Journal of Mathematical Systems, Estimation, and Control
Vol. 4, No. 2, 1994, pp. 1–26

Fixed Gain Off-line Estimators of ARMA Parameters*

László Gerencsér†

Abstract

Let $\widehat{\theta}_N^{\lambda}$ denote the estimator of the ARMA parameter vector θ^* using a fixed gain off-line prediction error method with gain or forgetting rate λ. We show that under certain conditions $\widehat{\theta}_N^{\lambda} - \theta^*$ is an L-mixing process and can be decomposed as the sum of an explicitly given L-mixing process of the order of magnitude $O(\lambda^{1/2})$ and of a residual term of the order of magnitude $O(\lambda) + O((1-\lambda)^N)$. Here the order of magnitude is measured as the $L_q(\Omega, \mathcal{F}, P)$ norm for any $q \geq 1$ where $(\Omega, \mathcal{F}, P)$ is the underlying probability space. The result of the paper is directly applicable to fixed gain recursive estimators of AR models. The result of this paper has been applied in the theory of stochastic complexity.

Key words: ARMA-processes, prediction error estimation, forgetting, strong approximation, L-mixing processes

AMS Subject Classifications: 60H10, 62L20, 93E12

1 Introduction

It has been shown in [5] that the estimation error of the parameters of an ARMA process can be approximated by the arithmetic mean of a martingale so that the approximation error is of the order of magnitude $O_M(N^{-1})$. The notation $O_M(\cdot)$ is to be described following Definition 4.1. That result played a key role in deriving asymptotic properties of the so-called predictive stochastic complexity for ARMA processes (cf. [4] and also [11]).

*October 1, 1992; received in final form February 22, 1993. Summary appeared in Volume 4, Number 2, 1994.

†This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant 01329, and by the National Science and Research Foundation of Hungary under Grant T 007336. The author wishes to thank Zsuzsanna Vágó for her work in the preparation of this document.

The purpose of this paper is to present an analogous representation of the estimation error process when the estimator is obtained using fixed gain estimation or in other words estimation with exponential forgetting.

Fixed gain estimation is typically used when we anticipate that the parameters may vary over time and in this case it would be more natural to consider recursive estimation methods. However the analysis of fixed gain recursive estimation methods in general is much harder than the analysis of the corresponding off-line estimation methods. An exception is the case when the model that we consider is an autoregressive process, namely in this case a suitable fixed gain recursive estimator is identical with the fixed gain off-line estimator. Thus in this case we have a practically useful result.

There is substantial evidence that the theory of stochastic complexity provides us with a new principle of statistical analysis with the help of which we can conveniently solve statistical problems, which were considered earlier very difficult. As an example we mention the problem of model structure selection (cf. [14] and for a recent survey [11]). Also it is well-known that the analysis of the so-called predictive stochastic complexity depends strongly on the fine asymptotic analysis of the estimator process which is used in the "encoding procedure"(cf. [4]). This explains the motivation of the present investigation. The results of the paper have actually been applied in [6] to derive the asymptotic properties of a new predictive stochastic complexity. The proposed estimation procedure and the associated predictive stochastic complexity is also relevant in a new change point detection method (cf. [9]). Also using "fixed-gain" estimators a new model-selection method has been developed (cf. [10, 11]).

The main result of the paper is Theorem 1.2 which gives the desired decomposition of the estimation error as the sum of an explicitly given L_0-mixing process of the order of magnitude $O_M(\lambda^{1/2})$ and of an L_0-mixing error term of the order of magnitude $O_M(\lambda) + O_M((1-\lambda)^N)$. Here λ denotes the "forgetting rate" (i.e. small λ means small rate of forgetting). The definition of L_0-mixing processes is given in Appendix II.

The results of the paper can easily be extended to multivariable finite dimensional linear stochastic systems if a certain uniqueness theorem analogous to the one given in [1] holds (cf. e.g. [16, 18]), or even to the general estimation problem described by Ljung's scheme. Some of these possible extensions are stated in [8]). Now we specify the notations and technical conditions for the first theorem of the present paper.

Let $(y_n), n = 0, \pm 1, \pm 2, \ldots$ be a second order stationary ARMA (p, q) process satisfying the following difference equation:

$$y_n + a_1^* y_{n-1} + \ldots + a_p^* y_{n-p} = e_n + c_1^* e_{n-1} + \ldots + c_q^* e_{n-q}. \qquad (1.1)$$

which we write in a shorthand notation as $A^* y = C^* e$ where A^*, C^* are polynomials of the backward shift operator. Define $A^*(z^{-1}) = \sum_{i=0}^{p} a_i^* z^{-i}$,

$C^*(z^{-1}) = \sum_{i=0}^{q} c_i^* z^{-i}$.

Condition 1.1 $A^*(z^{-1})$ and $C^*(z^{-1})$ have all their roots inside the unit circle, i.e. $A^*(z^{-1})$ and $C^*(z^{-1})$ are stable, moreover we assume that they are relative prime and $a_0^* = c_0^* = 1$.

Let $(\mathcal{F}_n)$, and $(\mathcal{F}_n^+)$, $n = 0, \pm 1, \pm 2 \ldots$ be an increasing and a decreasing family of σ-algebras, respectively, such that $\mathcal{F}_n$, and $\mathcal{F}_n^+$ are independent for all n.

Condition 1.2 (e_n) is a discrete-time, second order stationary, martingale-difference process with respect to $\mathcal{F}_n$, $n = 0, \pm 1, \pm 2 \ldots$ i.e. $E(e_n|\mathcal{F}_{n-1}) = 0$, and $\mathrm{E}(e_n^2|\mathcal{F}_{n-1}) = \sigma^{*2} = \text{const.}$ a.s. Moreover we assume that (e_n) is L-mixing.

The concept of L-mixing processes together with the conditions imposed onto $\mathcal{F}_n, \mathcal{F}_n^+$ are described in Appendix I. A detailed exposition is given in [3]. The significance of the class of L-mixing processes is that they are closed under all the operations which are usual in the estimation theory of linear stochastic systems. This invariance property is not shared by the class of ϕ-mixing processes or mixingales which are also potential candidates for the analysis of estimation methods. The concept of L-mixing processes has been used extensively in previous works (cf. [8, 11] for two recent surveys).

Let $G \subset \mathbb{R}^{p+q}$ denote the set of θ's such that the corresponding polynomials $A(z^{-1})$ and $C(z^{-1})$ are stable. G is an open set. Let D^* and D be compact domains such that $\theta^* \in \mathrm{int} D^* \subset \mathrm{int}\ D$ and $D \subset G$. Here int D denotes the interior of D. To estimate the unknown parameters $a_i^*, c_j^*, i = 1, \ldots p, j = 1, \ldots q$ and the unknown variance σ^{*2} we use the prediction-error method. (The prediction error method without forgetting is described e.g. in [2, 12] or in [15]). Let us take an arbitrary $\theta \in D$ and define an estimated prediction error process $(\varepsilon_n), n \geq 0$ by the equation

$$\varepsilon = (A/C)y$$

with initial values $\varepsilon_n = y_n = 0$ for $n \leq 0$. Let the coefficient s of $A(z^{-1})$ and $C(z^{-1})$ be denoted by a_i and c_j, respectively, and define the system parameter vector by $\theta = (a_1, \ldots, a_p, c_1, \ldots c_q)^T$. To stress the dependence of (ε_n) on θ and θ^* we shall write $\varepsilon_n = \varepsilon_n(\theta, \theta^*)$. Then the cost-function associated with the prediction-error method using forgetting is given by

$$V_N(\theta, \theta^*) = \frac{1}{2} \sum_{n=1}^{N} (1-\lambda)^{N-n} \lambda \varepsilon_n^2(\theta, \theta^*),$$

where $0 < \lambda < 1$ is the forgetting factor. The factor λ is included in the cost-function to ensure that the sum of the weights is approximately equal to 1.

It is easy to see that the cost function can be computed recursively as follows:

$$V_N(\theta,\theta^*) = (1-\lambda)V_{N-1}(\theta,\theta^*) + \lambda\varepsilon_N^2(\theta,\theta^*),$$

i.e. the correction term corresponding to the latest observation enters the cost function always with the same fixed weight. This representation of the cost function justifies our terminology "fixed gain estimation".

The estimate $\widehat{\theta}_N$ of θ^* may be defined as the solution of the equation

$$\frac{\partial}{\partial\theta}V_N(\theta,\theta^*) = V_{\theta N}(\theta,\theta^*) = 0. \tag{1.2}$$

(Here differentiation is taken both in the almost sure and in the M-sense. For the definition of the latter cf. Appendix I).

