– PANTHÉON - SORBONNE –
UNIVERSITÉ PARIS 1

Centre de Recherche

SAMOS

Statistique Appliquée &
Modélisation Stochastique

On least squares estimation for stable nonlinear AR processes

J.F. YAO

Prépublication du SAMOS N° 67 (Reversed version)

Novembre 1998

On least squares estimation for stable nonlinear AR processes

Jian-feng YAO

SAMOS, Université Paris I, Paris

Running head : Least squares for nonlinear AR processes

Correspondence address:

J.YAO,
SAMOS, Université Paris I,
90 rue de Tolbiac
F-75634 Paris Cedex 13
FRANCE
(e-mail: yao@univ-paris1.fr)

To appear in: *A.I.S.M*

Abstract

Following a Markov chain approach, this paper establishes asymptotic properties of the least squares estimator in nonlinear autoregressive (NAR) models. Based on conditions ensuring the stability of the model and allowing the use of a strong law of large number for a wide class of functions, our approach improves some known results on strong consistency and asymptotic normality of the estimator. The exact convergence rate is established by a law of the iterated logarithm. Based on this law and a generalized Akaike's information criterion, we build a strongly consistent procedure for selection of NAR models. Detailed results are given for familiar nonlinear AR models like exponential AR models, threshold models or multilayer feedforward perceptrons.

AMS Classification Code : 62 F 12, 62 M 10

Keywords: Nonlinear AR process; least squares estimation; law of the iterated logarithm; model selection; multilayer perceptron.

1 Introduction

Nonlinear models have become a standard tool for analysis of time series endowed with a complex dynamic. An important and widely used subclass is $\mathbb{R}^d$-valued NAR(p) processes defined by

$$X_t = f(X_{t-1}, \dots, X_{t-p}; \theta) + \varepsilon_t, \qquad t \geq 1. \tag{1.1}$$

Here (ε_t) is an i.i.d. error process and f a known measurable function depending on some parameter θ. As extension of familiar AR models, NAR(p) models include threshold AR models (Tong, 1983), exponential AR models (Haggan and Ozaki, 1981) or multilayer feedforward perceptrons, among others. Recent reviews on these models can be found in (Tong, 1990) and (Pötscher and Prucha, 1997).

For parameter estimation purpose, least squares estimation (LS) is a mostly used procedure for these models. Although the method is classical and well-known, its theoretical properties have been reported in recent years only. The literature can roughly be classified according to the type of dependence properties of the process which are exploited to derive the asymptotics of the LS estimator. One approach, based on stationarity and ergodicity conditions is proposed by (Tjøstheim, 1986). Another approach is based on mixing conditions or uniform mixing, see e.g. (White and Domowitz, 1984). A third approach, based on the concept of "near epoch dependence" and "L_p-approximability" of a stochastic process, has been proposed, see e.g. (Gallant, 1987; Gallant and White, 1988; Pötscher and Prucha, 1997). A fourth approach treats the process as a Markov chain (after rewriting it in companion form), and uses limit theorems for Markov chains. This approach seems to be initiated by (Tjøstheim, 1990).

In this paper, we follow the Markov chain approach and try to improve some results of (Tjøstheim, 1986) in the following way. While previous results require the associated Markov chain to be Harris ergodic (i.e. positive Harris recurrent and aperiodic), it is known that this requirement is stronger than needed for asymptotics of the LS estimator. Actually, we shall show that strong consistency and asymptotic normality both hold under a weaker condition, called *stability of order* a. Roughly speaking, such a stability holds when the chain has an unique invariant measure having moments up to order a and such that a strong law of large numbers holds for functions which are bounded at infinity by the moment function $|\mathrm{x}|^a$. As expected, a Harris ergodic chain with a suitable moment condition fulfills such stability condition. However, we shall show examples where we are able to establish asymptotic properties of the LS estimator, although the associated Markov chain is not Harris ergodic. On the other hand, for exponential AR models or threshold AR models, we found conditions on the error process which seem to be weaker than previously used ones.

Another contribution in this paper is a law of the iterated logarithm for the LS estimator which is established under the above stability condition. This law is not of theoretical interest only: it has an important application in building a strongly consistent procedure for selection of NAR(p) models.

It is worth noting that in the specific case of NAR(p) models, application of general results on conditional LS estimation as proposed in (Klimko and Nelson, 1978), and especially (Lai, 1994) is not obvious. Actually, the conditions given in these papers need to be explicited in such a way that they depend only on the regression function f and the error process.

An overview of the paper is as follows. In Section 2, the main assumption of stability of order a is stated and various known criteria are recalled for checking this stability. Section 3 is devoted to asymptotic properties of the LS estimator, including a law of the iterated logarithm (LIL). By using this LIL and Akaike's principle of parsimony, we give a strongly consistent procedure for selection of NAR models (Section 4). To illustrate our results, we treat some important examples in Section 5. Finally, Section 6 collects all proofs.

2 Stability of order a for the associated Markov chain

Let be $\mathbf{X}_t = (X^{\mathrm{T}}{}_t, \ldots, X^{\mathrm{T}}{}_{t-p+1})^{\mathrm{T}}$. Following the Markov chain approach (Tjøstheim, 1990), we rewrite the NAR(p) process in its companion form

$$\mathbf{X}_t = \begin{pmatrix} X_t \\ X_{t-1} \\ \vdots \\ X_{t-p+1} \end{pmatrix} = \begin{pmatrix} f(X_{t-1}, \ldots, X_{t-p}\,; \theta) \\ X_{t-1} \\ \vdots \\ X_{t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} =: F(\mathbf{X}_{t-1}\,; \theta) + \eta_t, \qquad t \geq 1 \tag{2.1}$$

where F and η are implicitly defined. Since (ε_t) is an i.i.d. sequence, the vectorized process $(\mathbf{X}_t)$ is an homogeneous Markov chain with initial (deterministic) state $\mathrm{x}_0 = (x^{\mathrm{T}}{}_0, \ldots, x^{\mathrm{T}}{}_{-p+1})^{\mathrm{T}} \in (\mathbb{R}^d)^p$.

Some notations are necessary. To any norm $||\cdot||$ on $\mathbb{R}^d$, we associate a norm on $(\mathbb{R}^d)^p$ by setting $|\mathrm{x}| := ||x_1|| + \cdots + ||x_p||$ for $\mathrm{x} := (x^{\mathrm{T}}{}_1, \ldots, x^{\mathrm{T}}{}_p)^{\mathrm{T}}$ in $(\mathbb{R}^d)^p$, where x_i are vectors in $\mathbb{R}^d$. The true value of the parameter is denoted θ_0 and $\mathbb{P}_{\theta_0}$ stands for the probability distribution of the chain $(\mathbf{X}_t)$ under the true model. Moreover, any convergence $\xrightarrow{a.s.}$ (resp. $\xrightarrow{\mathcal{D}}$) means an a.s. convergence (resp. convergence in distribution) under $\mathbb{P}_{\theta_0}$ which hold independently of the initial state x_0.

A basic tool for deriving asymptotic properties of the LS estimator is to exploit limit theorems of the Markov chain $(\mathbf{X}_t)$. As stated in Section 1, previous results mostly required Harris ergodicity for this chain. We shall show that an asymptotic theory is possible under the following weaker condition called *stability of order* a.

