Bernoulli **3**(3), 1997, 247–277

Efficient estimation in nonlinear autoregressive time-series models

HIRA L. KOUL[1] and ANTON SCHICK[2*]

[1]*Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824-1027, USA*
[2]*Department of Mathematical Sciences, State University of New York at Binghamton, Binghamton, NY 13902-6000, USA*

This paper discusses efficient estimation for a class of nonlinear time-series models with unknown error densities. It establishes local asymptotic normality in this semi-parametric setting. This is then used to describe efficient estimates and to discuss the question of adaptation. Stein's necessary condition for adaptive estimation is satisfied if the error densities are symmetric, but is also satisfied in some models with asymmetric error densities. The paper gives several methods of constructing efficient estimates. These results are then applied to construct efficient estimators in SETAR(2; 1, 1), EXPAR(1) and ARMA(1, 1) models. We observe that adaptation is not possible in the SETAR(2; 1, 1) model with asymmetric errors while the efficient estimators in the ARMA(1, 1) model are adaptive even for asymmetric error densities. Section 8 contains a result that is useful in verifying the continuity of the stationary density with respect to the underlying parameters.

Keywords: adaptivity; ergodicity; local asymptotic normality; semi-parametric time series; stationarity

1. Introduction

The construction of estimators that are asymptotically efficient in the presence of infinite-dimensional nuisance parameters has been the focus of numerous researchers in the last three decades – see, for example, the recent monograph by Bickel *et al*. (1993) and the references therein. The present paper is concerned with the construction of such estimators in a class of nonlinear time-series models.

To describe these models, let $\mathbb{R}$ denote the set of real numbers, and let m and p be positive integers. Let $\mathscr{F}$ be a class of Lebesgue densities, Θ be an open subset of $\mathbb{R}^m$, $\mathfrak{P} = \{P_{\vartheta,\phi}: (\vartheta, \phi) \in \Theta \times \mathscr{F}\}$ be a family of probability measures, $X_{1-p}, \ldots, X_0, X_1, X_2, \ldots$ be random variables and, for each $j = 1, 2, \ldots$, let h_j be a measurable map from $\mathbb{R}^{p+j-1} \times \Theta$ into $\mathbb{R}$. Let $\mathbf{X}_j = (X_{1-p}, \ldots, X_j)^{\mathrm{T}}$, $j = 0, 1, \ldots$, and

$$H_j(\vartheta) = h_j(\mathbf{X}_{j-1}, \vartheta), \qquad \vartheta \in \Theta, j = 1, 2, \ldots .$$

*To whom correspondence should be addressed. e-mail: anton@math.binghamton.edu

The time series $\{X_j: j \geq 1-p\}$ is assumed to have the following structure: under each $P_{\vartheta,\phi} \in \mathfrak{P}$, the random vector $\mathbf{X}_0$ has a Lebesgue density $g_{\vartheta,\phi}$, and the random variables

$$\varepsilon_j(\vartheta) = X_j - H_j(\vartheta), \qquad j = 1, 2, \ldots, \tag{1.1}$$

are independent with common density ϕ and independent of $\mathbf{X}_0$.

By selecting appropriate functions $\langle h_j \rangle$ one can obtain various models studied in the time-series literature such as the well-known ARMA models and the class of nonlinear autoregression (NLAR(p)) models, where

$$H_j(\vartheta) = h(X_{j-p}, \ldots, X_{j-1}, \vartheta)$$

for some known function h from $\mathbb{R}^p \times \Theta$. Examples of NLAR($p$) models are the SETAR(2; 1, 1) and EXPAR(1) models. The SETAR(2; 1, 1) model is obtained by taking $m = 2$, $p = 1$ and

$$h(x, \vartheta) = \vartheta_1 x I[x \leq 0] + \vartheta_2 x I[x > 0], \tag{1.2}$$

while the EXPAR(1) model is obtained by taking $m = 3$, $p = 1$ and

$$h(x, \vartheta) = (\vartheta_1 + \vartheta_2 \, \mathrm{e}^{-\vartheta_3 x^2})x. \tag{1.3}$$

Tong (1990) discusses these and many other nonlinear time-series models.

Suppose now that the true parameter is (θ, f). The problem of interest is the construction of efficient estimators of θ in the presence of the nuisance parameter f. Such estimates have been constructed for AR and ARMA models in two papers by Kreiss. In fact his estimates are *adaptive*, i.e., they are asymptotically as efficient as in the case of known f. Kreiss (1987a) provides adaptive estimates for parameters in ARMA models when the error densities are assumed to be symmetric; Kreiss (1987b) constructs adaptive estimates of parameters in AR models without this symmetry assumption. Jeganathan (1995) describes an extension of the construction of Kreiss (1987a) to the present models with symmetric errors and addresses various other inference issues for general time-series models. See also Koul and Pflug (1990) for adaptive estimation in explosive autoregression.

After the first draft of this paper, we became aware of the preprint by Drost *et al*. (1994). This preprint deals with adaptive estimation of (a part of) the parameter of interest in more general time-series models than considered here (our model is a subclass of their location–scale model) and shows how a sample splitting technique used in i.i.d. models by Schick (1986) can be used to construct adaptive estimates for these models. The authors then apply their construction to ARMA, TAR and ARCH models. In particular, they show that adaptive estimation of the full parameter of interest is possible in ARMA models. The possibility of adaptive estimation was already observed in an earlier paper (Drost *et al*. 1993) by these authors.

Since adaptive estimation is not always possible, Jeganathan (1995) imposed symmetry conditions and Drost *et al*. (1994) could only estimate the component of the parameter of interest which is adaptively estimable. In contrast, we consider efficient estimation in general. This allows us to estimate the full parameter of interest and gives us the freedom to consider general error models. Our estimates will be automatically adaptive if the necessary condition for adaptation is met.

We give three constructions of efficient estimates. The construction given in Section 4 is similar to that of Drost *et al*. (1994). It uses the sample splitting technique and shows that efficient estimation is possible under minimal assumptions. We feel that sample splitting should be avoided in moderate sample sizes. A small simulation study is included which supports our belief. Therefore we give two more constructions that avoid this technique at the expense of additional assumptions. The construction in Section 5 does so for adaptive estimation in symmetric error models. It adopts the construction from Kreiss (1987a) and fixes an erroneous argument in Jeganathan (1995). Section 6 gives a construction for error models whose densities have zero means and finite variances but are not necessarily symmetric. One encounters such error models in ARMA and NLAR models.

Our asymptotic considerations are based on the local asymptotic normality (LAN) of our models. Various LAN results have been proved in special cases by several authors – see Akritas and Johnson (1980), Swensen (1985), Kreiss (1987a), Hwang and Basawa (1993), Drost *et al*. (1994) and Jeganathan (1995). These results do not allow for a parametrization of the error density, and only Drost *et al*. (1994) and Jeganathan (1995) prove uniformity in the parameter of interest. However, the latter two papers give LAN for other time-series models. In contrast, we prove LAN in both the parameter of interest and the nuisance parameter, with uniformity in the former. This semi-parametric version is needed to characterize efficient estimates and to describe Stein's (1956) necessary condition for adaptive estimation. The uniformity is helpful in the construction of efficient estimates.

One of the assumptions used to obtain LAN is the assumption that the initial distribution has negligible effect. In our case this is guaranteed by the L_1-continuity of the map $(\vartheta, \phi) \mapsto g_{\vartheta,\phi}$ at (θ, f). The verification of this condition in stationary AR and ARMA models is rendered feasible because of the causality of these processes but its verification in general nonlinear time-series models is far from being routine. For this reason, Section 8 provides sufficient conditions for the L_1-continuity in stationary and ergodic NLAR(1) models. Neither Drost *et al*. (1994) nor Jeganathan (1995) address this issue.

Our paper is organized as follows. Section 2 proves LAN for the semi-parametric time-series models considered here. It includes a discussion about the verification of the sufficient conditions for NLAR models. Section 3 addresses the question of efficient estimation of θ. It begins by characterizing asymptotically efficient estimates for general error models. It then discusses Stein's (1956) necessary conditions for adaptive estimation. It is seen that this condition holds if $\mathscr{F}$ contains only symmetric error densities, but may fail otherwise. In particular, it is observed that without the symmetry assumption Stein's necessary condition is not satisfied in the SETAR(2; 1, 1) model, but is satisfied in the ARMA(1, 1) model when the error distributions have zero means and finite variances. The former is a new observation and the latter was already observed by Drost *et al*. (1993).

Section 4 constructs asymptotically efficient estimates using a sample splitting technique. The discussion on this construction is brief because this is the same approach as taken by Drost *et al*. (1994). This section then gives such estimates for SETAR(2; 1, 1), (restricted) EXPAR(1) and MA(1) models.

In Sections 5 and 6 we show that under additional assumptions efficient estimates can be constructed without the sample splitting technique. Section 5 does so for adaptive estimation in symmetric error models. It concludes with a simulation study that shows

superiority of these estimates over the ones based on the sample splitting technique. Section 6 gives a construction avoiding the sample splitting for error models whose densities have zero means and finite variances but are not necessarily symmetric. This construction is used in Section 7 to construct an efficient and adaptive estimate of the parameter of an ARMA(1, 1) model. Our construction differs from those in Kreiss (1987a) and Drost *et al.* (1994) in that it is based on the actual observations, while their constructions require observation of past error variables which are not available. Our proof gives also the argument for the continuity of the stationary densities, an issue omitted by both papers. Thus we give a *complete* argument for the construction of asymptotically efficient estimators in *truly* stationary ARMA models based on the actual observations under minimal assumptions.

Throughout this paper, θ and f are fixed and F denotes the distribution corresponding to f. The expectation under $P_{\vartheta,\phi}$ is denoted by $E_{\vartheta,\phi}$, $(\vartheta, \phi) \in \Theta \times \mathscr{F}$. For convenience, $P_{\vartheta,f}$ and $E_{\vartheta,f}$ are abbreviated by P_ϑ and E_ϑ, respectively. By a *local* sequence we mean a sequence $\langle \theta_n \rangle$ in Θ such that $\sqrt{n}(\theta_n - \theta)$ is bounded. For a local sequence $\langle \theta_n \rangle$ and a sequence $\{a_n\}$ of positive numbers, $o_{\theta_n}(a_n)$ $(O_{\theta_n}(a_n))$ denotes a sequence of random variables $\{\xi_n\}$ such that $a_n^{-1}\xi_n$ converges to 0 (is bounded) in P_{θ_n}-probability. The distribution of a random variable X under a probability measure P is denoted by $\mathfrak{L}(X|P)$. The multivariate normal distribution with mean μ and covariance matrix W will be denoted by $\mathscr{N}(\mu, W)$.

In what follows we shall often work with (submodels of) the error models $\mathscr{F}_0$ and $\mathscr{F}_0^+$, where $\mathscr{F}_0$ is the set of all Lebesgue densities that have zero means, finite variances and finite Fisher information for location, and $\mathscr{F}_0^+$ consists of all positive densities in $\mathscr{F}_0$.

2. Local asymptotic normality

In this section we provide sufficient conditions for the desired LAN of our model and discuss them in NLAR models. From now on we assume the following.

Assumption 2.1. *The density f has finite Fisher information for location, i.e., f is absolutely continuous with a.e.-derivative f' and*

$$J = \int \ell^2 \, dF < \infty, \qquad \text{where } \ell = -\frac{f'}{f}. \tag{2.1}$$

Moreover,

$$\int |g_{\vartheta,f}(\mathbf{x}) - g_{\theta,f}(\mathbf{x})| \, d\mathbf{x} \to 0, \qquad \text{as } \vartheta \to \theta. \tag{2.2}$$

Assumption 2.2. *There exist a $v \in \mathbb{R}^m$, a positive definite $m \times m$ matrix M and measurable functions $\dot{h}_j$ from $\mathbb{R}^{p+j-1} \times \Theta$ to $\mathbb{R}^m$, $j = 1, 2, \ldots$, such that for all local sequences $\langle \vartheta_n \rangle$ and $\langle \theta_n \rangle$*

$$\sum_{j=1}^{n} |H_j(\vartheta_n) - H_j(\theta_n) - (\vartheta_n - \theta_n)^{\mathrm{T}} \dot{H}_j(\theta_n)|^2 = o_{\theta_n}(1), \tag{2.3}$$

$$\max_{1 \leqslant j \leqslant n} \frac{1}{\sqrt{n}} \|\dot{H}_j(\theta_n)\| = o_{\theta_n}(1), \tag{2.4}$$

$$\frac{1}{n} \sum_{j=1}^{n} \dot{H}_j(\theta_n) = \nu + o_{\theta_n}(1), \tag{2.5}$$

$$\frac{1}{n} \sum_{j=1}^{n} \dot{H}_j(\theta_n) \dot{H}_j^{\mathrm{T}}(\theta_n) = M + o_{\theta_n}(1), \tag{2.6}$$

where $\dot{H}_j(\vartheta) = \dot{h}_j(\mathbf{X}_{j-1}, \vartheta)$ *for* $j = 1, 2, \ldots$ *and* $\vartheta \in \Theta$.

The quantities ν and M may depend on the parameter value θ. But since θ is fixed throughout, we have suppressed this dependence. Let us now introduce the parametrization of the error density.

Definition 2.3. *By an s-dimensional path we mean a map* $\eta \mapsto f_\eta$ *from a neighbourhood* Δ *of the origin in* $\mathbb{R}^s$ *into* $\mathscr{F}$ *such that* $f_0 = f$. *The path* $\eta \mapsto f_\eta$ *is said to be* ζ*-smooth if* ζ *is a measurable function from* $\mathbb{R}$ *to* $\mathbb{R}^s$ *such that* $\int \|\zeta\|^2 \, \mathrm{d}F < \infty$, $\int \zeta\zeta^{\mathrm{T}} \, \mathrm{d}F$ *is non-singular, and*

$$\int \left(\sqrt{f_\eta(x)} - \sqrt{f(x)} - \frac{1}{2} \eta^{\mathrm{T}} \zeta(x) \sqrt{f(x)} \right)^2 \mathrm{d}x = o(\|\eta\|^2). \tag{2.7}$$

The path $\eta \mapsto f_\eta$ *is said to be* ζ*-regular if it is* ζ*-smooth and if*

$$\int |g_{\vartheta, f_\eta}(\mathbf{x}) - g_{\theta, f}(\mathbf{x})| \, \mathrm{d}\mathbf{x} \to 0, \qquad \text{as } \vartheta \to \theta \text{ and } \eta \to 0. \tag{2.8}$$

Now let $\eta \mapsto f_\eta$ be an s-dimensional ζ-smooth path. Define $(m + s)$-dimensional random vectors

$$S_j(\vartheta, \zeta) = \begin{pmatrix} \dot{H}_j(\vartheta) \swarrow (\varepsilon_j(\vartheta)) \\ \zeta(\varepsilon_j(\vartheta)) \end{pmatrix}, \qquad j = 1, 2, \ldots,$$

and an $(m + s) \times (m + s)$ matrix

$$V(\zeta) = \begin{bmatrix} JM & \nu \int \swarrow \zeta^{\mathrm{T}} \, \mathrm{d}F \\ \int \swarrow \zeta \, \mathrm{d}F \nu^{\mathrm{T}} & \int \zeta\zeta^{\mathrm{T}} \, \mathrm{d}F \end{bmatrix}.$$

Let $P^n_{\vartheta,\eta}$ be the restriction of P_{ϑ, f_η} to the σ-field generated by $\mathbf{X}_n$. For $\vartheta_1, \vartheta_2 \in \Theta$ and $\eta \in \Delta$, let $\Lambda_n(\vartheta_1, \vartheta_2, \eta)$ denote the log-likelihood ratio of $P^n_{\vartheta_2,\eta}$ to $P^n_{\vartheta_1,0}$:

$$\Lambda_n(\vartheta_1, \vartheta_2, \eta) = \sum_{j=1}^{n} \log \frac{f_\eta(X_j - H_j(\vartheta_2))}{f(X_j - H_j(\vartheta_1))} + \log \frac{g_{\vartheta_2, f_\eta}(\mathbf{X}_0)}{g_{\vartheta_1, f}(\mathbf{X}_0)}.$$

We are now ready to state and prove the following LAN result.