It is easy to see that $\varepsilon_n(\theta,\theta^*)$ is a smooth function o f θ for all ω, and hence (1.2) can be written as

$$\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda\varepsilon_{\theta n}(\theta,\theta^*)\varepsilon_n(\theta,\theta^*) = 0. \tag{1.3}$$

In the special case when the process to be estimated is an AR process, i.e. when $C^* = 1$, equation (1.3) is linear, and its solutions, which we denote by $\widehat{\theta}_N$, can be computed recursively in N. Defining the regressor vector $\phi_N = (-y_{N-1}, -y_{N-2}, \ldots, -y_{N-p})^T$, the coefficient matrix of the normal equation is

$$R_N = \sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda\phi_n\phi_n^T.$$

With this notation we have the following recursion (cf. p.18. in [13]):

$$\widehat{\theta}_N = \widehat{\theta}_{N-1} + \lambda R_N^{-1}\phi_N(y_N - \phi_{N-1}^T\widehat{\theta}_{N-1}).$$

Assuming that (y_n) is actually stationary, and "freezing" $\widehat{\theta}_{N-1}$ at the value $\widehat{\theta}_{N-1} = \theta^*$, the process (R_N), and thus also the process $(R_N)^{-1}$ are stationary, too. We get that the correction term in the above recursion is λ times a stationary process, the above recursive estimation scheme is therefore called a fixed gain estimation method.

Since equation (1.2) or (1.3) may have no solution with positive and asymptotically non-negligible probability we will have to define $\widehat{\theta}_N$ with more care in the general case. But first we need some further definitions.

Let us introduce the asymptotic cost function defined by

$$W(\theta,\theta^*) = \lim_{n\to\infty}\frac{1}{2}\mathrm{E}\varepsilon_n^2(\theta,\theta^*).$$

It is easy to see that $W(\theta,\theta^*)$ is smooth in the interior of D and we have

$$W_\theta(\theta^*,\theta^*)=0 \quad \text{and} \quad R^* \stackrel{\Delta}{=} W_{\theta\theta}(\theta^*,\theta^*)>0$$

i.e. R^* is positive definite.

Let us now define the random variables

$$\delta V_{\theta N}=\sup_{\theta\in D,\theta^*\in D^*}|V_{\theta N}(\theta,\theta^*)-W_\theta(\theta,\theta^*)|$$

and

$$\delta V_{\theta\theta N}=\sup_{\theta\in D,\theta^*\in D^*}\|V_{\theta\theta N}(\theta,\theta^*)-W_{\theta\theta}(\theta,\theta^*)\|.$$

It is easy to see (cf. Lemma 4.7) that if $\delta V_{\theta N}<d'$ and $\delta V_{\theta\theta N}<d''$ where d' and d'' are sufficiently small then (1.2) or (1.3) have a unique solution in D. Let us choose $d'=d''$ and let us define the event

$$A_N=\{\omega:\delta V_{\theta N}<d'',\delta V_{\theta\theta N}<d''\},$$

and define a solution of (1.2) or (1.3) as a random variable which is the solution of (1.2) and (1.3) on A_N and arbitrary but D-valued otherwise.

With this notation the first result can be stated as follows:

Theorem 1.1 *Under Condition 1.1 and 1.2 we have for any solution of (1.2)*

$$\widehat{\theta}_N-\theta^*=-(R^*)^{-1}\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda\varepsilon_{\theta n}(\theta^*,\theta^*)e_n+r_N \qquad (1.4)$$

where $r_N=O_M(\lambda)+O_M((1-\lambda)^N)$. *The notation* $O_M(.)$ *is introduced after Definition 4.1.*

It will be seen that the dominant term on the right hand side of (1.4) is $O_M(\lambda^{1/2})$. Since it cannot be expected that $P(A_N)$ tends to zero it will be important to characterize the process χ_{A_N}, where χ_{A_N} denotes the characteristic function of the set A_N. For this we have to impose additional conditions on the input noise process.

Theorem 1.2 *If the input noise process is* L_0*-mixing then a suitable version of the solution of the process* $(\widehat{\theta}_N-\theta^*)$ *and the residual process* (r_N) *are* L_0*-mixing with respect to* $(\mathcal{F}_N,\mathcal{F}_N^+)$ *such that for all* $1\le q<\infty$ *and* $c>0$ *we have*

$$,_{q,c}(\widehat{\theta}_N-\theta^*)=O(\lambda^{-1+c/2(q+1)}) \quad \text{an d} \quad ,_{q,c}(r_N)=O(\lambda^{-1+c/2(q+1)}).$$

This theorem is useful to get pathwise results for the estimator process. E.g. it follows, that for all Lipschitz-continuous function f

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}(f(\widehat{\theta}_N)-f(\theta^*))\le c\lambda^{1/2}$$

almost surely, with some deterministic constant c.

2 The Proof of Theorem 1.1

Lemma 2.1 *For any fixed $d > 0$ and $s > 0$ equation (1.2) has a unique solution in D for $N > c/\lambda$, where c is a deterministic constant, with the property that it is also in the sphere $\{|\theta - \theta^*| < d\}$ with probability at least $1 - c'\lambda^s$. Here the constants depend only on the system and on d and s.*

Remark. The statement of the lemma can be weakened to saying that the probability in question can be bounded from below by $1 - c'\lambda^s - c''(1-\lambda)^N$.

Proof: We show first that the probability to have a solution outside the sphere $\{\theta : |\theta - \theta^*| < d\}$ is less than $c'\lambda^s$ with any $s > 0$ for $N > c/\lambda$. Indeed, the equation $W_\theta(\theta, \theta^*) = 0$ has a single solution $\theta = \theta^*$ in

D (cf. [1]), thus for any $d > 0$ we have

$$d' \stackrel{\Delta}{=} \inf\{|W_\theta(\theta, \theta^*)| : \theta \in D, \theta^* \in D^*, |\theta - \theta^*| \geq d\} > 0$$

since $W_\theta(\theta, \theta^*)$ is continuous in (θ, θ^*) and $D \times D$ is compact. Therefore if a solution of (1.2) exists outside the sphere $|\theta - \theta^*| < d$ then we have for

$$\delta V_{\theta N} = \sup_{\theta \in D, \theta^* \in D^*} |V_{\theta N}(\theta, \theta^*) - W_\theta(\theta, \theta^*)| \tag{2.1}$$

the inequality $\delta V_{\theta N} > d'$.

But the process

$$u_n(\theta, \theta^*) = \varepsilon_{\theta n}(\theta, \theta^*)\varepsilon_n(\theta, \theta^*) - \mathrm{E}\varepsilon_{\theta n}(\theta, \theta^*)\varepsilon_n(\theta, \theta^*) \tag{2.2}$$

is a zero-mean L-mixing process uniformly in θ, θ^* and the same holds for the process $(u_{\theta n}(\theta, \theta^*))$. By Theorem 4.3 we have

$$|\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda(\varepsilon_{\theta n}(\theta, \theta^*)\varepsilon_n(\theta, \theta^*) - \mathrm{E}\varepsilon_{\theta n}(\theta, \theta^*)\varepsilon_n(\theta, \theta^*))| = O_M(\lambda^{1/2}).$$

Note that $\mathrm{E}\varepsilon_{\theta n}(\theta, \theta^*)\varepsilon_n(\theta, \theta^*) = W_\theta(\theta, \theta^*) + \delta_n$, with some $\delta_n = O_M(c^n)$, where $0 < c < 1$, uniformly in θ. The error term δ_n is due to the nonstationary initial conditions $\varepsilon_{\theta 0}(\theta, \theta^*) = 0, \varepsilon_0(\theta, \theta^*) = 0$. It is easy to see that if we let the error process (δ_n) pass through an exponentially smoothing filter, then the output process at time N will be of the order of magnitude $O_M((1-\lambda)^N$ for small λ (cf. Lemma 4.5). Also note that

$$\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda = 1 - (1-\lambda)^{N+1}$$

hence

$$\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda \mathrm{E}\varepsilon_{\theta n}(\theta,\theta^*)\varepsilon_n(\theta,\theta^*) =$$

$$\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda(W_\theta(\theta,\theta^*)+\delta_n) = W_\theta(\theta,\theta^*)+O((1-\lambda)^N).$$

Thus we conclude that $\delta V_{\theta N} = O_M(\lambda^{1/2})+O((1-\lambda)^N)$, and here the second term on the right hand side is deterministic. Therefore we have with some $c>0$ that for $N > c/\lambda \quad O((1-\lambda)^N) < d'/2$ and hence $P(\delta V_{\theta N} > d') \leq P(O_M(\lambda^{1/2}) > d'/2) = O(\lambda^s)$ with any s by Markov's inequality, and thus the proposition at the beginning of the proof follows.

