On Recursive Parametric Estimation Theory

Teo Sharia

Department of Mathematics
Royal Holloway, University of London
Egham, Surrey TW20 0EX, England
http://www.rhul.ac.uk/mathematics/techreports

Abstract

The classical non-recursive methods to estimate unknown parameters of the model, such as the maximum likelihood method, the method of least squares etc. eventually require maximization procedures. These methods are often difficult to implement, especially for non i.i.d. models. If for every sample size n, when new data are acquired, an estimator has to be computed afresh, and if a numerical method is needed to do so, it generally becomes very laborious. Therefore, it is important to consider recursive estimation procedures which are appealing from the computational point of view. Recursive procedures are those which at each step allow one to re-estimate values of unknown parameters based on the values already obtained at the previous step together with new information. We propose a wide class of recursive estimation procedures for the general statistical model and study convergence, the rate of convergence, and the local asymptotic linearity. Also, we demonstrate the use of the results on some examples.

1   Introduction

Let $X_1, \ldots, X_n$ be independent identically distributed (i.i.d.) random variables (r.v.'s) with a common distribution function F_θ with a real (unknown) parameter θ. An M-estimator of θ is defined as a statistic $\theta_n = \theta_n(X_1, \ldots, X_n)$, which is a solution (w.r.t. v) of the estimating equation

$$(1.1) \qquad \sum_{i=1}^{n} \psi(X_i; v) = 0,$$

where ψ is a suitably chosen function. For example, if θ is a location parameter in the normal family of distribution functions, the choice $\psi(x, v) = x - v$ gives the MLE (maximum likelihood estimator). For the same problem, if $\psi(x, v) = \text{sign}(x - v)$, the solution of (1.1) reduces to the median of $X_1, \ldots, X_n$. Other forms of ψ for location or other models may be chosen from relevant considerations. If $F_\theta(x)$ has a density (w.r.t. a σ-finite measure μ) and the density is differentiable w.r.t. θ, then the choice $\psi(x, v) = f'(x, v)/f(x, v)$ yields the MLE.

Suppose now that $X_1, \ldots, X_n$ are not necessarily independent or identically distributed r.v's, with a joint distribution depending on a real parameter θ. Then an M-estimator of θ is defined as a solution of the estimating

1

equation

$$(1.2) \qquad \sum_{i=1}^{n} \psi_i(v) = 0,$$

where $\psi_i(v) = \psi_i(X_{i-k}^i; v)$ with $X_{i-k}^i = (X_{i-k}, \ldots, X_i)$. Therefore, the ψ-functions may now depend on the past observations as well. For instance, if X_i's are observations from a discrete time Markov process, then one can assume that $k = 1$. In general, if no restrictions are made on the dependence structure of the process X_i, one may need to consider ψ-functions depending on the vector of all past and present observations of the process (that is, $k = i - 1$). If the conditional probability density function of the observation X_i, given $X_{i-k}, \ldots, X_{i-1}$, is $f_i(x, v) = f_i(x, v | X_{i-k}, \ldots, X_{i-1})$, then one can obtain the MLE on choosing $\psi_i(v) = f_i'(X_i, v)/f_i(X_i, v)$. Besides MLEs, the class of M-estimators includes estimators with special properties such as *robustness*. An estimator is said to be robust if its behaviour is not "seriously affected" by violations of underlying assumptions. Under certain regularity and ergodicity conditions it can be proved that there exists a consistent sequence of solutions of (1.2) which has an asymptotic representation of the following type:

$$(1.3) \qquad a_n^{1/2}(\theta)\,(\theta_n - \theta) = a_n^{-1/2}(\theta) \sum_{i=1}^{n} \psi_i(\theta) + o_{p^\theta}(1),$$

where $a_n(\theta) = a_n(\theta, \psi_n)$ is a normalizing sequence. In the i.i.d. case this implies the asymptotic normality result. (See e.g., Huber (1981), Lehman (1983), Serfling (1980). A comprehensive bibliography can be found in Hampel at al (1986), Jurečková and Sen (1996), Launer and Wilkinson (1979), and Rieder (1994).)

If ψ-functions are nonlinear, it is rather difficult to work with the corresponding estimating equations. (It is well-known that, typically, robust M-estimators are nonlinear.) The situation is much more complex if a sequential estimation rule is required. If for every sample size n, when new data are acquired, an estimator has to be computed afresh, and if a numerical method is needed to do so, it generally becomes very laborious. Therefore one can consider recursive estimation procedures which are appealing from the computational point of view. Note that for a linear estimator, e.g., for the sample mean, $\theta_n = \bar{X}_n$ we have $\bar{X}_n = (n-1)\bar{X}_{n-1}/n + X_n/n$, that is $\theta_n = \theta_{n-1}(n-1)/n + X_n/n$, indicating that the estimator θ_n at each step n can be obtained recursively using the estimator at the previous step θ_{n-1} and the new information X_n. Such an exact recursive relation may not hold for nonlinear estimators (see, e.g., the case of median).

In general, the following heuristic argument can be used to establish a possible form of an approximate recursive relation (see also Jurečková and Sen (1996), and Lazrieva and Toronjadze (1987)). Since θ_n is defined as a root of the estimating equation (1.2), denoting the left hand side of (1.2) by $M_n(v)$ we have $M_n(\theta_n) = 0$ and $M_{n-1}(\theta_{n-1}) = 0$. Assuming the difference $\theta_n - \theta_{n-1}$ is "small" we can write

$$0 = M_n(\theta_n) - M_{n-1}(\theta_{n-1}) = M_n\left(\theta_{n-1} + (\theta_n - \theta_{n-1})\right) - M_{n-1}(\theta_{n-1})$$

$$\approx M_n(\theta_{n-1}) + M_n'(\theta_{n-1})(\theta_n - \theta_{n-1}) - M_{n-1}(\theta_{n-1}) = M_n'(\theta_{n-1})(\theta_n - \theta_{n-1}) + \psi_n(\theta_{n-1}).$$

Therefore,

$$\theta_n \approx \theta_{n-1} - \frac{\psi_n(\theta_{n-1})}{M_n'(\theta_{n-1})}.$$

Suppose now that the estimator θ_n is consistent (a.s. converges to the value of the unknown parameter θ as $n \to \infty$). Then we can replace $M_n'(\theta_{n-1})$ by $M_n'(\theta) = \sum_{i=1}^{n} \psi_i'(\theta)$, which in turn, depending on the nature of the underlying model, can be replaced by a simpler expression. For instant, in i.i.d. models with $\psi(x, v) = f'(x, v)/f(x, v)$ (a MLE case), by the strong law of large numbers,

$$\frac{1}{n} M_n'(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(f'(X_i, \theta)/f(X_i, \theta)\right)' \approx E_\theta\left[\left(f'(X_1, \theta)/f(X_1, \theta)\right)'\right] = -i(\theta)$$

for large n's, where $i(\theta)$ is the one-step Fisher information. So, in this case, one can use the recursion

$$\theta_n = \theta_{n-1} + \frac{1}{n\, i(\theta_{n-1})} \frac{f'(X_n, \theta_{n-1})}{f(X_n, \theta_{n-1})},$$

to construct an estimator which is "asymptotically equivalent" to the MLE.

In general, following the above argument, for a multidimensional parameter $\theta \in \mathbb{R}^m$ one can derive an estimator using the recursion

$$(1.4) \qquad \theta_n = \theta_{n-1} + \Gamma_n^{-1}(\theta_{n-1})\psi_n(\theta_{n-1}), \qquad t \geq 1,$$

where ψ_n is a suitably chosen vector process, Γ_n is a (possibly random) normalizing matrix process and θ_0 is some initial point.

In i.i.d. models, similar estimating procedures have been studied by a number of authors using methods of stochastic approximation theory (see, e.g., Khas'minskii and Nevelson (1972), Fabian (1978), Ljung, Pflug and Walk (1992), Titterington, Smith, and Makov (1985), and references therein).

Some work has been done for non i.i.d. models as well. In particular, Englund, Holst, and Ruppert (1989) give an asymptotic representation results for certain type of X_n processes. In Sharia (1998) theoretical results on convergence, the rate of convergence and the asymptotic representation are given under certain regularity and ergodicity assumptions on the model, in the one-dimensional parameter case with $\psi_n(x, \theta) = \frac{d}{d\theta}\log f_n(x, \theta)$ (see also Sharia (1992), Sharia (1997) and Lazrieva, Sharia and Toronjadze (1997)).

In the present paper we study multidimensional estimation procedures of type (1.4) for general statistical model. Section 2 introduces the basic model, objects and notation. In Section 3, imposing "global" restrictions on the processes ψ and Γ, we study "global" convergence of the recursive estimators, that is the convergence for an arbitrary starting point $\hat{\theta}_0$. In Section 4, assuming that $\hat{\theta}_t \to \theta$, we present results on rate of the convergence. In Section 5 we show that under certain regularity and continuity assumptions, the recursive estimators are locally asymptotically linear and therefore, corresponding asymptotic distributions can be determined using a suitable form of the central limit theorem. In Section 6 we demonstrate the use of these results on some examples. Namely, we consider the i.i.d model, the exponential family of Markov processes and robust estimation of parameters of AR(m) process together with a brief simulation study.

2   Basic model, notation and preliminaries

Let X_t, $t = 1, 2, \ldots$, 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 past observations $X_{t-1}, \ldots, X_2, X_1$, 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 function of the random variable X_1. Denote by $\mathcal{F}_t$ $(t = 1, 2, \ldots)$ the σ-field generated by the random variables $X_1, \ldots, X_t$, i.e.

$$\mathcal{F}_t = \sigma(X_1, \ldots, X_t).$$

There is no loss of generality in assuming that the basic space is the canonical space

$$(\Omega, \mathcal{F}) := (\mathbf{X}^\infty, \mathcal{B}(\mathbf{X}^\infty)),$$

4

where $\mathbf{X}^\infty = \{\mathbf{x} : \mathbf{x} = x_1, x_2, \ldots; \quad x_i \in \mathbf{X}\}$ and $\mathcal{B}(\mathbf{X}^\infty)$ is the σ-field generated by the cylindrical sets. Using Tulcea's theorem on extending a measure and the existence of a random sequence (see, e.g., Shiryayev (1984), Ch.II, §9, Theorem 2), we can construct the family $\{P^\theta, \ \theta \in \mathbb{R}\}$ of the corresponding distributions on $(\mathbf{X}^\infty, \mathcal{B}(\mathbf{X}^\infty))$ and identify $X = (X_t)_{t=1,2,\ldots}$ with the coordinate process on $(\mathbf{X}^\infty, \mathcal{B}(\mathbf{X}^\infty))$, that is, $X_t(\mathbf{x}) = x_t$, $t = 1, 2, \ldots$.

