



# Adaptive estimation of mean and volatility functions in (auto)-regressive models

Fabienne Comte, Yves Rozenholc

## ► To cite this version:

Fabienne Comte, Yves Rozenholc. Adaptive estimation of mean and volatility functions in (auto)-regressive models. *Stochastic Processes and their Applications*, 2002, 97 (1), pp.111-145. 10.1016/S0304-4149(01)00128-4 . hal-00748963

**HAL Id: hal-00748963**

**<https://hal.science/hal-00748963>**

Submitted on 6 Nov 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Adaptive estimation of mean and volatility functions in (auto-)regressive models

F. Comte\*, Y. Rozenholc†

First version July 2000, revised May 2001.

## Abstract

In this paper, we study the problem of non parametric estimation of the mean and variance functions  $b$  and  $\sigma^2$  in a model:  $X_{i+1} = b(X_i) + \sigma(X_i)\varepsilon_{i+1}$ . For this purpose, we consider a collection of finite dimensional linear spaces. We estimate  $b$  using a mean squares estimator built on a data driven selected linear space among the collection. Then an analogous procedure estimates  $\sigma^2$ , using a possibly different collection of models. Both data driven choices are performed via the minimization of penalized mean squares contrasts. The penalty functions are random in order not to depend on unknown variance-type quantities. In all cases, we state non asymptotic risk bounds in  $\mathbb{L}_2$  empirical norm for our estimators and we show that they are both adaptive in the minimax sense over a large class of Besov balls. Lastly, we give the results of intensive simulation experiments which show the good performances of our estimator.

**Keywords.** Nonparametric regression, Least-squares estimator, Adaptive estimation, Autoregression, Variance estimation, Mixing processes.

*1991 Mathematical Subject Classification.* Primary 62G07 Secondary 62J02.

## 1 Introduction

### 1.1 Presentation of the problem

In this paper, we study the following model:

$$X_{i+1} = b(X_i) + \sigma(X_i)\varepsilon_{i+1}, \quad (1.1)$$

with  $\varepsilon_i$  i.i.d. centered random variables with unit variance. It can be considered as a particular case of the standard regression model:

$$Y_i = b(X_i) + \sigma(X_i)u_i, \quad (1.2)$$

with i.i.d. centered  $u_i$ 's,  $\text{Var}(u_1) = 1$ , where the  $(X_i, Y_i)$ 's are not assumed to be independent but can be  $\beta$ -mixing. Our results hold for this model.

---

\*Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 6, comte@ccr.jussieu.fr

†Université du Maine and Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 7, rozen@math.jussieu.fr.

If  $\hat{f}$  is an estimator of  $f$ , where  $f$  is  $b$  or  $\sigma^2$ , then we measure the risk of  $\hat{f}$  via the  $\mathbb{L}_2$ -empirical norm:

$$\mathbb{E}[\|\hat{f} - f\|_n^2] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n (\hat{f}(X_i) - f(X_i))^2 \right].$$

For a discussion about the choice of this measure of risk, see Baraud et al. (1999). Roughly speaking, the reason for this choice is as follows: let  $\hat{f}_S$  be a minimizer of

$$\gamma_n(t) = \frac{1}{n} \sum_{i=1}^n [X_{i+1} - t(X_i)]^2,$$

for  $t$  in a linear space  $S \subset \mathbb{L}_2(\mathbb{R}, dx)$ ; then the vector  $(\hat{f}_S(X_1), \dots, \hat{f}_S(X_n))$  is uniquely defined, but of course not the global function  $\hat{f}_S$  at any point. We shall nevertheless talk about “the” mean squares estimator since only the associated vector  $(\hat{f}_S(X_1), \dots, \hat{f}_S(X_n))$  of  $\mathbb{R}^n$  is involved in the computations.

In addition, under suitable assumptions, this risk can be decomposed into bias + variance terms via:

$$\mathbb{E}[\|f - \hat{f}_S\|_n^2] \leq \kappa \left( \|f - f_S\|_\mu^2 + \frac{\dim(S)}{n} \right), \quad (1.3)$$

where  $f_S$  is the  $\mathbb{L}_2(dx)$ -orthogonal projection of  $f$  on  $S$ ,  $\|t\|_\mu^2 = \mathbb{E}(t^2(X_1))$  and  $\kappa$  depends on constants of the problem.