Theorem 2.4. *Suppose Assumptions 2.1 and 2.2 hold, the path $\eta \mapsto f_\eta$ is ζ-regular and $V(\zeta)$ is positive definite. Let $\langle \theta_n \rangle$ be a local sequence and $\langle v_n \rangle = \langle (t_n, u_n) \rangle$ be a bounded sequence in $\mathbb{R}^m \times \mathbb{R}^s$. Then*

$$\Lambda_n\left(\theta_n, \theta_n + \frac{1}{\sqrt{n}} t_n, \frac{1}{\sqrt{n}} u_n\right) = \frac{1}{\sqrt{n}} \sum_{j=1}^n v_n^{\mathrm{T}} S_j(\theta_n, \zeta) - \tfrac{1}{2} v_n^{\mathrm{T}} V(\zeta) v_n + o_{\theta_n}(1), \tag{2.9}$$

and

$$\mathfrak{L}\left(\frac{1}{\sqrt{n}} \sum_{j=1}^n S_j(\theta_n, \zeta) | P_{\theta_n}\right) \Rightarrow \mathcal{N}(0, V(\zeta)). \tag{2.10}$$

Consequently,

$$\frac{1}{\sqrt{n}} \sum_{j=1}^n (S_j(\theta_n, \zeta) - S_j(\theta, \zeta)) + \begin{bmatrix} JM \\ \int \dot{\ell}_\zeta \, \mathrm{d}F \nu^{\mathrm{T}} \end{bmatrix} \sqrt{n}(\theta_n - \theta) = o_\theta(1). \tag{2.11}$$

Proof. Let $\eta_n = n^{-1/2} u_n$, $\delta_n = n^{-1/2} t_n$ and $Z_{n,j} = n^{-1/2} S_j(\theta_n, \zeta)$, $j = 1, \ldots, n$. Our proof utilizes the martingale central limit theorem – see Corollary 3.1 in Hall and Heyde (1980) – and a proper application of Theorem 3.10 in Fabian and Hannan (1987). More precisely, we shall apply their theorem with $\Theta_n = \{t\colon t + \theta_n \in \Theta\} \times \Delta$, $\theta = 0$, $E_{n,(t,u)}(\cdot) = \int \cdot \, \mathrm{d}P^n_{\theta_n + t, u}$, $U_{n,j} = Z_{n,j}$ and $M_n = nI_{m+s}$, where I_{m+s} denotes the $(m+s) \times (m+s)$ identity matrix. In view of these results it suffices to verify

$$E_{\theta_n}(Z_{n,j} | \mathbf{X}_{j-1}) = 0, \qquad j = 1, \ldots, n, \qquad P_{\theta_n} \text{ a.s.}, \tag{2.12}$$

$$L_n(a) = \sum_{j=1}^n E_{\theta_n}(\|Z_{n,j}\|^2 I[\|Z_{n,j}\| > a] | \mathbf{X}_{j-1}) = o_{\theta_n}(1), \qquad a > 0, \tag{2.13}$$

$$\sum_{j=1}^n E_{\theta_n}(Z_{n,j} Z_{n,j}^{\mathrm{T}} | \mathbf{X}_{j-1}) = V(\zeta) + o_{\theta_n}(1), \tag{2.14}$$

$$\int \left(\sqrt{g_{\theta_n, f_{\eta_n}}(\mathbf{x})} - \sqrt{g_{\theta_n, f}(\mathbf{x})}\right)^2 \mathrm{d}\mathbf{x} + \sum_{j=1}^n \int w_{n,j}^2(y) \, \mathrm{d}y = o_{\theta_n}(1), \tag{2.15}$$

where

$$w_{n,j}(y) = [f_{\eta_n}(y - H_j(\theta_n + \delta_n))]^{1/2} - [f(y - H_j(\theta_n))]^{1/2} - \frac{1}{2\sqrt{n}} v_n^{\mathrm{T}} \dot{s}_{n,j}(y - H_j(\theta_n))$$

with

$$\dot{s}_{n,j}(y) = \begin{pmatrix} \dot{H}_j(\theta_n) \dot{\ell}(y) \\ \zeta(y) \end{pmatrix} \sqrt{f(y)}, \qquad y \in \mathbb{R}.$$

Straightforward calculations and Assumption 2.2 yield (2.12) and (2.14). Verify that

$$L_n(a) = \frac{1}{n}\sum_{j=1}^{n}\int \mathbf{1}_{\{\|\dot{s}_{n,j}\| > a\sqrt{nf}\}}(y)\|\dot{s}_{n,j}(y)\|^2\,\mathrm{d}y$$

$$\leqslant \frac{2}{n}\sum_{j=1}^{n}\|\dot{H}_j(\theta_n)\|^2\int \mathbf{1}_{\{2B_n|\dot{\ell}| > a\}}\dot{\ell}^2\,\mathrm{d}F + 2\int \mathbf{1}_{\{2\|\xi\| > a\sqrt{n}\}}\|\xi\|^2\,\mathrm{d}F, \qquad a > 0,$$

with $B_n = \max_{1\leqslant j\leqslant n} n^{-1/2}\|\dot{H}_j(\theta_n)\|$. Thus (2.13) follows from (2.1), (2.4), (2.6) and the finiteness of $\int ||\xi||^2\,\mathrm{d}F$. The first integral on the left-hand side of (2.15) tends to zero by (2.8). To deal with the second term, set $\zeta_* = \zeta\sqrt{f}/2$, $\xi = \dot{\ell}\sqrt{f}/2$ and $R_{n,j} = H_j(\theta_n + \delta_n) - H_j(\theta_n)$ and conclude from (2.3) and (2.4) that $R_n = \max_{1\leqslant j\leqslant n}\|R_{n,j}\| = o_{\theta_n}(1)$. Now bound the second term in (2.15) by $4(T_{n,1} + T_{n,2} + T_{n,3} + T_{n,4})$, where

$$T_{n,1} = \sum_{j=1}^{n}\int (f_{\eta_n}^{1/2} - f^{1/2} - \eta_n^{\mathrm{T}}\zeta_*)^2(y - H_j(\theta_n + \delta_n))\,\mathrm{d}y$$

$$= n\int (f_{\eta_n}^{1/2} - f^{1/2} - \eta_n^{\mathrm{T}}\zeta_*)^2(y)\,\mathrm{d}y \to 0$$

by the ζ-smoothness of the path $\eta \mapsto f_\eta$;

$$T_{n,2} = \frac{\|u_n\|^2}{n}\sum_{j=1}^{n}\int \|\zeta_*(y - H_j(\theta_n + \delta_n)) - \zeta_*(y - H_j(\theta_n))\|^2\,\mathrm{d}y$$

$$\leqslant \|u_n\|^2 \sup_{|t|\leqslant R_n}\int \|\zeta_*(y - t) - \zeta_*(y)\|^2\,\mathrm{d}y = o_{\theta_n}(1)$$

in view of $R_n = o_{\theta_n}(1)$ and Theorem 9.5 in Rudin (1974);

$$T_{n,3} = \sum_{j=1}^{n}\int (f^{1/2}(y - R_{n,j}) - f^{1/2}(y) - R_{n,j}\xi(y))^2\,\mathrm{d}y$$

$$\leqslant \sum_{j=1}^{n} R_{n,j}^2\int_0^1\int (\xi(y - tR_{n,j}) - \xi(y))^2\,\mathrm{d}y\,\mathrm{d}t$$

$$\leqslant \sum_{j=1}^{n} R_{n,j}^2 \sup_{|t|\leqslant R_n}\int (\xi(y - t) - \xi(y))^2\,\mathrm{d}y = o_{\theta_n}(1)$$

by Assumption 2.1, (2.3), (2.6) and Theorem 9.5 in Rudin (1974); and

$$T_{n,4} = \sum_{j=1}^{n}(R_{n,j} - \delta_n^{\mathrm{T}}\dot{H}_j(\theta_n))^2\int \xi^2(y)\,\mathrm{d}y = o_{\theta_n}(1)$$

by (2.3). This completes the proof. □

Remark 2.5. Inspection of the above proof shows that (2.5) is not needed if $u_n = 0$. Of course, the case $u_n = 0$ has already been obtained by Drost *et al*. (1994) and Jeganathan (1995).

Remark 2.6. *On Assumption 2.2.* If Assumption 2.1 holds and (2.3), (2.4) and (2.6) are met with $\theta_n = \theta$, then an application of Theorem 2.4 with $u_n = 0$ and $\theta_n = \theta$ yields that $\mathfrak{L}(\mathbf{X}_n|P_{\vartheta_n})$ and $\mathfrak{L}(\mathbf{X}_n|P_\theta)$ are mutually contiguous for each local sequence $\langle\vartheta_n\rangle$. Thus, under Assumption 2.1, to verify Assumption 2.2 it suffices to show that (2.3)–(2.6) hold with $\theta_n = \theta$ and that

$$\frac{1}{n}\sum_{j=1}^{n} \|\dot{H}_j(\theta_n) - \dot{H}_j(\theta)\|^2 = o_\theta(1). \tag{2.16}$$

In particular, consider a stationary and ergodic NLAR(1) process where $H_j(\vartheta) = h(X_{j-1}, \vartheta)$ for some function h from $\mathbb{R}\times\Theta$ to $\mathbb{R}$. Assume that there exists a function $\dot{h}$ from $\mathbb{R}\times\Theta$ into $\mathbb{R}^m$ such that $\mathrm{E}_\theta\|\dot{h}(X_0, \theta)\|^2 < \infty$, $\mathrm{E}_\theta \dot{h}(X_0, \theta)\dot{h}^{\mathrm{T}}(X_0, \theta)$ is positive definite,

$$\mathrm{E}_\theta(h(X_0, \vartheta) - h(X_0, \theta) - (\vartheta - \theta)^{\mathrm{T}}\dot{h}(X_0, \theta))^2 = o(\|\vartheta - \theta\|^2) \tag{2.17}$$

and

$$\mathrm{E}_\theta\|\dot{h}(X_0, \vartheta) - \dot{h}(X_0, \theta)\|^2 \to 0 \qquad \text{as } \vartheta \to \theta. \tag{2.18}$$

In the presence of Assumption 2.1, these conditions imply Assumption 2.2 with $\nu = \mathrm{E}_\theta \dot{h}(X_0, \theta)$ and $M = \mathrm{E}_\theta \dot{h}(X_0, \theta)\dot{h}^{\mathrm{T}}(X_0, \theta)$.

Remark 2.7. *On (2.2) and (2.8)*. Consider again a stationary and ergodic NLAR(1) process so that $H_j(\vartheta) = h(X_{j-1}, \vartheta)$ for some function h from $\mathbb{R}\times\Theta$ to $\mathbb{R}$. Assume that there is a positive constant A and a measurable non-negative function ψ such that

$$|h(x, \vartheta)| \leqslant A\psi(x), \qquad x \in \mathbb{R}, \tag{2.19}$$

and

$$|h(x, \vartheta) - h(x, \theta)| \leqslant \|\vartheta - \theta\| A\psi(x), \qquad x \in \mathbb{R}, \tag{2.20}$$

for all ϑ close to θ. It then follows from Lemma 8.2 in the Appendix, that (2.2) is implied by

$$\limsup_{\vartheta\to\theta} \mathrm{E}_\vartheta \psi(X_0) < \infty, \tag{2.21}$$

and, for a ξ-smooth path $\eta \mapsto f_\eta$, (2.8) is implied by

$$\limsup_{\vartheta\to\theta,\eta\to 0} \mathrm{E}_{\vartheta, f_\eta} \psi(X_0) < \infty. \tag{2.22}$$

In the AR(1) model one has $\Theta = (-1, 1)$, $h(x, \vartheta) = \vartheta x$, $\mathrm{E}_{\vartheta,\phi}|X_0| \leqslant \int |x|\phi(x)\,\mathrm{d}x/(1 - |\vartheta|)$ and $\mathrm{E}_{\vartheta,\phi}(X_0^2) = \int x^2\phi(x)\,\mathrm{d}x/(1 - \vartheta^2)$. Thus (2.19), (2.20) and (2.22) hold with $A = 1$ and $\psi(x) = |x|$ and one obtains (2.8) for every smooth path that also satisfies

$$\limsup_{\eta\to 0} \int |x| f_\eta(x)\,\mathrm{d}x < \infty. \tag{2.23}$$

The following result can be used to verify (2.22) if no closed form for $\mathrm{E}_{\vartheta,\phi}\psi(X_0)$ is available. Suppose the densities in $\mathscr{F}$ are positive, $\psi(x) \to \infty$ as $|x| \to \infty$ and

$$\int \psi(y + h(x, \vartheta))\phi(y)\,\mathrm{d}y \leqslant C + (1 - 2\delta)\psi(x), \qquad x \in \mathbb{R}, \tag{2.24}$$

for positive constants C and δ. Then for all sufficiently large K,

$$\mathrm{E}_{\vartheta,\phi}\psi(X_0) \leqslant \frac{1}{\delta}(C + (1 - \delta) \sup_{|x| \leqslant K} \psi(x)). \tag{2.25}$$

This can be derived from Theorem 1 in Tweedie (1983). Indeed, if K is large enough so that $C \leqslant \delta\psi(x)$ whenever $|x| > K$ and if $\sup_{|x| \leqslant K} \psi(x) < \infty$, then the assumptions of his theorem hold with $A = [-K, K]$ and $g = \delta^{-1}\psi$ and the desired result follows from the bound established in his proof.