Let us now consider the random variable

$$\delta V_{\theta\theta N} = \sup_{\theta\in D,\theta^*\in D^*} \|V_{\theta\theta N}(\theta,\theta^*) - W_{\theta\theta}(\theta,\theta^*)\|.$$

By the same argument as above we have for any $d''>0$ $P(\delta V_{\theta\theta N} > d'') = O(\lambda^s)$ for $N > c/\lambda$. Hence for the event $A_N = \{\omega : \delta V_{\theta N} < d'', \delta V_{\theta\theta N} < d''\}$ we have with any $s>0$ and $N > c/\lambda$

$$P(A_N) > 1 - O(\lambda^s).$$

But on A_N the equation (1.2) has a unique solution whenever d'' is sufficiently small. Indeed, the equation $W_\theta(\theta,\theta^*)$ has a unique solution $\theta=\theta^*$ in D by [1] hence the existence of a unique solution of (1.2) can easily be derived from the implicit function theorem (cf. Lemma 4.7). Thus the lemma follows. □

Let us now consider equation (1.2) and write it as:

$$0 = V_{\theta N}(\widehat{\theta}_N,\theta^*) = V_{\theta N}(\theta^*,\theta^*) + \overline{V}_{\theta\theta N}(\widehat{\theta}_N - \theta^*) \tag{2.3}$$

where

$$\overline{V}_{\theta\theta N} = \int_0^1 V_{\theta\theta N}((1-\mu)\theta^* + \mu\widehat{\theta}_N,\theta^*)d\mu.$$

Lemma 2.3 *We have* $\widehat{\theta}_N - \theta^* = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N)$.

Proof: First we prove that $V_{\theta N}(\theta^*,\theta^*) = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N)$. Indeed, the process $(1-\lambda)^{N-n}\lambda\varepsilon_{\theta n}(\theta^*,\theta^*)e_n$ is a martingale difference process with respect to the family of $\sigma-$fields $(\mathcal{F}_n)$. Hence by Burkholder's inequality for martingales, (cf. e.g. Theorem 3.3.6 in [17]) we get that for any $q>1$

$$\mathrm{E}^{1/q}|\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda\varepsilon_{\theta n}(\theta^*,\theta^*)e_n|^q \leq$$

$$\leq C_q \mathrm{E}^{1/q} \left((\sum_{n=1}^{N} ((1-\lambda)^{N-n} \lambda \varepsilon_{\theta n}(\theta^*, \theta^*) e_n)^2 \right)^{q/2}.$$

Taking the square of both sides and using the triangle inequality for the $L_{q/2}(\Omega, \mathcal{F}, P)$-norm of the right hand side we get that the square of the right hand side is majorated by

$$C_q^2 \sum_{n=1}^{N} (1-\lambda)^{2(N-n)} \lambda^2 M_q(\varepsilon_\theta(\theta^*, \theta^*) e) = O(\lambda) + O((1-\lambda)^{2N}).$$

Taking the square root of both sides and using the inequality $(a+b)^{1/2} < a^{1/2} + b^{1/2}$ for positive a, b, we get that

$$\sum_{n=1}^{N} (1-\lambda)^{N-n} \lambda \varepsilon_{\theta n}(\theta^*, \theta^*) e_n = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N).$$

Let us now replace e_n on the left hand side by $\varepsilon_n(\theta^*, \theta^*)$. Since $\varepsilon_n(\theta^*, \theta^*) = e_n + O_M(c^n)$ with some $0 < c < 1$, it follows as in the argument following (2.2) (cf. Lemma 4.5) that for small λ the contribution of the error term is $O_M((1-\lambda)^N)$, hence the statement at the beginning of the proof follows.

Let us now investigate $\overline{V}_{\theta\theta N}$. Define

$$\overline{W}_{\theta\theta N} = \int_0^1 W_{\theta\theta}((1-\mu)\theta^* + \mu \widehat{\theta}_N, \theta^*) d\mu. \tag{2.4}$$

Obviously $\overline{W}_{\theta\theta N} > cI$ with some positive c on A_N if d is sufficiently small. Indeed since W is smooth we have for $0 \leq \alpha \leq 1$

$$\|W_{\theta\theta}(\theta^* + \mu(\widehat{\theta}_N - \theta^*), \theta^*) - W_{\theta\theta}(\theta^*, \theta^*)\| \leq C|\widehat{\theta}_N - \theta^*| < Cd \tag{2.5}$$

where C is a system's constant in the sense that it depends only on the system parameters and on the noise characteristics $M_q(e), , _q(e)$. Hence if d is sufficiently small then the positive definiteness of $W_{\theta\theta}(\theta^*, \theta^*)$ and (2.4) imply that $\overline{W}_{\theta\theta N} > cI$ with some positive c. Since we have on A_N

$$\|\overline{V}_{\theta\theta N} - \overline{W}_{\theta\theta N}\| < d''$$

it follows that if d'' is sufficiently small then

$$\lambda_{\min}(\overline{V}_{\theta\theta N}) > c > 0 \tag{2.6}$$

on A_N where in general $\lambda_{\min}(B)$ denotes the minimal eigenvalue of the matrix B. Hence $\|\overline{V}_{\theta\theta N}^{-1}\| < CN^{-1}$ on A_N with some nonrandom constant C and we get from (2.3)

$$\chi_{A_N}(\widehat{\theta}_N - \theta^*) = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N). \tag{2.7}$$

Combining this inequality with the previous inequality $P(A_N^c) = O(\lambda^s) + O_M((1-\lambda)^N)$ for any $s > 0$ where A_N^c denotes the complement of A_N and using the fact that $|\widehat{\theta}_N - \theta^*|$ is bounded we get for any $s > 0$

$$\chi_{A_N^c}(\widehat{\theta}_N - \theta^*) = O_M(\lambda^s) + O_M((1-\lambda)^N) \tag{2.8}$$

for any $s > 0$. Adding this equality to (2.7) we get the lemma. □

Now we can complete the proof of Theorem 1.1. Using the result of the last lemma we can improve the inequality (2.5) by writing $O_M(\lambda^{1/2}) + O_M((1-\lambda)^N)$ on the right hand side. Thus we get after integration with respect to μ that

$$\|\overline{W}_{\theta\theta N} - W_{\theta\theta}(\theta^*, \theta^*)\| = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N). \tag{2.9}$$

On the other hand the inequality $\delta V_{\theta\theta N} = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N)$ implies that

$$\|\overline{V}_{\theta\theta N} - \overline{W}_{\theta\theta N}\| = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N) \tag{2.10}$$

hence we finally get

$$\|\overline{V}_{\theta\theta N} - W_{\theta\theta}(\theta^*, \theta^*)\| = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N). \tag{2.11}$$

Let us now focus on the event A_N, where we have the inequality (2.6). A simple calculation shows that (2.6) and (2.11) imply

$$\chi_{A_N}\|\overline{V}_{\theta\theta N}^{-1} - W_{\theta\theta}^{-1}(\theta^*, \theta^*)\| = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N). \tag{2.12}$$

Now we can get our final estimate for $\widehat{\theta}_N - \theta^*$ by substituting (2.12) into (2.3) to obtain

$$\chi_{A_N}(\widehat{\theta}_N - \theta^*) = -\chi_{A_N}\overline{V}_{\theta\theta N}^{-1} V_{\theta N}(\theta^*, \theta^*) \tag{2.13}$$

$$= -\chi_{A_N}(W_{\theta\theta}^{-1}(\theta^*, \theta^*) + O_M(\lambda^{1/2}) + O_M((1-\lambda)^N))V_{\theta N}(\theta^*, \theta^*)$$

$$= -\chi_{A_N} W_{\theta\theta}^{-1}(\theta^*, \theta^*)V_{\theta N}(\theta^*, \theta^*) + O_M(\lambda) + O_M((1-\lambda)^N)O_M(\lambda^{1/2}) +$$

$$+O_M((1-\lambda)^{2N}) = -W_{\theta\theta}^{-1}(\theta^*, \theta^*)V_{\theta N}(\theta^*, \theta^*) + O_M(\lambda) + O_M((1-\lambda)^2 N)$$

for $N \geq c/\lambda$. The last equality is obtained by taking into account that $1 - \chi_{A_N} = O_M(\lambda^s)$ with any $s > 0$ $N \geq c/\lambda$. For the error terms we used the inequality $(a+b)^2 \leq 2(a^2 + b^2)$.

Taking into account the estimate

$$V_{\theta N}(\theta^*, \theta^*) = \sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda\varepsilon_{\theta n}(\theta^*, \theta^*)e_n + O_M((1-\lambda)^N)$$

for small λ's, substituting this estimate into (2.13) and adding (2.13) and (2.8) we get the proposition of Theorem 1.1. □

3 The Proof of Theorem 1.2

Let us denote the first term on the right hand side of (1.4) by ξ_N. i.e let us set

$$\xi_N = -(R^*)^{-1} \sum_{n=1}^{N} (1-\lambda)^{N-n} \lambda \varepsilon_{\theta n}(\theta^*, \theta^*) e_n .$$

Theorem 5.2 implies that (ξ_N) is an L_0-mixing process, and we have for all $1 \le q < \infty$ and $c > 0$ the estimations $, _{q,c}(\xi) = O(\lambda^{-1+c/2})$. Thus for the dominant term of (2.14) the properties stated for $(\widehat{\theta}_N - \theta^*)$ and (r_N) in Theorem 1.2 are valid.

