Estimating and forecasting periodically partially linear models with exogeneous Variables

Xavier Brossat, Georges Oppenheim, Marie-Claude Viano

Pub. Inst. Stat. Univ. Paris
57, fasc. 3, 2013, 55-82

Estimating and forecasting periodically partially linear models with exogeneous variables

Xavier Brossat[1] Georges Oppenheim[2] and Marie-Claude Viano[3]

[1] Département OSIRIS – Service de Recherche et Développement – EDF
1, av. du Général de Gaulle 92140 Clamart, France

[2] Laboratoire d'Analyse et de Mathématiques Appliquées Université Paris-Est
5, bd. Descartes, Cité Descartes - Champs-sur-Marne
77454 Marne-la-Vallée cedex 2, France

[3] Laboratoire Paul Painlevé UMR CNRS 8524 – Bat M2.
Université Lille 1, Villeneuve d'Ascq, 59655 Cedex, France.

Abstract

This paper presents a backfitting-type method for estimating and forecasting a periodically correlated partially linear model with exogeneous variables and heteroskedastic input noise, in the situation of hidden periodicity. A rate of convergence of the estimator is given.

keywords: α-mixing, additive models, autoregressive models, cyclostationarity, backfitting, electricity consumption, forecasting interval, semiparametric regression smoothing.

1 Introduction

In this paper, we focus on partially linear models of the type

$$X_n = a_1 X_{n-1} \ldots + a_p X_{n-p} + b(e_n) + \sigma(e_n)\varepsilon_n, \qquad n \in \mathbb{Z}. \tag{1.1}$$

The parameter $p \geq 1$ is supposed known while the coefficients a_j as well as the functions b and σ are unknown. The sequence (ε_n) is an unobserved system noise. The aim is to estimate the model in order to predict X_{n+1}, from the observed set $((X_n, e_n), (X_{n-1}, e_{n-1}), \ldots)$ of past values available at date n.

The choice of the simple model (1.1) is made for convenience, and the algorithm presented below could easily be adapted to the general case

$$X_n = \sum_{j=1}^{p} a_j X_{n-j} + \sum_{j=0}^{q} b_j(e_{n-j}) + \sigma(e_n, \ldots, e_{n-q'})\varepsilon_n. \tag{1.2}$$

During the last 20 years, partially linear autoregressive models such as (1.2) have gained attention, as being a good compromise between linear models and purely non parametric ones. Such models, proposed in [7] to represent the relationship between weather and electricity consumption are now widely used in the literature. See for example [16] where a chapter is devoted to models including (1.2). The functions b_j are expanded on a suitable basis and the first coefficients of this expansion, together with the a_j's, are estimated via a Least Mean Square method as in [17]. With the same type of partially linear models, [8, 12] use wavelets in the

estimation scheme. In [2], the b_j's are treated as nuisance parameters. Let us also mention [13, 14, 15], devoted to models including purely autoregressive ones, where some past values operate in a linear form and the others in a functional one. These authors use an orthogonal series method, and propose a data based criterion to determine the truncation parameters. See also chapter 8 in [10], where models like

$$X_{n+1} = f(X_n) + aX_{n-1} + \sigma(X_n)\varepsilon_n.$$

include linear and non linear autoregressive summands together with some volatility. The functional parts are estimated via local linear estimators and gaussian limits for the renormalized errors are obtained.

Model (1.2) presents several advantages. Firstly, the additive form reduces the so-called curse of dimensionality. Secondly, linear autoregression is preserved when expressing the future values $(X_{n+h})_{h=1,\ldots}$ from the past ones $(X_{n-h}, e_{n-h})_{h=0,\ldots}$, which makes it easier, and in some sense coherent, forecast at lags greater than 1. Lastly, model (1.2) is specially well adapted to the situation where the output X_n is electricity consumption at date n and the input e_n is the temperature at the same date, since it is well-known that the effect of temperature on electricity sales is highly non-linear at extreme temperatures, while linearity of the autoregression seems to be a reasonable assumption. Notice that, in practical situations, the temperature at date n is either measured or forecasted by Météo-France. In both cases, the value of the exogeneous variable e_n is known. Accurate electrical load forecasting is essential for power utilities. Electricité de France (EDF) performs a climatic correction. The influence of a smoothed version of temperature on electric demand is widely reported.

2 Elements of discussion

2.1 Backfitting

Backfitting methods, first proposed by [4], are usually recommended for additive models which involve several explanatory variables, each having an unknown functional form. The method is well described in [10, 18]. See also [9, 22, 23] where the estimation algorithms use local polynomial regression and [21] based on projections on polynomial spaces. The performances of backfitting procedures when autoregression is involved are less well understood. In [29], for the non linear stationary autoregressive model with exogeneous variables

$$X_n = a(X_{n-1}) + b(e_n) + \varepsilon_n,$$

the algorithm works in two steps: the first step builds a preliminary estimator of a et b by piecewise constant functions. Then, from the obtained pseudo remainders, the second step builds kernel estimators of the same functions. The author obtains the limit law for the estimation error.

In the present paper, within the backfitting iterations, kernel-based statistics estimate the functional part of the model.

2.2 Why use backfitting?

Return now to the model (1.2). If the period T of s_n, when the input is $e_n = s_n + \eta_n$ (see hypothesis $\mathcal{H}_1$ below), was known, a simple estimation scheme would consist in splitting the data in T subsamples, each of them being a trajectory of a stationary process. Then the parameter $\theta = {}^t(a_1, \ldots, a_p)$ and the function b could be estimated separately, the first one at the usual parametric rate, and the second one at the slower usual functional rate (see [10, 27] for remarks on this question).

We try a backfitting scheme which presents the advantage of allowing the period to remain unknown, a situation in which the above method is impracticable.

Moreover, we conjecture that this backfitting scheme still works even when the period shows slight variations. This is indicated in a set of current simulations. This simultaions are not presented below since the theoretical developments on this point are still in work.

As it will be shown in the sequel, the price to pay for using this backfitting procedure is a slower rate in the estimation of θ.

3 Estimation of the parametric and non parametric components

The aim is to estimate the functions b and σ and the vector parameter

$$\theta = {}^t(a_1, \ldots, a_p).$$

Denoting

$$\phi_k = {}^t(X_{k-1}, \ldots, X_{k-p}),$$

the model can be written

$$X_n = {}^t\phi_n\theta + b(e_n) + \sigma(e_n)\varepsilon_n. \tag{3.3}$$

We choose a kernel K, and a smoothing parameter h_n.

Starting from an initial value of θ and a stopping rule, the iterative method consists of estimating θ (resp. b) by using an estimation of the residual calculated from the previous estimation of b (resp. θ).

- Initialisation. Fix the first value $\hat{\theta}^{(1)}$
- Step 1. Estimate the function b by a kernel estimator based on the partial residuals

$$\hat{b}_n^{(1)}(e) = \frac{\sum_{l=p+1}^{n}\left(X_l - {}^t\phi_l\hat{\theta}^{(1)}\right)K_n(e-e_l)}{\sum_{l=p+1}^{n-1}K_n(e-e_l)}$$

where

$$K_n(e) := K\left(\frac{e}{h_n}\right).$$

- Step 2. Update the estimation of θ by a least mean squares estimator based on the new partial residuals

$$\begin{aligned}\hat{\theta}_n^{(2)} &= \mathrm{Argmin}_\theta \sum_{l=p+1}^{n}(X_l - {}^t\phi_l\theta - \hat{b}_n^{(1)}(e_l))^2 \\ &= \Sigma_n^{-1}\sum_{l=p+1}^{n-1}\phi_l(X_l - \hat{b}_n^{(1)}(e_l))\end{aligned}$$

with

$$\Sigma_n = \sum_{l=p+1}^{n}\phi_l^t\phi_l. \tag{3.4}$$

Finally, the transition from step $k-1$ to step k can be expressed as

$$\hat{b}_n^{(k-1)}(e) = \frac{\sum_{l=p+1}^{n}\left(X_l - {}^t\phi_l\hat{\theta}^{(k-1)}\right)K_n(e-e_l)}{\sum_{l=p+1}^{n-1}K_n(e-e_l)} \tag{3.5}$$

$$\hat{\theta}_n^{(k)} = \Sigma_n^{-1}\sum_{l=p+1}^{n}\phi_l(X_l - \hat{b}_n^{(k-1)}(e_l)). \tag{3.6}$$

- Chosing a stopping time k for the iterations, the variance $\sigma^2(e)$ is then estimated via a kernel method using the partial residuals based on the estimates $\hat{\theta}_n^{(k)}$ and $\hat{b}_n^{(k-1)}$

$$\hat{\sigma}_{n,k}^2(e) = \frac{\sum_{l=p+1}^{n-1}\left(X_l - {}^t\phi_l\hat{\theta}_n^{(k)} - \hat{b}_n^{(k-1)}(e_l)\right)^2 K_n(e-e_l)}{\sum_{l=p+1}^{n-1}K_n(e-e_l)} \tag{3.7}$$

As in the case of linear regression, estimating θ and b does not require any estimation of σ, implying that $\hat{\sigma}_{n,k}$ is obtained at the end of the iterative scheme. See [10] for remarks on this so-called oracle effect.

4 Main results

4.1 Hypotheses

We adopt the following basic hypotheses ($\mathcal{H}$).

- $\mathcal{H}_1$: Periodicity. The exogeneous sequence (e_n) is the sum of a periodic deterministic sequence (s_n) and a bounded zero-mean strong white noise

$$e_n = s_n + \eta_n \quad \forall n \tag{4.8}$$

- $\mathcal{H}_2$: Whiteness of the system noise. (ε_n) is an i.i.d sequence of zero-mean variables, and $\mathrm{Var}(\varepsilon_n) = 1$.

- $\mathcal{H}_3$: Stability. The autoregressive dynamic is stable. In other words, the polynomial

$$A(z) = z^p - \left(\sum_{j=1}^{p} a_j z^{p-j} \right)$$

does not vanish on the domain $|z| \geq 1$.

- $\mathcal{H}_4$: Independence of the inputs. The two sequences (ε_n) and (η_n) are independent.

- $\mathcal{H}_5$: On the distributions of input sequences. The distributions of ε_1 and η_1 both have a density. The density f of η_1 is continuous and non-vanishing on the support $[-m_\eta, m_\eta]$ of η_1. The density g of ε_1 is C_1 on $\mathbb{R}$.