We assume that all random variables are defined on the probability space $(\Omega, \mathcal{F})$. 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.

Suppose that
$$h : \Theta \to \mathbb{R}^k, \qquad \Theta \subset \mathbb{R}^m,$$

and θ is an interior point of Θ. Then the (total) differential of h at θ is a $k \times m$ real valued matrix H such that

$$h(u) = h(\theta) + H(u - \theta) + \|u - \theta\|\varepsilon(u),$$

with a function ε satisfying $\lim_{u \to \theta} \varepsilon(u) = 0$.

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

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

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

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

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

with the convention $0/0 = 0$.

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

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

We shall use the notation

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

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

Note that the process $i_t(\theta)$ is *"predictable"*, that is, the random variable $i_t(\theta)$, is $\mathcal{F}_{t-1}$ measurable for each $t \geq 1$.

Note also that by the 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

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

If the X_t's are independent random variables, $I_t(\theta)$ reduces to the standard Fisher information matrix. Sometimes $I_t(\theta)$ is referred as the incremental expected Fisher information. Detailed discussion of this concept and related work, appears in Barndorff-Nielsen and Sorensen (1994) and Prakasa-Rao (1999), Ch. 3.

Let for each $t = 1, 2, \ldots$

$$\psi_t(\theta, x_t, x_{t-1}, \ldots, x_1) \ : \ \Theta \times \mathbf{X}^t \ \rightarrow \ \mathbb{R}^m$$

be Borel functions.

We say that the sequence $\psi = \{\psi_t(\theta, x_t, x_{t-1}, \ldots, x_1)\}_{t \geq 1}$ is a sequence of estimating functions and write $\psi \in \Psi$, if for each $t \geq 1$,

(2.1)
$$E_\theta \left\{ \psi_t(\theta) \mid \mathcal{F}_{t-1} \right\} = 0$$

where

$$\psi_t(\theta) = \psi_t(\theta, X_t, X_{t-1}, \ldots, X_1).$$

(we assume that the conditional expectation in (2.1) is well-defined and $\mathcal{F}_0$ is the trivial σ-algebra).

Note that if the derivative of $f_t(\theta)$ exists and differentiation of the equation

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

is allowed under the integral sign, then $\{l_t(\theta, x_t \mid x_1^{t-1})\}_{t \geq 1} \in \Psi$.

An estimator θ_t is said to be *locally asymptotically linear* if for each $\theta \in \Theta$ there exists a sequence of estimating functions ψ and a real valued predictable

matrix process $A_t(\theta)$ (i.e., a matrix with predictable components $A_t^{ij}(\theta)$) such that $\det A_t(\theta) \neq 0$ and

$$\theta_t = \theta + A_t^{-1}(\theta) \sum_{s=1}^{t} \psi_s(\theta) + R_t^\theta,$$

where $R_t^\theta \to 0$ in probability P^θ.

Asymptotic behaviour of a locally asymptotic linear estimator can be studied using a suitable form of the central limit theorem for martingales (see, e.g., Shiryayev (1984), Ch.VII, §8, Theorem 4).

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 larger than some $t_0(\omega) < \infty$.*

3   Strong consistency

Suppose that ψ is a sequence of estimating functions and $\Gamma_t(\theta)$, for each $\theta \in \mathbb{R}^m$, is a predictable matrix process with $\det \Gamma_t(\theta) \neq 0$, $t \geq 1$. Consider the estimator $\hat{\theta}_t$ defined recursively by the equation

(3.1) $$\hat{\theta}_t = \hat{\theta}_{t-1} + \Gamma_t^{-1}(\hat{\theta}_{t-1}) \psi_t(\hat{\theta}_{t-1}), \qquad t \geq 1,$$

where $\hat{\theta}_0 \in \mathbb{R}^m$ is some 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

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

Theorem 3.1 *Suppose that*

(C1) $u' \Gamma_t^{-1}(\theta + u) b_t(\theta, u) < 0 \qquad if \qquad u \neq 0;$

(C2) *for each $\varepsilon \in (0, 1)$,*

$$\sum_{t=1}^{\infty} \inf_{\varepsilon \leq \|u\| \leq 1/\varepsilon} |u' \Gamma_t^{-1}(\theta + u) b_t(\theta, u)| = \infty;$$

(C3) *there exists a predictable scalar process* $(B_t^\theta)_{t \geq 1}$ *such that*

$$E_\theta \left\{ \|\Gamma_t^{-1}(\theta + u)\psi_t(\theta + u)\|^2 \mid \mathcal{F}_{t-1} \right\} \leq B_t^\theta(1 + \|u\|^2)$$

for each $u \in \mathbb{R}^m$, *and*

$$\sum_{t=1}^\infty B_t^\theta < \infty.$$

Then $\hat\theta_t$ *is strongly consistent (i.e.,* $\hat\theta_t \to \theta$ P^θ*-a.s.) for any initial value* $\hat\theta_0$.

We will prove a more general results on the convergence of the procedure (3.1) (see Theorem 3.2 below), which implies Theorem 3.1. Let us first comment on the conditions used in Theorem 3.1.

Remark 3.1 Conditions (C1), (C2), and (C3) are natural analogues of the corresponding assumptions in the theory of stochastic approximation. Indeed, start with the i.i.d. case with

$$f_t(\theta, z \mid x_1^{t-1}) = f(\theta, z), \quad \psi_t(\theta) = \psi(\theta, z)|_{z=X_t},$$

where $\int \psi(\theta, z)f(\theta, z)\mu(dz) = 0$ and $\Gamma_t(\theta) = t\gamma(\theta)$ for some invertible non-random matrix $\gamma(\theta)$. In this case,

$$b_t(\theta, u) = b(\theta, u) = \int \psi(\theta + u, z)f(\theta, z)\mu(dz).$$

Denote $\Delta_t = \hat\theta_t - \theta$ and rewrite (3.1) in the form

$$(3.2) \qquad \Delta_t = \Delta_{t-1} + \frac{1}{t}\gamma^{-1}(\theta + \Delta_{t-1})b(\theta, \Delta_{t-1}) + \varepsilon_t^\theta,$$

where

$$\varepsilon_t^\theta = \frac{1}{t}\gamma^{-1}(\theta + \Delta_{t-1})\left\{\psi(\theta + \Delta_{t-1}, X_t) - b(\theta, \Delta_{t-1})\right\}.$$

Equation (3.2) defines a Robbins-Monro stochastic approximation procedure that locates the solution of the equation

$$R^\theta(u) := \gamma^{-1}(\theta + u)b(\theta, u) = 0,$$

when the values of the function $R^\theta(u)$ can only be observed with zero expectation errors ε_t^θ. Note that in the general, recursion (3.1) cannot be considered in the framework of classical stochastic approximation theory (see Lazrieva, Sharia, and Toronjadze (1997, 2003) for the generalized Robbins-Monro stochastic approximations procedures). For the i.i.d. case, conditions

8

(C1), (C2) and (C3) can be written as **(I)** and **(II)** in Section 6 which are usual assumptions for stochastic approximation procedures of type (3.2) (see, e.g., Robbins and Monro (1951), Khas'minskii and Nevelson (1972), Ljung, Pflug and Walk (1992)).

Remark 3.2 Let us consider the maximum likelihood type recursive estimator

$$(3.3) \qquad \hat{\theta}_t = \hat{\theta}_{t-1} + I_t^{-1}(\hat{\theta}_{t-1}) l_t(\hat{\theta}_{t-1}), \qquad t \geq 1,$$

where $l_t(\theta)$ is the likelihood function of the model and $I_t(\theta)$ is the conditional Fisher information with $\det I_t(\theta) \neq 0$. As we will see in Section 5, $I_t(\theta)$ is a "suitable" normalizing sequence for the recursion with the influence process $l_t(\theta)$. By Theorem 3.1, $\hat{\theta}_t$ is strongly consistent if conditions (C1), (C2) and (C3) are satisfied with $l_t(\theta)$ and $I_t(\theta)$ replacing $\psi_t(\theta)$ and $\Gamma_t(\theta)$ respectively. On the other hand, if $b_t(\theta, u)$ is differentiable at $u = 0$ and the differentiation is allowed under the integral sign, then

$$\dot{b}_t(\theta, 0) = E_\theta \left\{ \dot{l}_t(\theta) \mid \mathcal{F}_{t-1} \right\}.$$

Also, if the differentiation w.r.t. θ of $E_\theta \left\{ \dot{l}_t(\theta) \mid \mathcal{F}_{t-1} \right\} = 0$ is allowed under the integral sign and the corresponding integrals exist (see (5.7) with $\psi = l$), we obtain that

$$\dot{b}_t(\theta, 0) = -i_t(\theta)$$

implying that (C1) always holds for small u's.

Condition (C2) in the i.i.d. case is a requirement of the separateness of the function $\gamma^{-1}(\theta + u) b(\theta, u)$ from zero on each finite interval that does not contain 0. For the i.i.d. case with continuous w.r.t u functions $b(\theta, u)$ and $i(\theta + u)$, condition (C2) is an easy consequence of (C1).

Condition (C3) is a boundedness type assumption which restricts the growth of the estimating function $\psi_t(\theta)$ w.r.t. θ with certain uniformity w.r.t. t.

Denote by η^+ (respectively η^-) the positive (respectively negative) part of η.