Example 2.8. *SETAR(2; 1, 1) model.* Take $\mathscr{F}$ to be a subset of $\mathscr{F}_0^+$ so that $\mathscr{F}$ is a set of positive Lebesgue densities with zero means, finite variances and finite Fisher information for location, and take

$$\Theta = \{\vartheta \in \mathbb{R}^2 \colon \vartheta_1 < 1,\ \vartheta_2 < 1,\ \vartheta_1\vartheta_2 < 1\}.$$

Petrucelli and Woolford (1984) have shown that the SETAR(2; 1, 1) model defined by $H_j(\vartheta) = h(X_{j-1}, \vartheta)$, where

$$h(x, \vartheta) = \vartheta_1 x I[x \leqslant 0] + \vartheta_2 x I[x > 0], \qquad x \in \mathbb{R},$$

is ergodic for each $(\vartheta, \phi) \in \Theta \times \mathscr{F}$. Thus we take $\{g_{\vartheta,\phi} \colon (\vartheta, \phi) \in \Theta \times \mathscr{F}\}$ to be the stationary densities. Chan *et al*. (1985) have shown that the finiteness of the error variance implies that $\mathrm{E}_\theta(X_0^2) < \infty$. From this one easily derives (2.17) and (2.18) with

$$\dot{h}(x, \vartheta) = \begin{pmatrix} xI[x \leqslant 0] \\ xI[x > 0] \end{pmatrix}.$$

One also finds that

$$M = \mathrm{E}_\theta \dot{h}(X_0, \theta)\dot{h}^{\mathrm{T}}(X_0, \theta) = \begin{bmatrix} \mathrm{E}_\theta X_0^2 I[X_0 \leqslant 0] & 0 \\ 0 & \mathrm{E}_\theta X_0^2 I[X_0 > 0] \end{bmatrix}$$

is positive definite.

Now let a, b, c be positive numbers, $c < 1$, such that $\theta \in U$, where $U = (-ac/b, c) \times (-bc/a, c) \subset \Theta$, and set

$$\psi(x) = \begin{cases} a|x|, & x \leqslant 0, \\ bx, & x > 0. \end{cases}$$

Then, for all $\vartheta \in U$ and $\phi \in \mathscr{F}$, one verifies (2.19) and (2.20) for some $A > 0$ and calculates

$$\int \psi(y + h(x, \vartheta))\phi(y)\,\mathrm{d}y \leqslant (a + b)\int |y|\phi(y)\,\mathrm{d}y + c\psi(x), \qquad x \in \mathbb{R}.$$

Thus, in view of Remark 2.7, for each positive C there is a positive constant K_C such that

$$\mathrm{E}_{\vartheta,\phi}\psi(X_0) \leqslant \frac{2}{1-c}\left((a+b)C + \frac{1+c}{2}(a+b)K_C\right)$$

for all $\vartheta \in U$ and $\phi \in \mathscr{F}$ with $\int |y|\phi(y)\,\mathrm{d}y \leqslant C$, and this implies (2.2) and (2.8) for all smooth paths which satisfy (2.23). As $f \in \mathscr{F}_0$ satisfies (2.1), we see that Assumptions 2.1 and 2.2 hold and that every smooth path which satisfies (2.23) is regular.

Example 2.9. *EXPAR(1) model*. Let $\mathscr{F}$ again be a subset of $\mathscr{F}_0^+$, $\Theta = \{\vartheta \in \mathbb{R}^3\colon |\vartheta_1| < 1, \vartheta_3 > 0\}$ and

$$h(x, \vartheta) = (\vartheta_1 + \vartheta_2\,\mathrm{e}^{-\vartheta_3 x^2})x, \qquad x \in \mathbb{R}.$$

Chan and Tong (1985) have shown that the EXPAR(1) model defined by $H_j(\vartheta) = h(X_{j-1}, \vartheta)$ is geometrically ergodic for each $(\vartheta, \phi) \in \Theta \times \mathscr{F}$. We take $\{g_{\vartheta,\phi}\colon (\vartheta, \phi) \in \Theta \times \mathscr{F}\}$ to be the stationary densities. Let a, b, c be positive numbers, $a < 1$, $c < b$, such that $\theta \in U$, where $U = (-a, a) \times (-b, b) \times (c, b)$. Then one verifies (2.19) and (2.20) for all $\vartheta \in U$ with $\psi(x) = |x|$ and some $A > 0$. Furthermore, one calculates for $\vartheta \in U$ and $\phi \in \mathscr{F}$ that

$$\int |y + h(x, \vartheta)|^2\phi(y)\,\mathrm{d}y \leqslant \int y^2\phi(y)\,\mathrm{d}y + a^2x^2 + (2ab + b^2)\sup_{t\in\mathbb{R}} t^2\,\mathrm{e}^{-ct^2}, \qquad x \in \mathbb{R}.$$

Thus, in view of Remark 2.7, for each $B > 0$ there exists a $K_B > 0$ such that

$$\mathrm{E}_{\vartheta,\phi}X_0^2 \leqslant \frac{2}{1-a^2}\left(B + (2ab + b^2)\sup_{t\in\mathbb{R}} t^2\,\mathrm{e}^{-ct^2} + \frac{1+a^2}{2}K_B\right)$$

for all $\vartheta \in U$ and $\phi \in \mathscr{F}$ with $\int y^2\phi(y)\,\mathrm{d}y \leqslant B$. By the choice of $\mathscr{F}$, f has finite Fisher information for location. Thus one verifies with the aid of Remark 2.6 that Assumptions 2.1 and 2.2 hold with

$$\dot{h}(x, \vartheta) = \begin{pmatrix} x \\ x\mathrm{e}^{-\vartheta_3 x^2} \\ -\vartheta_2 x^3\,\mathrm{e}^{-\vartheta_3 x^2} \end{pmatrix}$$

provided $M = \mathrm{E}_\theta \dot{h}(X_0, \theta)\dot{h}^{\mathrm{T}}(X_0, \theta)$ is invertible, and obtains from Remark 2.7 that every smooth path which satisfies

$$\limsup_{\eta\to 0} \int x^2 f_\eta(x)\,\mathrm{d}x < \infty \tag{2.26}$$

is regular. It is easy to see that the matrix M is singular if $\theta_2 = 0$. Of course, this is intuitively clear; if $\theta_2 = 0$, then θ_3 is not identifiable. To avoid this singularity, we shall work in the following mainly with the EXPAR(1) model in which θ_3 is known, say $\theta_3 = \gamma$. For this model $\Theta = (-1, 1) \times \mathbb{R}$ and

$$h(x, \vartheta) = (\vartheta_1 + \vartheta_2\,\mathrm{e}^{-\gamma x^2})x, \qquad x \in \mathbb{R}.$$

We refer to this model as the *restricted* EXPAR(1) model.

3. Efficiency considerations

Throughout this section we assume again that Assumptions 2.1 and 2.2 hold. We shall now discuss efficient estimation of θ. For this purpose, let $\mathcal{Q}$ denote a set of regular paths. For a path q in $\mathcal{Q}$, we let s_q denote its dimension, ζ_q its smoothness parameter,

$$\tau_q = \int \dot{\ell} \zeta_q^{\mathrm{T}} \, \mathrm{d}F \left(\int \zeta_q \zeta_q^{\mathrm{T}} \, \mathrm{d}F \right)^{-1} \zeta_q$$

the projection of $\dot{\ell}$ onto the linear span $T_q = \{a^{\mathrm{T}} \zeta_q : a \in \mathbb{R}^{s_q}\}$ generated by the components of ζ_q, and

$$I(q) = JM - \nu \left(\int \dot{\ell} \zeta_q^{\mathrm{T}} \, \mathrm{d}F \left(\int \zeta_q \zeta_q^{\mathrm{T}} \, \mathrm{d}F \right)^{-1} \int \zeta_q \dot{\ell} \, \mathrm{d}F \right) \nu^{\mathrm{T}} = JM - \nu\nu^{\mathrm{T}} \int \tau_q^2 \, \mathrm{d}F$$

its information matrix for estimating θ. In view of a well-known formula for the determinant of partitioned matrices, $\det(V(\zeta_q)) = \det(I(q)) \det(\int \zeta_q \zeta_q^{\mathrm{T}} \, \mathrm{d}F)$. This shows that $V(\zeta_q)$ is invertible if and only if $I(q)$ is. As $I(q)$ can be written as $M \int (\dot{\ell} - \tau_q)^2 \, \mathrm{d}F + (M - \nu\nu^{\mathrm{T}}) \int \tau_q^2 \, \mathrm{d}F$, we see that $I(q)$ is positive definite if and only if $\dot{\ell} \neq \tau_q$ or $M - \nu\nu^{\mathrm{T}}$ is invertible. Thus $V(\zeta_q)$ is invertible if $\dot{\ell}$ does not belong to T_q. The invertibility of $V(\zeta_q)$ is required for LAN.

Now let $T_{\mathcal{Q}}$ denote the closed linear span generated by $\cup_{q \in \mathcal{Q}} T_q$ and $\dot{\ell}_*$ denote the projection of $\dot{\ell}$ onto $T_{\mathcal{Q}}$. We make the following additional assumption.

Assumption 3.1. *The score function $\dot{\ell}$ does not belong to $T_{\mathcal{Q}}$. There exists a path q_* in $\mathcal{Q}$ such that $\dot{\ell}_* \in T_{q_*}$.*

In view of the above discussion, this assumption guarantees the invertibility of $V(\zeta_q)$ for each $q \in \mathcal{Q}$; consequently, each path in $\mathcal{Q}$ generates a LAN subproblem. Abbreviate $I(q_*)$ by I_* so that

$$I_* = JM - \nu\nu^{\mathrm{T}} \int \dot{\ell}_*^2 \, \mathrm{d}F.$$

By the definition of $\dot{\ell}_*$, the difference $I(q) - I_*$ is non-negative definite. Thus the path q_* contains the least amount of information about θ and is hence a *least favourable* path for estimating θ. The matrix I_* will be called the *efficient information* (*matrix*) for estimating θ.

By a *loss function* we mean a Borel measurable function L from $\mathbb{R}^m$ into $[0, \infty)$ such that $L(x) = L(-x)$ for all $x \in \mathbb{R}^m$ and the set $\{L \leqslant u\}$ is convex for each $u > 0$. By an estimate of θ we mean a sequence $\langle Z_n \rangle$ of m-dimensional random vectors with Z_n a measurable function of $\mathbf{X}_n$.

Theorem 3.2. *Let Assumptions 2.1, 2.2 and 3.1 hold. Let $\langle Z_n \rangle$ be an estimate of θ. Then*

$$\sup_{q \in \mathcal{Q}} \lim_{C \to \infty} \liminf_{n \to \infty} \sup_{\|\vartheta - \theta\| + \|\eta\| \leqslant C/\sqrt{n}} \mathrm{E}_{\vartheta, q(\eta)} L(\sqrt{n}(Z_n - \vartheta)) \geqslant \int L \, \mathrm{d}\mathcal{N}(0, I_*^{-1}) \tag{3.1}$$

for every loss function L. Moreover, if $\langle Z_n \rangle$ *satisfies*

$$\sqrt{n}(Z_n - \theta) - \frac{1}{\sqrt{n}} \sum_{j=1}^{n} I_*^{-1}(\dot{H}_j(\theta)\ell(\varepsilon_j(\theta)) - \nu\ell_*(\varepsilon_j(\theta))) = o_\theta(1), \tag{3.2}$$

then

$$\mathfrak{L}(\sqrt{n}(Z_n - \theta_n)|P_{\theta_n, q(u_n/\sqrt{n})}) \Rightarrow \mathcal{N}(0, I_*^{-1}) \tag{3.3}$$

for every local sequence $\langle \theta_n \rangle$, *every* $q \in \mathcal{Q}$ *and every bounded sequence* u_n *in* $\mathbb{R}^{s_q}$, *and the latter implies*

$$\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{\|\vartheta-\theta\|+\|\eta\| \leqslant C/\sqrt{n}} \mathrm{E}_{\vartheta, q(\eta)} L(\sqrt{n}(Z_n - \vartheta)) \leqslant \int L \, \mathrm{d}\mathcal{N}(0, I_*^{-1}) \tag{3.4}$$

for every bounded loss function L and every path $q \in \mathcal{Q}$.

Proof. The above can be deduced from the results in Schick (1988). Alternatively and more directly, we can proceed as follows. It follows from Theorem 6 in Fabian and Hannan (1982) that

$$\lim_{C\to\infty} \liminf_{n\to\infty} \sup_{\|\vartheta-\theta\|+\|\eta\| \leqslant C/\sqrt{n}} \mathrm{E}_{\vartheta, q_*(\eta)} L(\sqrt{n}(Z_n - \vartheta)) \geqslant \int L \, \mathrm{d}\mathcal{N}(0, I_*^{-1})$$

for every loss function L. This immediately implies the lower bound (3.1).

To verify (3.3) fix a path q in $\mathcal{Q}$. It follows from (3.2) that

$$\mathfrak{L}\left(\begin{pmatrix} \sqrt{n}(Z_n - \theta) \\ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} S_j(\theta, \zeta_q) \end{pmatrix} | P_\theta \right) \Rightarrow \mathcal{N}\left(0, \begin{bmatrix} I_*^{-1} & C \\ C^{\mathrm{T}} & V(\zeta_q) \end{bmatrix}\right),$$

where $C = [I \ 0]$ with I the $m \times m$ identity matrix. The desired result now follows from Theorem 2.4 and an application of Le Cam's third lemma – see Le Cam (1960, Theorem 2.1) or Hájek and Šidák (1967). Since loss functions are almost surely continuous with respect to $\mathcal{N}(0, I_*^{-1})$, as shown in Fabian and Hannan (1982, p. 467), (3.3) implies (3.4). □

Definition 3.3. *In view of the above result an estimate* $\langle Z_n \rangle$ *of* θ *that satisfies (3.2) will be called* $\mathcal{Q}$*-efficient or simply efficient if* $\mathcal{Q}$ *is clear from the context.*

Remark 3.4. *Adaptive estimation*. Suppose ℓ is orthogonal to $T_{\mathcal{Q}}$. Then $\ell_* = 0$, the efficient information I_* reduces to JM, and every path in $\mathcal{Q}$ is least favourable. Note that JM is the information matrix if the error density is known. This means that there is no loss of information for not knowing the error density f. To stress this special fact, efficient estimates are called *adaptive*, or more precisely $\mathcal{Q}$*-adaptive*. A necessary condition for adaptive estimation is

$$\nu \int \ell \xi_q^{\mathrm{T}} \, \mathrm{d}F = 0, \qquad \text{for each } q \in \mathcal{Q}. \tag{3.5}$$

This condition goes back to Stein (1956); see also Fabian and Hannan (1982). Note that (3.5) is satisfied if either $\nu = 0$ or

$$\int \dot{\ell} \zeta_q \, \mathrm{d}F = 0, \qquad \text{for each } q \in \mathcal{Q}. \tag{3.6}$$

In stationary AR(m) models with centred and square-integrable innovations, one has $\nu = 0$, and adaptive estimates were constructed by Kreiss (1987b) under additional assumptions. If $\mathcal{F}$ contains only densities that are symmetric about zero, then $\dot{\ell}$ is odd and each ζ_q is even, hence (3.6) holds. If $\mathcal{F}$ includes asymmetric densities, then typically (3.6) fails to hold and Stein's condition is equivalent to $\nu = 0$. For such error models adaptive estimation is ruled out if $\nu \neq 0$. For example, in the SETAR(2; 1, 1) model one has

$$\nu = \begin{pmatrix} \mathrm{E}_\theta X_0 I[X_0 \leqslant 0] \\ \mathrm{E}_\theta X_0 I[X_0 > 0] \end{pmatrix} \neq 0,$$

and adaptive estimation is *not possible* for $\mathcal{F}$ and $\mathcal{Q}$ as defined in the next example.