- $\mathcal{H}_6$: On the moments. The noise ε_1 has a finite moment of order $\nu > 2$

$$\int |u|^\nu g(u) du < \infty \tag{4.9}$$

- $\mathcal{H}_7$: On the functions. Let $\mathcal{E} = \cup_{j=1}^{T} [s_j - m_\eta, s_j + m_\eta]$ denote the union of the T compact supports of the variables e_j.

  1. The function b is γ-Hölderian on $\mathcal{E}$, for some $0 < \gamma \leq 1$, which means that

$$\sup_{e_1, e_2 \in \mathcal{E}} \frac{|b(e_1) - b(e_2)|}{|e_1 - e_2|^\gamma} < \infty \tag{4.10}$$

  2. The variance $\sigma^2(e)$ of the input noise is γ_1-Hölderian on $\mathcal{E}$, for some $0 < \gamma_1 \leq 1$, and

$$\inf_{e \in \mathcal{E}} \sigma(e) > 0. \tag{4.11}$$

- $\mathcal{H}_8$: On the kernel. The kernel K is lipschitzian, and satisfies

$$\int K(u) du = 1$$

Keeping in mind the example of electricity consumption, hypothesis $\mathcal{H}_1$ allows some periodicity in the random structure of the input sequence (e_n). Boundedness of the noise η_n (hypothese $\mathcal{H}_1$) is useful to get uniform convergence in Theorem 2 (without this condition, uniformity only holds on compact sets) and hence to obtain results for the predictor. Hypothese $\mathcal{H}_5$ assures that the denominators of the kernel-type estimators of b and σ are not asymptotically vanishing. Hypothese $\mathcal{H}_5$ also allows to apply a Fuk-Nagaev-type inequality for mixing sequences. The Hölder exponents γ and γ_1 in hypothesis $\mathcal{H}_7$ govern the convergence rate of the estimation scheme.

In what follows, we work with the periodically correlated (or cyclostationary) solution of (3.3) defined in Section 7.1.

4.2 Existence of $\hat{\theta}_n^{(k)}$

The lemma below establishes that, almost surely, the matrix $\Sigma_n = \sum_{l=p+1}^{n} \phi_l^t \phi_l$ appearing in (3.4) and used in estimating the parameter θ is invertible at least for large enough n.

Lemma 1. *Under the hypotheses $\mathcal{H}_{1,2,3,4}$, the matrix Σ_n being defined in (3.4), as $n \to \infty$,*

(i)

$$\frac{\Sigma_n}{n} \xrightarrow{a.s.} M = \frac{1}{T}\sum_{l=0}^{T-1} \mathbb{E}\left(\phi_0^{(l)t}\phi_0^{(l)}\right) = \frac{1}{T}\sum_{l=0}^{T-1}\left[\mu^{(l)t}\mu^{(l)} + \Gamma^{(l)}\right]$$

where

$$\phi_k^{(l)} = {}^t(X_{kT+l}, X_{kT+l-1}, \ldots, X_{kT+l-(p-1)}). \tag{4.12}$$

and where $\mu^{(l)} = \mathbb{E}(\phi_0^{(l)})$ and $\Gamma^{(l)}$ is the covariance matrix of $\phi_0^{(l)}$.

(ii) The limit matrix M is regular.

The proof is in the Appendix.

4.3 Analysis of estimation errors

We first focus on the estimation errors

$$\tilde{\theta}_n^{(k)} = \theta - \hat{\theta}_n^{(k)} \quad \text{and} \quad \tilde{b}_n^{(k-1)}(e) = b(e) - \hat{b}_n^{(k-1)}(e). \tag{4.13}$$

From (3.3),

$$\tilde{\theta}_n^{(k)} = -\Sigma_n^{-1}\sum_{l=p+1}^{n} \phi_l\left(\tilde{b}_n^{(k-1)}(e_l) + \sigma(e_l)\varepsilon_l\right) \tag{4.14}$$

$$\tilde{b}_n^{(k-1)}(e) = \frac{\sum_{l=p+1}^{n}\left(-{}^t\phi_l\tilde{\theta}_n^{(k-1)} + b(e) - b(e_l) - \sigma(e_l)\varepsilon_l\right)K_n(e-e_l)}{\sum_{l=p+1}^{n} K_n(e-e_l)} \tag{4.15}$$

.

Thanks to the linearity of $\tilde{b}_n^{(k-1)}(e)$ with respect to $\tilde{\theta}_n^{(k-1)}$, this leads to the linear recursive equation

$$\tilde{\theta}_n^{(k)} = A_n\tilde{\theta}_n^{(k-1)} + R_n^{(1)} + R_n^{(2)} \tag{4.16}$$

where

$$A_n = \Sigma_n^{-1}\left(\sum_{l=p+1}^{n}\phi_l\frac{\sum_{j=p+1}^{n}{}^t\phi_j K_n(e_l-e_j)}{\sum_{j=p+1}^{n}K_n(e_l-e_j)}\right) \tag{4.17}$$

$$R_n^{(1)} = \Sigma_n^{-1}\sum_{l=p+1}^{n}\phi_l\frac{\sum_{j=p+1}^{n}(b(e_j)-b(e_l)+\sigma(e_j)\varepsilon_j)K_n(e_l-e_j)}{\sum_{j=p+1}^{n}K_n(e_l-e_j)} \tag{4.18}$$

$$R_n^{(2)} = \Sigma_n^{-1}\sum_{l=p+1}^{n}\phi_l\sigma(e_l)\varepsilon_l \tag{4.19}$$

4.4 Convergence results

Considering (4.16), we first prove that, as $n \to \infty$, the matrix operator A_n converges to a strictly shrinking matrix A. See Lemma 8 in the Appendix. As emphazised in [4] this is the key result implying that $\tilde{\theta}_n^{(k)}$ stabilizes as k increases. Then we prove that the remainder term $R_n^{(1)} + R_n^{(2)}$ tends to zero (see Lemma 6 and 7 in the Appendix). This implies that the stabilizing value $\tilde{\theta}_n^{(\infty)}$ in turn vanishes when $n \to \infty$. All this leads to the main result, whose detailed proof is in the Appendix.

Theorem 2. *With the assumptions of Section 4.1, if the smoothing parameter is such that, as* $n \to \infty$ $h_n \sim n^{\beta_1}(\ln n)^{\beta_2}$, *and if* $\beta_1 \in]-1, 0[$ *is linked with the parameter* ν *in (4.9) by*

$$\nu(1+\beta_1) > 2,$$

there exists $\beta \in]0, 1[$ *such that*

$$\left.\begin{array}{l} \|\hat{\theta}_n^{(k)} - \theta\|_2 \\ \sup_{e\in\mathcal{E}} |\hat{b}_n^{(k)}(e) - b(e)| \end{array}\right\} = O_{a.s.}\left(\sqrt{\frac{\ln n}{n h_n}}\right) + O_{a.s.}(h_n^{\gamma}) + O_{a.s.}(\beta^k) \tag{4.20}$$

and

$$\sup_{e\in\mathcal{E}} |\hat{\sigma}_{n,k}^2(e) - \sigma^2(e)| = O_{a.s.}\left(\sqrt{\frac{\ln n}{n h_n}}\right) + O_{a.s.}(h_n^{\min\{\gamma,\gamma'\}}) + O_{a.s.}(\beta^k) \tag{4.21}$$

where the $0(.)$*'s are uniform with respect to* k *and* n.

We see that the convergence rate of $\hat{\sigma}_{n,k}^2(e)$ cannot exceed the rate of the other parameters and can even be slower when $b(e)$ is smoother than $\sigma(e)$. The equality (4.21) is proved in the Appendix.

As a result, an optimal order of the rate is obtained by chosing suitable values for β_1 and β_2.

Corollary 3. *Under the same hypotheses, if* $h_n \sim (\ln n/n)^{\frac{1}{2\gamma+1}}$, *and provided that* $\nu > 2 + \frac{1}{\gamma}$, *there exists* $\beta \in]0, 1[$ *such that*

$$\left.\begin{array}{l} \|\hat{\theta}_n^{(k)} - \theta\|_2 \\ \sup_{e\in\mathcal{E}} |\hat{b}_n^{(k)}(e) - b(e)| \end{array}\right\} = O_{a.s.}\left(\frac{\ln n}{n}\right)^{\frac{\gamma}{2\gamma+1}} + O_{a.s.}(\beta^k)$$

and

$$\sup_{e\in\mathcal{E}} |\hat{\sigma}_{n,k}^2(e) - \sigma^2(e)| = O_{a.s.}\left(\frac{\ln n}{n}\right)^{\frac{\min\{\gamma,\gamma'\}}{2\gamma+1}} + O_{a.s.}(\beta^k)$$

It is clear that, provided β is not too close to 1, the convergence of the term β^k to zero is fast. In other words, stabilisation of the iterations is easily obtained while convergence of $\left(\frac{\ln n}{n}\right)^{\frac{\gamma}{2\gamma+1}}$ to zero requires large sample size. More precisely, taking $k = k(n) \geq C \ln n$ gives

Corollary 4. *Under the same hypotheses as in Corollary 3 and with* $h_n \sim (\ln n/n)^{\frac{1}{2\gamma+1}}$, *if the recursive scheme stops after* $k(n) \geq C \ln n$ *iterations*

$$\left.\begin{array}{l} \|\hat{\theta}_n^{(k(n))} - \theta\|_2 \\ \sup_{e\in\mathcal{E}} |\hat{b}_n^{(k(n))}(e) - b(e)| \end{array}\right\} = O_{a.s.}\left(\frac{\ln n}{n}\right)^{\frac{\gamma}{2\gamma+1}}$$

and

$$\sup_{e\in\mathcal{E}} |\hat{\sigma}_{n,k(n)}^2(e) - \sigma^2(e)| = O_{a.s.}\left(\frac{\ln n}{n}\right)^{\frac{\min\{\gamma,\gamma'\}}{2\gamma+1}}$$

Remark 1. With the above remark in mind, it is interesting to note that, when the autoregression is close to the instability domain, the value of β in Corollary 3 can approach 1. In such situations, a large number of iterations is needed before the stabilisation of the iterative scheme. For example consider the particular model

$$X_n = aX_{n-1} + b(e_n) + \varepsilon_n$$

where the sequence (e_n) is i.i.d. From Lemma 8, the limit matrix of A_n in (4.17) is

$$A = \frac{\mathbb{E}(X_n)^2}{\mathbb{E}(X_n)^2 + \sigma(0)} = \frac{1}{1 + \frac{\sigma(0)}{\mathbb{E}(X_n)^2}},$$

where $\sigma(0) = c^2/(1-a^2)$ and $\mathbb{E}(X_n) = c'/(1-a)$. Hence,

$$A = \frac{1}{1 + C\frac{1-a}{1+a}} \to 1 \quad \text{if} \quad a \to 1.$$

Consequently, the iterative scheme can be very slow if a is close to 1. On the opposite, when a is close to -1, the iterations stabilize very quickly.