Theorem 3.2 *Suppose that for $\theta \in \mathbb{R}^m$ there exists a real valued nonnegative function $V_\theta(u) : \mathbb{R}^m \longrightarrow \mathbb{R}$ having continuous and bounded partial second derivatives and*

(G1) *$V_\theta(0) = 0$ and for each $\varepsilon \in (0, 1)$,*

$$\inf_{\|u\| \geq \varepsilon} V_\theta(u) > 0;$$

9

(G2) *for each $\varepsilon \in (0, 1)$,*

$$\sum_{t=1}^{\infty} \inf_{\varepsilon \leq V_\theta(u) \leq 1/\varepsilon} [\mathcal{N}_t(u)]^- = \infty;$$

(G3) *for $\Delta_t = \hat{\theta}_t - \theta$,*

$$\sum_{t=1}^{\infty} (1 + V_\theta(\Delta_{t-1}))^{-1} [\mathcal{N}_t(\Delta_{t-1})]^+ < \infty,$$

where

$$\mathcal{N}_t(u) = (V_\theta'(u))^T \Gamma_t^{-1}(\theta+u) b_t(\theta, u) + \frac{1}{2} \sup_v \|V_\theta''(v)\| E_\theta \left\{ \|\Gamma_t^{-1}(\theta + u)\psi_t(\theta + u)\|^2 \mid \mathcal{F}_{t-1} \right\},$$

$u \in \mathbb{R}^m$.

Then $\hat{\theta}_t \to \theta$ P^θ-a.s. for any initial value $\hat{\theta}_0$.

Proof. Rewrite (3.1) in the form

$$\Delta_t = \Delta_{t-1} + \Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1}).$$

By the Taylor expansion,

(3.4) $\qquad V_\theta(\Delta_t) = V_\theta(\Delta_{t-1}) + (V_\theta'(\Delta_{t-1}))^T \Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1})$

$$+ \frac{1}{2} \left[\Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1}) \right]^T V_\theta''(\tilde{\Delta}_t) \Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1}),$$

where $\tilde{\Delta}_t \in \mathbb{R}^m$. Taking the conditional expectations w.r.t. $\mathcal{F}_{t-1}$ yields

(3.5) $\qquad E_\theta \{ V_\theta(\Delta_t) \mid \mathcal{F}_{t-1} \} \leq V_\theta(\Delta_{t-1}) + \mathcal{N}_t(\Delta_{t-1}).$

Using the obvious decomposition $\mathcal{N}_t(\Delta_{t-1}) = [\mathcal{N}_t(\Delta_{t-1})]^+ - [\mathcal{N}_t(\Delta_{t-1})]^-$, the previous inequality can be rewritten as

$$E_\theta \{ V_\theta(\Delta_t) \mid \mathcal{F}_{t-1} \} \leq V_\theta(\Delta_{t-1})(1 + B_t) + B_t - [\mathcal{N}_t(\Delta_{t-1})]^-,$$

where

$$B_t = (1 + V_\theta(\Delta_{t-1}))^{-1} [\mathcal{N}_t(\Delta_{t-1})]^+.$$

By condition (G3), $\sum_{t=1}^{\infty} B_t < \infty$. The application of Lemma A1 in Appendix gives that the processes $V_\theta(\Delta_t)$ and

$$Y_t = \sum_{t=1}^{t} [\mathcal{N}_t(\Delta_{t-1})]^-$$

converge to some finite limits. It therefore follows that $V_\theta(\Delta_t) \to r \geq 0$. Because of (G2), $\{r > 0\}$ would imply that $\{Y_t \to \infty\}$. Therefore, $r = 0$ and so, $V_\theta(\Delta_t) \to 0$. The assertion now follows from (G1). $\diamond$

4 Rate of convergence

Throughout this section, $\hat{\theta}_t$ is a sequence defined by (3.1) and $\Delta_t = \hat{\theta}_t - \theta$.

Lemma 4.1 *Let $\{C_t(\theta)\}$ be a predictable matrix process such that $C_t(\theta)$ is positive definite for $t = 1, 2, \ldots$. Denote $V_t(u) = (C_t(\theta)u, u)$ and $\triangle V_t(u) = V_t(u) - V_{t-1}(u)$. Suppose that*

$$(4.1) \qquad \sum_{t=1}^{\infty} (1 + V_{t-1}(\Delta_{t-1}))^{-1} [\mathcal{K}_t(\theta)]^+ < \infty$$

where

$$(4.2) \qquad \mathcal{K}_t(\theta) = \triangle V_t(\Delta_{t-1}) + 2 \left(C_t(\theta)\Delta_{t-1}, \Gamma_t^{-1}(\theta + \Delta_{t-1})b_t(\theta, \Delta_{t-1}) \right)$$
$$+ E_\theta \left\{ \left[\Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1}) \right]^T C_t(\theta)\Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1}) \mid \mathcal{F}_{t-1} \right\}.$$

Then $V_t(\Delta_t)$ converges to a finite limit.

Proof. To simplify notation we drop the argument or the index θ in some of the expressions below. Use of $V_t(u) = (C_t u, u)$ in (3.4) yields

$$(4.3) \quad V_t(\Delta_t) = V_t(\Delta_{t-1}) + 2 \left(C_t \Delta_{t-1}, \Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1}) \right)$$
$$+ \left[\Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1}) \right]^T C_t \Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1}).$$

Since

$$(4.4) \qquad V_t(\Delta_{t-1}) = V_{t-1}(\Delta_{t-1}) + \triangle V_t(\Delta_{t-1}),$$

we have

$$E_\theta \left\{ V_t(\Delta_t) \mid \mathcal{F}_{t-1} \right\} = V_{t-1}(\Delta_{t-1}) + \mathcal{K}_t.$$

Then, using the obvious decomposition $\mathcal{K}_t = [\mathcal{K}_t]^+ - [\mathcal{K}_t]^-$, the previous inequality can be rewritten as

$$E_\theta \left\{ V_t(\Delta_t) \mid \mathcal{F}_{t-1} \right\} = V_{t-1}(\Delta_{t-1})(1 + C_t) + C_t - [\mathcal{K}_t]^-,$$

where

$$C_t = (1 + V_{t-1}(\Delta_{t-1}))^{-1} [\mathcal{K}_t]^+.$$

Now, the assertion of the theorem follows immediately from Lemma A1 in Appendix. $\diamondsuit$

Corollary 4.1 *Let $\{a_t(\theta)\}$ be a predictable non-decreasing scalar process such that $a_t(\theta) \to \infty$ as $t \to \infty$. Denote $\triangle a_t(\theta) = a_t(\theta) - a_{t-1}(\theta)$ and suppose that*

(R1)
$$\lim_{t\to\infty} \frac{\triangle a_t(\theta)}{a_{t-1}(\theta)} = 0;$$

(R2) *there exists a symmetric and positive definite matrix C_θ such that*

$$\left(C_\theta\Delta_{t-1}, \Gamma_t^{-1}(\theta + \Delta_{t-1})b_t(\theta, \Delta_{t-1})\right) \leq -\lambda_t(\theta)\left(C_\theta\Delta_{t-1}, \Delta_{t-1}\right),$$

eventually, where $\{\lambda_t(\theta)\}$ is a predictable scalar process, satisfying

$$\sum_{s=1}^{\infty}\left[\frac{\triangle a_t(\theta)}{a_t(\theta)} - 2\lambda_t(\theta)\right]^+ < \infty;$$

(R3) *for each $0 < \varepsilon < 1$*

$$\sum_{s=1}^{\infty} a_t^\varepsilon(\theta) E_\theta\left\{\|\Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1})\|^2 \mid \mathcal{F}_{t-1}\right\} < \infty.$$

Then
$$a_t(\theta)^\delta(\hat\theta_t - \theta) \to 0$$

for any $\delta \in]0, 1/2[$.

Proof. Let us check the conditions of Lemma 4.1 for $C_t(\theta) = C_\theta(a_t(\theta))^{2\delta}$, $\delta \in]0, 1/2[$. To simplify notation we drop the fixed argument or the index θ in some of the expressions below. Denote

$$\mathcal{P}_t = a_t^{2\delta} E_\theta\left\{\left[\Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1})\right]^T C\left[\Gamma_t^{-1}(\theta + \Delta_{t-1})\psi_t(\theta + \Delta_{t-1})\right] \mid \mathcal{F}_{t-1}\right\}$$

and
$$r_t = \frac{\triangle a_t^{2\delta} - 2a_t^{2\delta}\lambda_t}{a_{t-1}^{2\delta}}.$$

Then, by (R2), for the process $\mathcal{K}_t$ defined in (4.2) we have

$$
\begin{aligned}
\mathcal{K}_t &= \triangle a_t^{2\delta}\left(C\Delta_{t-1}, \Delta_{t-1}\right) + 2a_t^{2\delta}\left(C\Delta_{t-1}, \Gamma_t^{-1}(\theta + \Delta_{t-1})b_t(\theta, \Delta_{t-1})\right) + \mathcal{P}_t \\
&\leq \left(\triangle a_t^{2\delta} - 2a_t^{2\delta}\lambda_t\right)\left(C\Delta_{t-1}, \Delta_{t-1}\right) + \mathcal{P}_t \\
&\leq r_t\left(a_{t-1}^{2\delta} C\Delta_{t-1}, \Delta_{t-1}\right) + \mathcal{P}_t
\end{aligned}
$$

Since C is positive definite,

$$\left(1 + V_{t-1}(\Delta_{t-1})\right)^{-1}\left[\mathcal{K}_t\right]^+ = \left(1 + \left(a_{t-1}^{2\delta} C\Delta_{t-1}, \Delta_{t-1}\right)\right)^{-1}\left[\mathcal{K}_t\right]^+ \leq [r_t]^+ + \mathcal{P}_t.$$

The finiteness of $\sum_{t=1}^{\infty} \mathcal{P}_t^{\theta}$ is guaranteed by (R3) and so, (4.1) reduces to

$$\sum_{t=1}^{\infty} [r_t]^+ < \infty.$$

Since $\triangle a_t^{2\delta} = a_t^{2\delta} - a_{t-1}^{2\delta}$, we can rewrite r_t as

$$r_t = \left(a_t a_{t-1}^{-1}\right)^{2\delta} (1 - 2\lambda_t) - 1.$$

Also, since $(1+x)^{2\delta} = 1 + 2\delta x + O(x^2)$, we have

$$(a_t a_{t-1}^{-1})^{2\delta} = \left(1 + \frac{\triangle a_t}{a_{t-1}}\right)^{2\delta} = 1 + 2\delta\frac{\triangle a_t}{a_{t-1}} + \delta_t^{(1)},$$

where, by (R1), $\delta_t^{(1)} = O\left(\triangle a_t / a_{t-1}\right)^2 \to 0$ as $t \to \infty$. Denote

$$\eta_t = \frac{\triangle a_t}{a_t} - 2\lambda_t.$$

Then simple calculations show that

$$
\begin{aligned}
r_t \quad &\leq \left(a_t a_{t-1}^{-1}\right)^{2\delta}\left(1 + \eta_t^+ - \frac{\triangle a_t}{a_t}\right) - 1 \\
&= -(1 - 2\delta)\frac{\triangle a_t}{a_{t-1}} + \delta_t^{(1)} + \eta_t^+ + 2\delta\eta_t^+\frac{\triangle a_t}{a_{t-1}} + \eta_t^+\delta_t^{(1)} + (1 - 2\delta)\frac{\triangle a_t}{a_t}\frac{\triangle a_t}{a_{t-1}} - \frac{\triangle a_t}{a_t}\delta_t^{(1)} \\
&= \frac{\triangle a_t}{a_{t-1}}\left(-(1 - 2\delta) + \delta_t^{(2)}\right) + \delta_t^{(3)}
\end{aligned}
$$

where

$$\delta_t^{(2)} = \left(\frac{\triangle a_t}{a_{t-1}}\right)^{-1}\delta_t^{(1)}(1 - \frac{\triangle a_t}{a_t}) + (1 - 2\delta)\frac{\triangle a_t}{a_t}$$

and

$$\delta_t^{(3)} = \eta_t^+ + 2\delta\eta_t^+\frac{\triangle a_t}{a_{t-1}} + \eta_t^+\delta_t^{(1)}.$$