Example 3.5. Let $\mathcal{F} = \mathcal{F}_0^+$, and let $\mathcal{Q}$ be the set of all smooth paths $\eta \to f_\eta$ which also satisfy

$$\int x^2 f_\eta \, \mathrm{d}x \to \int x^2 f(x) \, \mathrm{d}x, \qquad \text{as } \eta \to 0. \tag{3.7}$$

We have already seen that such paths are regular in the SETAR(2; 1, 1) and restricted EXPAR(1) model. For this given class $\mathcal{Q}$ of paths we obtain

$$T_{\mathcal{Q}} = \left\{ a \in L_2(F) \colon \int a(x) f(x) \, \mathrm{d}x = 0 \text{ and } \int x a(x) f(x) \, \mathrm{d}x = 0 \right\}.$$

Indeed, for every $a \in T_{\mathcal{Q}}$ one can construct a one-dimensional path $\eta \to f_\eta$ which is a-smooth and satisfies (3.7). Utilizing the fact that $\int x \dot{\ell}(x) f(x) \, \mathrm{d}x = 1$, it is easy to verify that the projection $\dot{\ell}_*$ of $\dot{\ell}$ onto $T_{\mathcal{Q}}$ is given by

$$\dot{\ell}_*(x) = \dot{\ell}(x) - \frac{x}{\sigma^2}, \qquad x \in \mathbb{R}, \tag{3.8}$$

where σ^2 denotes the variance of f. The above shows that the class $\mathcal{Q}$ satisfies Assumption 3.1. The above also holds for $\mathcal{F} = \mathcal{F}_0$.

Remark 3.6. If we let

$$\bar{Z}_n(\vartheta) = \vartheta + I_*^{-1} \frac{1}{n} \sum_{j=1}^{n} (\dot{H}_j(\vartheta) \dot{\ell}(\varepsilon_j(\vartheta)) - \nu \dot{\ell}_*(\varepsilon_j(\vartheta))), \qquad \vartheta \in \Theta,$$

then we can express (3.2) as $\sqrt{n}(Z_n - \bar{Z}_n(\theta)) = o_\theta(1)$. It follows from (2.11) applied with the least favourable path q_* that

$$\sqrt{n}(\bar{Z}_n(\theta_n) - \bar{Z}_n(\theta)) = o_\theta(1) \tag{3.9}$$

for every local sequence $\langle \theta_n \rangle$. Consequently, (3.2) is implied by

$$\sqrt{n}(Z_n - \bar{Z}_n(\theta_n)) = o_{\theta_n}(1)$$

with $\langle \theta_n \rangle$ a local sequence for θ. This fact will be exploited in the construction of efficient estimates.

4. On the existence of efficient estimates

Throughout this section we assume that Assumptions 2.1, 2.2 and 3.1 hold. We shall now show how to construct efficient estimates if we have available preliminary $\sqrt{n}$-consistent estimates of the parameter θ and appropriate estimates of the score function $\dot{\ell}$ and its projection $\dot{\ell}_*$ onto $T_{\mathcal{Q}}$. Our construction will adapt the methods proposed by Schick (1986) in the i.i.d. case. This includes a sample splitting technique and the use of discretized versions of the preliminary estimate. The idea of discretization goes back to Le Cam (1960) and has become an important technical tool in the construction of efficient estimators in semi-parametric models; see Bickel *et al.* (1993) and references therein.

Let $\langle \tilde{\theta}_n \rangle$ be a preliminary estimate of θ and set $\varepsilon_{n,j} = \varepsilon_j(\tilde{\theta}_n)$, $j = 1, \ldots, n$. Let $\langle d_n \rangle$ and $\langle m_n \rangle$ be sequences of positive integers such that $d_n \leqslant m_n \leqslant n$, $d_n/n \to 0$ and $m_n/n \to 1/2$; set $N_n' = m_n - d_n + 1$ and $N_n'' = n - m_n$. We shall estimate $\dot{\ell}$ and $\dot{\ell}_*$ using only the observations $\mathbf{e}_{n,2} = (\varepsilon_{n,m_n+1}, \ldots, \varepsilon_{n,n})$ if we want to evaluate these estimates at $\varepsilon_{n,j}$ with $j \leqslant m_n$ and only the observations $\mathbf{e}_{n,1} = (\varepsilon_{n,d_n}, \ldots, \varepsilon_{n,m_n})$ if we want to evaluate these estimates at $\varepsilon_{n,j}$ with $j > m_n$. Set

$$\hat{\psi}_{n,j} = \begin{cases} \dot{H}_j(\tilde{\theta}_n) L_{N_n''}(\varepsilon_{n,j}, \mathbf{e}_{n,2}) - \hat{\nu}_{2,n} L_{*,\, N_n''}(\varepsilon_{n,j}, \mathbf{e}_{n,2}), & j = d_n, \ldots, m_n, \\ \dot{H}_j(\tilde{\theta}_n) L_{N_n'}(\varepsilon_{n,j}, \mathbf{e}_{n,1}) - \hat{\nu}_{1,n} L_{*,\, N_n'}(\varepsilon_{n,j}, \mathbf{e}_{n,1}), & j = m_n + 1, \ldots, n, \end{cases}$$

where

$$\hat{\nu}_{1,n} = \frac{1}{N_n'} \sum_{j=d_n}^{m_n} \dot{H}_j(\tilde{\theta}_n), \qquad \hat{\nu}_{2,n} = \frac{1}{N_n''} \sum_{j=m_n+1}^{n} \dot{H}_j(\tilde{\theta}_n),$$

and L_N and $L_{*,N}$ are measurable functions from $\mathbb{R} \times \mathbb{R}^N$ to $\mathbb{R}$ for each positive integer N. Finally, define the estimate $\langle \hat{\theta}_n \rangle$ by

$$\hat{\theta}_n = \tilde{\theta}_n + \left(\frac{1}{n} \sum_{j=d_n}^{n} \hat{\psi}_{n,j} \hat{\psi}_{n,j}^{\mathrm{T}} \right)^{-1} \frac{1}{n} \sum_{j=d_n}^{n} \hat{\psi}_{n,j}.$$

Theorem 4.1. *Let Assumptions 2.1, 2.2 and 3.1 hold. Suppose that $\langle \tilde{\theta}_n \rangle$ is a discretized $\sqrt{n}$-consistent estimate of θ and that the functions L_n and $L_{*,n}$ are such that for independent random variables $Y_1, \ldots, Y_n$ with density f*

$$\int (L_n(x, Y_1, \ldots, Y_n) - \dot{\ell}(x))^2 f(x)\, \mathrm{d}x \to 0 \text{ in probability,} \tag{4.1}$$

$$\int (L_{*,n}(x, Y_1, \ldots, Y_n) - \ell_*(x))^2 f(x)\,dx \to 0 \text{ in probability,} \tag{4.2}$$

$$\sqrt{n}\int (L_n(x, Y_1, \ldots, Y_n) - L_{*,n}(x, Y_1, \ldots, Y_n)) f(x)\,dx \to 0 \text{ in probability.} \tag{4.3}$$

Then $\langle \hat{\theta}_n \rangle$ *satisfies (3.2) and hence is efficient.*

Remark 4.2. The proof of this theorem is similar to that of Theorem 3.1 in Drost *et al.* (1994) and will not be given here. The fact that d_n may not be 1 poses no problems. Our conditions (4.1)–(4.3) correspond to Condition F of their paper. Indeed, if we interpret their $\bar{\psi}$ to be our pair $(L_n, L_{*,n})$ and their matrix C to be $[1 \ -1]$, then our (4.1), (4.2) become their (3.1) and our (4.3) becomes their (3.2). Note also that their Condition H corresponds to our (2.5).

If $d_n > 1$, our procedure does not utilize the variables $\varepsilon_{n,1}, \ldots, \varepsilon_{n,d_n-1}$. Typically, one wants $d_n = 1$, but there are cases where it is useful to let $d_n \to \infty$. This is explained in Remark 4.6 below in the case of an MA(1) process.

Example 4.3. Let us now exhibit functions L_n and $L_{*,n}$ as required in Theorem 4.1. We shall do so for the symmetric error models and for the error model $\mathscr{F}_0$ and $\mathscr{F}_0^+$. In what follows $\langle a_n \rangle$ and $\langle b_n \rangle$ are sequences of positive numbers converging to zero, k is a symmetric density that satisfies Condition K of Schick (1993) such as the logistic density, and f_n and f'_n denote the maps defined by

$$f_n(x, y_1, \ldots, y_n) = \frac{1}{na_n}\sum_{j=1}^{n} k\left(\frac{x - y_j}{a_n}\right) \text{ and } f'_n(x, y_1, \ldots, y_n) = \frac{1}{na_n^2}\sum_{j=1}^{n} k'\left(\frac{x - y_j}{a_n}\right),$$

for $x, y_1, \ldots, y_n \in \mathbb{R}$.

Let us begin with the error model $\mathscr{F}_0^+$ and the class $\mathscr{Q}$ considered in Example 3.5. It was shown in Schick (1987, pp. 99–100) that the choice

$$L_n(x, y_1, \ldots, y_n) = -\frac{f'_n(x, y_1, \ldots, y_n)}{f_n(x, y_1, \ldots, y_n) + b_n} \tag{4.4}$$

satisfies (4.1) if $n^{-1}a_n^{-3}b_n^{-1} \to 0$. If we now take

$$L_{*,n}(x, y_1, \ldots, y_n) = L_n(x, y_1, \ldots, y_n) - \frac{x}{\frac{1}{n}\sum_{j=1}^{n} y_j^2}, \tag{4.5}$$

then (4.3) holds and (4.2) follows from (4.1). The same is true for the larger error model $\mathscr{F}_0$.

Now consider the symmetric error model. In this case f is an even function, ℓ is an odd function and $\ell_* = 0$. Thus we can take $L_{*,n} = 0$ and take a symmetrized version of the above choice, namely

$$L_n(x, y_1, \ldots, y_n) = -\frac{f'_n(x, y_1, \ldots, y_n) - f'_n(-x, y_1, \ldots, y_n)}{b_n + f_n(x, y_1, \ldots, y_n) + f_n(-x, y_1, \ldots, y_n)}. \tag{4.6}$$

Again, (4.1) holds if $n^{-1}a_n^{-3}b_n^{-1} \to 0$, and (4.3) holds as L_n is odd in its first argument. Of course, (4.2) is automatically satisfied.

Remark 4.4. Candidates for $\sqrt{n}$-consistent estimates are conditional M-estimates. These estimates are minimizers of the random function

$$\vartheta \mapsto Q_n(\vartheta) = \frac{1}{n}\sum_{j=1}^{n} \rho(X_j - H_j(\vartheta)),$$

for some given (smooth) function ρ. If $\rho(x) = x^2$, then the resulting M-estimator is the conditional least-squares estimator (CLSE) studied by Klimko and Nelson (1978) and Tjøstheim (1986).

In the SETAR(2; 1, 1) model the CLSE can be written down explicitly as $\bar{\theta}_n = (\bar{\theta}_{n,1}, \bar{\theta}_{n,2})^{\mathrm{T}}$, where

$$\bar{\theta}_{n,1} = \frac{\sum_{j=1}^n X_j X_{j-1} I[X_{j-1} \leqslant 0]}{\sum_{j=1}^n X_{j-1}^2 I[X_{j-1} \leqslant 0]} \text{ and } \bar{\theta}_{n,2} = \frac{\sum_{j=1}^n X_j X_{j-1} I[X_{j-1} > 0]}{\sum_{j=1}^n X_{j-1}^2 I[X_{j-1} > 0]},$$

and it is easily checked that this estimator is $\sqrt{n}$-consistent. Thus the above construction yields efficient estimates for θ in the error models $\mathscr{F}_0^+$ for the choices of L_n and $L_{*,n}$ given by (4.4) and (4.5) and an adaptive estimate in the symmetric error model $\mathscr{F}_S = \{\phi \in \mathscr{F}_0^+ : \phi$ symmetric about $0\}$ with L_n as in (4.6) and with $L_{*,n} = 0$.

Similarly, in the restricted EXPAR(1) model the CLSE is

$$\bar{\theta}_n = \begin{bmatrix} \sum_{j=1}^n X_{j-1}^2 & \sum_{j=1}^n X_{j-1}^2 \, \mathrm{e}^{-\gamma X_{j-1}^2} \\ \sum_{j=1}^n X_{j-1}^2 \, \mathrm{e}^{-\gamma X_{j-1}^2} & \sum_{j=1}^n X_{j-1}^2 \, \mathrm{e}^{-2\gamma X_{j-1}^2} \end{bmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^n X_{j-1} X_j \\ \sum_{j=1}^n X_{j-1} \, \mathrm{e}^{-\gamma X_{j-1}^2} X_j \end{pmatrix},$$

which is easily checked to be $\sqrt{n}$-consistent. Thus the above construction with appropriate choices of L_n and $L_{*,n}$ yields an efficient estimate of θ in the error model $\mathscr{F}_0^+$ and an adaptive estimate in the symmetric error model $\mathscr{F}_S$.

Remark 4.5. The above remark shows that efficient estimates can be constructed in the SETAR(2; 1, 1) model. One can easily extend our results to more general SETAR models defined by

$$h(x, \theta) = \sum_{i=1}^{k} (\mu_i + \rho_i x) I[x \in A_i], \qquad x \in \mathbb{R}, \, \theta = (\mu^{\mathrm{T}}, \rho^{\mathrm{T}})^{\mathrm{T}},$$

where the intervals $A_1, \ldots, A_k$ form a partition of $\mathbb{R}$. See Chan *et al.* (1985) for sufficient conditions for ergodicity and stationarity and for $\sqrt{n}$-consistent preliminary estimates in this model. Thus efficient estimates of the full parameter θ can be constructed in these models as well. These estimates, however, will not be adaptive for error models with asymmetric densities. Drost *et al.* (1994) show in their Example 4.3 that $(\mu_1 - \overline{\mu}, \ldots, \mu_k - \overline{\mu}, \rho^{\mathrm{T}})^{\mathrm{T}}$ can be adaptively estimated where $\overline{\mu} = (\mu_1 + \cdots + \mu_k)/k$.