We shall prove directly that the proposition of Theorem 1.2 holds for process $(\widehat{\theta}_N - \theta^*)$, i.e. for a suitable version of $\widehat{\theta}_N$ the process $(\widehat{\theta}_N - \theta^*)$ is an L_0-mixing process such that for all $1 \le q < \infty$ and $c > 0$ we have $, _{q,c}(\widehat{\theta}_N - \theta^*) = O(\lambda^{-1+c/2(q+1)})$. Combining these two statements the proposition of the theorem for the process (r_N) follows.

First let us define the event A_N in a different way as follows: define

$$\delta_N = \max\{\delta V_{\theta N}, \delta V_{\theta\theta N}\}$$

and write

$$A_N = \{\omega : \delta_N < d''\} \quad \text{and} \quad B_N = \{\omega : \delta_N \ge d''\}.$$

The "suitable" version of $\widehat{\theta}_N$ will be defined as follows. If d'' is sufficiently small then on the event A_N the nonlinear algebraic equation $V_{\theta N}(\theta) = 0$ has a unique solution which we take for $\widehat{\theta}_N$. We shall extend the definition of $\widehat{\theta}_N$ to B_N by setting $\widehat{\theta}_N = \theta_0 \in D_0$ on B_N where θ_0 is fixed in advance. We prove that with suitable choice of $d'' = d''_N$ this version of the solution is L_0-mixing with "mixing rate" indicated above.

Lemma 3.1 *The process (δ_N) is L_0-mixing and we have for all $q \ge 1$ and $c > 0$*

$$, _{q,c}(\delta) = O(\lambda^{-1+c/2}). \tag{3.1}$$

Proof: Note that the process $\varepsilon_n(\theta, \theta^*)$ and all its derivatives with respect to θ are L_0-mixing. Indeed this follows from the fact that $\varepsilon_n(\theta, \theta^*)$ and its derivatives are obtained from (e_n) by the application of an exponentially stable linear filter, hence Theorem 5.2 implies our claim. It follows that $\varepsilon_n(\theta, \theta^*)\varepsilon_n(\theta, \theta^*)$ is L_0-mixing uniformly in θ for $\theta \in D$, and also it follows that $V_N(\theta, \theta^*)$ is L_0-mixing uniformly in θ for $\theta \in D$. Similarly all derivatives of $V_N(\theta, \theta^*)$ can be shown to be L_0-mixing uniformly in θ. Moreover we have by Theorem 5.2 for all $q > 1$ and $c > 0$ $, _{q,c}(V(\theta, \theta^*)) =$

$O(\lambda^{-1+c/2})$, uniformly in θ for all $\theta \in D$, and similar estimates hold for all derivatives of $V_N(\theta, \theta^*)$ with respect to θ.

By Theorem 5.3 we conclude that for all $q \geq 1$ and $c > 0$

$$, _{q,c}(\delta V_{\theta N}) = O(\lambda^{-1+c/2})$$

and

$$, _{q,c}(\delta V_{\theta\theta N}) = O(\lambda^{-1+c/2}).$$

But then the lemma follows by the remark after Definition 5.1 since $\delta_N \stackrel{=}{\square}$ $\max(\delta V_{\theta N} \delta V_{\theta\theta N})$.

Let us now take an M such that $0 < M < N$. To find a good $\mathcal{F}_M^+$-measurable approximation of $\widehat{\theta}_N$ we consider the function

$$V_{N,M}^+(\theta) = \mathrm{E}(V_N(\theta)|\mathcal{F}_M^+).$$

Since $V_N(\theta)$ is smooth in the strong $L_2(\Omega, \mathcal{F}, P)$ topology, it follows that

$$\frac{\partial}{\partial\theta} V_{N,M}^+(\theta) = \mathrm{E}(V_{\theta N}(\theta)|\mathcal{F}_M^+) \stackrel{\Delta}{=} V_{\theta N,M}^+(\theta),$$

and similar conclusion is valid for the second derivatives. Let us now consider the equation

$$V_{\theta N,M}^+(\theta) = 0 \tag{3.2}$$

and proceed with its analysis in the same way as we did with the equation $V_{\theta N}(\theta) = 0$. So let us define

$$\delta V_{\theta N,M}^+ = \sup_{\theta \in D_0} |V_{\theta N,M}^+(\theta) - W_\theta(\theta)| \tag{3.3}$$

and

$$\delta V_{\theta\theta N,M}^+ = \sup_{\theta \in D_0} |V_{\theta\theta N,M}^+(\theta) - W_{\theta\theta}(\theta)| \tag{3.4}$$

and set

$$\delta_{N,M}^{++} = \max(\delta V_{\theta N,M}^+, \delta V_{\theta\theta N,M}^+). \tag{3.5}$$

Since $V_{\theta N}(\theta)$ is M-Lipschitz-continuous in θ the same holds for $V_{\theta N,M}^+(\theta)$, due to Jensen's inequality. The same argument applies to higher order derivatives of V. Since by convention we take separable versions of $V_{\theta N,M}^+$ and $V_{\theta\theta N,M}^+(\theta)$, the random variables $\delta V_{\theta N,M}^+$ and $\delta V_{\theta\theta N,M}^+$ are well defined, hence $\delta_{N,M}^{++}$ is also well-defined.

Let us now consider the events

$$A_{N,M}^+ = \{\omega : \delta_{N,M}^{++} < d''\} \quad \text{and} \quad B_{N,M}^+ = \{\omega : \delta_{N,M}^{++} \geq d''\}$$

where $d'' > 0$. Now an $\mathcal{F}_M^+$-measurable approximation $\widehat{\theta}_{N,M}^{++}$ of $\widehat{\theta}_N$ is defined as follows. If c is sufficiently small then the equation (3.1) has a unique solution in D for all $\omega \in A_{N,M}^+$, which we take for $\widehat{\theta}_{N,M}^{++}$. We extend the definition of $\widehat{\theta}_{N,M}^{++}$ onto $B_{N,M}$ setting $\widehat{\theta}_{N,M}^{++} = \theta_0$ there. Obviously $\widehat{\theta}_{N,M}^{++}$ is $\mathcal{F}_M^+$-measurable and we have

$$V_{\theta N,N}^+(\widehat{\theta}_{N,M}^{++}) = 0 \quad \text{on} \quad A_{N,M}^+ \quad \text{and} \quad \widehat{\theta}_N^{++} = \theta_0 \quad \text{on} \quad B_{N,M}^+ .$$

Let us now estimate $\widehat{\theta}_N - \widehat{\theta}_{N,M}^{++}$. First we consider this approximation error on the set $A_N \cap A_{N,M}^+$.

Lemma 3.2 *We have for any* $s > 0$

$$\chi_{A_N \cap A_{N,M}^+} \cdot (\widehat{\theta}_N - \widehat{\theta}_{N,M}^{++}) = O_M((1-\lambda)^{N-M}\lambda^{1/2}) + O_M(N-M)^{-s}.$$

Proof: Let us define

$$V_{\theta N,M}^{0*} = \sup_{\theta \in D_0} |V_{\theta N}(\theta) - V_{\theta N,M}^+(\theta)|.$$

We have on $A_N \cap A_{N,M}^+$ the inequality $|\widehat{\theta}_N - \widehat{\theta}_{N,M}^{++}| \le C V_{\theta N,M}^{0*}$ with some systems constant C. (cf. Lemma 4.7).

To estimate $V_{\theta N,M}^{0*}$ note that $V_{\theta N}(\theta)$ satisfies

$$V_{\theta N}(\theta, \theta^*) = \frac{1}{2} \sum_{n=1}^{N} (1-\lambda)^{N-n} \lambda \varepsilon_{\theta n}(\theta, \theta^*) \varepsilon_n(\theta, \theta^*)$$

where $\varepsilon_{\theta n}(\theta, \theta^*)$ and $\varepsilon_n(\theta, \theta^*)$ are L_0-mixing. Therefore this product is also L_0-mixing and thus we can apply the estimate (4.2) obtained in the derivation of Theorem 5.2 Thus we get for any $q \ge 1$ and $s > 0$ that

$$\mathrm{E}^{1/q} |V_{\theta N}(\theta, \theta^*) - V_{\theta N,M}^+(\theta, \theta^*)|^q = O((1-\lambda)^{N-M}\lambda^{1/2}) + O((N-M)^{-s})$$

uniformly in N and M and θ.

