) - \psi(\theta, X_s)\|^2\\ &\le 2\|\gamma(\theta)\gamma^{-1}(\theta + \Delta_{s-1})\psi(\theta + \Delta_{s-1}, X_s) - \psi(\theta, X_s)\|^2\\ &\quad +2\|\gamma(\theta)\gamma^{-1}(\theta + \Delta_{s-1})b(\theta, \Delta_{s-1})\|^2.\end{aligned}$$

(III) and (V) imply that $\left(\gamma(\theta)\gamma^{-1}(\theta + \Delta_{s-1}) - \mathbf{1}\right) \to 0$ and $b(\theta, \Delta_{s-1}) \to 0$ as $s \to \infty$. So, using (IV), it is easy to see that $E_\theta\left\{\left(\mathcal{E}_s^{(j)}(\theta)\right)^2 \mid \mathcal{F}_{s-1}\right\} \to 0$. Since $(A_t^{(jj)}(\theta))^2 = t$, (L2) follows from the Toeplitz lemma.

Therefore, the conditions of Lemma 1 hold for $A_t(\theta) = \sqrt{t}\mathbf{1}$. This implies that $\sqrt{t}(\hat{\theta}_t - \theta_t^*) \to 0$ in probability P^θ, where

$$\theta_t^* = \frac{1}{t\gamma(\theta)}\sum_{s=1}^{t}\psi_s(\theta, X_s).$$

The asymptotic normality now obviously follows from the central limit theorem for i.i.d. random variables. $\diamondsuit$

Corollary 9 *Suppose that $H_t \to \infty$ and $\hat{\theta}_t$ is derived by* (20). *Denote $\Delta_t = \hat{\theta}_t - \theta$, $l_t(\theta) = m(X_t, X_{t-1}) - \dot{\gamma}(\theta)h(X_{t-1})$, and suppose also that*

(I)

$$H_t^{-1/2}\sum_{s=1}^{t}\mathcal{E}_s(\theta)\to 0,$$

where

$$\mathcal{E}_s(\theta)=\frac{\ddot{\gamma}(\theta+\Delta_{s-1})-\ddot{\gamma}(\theta)}{\ddot{\gamma}(\theta+\Delta_{s-1})}l_s(\theta);$$

(II) one of the following two conditions are satisfied;

$$H_t^{-1/2}\sum_{s=1}^{t}\triangle H_s\mathcal{C}_s(\theta)\to 0,$$

OR

$$H_t^{-1}\sum_{s=1}^{t}\triangle H_s\ H_s^{1/2}\mathcal{C}_s(\theta)\to 0,$$

where

$$\mathcal{C}_s(\theta)=\frac{\ddot{\gamma}(\theta+\Delta_{s-1})-\ddot{\gamma}(\theta+\tilde{\Delta}_{s-1})}{\ddot{\gamma}(\theta+\Delta_{s-1})}\Delta_{s-1}$$

and $\tilde{\Delta}_t$ is a predictable process with $|\tilde{\Delta}_t|\le|\Delta_t|$.

Then (21) *holds, i.e., the estimator $\hat{\theta}_t$ is asymptotically linear.*

Proof. Let us check the conditions of Lemma 1 for $\psi_t(\theta)=l_t(\theta)$,

$$\Gamma_t(\theta)=I_t(\theta)=\ddot{\gamma}(\theta)H_t \tag{24}$$

and $A_t(\theta)=H_t^{1/2}$. Since $l_t(\theta)$ is a martingale-difference, we have $E_\theta\left\{m(X_t,X_{t-1})\mid\mathcal{F}_{t-1}\right\}=\dot{\gamma}(\theta)h(X_{t-1})$ and so

$$b_t(\theta,u)=E_\theta\left\{l_t(\theta+u)\mid\mathcal{F}_{t-1}\right\}=h(X_{t-1})\left(\dot{\gamma}(\theta)-\dot{\gamma}(\theta+u)\right) \tag{25}$$

and

$$R_t(\theta,u)=\frac{\ddot{\gamma}(\theta)}{\ddot{\gamma}(\theta+u)}h(X_{t-1})(\dot{\gamma}(\theta)-\dot{\gamma}(\theta+u))=-\frac{\ddot{\gamma}(\theta)}{\ddot{\gamma}(\theta+u)}h(X_{t-1})\dddot{\gamma}(\theta+\tilde{u})u$$

where $|\tilde{u}|\le|u|$. Then, since $\triangle\Gamma_t(\theta)=\triangle I_t(\theta)=h(X_{t-1})\ddot{\gamma}(\theta)$ we have

$$\triangle\Gamma_t(\theta)u+R_t(\theta,u)=h(X_{t-1})\ddot{\gamma}(\theta)\frac{\ddot{\gamma}(\theta+u)-\ddot{\gamma}(\theta+\tilde{u})}{\ddot{\gamma}(\theta+u)}u.$$