Remark 4.6. Let us now show why it is sometimes useful to have the result available for

other choices than $d_n = 1$. For this we shall consider a stationary MA(1) process. We shall show that such a process fits our model if we pretend that we can also observe the initial error variable. This idea goes back to Kreiss (1987a). We then show how the estimate constructed under this assumption can be modified to use only the actual observations. This argument requires the fact that $d_n \to \infty$. We should point out that the construction of Kreiss (1987a) utilizes the initial error variable.

The stationary MA(1) process $\{Y_t\colon t \in \mathbb{Z}\}$ satisfies the structural relation

$$Y_t = \eta_t + \theta\eta_{t-1}, \qquad t \in \mathbb{Z},$$

for some $\theta \in (-1, 1)$ and for independent and identically distributed innovations $\{\eta_t\colon t \in \mathbb{Z}\}$ with zero means and finite variances. For our purposes we also assume that these innovations possess a density f in $\mathscr{F}_0$. If we take $X_0 = \eta_0$ and $X_j = Y_j$ for $j = 1, 2, \ldots,$ and set

$$H_j(\vartheta) = \sum_{i=1}^{j}(-1)^{i-1}\vartheta^i X_{j-i}, \qquad \vartheta \in (-1, 1),$$

then we arrive at our basic model with $\Theta = (-1, 1)$, $\mathscr{F} = \mathscr{F}_0$ and initial densities $g_{\vartheta,\phi} = \phi$. One verifies that Assumptions 2.1 and 2.2 hold with

$$\dot{H}_j(\vartheta) = \sum_{i=1}^{j-1}(-1)^{i-1} i\vartheta^{i-1} X_{j-i}, \qquad \vartheta \in (-1, 1),$$

and $\nu = 0$. Thus the necessary condition for adaptive estimation is satisfied. Note also that we have chosen the random variables $\dot{H}_j(\vartheta)$ so that they do not depend on X_0. Of course, preliminary $\sqrt{n}$-consistent estimates exist in this case and can be constructed from the data $X_1, \ldots, X_n$ only. The above construction with L_n as in (4.4) leads to adaptive estimates. Since in the present case one can show that $\hat{\nu}_{i,n} = o_\theta(1)$, $i = 1, 2$, one can simplify the construction by replacing $L_{*,n}$ by L_n. The resulting estimate depends on the initial innovation $X_0 = \eta_0$. In practice, one does not observe X_0. To overcome this hurdle one replaces the variables $\varepsilon_{n,j}$ in the construction by the variables $\tilde{\varepsilon}_{n,j}$ which are obtained by substituting 0 for X_0 in the definition of $\varepsilon_{n,j}$ and chooses a preliminary estimate $\tilde{\theta}_n$ that does not require the knowledge of X_0. Since

$$\sum_{j=d_n}^{n}(\tilde{\varepsilon}_{n,j} - \varepsilon_{n,j})^2 = \sum_{j=d_n}^{n}\tilde{\theta}_n^{2j}X_0^2 \leqslant \frac{\tilde{\theta}_n^{2d_n}}{1+\tilde{\theta}_n^2}X_0^2,$$

one can now use Lemma 10.1 in Schick (1993) to conclude that the estimate based on the variables $\tilde{\varepsilon}_{n,j}$ is also adaptive provided $d_n \to \infty$ fast enough such that $n\rho^{d_n} \to 0$ for every $0 < \rho < 1$. The resulting estimate can be written as

$$\tilde{\theta}_n + \frac{\sum_{j=d_n}^{n}\tilde{\psi}_{n,j}}{\sum_{j=d_n}^{n}\tilde{\psi}_{n,j}^2},$$

where

$$\tilde{\psi}_{n,j} = \begin{cases} (\dot{H}_j(\tilde{\theta}_n) - \hat{\nu}_{1,n})L_{N_n''}(\tilde{\varepsilon}_{n,j}, \tilde{\varepsilon}_{n,m_n+1}, \ldots, \tilde{\varepsilon}_{n,n}), & j = d_n, \ldots, m_n, \\ (\dot{H}_j(\tilde{\theta}_n) - \hat{\nu}_{2,n})L_{N_n'}(\tilde{\varepsilon}_{n,j}, \tilde{\varepsilon}_{n,d_n}, \ldots, \tilde{\varepsilon}_{n,m_n}), & j = m_n + 1, \ldots, n. \end{cases}$$

5. Adaptive estimation in symmetric error models

In this section we shall show that under an additional assumption one can construct adaptive estimates for symmetric error models without splitting the sample. Our construction is essentially the same as in Jeganathan (1995), but we need to truncate the $\dot{H}_j$ in order to overcome a mistake in his proof. (He erroneously assumes that his variables $U_{nt}\hat{\psi}^*_{nt}$ form a martingale difference; but these variables are not measurable with respect to the given filtration.)

One expects that not splitting the sample should result in estimates with a better performance for moderate sample sizes. A small simulation at the end of this section supports this in the case considered.

Throughout this section we assume that Assumptions 2.1 and 2.2 hold and that f is symmetric. In addition, we impose the following condition which allows for the truncation of $\dot{H}_j$.

Condition 5.1. *For every local sequence $\langle\theta_n\rangle$ for θ and every sequence $\langle c_n\rangle$ tending to infinity,*

$$\frac{1}{n}\sum_{j=1}^{n}\|\dot{H}_j(\theta_n)\|^2 I[\|\dot{H}_j(\theta_n)\| > c_n] = o_{\theta_n}(1). \tag{5.1}$$

Let $\langle c_n\rangle$ be a sequence of positive numbers converging to infinity and χ_n denote the map from $\mathbb{R}^m$ into $\mathbb{R}^m$ defined by

$$\chi_n(x) = xI[\|x\| \leqslant c_n] + c_n\frac{x}{\|x\|}I[\|x\| > c_n], \qquad x \in \mathbb{R}^m.$$

Let $\langle d_n\rangle$ be a sequence of positive integers such that $d_n/n \to 0$. Set $N_n = n - d_n + 1$. Let $\langle\tilde{\theta}_n\rangle$ be a preliminary estimate of θ. Set

$$\varepsilon_{n,j} = \varepsilon_j(\tilde{\theta}_n) \text{ and } \dot{H}_{n,j} = \chi_{N_n}(\dot{H}_j(\tilde{\theta}_n)), \qquad j = 1, \ldots, n.$$

We estimate the score function $\dot{f}/f$ by

$$\hat{\dot{f}}_n(x) = L_{N_n}(x, \varepsilon_{n,d_n}, \ldots, \varepsilon_{n,n}), \qquad x \in \mathbb{R},$$

where L_n is defined in (4.6). Define the estimate

$$\hat{\theta}_n = \tilde{\theta}_n + (\hat{J}_n M_n)^{-1}\frac{1}{N_n}\sum_{j=d_n}^{n}\dot{H}_{n,j}\hat{\dot{f}}_n(\varepsilon_{n,j}),$$

where

$$\hat{J}_n = \frac{1}{N_n}\sum_{j=d_n}^{n}\hat{\dot{f}}^2_n(\varepsilon_{n,j}) \text{ and } M_n = \frac{1}{N_n}\sum_{j=d_n}^{n}\dot{H}_{n,j}\dot{H}^{\mathrm{T}}_{n,j}.$$

The efficiency of this estimator is proved in the following theorem.

Theorem 5.2. *Let Assumptions 2.1 and 2.2 and Condition 5.1 hold, and let f be symmetric about zero. Suppose $\langle \tilde{\theta}_n \rangle$ is a discretized $\sqrt{n}$-consistent estimator of θ and the sequences $\langle a_n \rangle$, $\langle b_n \rangle$ and $\langle c_n \rangle$ satisfy in addition*

$$n^{-1} a_n^{-3} b_n^{-1} c_n^2 \to 0. \tag{5.2}$$

Then $\langle \hat{\theta}_n \rangle$ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) - \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (JM)^{-1} \dot{H}_j(\theta) \ell(\varepsilon_j(\theta)) = o_\theta(1), \tag{5.3}$$

and is thus adaptive for symmetric error models.

Proof. For simplicity we shall give the proof only for the case $d_n = 1$. The case $d_n > 1$ is similar. By the properties of $\langle \tilde{\theta}_n \rangle$ it suffices to verify (5.3) if $\langle \tilde{\theta}_n \rangle$ is a local sequence. Therefore, throughout this proof, $\langle \tilde{\theta}_n \rangle$ will be assumed to be a local sequence. In view of (2.6), (3.9) and the mutual contiguity of $\langle \mathfrak{L}(\mathbf{X}_n | P_{\tilde{\theta}_n}) \rangle$ and $\langle \mathfrak{L}(\mathbf{X}_n | P_\theta) \rangle$, it suffices to verify

$$\hat{J}_n = J + o_{\tilde{\theta}_n}(1) \tag{5.4}$$

$$D_{n,1} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\dot{H}_{n,j} \bar{\ell}_n(\varepsilon_{n,j}) - \dot{H}_j(\tilde{\theta}_n) \ell(\varepsilon_{n,j})) = o_{\tilde{\theta}_n}(1), \tag{5.5}$$

$$D_{n,2} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{H}_{n,j} (\hat{\ell}_n(\varepsilon_{n,j}) - \bar{\ell}_n(\varepsilon_{n,j})) = o_{\tilde{\theta}_n}(1), \tag{5.6}$$

where $\bar{\ell}_n(x) = -\bar{f}_n'(x)/(b_n + \bar{f}_n(x))$ and $\bar{f}_n(x = \int f(x - a_n t) k(t)\, \mathrm{d}t$, $x \in \mathbb{R}$. Note that the random variables $\varepsilon_{n,1}, \ldots, \varepsilon_{n,n}$ are independent under $P_{\tilde{\theta}_n}$. Arguing as in Schick (1987, pp. 99–100) and utilizing (5.2), one obtains

$$\int |\bar{\ell}_n(y) - \ell(y)|^2 \, \mathrm{d}F(y) = o(1) \tag{5.7}$$

and

$$\frac{1}{n} \sum_{j=1}^{n} E_{\tilde{\theta}_n} |\hat{\ell}_n(\varepsilon_{n,j}) - \bar{\ell}_n(\varepsilon_{n,j})|^2 = O(n^{-1} a_n^{-3} b_n^{-1}). \tag{5.8}$$

Of course, this yields (5.4). For each $a \in \mathbb{R}^m$, $a^{\mathrm{T}} D_{n,1}$ is a martingale (for the σ-fields generated by $\mathbf{X}_j$) with quadratic variation $(1/n) \sum_{j=1}^{n} \int (a^{\mathrm{T}} \dot{H}_{n,j} \bar{\ell}_n(y) - a^{\mathrm{T}} \dot{H}_j(\tilde{\theta}_n) \ell(y))^2 \, \mathrm{d}F(y)$. In view of Condition 5.1 and (5.7) this variation tends to zero in $P_{\tilde{\theta}_n}$-probability. This yields (5.5).

To verify (5.6), let $\mathscr{A}_{n,j}$ denote the σ-field generated by $\mathbf{X}_j$ and $|\varepsilon_{n,1}|, \ldots, |\varepsilon_{n,n}|$. By the symmetry of f, under $P_{\tilde{\theta}_n}$, and given $\mathscr{A}_{n,j-1}$, the random variable $\operatorname{sign}(\varepsilon_{n,j})$ takes values -1 and 1 with probability $1/2$. By construction, $\hat{\ell}_n - \bar{\ell}_n$ is odd so that $\mathrm{E}_{\tilde{\theta}_n}[\hat{\ell}_n(\varepsilon_{n,j}) -$

$\bar{\ell}_n(\varepsilon_{n,j})|\mathscr{A}_{n,j-1}] = 0$, $j = 1, \ldots, n$. Thus, for each $a \in \mathbb{R}^m$ of length 1, $a^{\mathrm{T}} D_{n,2}$ is a martingale for the filtration $\{\mathscr{A}_{n,j}: j = 0, \ldots, n\}$ and its quadratic variation

$$\frac{1}{n}\sum_{j=1}^{n}(a^{\mathrm{T}}\dot{H}_{n,j})^2(\hat{\ell}_n(|\varepsilon_{n,j}|) - \bar{\ell}_n(|\varepsilon_{n,j}|))^2 \leqslant \frac{1}{n}\sum_{j=1}^{n} c_n^2(\hat{\ell}_n(\varepsilon_{n,j}) - \bar{\ell}_n(\varepsilon_{n,j}))^2 = o_{\tilde{\theta}_n}(1), \qquad (5.9)$$

in view of (5.8) and (5.2). This yields (5.6) and completes the proof. □

Remark 5.3. If the random variables $\dot{H}_1(\vartheta), \ldots, \dot{H}_n(\vartheta)$ are independent of the random variables $\varepsilon_1(\vartheta), \ldots, \varepsilon_n(\vartheta)$ under $P_{\vartheta,f}$ for every ϑ and n, then one does not need to truncate the random variables $\dot{H}_j(\tilde{\theta}_n)$ and Condition 5.1 is not required. Indeed, in this case one can replace (5.9) by

$$E_{\tilde{\theta}_n}\left(\frac{1}{n}\sum_{j=1}^{n}(a^T H_j(\tilde{\theta}_n))^2(\hat{\ell}_n(|\varepsilon_{n,j}|) - \bar{\ell}_n(|\varepsilon_{n,j}|)^2|\,\dot{H}_1(\tilde{\theta}_n), \ldots, \dot{H}_n(\tilde{\theta}_n)\right)$$

$$\leqslant \frac{1}{n}\sum_{j=1}^{n}(a^T H_j(\tilde{\theta}_n))^2 E_{\tilde{\theta}_n}(\hat{\ell}_n(\varepsilon_{n,1}) - \bar{\ell}_n(\varepsilon_{n,1}))^2 = o_{\tilde{\theta}_n}(1),$$

where $\langle\tilde{\theta}_n\rangle$ is thought of as a local sequence.