Definition [S]: stability of order a. *Let be $a \geq 1$. We say that under $\mathbb{P}_{\theta_0}$, the chain $(\mathbf{X}_t)$ has a stability of order a if*

(i) *The chain has an unique invariant measure μ_{θ_0}.*

(ii) Moment conditions. *The marginal distributions of $\mathbf{X}_t$, $t \geq 1$ as well as the invariant measure μ_{θ_0} has a moment of order a, that is*

$$\mathbb{E}_{\theta_0}\ |\mathbf{X}_t|^a < \infty,\ t \geq 1; \qquad \mu_{\theta_0}(|\cdot|^a) := \int_{(\mathbb{R}^d)^p} |\mathrm{x}|^a \mu_{\theta_0}(d\mathrm{x}) < \infty.$$

(iii) Strong law of large numbers (SLLN). *For any scalar function ϕ on $(\mathbb{R}^d)^p$ which is μ_{θ_0}-a.s. continuous and satisfies $|\phi(\cdot)| \leq$ const. $(1 + |\cdot|^a)$, it holds*

$$\frac{1}{n}\sum_{t=1}^{n} \phi(\mathbf{X}_t) \xrightarrow{a.s.} \int_{(\mathbb{R}^d)^p} \phi(\mathrm{x})\mu_{\theta_0}(d\mathrm{x}). \qquad \square$$

In a model where such a stability holds, the above SLLN will be fundamental to derive asymptotic properties of the LS estimator. Actually, we shall successively apply this law to the LS criterion function and its first and second-order derivatives.

Consequently, model assumptions will be set in such a way that these functions are bounded by a polynomial of type const. $(1 + |\cdot|^a)$.

An immediate question from this definition is to find conditions on a NAR(p) model to guarantee such a stability. For general stability theory of Markov chains, we refer to representative monographs (Meyn and Tweedie, 1993; Duflo, 1997) and papers from (Borovkov, 1991; Borovkov and Korshunov, 1993). Here we shall emphasize on specific criteria for a NAR(p) model. A clear classification of existing criteria can be obtained according to whether or not the error process is a *Lebesgue noise*: we shall call an i.i.d. error sequence (ε_t) a Lebesgue noise if ε_1 has an everywhere positive density function with respect to the Lebesgue measure.

Criterion [C.1] for a Lebesgue noise. *Assume for the NAR(p) model (1.1)*

(i) (ε_t) *is a Lebesgue noise and* $\mathbb{E}||\varepsilon_1||^a < \infty$ *for some* $a \geq 1$.

(ii) The function $\mathrm{x} \mapsto f(\mathrm{x};\theta_0)$ *is continuous and there exists positive numbers* $\lambda_1, \ldots, \lambda_p$ *satisfying* $\lambda_1 + \cdots + \lambda_p < 1$, *and a constant* $\kappa \geq 0$ *such that for some norm* $||\cdot||$ *on* $\mathbb{R}^d$,

$$||f(\mathrm{x};\theta_0)|| \leq \lambda_1||x_1|| + \cdots + \lambda_p||x_p|| + \kappa, \quad \mathrm{x} \in (\mathbb{R}^d)^p.$$

Then, the NAR(p) model under θ_0 *has the stability* **[S]** *of order* a. □

In the case $p = 1$, the criterion **[C.1]** is well-known, see e.g. (Doukhan and Ghindès, 1980; Mokkadem, 1987; Tjøstheim, 1990). Based on Tweedie's results, these authors proved that under **[C.1]**, the chain $(\mathbf{X}_t)$ is Harris ergodic with an (unique) invariant measure μ_{θ_0} equivalent to Lebesgue measure. The moment condition in **[C.1-i]** ensures that $\mu_{\theta_0}(|\cdot|^a) < \infty$. The required SLLN thus follows from e.g. Theorem 17.1.7 in (Meyn and Tweedie, 1993). Extensions for general $p > 1$ are recent. We are only aware of results from (Duflo, 1997) and (Attali, 1998).

However, Condition **[C.1-ii]** is an approriate criterion only for those models which are basically nonlinear. To specify, assume in contrary $f(\mathrm{x};\theta_0)$ is close to a linear model in the sense that

$$f(\mathrm{x};\theta_0) = a_1x_1 + \cdots + a_px_p + \varphi(\mathrm{x}\,;\theta_0)\ ,$$

where φ is a *small* nonlinear component satisfying

$$\limsup_{|\mathrm{x}|\to\infty} \frac{||\varphi(\mathrm{x})||}{|\mathrm{x}|} = 0,$$

($\varphi = 0$ corresponds to AR models). For such models, Condition **[C.1-ii]** is too strong. Fortunately, the conclusion of **[C.1]** still holds if we replace **[C.1-ii]** by the following

(ii)' The function $\mathrm{x} \mapsto f(\mathrm{x};\theta_0)$ *is continuous and the polynomial* $1 - \sum_j a_j z^j$ *is causal.*

If the error process is no longer a Lebesgue noise, the situation is more intricate and very few is known. In general, a stronger contraction condition on f is necessary to ensure stability. The following Lipschitz condition is found in (Duflo, 1997).

Criterion [C.2] for arbitrary noise. *Assume for the NAR(p) model (1.1)*

(i) $\mathbb{E}||\varepsilon_1||^a < \infty$ *for some* $a \geq 1$.

(ii) *there are* p *positive numbers* $\lambda_1, \ldots, \lambda_p$ *such that* $\lambda_1 + \cdots + \lambda_p < 1$*, and for some norm* $||\cdot||$ *on* $\mathbb{R}^d$*,*

$$||f(\mathbf{x};\theta_0) - f(\mathbf{y};\theta_0)|| \leq \lambda_1||x_1 - y_1|| + \cdots + \lambda_p||x_p - y_p||, \quad \mathbf{x}, \mathbf{y} \in (\mathbb{R}^d)^p.$$

Then, the NAR(p) model under θ_0 *has the stability* **[S]** *of order* a. □

For illustration purpose, consider the following univariate AR(1) model

$$X_t = \frac{1}{2}X_{t-1} + \varepsilon_t, \quad t \geq 1$$

started with $X_0 = 0$ and where (ε_t) is an i.i.d. Rademacher sequence, i.e. $\mathbb{P}(\varepsilon_t = 1) = \mathbb{P}(\varepsilon_t = -1) = \frac{1}{2}$. It is known that (X_t) is not Harris ergodic. However, by applying Criterion **[C.2]**, we see that the model is stable with an order which can be arbitrarily high. Hence our asymptotic results on LS estimator are valid in such a case.

3 Asymptotic properties of the LS estimator

In the sequel, $||\cdot||$ denotes the usual Euclidian norm with associated inner product $\langle\cdot,\cdot\rangle$. Let $(X_{-p+1}, \ldots, X_n)$ be observations from the NAR(p) model (1.1). The (normalized) sum of squares (U_n) is

$$U_n(\theta) := \frac{1}{n}\sum_{t=1}^{n} ||X_t - f(X_{t-1}, X_{t-2}, ..., X_{t-p}\,; \theta)||^2. \tag{3.1}$$

and the *LS estimator* is defined by

$$\widehat{\theta}_n := \text{Arg}\min_{\theta\in\Theta} U_n(\theta). \tag{3.2}$$

We shall derive successively strong consistency, asymptotic normality and a law of the iterated logarithm for this estimator.

3.1 Strong consistency

We shall call *continuity modulus* any increasing function g satisfying $\lim_{x\to 0} g(x) = g(0) = 0$. Let us make the following assumptions.

ASSUMPTION **[M]**:

(i) *The parameter θ belongs to a compact subset Θ in $\mathbb{R}^s$. The error process $(\varepsilon_t)_{t>0}$ is centered and i.i.d., with a known covariance matrix Γ.*

(ii) *Under the true model θ_0, the Markov chain $(\mathbf{X}_t)$ has a stability of order $a \geq 1$ according to Definition* **[S]**.