4.5 Improvement of the rate for smooth functions b

As well-known in functional estimation, a smoother b induces, with some extra conditions on the kernel K, a better rate of convergence of the estimators.

Corollary 5. *If the function b is C_ℓ for some integer $\ell > 1$ and if the kernel satisfies*

$$\int e^k K(e)de = 0 \quad \forall k \in 1, \ldots, \ell \tag{4.22}$$

$$\int K(e)de = 1 \tag{4.23}$$

(i) if the smoothing parameter is such that, as $n \to \infty$, $h_n \sim n^{\beta_1}(\ln n)^{\beta_2}$ there exists $\beta \in]0,1[$ such that

$$\|\hat{\theta}_n^{(k)} - \theta\|_2 = O_{a.s.}\left(\sqrt{\frac{\ln n}{nh_n}}\right) + O(h_n^\ell) + O_{a.s.}(\beta^k)$$

$$\sup_{e\in\mathcal{E}} |\hat{b}_n^{(k)}(e) - b(e)| = O_{a.s.}\left(\sqrt{\frac{\ln n}{nh_n}}\right) + O(h_n^\ell) + O_{a.s.}(\beta^k)$$

(ii) if $h_n \sim (\ln n/n)^{\frac{1}{2\ell+1}}$ the rate of the two first terms is optimal and becomes

$$O_{a.s.}\left(\frac{\ln n}{n}\right)^{\frac{\ell}{2\ell+1}},$$

The proof, based on the fact that, using (4.22), $\int (b(vh_n + e) - b(e))K(v)f(vh_n + e)dv = O(h_n^\ell)$, is omitted.

5 Forecasting intervals

The natural predictor for X_{n+1},

$$\mathbb{E}(X_{n+1}|e_{n+1}, e_n, \ldots, e_1, X_n, \ldots, X_1) = {}^t\phi_{n+1}\theta + b(e_{n+1})$$

can be evaluated via the estimates of θ and b based on the observations up to time n with a stopping rule as in Corollary 4. In other words, we propose with notations of Corollary 4, the predictor

$$\hat{X}_{n+1} = {}^t\phi_{n+1}\hat{\theta}_n^{(k(n))} + \hat{b}_n^{(k(n))}(e_{n+1}).$$

It should be clear that, under the conditions of Corollary 3,

$$\frac{\hat{X}_{n+1} - X_{n+1}}{\hat{\sigma}^2_{n,k(n)}(e_{n+1})} \xrightarrow{\mathcal{L}} \varepsilon_1,$$

and, consequently, building a prediction interval requires an estimation of the noise's quantile function $Q(t)$. The inverse of Q can be consistently estimated by

$$\hat{Q}_n^{-1}(a) = \frac{1}{n}\sum_{j=1}^{n-1} \mathbb{I}_{\left|\frac{\hat{X}_{j+1,n} - X_{j+1}}{\hat{\sigma}^2_{n,k(n)}(e_{j+1})}\right| > a},$$

based on the set of retroactive predictions $\hat{X}_{j+1,n} = {}^t\phi_{j+1}\hat{\theta}_n^{(k(n))} + t\hat{b}_n^{(k(n))}(e_{j+1}), \quad j \leq n-1$. which use the estimates available at time n.

Summarizing, for X_{n+1} we obtain the prediction interval at asymptotic level α

$$\left[\hat{X}_{n+1} - \hat{\sigma}^2_{n,k(n)}(e_{n+1})\hat{Q}_n(\alpha)\ ,\ \hat{X}_{n+1} + \hat{\sigma}^2_{n,k(n)}(e_{n+1})\hat{Q}_n(\alpha)\right]$$

6 Simulation examples: results and comments

Let us mention some questions of interest when performing the data simulations part.

• Question 1. What is the influence of a single irregular point of the function b on the estimation at the regular points?

• Question 2. The results are proved when both $k \to \infty$ and $n \to \infty$. Could we use a simplified procedure base on one iteration $k = 1$ of the inner loop? Do we actually get benefit from the iteration process? Remember we do not use any preliminary estimator. From a practical point of view, can we get any information linking the stochastic process dependencies to a good value of k?

• Question 3. When the EDF engineers estimate a model, the question of the sample size n is a recurrent one. A sample could be said to be large when either its cost is high or when the distance of the statistic distribution (computed with n observations) to its limit distribution is small. From this point of view, Theorem 2 and Corollaries 4 and 5 provide unsatisfactory bounds since the constants in the r.h.s are, as often, missing.

The simulations below can be viewed as an attempt to answer some of these questions.

Three type of autoregressions are chosen, two of order one and one of order 4.

1. An AR1 process with a positive coefficient

$$X_{n+1} = 0.7X_n + b(e_n) + \sigma(e_n)\varepsilon_{n+1} \tag{6.24}$$

2. An AR1 process with a negative coefficient

$$X_{n+1} = -0.7X_n + b(e_n) + \sigma(e_n)\varepsilon_{n+1} \tag{6.25}$$

3. An AR4 process

$$X_{n+1} = a_1X_n + a_2X_{n-1} + a_3X_{n-2} + a_4X_{n-3} + b(e_n) + \sigma(e_n)\varepsilon_{n+1} \tag{6.26}$$

where the roots of the characteristic polynomial are ± 0.5 and 0.5 ± 0.25.

The functions b and σ are the same in all examples

$$b(e) = \sqrt{|e|}, \qquad \sigma(e) = 1 + \frac{e^2}{24}.$$

The input white noise ε_n has a standard gaussian distribution, and the exogeneous $e_n = s_n + \eta_n$ where s_n is a 6-periodic sequence with $s_0 = -1.2$, $s_1 = 3.1$, $s_2 = 1.80$, $s_3 = -2.51$,; $s_4 = -3.2$, $s_5 = -0.25$ and where the noise η_n is i.i.d. with marginal distribution uniform on $[-3, +3]$.

6.1 Three examples

For each of the three models (6.24), (6.25) and (6.26), a trajectory of size $n = 5000$ is simulated and the estimations of the parameter θ and of the functional parameter b are carried over through a number k of iterations varying from 1 to 50. Having reached the last iteration, the estimation of σ^2 is then calculated. The kernel K is the gaussian kernel and the smoothing parameters h_n and h'_n, used in the estimations of b and σ, are

$$h_n = 1.5\hat{S}_e n^{-1/2} \quad \text{and} \quad h'_n = 0.15\hat{S}_e n^{-1/3} \tag{6.27}$$

where $\hat{S}_e$ is the empirical standard deviation of the e_j's.

The results are depicted in Figures 1, 2 and 3. The upper-left graphic shows the function $b(e) = \sqrt{|e|}$, its estimate after 50 iterations together with the cloud of partial residuals used to calculate the estimate (see formula (3.5)). The upper-right graphic shows b and the evolution of its estimations $\hat{b}_n^{(k)}$ as k varies from 1 to 50. The lower-left graphic shows the evolution of the estimator of the AR parameter θ as a function of the number k of iterations, and the lower-right one presents the standard deviation $\sigma(e)$ and its final estimation.

6.1.1 Model (6.24)

Four main effects are noticeable.

- As the number of iterations increases, the estimates of b and of θ improve.
- The iterations stabilize very slowly. This is not surprising since the value of the parameter $\theta = 0.7$ is close to 1 (see Remark 1 just after Corollary 4).
- As expected, for fixed k, the convergence of $\hat{b}_n^{(k)}(e)$ is far worse in the neighbourhood of $e = 0$, discontinuity point of b'. This effect is even still visible for the estimator of $\sigma(e)$ (lower-right graphic), despite the smoothness of this function at this point.

6.1.2 Model (6.25)

Compared with the first example, there are only two differences

- The iterations stabilize quickly (4 iterations are enough), due to the fact that $\theta = -0.7$ is close to -1,
- But the obtained limit value of $\hat{\theta}_n^{(4)}$ is not very close to the true value, meaning that in this case, more observations are needed for a good estimation. However, the estimate of $\sigma(e)$ seems quite good.

6.1.3 Model (6.26)

In this example θ has 4 components. They are indicated, in the lower-left graphic, by 4 horizontal lines. The stabilisation point of the iterations is between those obtained in the two other examples (40 iterations are enough), perhaps due to the presence of the positive root 0.5. The sample size is large enough to get good estimations. It seems that the order of the autoregression, at least for moderate orders, has no significant effect on the quality of the method.

6.2 Evolution of the estimation errors as functions of the sample size

In the two last sections, we take model (6.25) and we simulate sample paths for sizes going from 200 to 10000. For each sample path, the estimations of the three parameters θ, b and σ^2 and of the distribution of the noise are computed, based on $k = 20$ iterations. Then the estimation errors are calculated. Except for the error on θ, we compute three sorts of errors, based on L_1, L_2 and L_∞ norms:

- For the functional parameters b and σ, denoting by d the length of the domain of e, we choose

$$N_1(h) = \frac{1}{\sqrt{d}} \int |h(e)|de, \qquad N_2(h) = \sqrt{\int h^2(e)de} \quad \text{and} \quad N_\infty(h) = \sqrt{d}\|h\|_\infty$$

  which satisfy $N_1 \leq N_2 \leq N_\infty$

- For the noise distribution, we compute the total variation, the Hellinger and the Kolmogorov distances.

Moreover, in order to reduce fluctuations, we simulate fifty independent trajectories for each sample size, and compute the average of the errors obtained from these trajectories.

The averaged errors are presented in Figures 4 and 5 which show, from top to bottom and left to right, the error on θ, b, σ, and on the noise distribution (three curves in each of the three last graphics, corresponding to different distances). The abscissa is the sample size n.

Figure 4 presents clearly the fact that the convergence to zero of all the errors becomes very slow when n is larger than 2000, meaning that the asymptotic speed $(\ln n/n)^{1/4}$ is reached. Errors seem to quickly decrease for small sizes.

Figure 5 is a log log set of graphics. The (nearly!) straight lines represent $c(\ln n/n)^{1/4}$ for five values of c. Except for the error on the noise distribution, which decreases faster, the theoretical bound $n^{-1/4}$ (see Corollary 4 with $\gamma = 1/2$) looks exact.