Now the same procedure can be repeated for the second derivatives of $V_N(\theta, \theta^*)$. Combining these two estimates and applying the maximal inequality given as Theorem 4.2, we get that for any $q \ge 1$ and $s > 0$

$$\mathrm{E}^{1/q} |V_{\theta N,M}^{0*}|^q = O((1-\lambda)^{N-M}\lambda^{1/2}) + O((N-M)^{-s}$$

uniformly in N and M, thus the lemma follows. □

On the set $B_N \cap B^{++}_{N,M}$ we have $|\widehat{\theta}_N - \widehat{\theta}^{++}_{N,M}| = 0$. Finally on the set

$$D^0_{N,M} = (A_N \cap B^+_{N,M}) \cup (A^+_{N,M} \cap B_N) \tag{3.6}$$

we use the trivial inequalities

$$|\widehat{\theta}_N - \widehat{\theta}^{++}_{N,M}| \leq K \chi_{D^0_{N,M}}$$

where K is the diameter of the set D. Hence we have altogether the following inequality:

$$|\widehat{\theta}_N - \widehat{\theta}^{++}_{N,M}| \leq C V^{0*}_{\theta N,M} + K \chi_{D^0_{N,M}} \tag{3.7}$$

Lemma 3.3 *For any $d > 0$ there exists a d''_N such that $0 < d''_N < d$ and that with $d'' = d''_N$ we have for any $q \geq 1$ and $c > 0$*

$$\sum_{M=1}^{N} \mathrm{E}^{1/q} |\chi_{x<d''_N}(\delta_N) - \chi_{x<d''_N}(\delta^{++}_{N,M})|^{qc} = O(\lambda^{-1+c/2(q+1)})$$

uniformly in N.

Proof: We can write

$$D^0_{N,M} = A_N \Delta A^+_{N,M}$$

where Δ denotes the symmetric difference operator. Now remember the definitions of A_N and $A^+_{N,M}$:

$$A_N = \{\omega : \delta_N < d''\} \quad \text{and} \quad A^+_{N,M} = \{\omega : \delta^{++}_{N,M} < d''\}.$$

We have for any $q \geq 1$

$$\mathrm{E}^{1/q} |\delta_N - \delta^{++}_{N,M}|^q \leq C \gamma'''_q(N - M, V)$$

where

$$\gamma'''_q(\tau, V) = \gamma_q(t, V) + \gamma_q(\tau, V_\theta) + \gamma_q(\tau, V_{\theta\theta}) + \gamma_q(\tau, V_{\theta\theta\theta}).$$

Now since (e_n) is L_0-mixing we conclude as in Lemma 3.2 that

$$\gamma'''_q(\tau, v) = O((1-\lambda)^\tau \lambda^{1/2}) + O(\tau)^{-s}$$

for any $s > 0$. Thus (δ_N) is an L_0-mixing process , $_{q,c}(\delta) = O(\lambda^{-1+c/2})$ and Theorem 5.7 would imply that for some $0 < d''_N < d$ the process $\kappa_N = \chi_{x<d''_N}(\delta_N)$ is L_0-mixing and for any $q \geq 1$ and $c > 0$

$$, _{q,c}(\kappa) = \sum_{M=1}^{N} \mathrm{E}^{1/q} |\chi_{x<d''_N}(\delta_N) - \chi_{x<d''_N}(\delta^+_{N,m})|^q \leq$$

$$\leq O(\lambda^{-1+c/2(q+1+c)})^{(q+1+c)/(q+1)} = O(\lambda^{-1+c/2(q+1)}). \tag{3.8}$$

Now it is easy to check that in the proof of Theorem 5.7 we can replace $x^+_{n,m} = \mathrm{E}(x_n|\mathcal{F}^+_m)$ by any $\mathcal{F}^+_M$-measurable random variable $x^{++}_{n,m}$ and to approximate $\chi_{x<d_n}(x_n)$ by $\chi_{x<d_n}(x^{++}_{n,m})$. Then in (5.6) $\gamma_r(n-m,x)$ will be replaced by $\sup_n \mathrm{E}^{1/r}|x_n - x^{++}_{n,m}|^r$. In our case x_n is replaced by δ_N and m is replaced by M, and $x^{++}_{n,m}$ is replaced by $\delta^{++}_{N,M}$. Thus Theorem 5.7 implies the claim of the lemma. □

APPENDIX I: L-mixing Processes

We summarize a few results published in [3] and used in this paper. The set of real numbers will be denoted by $\mathbb{R}$, the p-dimensional Euclidean space will be denoted by $\mathbb{R}^p$. Let $D \subset \mathbb{R}^p$ be compact domain and let the stochastic process $(x_n(\theta))$ be defined on $\mathbb{Z} \times D$, where $\mathbb{Z}$ denotes the set of natural numbers.

Definition I.1 We say that $(x_n(\theta))$ is M-bounded if for all $1 \leq q < \infty$

$$M_q(x) = \sup_{\substack{n \geq 0 \\ \theta \in D}} \mathrm{E}^{1/q}|x_n(\theta)|^q < \infty.$$

We shall also use the notation $x_n = O_M(1)$ to indicate that (x_n) is M-bounded. Moreover if c_n is a sequence of positive numbers, we write $x_n = O_M(c_n)$ if $x_n/c_n = O_M(1)$. Analogously if (x^λ_n) is a parametric family of stochastic processes, parametrized by the real-valued and positive parameter λ, then we write $x^\lambda_n = O_M(c(\lambda))$,where $c(\lambda)$ is a real-valued, positive function of λ, if for all $q \geq 1$ we have $M_q(x^\lambda) = O(c(\lambda))$. Finally, we can combine the above two notations, i.e. we can write $x^\lambda_n = O_M(c_n(\lambda))$,the definition of which is self-explanatory.

We say that a sequence of r.v. x_n tends to a r.v. x in the M-sense if for all $q \geq 1$ we have

$$\lim_{n\to\infty} \mathrm{E}^{1/q}|x_n - x|^q = 0.$$

Similarly we can define differentiation in the M-sense.

Let $(\mathcal{F}_n), n \geq 0$ be a family of monotone increasing σ-algebras, and $(\mathcal{F}^+_n), n \geq 0$ be a monotone decreasing family of σ-algebras. We assume that for all $n \geq 0, \mathcal{F}_n$ and $\mathcal{F}^+_n$ are independent. For $n \leq 0$ $\mathcal{F}^+_n = \mathcal{F}^+_0$. A typical example is provided by the σ-algebras

$$\mathcal{F}_n = \sigma\{e_i : i \leq n\} \qquad \mathcal{F}^+_n = \sigma\{e_i : i > n\}$$

where (e_i) is an i.i.d. sequence of r.v.'s.

Definition I.2 A stochastic process $(x_n(\theta)), n \geq 0$ is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ uniformly in θ if it is $\mathcal{F}_n$-progressively measurable, M-bounded and with τ being a positive integer and

$$\gamma_q(\tau, x) = \gamma_q(\tau) = \sup_{\substack{n \geq \tau \\ \theta \in D}} \mathrm{E}^{1/q} |x_n(\theta) - \mathrm{E}(x_n(\theta) | \mathcal{F}_{n-\tau}^+)|^q$$

we have for any $1 \leq q < \infty$.

$$, _q = , _q(x) = \sum_{\tau=1}^{\infty} \gamma_q(\tau) < \infty.$$

If we consider a single stochastic process (x_n) which does not depend on a parameter θ we can still use the above definition but the phrase "uniformly in θ" will be omitted.

Example Discrete time stationary Gaussian ARMA processes are L-mixing. (This can be seen using a state space representation).

Theorem I.1. (cf. Theorem 1.1 in [3]) *Let $(u_n), n \geq 0$ be an L-mixing process with $\mathrm{E}u_n = 0$ for all n and let (f_n) be a deterministic sequence. Then we have for all $1 \leq m < \infty$*

$$\mathrm{E}^{1/2m} |\sum_{n=1}^{N} f_n u_n|^{2m} \leq C_m \left(\sum_{n=1}^{N} f_n^2 \right)^{1/2} M_{2m}^{1/2}(u), _{2m}^{1/2}(u)$$

where $C_m = 2(2m-1)^{1/2}$.

Corollary *Let us take $f_n = (1-\lambda)^{N-n}\lambda$ where $0 < \lambda < 1$. Then we get*

$$E^{1/2m} \left| \sum_{n=1}^{N} (1-\lambda)^{N-n} \lambda u_n \right|^{2m} \leq C\lambda^{1/2}$$

where C depends only on m and $M_{2m}(u)$ and $, _{2m}(u)$.

Define

$$\Delta x / \Delta^\alpha \theta = |x_n(\theta + h) - x_n(\theta)| / |h|^\alpha$$

defined for $n \geq 0, \theta \neq \theta + h \in D$ with $0 < \alpha \leq 1$.

Definition I.3 The stochastic process $x_n(\theta)$ is M-Hölder-continuous in θ with exponent α if the process $\Delta x / \Delta^\alpha \theta$ is M-bounded, i.e. if for all $1 \leq q < \infty$ we have

$$M_q(\Delta x / \Delta^\alpha \theta) = \sup_{\substack{n \geq 0 \\ \theta \neq \theta + h \in D}} \mathrm{E}^{1/q} |x_n(\theta + h) - x_n(\theta)|^q / |h|^\alpha < \infty.$$

Example If $(x_n(\theta))$ is absolutely continuous with respect to θ a.s. and the gradient process $(x_n(\theta))$ is M-bounded, then $(x_n(\theta))$ is M-Hölder-continuous with $\alpha = 1$, in other words $(x_n(\theta))$ is M-Lipschitz-continuous.