(iii) *(a). For all θ, $\mathrm{x} \mapsto f(\mathrm{x}\,;\theta)$ is μ_{θ_0}-a.s. continuous. (b). $||f(\mathrm{x}\,;\theta_0)|| \leq$ const. $(1 + |\mathrm{x}|^{a/2})$. (c). There exists a continuity modulus G such that:*

$$\forall x \in (\mathbb{R}^d)^p, \quad \forall(\alpha,\beta) \in \Theta^2, \quad ||f(\mathrm{x}\,;\alpha) - f(\mathrm{x}\,;\beta)|| \leq G(||\alpha-\beta||)(1+|\mathrm{x}|^{a/2}). \qquad \Box$$

Condition (i) is standard. Condition (ii) is the basic requirement that we need on the stochastic behaviour of the true model. Condition (iii) roughly means that the autoregression function f is continuous, and with respect to x it is bounded by $1 + |\mathrm{x}|^{a/2}$ (up to a constant factor). Such a control together with the stability assumption (ii) guarantee a SLLN for functions like $||f||$ or $||f||^2$.

First we identify the limit of the estimating function U_n.

Proposition 1 *Assume that* **[M]** *holds. Then for any fixed $\theta \in \Theta$,*

$$U_n(\theta) - U_n(\theta_0) \xrightarrow{a.s.} \int_{(\mathbb{R}^d)^p} ||f(\mathrm{x}\,;\theta) - f(\mathrm{x}\,;\theta_0)||^2 \mu_{\theta_0}(d\mathrm{x}) =: K(\theta,\theta_0) \quad . \tag{3.3}$$

Moreover, the limit function $K(\theta,\theta_0)$ is continuous in θ.

Clearly, θ_0 is a global minimum of the limit function K. Whether or not it is the unique one depends on the identifiability of the model. We shall use the following

CONDITION OF IDENTIFIABILITY **[D]**:

$$\textit{for any} \quad \theta \in \Theta, \quad f(\cdot\,;\theta) = f(\cdot\,;\theta_0) \;\; \mu_{\theta_0} - \textit{a.s.} \quad \textit{implies} \quad \theta = \theta_0. \quad \Box.$$

The LS estimator is then strongly consistent if both the assumptions **[M]** and the identifiability condition **[D]** hold.

Theorem 1 *Assume that both Conditions* **[M]** *and* **[D]** *hold. Then the least squares estimator $\widehat{\theta}_n$ is strongly consistent.*

3.2 Asymptotic normality

For asymptotic normality of $\widehat{\theta}_n$, we typically need some additional conditions on second order differentiability of the process (U_n). We make the following assumptions, where partial derivatives of a scalar function $g(\theta)$ are denoted $D_i g = \partial g/\partial\theta_i$, $D^2_{ij}g = \partial^2 g/(\partial\theta_i\partial\theta_j)$.

ASSUMPTION **[N]** *Assume that $\theta_0 \in \overset{\circ}{\Theta}$ and there exists a neighbourhood V of θ_0 where for any $\mathrm{x} \in (\mathbb{R}^d)^p$, the d coordinate functions $f_1, \dots, f_d$ of $\theta \mapsto f(\mathrm{x}\,;\theta)$ are twice continuously differentiable such that, for all $k = 1, \dots, d$ and $i, j = 1, \dots, s$, we have:*

(i) *for all $\theta \in V$, $\mathrm{x} \mapsto D_i f_k(\mathrm{x}\,;\theta)$ and $\mathrm{x} \mapsto D^2_{ij} f_k(\mathrm{x}\,;\theta)$ are μ_{θ_0}-a.s. continuous.*

(ii) $|D_i f_k(\mathrm{x}\,;\theta_0)| \le$ *const.* $(1 + |\mathrm{x}|^{a/2})$, $\quad \left|D^2_{ij} f_k(\mathrm{x}\,;\theta_0)\right| \le$ *const.* $(1 + |\mathrm{x}|^{a/2})$, $\quad \mathrm{x} \in (\mathbb{R}^d)^p$.

(iii) *there exists a continuity modulus σ_{ijk} such that*

$$\left|D^2_{ij} f_k(\mathrm{x}\,;\theta) - D^2_{ij} f_k(\mathrm{x}\,;\theta_0)\right| \le \sigma_{ijk}(||\theta-\theta_0||)(1+|\mathrm{x}|^{a/2}), \qquad \theta \in V,\ \mathrm{x} \in (\mathbb{R}^d)^p. \qquad \square \tag{3.4}$$

It is worth noting that Conditions **[N]**-(ii)-(iii) are similar to **[M]**-(iii). In particular, they also guarantee a SLLN for functions involving first or second order derivatives of f.

Let us denote the matrices:

$$\begin{aligned} Df(\mathrm{x}\,;\theta) &:= [D_j f_k(\mathrm{x}\,;\theta)]\,, && d \times s \text{ matrix,} \\ M(\mathrm{x}\,;\theta) &:= \{Df(\mathrm{x}\,;\theta)\}^{\mathrm{T}} Df(\mathrm{x}\,;\theta), && s \times s \text{ matrix,} \qquad (3.5) \\ D^2_{ij} f(\mathrm{x}\,;\theta) &:= \left[D^2_{ij} f_k(\mathrm{x}\,;\theta)\right]\,, && d \times 1 \text{ vector,} \qquad (3.6) \end{aligned}$$

with $1 \le i, j \le s$ and $1 \le k \le d$. The gradient vector and the Hessian matrix of U_n are respectively:

$$DU_n(\theta) = -\frac{2}{n} \sum_{0 \le t < n} [X_{t+1} - f(\mathbf{X}_t\,;\theta)]^{\mathrm{T}} Df(\mathbf{X}_t\,;\theta), \tag{3.7}$$

$$\frac{1}{2} D^2 U_n(\theta) = \frac{1}{n} \sum_{0 \le t < n} M(\mathbf{X}_t\,;\theta) - \frac{1}{n} \left[\sum_{0 \le t < n} [X_{t+1} - f(\mathbf{X}_t\,;\theta)]^{\mathrm{T}} D^2_{ij} f(\mathbf{X}_t\,;\theta) \right]_{1 \le i,j \le s} \tag{3.8}$$

First we prove two results on $[DU_n(\theta_0)]$ and $[D^2 U_n(\theta_0)]$.

Proposition 2 *Assume that Conditions* **[M]**, **[D]** *and* **[N]** *hold. Then*

$$D^2 U_n(\theta_0) \xrightarrow{a.s.} I_0 \quad \textit{with} \quad I_0 := 2 \int_{(\mathbb{R}^d)^p} M(\mathrm{x}\,;\theta_0)\mu_{\theta_0}(d\mathrm{x}), \tag{3.9}$$

$$\sqrt{n} DU_n(\theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_0) \quad \textit{with} \quad J_0 := 4 \int_{(\mathbb{R}^d)^p} \{Df(\mathrm{x}, \theta_0)\}^T \Gamma Df(\mathrm{x}, \theta_0)\mu_{\theta_0}(d\mathrm{x}) \tag{3.10}$$

We now establish the asymptotic normality of the LS estimator.

Theorem 2 *Assume that Conditions* **[M]**, **[D]**, **[N]** *hold and in addition* I_0 *is regular. Then*

$$\sqrt{n}\left[\widehat{\theta}_n - \theta_0\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_0^{-1} J_0 I_0^{-1}).$$

One may note that in the univariate case ($d = 1$), the two matrices I_0 and J_0 are proportional: $J_0 = 2\sigma^2 I_0$ with the noise variance $\sigma^2 = \Gamma$. In this case, the above asymptotic covariance matrix is reduced to $2\sigma^2 I_0^{-1}$. It is also worth noting that Theorem 2 can be applied to subhypothesis testing.

3.3 A law of the iterated logarithm

The following law of the iterated logarithm gives exact a.s. convergence rate of $\widehat{\theta}_n$. In addition of its own interest, such a law would be a basic step in search of a strongly consistent procedure for selection of NAR models (see Section 4).

Theorem 3 *Assume*

(i) Conditions **[M]**, **[D]** *and* **[N]** *hold with some* $a \geq 1$ *where the condition **[M]**-(ii) is strenthened with* a *replaced by some* $a' > a$.