Now, since $\triangle H_t=h(X_{t-1})$, it is easy to see that the first condition in (II) implies (S1) in Lemma 1 and the second condition in (II) implies (L1) in Proposition 2. Therefore, (S1) holds.

To verify (S2), consider the process $\mathcal{E}_s(\theta)$ defined in (S2). Using (24) and (25), it is easy to see that

$$\mathcal{E}_s(\theta)=\left(1-\frac{\ddot{\gamma}(\theta)}{\ddot{\gamma}(\theta+\Delta_{s-1})}\right)\left(m(X_s,X_{s-1})-\dot{\gamma}(\theta)h(X_{s-1})\right)$$

$$= \frac{\ddot{\gamma}(\theta + \Delta_{s-1}) - \ddot{\gamma}(\theta)}{\ddot{\gamma}(\theta + \Delta_{s-1})} l_s(\theta). \tag{26}$$

This shows that (I) implies (S2). $\diamondsuit$

Proof of Proposition 8 Since, by (iii), $\ddot{\gamma}(\cdot)$ is obviously a continuous function, condition (M2) of Proposition 4.1 in Sharia (2007b) holds. Also, (M1) in the same proposition obviously follows from (i). So, it follows that all the conditions of Proposition 4.1 and Corollary 4.2 in Sharia (2007b) are satisfied implying that $H_t^{\delta}(\hat{\theta}_t - \theta) \to 0$ (P^{θ}-a.s.). Also, by (i), $\triangle H_t / H_{t-1} = h(X_{t-1})/H_{t-1} \to 0$ implying that $H_t/H_{t-1} = 1 + \triangle H_t/H_{t-1} \to 1$. So,

$$H_t^{\delta} \Delta_{t-1} = H_t^{\delta}(\hat{\theta}_{t-1} - \theta) \to 0. \tag{27}$$

To establish asymptotic linearity let us verify the conditions of Corollary 9. Since $\Delta_{s-1} = \hat{\theta}_{s-1} - \theta \to 0$ (P^{θ}-a.s.) and $|\tilde{\Delta}_{s-1}| \le |\Delta_{s-1}|$, by (iii) we obtain that $|\ddot{\gamma}(\theta + \Delta_{s-1}) - \ddot{\gamma}(\theta + \tilde{\Delta}_{s-1})| \le 2K_{\theta}|\Delta_{s-1}|^{\varepsilon_{\theta}}$ eventually. So,

$$|H_s^{\frac{1}{2}} \mathcal{C}_s(\theta)| = H_s^{\frac{1}{2}} \frac{|\ddot{\gamma}(\theta + \Delta_{s-1}) - \ddot{\gamma}(\theta + \tilde{\Delta}_{s-1})||\Delta_{s-1}|}{\ddot{\gamma}(\theta + \Delta_{s-1})} \le \frac{2K_{\theta} H_s^{\frac{1}{2}} |\Delta_{s-1}|^{1+\varepsilon_{\theta}}}{\ddot{\gamma}(\theta + \Delta_{s-1})}$$

eventually. Now,

$$H_s^{\frac{1}{2}} |\Delta_{s-1}|^{1+\varepsilon_{\theta}} = |H_s^{\frac{1}{2(1+\varepsilon_{\theta})}} (\hat{\theta}_{s-1} - \theta)|^{1+\varepsilon_{\theta}} \to 0,$$

by (27) since $\frac{1}{2(1+\varepsilon_{\theta})} < \frac{1}{2}$. So, since the function $\ddot{\gamma}(\cdot)$ is continuous, we obtain that $|H_s^{\frac{1}{2}} \mathcal{C}_s(\theta)| \to 0$. Therefore, by the Toeplits Lemma, the second condition of (II) holds.