Remark 5.4. In stationary and ergodic NLAR(1) models one can provide simple sufficient conditions for Condition 5.1. Suppose there is a map ψ such that $E_\theta\psi(X_0) < \infty$ and $\psi(x) \geqslant \sup_{\|\vartheta-\theta\|<\delta}\|\dot{h}(x,\vartheta)\|^2$, $x \in \mathbb{R}$, for some $\delta > 0$. Then Condition 5.1 holds. To see this, fix a local sequence $\langle\theta_n\rangle$ and a sequence $\langle c_n\rangle$ tending to infinity. It follows from the ergodic theorem that

$$\limsup_n \frac{1}{n}\sum_{j=1}^{n}\|\dot{H}_j(\theta_n)\|^2 I[\|\dot{H}_j(\theta_n)\| > c_n] \leqslant \mathrm{E}_\theta\psi(X_0) I[\psi(X_0) > c]$$

P_θ-almost surely for every $c > 0$. As $\lim_{c\to\infty}\mathrm{E}_\theta\psi(X_0)I[\psi(X_0) > c] = 0$ we find

$$\frac{1}{n}\sum_{j=1}^{n}\|\dot{H}_j(\theta_n)\|^2 I[\|\dot{H}_j(\theta_n)\| > c_n] = o_\theta(1).$$

This and a contiguity argument yield Condition 5.1.

In the SETAR(2; 1, 1) model take $\psi(x) = x^2$ to obtain Condition 5.1 from $\mathrm{E}_\theta(X_0^2) < \infty$. In the EXPAR(1) model take $\psi(x) = Ax^2$ for some large positive A to obtain Condition 5.1 from $\mathrm{E}_\theta(X_0^2) < \infty$.

Thus, in the SETAR(2; 1, 1) and the restricted EXPAR(1) model, Condition 5.1 holds and the above construction, with $\tilde{\theta}_n$ a discretized version of the CLSE, produces adaptive estimates of θ under the symmetric error model $\mathscr{F} = \mathscr{F}_S = \{\phi \in \mathscr{F}_0^+: \phi$ is symmetric about zero$\}$.

Remark 5.5. To see whether not splitting the sample is superior in moderate sample sizes, we have performed a small simulation study in S-PLUS for the AR(1) model with $\theta = 1/2$. We

considered two error densities, the double exponential density ($f(x) = \exp(-|x|)/2$) and the $t(4)$ density. We performed the simulations for sample sizes $n = 100$ and $n = 200$. In all simulations we took the kernel $k(y) = c/(1 + y^6)$ and $b_n = 0.02$, but varied the values of the window length among the values 0.5, 0.6, 0.7, 0.8, 0.9. As preliminary estimate we took the least-squares estimate whose asymptotic variance is $0.75/n$. For each of the different values of the window length, we simulated N pseudo-samples and constructed the preliminary estimate and five efficient estimates $\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,5}$ in each pseudo-sample. We took $N = 10\,000$ when $n = 100$ and $N = 5000$ when $n = 200$. The estimates $\hat{\theta}_{n,1}$, $\hat{\theta}_{n,3}$ and $\hat{\theta}_{n,5}$ are of the form

$$\tilde{\theta}_n + \frac{\frac{1}{n}\sum_{j=1}^{n} X_{j-1}\hat{\ell}_{n,j}}{\frac{1}{n}\sum_{j=1}^{n} X_{j-1}^2 \frac{1}{n}\sum_{j=1}^{n} \hat{\ell}_{n,j}^2}, \tag{5.10}$$

and the estimates $\hat{\theta}_{n,2}$ and $\hat{\theta}_{n,4}$ are of the form

$$\tilde{\theta}_n + \frac{\sum_{j=1}^{n} X_{j-1}\hat{\ell}_{n,j}}{\sum_{j=1}^{n} X_{j-1}^2 \hat{\ell}_{n,j}^2}, \tag{5.11}$$

where $\hat{\ell}_{n,j} = \hat{\ell}_n(\varepsilon_{n,j})$ for $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$; $\hat{\ell}_{n,j} = \ell(\varepsilon_{n,j})$ for $\hat{\theta}_{n,5}$; $\hat{\ell}_{n,j} = L_{n/2}(\varepsilon_{n,j}, \varepsilon_{n,n/2+1}, \ldots, \varepsilon_{n,n})$ if $j \leqslant n/2$ and $\hat{\ell}_{n,j} = L_{n/2}(\varepsilon_{n,j}, \varepsilon_{n,1}, \ldots, \varepsilon_{n,n/2})$ if $j > n/2$, for $\hat{\theta}_{n,3}$ and $\hat{\theta}_{n,4}$ with L_n as in (4.6). Thus $\hat{\theta}_{n,1}$ is as described in this section with $d_n = 1$ and $c_n = \infty$ (no truncation), $\hat{\theta}_{n,2}$ is similar, but uses a different estimate of I_*, $\hat{\theta}_{n,3}$ and $\hat{\theta}_{n,4}$ use the sample splitting technique and only differ by the type of estimate of I_*, and $\hat{\theta}_{n,5}$ uses the actual score function and corresponds to the case of known f.

Tables 1–4 list the sample mean square errors of these estimates. In the cases considered, sample splitting does not fare as well as not splitting. As expected, the estimate $\hat{\theta}_{n,5}$ which uses the actual score function performs the best. In the case of the double exponential density the improvement of this estimate over the preliminary estimate is about 30% if $n = 100$ and 36% if $n = 200$ (the theoretical asymptotic improvement is 50%); the best improvements of the estimate $\hat{\theta}_{n,1}$ over the preliminary estimate are 23% if $n = 100$ and 31% if $n = 200$; while the best improvements of $\hat{\theta}_{n,3}$ are 15% if $n = 100$ and 26% if $n = 200$.

Table 1. Sample mean square errors, double exponential density, $n = 100$

Window length	$\tilde{\theta}_n$	$\hat{\theta}_{n,1}$	$\hat{\theta}_{n,2}$	$\hat{\theta}_{n,3}$	$\hat{\theta}_{n,4}$	$\hat{\theta}_{n,5}$
0.5	0.007 49	0.005 86	0.005 60	0.006 50	0.006 58	0.005 30
0.6	0.007 42	0.005 75	0.005 90	0.006 31	0.006 42	0.005 08
0.7	0.007 52	0.005 84	0.006 02	0.006 36	0.006 49	0.005 29
0.8	0.007 35	0.005 79	0.005 95	0.006 22	0.006 35	0.005 17
0.9	0.007 64	0.006 20	0.006 37	0.006 58	0.006 71	0.005 38

Table 2. Sample mean square errors, $t(4)$ density, $n = 100$

Window length	$\tilde{\theta}_n$	$\hat{\theta}_{n,1}$	$\hat{\theta}_{n,2}$	$\hat{\theta}_{n,3}$	$\hat{\theta}_{n,4}$	$\hat{\theta}_{n,5}$
0.5	0.007 32	0.006 60	0.006 78	0.007 02	0.007 16	0.005 82
0.6	0.007 13	0.006 22	0.006 41	0.006 66	0.006 81	0.005 66
0.7	0.007 32	0.006 28	0.006 44	0.006 66	0.006 80	0.005 76
0.8	0.007 26	0.006 12	0.006 32	0.006 49	0.006 64	0.005 72
0.9	0.007 36	0.006 23	0.006 44	0.006 56	0.006 73	0.005 81

Table 3. Sample mean square errors, double exponential density, $n = 200$

Window length	$\tilde{\theta}_n$	$\hat{\theta}_{n,1}$	$\hat{\theta}_{n,2}$	$\hat{\theta}_{n,3}$	$\hat{\theta}_{n,4}$	$\hat{\theta}_{n,5}$
0.5	0.003 89	0.002 67	0.002 72	0.002 90	0.002 93	0.002 49
0.6	0.003 72	0.002 60	0.002 63	0.002 80	0.002 83	0.002 34
0.7	0.003 79	0.002 70	0.002 74	0.002 82	0.002 86	0.002 37
0.8	0.003 68	0.002 72	0.002 77	0.002 81	0.002 86	0.002 39
0.9	0.003 75	0.002 84	0.002 88	0.002 96	0.002 99	0.002 43

Table 4. Sample mean square errors, $t(4)$ density, $n = 200$

Window length	$\tilde{\theta}_n$	$\hat{\theta}_{n,1}$	$\hat{\theta}_{n,2}$	$\hat{\theta}_{n,3}$	$\hat{\theta}_{n,4}$	$\hat{\theta}_{n,5}$
0.5	0.003 70	0.003 03	0.003 08	0.003 23	0.003 26	0.002 79
0.6	0.003 63	0.002 95	0.003 03	0.003 12	0.003 17	0.002 75
0.7	0.003 65	0.002 84	0.002 90	0.002 97	0.003 02	0.002 71
0.8	0.003 66	0.002 93	0.002 99	0.003 03	0.003 08	0.002 77
0.9	0.003 73	0.002 95	0.003 02	0.003 03	0.003 09	0.002 78

In the case of the $t(4)$ density the improvement of the estimate $\hat{\theta}_{n,5}$ over the preliminary estimate is about 21% if $n = 100$ and 25% if $n = 200$ (the theoretical asymptotic improvement is 30%); the best improvements of the estimate $\hat{\theta}_{n,1}$ over the preliminary estimate are 16% if $n = 100$ and 22% if $n = 200$; while the best improvements of $\hat{\theta}_{n,3}$ are 11% if $n = 100$ and 19% if $n = 200$. Thus the performance of the unsplit estimate $\hat{\theta}_{n,1}$ lies between that of $\hat{\theta}_{n,5}$ and $\hat{\theta}_{n,3}$.

From the tables we also see that the estimates of type (5.10), namely $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,3}$, perform slightly better than the corresponding estimates $\hat{\theta}_{n,2}$ and $\hat{\theta}_{n,4}$ of type (5.11).

6. Efficient estimation in the error models $\mathscr{F}_0$ and $\mathscr{F}_0^+$

Throughout this section we assume that Assumptions 2.1 and 2.2 hold and that f has zero mean and finite variance σ^2. We shall now avoid the sample splitting technique and construct an estimate $\langle\hat{\theta}_n\rangle$ that satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) - \frac{1}{\sqrt{n}}\sum_{j=1}^{n} I_*^{-1}\left((\dot{H}_j(\theta) - \nu)\not{\ell}(\varepsilon_j(\theta)) + \nu\frac{\varepsilon_j(\theta)}{\sigma^2}\right) = o_\theta(1) \tag{6.1}$$

with

$$I_* = JM - \nu\nu^{\mathrm{T}}\left(J - \frac{1}{\sigma^2}\right) = J(M - \nu\nu^{\mathrm{T}}) + \nu\nu^{\mathrm{T}}\frac{1}{\sigma^2}.$$

Such an estimate is efficient for the error models $\mathscr{F}_0$ and $\mathscr{F}_0^+$.

Let c_n, d_n, N_n, $\tilde{\theta}_n$, $\varepsilon_{n,j}$, $\dot{H}_{n,j}$, M_n, $\hat{J}_n$ and $\hat{\not{\ell}}_n$ be as defined in the previous section, but with L_n now as in (4.4). In addition, set

$$\hat{\nu}_n = \frac{1}{N_n}\sum_{j=d_n}^{n}\dot{H}_{n,j}, \quad \hat{\sigma}_n^2 = \frac{1}{N_n}\sum_{j=d_n}^{n}\varepsilon_{n,j}^2 \quad \text{and} \quad \hat{I}_{*,n} = \frac{\hat{\nu}_n\hat{\nu}_n^{\mathrm{T}}}{\hat{\sigma}_n^2} + \hat{J}_n(M_n - \hat{\nu}_n\hat{\nu}_n^{\mathrm{T}}).$$

Finally, let

$$\hat{\theta}_n = \tilde{\theta}_n + \frac{1}{N_n}\sum_{j=d_n}^{n}\hat{I}_{*,n}^{-1}\left((\dot{H}_{n,j} - \hat{\nu}_n)\hat{\not{\ell}}_n(\varepsilon_{n,j}) + \hat{\nu}_n\frac{\varepsilon_{n,j}}{\hat{\sigma}_n^2}\right).$$

To prove the efficiency of this estimator we need Condition 5.1 and the following additional condition

Condition 6.1. *For every local sequence $\langle\theta_n\rangle$ and some sequence $\langle m_n\rangle$ of positive integers tending to infinity,*

$$\frac{1}{n}\sum_{1\leqslant i,j\leqslant n,|j-i|>m_n}\sum \mathrm{E}_{\theta_n}(\|\dot{H}_j(\theta_n) - \mathrm{E}_{\theta_n}(\dot{H}_j(\theta_n)|\mathbf{B}_{n,i}(\theta_n))\|^2) = o(a_n^2), \tag{6.2}$$

where $\mathbf{B}_{n,i}(\vartheta)$ is the σ*-field generated by $\{\mathbf{X}_0, \varepsilon_1(\vartheta), \ldots, \varepsilon_{i-1}(\vartheta), \varepsilon_{i+1}(\vartheta), \ldots, \varepsilon_n(\vartheta)\}$, $\vartheta \in \Theta$, $i = 1, \ldots, n$.*

Theorem 6.2. *Suppose Assumptions 2.1 and 2.2 and Conditions 5.1 and 6.1 hold, f has zero mean and finite variance σ^2, $\langle\tilde{\theta}_n\rangle$ is a discretized $\sqrt{n}$-consistent estimator of θ and the sequences $\langle a_n\rangle$, $\langle b_n\rangle$, $\langle c_n\rangle$ and $\langle m_n\rangle$ satisfy in addition*

$$n^{-1}a_n^{-4}b_n^{-2}c_n^2 + n^{-1}a_n^{-3}b_n^{-1}c_n^2 m_n \to 0. \tag{6.3}$$

Then $\langle\hat{\theta}_n\rangle$ satisfies (6.1).