(ii) Both the matrices I_0 *and* J_0 *are regular.*

Then, for all $u \in \mathbb{R}^s$, $u \neq 0$, *it holds a.s.*

$$\limsup_n \sqrt{\frac{n}{2\log\log n}} \langle DU_n(\theta_0), u\rangle = \sqrt{u^T J_0 u} = -\liminf_n \sqrt{\frac{n}{2\log\log n}} \langle DU_n(\theta_0), u\rangle, \tag{3.11}$$

$$\limsup_n \sqrt{\frac{n}{2\log\log n}} \langle \widehat{\theta}_n - \theta_0, u\rangle = \sqrt{u^T I_0^{-1} J_0 I_0^{-1} u} = -\liminf_n \sqrt{\frac{n}{2\log\log n}} \langle \widehat{\theta}_n - \theta_0, u\rangle. \tag{3.12}$$

4 A strongly consistent procedure for selection of NAR models

For model selection, (Akaike, 1969) and (Schwarz, 1978) introduced the method of penalized quasi-likelihood. There is a huge literature on selection of linear models, see e.g. (Hannan, 1980; Quinn, 1980; Tsay, 1984). In contrast, few well-established results are known for nonlinear models, despite the widely-spread use of the method in practice. Some related works can be found in (Nishii, 1984; Haughton, 1991). An approach based on the accumulated prediction errors has been recently proposed by (Lai and Lee, 1997). We establish below the strong consistency of a generalized information criterion based on the LS estimates.

Let us denote by $\lambda_{max}A$ and $\lambda_{min}A$ the greatest and the smallest eigenvalue of a real symetric matrix A, respectively. Here we follow the presentaion given in (Guyon, 1995) and consider a generalized information criterion defined in Eqns. (3.17)-(3.18) there with the sum of squares $U_n(\theta)$. Therefore $[c(n)]$ denotes some penalization rate and $\widehat{\delta}_n$ the selected model based on the observations $(X_t)_{-p<t\leq n}$. Since a law of the iterated logarithm is established for the LS estimator (Theorem 3), straightforward application of Theorem (3.4.8) from (Guyon, 1995) yields the following

Proposition 3 *Within the theorem 3 framework, if the penalization rate $c(n)$ is such that:*

$$\lim_n \frac{c(n)}{n} = 0, \qquad \liminf_n \frac{c(n)}{2\log\log n} > \frac{\lambda_{max} J_0}{2\lambda_{min} I_0}, \tag{4.1}$$

then $\widehat{\delta}_n$ converges to the true model δ_0 $\mathbb{P}_{\theta_0}$-almost surely.

A popular choice for the penalization rate is a BIC-like rate $c(n) = \text{const.} \cdot \log n$. Clearly it satisfies Conditions (4.1). Hence a BIC-style procedure is strongly consistent.

5 Examples

5.1 Threshold-exponential AR process

Let I_i, $i = 1, \ldots, K$ be non-overlapping and non-empty intervals of $\mathbb{R}$ such that $\cup_i I_i = \mathbb{R}$. A combined threshold-exponential AR process is defined by

$$X_t = \sum_{i=1}^{K} (\alpha_i + \beta_i X_{t-1}) \, \mathbb{1}_{X_{t-1} \in I_i} + c e^{-\gamma X_{t-1}^2} X_{t-1} + \varepsilon_t, \tag{5.1}$$

with $X_0 = x_0$, and where (ε_t) is a sequence of i.i.d and zero-mean variables. The parameters are $\theta = (\alpha_i, \beta_i, c, \gamma)$ of number $s = 2K + 2$. We shall denote the trues values by $\theta_0 = (\alpha_i^*, \beta_i^*, c_*, \gamma_*)$.

Note that when (ε_t) is a Gaussian noise, (Tjøstheim, 1990) has proved that the maximum likelihood estimator is strongly consistent and asymptotic normal. Application of previous results will prove the same for the LS estimator with more general noise. We also give an LIL for this model. One should remark that the likelihood method is feasible only if the density function of the noise is available.

Theorem 4 *Assume*

(i) $(\varepsilon_t)_{t>0}$ *is an i.i.d., zero-mean Lebesgue noise with* $\sigma^2 := \mathbb{E}\varepsilon_1^2 < \infty$*;*

(ii) $c_* \neq 0$, $\gamma_* > 0$ *and* $|\beta_i^*| < 1$ *for all* $i = 1, \ldots, K$.

(iii) $\theta \in \Theta$, *a compact set of* $\mathbb{R}^{2K+2}$ *such that* $\theta_0 \in \overset{\circ}{\Theta}$.

Then,

(a) $\widehat{\theta}_n \xrightarrow{a.s.} \theta_0$.

(b) $\sqrt{n}\left[\widehat{\theta}_n - \theta_0\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\sigma^2 I_0^{-1})$.

Moreover, if $\mathbb{E}\varepsilon_1^{2+\delta} < \infty$ *for some* $\delta > 0$, *then the LIL from Theorem 3 holds.*

It may be useful to explicit for this model the information matrix I_0 defined Eq. (3.9). Let Y be some real random variable with probability distribution μ_{θ_0} and set

$$W = \left(\left[\mathbb{1}_{I_i}(Y)\right]_{1\leq i\leq K},\ Y\ \left[\mathbb{1}_{I_i}(Y)\right]_{1\leq i\leq K},\ Ye^{-\gamma_* Y^2},\ -c_* Y^3 e^{-\gamma_* Y^2}\right)^{\mathrm{T}}.$$

Straightforward calculus give

$$I_0 = \ \mathbb{E}WW^{\mathrm{T}}\ . \tag{5.2}$$

5.2 Multilayer perceptrons

Multilayer perceptrons (MP) have become popular in nonlinear modelling due to its universal approximation ability, see e.g. (Hertz et al., 1991). Such a example is the model described Eq. (5.3) which has p input units feeding by the variables $X_{t-1}, \ldots, X_{t-p}$ at time t, a hidden layer with K units and one ouput unit which provides the variable X_t:

$$X_t = \sum_{j=1}^{K} \alpha_j \psi(\sum_{i=1}^{p} \beta_{ij} X_{t-i} + \beta_{0j}) + \alpha_0 + \varepsilon_t. \tag{5.3}$$

Here (ε_t) is the system noise. Parameters are $\theta = (\alpha_0, \ldots, \alpha_K\,; \beta_{ij}, 0 \leq i \leq p, 1 \leq j \leq K)^{\mathrm{T}}$ with a parametric dimension $s = 1 + K(p+2)$. Their true values are denoted by $\theta_0 = (\alpha_k^*, \beta_{ij}^*)$. For the so-called *activation function* ψ, there are two widely spread choices: the sigmoid map $\psi(x) = \tanh(x)$ or the logistic map $\psi(x) = 1/(1 + e^{-x})$. (Cottrell et al., 1995) describes an interesting use of this model in time series forecasting.

To simplify, we fix $\psi(x) = \tanh(x)$ and shall assume the univariate case. Application of previous results yields

Theorem 5 *Consider an univariate MP model (5.3) with* $\psi(x) = \tanh(x)$. *Assume*

(i) $(\varepsilon_t)_{t>0}$ *is an i.i.d., zero-mean Lebesgue noise such that* $\mathbb{E}\varepsilon_1^{6+\delta} < \infty$ *for some* $\delta > 0$;

(ii) *$\theta \in \Theta$, a compact subset of $\mathbb{R}^s$ such that $\theta_0 \in \overset{\circ}{\Theta}$.*

(iii) *For all θ different from θ_0, there exists $\mathrm{x} \in \mathbb{R}^p$ such that $f(\mathrm{x}, \theta) \neq f(\mathrm{x}, \theta_0)$.*

(iv) *The matrix I_0, defined (3.9), is regular.*

Then, with $\sigma^2 = \mathbb{E}\varepsilon_1^2$,

(a) $\widehat{\theta}_n \xrightarrow{a.s.} \theta_0$.