To see how (1.3) is obtained, consider a strictly stationary sequence  $(X_i)$  drawn from model (1.1) with  $\sigma \equiv 1$  and stationary  $[0, 1]$ -supported density, and let  $S$  be generated by  $\varphi_1, \dots, \varphi_D$ , the histogram orthonormal basis of  $\mathbb{L}_2([0, 1])$ :  $\varphi_j(x) = \sqrt{D} \mathbf{I}_{[(j-1)/D, j/D]}(x)$ . Simple algebra leads to

$$\gamma_n(t) - \gamma_n(s) = \|b - t\|_n^2 - \|b - s\|_n^2 + 2\langle s - t, \varepsilon \rangle_n \quad (1.4)$$

where  $\langle t, \varepsilon \rangle_n = (1/n) \sum_{i=1}^n t(X_i) \varepsilon_{i+1}$ . Then we find from (1.4)

$$\begin{aligned} \|\hat{b}_S - b\|_n^2 &\leq \|b_S - b\|_n^2 + 2\langle b_S - \hat{b}_S, \varepsilon \rangle_n \\ &\leq \|b_S - b\|_n^2 + 2\|b_S - \hat{b}_S\| \sup_{t \in S, \|t\|=1} |\langle t, \varepsilon \rangle_n| \\ &\leq \|b_S - b\|_n^2 + \frac{1}{4a} \|b_S - \hat{b}_S\|^2 + 4a \sup_{t \in S, \|t\|=1} \langle t, \varepsilon \rangle_n^2. \end{aligned}$$

Assume that, for some  $a > 1$ ,

$$\forall t \in S, \quad \|t\|^2 \leq a \|t\|_n^2, \quad (1.5)$$

then

$$\mathbb{E}(\|\hat{b}_S - b\|_n^2) \leq 3\|b_S - b\|_\mu^2 + 8a \mathbb{E} \left( \sup_{t \in S, \|t\|=1} \langle t, \varepsilon \rangle_n^2 \right). \quad (1.6)$$

Besides, using Cauchy Schwarz inequality yields

$$\mathbb{E} \left( \sup_{t \in S, \|t\|=1} \langle t, \varepsilon \rangle_n^2 \right) = \mathbb{E} \left[ \sup_{\sum_j a_j^2 \leq 1} \left( \sum_{j=1}^D a_j \langle \varphi_j, \varepsilon \rangle_n \right)^2 \right] \leq \sum_{j=1}^D \mathbb{E} \langle \varphi_j, \varepsilon \rangle_n^2$$

$$\begin{aligned}
&= \frac{1}{n^2} \sum_{j=1}^D \mathbb{E} \left( \sum_{i=1}^n \varphi_j(X_i) \varepsilon_{i+1} \right)^2 \\
&= \frac{1}{n^2} \sum_{j=1}^D \sum_{i=1}^n \mathbb{E}(\varphi_j^2(X_i)) \mathbb{E}(\varepsilon_{i+1}^2) \\
&= \frac{1}{n} \mathbb{E} \left( \sum_{j=1}^D \varphi_j^2(X_1) \right) = \frac{D}{n}.
\end{aligned} \tag{1.7}$$

Therefore (1.6) and (1.7) lead to (1.3), provided that (1.5) holds, which is generally true with large probability.

In view of these considerations, here is now our estimation procedure. We start with two finite collections of models denoted by  $\{S_m^{(i)}, m \in \mathcal{M}_n^{(i)}\}$  for  $b$  if  $i = 1$  and  $\sigma^2$  if  $i = 2$ ; each  $S_m^{(i)}$  is a finite dimensional subspace of  $\mathbb{L}_2(\mathbb{R}, dx)$ . The functions  $b$  and  $\sigma^2$  are not required to belong to any of the models. Let  $\hat{b}_m$  denote the least squares estimator of  $b$  on  $S_m^{(1)}$  associated to

$$\gamma_n^{(1)}(t) = \frac{1}{n} \sum_{i=1}^n (X_{i+1} - t(X_i))^2$$

based on the observations  $X_1, \dots, X_{n+1}$  arising from model (1.1). We use a procedure that chooses  $\hat{m}_1$  in  $\mathcal{M}_n^{(1)}$  as the minimizer  $\gamma_n^{(1)}(\hat{b}_m) + \text{pen}^{(1)}(m)$  among all  $m$  in  $\mathcal{M}_n^{(1)}$ , where  $\text{pen}^{(1)}$  is a known penalty function specified later. The key point is that this procedure is entirely based on the data and not on any prior information on  $b$ , and that it realizes a good trade-off between the bias and variance terms, namely:

$$\mathbb{E} \left[ \|b - \hat{b}_{\hat{m}}\|_n^2 \right] \leq C \min_{m \in \mathcal{M}_n} \left\{ \|b - b_m\|_\mu^2 + \frac{\dim(S_m^{(1)})}{n} \right\}, \tag{1.8}$$

where  $b_m$  is the  $\mathbb{L}^2(dx)$ -orthogonal projection of  $b$  on  $S_m^{(1)}$  and  $C$  is a multiplicative constant depending on some quantities of the problem. This means that, up to the constant  $C$ , the estimator chooses an optimal model among the collection.

In the second step,  $\hat{\sigma}^2$  is based in an analogous way on the contrast:

$$\gamma_n^{(2)}(t) = \frac{1}{n} \sum_{i=1}^n \left[ X_{i+1}^2 - \hat{b}_{\hat{m}_1}^2(X_i) - t(X_i) \right]^2,$$

for  $t \in S_m^{(2)}$ , with an aim similar to (1.8) and  $b$  replaced by  $\sigma^2$ , using a penalty function  $\text{pen}^{(2)}(m)$ .