From (R1) and (R2),

$$\delta_t^{(2)} \to 0 \quad \text{and} \quad \sum_{t=1}^{\infty} |\delta_t^{(3)}| < \infty.$$

Then, since $1 - 2d > 0$,

$$[r_t]^+ \leq |\delta_t^{(3)}|.$$

It therefore follows that the conditions of Lemma 4.1 are satisfied implying that $a_t^{2\delta}\|\hat{\theta}_t - \theta)\|$ converges to a finite limit. Finally, since this holds for an arbitrary $\delta \in]0, 1/2[$ and $a_t \to \infty$, the result follows. $\diamond$

13

Corollary 4.2 *Consider the i.i.d. case with*

$$\psi_t(\theta) = \psi(\theta, X_t) \quad \text{and} \quad \Gamma_t(\theta) = t\gamma(\theta).$$

Suppose that $\hat\theta \to \theta$ and

(B1) *there exists a symmetric and positive definite matrix C_θ such that*

$$\left(C_\theta u, \gamma^{-1}(\theta+u)E^\theta\psi(\theta+u, X_1)\right) \le -\frac{1}{2}\left(C_\theta u, u\right),$$

for small u 's;

(B2) $\qquad\qquad E_\theta\|\gamma^{-1}(\theta+u)\psi(\theta+u)\|^2 = O(1) \ \text{as } u \to 0.$

Then, for any $\delta \in\,]0, 1/2[$,
$$t^\delta(\hat\theta_t - \theta) \to 0.$$

Proof. The result follows immediately if we take $a_t(\theta) = t$ and $\lambda_t(\theta) = 1/(2t)$ in Corollary 4.1. $\diamondsuit$

Remark 4.1 As it was mentioned in Remark 3.1, for the i.i.d. case the recursive procedures can be studied in the framework of stochastic approximation theory. For stochastic approximation procedures of this type, conditions that guarantee a good rate of the convergence are expressed in terms of stability of matrices. Recall that a matrix A is called stable if the real parts of its eigenvalues are negative. A common requirement in stochastic approximation theory is existence of the representation (see Remark 3.1 for the notation)

$$R^\theta(u) = B^\theta u + o(\|u\|) \ \text{ as } \ u \to 0,$$

where the matrix

$$S^\theta = B^\theta + \frac{1}{2}\mathbf{1}$$

is stable. It is easy to see that this assumption implies (B1). Indeed, it follows from the stability of S^θ that the maximum of the real parts of the eigenvalues of B^θ is less that $-1/2$. This implies (see, e.g., Khas'minskii and Nevelson (1972), Ch.6, §3, Corollary 3.1), that their exists a symmetric and positive definite matrix C_θ such that

$$(C_\theta u, B_\theta u) < -\frac{1}{2}\left(C_\theta u, u\right),$$

and therefore (B1) follows.

5 Asymptotic representation

Define

$$(5.1) \qquad c_t(\theta, u) = \begin{cases} -\Gamma_t(\theta)\Gamma_t^{-1}(\theta + u)b_t(\theta, u)u'/\|u\|^2 & \text{if } u \neq 0 \\[2mm] \triangle\Gamma_t(\theta) & \text{if } u = 0. \end{cases}$$

(A more revealing form for $c_t(\theta, u)$ is given in Remark 5.2 (iii).) Denote as before $\Delta_t = \hat{\theta}_t - \theta$. Then (3.1) can be rewritten as

$$(5.2) \qquad \Delta_t = \left(1 - \Gamma_t^{-1}(\theta)c_t(\theta, \Delta_{t-1})\right)\Delta_{t-1} + \Gamma_t^{-1}(\theta)\varepsilon_t^\theta,$$

where

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

is a P^θ-martingale difference.

The aim of this section is to show that $\hat{\theta}_t$ is asymptotically linear, that is, the main term in the asymptotic representation of $\hat{\theta}_t$ is a linear statistic

$$(5.3) \qquad \hat{\theta}_t^* = \theta + \Gamma_t^{-1}(\theta)\sum_{s=1}^{t}\psi_s(\theta).$$

Let $\Delta_0^* = 0$ and for $t \geq 1$ denote $\Delta_t^* = \hat{\theta}_t^* - \theta$. It is easy to verify, by inspection of the difference $\Delta_t^* - \Delta_{t-1}^*$, that Δ_t^* satisfies the recursive relation given by

$$(5.4) \qquad \Delta_t^* = \left(1 - \Gamma_t^{-1}(\theta)\triangle\Gamma_t(\theta)\right)\Delta_{t-1}^* + \Gamma_t^{-1}(\theta)\varepsilon_t^*$$

where $t \geq 1$ and $\varepsilon_t^* = \psi_t(\theta)$ is a P^θ-martingale difference. By comparing equations (5.2) and (5.4), one can obtain the following result on the asymptotic relationship between $\hat{\theta}_t$ and $\hat{\theta}_t^*$.

Theorem 5.1 *Suppose there exists a non-random sequence of diagonal invertible matrices $A_t(\theta)$ such that*

(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 $\det \eta(\theta) \neq 0$;

(L1)

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

in probability P^θ;

(L2)

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

in probability P^θ, where

$$\mathcal{E}_s(\theta) = \Gamma_s(\theta) \Gamma_s^{-1}(\theta + \Delta_{s-1}) \left(\psi_s(\theta + \Delta_{s-1}) - b_s(\theta, \Delta_{s-1}) \right) - \psi_s(\theta).$$

Then, in probability P^θ,

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

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

$$\delta_t := \hat{\theta}_t - \hat{\theta}_t^*.$$

Subtraction (5.4) from (5.2) yields the recursive relation

$$\delta_t = \left(1 - \Gamma_t^{-1} \triangle \Gamma_t \right) \delta_{t-1} + \Gamma_t^{-1}(\varepsilon_t - \varepsilon_t^*) + \Gamma_t^{-1}(\triangle \Gamma_t - c_t(\theta, \Delta_{t-1})) \Delta_{t-1}.$$

Denote

$$M_t := \sum_{s=1}^{t} [\varepsilon_s - \varepsilon_s^*] \qquad \text{and} \qquad \mathcal{H}_t := \sum_{s=1}^{t} [\triangle \Gamma_s - c_s(\theta, \Delta_{s-1})] \Delta_{s-1}.$$

Then the expression

$$\delta_t = \Gamma_t^{-1} \left\{ M_t + \mathcal{H}_t + \delta_0 \right\}, \qquad t \geq 1$$

can easily be obtained by inspecting the difference between t'th and $(t-1)$'th term of this sequence to check that it satisfies the above recursive relation.

So, we have to prove that $A_t \delta_t \to 0$ in probability P^θ. Condition (L1) implies that $A_t^{-1} \mathcal{H}_t \to 0$ in probability P^θ. Because of (E), it remains only to prove that $A_t^{-1} M_t \to 0$ in probability P^θ. Denote by $M_t^{(j)}$ the j-th component of the martingale M_t. Then the square characteristic $\langle M^{(j)} \rangle_t$ of the martingale $M_t^{(j)}$ is

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

where $\mathcal{E}_s$ is defined in (L2). From (L2), since A_t is diagonal,

$$\left(A_t^{-1(jj)} \right)^2 \langle M^{(j)} \rangle_t \to 0$$

16

in probability P^θ. Now, use of the Lenglart-Rebolledo inequality (see, e.g., Liptser and Shiryayev (1989), Ch.1, §9) yields

$$P^\theta \left\{ (M_t^{(j)})^2 \geq K^2 \left(A_t^{-1(jj)} \right)^{-2} \right\} \leq \frac{\varepsilon}{K} + P^\theta \left\{ \langle M^{(j)} \rangle_t \geq \varepsilon \left(A_t^{-1(jj)} \right)^{-2} \right\}$$

for each $K > 0$ and $\varepsilon > 0$. This implies that $A_t^{-1(jj)} M_t^{(j)} \to 0$ in probability P^θ and since A_t is diagonal, the result follows. $\diamondsuit$

Remark 5.1 Results in Section give sufficient conditions for convergence of sequences of the form $A_t(\theta)\Delta_t$. It may therefore be useful to have a sufficient for (L1) condition written in the following form.

(LL) The elements of $A_t(\theta)$ are non-decreasing processes with $A_t(\theta)_{jj} \to \infty$ and

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

Proposition (LL) implies (L1).
Proof To simplify notation we drop the fixed argument or the index θ in some of the expressions below. Denote

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

and

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

Applying the formula

$$(5.5) \qquad \sum_{s=1}^{t} D_s\Delta C_s = D_t C_t - \sum_{s=1}^{t}\Delta D_s C_{s-1}, \quad C_0 = 0 = D_0,$$

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,

$$\triangle A_s^{-1} = A_s^{-1} - A_{s-1}^{-1} = -A_s^{-1}(A_s - A_{s-1})A_{s-1}^{-1} = -\triangle A_s A_s^{-1} A_{s-1}^{-1},$$

17

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, $\mathcal{G}_t \to 0$ follows on application of the Toeplits Lemma to the elements of $\mathcal{G}_t$.