Let us consider the case when $(x_n(\theta))$ is a stochastic process which is measurable, separable, M-bounded and M-Hölder-continuous in θ with exponent α for $\theta \in D$. By Kolmogorov's theorem the realizations of $(x_n(\theta))$ are continuous in θ with probability 1, hence we can define for almost all ω

$$x_n^* = \max_{\theta \in D_0} |x_n(\theta)|$$

where $D_0 \subset \mathrm{int}D$ is a compact domain. As the realizations of $x_n(\theta)$ are continuous, x_n^* is measurable with respect to $\mathcal{F}$, that is x_n^* is a random variable. We shall estimate its moments.

Theorem I.2 (cf. Theorem 3.4 in [3]) *Assume that $(x_n(\theta))$ is a stochastic process which is measurable, separable, M-bounded and M-Hölder continuous in θ with exponent α for $\theta \in D$. Let x_n^* be the random variable defined above. Then we have for all positive integers q and $s > p/\alpha$*

$$M_q(x^*) \le C(M_{qs}(x) + M_{qs}(\Delta x/\Delta^\alpha \theta))$$

where C depends only on p, q, s, α and D_0, D.

Combining theorems I.1 and I.2 we get the following theorem when $f_n = 1$ and $\alpha = 1$.

Theorem I.3 *Let $u_n(\theta)$ be an L-mixing process uniformly in $\theta \in D$ such that $\mathrm{E}u_n(\theta) = 0$ for all $n \ge 0$, $\theta \in D$ and assume that $\Delta u/\Delta\theta$ is also L-mixing, uniformly in θ, $\theta + h \in D$. Let $0 < \lambda < 1$, then we get*

$$\sup_{\theta \in D_0} |\sum_{n=1}^{N} (1-\lambda)^{N-n} \lambda u_n(\theta)| = O_M(\lambda^{1/2}).$$

Theorem I.4 *Let (u_n), $n = 0, 1 \cdots$ be an L-mixing process with respect to a pair of families of σ-algebras $(\mathcal{F}_n, \mathcal{F}_n^+)$. Define the process (x_n) by*

$$x_n = \sum_{r=0}^{n} (1-\lambda)^{n-r} \lambda u_q$$

where $0 < \lambda < 1$. Then (x_n) is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ and we have for $q \ge 1$

$$, _q(x) \le O(\lambda^{-1/2}). \qquad (I.1)$$

Proof: We can assume $\mathrm{E}u_q = 0$ for all r since this normalization of the mean does not effect $,_q(x)$. Let τ be a fixed positive integer and approximate x_n by

$$x^*_{n,n-\tau} = \sum_{r=0}^{n} (1-\lambda)^{n-r} \lambda u^+_{r,n-\tau}$$

where $u^+_{r,n-\tau} = \mathrm{E}(u_q|\mathcal{F}^+_{n-\tau})$. Then obviously $x^*_{n,n-\tau}$ is $\mathcal{F}^+_{n-\tau}$ measurable and the summation actually goes from $n-\tau$ to n since for $r < n-\tau$ we have $\mathrm{E}(u_r|\mathcal{F}^+_{n-\tau}) = 0$. Furthermore it is easy to see that

$$|x_n - x^*_{n,n-\tau}| \leq \sum_{r=0}^{n-\tau-1} (1-\lambda)^{n-r}\lambda u_q + \sum_{r=n-\tau}^{n} (1-\lambda)^{n-r}\lambda |u^0_{r,n-\tau}|$$

where $u^0_{r,n-\tau} = u_r - u^+_{r,n-\tau}$. Taking the $L_q(\Omega, \mathcal{F}, P)$ norm of the right hand side and using the corollary after Theorem I.1 for the first term we get

$$E^{1/q}|x_n - x^*_{n,n-\tau}|^q \leq (1-\lambda)^{\tau+1} O_M(\lambda^{1/2}) + (\rho * \gamma_q)(\tau)$$

where $*$ denotes discrete time convolution which is now applied to the discrete sequences $\rho = ((1-\lambda)^{\tau-1}\lambda_{\tau=1}^{\infty})$ and $(\gamma_q(\tau,u))_{\tau=1}^{\infty}$. Applying Lemma 2.1 of [3] gives

$$\gamma_q(\tau, x) \leq 2(1-\lambda)^{\tau+1} O_M(\lambda^{1/2}) + 2(\rho * \gamma)(\tau). \qquad (I.2)$$

Let us perform summation over τ form 1 to ∞. The contribution of the first term on the right hand side is $\lambda^{-1} O_M(\lambda^{1/2}) = O_M(\lambda^{-1/2})$. For the second term we apply the inequality

$$\sum_{\tau=1}^{\infty} (\tau * \gamma_q)(\tau) \leq \sum_{\tau=1}^{\infty} ((1-\lambda)^{\tau-1}\lambda) \cdot \sum_{\tau=1}^{\infty} \gamma_q(\tau, u) = ,_q(u) = O(1).$$

Thus the proof of the lemma is complete. □

Lemma I.5 (cf. Lemma 7.4 in [7]). *Let $(u_n), n \geq 0$ be an M-bounded process and define a process (x_n) by*

$$x_{n+1} = (1 - \lambda x_n) + \lambda \rho^n u_n \qquad x_0 = 0$$

where $0 < (1-\lambda) < \rho < 1$. If $0 < \rho < (1-\lambda) < 1$ then we have

$$\mathrm{E}^{1/m}|x_n|^m \leq (1-\lambda)^{n-1} \cdot M_m(u).$$

Lemma I.6 (cf. Lemma 7.5 in [7]). *Let $v_i, i = 1, 2 \dots$ be an $\mathbb{R}$ valued stochastic process such that $v_i = O_M(g_i)$, where $g_i > 0$ satisfies $\lim_{i\to\infty} g_i/g_{i-1} = 1$. Define x by*

$$x_N = \lambda x_{N-1} + v_i \qquad\qquad x_0 = 0$$

where $|\lambda| < 1$. Then $x_N = O_M(g_N)$.

Finally we formulate a simple analytical lemma which is easily derived from the implicit function theorem:

Lemma I.7 *Let $G(\theta)$ and $\delta G(\theta)$ be $\mathbb{R}^p$ -valued continuously differentiable functions, $\theta \in D \subset \mathbb{R}^p$. Assume that $G(\theta) = 0$ has a unique solution $\theta = \theta^*$ in $\mathrm{int} D_0$, where D_0 is a compact subset of D. Assume that $G_\theta(\theta^*)$ is nonsingular. Then for any $d > 0$ there exists a positive number d' such that if $|\delta G(\theta)| \le d'$ and $|\delta G_\theta(\theta)| \le d'$ for all $\theta \in D_0$ then $G(\theta) + \delta G(\theta) = 0$ has a unique solution $\widehat{\theta} \in D_0$, moreover $|\theta^* - \widehat{\theta}| \le d$ and also the inequality $|\widehat{\theta} - \theta^*| < Cd'$ holds where C depends only on G and G_0.*

Appendix II: L_0-mixing Process

Let us now define a class of stochastic processes which is smaller than the class of L-mixing processes. The notations are the same as in Definition I.2.

Definition II.1 A stochastic process $(x_n(\theta))$, $n \ge 0$ is L_0-mixing uniformly in θ (with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$) if we have for any $q \ge 1$ and any $c > 0$

$$, _{q,c} = \sum_{\tau=1}^{\infty} \gamma_q^c(\tau) < \infty.$$

Obviously the defintion extends to processes which are not parameter dependent.

Remark It is easy to see that if (x_n) is L_0-mixing processes and $(F(x))$ is a function which grows at most polynomially together with its first derivatives, then $(F(x_n))$ is also L_0-mixing. Thus e.g. if (x_n) and (y_n) are L_0-mixing then $(x_n \cdot y_n)$ is also L_0-mixing. Or if x, y are L_0-mixing then setting $z_n = \max(x_n, y_n)$ yields an L_0-mixing process. It is easy to see that $, _{q,c}(z) \le 4^c(, _{q,c}(x) + , _{q,c}(y)$.

Theorem II.1 *If a stochastic process (x_n) is L_0-mixing then we have for all $s \ge 1$*

$$\gamma_q(\tau, x) \le 4^s, _{q,1/s}(x) n^{-s}. \qquad (II.1)$$

Conversely if for all $s \ge 1$ we have $\gamma_q(\tau, x) \le C_s n^{-s}$ then (x_n) is L_0-mixing and for any $c > 0$ and $s > c$ we have

$$, _{q,c}(x) \le C_s^{1/c} s/(s-c). \qquad (II.2)$$

Proof: Define the sequence $\gamma_q^*(\tau,\mu)$ by

$$\gamma_q^*(\tau,\mu) = \min_{0\le\tau'\le\tau}(\tau',u)$$

and define for $0<c<1$

$$,\ {}^*_{q,c}(x) = \sum_{\tau=1}^{\infty}(\gamma_q^*(\tau,x))^c.$$

Obviously $\gamma^*_{q,c}(x) \le \gamma_{q,c}(x) < \infty$. Writing $(\gamma_q^*(\tau,\mu))^c = a_\tau$ the sequence a_τ is nonnegative and monotone decreasing. Since

$$A = \sum_{\tau=1}^{\infty} a_\tau = ,\ {}^*_{q,c}(x) < \infty$$

a well-known elementary result implies that $a_n \le 2A/n$. Hence we conclude that

$$(\gamma_q^*(n,x))^c \le 2,\ {}^*_{q,c}(x)/n \le 2,\ {}_{q,c}(x). \qquad (II.3)$$

Now note that for $0\le\tau'\le\tau$ we have $\gamma_q(\tau,x)\le 2\gamma_q(\tau',x)$ (cf. Lemma 2.1 in [3]), and hence we have

$$(\gamma_q(\tau,x))^c \le 2\gamma_q^*(\tau,x).$$

Combining this inequality with (II.3) and setting $c=1/s$ we get the first part of the lemma.