Proof. For simplicity in notation, we shall give the proof only for the case $d_n = 1$. The case $d_n > 1$ is similar. As in the proof of Theorem 5.2, we may and shall assume that $\langle \tilde{\theta}_n \rangle$ is a local sequence. Then the random variables $\varepsilon_{n,1}, \ldots, \varepsilon_{n,n}$ are independent under $P_{\tilde{\theta}_n}$ and consequently the following facts are obtainable from the arguments of Schick (1987, pp. 99–100). There is a constant c such that

$$\|\hat{\ell}_n\|_\infty \leqslant c a_n^{-1}, \qquad \max_{1 \leqslant j \leqslant n} \|\hat{\ell}_{n,j}\|_\infty \leqslant c a_n^{-1}, \tag{6.4}$$

$$\|\hat{\ell}_n - \hat{\ell}_{n,j}\|_\infty \leqslant c a_n^{-2} b_n^{-1} n^{-1}, \qquad \|\hat{\ell}_{n,j} - \hat{\ell}_{n,j,i}\|_\infty \leqslant c a_n^{-2} b_n^{-1} n^{-1}, \qquad i \neq j \tag{6.5}$$

$$\mathrm{E}_{\tilde{\theta}_n} \int |\hat{\ell}_n - \bar{\ell}_n|^2 \, \mathrm{d}F = O(n^{-1} b_n^{-1} a_n^{-3}), \tag{6.6}$$

$$\int (\bar{\ell}_n - \ell)^2 \, \mathrm{d}F \to 0, \tag{6.7}$$

$$\frac{1}{n} \sum_{j=1}^{n} \hat{\ell}_n^2(\varepsilon_{n,j}) = J + o_{\tilde{\theta}_n}(1), \tag{6.8}$$

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\hat{\ell}_n(\varepsilon_{n,j}) - \ell(\varepsilon_{n,j})) = \sqrt{n} \int \hat{\ell}_n \, \mathrm{d}F + o_{\tilde{\theta}_n}(1), \tag{6.9}$$

where $\bar{\ell}_n(y) = \mathrm{E}_{\tilde{\theta}_n}(\hat{\ell}_n(y))$,

$$\hat{\ell}_{n,j}(y) = -\frac{\dfrac{1}{n a_n^2} \displaystyle\sum_{r:r \neq j} k'\left(\frac{y - \varepsilon_{n,r}}{a_n}\right)}{b_n + \dfrac{1}{n a_n} \displaystyle\sum_{r:r \neq j} k\left(\frac{y - \varepsilon_{n,r}}{a_n}\right)}, \text{ and } \hat{\ell}_{n,j,i}(y) = -\frac{\dfrac{1}{n a_n^2} \displaystyle\sum_{r:r \neq i,j} k'\left(\frac{y - \varepsilon_{n,r}}{a_n}\right)}{b_n + \dfrac{1}{n a_n} \displaystyle\sum_{r:r \neq i,j} k\left(\frac{y - \varepsilon_{n,r}}{a_n}\right)},$$

for $y \in \mathbb{R}$ and $i \neq j$.

The independence of $\varepsilon_{n,1}, \ldots, \varepsilon_{n,n}$ under $P_{\tilde{\theta}_n}$, (2.5) and (5.1) imply that

$$\hat{\sigma}_n^2 = \sigma^2 + o_{\tilde{\theta}_n}(1) \text{ and } \hat{\nu}_n = \nu + o_{\tilde{\theta}_n}(1). \tag{6.10}$$

Using this, (2.6), (5.1), (6.8) and the non-singularity of I_*, one verifies that

$$\hat{I}_{*,n}^{-1} = I_*^{-1} + o_{\tilde{\theta}_n}(1). \tag{6.11}$$

Let

$$\tilde{Z}_n = \tilde{\theta}_n + I_*^{-1} \frac{1}{n} \sum_{j=1}^{n} \left((\dot{H}_{n,j} - \nu) \ell(\varepsilon_{n,j}) + \nu \frac{\varepsilon_{n,j}}{\sigma^2} \right)$$

and

$$\hat{Z}_n = \tilde{\theta}_n + \hat{I}_{*,n}^{-1} \frac{1}{n} \sum_{j=1}^{n} \left((\dot{H}_{n,j} - \hat{\nu}_n) \ell(\varepsilon_{n,j}) + \hat{\nu}_n \frac{\varepsilon_{n,j}}{\hat{\sigma}_n^2} \right).$$

It follows from Remark 3.6 and Condition 5.1 that $\langle \tilde{Z}_n \rangle$ satisfies (6.1). From (6.10) and (6.11) one concludes $\sqrt{n}(\hat{Z}_n - \tilde{Z}_n) = o_{\tilde{\theta}_n}(1)$. In view of the mutual contiguity of the sequences $\langle \mathfrak{L}(\mathbf{X}_n | P_{\tilde{\theta}_n}) \rangle$ and $\langle \mathfrak{L}(\mathbf{X}_n | P_\theta) \rangle$ and the above results, it suffices to prove that

$$\sqrt{n}(\hat{\theta}_n - \hat{Z}_n) = o_{\tilde{\theta}_n}(1).$$

But, by (6.9), (6.10) and (6.11), this is implied by

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{H}_{n,j}(\hat{\ell}_n(\varepsilon_{n,j}) - \ell(\varepsilon_{n,j})) = \sqrt{n}\hat{\nu}_n \int \hat{\ell}_n \, \mathrm{d}F + o_{\tilde{\theta}_n}(1). \tag{6.12}$$

From a martingale argument, (2.6), (5.1) and (6.7) we obtain

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{H}_{n,j}(\bar{\ell}_n(\varepsilon_{n,j}) - \ell(\varepsilon_{n,j})) = \sqrt{n}\hat{\nu}_n \int \bar{\ell}_n \, \mathrm{d}F + o_{\tilde{\theta}_n}(1).$$

In view of this, (6.12) is now a consequence of

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{H}_{n,j}(\hat{\ell}_n(\varepsilon_{n,j}) - \bar{\ell}_n(\varepsilon_{n,j})) = \sqrt{n}\hat{\nu}_n \int (\hat{\ell}_n - \bar{\ell}_n) \, \mathrm{d}F + o_{\tilde{\theta}_n}(1).$$

By (6.5), (2.6) and (5.1), the latter follows from

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \dot{H}_{n,j} \left(\hat{\ell}_{n,j}(\varepsilon_{n,j}) - \bar{\ell}_n(\varepsilon_{n,j}) - \int (\hat{\ell}_{n,j} - \bar{\ell}_n) \, \mathrm{d}F \right) = o_{\tilde{\theta}_n}(1). \tag{6.13}$$

To prove (6.13), let D_j denote the jth summand in the left-hand side of (6.13) and $D_{j,i} = \mathrm{E}_{\tilde{\theta}_n}(D_j | \mathbf{B}_{n,i}(\tilde{\theta}_n))$. Using conditioning arguments one verifies for $i \neq j$ that $\mathrm{E}_{\tilde{\theta}_n} D_{i,j}^{\mathrm{T}} D_j = \mathrm{E}_{\tilde{\theta}_n} D_i^{\mathrm{T}} D_{j,i} = \mathrm{E}_{\tilde{\theta}_n} D_{i,j}^{\mathrm{T}} D_{j,i} = 0$, which yields $\mathrm{E}_{\tilde{\theta}_n} D_i^{\mathrm{T}} D_j = \mathrm{E}_{\tilde{\theta}_n}(D_i - D_{i,j})^{\mathrm{T}}(D_j - D_{j,i})$. This and the Cauchy–Schwarz inequality yield

$$\begin{aligned} \mathrm{E}_{\tilde{\theta}_n} \left\| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} D_j \right\|^2 &\leqslant \frac{1}{n} \sum_{j=1}^{n} \mathrm{E}_{\tilde{\theta}_n} \|D_j\|^2 + \frac{1}{n} \sum_{i \neq j} \mathrm{E}_{\tilde{\theta}_n} \|D_j - D_{j,i}\|^2 \\ &\leqslant \frac{1 + 2m_n}{n} \sum_{j=1}^{n} \mathrm{E}_{\tilde{\theta}_n} ||D_j||^2 + \frac{1}{n} \sum_{1 \leqslant i,j \leqslant n, |j-i| > m_n} \sum \mathrm{E}_{\tilde{\theta}_n} \|D_j - D_{j,i}\|^2. \end{aligned}$$

Verify that $\mathrm{E}_{\tilde{\theta}_n} \|D_j\|^2 \leqslant c_n^2 \mathrm{E}_{\tilde{\theta}_n} \int (\hat{\ell}_{n,j} - \bar{\ell}_n)^2 \, \mathrm{d}F$ and use (6.5) and (6.6) to conclude that

$$\frac{1}{n} \sum_{j=1}^{n} \mathrm{E}_{\tilde{\theta}_n} \|D_j\|^2 \leqslant 2Cc_n^2(4a_n^{-4}b_n^{-2}n^{-2} + n^{-1}a_n^{-3}b_n^{-1}) \tag{6.14}$$

for some $C > 0$. Let

$$\tilde{D}_{j,i} = \dot{H}_{n,j,i}(\hat{\ell}_{n,j,i}(\varepsilon_{n,j}) - \bar{\ell}_n(\varepsilon_{n,j}) - \int (\hat{\ell}_{n,j,i} - \bar{\ell}_n) \, \mathrm{d}F),$$

where $\dot{H}_{n,j,i} = \mathrm{E}_{\tilde{\theta}_n}(\dot{H}_{n,j}|\mathbf{B}_{n,i}(\tilde{\theta}_n))$. Then by (6.4)

$$\mathrm{E}_{\tilde{\theta}_n}\|D_j - D_{j,i}\|^2 \leqslant \mathrm{E}_{\tilde{\theta}_n}\|D_j - \tilde{D}_{j,i}\|^2$$

$$\leqslant 32c^2 a_n^{-2}\mathrm{E}_{\tilde{\theta}_n}\|\dot{H}_{n,j} - \dot{H}_{n,j,i}\|^2 + 2c_n^2\mathrm{E}_{\tilde{\theta}_n}\int(\hat{\ell}_{n,j,i} - \hat{\ell}_{n,j})^2\,\mathrm{d}F.$$

Therefore by (6.2) and (6.5) we obtain

$$\frac{1}{n}\sum_{1\leqslant i,j\leqslant n,|j-i|>m_n}\sum \mathrm{E}_{\tilde{\theta}_n}\|D_j - D_{j,i}\|^2 = o(1) + O(c_n^2 n^{-1} a_n^{-4} b_n^{-2}).$$

Combining the above with (6.3), we obtain that $\mathrm{E}_{\tilde{\theta}_n}\|(1/\sqrt{n})\sum_{j=1}^n D_j\|^2 \to 0$. Thus (6.13) holds, and this completes the proof. □

7. Adaptive estimation in ARMA models

We shall now apply the construction of the previous section in stationary and ergodic ARMA processes with error model $\mathscr{F} = \mathscr{F}_0$. The resulting estimate will also be adaptive as the necessary condition for adaptation holds in this case. Our construction avoids the sample splitting trick, does not require the knowledge of initial innovations, and shows that some assumptions imposed by Kreiss (1987a), namely (A.5) and positivity of the error density, are not required for the existence of adaptive estimates, nor is symmetry. The latter was also observed by Drost *et al.* (1994). Since the constructions in Kreiss (1987a) and Drost *et al.* (1994) use the initial innovations, they do not solve the question of adaptive estimation for truly stationary and ergodic ARMA models, but only for a closely related model.

For simplicity, we only consider the ARMA(1, 1) process. This process $\{Y_t\colon t \in \mathbb{Z}\}$ is described by the difference equation

$$Y_t - \alpha Y_{t-1} = \eta_t - \beta\eta_{t-1}, \qquad t \in \mathbb{Z},$$

where $\alpha,\ \beta \in (-1, 1)$, $\alpha \neq \beta$, and $\{\eta_t\colon t \in \mathbb{Z}\}$ are independent random variables with common density f. We assume that $f \in \mathscr{F}_0$. Then, for each $t \in \mathbb{Z}$, one has

$$Y_t = \eta_t + (\alpha - \beta)\sum_{i=1}^{\infty}\alpha^{i-1}\eta_{t-i}, \tag{7.1}$$

$$\eta_t = Y_t + (\beta - \alpha)\sum_{i=1}^{\infty}\beta^{i-1}Y_{t-i}, \tag{7.2}$$

$$\sum_{i=0}^{\infty}\beta^i Y_{t-i} = \sum_{i=0}^{\infty}\alpha^i\eta_{t-i}, \tag{7.3}$$

where the series converge almost surely and in mean square. One arrives at (1.1) by setting $X_{-1} = \eta_0$, $X_j = Y_j$ for $j = 0, 1, \ldots,$ and

$$H_j(\vartheta) = (\vartheta_1 - \vartheta_2)\sum_{i=1}^{j-1} \vartheta_2^{i-1} X_{j-i} + \vartheta_2^{j-1}(\vartheta_1 X_0 - \vartheta_2 X_{-1}), \qquad j = 1, 2, \ldots .$$

We take $\mathscr{F} = \mathscr{F}_0$, $\Theta = \{(a, b) \in (-1, 1)^2 \colon a \neq b\}$ and $\theta = (\alpha, \beta)$. The initial density has the following form:

$$g_{(a,b),\phi}(x, y) = \frac{1}{|a-b|}\gamma_{a,\phi}\left(\frac{y-x}{a-b}\right)\phi(x), \qquad x, y \in \mathbb{R}, (a, b) \in \Theta,$$

where $\gamma_{a,\phi}$ is the stationary density of the AR(1) model with parameter a and error density ϕ. It was shown in Remark 2.7 that $(a, \eta) \mapsto \gamma_{a,f_\eta}$ is L_1-continuous at $(\alpha, 0)$ for every smooth path $\eta \to f_\eta$ satisfying (3.7). From this one derives (2.2) and (2.8) for each such path. Consequently, every smooth path satisfying (3.7) is regular.

Let $\dot{H}_j$ denote the gradient of the map $(a, b) \mapsto \sum_{i=1}^{j-1}(a-b)b^{i-1}X_{j-i}$, i.e.,

$$\dot{H}_j(a, b) = \begin{pmatrix} \sum_{i=0}^{j-2} b^i X_{j-1-i} \\ \sum_{i=0}^{j-3} a(i+1)b^i X_{j-2-i} - \sum_{i=0}^{j-2}(i+1)b^i X_{j-1-i} \end{pmatrix},$$

and

$$V_j = \begin{pmatrix} \sum_{i=0}^{\infty}\beta^i Y_{j-1-i} \\ \sum_{i=0}^{\infty}\alpha(i+1)\beta^i Y_{j-2-i} - \sum_{i=0}^{\infty}(i+1)\beta^i Y_{j-1-i} \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{\infty}\alpha^i \eta_{j-1-i} \\ -\sum_{i=0}^{\infty}\beta^i \eta_{j-1-i} \end{pmatrix}.$$

By the ergodic theorem $n^{-1}\sum_{j=1}^{n} V_j \to \mathrm{E}_\theta(V_1)$ and $n^{-1}\sum_{j=1}^{n} V_j V_j^{\mathrm{T}} \to \mathrm{E}_\theta(V_1 V_1^{\mathrm{T}})$ almost surely. Note also that $\mathrm{E}_\theta\|\dot{H}_n(\theta) - V_n\| \to 0$. Using the above and Remark 2.6 one verifies that Assumption 2.2 holds with

$$\nu = \mathrm{E}_\theta V_1 = 0 \text{ and } M = \mathrm{E}_\theta V_1 V_1^{\mathrm{T}} = \sigma^2 \begin{bmatrix} \dfrac{1}{1-\alpha^2} & -\dfrac{1}{1-\alpha\beta} \\ -\dfrac{1}{1-\alpha\beta} & \dfrac{1}{1-\beta^2} \end{bmatrix}.$$

As $\nu = 0$, the necessary condition for adaptive estimation holds.