(b) $\sqrt{n}\left[\widehat{\theta}_n - \theta_0\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\sigma^2 I_0^{-1})$.

(c) *The LIL from Theorem 3, as well as the strong consistency of the model selection procedure from Proposition 3 both hold with $d = 1$ and $J_0 = 2\sigma^2 I_0$.*

It is worth to point out that the strong consistency of the estimator, statement (a), is obtained as soon as the noise has a moment of second order.

6 Proofs

The following definitions and notations will be used in proofs. Let $\mathcal{F} = (\mathcal{F}_n)_{n \geq 0}$ be the natural filtration associated to the NAR(p) process where $\mathcal{F}_n = \sigma(\varepsilon_t,\ 1 \leq t \leq n)$ for $n \geq 1$, and $\mathcal{F}_0$ is the degenerated σ-algebra. If (M_n) is a square integrable martingale w.r.t. $\mathcal{F}$, we denote by $(\langle M \rangle_n)$ its *increasing process* defined by:

$$\langle M \rangle_0 = 0, \qquad \langle M \rangle_n = \langle M \rangle_{n-1} + \mathbb{E}\left[||M_n - M_{n-1}||^2 \,\middle|\, \mathcal{F}_{n-1}\right] \qquad \text{for } n \geq 1.$$

Proof of Proposition 1

We denote $\Delta f_t = f(\mathbf{X}_t\,; \theta_0) - f(\mathbf{X}_t\,; \theta)$. We have:

$$U_n(\theta) - U_n(\theta_0) = \frac{B_n}{n} + \frac{C_n}{n},$$

with:

$$B_n = \sum_{0 \leq t < n} ||\Delta f_t||^2, \quad C_n = 2 \sum_{0 \leq t < n} \langle \varepsilon_{t+1}, \Delta f_t \rangle.$$

From **[M]**-(iii),

$$||f(\mathrm{x}\,; \theta_0) - f(\mathrm{x}\,; \theta)||^2 \leq \text{const. } (1 + |\mathrm{x}|^a), \qquad \mathrm{x} \in (\mathbb{R}^d)^p. \tag{6.4}$$

Since the true model under θ_0 meets the assumption of stability **[S]**, the SLLN **[S]**-(iii) ensures that:

$$\frac{B_n}{n} \xrightarrow{a.s.} \int_{(\mathbb{R}^d)^p} ||f(\mathrm{x}\,; \theta) - f(\mathrm{x}\,; \theta_0)||^2 \mu_{\theta_0}(d\mathrm{x}).$$

$M_n := C_n/2$ is a square integrable martingale (**[S]**-(i)). Its increasing process $\langle M\rangle_n$ is equal to:

$$\langle M\rangle_n = \sum_{0\le t<n} \Delta f_t{}^{\mathrm{T}}\Gamma\Delta f_t \quad ,$$

and tends to some positive variable $\langle M\rangle_\infty \le \infty$. From the SLLN for square integrable martingale, we know that on $\{\langle M\rangle_\infty < \infty\}$, (see e.g. (Duflo, 1997), Theorem 1.3.15, p. 20), M_n converges to a finite variable, and so M_n/n tends to 0. On $\{\langle M\rangle_\infty = \infty\}$, $M_n/\langle M\rangle_n$ converges to 0. As almost surely,

$$\frac{1}{n}\langle M\rangle_n \longrightarrow \int_{(\mathbb{R}^d)^p} [f(\mathbf{x}\,;\theta) - f(\mathbf{x}\,;\theta_0)]^{\mathrm{T}}\,\Gamma\,[f(\mathbf{x}\,;\theta) - f(\mathbf{x}\,;\theta_0)]\,\mu_{\theta_0}(d\mathbf{x}) \ge 0 \ ,$$

again $M_n/n \to 0$. Hence C_n/n tends to 0 in both cases.

On the other hand, the assumption **[M]**-(iii) and the inequality (6.4) makes the map $\theta \mapsto K(\theta,\theta_0)$ continuous. ∎

Proof of Theorem 1

If we denote W_n the uniform continuity modulus of U_n, i.e.

$$W_n(\eta) := \sup_{\substack{\alpha,\beta\in\Theta\\ ||\alpha-\beta||\le\eta}} |U_n(\alpha) - U_n(\beta)|, \qquad \eta > 0.$$

a sufficient condition ensuring the strong consistency of $(\widehat{\theta}_n)$ is (see (Guyon, 1995), §3.4) the existence of a deterministic sequence (u_k), decreasing to 0, such that for all k,

$$P_{\theta_0}\left[\limsup_{n\longrightarrow\infty}\left\{W_n\left(\frac{1}{k}\right) \ge u_k\right\}\right] = 0 \quad . \tag{6.5}$$

For $\alpha,\beta\in\Theta$, set $\delta(\mathbf{x}\,;\alpha,\beta) := f(\mathbf{x}\,;\alpha) - f(\mathbf{x}\,;\beta)$. From **[M]**-(iii), we have:

$$\begin{aligned}
&n|U_n(\alpha) - U_n(\beta)|\\
&= \left|\sum_{0\le t<n} \langle \delta(\mathbf{X}_t\,;\theta_0,\alpha) + \delta(\mathbf{X}_t\,;\theta_0,\beta) + 2\varepsilon_{t+1}\ ,\ \delta(\mathbf{X}_t\,;\alpha,\beta)\rangle\right|\\
&\le G(||\alpha-\beta||)(1+|\mathbf{X}_t|^{a/2}) \sum_{0\le t<n} \left[\text{const. } (1+|\mathbf{X}_t|^{a/2}) + 2||\varepsilon_{t+1}||\right]\\
&\le G(||\alpha-\beta||) \sum_{0\le t<n} \left[||\varepsilon_{t+1}||^2 + \text{const. } (1+|\mathbf{X}_t|^{a})\right] .
\end{aligned} \tag{6.6}$$

We denote S_n the sum from the last inequality. Applying the SLLN to the integrable i.i.d. sequence $(||\varepsilon_{t+1}||^2)$ on one hand, and to the function $(1+|\cdot|^a)$ on the other hand, S_n/n tends a.s., to a constant limit $\ell > 0$.

By (6.6), we find $W_n(\eta) \leq G(\eta)S_n/n$. For any positive integer k, let us define $u_k = 2\ell G(1/k)$. This sequence decreases down to 0. Then, for fixed k (where *i.o.* means *infinitely often*),

$$\limsup_n \left\{ W_n(\frac{1}{k}) \geq u_k \right\} = \left\{ W_n(\frac{1}{k}) \geq u_k \ \ i.o. \ \right\}$$
$$\subset \left\{ G(\frac{1}{k})\frac{S_n}{n} \geq u_k \ \ i.o. \ \right\} = \left\{ \frac{S_n}{n} \geq 2\ell \ \ i.o. \ \right\}.$$

On $A := \{\frac{1}{n}S_n \geq 2\ell \ \ i.o. \ \}$, $\frac{1}{n}S_n$ can not converge to ℓ ; then A is a null event. The condition (6.5) is satisfied, and the strong consistency established. ∎

Some preliminary computations are useful for next proofs. Since Θ is compact and by Conditions **[N]**-(ii)-(iii), there exists $\gamma > 0$ such that:

$$\forall i,j,k, \ \forall \theta \in V, \ \forall x \in (\mathbb{R}^d)^p, \qquad \left| D^2_{ij} f_k(\mathrm{x}\,;\theta) \right| \leq \gamma(1 + |\mathrm{x}|^{a/2}). \tag{6.7}$$

It follows an estimate of increasing rate of first order derivatives

$$\forall i,k, \ \forall \theta \in V, \ \forall x \in (\mathbb{R}^d)^p, \qquad |D_i f_k(\mathrm{x}\,;\theta) - D_i f_k(\mathrm{x}\,;\theta_0)| \leq \gamma ||\theta - \theta_0||(1 + |\mathrm{x}|^{a/2}). \tag{6.8}$$