Remark 5.2
(i) Condition (E) is an ergodicity type assumption on the statistical model It trivially holds if $\Gamma_t(\theta)$ is non-random or, e.g., if $\Gamma_t(\theta)/t \to \gamma(\theta)$ as $t \to \infty$ for some non-random invertible matrix $\gamma(\theta)$. Further discussion of this concept and related work appears in Hall and Heyde (1980), § 6.2., Barndorff-Nielsen and Sorensen (1994) and Basawa and Scott (1983).

(ii) Condition (L2) is a continuity type assumption on functions $\psi_t(\theta)$ and $\Gamma_t(\theta)$ w.r.t. θ.

(iii) In Section 3, while studying the convergence problem, no connections have been assumed between the estimating functions and the normalizing sequence. In Sections 4, to ensure a good rate of convergence, one has to assume certain asymptotic relationship between these two quantities by imposing conditions (4.1) or (R2). While these conditions still leave a lot of room for flexible choice of the normalizing sequence, the condition (L1) assumes more asymptotic balance between $\psi_t(\theta)$ and $\Gamma_t(\theta)$. To understand this relationship, let us first explore the behaviour of $c_t(\theta, u)$ as $u \to 0$. Since $b_t(\theta, 0) = 0$ (see (2.1)), the function $c_t(\theta, u)$ for $u \neq 0$ can be expressed as

$$(5.6) \qquad c_t(\theta, u) = -\Gamma_t(\theta)\Gamma_t^{-1}(\theta + u)\frac{(b_t(\theta, u) - b_t(\theta, 0))u^T}{\|u\|^2}.$$

Suppose now that $\Gamma_t(\theta)$ is continuous w.r.t. θ, so that $\Gamma_t(\theta)\Gamma_t^{-1}(\theta + u) \to 1$. Then, $-\lim_{u \to 0} c_t(\theta, u)$, if it exists, is the total differential $\dot{b}_t(\theta, 0)$ of $b_t(\theta, u)$ at $u = 0$. Let us now examine conditions (L1) and (LL). 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 $(\triangle\Gamma_t(\theta) - c_t(\theta, \Delta_{t-1})) \to 0$. Since $\Delta_{t-1} \to 0$, if t is sufficiently large, we expect $\triangle\Gamma_t(\theta) \approx -\dot{b}_t(\theta, 0)$. If, e.g., $\psi_t(\theta)$ 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 $\dot{b}_t(\theta, 0) = E_\theta\{\dot{\psi}_t(\theta) \mid \mathcal{F}_{t-1}\}$. This implies that, for a given sequence of estimating functions $\psi_t(\theta)$, a natural choice of the normalizing sequence would be

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

On the other hand, if we suppose that the differentiation w.r.t. θ of

$$0 = E_\theta\{\psi_t(\theta) \mid \mathcal{F}_{t-1}\} = \int \psi_t(\theta, z \mid X_1^{t-1}) f_t(\theta, z \mid X_1^{t-1}) \mu(dz)$$

is allowed under the integral sign, then

$$
\begin{aligned}
E_\theta\{\dot{\psi}_t(\theta) \mid \mathcal{F}_{t-1}\} &= \int \dot{\psi}_t(\theta, z \mid X_1^{t-1}) f_t(\theta, z \mid X_1^{t-1}) \mu(dz) \\
&= - \int \psi_t(\theta, z \mid X_1^{t-1}) \dot{f}_t(\theta, z \mid X_1^{t-1}) \mu(dz) \\
(5.7) \qquad &= - \int \psi_t(\theta, z \mid X_1^{t-1}) \dot{l}_t^T(\theta, z \mid X_1^{t-1}) f_t(\theta, z \mid X_1^{t-1}) \mu(dz) \\
&= - E_\theta\{\psi_t(\theta) l_t^T(\theta) \mid \mathcal{F}_{s-1}\}.
\end{aligned}
$$

Therefore, denoting

$$\gamma_t^\psi(\theta) = E_\theta\{\psi_t(\theta) l_t^T(\theta) \mid \mathcal{F}_{t-1}\},$$

another possible choice of the normalizing sequence is

$$\Gamma_t(\theta) = \sum_{s=1}^t \gamma_s^\psi(\theta).$$

Note that for $\psi_t(\theta) = l_t(\theta)$ (a MLE case), $\gamma_t^\psi(\theta) = i_t(\theta)$ and therefore, the suggested normalizing sequence in this case is the conditional Fisher information $I_t(\theta)$.

6 SPECIAL MODELS AND EXAMPLES

Example 1 The i.i.d. scheme. Consider the classical scheme of independent and identically distributed observations $X_1, X_2, \ldots$, with a common probability density/mass function $f(\theta, x)$, $\theta \in \mathbb{R}^m$. Suppose that $\psi(\theta, z)$ is an estimating function, i.e

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

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

$$(6.1) \qquad \hat{\theta}_t = \hat{\theta}_{t-1} + \frac{1}{t} \gamma^{-1}(\hat{\theta}_{t-1}) \psi(\hat{\theta}_{t-1}, X_t), \qquad t \geq 1.$$

where $\gamma(\theta)$ is a non-random matrix such that $\gamma^{-1}(\theta)$ exists for any $\theta \in \mathbb{R}^m$.

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 $0 < \varepsilon < 1$,

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

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

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

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) for some $\varepsilon > 0$,

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

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

Corollary 6.1 *Suppose that for any $\theta \in \mathbb{R}^m$ conditions **(I)** - **(V)** are satisfied. Then the estimator $\hat{\theta}_t$ is strongly consistent and*

$$t^\delta(\hat{\theta}_t - \theta) \to 0 \quad (P^\theta - a.s.)$$

for any $0 < \delta < 1/2$. Furthermore, $\hat{\theta}_t$ is asymptotically normal with parameters $(0, \ \gamma^{-1}(\theta)j(\theta, 0)\gamma^{-1}(\theta))$, that is,

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

Proof. Since $\Gamma_t(\theta) = t\gamma(\theta)$ and

$$b_t(\theta, u) = b(\theta, u) = \int \psi(\theta + u, z) f(\theta, z) \mu(dz),$$

it is easy to see that (I) and (II) imply (C1), (C2) and (C3) from Theorem 3.1 which yields $(\hat{\theta}_t - \theta) \to 0$.

Then, (II) implies (B2) from Corollary 4.2. Condition (V) implies that (B1) in Corollary 4.2 holds with $C_\theta = \mathbf{1}$.

Let us check that conditions of Theorem 5.1 are also satisfied with $A_t = \sqrt{t}\mathbf{1}$. Condition (E) trivially holds. From Proposition 5.1, to check (L1), it is sufficient to show that

(6.2)
$$\frac{1}{t} \sum_{s=1}^{t} [\gamma(\theta) - c(\theta, \Delta_{s-1})] \sqrt{s} \Delta_{s-1} \to 0,$$

where, by (5.1),

$$c_t(\theta, u) = c(\theta, u) = -\gamma(\theta)\gamma^{-1}(\theta + u) \int \psi(\theta + u, z) f(\theta, z) \mu(dz) \, u^T / \|u\|^2$$

for $u \neq 0$ and $c(\theta, 0) = \gamma(\theta)$. Condition (V) yields

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

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, since $1/(2(1+\varepsilon)) < 1/2$, converges to zero. Therefore, (6.2) is now a consequence of the Toeplits Lemma.

For the process $\mathcal{E}_s(\theta)$ from (L2) we have

$$\|\mathcal{E}_s(\theta)\|^2 = \|\gamma(\theta)\gamma^{-1}(\theta + \Delta_{s-1})(\psi(\theta + \Delta_{s-1}, X_s) - b(\theta, \Delta_{s-1})) - \psi(\theta, X_s)\|^2$$

$$\leq 2\|\gamma(\theta)\gamma^{-1}(\theta + \Delta_{s-1})\psi(\theta + \Delta_{s-1}, X_s) - \psi(\theta, X_s)\|^2 + 2\|\gamma(\theta)\gamma^{-1}(\theta + \Delta_{s-1})b(\theta, \Delta_{s-1})\|^2$$

From (III) and (V) we obtain that $(\gamma(\theta)\gamma^{-1}(\theta + \Delta_{s-1}) - \mathbf{1}) \to 0$ and $b(\theta, \Delta_{s-1}) \to 0$ as $s \to \infty$. Therefore, (IV) implies that

$$E_\theta\left\{\|\mathcal{E}_s(\theta)\|^2 \mid \mathcal{F}_{s-1}\right\} \to 0.$$

Since

$$\sum_{s=1}^{t} E_\theta\left\{\|A_t^{-1}(\theta)\mathcal{E}_s(\theta)\|^2 \mid \mathcal{F}_{s-1}\right\} = \frac{1}{t}\sum_{s=1}^{t} E_\theta\left\{\|\mathcal{E}_s(\theta)\|^2 \mid \mathcal{F}_{s-1}\right\},$$

21

(L2) follows from the Toeplitz lemma.

Therefore, the conditions of Theorem 5.1 hold for $A_t(\theta) = \sqrt{t}$. This implies that $\hat{\theta}_t$ is asymptotically linear. The asymptotic normality obviously follows from the central limit theorem for i.i.d. random variables. $\diamondsuit$

Similar results (for i.i.d. schemes) were obtained by Khas'minskii and Nevelson (1972) (when $\psi(\theta, x) = l(\theta, x)$ and $\gamma(\theta) = i(\theta)$, Ch.8, §4) and Fabian (1978).

Note also that conditions (I) and (II) are derived from Theorem 3.1 and are sufficient conditions for the convergence of (6.1). Applying Theorem 3.2 to ψ and Γ, one can obtain various alternative sufficient conditions analogous to those given in Fabian (1978).