6.3 Stopping rule for the iterations

We chose to stop the backfitting iterations when the estimations are stabilized: namely, after the first k such that

$$\max\left\{\|\hat{\theta}_n^{(k)} - \hat{\theta}_n^{(k-1)}\|_2, N_1(b_n^{(k)} - \hat{b}_n^{(k-1)})\right\} \leq 10^{-3}$$

Let us denote by $k(n)$ the obtained stopping point. As pointed out in Corollary 4, $k(n)$ should be of order $\ln n$, hence hardly varying in the domain $n \leq 1000$.

For each model, and each sample size n, five independent trajectories are simulated. This is illustrated in Figure 6, for the three models (6.24), (6.25), (6.26). The sample size varies between 100 and 1000. There are three groups of 5 piecewise linear lines. Model (6.25) is represented by the lines in the lower part of the graphic. For this model, the stopping point is almost constantly equal to 7 and 8. Models (6.26) (darkest lines) and (6.24) occupy the upper part. This illustrates the asymptotic theory and the observations in Figures 1, 2 and 3.

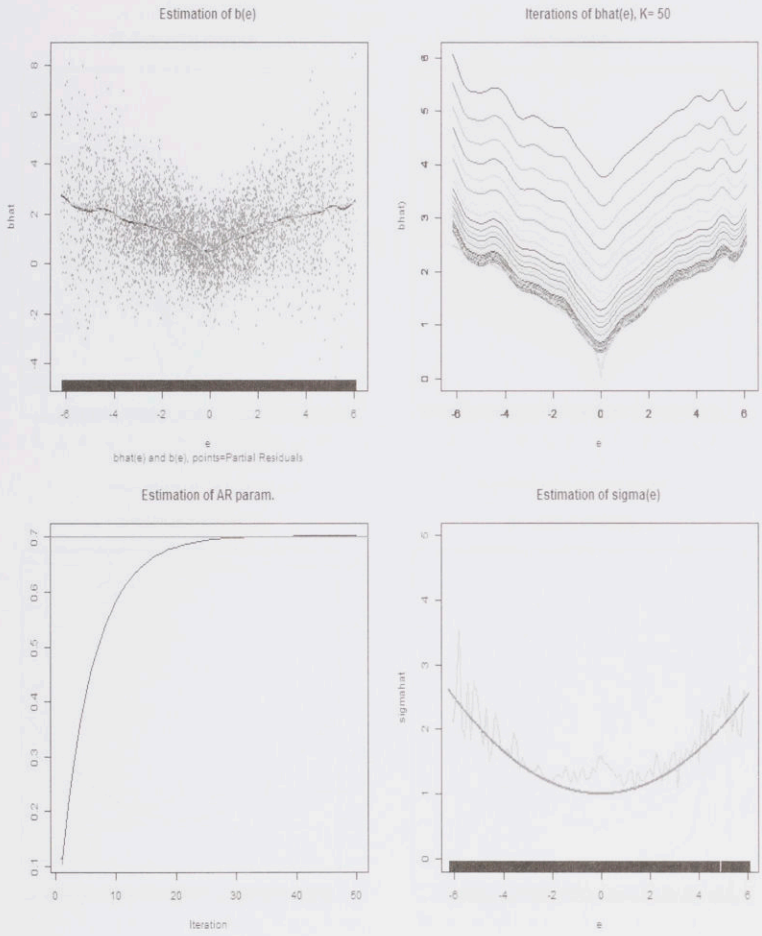

Figure 1: Estimation results for model (6.24). The trajectory length is $n = 5000$ and $k = 50$ iterations are performed. The upper-right figure shows the evolution of $\hat{b}^{(k)}(e)$ for $k = 1 : 50$. The lower-left figure presents the evolution of the estimator of θ for $k = 1 : 50$. The iterations stabilize very slowly, but the size of the sample is enough to obtain good estimation. The lower-right figure presents the estimator of $\sigma(e)$ for $k = 50$.

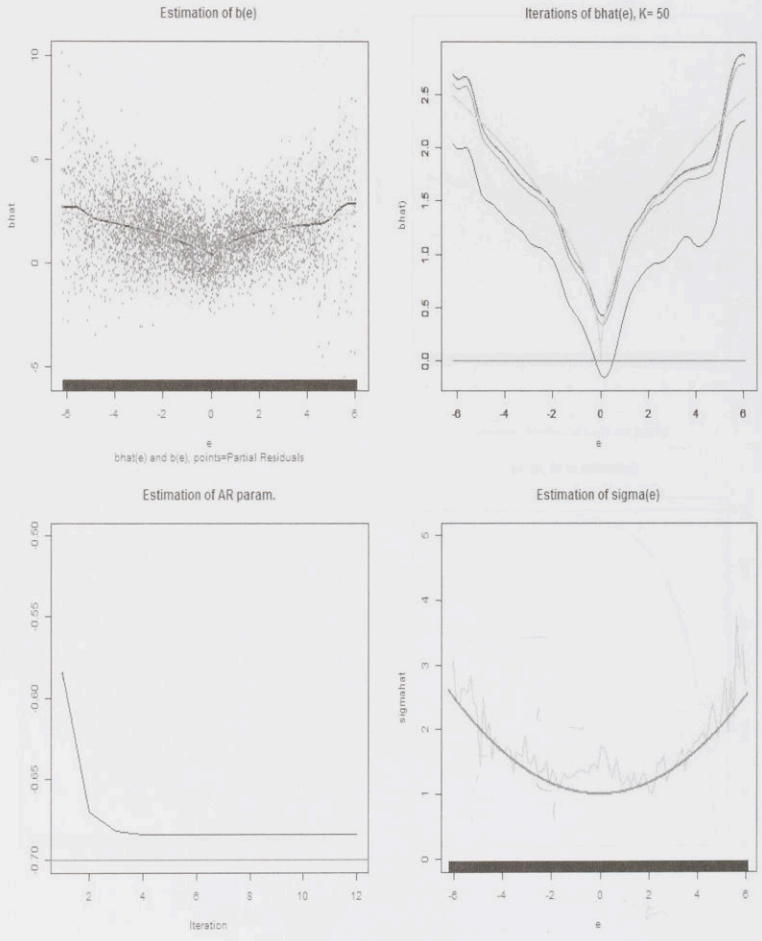

Figure 2: Estimation results for model (6.25). Few iterations are needed, but the size sample seems to be too small to obtain a good estimation of θ. Nevertheless, the estimation of the functions looks satisfactory.

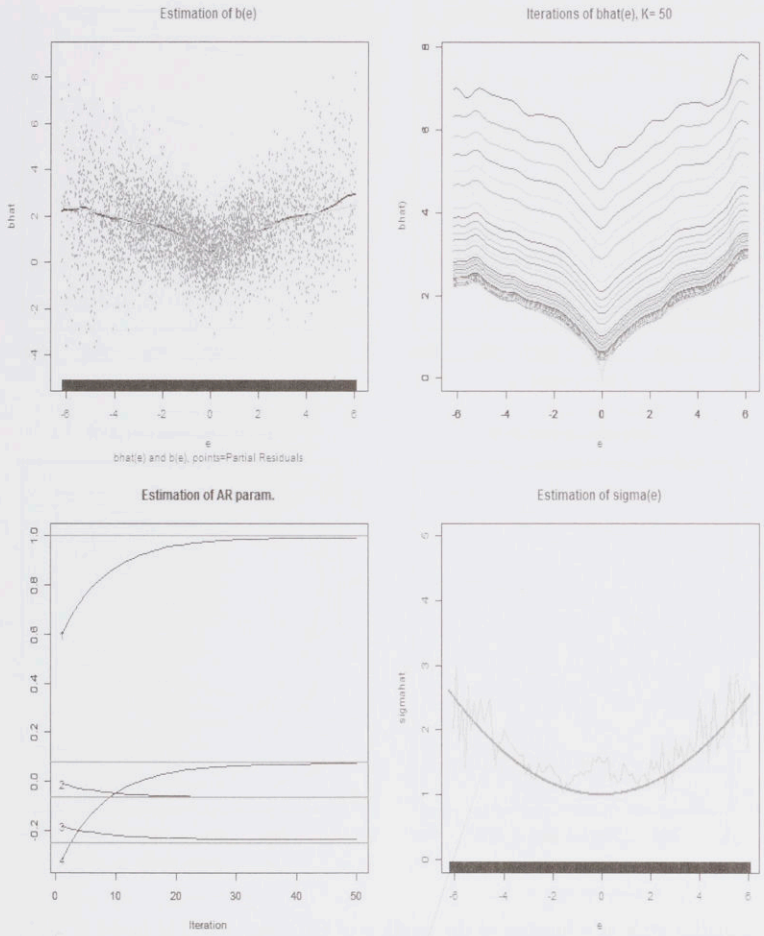

Figure 3: Estimation results for model (6.26). The coordinates of θ are the horizontal lines on the lower-left figure. The iterations converge slowly (40 iterations), and the estimations are rather good.

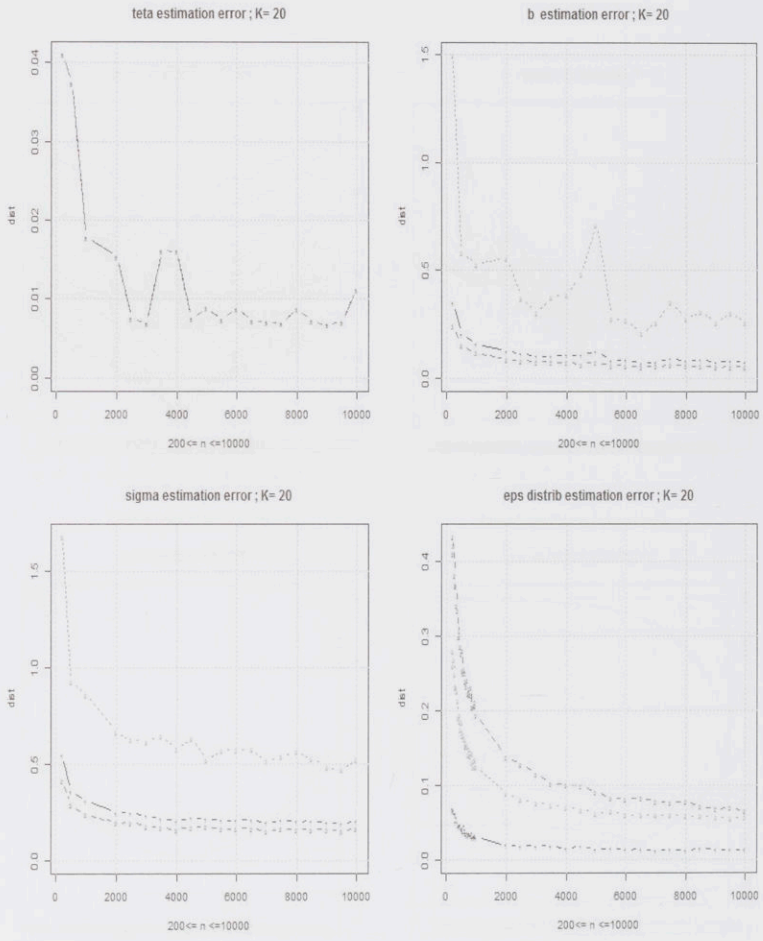

Figure 4: Estimation error as a function of the length n of the trajectory. The model is (6.25). From top to bottom and left to right: errors on the estimation of θ, b, σ, and the distance between the estimated distribution of the noise and the Gaussian distribution. In abscissa, the two first values are 200 and 500. Then, the lag remains equal to 500. In graphics 2 and 3, the positions of the three curves are conform to inequalities $N_1 \leq N_2 \leq N_\infty$. For the lower-right figure, the distances are (top to bottom) are Total-Variation, Hellinger and Kolmogorov-Smirnov distances.