To prove the second part note that

$$\sum_{\tau=1}^{\infty}\gamma_q^c(\tau,c) \le \sum_{\tau=1}^{\infty} C_s^{1/c}\tau^{-s/c} \le C_s^{1/c}(1+\int_1^{\infty}\tau^{-s/c}d\tau) = C_s^{1/c}(1+\frac{1}{s/c-1})$$

from which (II.2) follows. □

We show that the class of L_0-mixing processes is closed under the operations we need in system identification. First we prove the following lemma.

Theorem II.2 *Let (u_n) be an L_0-mixing process and define (x_n) by*

$$x_n = (1-\lambda)x_{n-1} + \lambda u_n$$

with $x_0 = 0$ and $0<\lambda<1$. Then (x_n) is L_0-mixing, and we have for all $q\ge 1$ and $c>0$

$$,\ {}_{q,c}(x) = O(\lambda^{-1+c/2}).$$

Proof: We proved in (I.2) that

$$\gamma_q(\tau, x) \le 2(1-\lambda)^{\tau+1} O(\lambda^{1/2}) + 2(\rho * \gamma)(\tau) \qquad (II.4)$$

(cf. (I.2)). Write the right hand side as $a+b$, and apply the inequality $(a+b)^c \le 2^c(a^c+b^c)$ for $a, b, c > 0$. Now $a^c(\tau) = 2(1-\lambda)^{c(\tau+1)} O(\lambda^{c/2})$, hence summation over τ gives the upper bound

$$2(1-(1-\lambda)^c)^{-1} \cdot O(\lambda^{c/2}).$$

Since $(1-\lambda)^c = (1-c\lambda) + O(\lambda^2)$ for small λ's we finally get the upper bound $\lambda^{-\tau} O(\lambda^{c/2})$, i.e.

$$\sum_{\tau=1}^{\infty} (a(\tau))^c = O(\lambda^{-1+c/2}).$$

For the second term in (II.4) we apply Lemma II.1 twice. First take an s to be fixed later and consider the inequality (II.1). Then consider the sequence $(b'(\tau))$ defined by the convolution of the sequences $\rho = ((1-\lambda)^{\tau-1} \cdot \lambda)_{\tau=1}^{\infty}$ and $(\tau^{-s})_{\tau=1}^{\infty}$. Lemma I.6 implies that $b'(\tau) = O(\tau^{-s})$. Hence for fixed $c > 0$ taking $s > c$ we get that

$$\sum_{\tau=1}^{\infty} b'(\tau) = O(1)$$

and the claim of the lemma follows. □

Theorem II.3 *Let $x = (x_n(\theta))$ be as in Theorem I.2 and assume that $(x(\theta))$ and $\Delta x/\Delta\theta$ are L_0-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$. Then the process $x^* = (x_n^*)$ is also L_0-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ and we have for any $c > 0$ and $r > p$*

$$, _{q,c}(x^*) \le 2(, _{rq,c}(x)) + , _{rq,c}(\Delta x/\Delta\theta)).$$

Proof: We have for any $r > p$

$$\gamma_q(x^*, \tau) \le 2(\gamma_{rq}(x) + \gamma_{rq}(\Delta x/\Delta\theta)).$$

This inequality easily follows from Theorem I.2. Applying the inequality $(a+b)^c \le 2(a^c+b^c)$ for $a, b, c, > 0$ we arrive at the proposition of the theorem. □

It is easy to see that if (x_n) is an L-mixing process and $(f(x))$ is an absolutely continuous function which grows at most polynomially together with its first derivatives then the process (y_n) defined by $y_n = f(x_n)$ is also L-mixing. However if we take a discontinuous function f say $f(x) = \chi_{x>\delta}$, i.e, f is the characteristic function of the set $\{x > \delta\}$ then the proposition above is no longer true. In other words it is not necessarily true that the level set

$$A_{\delta,n} = \{\omega : x_n > \delta\}$$

can be well approximated by sets which are $\mathcal{F}_m^+$-measurable. However we do have some results for L_0-mixing processes. But first we need a lemma. In this lemma we use the set-theoretic notation $A\Delta B = (A\backslash B) + (B\backslash A)$.

Lemma II.4 *Let x, y be the two random variables for which the q-th moments exist with some $q \geq 1$, and let us define the sets*

$$A_\delta = \{\omega : x > \delta\} \quad \textit{and} \quad B_\delta = \{\omega : y > \delta\}.$$

Then for any $K > 0$ we have

$$P(A_\delta \triangle B_\delta) \leq C_q K^{q/(q+1)} E^{1/(q+1)} |x - y|^q$$

for all δ-s except a set of δ-s of measure at most $1/K$.

Proof: Let $\varepsilon > 0$ and consider the sets $A_\delta \setminus B_{\delta-\varepsilon}$ and $B_\delta \setminus A_{\delta-\varepsilon}$. We have

$$A_\delta \setminus B_{\delta-\varepsilon} = \{\omega : x > \delta, y \leq \delta - \varepsilon\} \subset \{\omega : |x - y| > \varepsilon\},$$

hence we have for any $q \geq 1$

$$P(A_\delta \setminus B_{\delta-\varepsilon}) \leq E|x - y|^q / \varepsilon^q.$$

Similar estimation hold for $B_\delta \setminus A_{\delta-\varepsilon}$. Now note that

$$A_\delta \triangle B_\delta \subset (A_\delta \setminus B_{\delta-\varepsilon}) \cup (B_{\delta-\varepsilon} \setminus B_\delta) \cup (B_\delta \setminus A_{\delta-\varepsilon}) \cup (A_{\delta-\varepsilon} \setminus A_\delta).$$

Setting

$$P(B_{\delta-\varepsilon} \setminus B_\delta) = P(\delta - \varepsilon \leq y < \delta) = \phi(\delta)$$

we have

$$\int_{-\infty}^{\infty} \phi(\delta) d\delta = \varepsilon.$$

This can be easily seen by Fubini's theorem. Indeed we can write

$$\phi(\delta) = \mathrm{E}\chi_{[\delta-\varepsilon,\delta)}(y) = \int_\Omega \chi_{[\delta-\varepsilon,\delta)}(y) dP$$

hence

$$\int_{-\infty}^{+\infty} \phi(\delta)d\delta = \int_{-\infty}^{+\infty} \int_{\Omega} \chi_{[\delta-\epsilon,\delta)}(y)dPd\delta =$$

$$\int_{\Omega} \int_{-\infty}^{+\infty} \chi_{[\delta-\varepsilon,\delta)}(y)d\delta dP = \int_{\Omega} \varepsilon dP = \varepsilon.$$

Since $\phi(\delta) \geq 0$ we conclude that for any $K > 0$ we have $\phi(\delta) \leq K\varepsilon$ except perhaps on a set of δ-s, say N, of measure not greater that $1/K$.

Thus we get that for $\delta \notin N$

$$P(A_\delta \triangle B_\delta) \leq 2E|x-y|^q/\varepsilon^q + 2K\varepsilon.$$

Let us now minimize the right hand side with respect to ε. In general let us consider the function

$$\psi(\varepsilon) = a\varepsilon^{-\alpha} + b\varepsilon^{\beta}.$$

Lemma II.5 *The minimized value of $\psi(\varepsilon)$ is of the form*

$$Ca^{\beta/(\alpha+\beta)}b^{\alpha/(\alpha+\beta)},$$

where C depends only on α and β.