And finally, there exists another constant γ' such that:

$$\forall i,k, \ \forall \theta \in V, \ \forall x \in (\mathbb{R}^d)^p, \qquad |D_i f_k(\mathrm{x}\,;\theta)| \leq \gamma'(1 + |\mathrm{x}|^{a/2}). \tag{6.9}$$

For the matrix function $M(\mathrm{x};\theta_0)$ define Eq. (3.5), the estimates (6.8)-(6.9) lead to:

$$||M(\mathrm{x}\,;\theta) - M(\mathrm{x}\,;\theta_0)|| \leq \text{const. } ||\theta - \theta_0||(1 + |\mathrm{x}|^a), \quad \mathrm{x} \in (\mathbb{R}^d)^p, \theta \in V \tag{6.10}$$
$$||M(\mathrm{x}\,;\theta)|| \leq \text{const. } (1 + |\mathrm{x}|^a), \quad \mathrm{x} \in (\mathbb{R}^d)^p, \theta \in V \tag{6.11}$$

Proof of Proposition 2

Let us first prove (3.9). Within the expression (3.8) of $D^2 U_n(\theta_0)$, the first term converges a.s. to the matrix I_0. Indeed, the SLLN **[S]**-(iii) can be applied from the control (6.11) of the matrix function $M(\mathrm{x}, \theta_0)$.

For the second term, its element (i,j), say $M_n := \sum_{0 \leq t < n} \varepsilon^{\mathrm{T}}{}_{t+1} D^2_{ij} f(\mathbf{X}_t\,;\theta_0)$, is a square integrable martingale. Its increasing process $\langle M \rangle_n$ is equal to:

$$\langle M \rangle_n = \sum_{0 \leq t < n} \mathbf{tr} \left[\Gamma \cdot \left\{ D^2_{ij} f (D^2_{ij} f)^{\mathrm{T}} \right\} (\mathbf{X}_t\,;\theta_0) \right]$$

Given (6.7), an argument similar to the one used at the end of the proof of the proposition 1 ensures that M_n/n tends a.s. to 0. The conclusion (3.9) follows.

For (3.10), let us denote this time:

$$M_n := -\frac{n}{2} DU_n(\theta_0) = \sum_{0 \leq t < n} \varepsilon^{\mathrm{T}}{}_{t+1} Df(\mathbf{X}_t\,;\theta_0). \tag{6.12}$$

By (6.9), it is a square integrable vector martingale. And $\langle M \rangle_n$ is equal to:

$$\langle M \rangle_n = \sum_{0 \leq t < n} \{Df(\mathbf{X}_t\,;\theta_0)\}^{\mathrm{T}} \Gamma Df(\mathbf{X}_t\,;\theta_0). \tag{6.13}$$

Still by (6.9), each term of the matrix function $\mathrm{x} \mapsto J(\mathrm{x}\,;\theta_0) := \{Df(\mathrm{x}\,;\theta_0)\}^{\mathrm{T}} \Gamma Df(\mathrm{x}\,;\theta_0)$ is bounded (in norm) by const. $(1 + |\mathrm{x}|^a)$. So, from the SLLN **[S]**-(iii),

$$\frac{1}{n} \langle M \rangle_n \xrightarrow{a.s.} \int_{(\mathbb{R}^d)^p} J(\mathrm{x}\,;\theta_0) \mu_{\theta_0}(d\mathrm{x}) = \frac{1}{4} J_0. \tag{6.14}$$

The CLT (3.10) is proved if (M_n) fulfills the following Lindeberg's condition (see e.g. (Duflo, 1997), corollary 2.1.10):

$$\text{for all } \delta > 0,\;\; L_n := \frac{1}{n} \sum_{0 \leq t < n} \mathbb{E}\left[||M_t - M_{t-1}||^2 \mathbb{1}_{\{||M_t - M_{t-1}|| \geq \delta\sqrt{n}\}} \,\middle|\, \mathcal{F}_{t-1} \right] \xrightarrow{\mathbb{P}_{\theta_0}} 0. \tag{6.15}$$

Let be $A > 0$ and:

$$F_n(A) := \frac{1}{n} \sum_{0 \leq t < n} \mathbb{E}\left[||M_t - M_{t-1}||^2 \mathbb{1}_{||M_t - M_{t-1}|| \geq \delta A} \,\middle|\, \mathcal{F}_{t-1} \right] = \frac{1}{n} \sum_{0 \leq t < n} h(\mathbf{X}_t, A),$$

with:

$$h(\mathrm{x}, A) = \mathbb{E}\left[\{Df(\mathrm{x}\,;\theta_0)\}^{\mathrm{T}} \varepsilon_1 \varepsilon^{\mathrm{T}}{}_1 Df(\mathrm{x}\,;\theta_0) \mathbb{1}_{\{||\{Df(\mathrm{x}\,;\theta_0)\}^{\mathrm{T}} \varepsilon_1|| > A\}} \right].$$

It is clear that from (6.9),

$$h(\mathrm{x}, A) \leq \text{const. } (1 + |\mathrm{x}|^a). \tag{6.16}$$

Hence, by **[S]**-(iii) again,

$$F_n(A) \xrightarrow{a.s.} \phi(A) := \int_{(\mathbb{R}^d)^p} h(\mathrm{x}, A) \mu_{\theta_0}(d\mathrm{x}).$$

The last function ϕ is positive and decreasing. Moreover, by the dominated convergence theorem, $\phi(A)$ tends to 0 as A tends to ∞.

On the other hand, when A is fixed, we have $\delta\sqrt{n} > A$ for n large enough, and $L_n = F_n(\delta\sqrt{n}) \leq F_n(A)$. So a.s., $\limsup_n L_n \leq \phi(A)$. Since A is arbitrary, we have a.s., $\lim L_n = 0$. The Lindeberg's condition (6.15) is thus fulfilled and $M_n/\sqrt{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_0/4)$. ∎

Proof of Theorem 2

Since $\widehat{\theta}_n \xrightarrow{a.s.} \theta_0$, for almost all ω, there exists $n_0(\omega)$ such that $\widehat{\theta}_n \in V$ for all $n \geq n_0(\omega)$. By Taylor's formula

$$0 = DU_n(\widehat{\theta}_n) = DU_n(\theta_0) + \Delta_n(\widehat{\theta}_n)(\widehat{\theta}_n - \theta_0), \tag{6.17}$$

where

$$\Delta_n(\widehat{\theta}_n) := \int_0^1 D^2U_n\left[\widehat{\theta}_n + u(\widehat{\theta}_n - \theta_0)\right] du.$$

Taking Proposition 2 into account, we deduce Theorem 2 from the following lemma.

■

Lemma 1 *Within the context of the theorem 2, we have:*

$$\Delta_n(\widehat{\theta}_n) - D^2U_n(\theta_0) \xrightarrow{a.s.} 0, \qquad \Delta_n(\widehat{\theta}_n) \xrightarrow{a.s.} I_0. \tag{6.18}$$