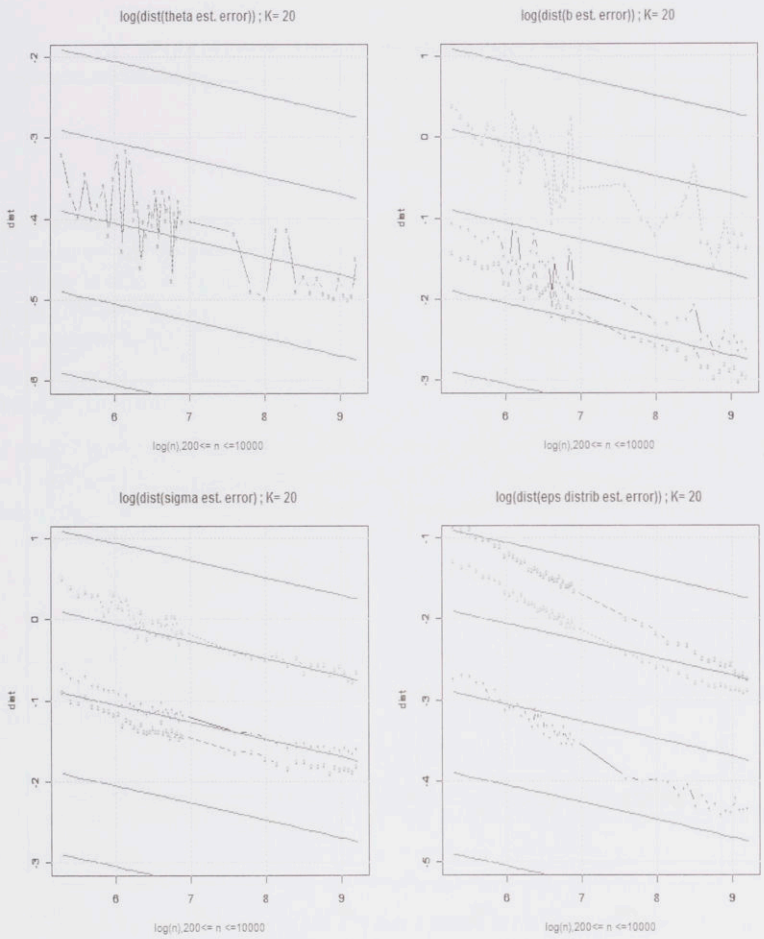

Figure 5: For model (6.25), the graphics present, in log-log coordinates, the error in estimating the parameters θ (top-left), b (top-right) and σ (bottom-left) and the distance between the estimated distribution of the noise and the Gaussian distribution (bottom-right). The straight lines (almost straight, because of the term $\ln n$) show the curves $c(\ln n/n)^{1/4}$ for several values of c.

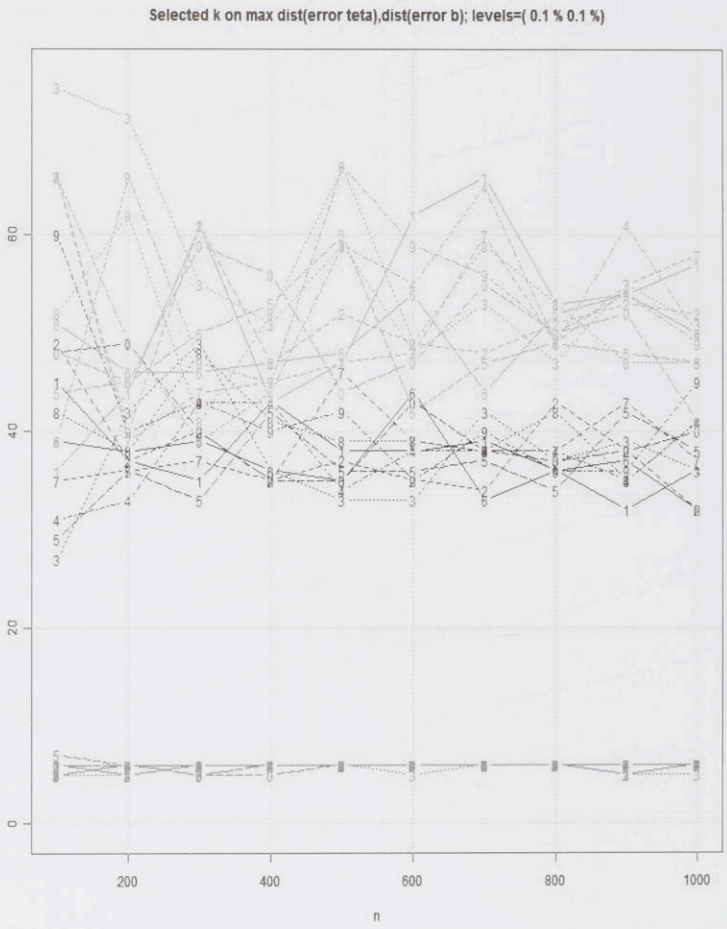

Figure 6: Value of $k(n)$ as a function of the sample size n for the three models (6.24) (the 5 upper lines), (6.25) (the 5 lowest nearly constant lines) and (6.26) (the 5 intermediate lines).

7 Appendix: Proofs

7.1 Preliminaries about the process (X_n) and its covariances

We consider the solution of (1.1) defined by the MA_∞ expansion

$$X_n = \sum_{j\geq 0} g_j b(e_{n-j}) + g_j \sigma(e_{n-j})\varepsilon_{n-j} \qquad n \in \mathbb{Z} \tag{7.28}$$

where the geometrically vanishing sequence (g_j) is defined by

$$\frac{1}{1 - a_1 z - \ldots - a_p z^p} = \sum_{j\geq 0} g_j z^j$$

Since the sequence s_n is T-periodic, and η_n is i.i.d., the T-dimensional vector sequence $Z_k = {}^t\left(X_{kT}, \ldots, X_{(k+1)T-1}\right)$ is a strictly stationary process, each coordinate being the sum of T linear scalar processes based on T independent white noises. In other words, the process (X_n) is periodically correlated (see for example [20] for a review on periodically correlated time series).

Hereafter, the stationarity of Z_k is the key for proving convergence results via ergodic theorem.

7.2 Proof of Lemma 1

The proof consists in separating the sequence (ϕ_k) into the T stationary and ergodic subsequences $\{(\phi_k^{(l)})|l = 0, \ldots, T-1\}$ defined in (4.12) and using the ergodic theorem. Details are omitted.

To check regularity of the limit M, consider the vector sequences (ψ_k) and $(\psi_k^{(l)})$ built from

$$Y_n = \sum_{j\geq 0} g_j \sigma(e_{n-j})\varepsilon_{n-j}$$

exactly as (ϕ_k) and $(\phi_k^{(l)})$ are built from (X_n). Similarly, consider the sequences (ψ'_k) and $(\psi_k'^{(l)})$ built from $Y'_n = \sum_{j\geq 0} g_j b(e_{n-j})$. Denoting by $\Gamma'^{(l)}$ the covariance matrix of $(\psi_k'^{(l)})$, and noticing that the sequences (ψ_k) and (ψ'_k) are orthogonal,

$$\Gamma^{(l)} = \Gamma'^{(l)} + \mathbb{E}\left(\psi_0^{(l)t}\psi_0^{(l)}\right)$$

Hence, if M is singular, the same holds for $\sum_{l=0}^{T-1} \mathbb{E}\left(\psi_0^{(l)t}\psi_0^{(l)}\right)$. This in turn implies that there exists $(c_1, \ldots, c_p)$ such that, for every k,

$$c_1 S_k + \cdots + c_p S_{k-p+1} =_{as} 0 \tag{7.29}$$

where $S_k = Y_k + \ldots + Y_{k-T+1}$ is the sum of the Y's over a period of the input e_n. Now, it is clear that S_k is a stationary ARMA process having the representation

$$S_k = a_1 S_{k-1} + \ldots + a_p S_{k-p+1} + \sum_{j=0}^{T-1} \sigma(e_{k-j})\varepsilon_{k-j},$$

where, from (4.11), the variance of the noise is not zero. This contradicts (7.29).

7.3 Proof of Theorem 2

Most proofs below are classical in the field of kernel functional estimation. This is why some details are omitted. The reader can refer to [3], [10] or [11] for complete developments.