Utilizing the identities (7.1)–(7.3) one verifies that

$$\dot{H}_j(a, b) = \sum_{i=0}^{j-2}\begin{pmatrix} a^i \\ -b^i \end{pmatrix}\varepsilon_{j-1-i}(a, b) + \begin{pmatrix} (b-a)^{-1}(b^{j-1}aX_0 + ba^{j-1}X_{-1}) \\ (1-j)b^{j-2}(aX_0 - bX_{-1}) \end{pmatrix}.$$

From this one derives that Conditions 5.1 and 6.1 hold for each sequence $\langle m_n \rangle$ tending to infinity. Thus we can use the construction of Section 5. However, we shall slightly modify this construction to obtain an estimator that depends on $X_1, \ldots, X_n$ only, and not on X_0 and X_{-1}. This is necessitated by the fact that one does not observe X_0 and X_{-1} in practice. To this end, set

$$\tilde{\varepsilon}_j(\vartheta) = X_j - \sum_{i=0}^{j-1}(\vartheta_1 - \vartheta_2)\vartheta_2^{i-1}X_{j-i}, \qquad j = 1, 2, \ldots, \vartheta \in \Theta.$$

Note that $\varepsilon_j(\vartheta)$ and $\tilde{\varepsilon}_j(\vartheta)$ differ only by the term $\vartheta_2^{j-1}(\vartheta_1 X_0 - \vartheta_2 X_{-1})$. Now let $\tilde{\theta}_n$ be an estimate of θ based on $X_1, \ldots, X_n$ only and define the estimate

$$\tilde{\theta}_n^* = \tilde{\theta}_n + \frac{1}{N_n} \sum_{j=d_n}^{n} \tilde{I}_{n,*}^{-1} (\dot{H}_{n,j} - \hat{\nu}_n) \tilde{\ell}_n(\tilde{\varepsilon}_{n,j}), \tag{7.4}$$

where $\dot{H}_{n,j}$ and $\hat{\nu}_n$ are as in Section 6, $\tilde{\varepsilon}_{n,j} = \tilde{\varepsilon}_j(\tilde{\theta}_n)$, $\tilde{\ell}_n$ is defined as $\hat{\ell}_n$ but with $\tilde{\varepsilon}_{n,j}$ replacing $\varepsilon_{n,j}$, and

$$\tilde{I}_{n,*} = \frac{1}{N_n} \sum_{j=d_n}^{n} \tilde{\ell}_n^2(\tilde{\varepsilon}_{n,j}) \frac{1}{N_n} \sum_{j=d_n}^{n} (\dot{H}_{n,j} - \hat{\nu})(\dot{H}_{n,j} - \hat{\nu}_n)^{\mathrm{T}}.$$

Theorem 7.1. *Suppose $\langle \tilde{\theta}_n \rangle$ is a discretized $\sqrt{n}$-consistent estimator of θ and the sequences $\langle a_n \rangle$, $\langle b_n \rangle$, $\langle c_n \rangle$ and $\langle d_n \rangle$ are such that*

$$n^{-1} c_n^2 (a_n^{-4} b_n^{-2} + a_n^{-3} b_n^{-1} d_n) \to 0 \quad \text{and} \quad a_n^{-5} b_n^{-1} \rho^{2d_n} \to 0 \quad \text{for every } \rho \in (0, 1). \tag{7.5}$$

Then $\langle \tilde{\theta}_n^* \rangle$ defined by (7.4) is adaptive.

Proof. Note that the assumptions of Theorem 6.2 hold with $m_n = d_n$. Thus we only need to show that $\tilde{\theta}_n^* - \hat{\theta}_n = o_\theta(n^{-1/2})$. As in the proof of Theorem 6.2, we can assume that $\langle \tilde{\theta}_n \rangle$ is a local sequence. Let θ_n^* be defined as $\tilde{\theta}_n^*$ but with $\varepsilon_{n,j}$ replacing $\tilde{\varepsilon}_{n,j}$. It is easy to prove that $\theta_n^* - \hat{\theta}_n = o_{\tilde{\theta}_n}(n^{-1/2})$. Easy calculations show that $U_n = \sum_{j=d_n}^{n} \mathrm{E}_{\tilde{\theta}_n}(\tilde{\varepsilon}_{n,j} - \varepsilon_{n,j})^2 = O(\rho^{2d_n})$ for some $\rho \in (0, 1)$. From this and Lemma 10.1 in Schick (1993) one obtains that

$$\sum_{j=d_n}^{n} \mathrm{E}_{\tilde{\theta}_n}(\tilde{\ell}_n(\tilde{\varepsilon}_{n,j}) - \hat{\ell}_n(\varepsilon_{n,j}))^2 \leqslant c_0(a_n^{-4} + a_n^{-5} b_n^{-1}) U_n \to 0.$$

This lets us conclude that $\tilde{\theta}_n^* - \theta_n^* = o_{\tilde{\theta}_n}(n^{-1/2})$. The desired result now follows from contiguity. □

Remark 7.2. Of course, $\sqrt{n}$-consistent estimates do exist for θ and can be calculated from the sample autocovariances.

8. Appendix

Let λ denote the Lebesgue measure. Let v be a measurable function from $\mathbb{R}$ to $[1, \infty)$. Let μ denote the measure with Lebesgue density v. Let Ξ be an open subset of $\mathbb{R}^m$ containing 0. Let $\{h_\xi : \xi \in \Xi\}$ be a collection of measurable functions from $\mathbb{R}$ to $\mathbb{R}$ such that for some $A > 0$ and all $\xi \in \Xi$ and $x \in \mathbb{R}$:

(1) $|h_\xi(x)| \leqslant A v(x)$;
(2) $|h_\xi(x) - h_0(x)| \leqslant A \|\xi\| v(x)$.

Let $\{f_\xi\colon \xi \in \Xi\}$ be a family of Lebesgue probability densities such that

$$\int |f_\xi - f_0|\, d\lambda \to 0. \tag{8.1}$$

For every $\xi \in \Xi$, define a bounded linear operator T_ξ from $L_1(\mu)$ to $L_1(\lambda)$ by

$$T_\xi g(x) = \int f_\xi(x - h_\xi(y)) g(y)\, d\lambda(y).$$

Recall that the norm of the operator T_ξ is defined by

$$\|T_\xi\| = \sup\left\{\int |T_\xi g|\, d\lambda\colon\ g \in L_1(\mu),\ \int |g|\, d\mu = 1\right\}.$$

Lemma 8.1. *The operators* $\{T_\xi\colon \xi \in \Xi\}$ *are contractions, i.e.,* $\|T_\xi\| \leqslant 1$, *and are norm-continuous at 0, i.e.,* $\lim_{\xi\to 0} \|T_\xi - T_0\| = 0$.

Proof. The former follows from the bound $\int |T_\xi g|\, d\lambda \leqslant \int |g|\, d\lambda \leqslant \int |g|\, d\mu$. For the latter, fix a positive constant K and derive the bound

$$\|T_\xi - T_0\| \leqslant \int |f_\xi - f_0|\, d\lambda + \sup_{|t|\leqslant K\|\xi\|} \int |f_0(x - t) - f_0(x)|\, d\lambda(x) + \frac{2A}{K},$$

where we use the fact that $\{|h_\xi - h_0| > K\|\xi\|\} \subset \{Av > K\}$ for $\xi \neq 0$. Now use (8.1) and the pointwise continuity of the translation operator on $L_1(\lambda)$ – see Theorem 9.5 in Rudin (1974) – to conclude that $\lim_{\xi\to 0} \|T_\xi - T_0\| \leqslant 2A/K$. As K was arbitrary, this yields the desired result. □

Lemma 8.2. *Suppose for each* $\xi \in \Xi$ *there exists a unique Lebesgue probability density* g_ξ *such that* $T_\xi g_\xi = g_\xi$. *Then* $\sup\{\int |g_\xi|\, d\mu\colon \xi \in \Xi\} < \infty$ *implies that* $\lim_{\xi\to 0} \int |g_\xi - g_0|\, d\lambda = 0$.

Proof. Choose a sequence $\{\xi_n\}$ in Ξ which converges to 0. We shall show that

$$\sup_n \int |T_{\xi_n} g_{\xi_n}(x)|\, d\lambda(x) < \infty, \tag{8.2}$$

$$\lim_{t\to 0} \sup_n \int |T_{\xi_n} g_{\xi_n}(x + t) - T_{\xi_n} g_{\xi_n}(x)|\, d\lambda(x) = 0, \tag{8.3}$$

$$\lim_{K\to\infty} \sup_n \int_{|x|\geqslant 2K} |T_{\xi_n} g_{\xi_n}(x)|\, d\lambda(x) = 0. \tag{8.4}$$

The Fréchet–Kolmogorov theorem (see Yosida 1971, p. 275) implies that the sequence $\{T_{\xi_n} g_{\xi_n}\}$ is sequentially compact in $L_1(\lambda)$. Thus, in view of the identity $T_{\xi_n} g_{\xi_n} = g_{\xi_n}$, the sequence $\{g_{\xi_n}\}$ is sequentially compact in $L_1(\lambda)$. Let g be an $L_1(\lambda)$ limit point of this sequence. Without loss of generality, assume that $\int |g_{\xi_n} - g|\, d\lambda \to 0$, otherwise consider a

subsequence. By Lemma 8.1, $\int |T_{\xi_n} g_{\xi_n} - T_0 g| \, d\lambda \to 0$. This leads to the identity $g = T_0 g$. As g is a Lebesgue probability density, one obtains $g = g_0$.

Let us now show that (8.2)–(8.4) hold. Clearly, (8.2) holds. The statements (8.3) and (8.4) follow from the bounds

$$\int |T_\xi g(x+t) - T_\xi g(x)| \lambda(x) \leqslant \int |g| \, d\mu \int |f_\xi(x+t) - f_\xi(x)| \, d\lambda(x),$$

$$\begin{aligned}\int_{|x| \geqslant 2K} |T_\xi g(x)| \, d\lambda(x) &\leqslant \int_{|x| \geqslant 2K} \int f_\xi(x - h_\xi(y)) |g(y)| \, d\lambda(y) \, d\lambda(x) \\ &\leqslant \int_{|x| \geqslant K} f_\xi(x) \, d\lambda(x) \int |g| \, d\lambda + \int_{|h_\xi(y)| \geqslant K} |g(y)| \, d\lambda(y) \\ &\leqslant \int_{|x| \geqslant K} f_\xi(x) \, d\lambda(x) \int |g| \, d\mu + \frac{A}{K} \int |g| \, d\mu,\end{aligned}$$

and the following facts

$$\lim_{t \to 0} \sup_n \int |f_{\xi_n}(x+t) - f_{\xi_n}(x)| \, d\lambda(x) = 0,$$

$$\lim_{K \to \infty} \sup_n \int_{|x| > K} |f_{\xi_n}(x)| \, d\lambda(x) = 0,$$

which are consequences of (8.1). □

Acknowledgements

Koul's research was partially supported by NSF grant DMS-9402904; Schick's research was partially supported by NSF grant DMS-9206138. Thanks are due to the two referees for their constructive comments.

References

Akritas, M.G. and Johnson, R.A. (1980) Efficiencies of tests and estimators of p-order autoregressive processes when the error distribution is nonnormal. *Ann. Inst. Statist. Math.*, **34**, 579–589.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

Chan, K.S. and Tong, H. (1985) On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. *Adv. Appl. Probab.*, **17**, 666–678.

Chan, K.S., Petrucelli, J.D., Tong, H. and Woolford, S.W. (1985) A multiple-threshold AR(1) model. *J. Appl. Probab.*, **22**, 267–279.

Drost, F.C., Klaassen, C.A.J. and Werker, B.J.M. (1993) Adaptiveness in time series models. In P. Mandl and M. Hušková (eds), *Asymptotic Statistics, Proceedings of the Fifth Prague Symposium*, pp. 203–211. Heidelberg: Physica-Verlag.

Drost, F.C., Klaassen, C.A.J. and Werker, B.J.M. (1994) Adaptive estimation in time-series models. Discussion Paper Series, No. 9488, CentER, Tilburg University, The Netherlands.

Fabian, V. and Hannan, J. (1982) On estimation and adaptive estimation for locally asymptotically normal families. *Z. Wahrscheinlichkeitstheorie Verw. Geb.,* **59**, 459–478.

Fabian, V. and Hannan, J. (1987) Local asymptotic behavior of densities. *Statist. Decisions*, **5**, 105–138. Correction: **6** (1988), 195.

Hájek, J. and Šidák, Z. (1967) *Theory of Rank Tests*. New York: Academic Press.

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

Hwang, S.Y. and Basawa, I.V. (1993) Asymptotic optimal inference for a class of nonlinear time series. *Stochastic Process. Appl*., **46**, 91–114.

Jeganathan, P. (1995) Some aspects of asymptotic theory with applications to time series models. *Econometric Theory*, **11**, 818–887.

Klimko, L.A. and Nelson, P.I. (1978) On conditional least squares estimation for stochastic processes. *Ann. Statist*., **6**, 629–642.

Kreiss, J.-P. (1987a) On adaptive estimation in stationary ARMA processes. *Ann. Statist*. **15**, 112–133.

Kreiss, J.-P. (1987b) On adaptive estimation in autoregressive models when there are nuisance functions. *Statist. Decisions*, **5**, 59–76.

Koul, H.L. and Pflug, G. (1990) Weakly adaptive estimators in explosive autoregression. *Ann. Statist*., **18**, 939–960.

Le Cam, L. (1960) Locally asymptotically normal families of distributions. *Univ. California Publ. Statist*., **3**, 37–98.

Petrucelli, J.D. and Woolford, S.W. (1984) A threshold AR(1) model. *J. Appl. Probab*., **21**, 270–286.

Rudin, W. (1974) *Real and Complex Analysis* (2nd edition). New York: McGraw-Hill.

Schick, A. (1986) On asymptotically efficient estimation in semi-parametric models. *Ann. Statist*., **14**, 1139–1151.

Schick, A. (1987) A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference*, **16**, 89–105. Correction: **22** (1989), 269–270.

Schick, A. (1988) On estimation in LAMN families when there are nuisance parameters present. *Sankhyá Ser. A*, **50**, 249–268.

Schick, A. (1993) On efficient estimation in regression models. *Ann. Statist.*, **21**, 1486–1521.

Stein, C. (1956) Efficient nonparametric estimation and testing. In J. Neyman (ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 187–195. Berkeley: University of California Press.

Swensen, A.R. (1985) The asymptotic distribution of the likelihood ratio for autoregressive time series with a regression trend. *J. Multivariate Anal*., **16**, 54–70.

Tjøstheim, D. (1986) Estimation in nonlinear time series models. *Stochastic Process. Appl.*, **21**, 251–273.

Tong, H. (1990) *Nonlinear Time Series: A Dynamical Approach*. New York: Oxford University Press.

Tweedie, R.L. (1983) The existence of moments for stationary Markov chains. *J. Appl. Probab.,* **20**, 191–196.

Yosida, K. (1971) *Functional Analysis*. Berlin: Springer-Verlag.

Received March 1995 and revised December 1996.