Now, since $\mathcal{E}_s(\theta)$ is a martingale-difference, to verify (I), it is sufficient to show that (see e.g., Shiryayev (1984), Ch.VII, §5, Theorem 4),

$$\sum_{s=1}^{\infty} \frac{E_{\theta}\left\{\mathcal{E}_s^2(\theta) \mid \mathcal{F}_{s-1}\right\}}{H_s} < \infty.$$

Since $E_{\theta}\{l_s^2(\theta) \mid \mathcal{F}_{s-1}\} = \ddot{\gamma}(\theta) h(X_{s-1}) = \ddot{\gamma}(\theta) \triangle H_s$, the above series can be rewritten as

$$\sum_{s=1}^{\infty} \frac{\triangle H_s}{H_s} \ddot{\gamma}(\theta) \left(\frac{\ddot{\gamma}(\theta + \Delta_{s-1}) - \ddot{\gamma}(\theta)}{\ddot{\gamma}(\theta + \Delta_{s-1})} \right)^2 = \ddot{\gamma}(\theta) \sum_{s=1}^{\infty} \frac{\triangle H_s}{H_s^{1+\varepsilon_{\theta}/2}} r_s$$

where, by (iii),

$$r_s = \frac{\left(\ddot{\gamma}(\theta + \Delta_{s-1}) - \ddot{\gamma}(\theta)\right)^2 H_s^{\varepsilon_{\theta}/2}}{\ddot{\gamma}^2(\theta + \Delta_{s-1})} \le K_{\theta}^2 \frac{|\Delta_{s-1}|^{2\varepsilon_{\theta}} H_s^{\varepsilon_{\theta}/2}}{\ddot{\gamma}^2(\theta + \Delta_{s-1})} = K_{\theta}^2 \frac{(|\Delta_{s-1}| H_s^{1/4})^{2\varepsilon_{\theta}}}{\ddot{\gamma}^2(\theta + \Delta_{s-1})}.$$

Now, using (27) and continuity of $\ddot{\gamma}(\cdot)$ we deduce that $r_s \to 0$. Also,

$$\sum_{s=1}^{\infty} \frac{\triangle H_s}{H_s^{1+\varepsilon_{\theta}/2}} < \infty$$

(see Sharia (2007b), Appendix A, Proposition A2), implying that the above series converge which completes the proof. $\diamondsuit$

Acknowledgements I am grateful to the referee for constructive and helpful comments and suggestions.

References

Barndorff-Nielsen, O.E., Sorensen, M. (1994). A review of some aspects of asymptotic likelihood theory for stochastic processes. *International Statistical Review.* **62,** 1, 133-165.

Basawa, I.V., Scott, D.J. (1983). *Asymptotic Optimal Inference for Non-ergodic Models.* Springer-Verlag, New York.

Fabian, V. (1978). On asymptotically efficient recursive estimation. *Annals of Statistics.* **6**, 854-867.

Feigin, P.D. (1981). Conditional exponential families and a representation theorem for asymptotic inference. *Annals of Statistics.* **9**, 597-603.

Feigin, P.D. (1985). Stable convergence for semimartingales. *Stochastic Processes and Their Applications.* **19**, 125–134.

Hall, P., Heyde, C.C. (1980). *Martingale Limit Theory and Its Application.* Academic Press, New York.

Heyde, C.C. (1997). *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter estimation.* Springer-Verlag, New York.

Hutton, J.E., Nelson, P.I. (1986). Quasi-likelihood estimation for semimartingales. *Stochastic Processes and Their Applications.* **22**, 245–257.

Jacod, J., Shiryayev, A.N. (1987). *Limit Theorems for Stochastic Processes.* Heidelberg, Springer.

Khas'minskii, R.Z., Nevelson, M.B. (1972). *Stochastic Approximation and Recursive Estimation.* Nauka, Moscow.

Liptser, R.Sh., Shiryayev, A.N. (1989). *Theory of Martingales.* Kluwer, Dordrecht.

Prakasa Rao, B.L.S. (1999). *Semimartingales and their Statistical Inference.* Chapman & Hall, New York.

Sharia, T. (2007a). Recursive parameter estimation: Convergence. *Statistical Inference for Stochastic Processes* (in press). DOI: 10.1007/s11203-007-9008-x

Sharia, T. (2007b). Rate of convergence in recursive parameter estimation procedures. *Georgian mathematical Journal* (in press) (see also *http://personal.rhul.ac.uk/UkAH/113/GmjA.pdf*).

Shiryayev, A.N. (1984). *Probability.* Springer-Verlag, New York.