7.3.1 Convergence of $R_n^{(1)}$ and $R_n^{(2)}$ defined in (4.18) and (4.19)

Lemma 6. *Under the assumptions $\mathcal{H}_{1,\dots,8}$, and if the smoothing parameter h_n is such that $h_n \sim n^{\beta_1}(\ln n)^{\beta_2}$ with some $\beta_1 < 0$ we have*

$$R_n^{(1)} = O_{a.s.}\left(\sqrt{\frac{\ln n}{nh_n}}\right) + O(h_n^{\gamma})$$

Proof. Given the convergence of Σ_n/n to a regular matrix, it is enough to prove the wanted result for

$$\frac{1}{n}\sum_{l=p+1}^{n} \phi_l \frac{\sum_{j=p+1}^{n} (b(e_j) - b(e_l) + \sigma(e_j)\varepsilon_j) K_n(e_l - e_j)}{\sum_{j=p+1}^{n} K_n(e_l - e_j)}. \tag{7.30}$$

We prove the uniform convergence

$$\sup_e \left| \frac{\sum_{j=p+1}^{n} (b(e_j) - b(e) + \sigma(e_j)\varepsilon_j) K_n(e - e_j)}{\sum_{j=p+1}^{n} K_n(e - e_j)} \right| = O_{a.s.}\left(\sqrt{\frac{\ln n}{nh_n}}\right) + O(h_n^{\gamma}). \tag{7.31}$$

The result will then follow from the fact that, thanks to the ergodic theorem applied to each subsequence $(\phi_k^{(l)})_k$, $(k = 0, \dots, T-1)$, the arithmetic mean $n^{-1}\sum_{p+1}^{n} \phi_k$ almost surely converges. □

In order to prove (7.31) we only consider

$$\frac{\sum_{j=p+1}^{n} (b(e_j) - b(e)) K_n(e - e_j)}{\sum_{j=p+1}^{n} K_n(e - e_j)} = \frac{\frac{1}{nh_n}\sum_{j=p+1}^{n} (b(e_j) - b(e)) K_n(e - e_j)}{\frac{1}{nh_n}\sum_{j=p+1}^{n} K_n(e - e_j)}. \tag{7.32}$$

The treatment of the other part in (7.30) is simpler since $\mathbb{E}(\sigma(e_j)\varepsilon_j K_n(e - e_j)) = 0$ for every j.

• Consider first the numerator of (7.32) conveniently splitted in two parts: a variance term and a bias term

$$N_1(e) = \frac{1}{nh_n}\sum_{j=p+1}^{n} (b(e_j) - b(e)) K_n(e - e_j) - \mathbb{E}\left[(b(e_j) - b(e)) K_n(e - e_j)\right]$$

and

$$N_2(e) = \frac{1}{nh_n}\sum_{j=p+1}^{n} \mathbb{E}\left[(b(e_j) - b(e)) K_n(e - e_j)\right].$$

For the so-called variance term $N_1(e)$, the basic tool is the exponential inequality

$$P\left(\left|\frac{\sum_{j=1}^{n} U_j}{n}\right| > \varepsilon\right) \leq 2e^{-\frac{n\varepsilon^2}{4\delta^2}} \quad \forall \varepsilon \in]0, 3\delta^2/d[, \tag{7.33}$$

which holds for every set $(U_1, \dots, U_n)$ of independent zero-mean variables such that $|U_j| \leq d$ and $\mathbb{E}(U_j^2) \leq \delta^2$ $(j = 1, \dots, n)$. This inequality is easily deduced from Bernstein's one as noticed in [19], page 17.

Looking at the independent sequence

$$U_j = \frac{1}{h_n}\left((b(e_j) - b(e)) K_n(e_j - e) - \mathbb{E}((b(e_j) - b(e)) K_n(e_j - e))\right),$$

firstly, since b and K are bounded, it is clear that

$$|U_j| \leq \frac{c}{h_n},$$

and secondly

$$\begin{aligned}\mathbb{E}(U_j^2) &\leq \frac{1}{h_n^2}\int (b(u)-b(e))^2 K^2\left(\frac{u-e}{h_n}\right) f(u)du \\ &= \frac{1}{h_n}\int (b(vh_n+e)-b(e))^2 K^2(v) f(vh_n+e)dv \leq \frac{c}{h_n}.\end{aligned}$$

Applying inequality (7.33) with $d=\delta^2=c/h_n$ we obtain

$$P\left(|N_1(e)|>\varepsilon\right) \leq 2e^{-\frac{nh_n\varepsilon^2}{4c}} \quad 0<\varepsilon<1$$

and then

$$P\left(|N_1(e)|>\varepsilon_0\sqrt{\frac{\ln n}{nh_n}}\right) \leq 2e^{-\frac{\varepsilon_0^2 \ln n}{4c}}.$$

A suitable choice of ε_0 yields summability of the r.h.s. and finally, by Borel Cantelli Lemma

$$N_1(e) = O_{a.s.}\left(\sqrt{\frac{\ln n}{nh_n}}\right).$$

We now turn to the bias term $N_2(e)$. From (4.10),

$$\begin{aligned}N_2(e) &= \frac{1}{h_n}\int (b(u)-b(e))K\left(\frac{u-e}{h_n}\right) f(u)du \\ &= \int (b(vh_n+e)-b(e))K(v) f(vh_n+e)dv = O(h_n^\gamma).\end{aligned}$$

We have thus proved that

$$\begin{aligned}N_1(e)+N_2(e) &= \frac{1}{nh_n}\sum_{j=p+1}^{n}\left(b(e_j)-b(e)\right)K_n(e-e_j) \\ &= O_{a.s.}\left(\sqrt{\frac{\ln n}{nh_n}}\right)+O(h_n^\gamma).\end{aligned}$$

The same rate for $\sup_e(|N_1(e)+N_2(e)|)$ is obtained by covering the domain of e by well chosen intervals and using Lipschitz property of the kernel. See [3] and [11] among others for the details.

- A similar treatment leads to

$$\sup_{e\in\mathcal{E}}\left|\frac{\sum_{j=p+1}^{n}K_n(e-e_j)}{nh_n}-\frac{\sum_{l=0}^{T-1}f(e-s_l)}{T}\right| \xrightarrow{a.s.} 0 \tag{7.34}$$

This, together with the fact that $\inf_{e\in\mathcal{E}} f(e)>0$, leads to

$$\begin{aligned}&\sup_{l}\left|\frac{\sum_{j=p+1}^{n}\left(b(e_j)-b(e_l)\right)K_n(e_l-e_j)}{\sum_{j=p+1}^{n}K_n(e_l-e_j)}\right| \\ \leq\ &\sup_{e\in\mathcal{E}}\left|\frac{\sum_{j=p+1}^{n}\left(b(e_j)-b(e)\right)K_n(e-e_j)}{\sum_{j=p+1}^{n}K_n(e-e_j)}\right| = O_{a.s.}\left(\sqrt{\frac{\ln n}{nh_n}}\right)+O(h_n^\gamma)\end{aligned}$$

and the proof of the convergence of $R_n^{(1)}$ is over.

Let us now consider the convergence of $R_n^{(2)}$:

Lemma 7. *Under the assumptions* $\mathcal{H}_{1,2,3,4}$,

$$R_n^{(2)} = o_{a.s.}\left(n^{\gamma}\right) \quad \forall \gamma > -1/2$$

Proof. The vector sequence $\phi_k \sigma(e_k)\varepsilon_k$ is a martingale difference sequence since $\mathbb{E}(\varepsilon_k) = 0$ and since $\phi_k e_k$ and ε_k are independent. Moreover,

$$\mathbb{E}(\|\phi_k \sigma(e_k)\varepsilon_k\|_2^2) = \sigma^2 \mathbb{E}\left(b(e_k)^2\right) \mathbb{E}(\|\phi_k\|_2^2).$$

where $\mathbb{E}(\|\phi_k\|_2^2)$ and $\mathbb{E}\left(b(e_k)^2\right)$ are periodic. Hence, for every $\beta > 1/2$

$$\sum \frac{\mathbb{E}(\|\phi_k \sigma(e_k)\varepsilon_k\|_2^2)}{k^{2\beta}} < \infty,$$

implying, from theorem 3.3.1 of [28],

$$n^{-\beta} \sum_{p+1}^{n} \phi_k \sigma(e_k)\varepsilon_k \xrightarrow{a.s.} 0.$$

Finally, the convergence of Σ_n/n leads to the conclusion. □

7.3.2 Convergence of the matrix coefficient A_n defined in (7.45)

We prove the convergence of A_n, the matrix coefficient of $\tilde{\theta}_n^{(k-1)}$ in (4.16).

Lemma 8. *Under the assumptions* $\mathcal{H}_{1,\ldots,8}$,
(i) As $n \to \infty$,

$$\begin{aligned} A_n &= \Sigma_n^{-1} \left(\sum_{l=p+1}^{n} \phi_l \frac{\sum_{j=p+1}^{n} {}^t\phi_j K_n(e_l - e_j)}{\sum_{j=p+1}^{n} K_n(e_l - e_j)} \right) \\ &\xrightarrow{a.s.} M^{-1} \sum_{l,j=0}^{T-1} \mu^{(l)} {}^t\mu^{(j)} \int \frac{f(u - s_j) f(u - s_l)}{\sum_{i=0}^{T-1} f(u - s_i)} du =: A \end{aligned}$$

where M *is defined in Lemma 1.*
(ii) Moreover $\|A_n - A\| = 0_{as}\left(\sqrt{\frac{\ln n}{n h_n}}\right) + O(h_n^{\gamma})$.

Proof. We consider first

$$R_n(e) := \frac{\sum_{j=p+1}^{n} {}^t\phi_j K_n(e - e_j)}{\sum_{j=p+1}^{n} K_n(e - e_j)} = \frac{\frac{\sum_{j=p+1}^{n} {}^t\phi_j K_n(e - e_j)}{n h_n}}{\frac{\sum_{j=p+1}^{n} K_n(e - e_j)}{n h_n}}.$$

The denominator has been already treated in the proof of Lemma 6 (see (7.34)), so we focus on the numerator and successively show that

$$\sup_{e \in \mathcal{E}} \left| \frac{\sum_{j=p+1}^{n} {}^t\phi_j K_n(e - e_j) - \mathbb{E}({}^t\phi_j K_n(e - e_j))}{n h_n} \right| = O_{a.s.}\left(\sqrt{\frac{\ln n}{n h_n}}\right), \tag{7.35}$$

then, with $\phi_k^{(l)}$ defined in (4.12),

$$\sup_{e \in \mathcal{E}} \left| \frac{\sum_{j=p+1}^{n} \mathbb{E}({}^t\phi_j^{(l)} K_n(e - e_j))}{n h_n} - \frac{\sum_{l=0}^{T-1} {}^t\mu^{(l)} f(e - s_l)}{T} \right| = O(h_n^{\gamma}) \tag{7.36}$$

The proof of (7.36) uses $\int K(e)de = 1$. The details are omitted. The proof of (7.35) follows the lines of the proof of (7.31), the difference coming from the fact that the $(\phi_j K_n(e_j - e))_k$ are not independent. In fact, they are weakly dependent in so far as, conditionally to the exogeneous sequence, they are mixing.

Lemma 9.

(i) For every $e \in \mathcal{E}$ and every h, the vector sequence $(\phi_j K(\frac{e_j - e}{h}))_j$ is, conditionally to the sequence $(e_j)_j =: \overline{E}$, geometrically α-mixing.