Proof. For $\theta \in V$ and by (3.8), we have

$$D^2U_n(\theta) - D^2U_n(\theta_0) = \frac{2}{n}\left[A_n(\theta) + B_n(\theta) + C_n(\theta)\right]$$

with

$$\begin{aligned}
A_n(\theta) &= \sum_{0 \leq t < n} \left[M(\mathbf{X}_t\,;\theta) - M(\mathbf{X}_t\,;\theta_0)\right], \\
B_n(\theta) &= \sum_{0 \leq t < n} \left[f(\mathbf{X}_t\,;\theta) - f(\mathbf{X}_t\,;\theta_0)\right]^{\mathrm{T}} \left[D^2_{ij} f(\mathbf{X}_t\,;\theta)\right]_{1 \leq i,j \leq s}, \\
C_n(\theta) &= -\sum_{0 \leq t < n} \varepsilon^{\mathrm{T}}{}_{t+1} \left[D^2_{ij} f(\mathbf{X}_t\,;\theta) - D^2_{ij} f(\mathbf{X}_t\,;\theta_0)\right]_{1 \leq i,j \leq s}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\|A_n(\theta)\| &\leq \text{const. } \|\theta - \theta_0\| \sum_{0 \leq t < n} (1 + |\mathbf{X}_t|^a), \quad \text{by (6.11)} \\
\|B_n(\theta)\| &\leq \text{const. } G\left(\|\theta - \theta_0\|\right) \sum_{0 \leq t < n} (1 + |\mathbf{X}_t|^a), \quad \text{by } \textbf{[M]}\text{-iii and (6.7)} \\
\|C_n(\theta)\| &\leq \text{const. } \left[\sum_{i,j,k} \sigma_{ijk}(z)\right] \sum_{0 \leq t < n} \|\varepsilon_{t+1}\|(1 + |\mathbf{X}_t|^{a/2}) \\
&\leq \text{const. } \left[\sum_{i,j,k} \sigma_{ijk}(z)\right] \left[\sum \|\varepsilon_{t+1}\|^2 + \sum (1 + |\mathbf{X}_t|^{a/2})\right].
\end{aligned}$$

On the other hand, by (6.17)

$$\begin{aligned}
&\|\Delta_n(\widehat{\theta}_n) - D^2U_n(\theta_0)\| \\
&= \left\| \int_0^1 \left\{ D^2U_n\left[\widehat{\theta}_n + u(\widehat{\theta}_n - \theta_0)\right] - D^2U_n(\theta_0)\right\} du \right\| \\
&\leq \frac{2}{n}\left\| A_n(\widehat{\theta}_n) + B_n(\widehat{\theta}_n) + C_n(\widehat{\theta}_n)\right\| .
\end{aligned}$$

Since both $\frac{1}{n}\sum ||\varepsilon_{t+1}||^2$ and $\frac{1}{n}\sum(1 + ||\mathbf{X}_t||^a)$ converge a.s., and $\widehat{\theta}_n \xrightarrow{a.s.} \theta_0$, $\Delta_n(\widehat{\theta}_n) - D^2U_n(\theta_0)$ converges to 0 a.s. The second result is a consequence from Proposition 2. ∎

Proof of Theorem 3

We shall apply Lemma 2 below to the regressive series

$$M_n := -\frac{n}{2}\langle DU_n(\theta_0), u\rangle = \sum_{0\leq t<n} \varepsilon^{\mathrm{T}}{}_{t+1} Df(\mathbf{X}_t\,;\theta_0)u \ .$$

Following the notaions used there, let $\phi_t = Df(\mathbf{X}_t\,;\theta_0)u$ and we check Conditions (i), (ii) and (iii) of Lemma 2. We have,

$$\Gamma_n \equiv \Gamma, \quad T_n^2 = u^{\mathrm{T}}\{Df(\mathbf{X}_n\,;\theta_0)\}^{\mathrm{T}}\Gamma Df(\mathbf{X}_n\,;\theta_0)u, \quad s_n^2 = \sum_{0\leq t<n} T_t^2.$$

Let α be a positive number such that $\alpha < \min(1, a'/a - 1)$. The conditions (i)-(ii) are clearly fulfilled. For (iii), first note that by SLLN, s_n^2/n tends a.s. to $\frac{1}{4}u^{\mathrm{T}}J_0u$, which is strictly positive (see assumption). It is thus sufficient to prove that there exists an $\eta \in (0,1)$ for which

$$\sum \frac{T_n^{2+2\alpha}}{s_n^{2(1-\eta)(1+\alpha)}} \quad \text{converge a.s.} \tag{6.19}$$

Since $s_n^2 \sim$ const. n, we have to prove

$$\sum \frac{T_n^{2+2\alpha}}{n^{(1-\eta)(1+\alpha)}} \quad \text{converge a.s.}$$

Set $\Sigma_n := T_1^{2+2\alpha} + \cdots + T_n^{2+2\alpha}$. The choice of α ensures that

$$\left|u^{\mathrm{T}}\{Df(\mathrm{x}\,;\theta_0)\}^{\mathrm{T}}\,\Gamma\,\{Df(\mathrm{x}\,;\theta_0)\}u\right|^{1+\alpha} \leq \text{const. } (1 + |\mathrm{x}|^{a'}) \ . \tag{6.20}$$

Therefore Σ_n/n converges a.s. towards some constant $\gamma \geq 0$. By Abel's transformation rule,

$$\sum_{k=1}^{n} \frac{T_k^{2+2\alpha}}{k^{(1-\eta)(1+\alpha)}} = \frac{\Sigma_n}{n^{(1-\eta)(1+\alpha)}} + \sum_{k=1}^{n-1}\left[\frac{1}{k^{(1-\eta)(1+\alpha)}} - \frac{1}{(k+1)^{(1-\eta)(1+\alpha)}}\right]\Sigma_k.$$

Furthermore,

$$\left[\frac{1}{k^{(1-\eta)(1+\alpha)}} - \frac{1}{(k+1)^{(1-\eta)(1+\alpha)}}\right] \Sigma_k \sim \frac{(1-\eta)(2+\alpha)}{k^{(1-\eta)(1+\alpha)}} \cdot \frac{\Sigma_k}{k}.$$

Now choosing $\eta < \alpha/(1+\alpha)$ yields $(1-\eta)(1+\alpha) > 1$. The last series converges and $\Sigma_n/n^{(1-\eta)(1+\alpha)}$ tends to 0. So the convergence (6.19) holds. Applying Lemma 2 ends the proof of (3.11). Finally, (3.12) follows from (6.17) and Lemma 1. ■

Lemma 2 (Law of the iterated logarithm for regressive series) *Let $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$ be some filtration defined on a probability space $(\Omega, \mathcal{A}, P)$, and $(\varepsilon_n)_{n\geq 0}$, $(\phi_n)_{n\geq 0}$ two $\mathcal{F}$-adapted sequences of $\mathbb{R}^d$-valued random vectors. Set for $n \geq 1$,*

$$M_n := \sum_{t=1}^{n} \langle \phi_{t-1}, \varepsilon_t \rangle, \quad \Gamma_n := \mathbb{E}(\varepsilon_{n+1} \varepsilon^T{}_{n+1} \,|\, \mathcal{F}_n), \quad T_n^2 := \phi^T{}_n \Gamma_n \phi_n, \quad \textit{and} \quad s_n^2 := \sum_{t=1}^{n} T_t^2. \tag{6.21}$$

Assume that there exists some $\alpha \in (0,1)$ such that a.s.

(i) *for all $n \geq 0$, $\mathbb{E}(\varepsilon_{n+1}|\mathcal{F}_n) = 0$; $\sup_n \mathbb{E}(||\varepsilon_{n+1}||^{2+2\alpha}|\mathcal{F}_n) < \infty$.*

(ii) $\liminf_n \lambda_{min}(\Gamma_n) > 0$.

(iii) $s_n^2 \to \infty$, $\sum T_n^{2+2\alpha}/s_n^{2+2\alpha} < \infty$ *and* $T_n^2 = o[s_n^2\{\log\log(s_n^2)\}^{-1/\alpha}]$.

Then,

$$\limsup \frac{M_n}{\sqrt{s_{n-1}^2 \log\log s_{n-1}^2}} = 1 = -\liminf \frac{M_n}{\sqrt{s_{n-1}^2 \log\log s_{n-1}^2}}, \quad a.s. \tag{6.22}$$

We do not go into more details since this LIL can be deduced from Stout's LLI for martingales, (Stout, 1970) and a troncature technique developped in (Duflo et al., 1990). It is worth noting that this LIL does not require any moment conditions on the regressors (ϕ_n).