Example 2 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 m-dimensional time homogeneous Markov chain with the one-step transition density (with respect to some dominating measure on $\mathbb{R}^m$)

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

where $m(y, x)$ is a m-dimensional vector and $\beta(\theta; x)$ is one dimensional. It is assumed that the canonical parameter space Θ does not depend on on x. Suppose we observe the Markov chain X at times $1, 2, \ldots, t$. Then in our notation $f_t(\theta) = f(X_t; \theta, X_{t-1})$ and

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

where $\dot{\beta}(\theta; X_{t-1})$, as before denotes the m-dimensional raw-vector of partial derivatives of $\beta(\theta; X_{t-1})$ with respect to the coordinates of θ. 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

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

So, a MLE type recursive procedure can be defined as

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

Let us find the functions appearing in the conditions of our theorems for the case $\psi_t = l_t$ and $\Gamma_t = I_t$. Since $E_\theta \{l_t(\theta) \mid \mathcal{F}_{t-1}\} = 0$ we have

$$E_\theta \{m(X_t, X_{t-1}) \mid \mathcal{F}_{t-1}\} = \dot{\beta}^T(\theta; X_{t-1})$$

and also,

$$\ddot{\beta}(\theta; X_{t-1}) = i_t(\theta) = E_\theta \left\{l_t(\theta)l_t^T(\theta) \mid \mathcal{F}_{t-1}\right\}$$
$$= E_\theta \left\{m(X_t, X_{t-1})m^T(X_t, X_{t-1}) \mid \mathcal{F}_{t-1}\right\} - \dot{\beta}^T(\theta; X_{t-1})\dot{\beta}(\theta; X_{t-1}),$$

which implies that

$$E_\theta \left\{m(X_t, X_{t-1})m^T(X_t, X_{t-1}) \mid \mathcal{F}_{t-1}\right\} = \ddot{\beta}(\theta; X_{t-1}) + \dot{\beta}^T(\theta; X_{t-1})\dot{\beta}(\theta; X_{t-1}).$$

Now, it is a simple matter to check that

$$b_t(\theta, u) = E_\theta \{l_t(\theta + u) \mid \mathcal{F}_{t-1}\} = \dot{\beta}^T(\theta; X_{t-1}) - \dot{\beta}^T(\theta + u; X_{t-1}),$$

and

$$E_\theta \left\{\|l_t(\theta + u)\|^2 \mid \mathcal{F}_{t-1}\right\} = \text{trace}\left(\ddot{\beta}(\theta; X_{t-1})\right) + \|\dot{\beta}^T(\theta; X_{t-1}) - \dot{\beta}^T(\theta + u; X_{t-1})\|^2.$$

Using these expressions one can check conditions of the relevant theorems for different choices of functions m and β.

A particular example of a conditional exponential family is the Gaussian autoregressive model defined by

$$X_t = \theta X_{t-1} + Z_t, \qquad t = 1, 2, \ldots,$$

where $\theta \in \mathbb{R}$, $X_0 = 0$ and Z_t's are independent random variables with the standard normal distribution. In this model $m(y, x) = xy$ and $\beta(\theta, x) = \frac{1}{2}x^2\theta^2$. Since $\dot{\beta}(\theta, x) = x^2\theta$ and $\ddot{\beta}(\theta, x) = x^2$,

$$l_t(\theta) = X_t X_{t-1} - X_{t-1}^2\theta, \qquad I_t = I_t(\theta) = \sum_{s=1}^{t} X_{s-1}^2,$$

$$b_t(\theta, u) = -X_{t-1}^2 u, \qquad E_\theta \left\{l_t^2(\theta + u) \mid \mathcal{F}_{t-1}\right\} = X_{t-1}^2 + X_{t-1}^4 u^2.$$

The recursive procedure in this case is

$$(6.3) \qquad \hat{\theta}_t = \hat{\theta}_{t-1} + \frac{1}{I_t}\left(X_t X_{t-1} - X_{t-1}^2\hat{\theta}_{t-1}\right)$$
$$I_t = I_{t-1} + X_{t-1}^2.$$

Note that the rate of the conditional Fisher information I_t varies for the different values of θ. Suppose

$$(6.4) \qquad a_t(\theta) = \begin{cases} t(1-\theta^2)^{-1} & \text{for } |\theta| < 1 \\ \frac{1}{2}t^2 & \text{for } |\theta| = 1 \\ \theta^{2t}(\theta^2-1)^{-2} & \text{for } |\theta| > 1. \end{cases}$$

For $|\theta| < 1$ the process is stable and $I_t(\theta)/a_t(\theta) \to 1$ in probability as $t \to \infty$, whereas $I_t(\theta)/a_t(\theta) \to W \sim \chi^2(1)$ almost surely in the unstable case $|\theta| > 1$ (non-ergodic case). In the critical case, $|\theta| = 1$, the ratio $I_t(\theta)/a_t(\theta)$ converges in distribution, but not in probability (for details, see White (1958) and Anderson (1959)). It is also well known that $I_t \to \infty$ almost surely for any $\theta \in \mathbb{R}$ (see, e.g, Shiryayev (1984), Ch.VII, 5.5). Note that if d_n is a sequence of positive numbers, then $d_n \to \infty$ implies that $\sum_{k=1}^{\infty} \triangle d_k/d_k = \infty$ and $\sum_{n=1}^{\infty} \triangle d_n/d_n^2 < \infty$ (where, as before, $\triangle d_n = d_n - d_{n-1}$). Therefore, almost surely,

$$\sum_{t=1}^{\infty} \frac{X_{t-1}^2}{I_t} = \infty \qquad \text{and} \qquad \sum_{t=1}^{\infty} \frac{X_{t-1}^2}{I_t^2} < \infty.$$

Let us check that the conditions of Theorem 3.2 hold for for $V_\theta(u) = u^2$. We have

$$N_t(u) = -2u^2\frac{X_{t-1}^2}{I_t} + u^2\frac{X_{t-1}^4}{I_t^2} + \frac{X_{t-1}^2}{I_t^2} = -u^2\frac{X_{t-1}^2}{I_t} - \frac{X_{t-1}^2}{I_t^2}\left(u^2 I_{t-1} - 1\right)$$

To check (G2) note that since the infimum is taken over $\varepsilon \leq u^2 \leq 1/\varepsilon$, the value of $(u^2 I_{t-1} - 1)$ is positive eventually and therefore, there exists an almost surely finite r.v. ξ such that

$$\sum_{t=1}^{\infty} \inf_{\varepsilon \leq u^2 \leq 1/\varepsilon} [\mathcal{N}_t(u)]^- \geq \xi + \varepsilon^2 \sum_{t=1}^{\infty} \frac{X_{t-1}^2}{I_t} = \infty.$$

To check (G2) we rewrite

$$N_t(u) = -u^2\left(\frac{X_{t-1}^2}{I_t} + \frac{X_{t-1}^2}{I_t^2}I_{t-1}\right) + \frac{X_{t-1}^2}{I_t^2}$$

and

$$\sum_{t=1}^{\infty}(1+\Delta_{t-1}^2)^{-1}[\mathcal{N}_t(\Delta_{t-1})]^+ \leq \sum_{t=1}^{\infty}[\mathcal{N}_t(\Delta_{t-1})]^+ \leq \sum_{t=1}^{\infty}\frac{X_{t-1}^2}{I_t^2} < \infty.$$

Therefore, we conclude for any $\theta \in \mathbb{R}$, the recursive estimator $\hat{\theta}_t$ is strongly consistent for any choice of the initial $\hat{\theta}_0$. This result, of course, could have

been obtained directly for the linearity of the model allows to solve the recursion equation analytically. Indeed, in this case,

$$\hat{\theta}_t = \frac{1}{I_t}\left(\hat{\theta}_0 + \sum_{s=1}^{t} X_s X_{s-1}\right)$$

which, in the case of $\hat{\theta}_0 = 0$ coincides with the MLE. Since $\hat{\theta}_t$ is obviously asymptotically linear, we expect the conditions of Theorem 5.1 to hold without any additional assumptions. This is indeed the case. Condition (E) is satisfied with $A_t(\theta) = \sqrt{a_t(\theta)}$ where $a_t(\theta)$ is defined in (6.4). Conditions (L1) and (L2) trivially hold since in this case $(c_s(\theta, \Delta_{s-1}) - \triangle\Gamma_s(\theta)) = 0$ and $\mathcal{E}_s(\theta) = 0$ for any s.

Example 3 AR(m) process Consider an AR(m) process

$$X_i = \theta_1 X_{i-1} + \cdots + \theta_m X_{i-m} + \xi_i = \theta^T X_{i-m}^{i-1} + \xi_i,$$

where $X_{i-m}^{i-1} = (X_{i-1}, \ldots, X_{i-m})^T$, $\theta = (\theta_1, \ldots, \theta_m)^T$ and ξ_i is a sequence of i.i.d. random variables.

A reasonable class of procedures in this model should have a form

(6.5) $$\hat{\theta}_t = \hat{\theta}_{t-1} - \Gamma_t^{-1}(\hat{\theta}_{t-1})\psi_t(X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1}),$$

where $\psi_t(z)$ and $\Gamma_t^{-1}(z)$ $(z \in \mathbb{R}^m)$ are respectively vector and matrix processes meeting conditions of the previous sections. Suppose that the probability density function of ξ_t w.r.t. Lebesgue's measure is $g(x)$. Then, if

$$\psi_t(z) = -\frac{\dot{g}^T(z)}{g(z)}X_{t-m}^{t-1}$$

then $\psi_t(X_t - \hat{\theta}_{t-1}X_{t-1})$ is a score vector and (6.5) becomes the likelihood recursive procedure. A possible choice of $\Gamma_t(z)$ in this case would be the conditional Fisher information matrix

$$I_t = i^g \sum_{s=1}^{t} X_{t-m}^{t-1}(X_{t-m}^{t-1})^T$$

where

$$i^g = \int \left(\frac{\dot{g}^T(z)}{g(z)}\right)^2 g(z)\,dz.$$

An interesting class of recursive estimators for strongly stationary AR(m) processes is studied in Campbell (1982). These estimators are recursive versions of robust modifications of the least squares method and are defined as

(6.6) $$\hat{\theta}_t = \hat{\theta}_{t-1} - a_t \gamma(X_{t-m}^{t-1}) \phi(X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1}),$$

where a_t is a sequence of a positive numbers with $a_n \to 0$, ϕ is a bounded scalar function and $\gamma(u)$ is a vector function of the form $uh(u)$ for some non-negative function h of u (See also Leonov (1988)). The class of procedures of type (6.6) is clearly a subclass of that defined by (6.5) and therefore can be studies using the results of the previous sections.