(ii) This property holds uniformly with respect to $\overline{E}$: there exists a constant C and $\alpha \in]0,1[$ such that, $\alpha^{\overline{E}}(n)$ being the conditional mixing sequence,

$$\alpha^{\overline{E}}(n) \leq C\alpha^n \qquad \forall n.$$

Proof. Consider for example the first coordinate $K(\frac{e_j - e}{h})X_{j-1}$ of the vector sequence. Conditionally to $\overline{E}$, the sequence $K(\frac{e_j - e}{h})$ is deterministic, and it is enough to consider the sequence X_j which has the same conditional mixing coefficients as $K(\frac{e_j - e}{h})X_{j-1}$. From (7.28)

$$X_n = \sum_{j \geq 0} g_j b(e_{n-j}) + g_j \sigma(e_{n-j})\varepsilon_{n-j}$$

is a linear time series based on the noise $b(e_j) + \sigma(e_j)\varepsilon_j$, where $b(e_j)$ and $\sigma(e_j)$ are deterministic trend and variance, while ε_j is i.i.d.. This white noise has a conditional density $g\left(\frac{u - b(e_j)}{\sigma(e_j)}\right)$ and a finite conditional moment of order ν. Hence,

$$\mathbb{E}\left(|X_n|^\nu \middle| \overline{E}\right) < \infty$$

Then, applying Theorem 2.1 of [5] (see also [1]) the autoregressive sequence X_n is (conditionnally) geometrically strong mixing. □

The reader is referred to [6] for definitions and properties of mixing sequences. Hereafter we need to replace inequality (7.33) by the following one, a direct consequence (see Rio's remark on page 87) of theorem 6.2 in [24]:

Lemma 10. *Let (V_j) be a strong mixing sequence of centered random variables such that for some $0 < \alpha < 1$ and $\nu > 2$,*

$$\alpha(n) \leq c\alpha^n, \ \forall n \quad \text{and} \quad M := \sup_j \mathbb{E}|V_j|^\nu < \infty.$$

Denote $s_n^2 = \sum_{1 \leq j,k \leq n} |\mathrm{Cov}(V_j, V_k)|$. For any $r > 1$ and $\lambda > 0$ and for all $a > 0$,

$$P\left(\left|\sum_{j=1}^n V_j\right| > 4\lambda\right) \leq 4\left(1 + \frac{\lambda^2}{r s_n^2}\right)^{-r/2} + C\frac{n}{r}\left(\frac{r}{\lambda}\right)^{(a+1)\nu/(a+\nu)} \tag{7.37}$$

where C is a constant depending on a and on M.

This inequality applies, conditionally to $\overline{E}$, to

$$V_j = {}^t\phi_j K_n(e - e_j) - \mathbb{E}({}^t\phi_j K_n(e - e_j)), \quad j \geq p+1.$$

For this sequence V_j, the conditional variance s_n^2 satisfies

$$s_n^2 = O(nh_n) \tag{7.38}$$

where the O is uniform with respect to $\overline{E}$. Indeed,

$$\begin{cases} \mathrm{Var}^{\overline{E}}(V_j) \leq c_1 h_n \\ |\mathrm{Cov}^{\overline{E}}(V_j, V_l)| \leq c_2 h_n^2 & \text{if} \quad |j-l| \leq \delta_n \\ |\mathrm{Cov}^{\overline{E}}(V_j, V_l)| \leq C\alpha^{\frac{2\nu}{2\nu-1}|j-l|} = C\alpha_1^{|j-l|} & \text{if} \quad |j-l| > \delta_n \end{cases}$$

For the last bound, the reader can refer to [6]. The two first ones are directly obtained. The functions b and σ being bounded, the constant term in each r.h.s. above does not depend on $\overline{E}$. Taking then $\delta_n = 1/(h_n \ln n)$ easily leads to (7.38).

Now, (7.37) leads to

$$P^{\overline{E}}\left(\left|\sum_{j=p+1}^{n} {}^t\phi_j K_n(e-e_j) - \mathbb{E}({}^t\phi_j K_n(e-e_j))\right| > 4\lambda\right) \leq 4\left(1+\frac{c\lambda^2}{rnh_n}\right)^{-r/2} + C_1\frac{n}{r}\left(\frac{r}{\lambda}\right)^{(a+1)\nu/(a+\nu)},$$

and then, if $\ln n = o(r_n)$

$$\begin{aligned} & P^{\overline{E}}\left(\left|\frac{\sum_{j=p+1}^{n} {}^t\phi_j K_n(e-e_j) - \mathbb{E}({}^t\phi_j K_n(e-e_j))}{nh_n}\right| > \lambda_0\sqrt{\frac{\ln n}{nh_n}}\right) \\ \leq\ & 4\left(1+\frac{c\lambda_0^2 \ln n}{16 r_n}\right)^{-r_n/2} + C_1\frac{n}{r}\left(\frac{r}{\lambda_0\sqrt{nh_n \ln n}}\right)^{(a+1)\nu/(a+\nu)} \\ \leq\ & 4e^{-\frac{c\lambda_0^2 \ln n}{32}} + C_1\frac{n}{r}\left(\frac{r}{\lambda_0\sqrt{nh_n \ln n}}\right)^{(a+1)\nu/(a+\nu)}. \end{aligned}$$

Now, if $h_n \sim n^{\beta_1}\ln n^{\beta_2}$ with $\beta_1 > -1$, $r_n = (\ln n)^\beta$ we get, for n large enough,

$$\begin{aligned} & P^{\overline{E}}\left(\left|\frac{\sum_{j=p+1}^{n} {}^t\phi_j K_n(e-e_j) - \mathbb{E}({}^t\phi_j K_n(e-e_j))}{nh_n}\right| > \lambda_0\sqrt{\frac{\ln n}{nh_n}}\right) \\ \leq\ & 4n^{-c\lambda_0^2} + C_2 n^{1-(1+\beta_1)\frac{(a+1)\nu}{2(a+\nu)}}(\ln n)^{\beta_3} \qquad (7.39) \\ & \qquad (7.40) \end{aligned}$$

for some β_3. Now, the constant C_2 in (7.39) does not depend on $\overline{E}$, implying that

$$\begin{aligned} & P\left(\left|\frac{\sum_{j=p+1}^{n} {}^t\phi_j K_n(e-e_j) - \mathbb{E}({}^t\phi_j K_n(e-e_j))}{nh_n}\right| > \lambda_0\sqrt{\frac{\ln n}{nh_n}}\right) \\ \leq\ & 4n^{-c\lambda_0^2} + C_2 n^{1-(1+\beta_1)\frac{(a+1)\nu}{2(a+\nu)}}(\ln n)^{\beta_3}. \end{aligned}$$

Select $\lambda_0^2 > 1/c$. Then, thanks to the inequality $\nu(1+\beta_1)$ we obtain, for a large enough,

$$\sum 4n^{-c\lambda_0^2} + C_2 n^{1-(1+\beta_1)\frac{(a+1)\nu}{2(a+\nu)}}(\ln n)^{\beta_3} < \infty$$

So we have proved that, for fixed e,

$$\frac{\sum_{j=p+1}^{n} {}^t\phi_j K_n(e-e_j) - \mathbb{E}({}^t\phi_j K_n(e-e_j))}{nh_n} = O_{a.s.}\left(\sqrt{\frac{\ln n}{nh_n}}\right).$$

The same speed is obtained for the sup-norm.
From (7.35), (7.36) and (7.34) it follows that, with

$$\tilde{R}(e) := \frac{\sum_{j=0}^{T-1} {}^t\mu_j f(e-s_j)}{\sum_{j=0}^{T-1} f(e-s_j)}$$

$$\sup_e \left| R_n(e) - \tilde{R}(e) \right| = O_{a.s.}\left(\sqrt{\frac{\ln n}{n h_n}} \right) + O_{a.s.}(h_n^\gamma) \tag{7.41}$$

implying in turn

$$\begin{aligned} A_n &= n\Sigma_n^{-1} \frac{1}{n} \sum_{l=p+1}^{n} \phi_l R_n(e_l) \\ &= n\Sigma_n^{-1} \frac{1}{n} \sum_{l=p+1}^{n} \phi_l (R_n(e_l) - \tilde{R}(e_l)) + n\Sigma_n^{-1} \frac{1}{n} \sum_{l=p+1}^{n} \phi_l \tilde{R}(e_l) \\ &= O_{a.s.}\left(\sqrt{\frac{\ln n}{n h_n}} \right) + O_{a.s.}(h_n^\gamma) + n\Sigma_n^{-1} \frac{1}{n} \sum_{l=p+1}^{n} \phi_l \tilde{R}(e_l). \end{aligned} \tag{7.42}$$

In (7.42), the last sum is splitted in T sums

$$\frac{1}{n} \sum_{l=p+1}^{n} \phi_l \tilde{R}(e_l) = \sum_{l=0}^{T-1} \frac{1}{n} \sum_{kT+l \le n}^{n} \phi_k^{(l)} \tilde{R}(s_l + \eta_{kT+l}),$$

which almost surely converges to

$$\begin{aligned} \frac{1}{T} \sum_{l=0}^{T-1} \mathbb{E}\left(\phi_0^{(l)} \right) E(\tilde{R}(s_l + \eta_0)) &= \frac{1}{T} \sum_{l=0}^{T-1} \mu^{(l)} E(\tilde{R}(s_l + \eta_0)) \\ &= \frac{1}{T} \sum_{l,j=0}^{T-1} \mu^{(l)t} \mu^{(j)} \int \frac{f(v - s_j) f(v - s_l)}{\sum_{i=0}^{T-1} f(v - s_i)} dv \end{aligned}$$

Moreover, this convergence rate, being the rate in the law of large numbers for i.i.d sequences, is faster than the first two terms in (7.42). This, together with (7.42) and the almost sure convergence of $n\Sigma_n^{-1}$, leads to the desired result. Lemma 8 is proved.

□

Lemma 8, together with Lemma 11 below, shows that the passage (4.16) from step $k-1$ to step k is a fixed point iteration, at least for n large enough.

Lemma 11. *There exists $k_0 \ge 1$ such that*

$$\sup_v \frac{\|A^{k_0} v\|_2}{\|v\|_2} < 1. \tag{7.43}$$

Moreover, $k_0 = 1$ when $p = 1$.