Proof of Theorem 4

First by applying results from (Tjøstheim, 1990) (or applying Criterion **[C.1]**), we know that, under Conditions (i)-(ii) of Theorem 4, the process (X_t) under the true model $f(\cdot\,;\theta_0)$ has a stability of order 2 (resp. $2+\delta$) if $\mathbb{E}|\varepsilon_1|^2 < \infty$ (resp. if $\mathbb{E}|\varepsilon_1|^{2+\delta} < \infty$). Moreover, its invariant measure μ_{θ_0} has a everywhere positive density with respect to the Lebesgue measure. In particular, taking into account (5.2) and the fact $\gamma_* > 0$, this makes I_0 regular.

The remaining conditions in **[M]-[N]-[D]** could be readily checked. This is mainly based on the following estimates which hold since the parameter space Θ is compact:

$$\begin{aligned} |f(x\,;\theta)| &\leq \text{ const. } (1+|x|), \qquad \qquad \qquad \qquad \qquad \qquad \qquad \text{(6.23)} \\ \Big|\frac{\partial^m}{\partial\theta_{i_1}\dots\theta_{i_m}} f(x\,;\theta)\Big| &\leq \text{ const. } (1+|x|),\ (i_1,\dots,i_m)\in\{1,\dots,s\}^m,\ m=1,2,.\text{(6.24)} \end{aligned}$$

for all $x \in \mathbb{R}$ and $\theta \in \Theta$. ∎

Proof of Theorem 5

Note first that the map $\psi(x) = \tanh(x)$ is $\mathcal{C}^\infty$, and all its derivatives are bounded. In particular, we have for $x \in \mathbb{R}, 0 \leq \psi(x) \leq 1,\ 0 \leq \psi'(x) \leq 1$ et $-2 \leq \psi''(x) \leq 0$.

Assumption [M] : **[M]**-(i) is clearly fulfilled. Since ψ is bounded, $\mathrm{x} \mapsto f(\mathrm{x}\,;\theta_0)$ is bounded too. The true model under θ_0 fulfills the stability criterion **[C.1]** from §2. The model has the stability property **[S]** with $a = 6$ and **[M]**-(ii) is proved. For **[M]**-(iii), let be $\theta = (\alpha_j, \beta_{ij})$, and $\theta' = (\alpha'_j, \beta'_{ij})$. A straightforward calculus shows that for all $\mathrm{x} \in \mathbb{R}^p$,

$$|f(\mathrm{x}\,;\theta) - f(\mathrm{x}\,;\theta')| \leq \text{const. } ||\theta - \theta'||(1 + ||\mathrm{x}||).$$

Identifiability [D] : Because (ε_t) has an everywhere positive density with respect to Lebesgue measure, the invariant probability μ_{θ_0} of the vectorized chain $X^{(p)}$ (under $\mathbb{P}_{\theta_0}$) is also equivalent to Lebesgue measure. The condition **[D]** is met taking into account the assumption (iii).

Assumption [N] : Consider $V = \overset{\circ}{\Theta}$. **[N]**-(i) is straightforward. For **[N]**-(ii)-(iii), we can show easily that, for all $\mathrm{x} \in \mathbb{R}^p$, we have:

$$\begin{aligned} |D_i f(\mathrm{x}\,;\theta_0)| &\leq \text{ const. } (1+|\mathrm{x}|), \qquad i=1,\dots,s \\ \left|D^2_{ij} f(\mathrm{x}\,;\theta_0)\right| &\leq \text{ const. } (1+|\mathrm{x}|^2), \qquad i,j=1,\dots,s \\ \left|D^2_{ij} f(\mathrm{x}\,;\theta) - D^2_{ij} f(\mathrm{x}\,;\theta_0)\right| &\leq \text{ const. } ||\theta-\theta_0||(1+|\mathrm{x}|^3), \qquad \theta \in \overset{\circ}{\Theta}, \quad i,j=1,\dots,s. \end{aligned}$$

The upper bound in the last inequality involves a polynomial of degree 3, that is why we need a moment of order larger than 6 for the noise.

At last, the required conditions in theorem 3 are directly fullfilled. ∎

ACKNOWLEDGMENTS. *The author is grateful to* Xavier GUYON *and* Marie DUFLO *for their helpful suggestions on this work. Lemma 2 is due to* Marie DUFLO.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, 21:243–247.

Attali, J. (1998). *Chaînes de Markov Stables*. PhD thesis, Université Paris I.

Borovkov, A. A. (1991). Lyapounov functions and ergodicity of multidimensional Markov chains. *Theory Probab. Appl.*, 36(1):1–18.

Borovkov, A. A. and Korshunov, D. (1993). Ergodicity in a sense of weak convergence, equilibrium-type identities and large deviations for Markov chains. In et al., B. G., editor, *Probability theory and mathematical statistics*, Vilnius.

Cottrell, M., Girard, B., Girard, Y., Mangeas, M., and Muller, C. (1995). Neural modeling for time series : a statistical stepwise method for weight elimination. *I.E.E.E. Trans. Neural Networks*, 6:1355–1364.

Doukhan, P. and Ghindès, M. (1980). Etude du processus $X_n = f(X_{n-1}) + \varepsilon_n$. *C.R.A.S.*, 290:921–923.

Duflo, M. (1997). *Random Iterative Models*. Springer-Verlag.

Duflo, M., Senoussi, R., and Touati, A. (1990). Sur la loi des grands nombres pour les martingales vectorielles et l'estimateur des moindres carrés d'un modèle de regression. *Ann. I.H.P.*, 26:549–566.

Gallant, A. R. (1987). *Nonlinear Statistical Models*. Wiley.

Gallant, A. R. and White, H. (1988). *A Unified Theory for estimation and Inference for Nonlinear Dynamic Models*. B. Blackwell.

Guyon, X. (1995). *Random Fields on a Network – Modeling, Statistics, and Applications*. Springer-Verlag, Berlin.

Haggan, V. and Ozaki, T. (1981). Modeling nonlinear random vibrations using an amplititude-dependent autoregressive time series model. *Biometrika*, 68:189–196.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.*, 8:1071–1081.

Haughton, D. (1991). Consistency of a class of information criteria for model selection in nonlinear regression. *Commun. Statist. Theory and Methods*, 20:1619–1629.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley Pub. Co., Redwood City.

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

Lai, T. (1994). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Ann. Statist.*, 22:1917–1930.

Lai, T. and Lee, C. (1997). Information and prediction criteria for model selection in stochastic regression and arma models. *Statistica Sinica*, 7:285–309.

Lai, T. and Zhu, G. (1991). Adaptive prediction in non-linear autoregressive models and control systems. *Statistica Sinica*, 1:309–314.

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, Berlin.

Mokkadem, A. (1987). Sur un modèle autorégressif non linéaire: ergodicité et ergodicité géométrique. *J. Time Series Analysis*, 8(2):195–204.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, 12:758–765.

Pötscher, B. M. and Prucha, L. R. (1997). *Dynamic Nonlinear Econometric Models*. Springer-Verlag.

Quinn, B. G. (1980). Order determination for a multivariate autoregression. *J. Roy. Statist. Soc. Ser. B*, 42:182–185.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6:461–464.

Stout, W. (1970). A martingale analogue of Kolmogorov's law of the iterated logarithm. *Z. Wahr. Verv. Gebiete.*, 15:279–290.

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

Tjøstheim, D. (1990). Nonlinear time series and Markov chains. *Adv. Appl. Prob.*, 22:587–611.

Tong, H. (1983). *Threshold Models in Nonlinear Time Series*. Springer-Verlag.

Tong, H. (1990). *Non-linear Time Series (A Dynamic System Approach)*. Oxford Univ. Press, New York.

Tsay, R. S. (1984). Order selection in nonstationary autoregressive models. *Ann. Statist*, 12:1425–1433.

White, H. and Domowitz, I. (1984). Nonlinear regression with dependent observations. *Econometrica*, 52:143–161.