Proof: Differentiation with respect to ε gives

$$\frac{d}{d\varepsilon}\psi(\varepsilon) = -\alpha a\varepsilon^{-\alpha-1} + \beta b\varepsilon^{\beta-1}.$$

Setting this expression equal to zero and solving the resulting equation for ε gives

$$\varepsilon = \left(\frac{\alpha a}{\beta b}\right)^{1/(\alpha+\beta)}$$

and thus

$$\psi(\varepsilon) = a\left(\frac{\beta b}{\alpha a}\right)^{\alpha/(\alpha+\underline{e}ta)} + b\left(\frac{\alpha a}{\beta b}\right)^{\beta/(\alpha+\beta)}.$$

Taking out the powers of a and b we get

$$\psi(\varepsilon) = Ca^{\beta/(\alpha+\beta)}b^{\alpha/(\alpha+\beta)}$$

where C depends only on α and β, which proves the lemma. □

Applying Lemma II.5 with $\alpha = q, \beta = 1$ $a = 2\mathrm{E}|x-y|^q$ and $b = 2k$ we get the proposition of Lemma II.4. □

Let $\|x-y\|_{L_q} = \mathrm{E}^{1/q}|x-y|^q$. Then we can write Lemma II.4 as follows:

$$P(A_\delta \Delta B_\delta) \le C_q K^{q/(q+1)} \|x-y\|_{L_q}^{q/(q+1)}.$$

Lemma II.6 *With the notation of Lemma II.4 we have for any $q \ge 1$ and $L > 0$ and $\varepsilon > 0$*

$$P(A_\delta \Delta B_\delta) \le C_q L^{q/(q+1)} \|x-y\|_{L_q}^{(1-\varepsilon)q/(q+1)}$$

except for a set of δ's of measure at most $\|x-y\|_{L_q}^{\varepsilon}/L$.

Proof: Setting $K = L\|x-y\|_{L_q}^{-\varepsilon}$ we get from Lemma II.4

$$P(A_\delta \Delta B_\delta) \le C_q L^{q/(q+1)} \|x-y\|_{L_q}^{-\varepsilon q/(q+1)} \|x-y\|_{L_q}^{q/(q+1)}$$

which simplifies to the expression given in the lemma. □

Theorem II.7 *Let (x_n), $n \ge 0$ be an L_0-mixing process, and let $I \subset \mathbb{R}$ be a fixed nonempty open interval. Then there exist a sequence of real numbers $\delta_n \varepsilon I$ such that the process $y_n = \chi_{x>\delta_n}(x_n)$ is L_0-mixing, and we have for any $r \ge 1$ and $c > 0$*

$$, _{r,c}(y) \le 2C_0, _{r,c/(r+1+c)}^{(r+1+c)/(r+1)}(x)$$

where C_0 depends only on I and r.

Proof: Let us apply Lemma II.4 with $x = x_n$, $y = x_{n,m}^+$ where $0 \le m \le n$. Then $\mathrm{E}|x-y|^r \le \gamma_r^r(n-m,x)$. To make sure that the estimation provided by Lemma II.4 is simple we choose with $K = L\gamma_r^{-\varepsilon}(n-m,x)$, where L and ε are to be fixed later. We shall approximate $\chi_{x>\delta}(x_n)$ by $\chi_{x>\delta}(x_{n,m}^+)$. Then we have with A_δ and B_δ defined as in Lemma II.4

$$\mathrm{E}|\chi_{x>\delta}(x_n) - \chi_{x>\delta}(x_{n,m}^+)| = P(A_\delta \Delta B_\delta) \le$$

$$\le C_r L^{-r/(r+1)} \gamma_r^{(1-\varepsilon)r/(r+1)}(n-m,x), \qquad (II.5)$$

except for a set of δ's of measure at most $1/K = \gamma_r^{\varepsilon}(n-m,x)/L$. For fixed n and running m the union of the sets of exceptional δ's has measure at most

$$\sum_{\tau=1}^{\infty} \gamma_r^{\varepsilon}(\tau,x) = , _{r,\varepsilon}(x)/L.$$

To ensure the existence of a non-exceptional $\delta_n \varepsilon I$ we choose L so that $, _{r,\varepsilon}(x)/L < |I|$. Let us set $L = 2, _{r,\varepsilon}(x)/|I|$.

Let us substitute this value of L into $(I.1)$ and take the r-th root of both sides with some $r \geq 1$. We get for some $\delta_n \in I$

$$E^{1/r}|\chi_{x>\delta_n}(x_n) - \chi_{x>\delta_n}(x^+_{n,m})| \leq C_r^{1/r}(2/|I|)^{1/(r+1)} \cdot$$
$$\cdot, \, {}^{1/(r+1)}_{r,\varepsilon}(x) \cdot \gamma_r^{(1-\varepsilon)/(r+1)}(n-m,x). \qquad (II.6)$$

Since $y^{00}_{n,m} \stackrel{\Delta}{=} \chi_{x>\delta_n}(x_n) - \chi_{x>\delta_n}(x^+_{n,m})$ is a random variable with values 0 or ± 1 the left hand side of (II.6) can also be written as $\mathrm{E}^{1/r}|y^{00}_{n,m}|^r$.

Now applying Lemma 2.1 of [3] we get that $\mathrm{E}^{1/r}|y_n - \mathrm{E}(y_n|\mathcal{F}^+_m)|^r$ and hence $\gamma_r(n-m,y)$ is majorated by twice the right hand side of (II.6). Hence for any $c > 0$ we get

$$\gamma_r^c(n-m,y) \leq 2^{1/c}C_r^{c/r}(2/|I|)^{c/(r+1)} \, , \, {}^{c/(r+1)}_{r,\varepsilon}$$
$$\cdot \gamma_r^{c(1-\varepsilon)/(r+1)}(n-m,x). \qquad (II.7)$$

Let us set

$$C_0 = 2^{1/c}C_r^{c/r}(2/|I|)^{c/(r+1)}$$

and let us now choose ε so that we have $\varepsilon = c(1-\varepsilon)r/r(r+1)$, i.e. we set $\varepsilon = \varepsilon^* = c/(r+1+c)$. Then summation of the inequalities (II.7) over $\tau = n - m$ gives

$$\sum_{\tau=1}^{\infty} \gamma_r^c(\tau,y) \leq C_0 \, , \, {}^{c/(r+1)}_{r,\varepsilon}(x) \cdot \, , \, {}_{r,\varepsilon}(x).$$

The exponent of $, \, {}_{r,\varepsilon}$ will be $c/(r+1)+1 = (r+1+c)/(r+1)$ which was the proposition of the theorem. □

References

[1] K.J. Åström and T. Söderström. Uniqueness of the maximum-likelihood estimates of the parameters of an ARMA model, *IEEE Transactions on Automatic Control* **AC-19** (1974), 769–773.

[2] P.E. Caines. *Linear Stochastic Systems*. New York: John Wiley, 1988.

[3] L. Gerencsér. On a class of mixing processes, *Stochastics* **26** (1989), 165–191.

[4] L. Gerencsér. On Rissanen's predictive stochastic complexity for stationary ARMA processes, *J. of Statistical Planning and Inference*, 1993. To appear.

[5] L. Gerencsér. On the martingale approximation of the estimation error of ARMA parameters, *Systems & Control Letters* **15** (1990), 417–423.

[6] L. Gerencsér. Predictive stochastic complexity associated with fixed gain estimators, in *Proceedings of the 2-nd European Control Conference*, (J.W. Nieuwenhuis, C. Praagman and H.L. Trentelman, eds.) (1993), 1673–1677.

[7] L. Gerencsér. Rate of convergence of recursive estimators, *SIAM J. of Control and Optimization* **30(5)** (1992), 1200–1227.

[8] L. Gerencsér. Strong approximation results in estimation and adaptive control, in (L. Gerencsér and P. Caines, eds.), *Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control*, 268–299, Lecture Notes in Control and Information Sciences, Springer, 1991.

[9] L. Gerencsér and J. Baikovicius. Change point detection using stochastic complexity, in *Preprint of the 9th IFAC-IFORS Symposium on Identification and System Parameter Estimation* (1991), 395–400.

[10] L. Gerencsér and J. Baikovicius. Model selection, stochastic complexity and badness amplification, in *Proceedings of the 30-th IEEE Conference on Decision and Control* (1991), 1999–2004.

[11] L. Gerencsér and J. Rissanen. Asymptotics of predictive stochastic complexity: from parametric to nonparametric models, in (D. Brillinger, C. Caines, J. Geweke, E. Parzen, M. Rosenblatt, and M. Taqqu, eds.), *New Directions in Time-series Analysis*, The IMA Volumes in Mathematics and its Applications, Volume 46, 1993.

[12] L. Ljung. *Identification. Theory for the user.* Englewood Cliffs, NJ: Prentice-Hall, Inc., 1987.

[13] L. Ljung and T. Söderström. *Theory and Practice of Recursive Identification.* Cambridge: The MIT Press, 1983.

[14] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publisher, 1989.

[15] T. Söderström and P. Stoica. *System Identification.* New York: Prentice Hall, 1989.

[16] T. Söderström and P. Stoica. Uniqueness of prediction error estimates of multivariable moving average models, *Automatica* **18** (1982), 617–620.

[17] W.F. Stout. *Almost Sure Convergence.* New York: Academic Press, 1974.

[18] Vágó Zs. and L. Gerencsér. Uniqueness of the Maximum-Likelihood estimates of the Kalman-gain matrix of a state space model, in *Proceedings of the IFAC/IFORS Conference on Dynamic Modelling of National Economics* (1985), 483–486, Budapest.

Computer and Automation Institute, Hungarian Academy of Sciences H-1111, Budapest Kende u 13-17, Hungary

Communicated by Clyde F. Martin