Proof. For the sake of simplicity, we take $T = 2$. The general case only brings more complicated formulas. Denoting

$$S = \Gamma^{(1)} + \Gamma^{(2)},$$

where the $\Gamma^{(j)}$s are defined in Lemma 1, and

$$\alpha_{jl} = \int \frac{f(v - s_j) f(v - s_l)}{\sum_{i=0}^{1} f(v - s_i)} dv,$$

$$A = \left[S + \mu_0^t \mu_0 + \mu_1^t \mu_1 \right]^{-1} \left(\alpha_{00} \mu_0^t \mu_0 + \alpha_{11} \mu_1^t \mu_1 + \alpha_{01} (\mu_0^t \mu_1 + \mu_1^t \mu_0) \right). \tag{7.44}$$

We then apply a popular matrix inversion formula:

$$\begin{aligned}\left[S+\mu_0^t\mu_0+\mu_1^t\mu_1\right]^{-1}\mu_1 &= \frac{[S+\mu_0^t\mu_0]^{-1}\mu_1}{1+^t\mu_1[S+\mu_0^t\mu_0]^{-1}\mu_1}=\frac{S_1^{-1}\mu_1}{1+^t\mu_1M_1^{-1}\mu_1}\\ \left[S+\mu_0^t\mu_0+\mu_1^t\mu_1\right]^{-1}\mu_0 &= \frac{[S+\mu_1^t\mu_1]^{-1}\mu_0}{1+^t\mu_0[S+\mu_1^t\mu_0]^{-1}\mu_1}=\frac{S_0^{-1}\mu_0}{1+^t\mu_0S_0^{-1}\mu_0}\end{aligned}$$

where

$$S_1=S+\mu_0^t\mu_0 \quad \text{et} \quad S_0=S+\mu_1^t\mu_1.$$

This leads to

$$A=\frac{S_0^{-1}\mu_0}{1+^t\mu_0S_0^{-1}\mu_0}\left(\alpha_{00}{}^t\mu_0+\alpha_{01}{}^t\mu_1\right)+\frac{S_1^{-1}\mu_1}{1+^t\mu_1S_1^{-1}\mu_1}\left(\alpha_{11}{}^t\mu_1+\alpha_{01}{}^t\mu_0\right)$$

and finally to

$$\begin{aligned}A &= \alpha_{00}\frac{S_0^{-1}\mu_0^t\mu_0}{1+^t\mu_0S_0^{-1}\mu_0}+\alpha_{11}\frac{S_1^{-1}\mu_1^t\mu_1}{1+^t\mu_1S_1^{-1}\mu_1}\\ &+ \alpha_{01}\left(\frac{S_0^{-1}\mu_0^t\mu_1}{1+^t\mu_0S_0^{-1}\mu_0}+\frac{S_1^{-1}\mu_1^t\mu_0}{1+^t\mu_1S_1^{-1}\mu_1}\right)\\ &= \alpha_{00}S_{00}+\alpha_{11}S_{11}+\alpha_{01}(S_{01}+S_{10})\end{aligned} \tag{7.45}$$

where the last line defines the S_{jl}'s.

It is easily checked that

$$\begin{aligned}S_{jj}^2 &= \frac{^t\mu_jS_j^{-1}\mu_j}{1+^t\mu_jS_j^{-1}\mu_j}S_{jj}=\beta_{jj}S_{jj}, \quad j=1,2\\ S_{jk}^2 &= \frac{^t\mu_kS_j^{-1}\mu_j}{1+^t\mu_jS_j^{-1}\mu_j}S_{jk}=\beta_{jk}S_{jk} \quad j\neq k\end{aligned}$$

Clearly, $0\leq\beta_{jj}<1$. Moreover,

$$|\beta_{jk}|\leq\frac{\sqrt{^t\mu_jS_j^{-1}\mu_j}}{1+^t\mu_jS_j^{-1}\mu_j}\sqrt{^t\mu_kS_j^{-1}\mu_k}<1$$

because the first factor is less than $1/2$, and

$$\begin{aligned}{}^t\mu_kS_j^{-1}\mu_k &= {}^t\mu_k[S+\mu_k^t\mu_k]^{-1}\mu_k={}^t\mu_k\left(S^{-1}-\frac{S^{-1}\mu_k^t\mu_kS^{-1}}{1+^t\mu_kS^{-1}\mu_k}\right)\mu_k\\ &= \frac{^t\mu_kM^{-1}\mu_k}{1+^t\mu_kM^{-1}\mu_k}<1.\end{aligned}$$

As $\alpha_{jl}\in[0,1]$ for every j,l, it results that

$$A^2=\alpha_{00}^{(2)}S_{00}+\alpha_{11}^{(2)}S_{11}+\alpha_{01}^{(2)}(S_{01}+S_{10})$$

where for every j,l, $|\alpha_{j,l}^{(2)}|\leq\beta_{jl}\alpha_{j,l}$, whence

$$A^k=\alpha_{00}^{(k)}M_{00}+\alpha_{11}^{(k)}M_{11}+\alpha_{01}^{(k)}(M_{01}+M_{10})$$

where for every j,l, $|\alpha_{j,l}^{(k)}|\leq(\beta_{jl})^{k-1}\alpha_{j,l}$. Lemma 11 is proved. □

It remains to prove (4.21), the rate of convergence of the error on the standard deviation. The estimation error $\tilde{\sigma}_{n,k}(e)$ is

$$\begin{aligned}
\tilde{\sigma}_{n,k}(e) &= \hat{\sigma}_{n,k}(e) - \sigma^2(e) \\
&= \frac{\sum_{l=p+1}^{n-1} \left(\left(X_l - {}^t \phi_l \hat{\theta}_n^{(k)} - \hat{b}_n^{(k-1)}(e_l) \right)^2 - \sigma^2(e) \right) K_n(e - e_l)}{\sum_{l=p+1}^{n-1} K_n(e - e_l)} \\
&= \frac{\sum_{l=p+1}^{n-1} \left(\sigma^2(e_l)\varepsilon_l^2 - \sigma^2(e) \right) K_n(e - e_l)}{\sum_{l=p+1}^{n-1} K_n(e - e_l)} + R_{n,k}(e) \qquad (7.46)
\end{aligned}$$

where, from the first part of the theorem,

$$R_{n,k}(e) = O_{a.s.}\left(\sqrt{\frac{\ln n}{n h_n}} \right) + O_{a.s.}(h_n^{\gamma}) + O_{a.s.}(\beta^k).$$

Now, since the variables $\sigma^2(e_l)\varepsilon_l^2 - \sigma^2(e)$ are independent and centered, the first term in (7.46) can be treated exactly as was (7.32), leading to

$$\frac{\sum_{l=p+1}^{n-1} \left(\sigma^2(e_l)\varepsilon_l^2 - \sigma^2(e) \right) K_n(e - e_l)}{\sum_{l=p+1}^{n-1} K_n(e - e_l)} = O_{a.s.}\left(\sqrt{\frac{\ln n}{n h_n}} \right) + O_{a.s.}(h_n^{\gamma'})$$

and the proof of (4.21) is completed.

References

[1] Athreya B., Pantula G. (1986). *A note on strong mixing of ARMA processes.* Statist. Probab. Lett. 4. 187–190.

[2] Bhattacharya P. K., Zhao P-L. (1997). Semiparametric inference in a partial linear model. *Ann. Statist.* 25-1. 244–262.

[3] Bosq D. (1998). *Nonparametric statistics for stochastic processes.* LNS. Springer.

[4] Buja A, Hastie T.J., Tibshirani R.J. (1989). Linear smoothers and additive models. *Ann. Statist.* 17. 453–555.

[5] Chanda K.C. (1974). *Strong mixing properties of linear stochastic processes.* J. Appl. Probab. 11. 401–408.

[6] Doukhan P. (1994). *Mixing: properties and examples.* LNS. Springer.

[7] Engel R.F., Granger C.W.J., Rice J., Weiss A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* 81. 310–320.

[8] Fadili J. M. (2005). Penalized partially linear models using sparse representation with an application to fMRI time series. *IEEE Trans. Sign. Proc.* 53-9 3436–3448.

[9] Fan J., Jiang J. (2005). Non parametric inference for additive models. *J. Amer. Statist. Assoc.* 100-471. 890–907.

[10] Fan J., Yao Q. (2003). *Non linear time series. Non parametric and parametric methods.* Springer.

[11] Ferraty F., Vieu Ph. (2001). *Statistique fonctionnelle: modèles de régression pour variables aléatoires uni, multi et infiniment dimensionnées.* Pub. n° LSP-2001-03.

[12] Gannaz I. (2007). *Estimation par ondelettes dans les modèles partiellement linéaires.* Thèse de l'Université Joseph Fourier.

[13] Gao J.(1998). Semiparametric regression smoothing of non-linear time-series. *Scand. J. Statist.* 25. 521–539.

[14] Gao J., Yee T. (2000). Adaptive estimation in partially linear models. *Canadian J. Statist.* 28-3. 571–588.

[15] Gao J., Tong H., Wolff R. (2002). Adaptive orthogonal series esttimation in additive stochastic regression models. *Statistica Sinica* 12-(2). 409–428.

[16] Härdle W. (1990). *Applied non-parametric regression.* Econom. Soc. Monographs.

[17] Härdle W., Liang H., Gao J. (2000). *Partially linear models*; Physica-Verlag. Heidleberg.

[18] Hastie T, Tibshirani R. (1991). *Generalized additive models.* Chapman and Hall. London.

[19] Hoeffding W. (1966). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58-301. 13–30.

[20] Lund, R. B., Basawa I. V. (1999). *Modeling and inference for periodically correlated time series* In Asymptotics, nonparametrics, and time series, vol. 158 of Statist. Textbooks Monogr. 37–62. Dekker, New York.

[21] Mammen E, Linton O, and Nielsen J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* 27. 1443–1490.

[22] Opsomer J. D. (2000). Asymptotic properties of backfitting estimators. *J. Mult. Anal.* 73-2. 166–179.

[23] Opsomer J. D., Ruppert D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* 25. 186-211.

[24] Rio E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants.* Maths et Applications SMAI. Springer.

[25] Robinson, P.M. (1988). Root-N-consistent semiparametric regression. *Econometrica.* 56, 931-954.

[26] Severini T. A., Staniswalis J. G. (1994). Quasi-likelihood Estimation in Semiparametric models. *J. Amer. Statist. Assoc.* 89-426. 501–511.

[27] Speckman P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* 50. 413–436.

[28] Stout W. F. (1974). *Almost sure convergence* Academic Press.

[29] Wang L., Yang L. (2007). Spline-backfitted kernel smoothing of non-linear autoregression model. *Ann. Statist.* 35. 2474–2503.