Suppose that ξ_i are i.i.d. random variables with a bell-shaped, symmetric about zero probability density function $g(z)$ (that is, $g(-z) = g(z)$, and $g \downarrow 0$ on $\mathbb{R}_+$). Suppose also that $\phi(x)$ is an odd, continuous in zero function. Let us write conditions of Theorem 3.1 for

$$\Gamma(\theta) = a_t \mathbf{1} \qquad \text{and} \qquad \psi_t(\theta) = X_{t-m}^{t-1} h\left(X_{t-m}^{t-1}\right) \phi\left(\xi_t - \theta^T X_{t-m}^{t-1}\right).$$

We have

$$E_\theta \left\{ \phi\left(X_t - (\theta + u)^T X_{t-m}^{t-1}\right) \mid \mathcal{F}_{s-1} \right\} = E_\theta \left\{ \phi\left(\xi_t - u^T X_{t-m}^{t-1}\right) \mid \mathcal{F}_{s-1} \right\}$$

$$= \int \phi\left(z - u^T X_{t-m}^{t-1}\right) g(z) dz.$$

It follows from Lemma A2 in Appendix that if $w \neq 0$,

$$G(w) = -w \int_{-\infty}^{\infty} \phi\left(z - w\right) g(z) dz > 0.$$

Therefore,

$$u^T \Gamma_t^{-1}(\theta + u) b_t(\theta, u) = a_t u^T X_{t-m}^{t-1} h\left(X_{t-m}^{t-1}\right) E_\theta \left\{ \phi\left(\xi_t - u^T X_{t-m}^{t-1}\right) \mid \mathcal{F}_{s-1} \right\}$$

$$= -a_t \, h\left(X_{t-m}^{t-1}\right) G(u^T X_{t-m}^{t-1}) \leq 0.$$

Also, since ϕ is a bounded function,

$$E_\theta \left\{ \|\Gamma_t^{-1}(\theta + u) \psi_t(\theta + u)\|^2 \mid \mathcal{F}_{t-1} \right\} \leq C^\theta a_t^2 \|X_{t-m}^{t-1}\|^2 h^2(X_{t-m}^{t-1})$$

for some positive constant C^θ. Therefore, conditions of Theorem 3.1 hold if

(6.7) $$\sum_{t=1}^{\infty} a_t h\left(X_{t-m}^{t-1}\right) \inf_{\varepsilon \leq \|u\| \leq 1/\varepsilon} G(u^T X_{t-m}^{t-1}) = \infty$$

and

$$(6.8) \qquad \sum_{t=1}^{\infty} a_t^2 \|X_{t-m}^{t-1}\|^2 h^2(X_{t-m}^{t-1}) < \infty.$$

If X_t is strongly stationary process, these conditions can be verified using limit theorems for strongly stationary processes. Suppose, e.g., that $a_t = 1/t$ and $\|\gamma(X_{t-m}^{t-1})\|^2 = \|X_{t-m}^{t-1}\|^2 h^2(X_{s-m}^{s-1})$ is integrable. Then it follows that $h\left(X_{t-m}^{t-1}\right) \inf_{\varepsilon \leq \|u\| \leq 1/\varepsilon} G(u^T X_{t-m}^{t-1})$ is integrable as well and if $h(\mathbf{x}) \neq 0$ for any $\mathbf{x} \neq 0$ then $h(\mathbf{x}) \inf_{\varepsilon \leq \|u\| \leq 1/\varepsilon} G(u^T \mathbf{x}) > 0$ for any $\mathbf{x} \neq 0$ (see Appendix, Lemma 2). Therefore, it follows from an ergodic theorem for strongly stationary processes that

$$\lim_{t \to \infty} \frac{1}{t} \sum_{s=1}^{t} h\left(X_{t-m}^{t-1}\right) \inf_{\varepsilon \leq \|u\| \leq 1/\varepsilon} G(u^T X_{s-m}^{s-1}) > 0$$

and

$$\frac{1}{t} \sum_{s=1}^{t} \|X_{t-m}^{t-1}\|^2 h^2(X_{s-m}^{s-1})$$

converges to a finite limit. Now, (C2) in Theorem 3.1 follows from the observation that for any nonnegative sequence c_t, the convergence $\frac{1}{t} \sum_{s=1}^{t} c_s \to c > 0$ implies $\sum_{t=1}^{\infty} c_t/t = \infty$. The convergence $\sum_{t=1}^{\infty} \gamma^2(X_{t-m}^{t-1})/t^2 < \infty$ follows since for any nonnegative sequence c_t, the convergence $\frac{1}{t} \sum_{i=1}^{t} c_i \to c$ implies $\sum_{t=1}^{\infty} c_t/t^2 < \infty$ (this can be easily verified using (5.5) with $C_s = \sum_{i=1}^{t} c_i$ and $D_t = 1/t^2$).

Examples of the procedures of type (6.6) as well as some simulation results are presented in Campbell (1982).

Sometimes the conditions of the theorems presented here are difficult to verify. The next example shows that even when this is the case, the results of the paper can be useful to construct recursive analogue of the estimators given by estimating equations. Consider for instance a robust generalized M-estimator of the parameter of an AR(1) process proposed by Denby and Martin (1979), which is a solutions of the equation

$$\sum_{s=1}^{t} c_x s_x \phi_H \left(\frac{X_{s-1}}{c_x s_x} \right) c_r s_r \phi_H \left(\frac{X_s - \theta X_{s-1}}{c_r s_r} \right) = 0$$

where

$$\phi_H(x) = \begin{cases} x, & \text{if } |x| \leq 1 \\ \text{sign}(x) & \text{if } |x| > 1 \end{cases}$$

27

is the Huber function and s_x, s_r are scale estimates. A recursive analogue of this estimator is a sequence defined by

$$(6.9) \qquad \hat{\theta}_t = \hat{\theta}_{t-1} - \Gamma_t^{-1} c_x s_x \phi_H \left(\frac{X_{t-1}}{c_x s_x} \right) c_r s_r \phi_H \left(\frac{X_t - \hat{\theta}_{t-1} X_{t-1}}{c_r s_r} \right),$$

where Γ_t is a random process with the increments $\Delta \Gamma_t = \Gamma_t - \Gamma_{t-1}$ defined by

$$(6.10) \qquad \Delta \Gamma_t = \begin{cases} c_x s_x \phi_H^2 \left(\frac{X_{t-1}}{c_x s_x} \right) & \text{if} \quad \left| \frac{X_t - \hat{\theta}_{t-1} X_{t-1}}{c_r s_r} \right| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and c_x, c_r are tuning constants. The form of Γ_t is suggested by the theorems presented in Sections 4 and 5 (see Remark 5.2 (iii)). Note that the direct application of Theorem 5.1 suggests using a normalising sequence with the increments defined in (6.10) but $c_x s_x \phi_H^2$ replaced by $c_x s_x \phi_H X_{t-1}$. Nevertheless, from the considerations of robustness we have preferred to truncate X_{t-1}, especially as, as simulation results showed, the truncated version works better.

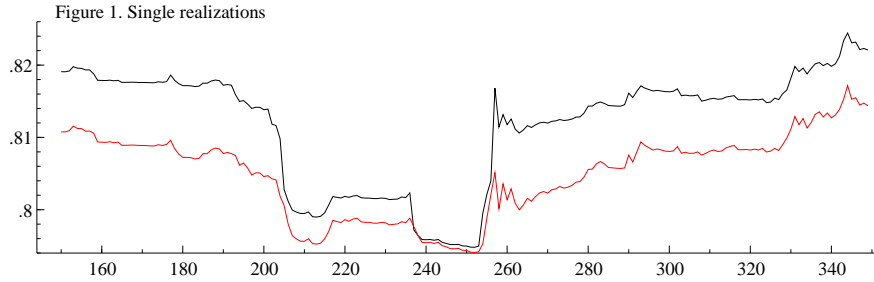Note also that (6.9) is clearly a procedure of type (6.6) with $m = 1$ and

$$\phi(x) = c_r s_r \phi_H \left(\right), \qquad \gamma(x) = c_x s_x \phi_H (x) = c_x s_x x h(x),$$

$$h(x) = \begin{cases} 1, & \text{if } |x| \leq 1 \\ 1/|x| & \text{if } |x| > 1 \end{cases}$$

but with a stochastic normalising sequence $a_t = \Gamma_t^{-1}$. It is easy to see that, as far as a_t is a predictable process, conditions (6.7) and (6.8) remain sufficient for the conditions of Theorem 3.1. For the recursion (6.9), condition (6.7) can be easily verified, but the difficulties arise with (6.8) since it requires the convergence $\sum_{t=1}^{\infty} \Gamma_t^{-2}$.

Further research is required to study behaviour of the procedures of type (6.9). Below we present a brief simulation study to compare the performance of the recursive procedure (6.9) to that of the least squares method which is equivalent to the procedure defined by (6.3). The time series were generated from the additive effect outliers (AO) model:

$$Y_t = \theta Y_{t-1} + \varepsilon_t$$
$$X_t = Y_t + v_t,$$

28

Figure 1. Single realizations

Figure 2. Averages over 30 replications

where innovations ε_t are i.i.d. Gaussian $N(0,1)$ and v_t are also i.i.d. with a Gaussian mixture distribution $(1-\delta)N(0,1) + \delta N(0,\sigma^2)$. The figures below show the performances of the estimators for $\theta = 0.8$, $\delta = 0.1$ and $\sigma = 10$. The estimators are computed for the series of length 300, with the additional 20 observations at the beginning on which initial estimates are based (as an estimates for s_x and s_r we take the median of the absolute values of the data and residuals respectively, divided by 0.6745). The tuning constants in (6.9) help to damp excess oscillation of the values of $\hat{\theta}_t$ and are surprisingly easy to adjust (in the simulations below their values vary from 0.5 to 4). Figure 1 shows single realizations of $\hat{\theta}_t$ and $\hat{\theta}_t^{LS}$ ($t = 125, \ldots, 300$), derived by (6.9) and (6.3) respectively. Figure 2 presents averages of $\hat{\theta}_t$ and $\hat{\theta}_t^{LS}$ ($t = 125, \ldots, 300$), over 30 replications and finally, Figure 3 shows the mean squared errors (over 30 replications) of $\hat{\theta}_t$ and $\hat{\theta}_t^{LS}$ ($t = 125, \ldots, 300$).

In all three figures the black and red lines correspond to $\hat{\theta}_t^{LS}$ and $\hat{\theta}_t$ respectively.

Figure 3. Mean squared errors
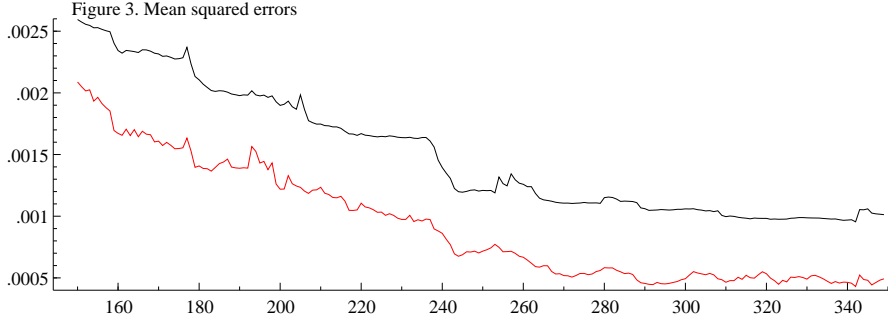
APPENDIX

Lemma A1(Robbins and Siegmund) *Let $\mathcal{F}_0, \mathcal{F}_1, \ldots$ be a non-decreasing sequence of σ-algebras and $X_n, \beta_n, \xi_n, \zeta_n \in \mathcal{F}_n, \quad n \geq 0$, are nonnegative r.v.'s such that*

$$E(X_n|\mathcal{F}_{n-1}) \leq X_{n-1}(1 + \beta_{n-1}) + \xi_{n-1} - \zeta_{n-1}, \quad n \geq 1$$

eventually. Then

$$\{\sum_{i=1}^{\infty} \xi_{i-1} < \infty\} \cap \{\sum_{i=1}^{\infty} \beta_{i-1} < \infty\} \subseteq \{X \to\} \cap \{\sum_{i=1}^{\infty} \zeta_{i-1} < \infty\} \quad (P\text{-}a.s.).$$

Remark Proof can be found in Robbins and Siegmund (1971). Note also that this lemma is a special case of the theorem on the convergence sets nonnegative semimartingales (see, e.g., Lazrieva, Sharia, and Toronjadze (1997)).

Lemma A2 *Suppose that $g \not\equiv 0$ is a nonnegative even function on $\mathbb{R}$ and $g \downarrow 0$ on $\mathbb{R}_+$. Suppose also that ϕ is a measurable odd function on $\mathbb{R}$ such that $\phi(z) > 0$ for $z > 0$ and $\int_{\mathbb{R}} |\phi(z-w)|g(z)dz < \infty$ for all $w \in \mathbb{R}$. Then for any $w \neq 0$,*

$$(A1) \qquad w \int_{-\infty}^{\infty} \phi(z-w)\, g(z)dz < 0.$$

Furthermore, if $g(z)$ is continuous, then for any $\varepsilon \in (0,1)$

$$(A2) \qquad \sup_{\varepsilon \leq |w| \leq 1/\varepsilon} w \int_{-\infty}^{\infty} \phi(z-w)\, g(z)dz < 0.$$

Proof Denote

$$(A3) \qquad \Phi(w) = \int_{-\infty}^{\infty} \phi(z-w)\, g(z)dz.$$

30

Using the change of variable $z \rightleftarrows -z$ in the integral over $(-\infty, 0)$ and the equalities $\phi(-z) = -\phi(z)$ and $g(-z+w) = g(z-w)$, we obtain

$$
\begin{aligned}
\Phi(w) &= \int_{-\infty}^{\infty} \phi(z) g(z+w) dz \\
&= \int_{-\infty}^{0} \phi(z) g(z+w) dz + \int_{0}^{\infty} \phi(z) g(z+w) dz \\
&= \int_{0}^{\infty} \phi(z) \left(g(z+w) - g(-z+w) \right) dz \\
&= \int_{0}^{\infty} \phi(z) \left(g(z+w) - g(z-w) \right) dz.
\end{aligned}
$$

Suppose now that $w > 0$. Then $z - w$ is closer to 0 than $z + w$, and the properties of g imply that $g(z+w) - g(z-w) \leq 0$. Since $\phi(z) > 0$ for $z > 0$, $\Phi(w) \leq 0$. The equality $\Phi(w) = 0$ would imply that $g(z+w) - g(z-w) = 0$ for *all* $z \in (0, +\infty)$ since, being monotone, g has right and left limits at each point of $(0, +\infty)$. The last equality, however, contradicts the restrictions on g. Therefore, (A1) holds true. Similarly, if $w < 0$, then $z + w$ is closer to 0 than $z-w$, and $g(z+w) - g(z-w) \geq 0$. Hence $w\left(g(z+w) - g(z-w) \right) \leq 0$, which yields (A1) as before.

To prove (A2) note that the continuity of g implies that $g(z+w) - g(z-w)$ is a continuous functions of w and (A2) will follow from (A1) if one proves that $\Phi(w)$ is also continuous in w. So, it is sufficient to show that the integral in (A3) is uniformly convergent for $\varepsilon \leq |w| \leq 1/\varepsilon$. It follows from the restrictions we have placed on g that there exists $\delta > 0$ such that $g \geq \delta$ in a neighbourhood of 0. Then the condition

$$
\int_{0}^{\infty} \phi(z) \left(g(z+w) + g(z-w) \right) dz = \int_{\mathbb{R}} |\phi(z-w)| g(z) dz < \infty, \quad \forall w \in \mathbb{R}
$$

implies that ϕ is locally integrable on $\mathbb{R}$. It is easy to see that, for any $\varepsilon \in (0, 1)$,

$$
g(z \pm w) \leq g(0) \chi_\varepsilon(z) + g(z - 1/\varepsilon), \quad z \geq 0, \quad \varepsilon \leq |w| \leq 1/\varepsilon,
$$

where χ_ε is the indicator function of the interval $[0, 1/\varepsilon]$. Since the function $\phi(\cdot) \left(g(0) \chi_\varepsilon + g(\cdot - 1/\varepsilon) \right)$ is integrable on $(0, +\infty)$ and does not depend on w, we conclude that the integral in (A3) is indeed uniformly convergent for $\varepsilon \leq |w| \leq 1/\varepsilon$. $\diamond$

REFERENCES

31

ANDERSON, T.W. (1959). On asymptotic distributions of estimates of parameters of stochastic difference equations. *Ann. Math. Statist.* **30**, 676–687.

BARNDORFF-NIELSEN, O.E. (1988). *Parametric Statistical Models and Likelihood.* Springer Lecture Notes in Statistics 50. Heidelberg, Springer.

BARNDORFF-NIELSEN, O.E. and 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. and SCOTT, D.J. (1983). *Asymptotic Optimal Inference for Non-ergodic Models.* Springer-Verlag, New York.

CAMPBELL, K. (1982). Recursive computation of M-estimates for the parameters of a finite autoregressive process. *Ann. Statist.* **10**, 442-453.

DENBY, L. and MARTIN, R.D. (1979). Robust estimation of the first order autoregressive parameter. *J. Amer. Statist. Assoc.* **74**, 140-146.

ENGLUND, J.-E., HOLST, U., and RUPPERT, D. (1989) Recursive estimators for stationary, strong mixing processes – a representation theorem and asymptotic distributions *Stochastic Processes Appl.* **31**, 203–222.

FABIAN, V. (1978). On asymptotically efficient recursive estimation, *Ann. Statist.* **6**, 854-867.

FEIGIN, P.D. (1981). conditional exponential families and a representation theorem for asymptotic inference. *Ann. Statist.* **9**, 597-603.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., and STAHEL, W. (1986). *Robust Statistics - The Approach Based on Influence Functions.* Wiley, New York

HUBER, P.J. (1981). *Robust Statistics.* Wiley, New York.

JUREČKOVÁ, J. and SEN, P.K. (1996). *Robust Statistical Procedures - Asymptotics and Interrelations.* Wiley, New York.

KHAS'MINSKII, R.Z. and NEVELSON, M.B. (1972). *Stochastic Approximation and Recursive Estimation.* Nauka, Moscow.

LAUNER, R.L. and WILKINSON, G.N. (1979). *Robustness in Statistics.* Academic Press, New York.

LAZRIEVA, N., SHARIA, T. and TORONJADZE, T. (1997). The Robbins-Monro type stochastic differential equations. I. Convergence of solutions. *Stochastics and Stochastic Reports* **61**, 67–87.

LAZRIEVA, N., SHARIA, T. and TORONJADZE, T. (2003). The Robbins-Monro type stochastic differential equations. II. Asymptotic behaviour of solutions. *Stochastics and Stochastic Reports* (in print).

LAZRIEVA, N. and TORONJADZE, T. (1987). Ito-Ventzel's formula for semimartingales, asymptotic properties of MLE and recursive estimation. *Lect. Notes in Control and Inform. Sciences, 96, Stochast. diff. systems, H.J, Engelbert, W. Schmidt (Eds.), Springer* 346–355.

LEHMAN, E.L. (1983). *Theory of Point Estimation.* Wiley, New York.

LEONOV, S.L. (1988). On recurrent estimation of autoregression parameters, *Avtomatika i Telemekhanika* **5**, 105-116.

LIPTSER, R.S. and SHIRYAYEV, A.N. (1989). *Theory of Martingales.* Kluwer, Dordrecht.

LJUNG, L. PFLUG, G. and WALK, H. (1992). *Stochastic Approximation and Optimization of Random Systems.* Birkhäuser, Basel.

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

RIEDER, H. (1994). *Robust Asymptotic Statistics.* Springer–Verlag, New York.

ROBBINS, H. MONRO, S. (1951) A stochastic approximation method, *Ann. Statist.* **22**, 400–407.

ROBBINS, SIEGMUND, H.D. (1971) A convergence theorem for nonnegative almost supermartingales and some applications, *Optimizing Methods in Statistics*, ed. J.S. Rustagi Academic Press, New York. 233–257.

SERFLING, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

SHARIA, T. (1992). On the recursive parameter estimation for general statistical model in discrete time. Bulletin of the Georgian Acad. Scienc. 145 **3**, 465–468.

SHARIA, T. (1998). On the recursive parameter estimation for the general discrete time statistical model. *Stochastic Processes Appl.* **73**, **2**, 151–172.

SHARIA, T. (1997). Truncated recursive estimation procedures. *Proc. A. Razmadze Math. Inst.* **115**, 149–159.

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

TITTERINGTON, D.M., SMITH, A.F.M. and MAKOV, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, New York.

WHITE, J.S. (1958). The limiting distribution of the serial correlation coefficient in the explosive case. *Ann. Math. Stat.* **29**, 1188–1197.