Both penalty functions  $\text{pen}^{(i)}(m)$ ,  $i = 1, 2$  are found of order  $\dim(S_m^{(i)})/n$ . This model selection criterion is closely related to the classical  $C_p$  criterion of Mallows (1973).

It is important to notice that estimators satisfying inequalities as (1.8) have interesting properties on the collections of models that we have in mind (piecewise polynomials, wavelets, trigonometric polynomials). In particular, such estimators are adaptive in the minimax sense with respect to many well known classes of smoothness (see Barron et al. (1999) and Birgé and Massart (1997)).

## 1.2 Some bibliographic remarks

The autoregressive model has been extensively studied in the literature in view of applications to Finance and Econometrics in particular. People first modeled the conditional mean of the variable of interest  $X_t$  given its past as a linear function of past  $X_t$ 's, the conditional variance being constant, see Lütkepohl (1992) and the ARMA (Autoregressive Moving Average) models of the time series literature. Then many financial variables were experimented to have non constant conditional variance, and specifications of it as a linear function of the squared values of the past innovations were developed with ARCH (Autoregressive Conditionally Heteroskedastic) models introduced by Engle (1982) and generalized by Bollerslev (1986). Lastly, nonlinear extensions of both types of functions (conditional mean and conditional variance) were studied: step functions in Gouriéroux and Monfort (1992), general non linear functions in Mc Keague and Zhang (1994) or Härdle and Tsybakov (1997). This is the reason why statistical methods for nonparametric estimation of variance functions were recently developed.

On the other hand, adaptive estimation methods have been studied in some frameworks that can be related to the present one. In particular, several studies related to penalization criteria as Akaike's or BIC criterion for regressive models, by Akaike (1973), Shibata (1976), Li (1987), Polyak and Tsybakov (1992), have lead to asymptotic results. More recently, a general approach to model selection has been developed by Birgé and Massart (1997) and Barron et al. (1999) with many applications to adaptive estimation. Their viewpoint is non asymptotic, and so is ours. The procedure we use has been studied for fixed design regressive models by Baraud (2000) and for  $\beta$ -mixing random design and autoregressive models by Baraud et al. (1999); the variance function is constant in all of these works and thus only the mean function is estimated. Our results here are an extension of the latter to the estimation of the mean when the variance function is not constant, and to the estimation of the variance function as well.

Variance estimation has been first studied in fixed design regression models, see for instance Müller and Stadtmüller (1987) who apply to this problem a difference-based estimator. Hall and Carroll (1989) build a residual-based estimator and show that they pointwise reach the optimal rate of convergence even with an unknown mean function  $b$ , provided that  $b$  has a smoothness order larger than  $1/2$ . Dependent models (autoregressive models or regressive models with mixing random design) have been handled by Härdle and Tsybakov (1997), Härdle et al. (1998) and Fan and Yao (1998). Härdle and Tsybakov (1997) study the estimation of  $b$  and  $\sigma^2$  using local polynomial estimators; they prove pointwise asymptotic normality with standard rates but their procedure is not adaptive. Fan and Yao (1998) describe a data driven procedure with automatic bandwidth selection but their theoretical results provide only a pointwise Central Limit Theorem for a non adaptive estimator.

Lastly, adaptive procedures for variance estimation have been studied by Neumann (1994) and Hoffmann (1999). Neumann (1994) builds an adaptive kernel (with random bandwidth) residual-based estimator, but in a fixed design model with a noise admitting moments of any order. He proves optimal rates for the mean integrated squared error of his estimator, provided that the mean function has a smoothness order  $\alpha_1$ . The framework the most related to the present work is Hoffmann (1999)'s who proposes an adaptive wavelet thresholding procedure in an autoregressive framework. He requires that the noise admits moments of any order and obtains for the general  $\mathbb{L}^p$ -integrated risk the optimal rates up to some logarithmic factors. The rates for  $b$  and  $\sigma^2$  do not depend on each

other, but he assumes that both orders of smoothness are larger than  $3/2$ . To enhance the comparison, let us say that our procedure is adaptive, deals with random and dependent regression variables including the autoregressive framework, requires for the noise the finiteness of moments of a given order, 16 in many cases (and not any order  $p$ ) and reaches the optimal rate (without any loss) provided that the mean function is smoother than the variance function (namely,  $\alpha \geq 2\beta + 1/2$  if  $\alpha$  and  $\beta$  are the smoothness orders of  $b$  and  $\sigma^2$  respectively). On the one hand, this condition is less attractive than Neumann's ( $\alpha > 1$ ) in his independent framework or Hoffmann's and is only a technical loss with no other structural reason than the use of a unique first step estimator of  $b$  to estimate  $\sigma^2$ . Note that, contrary to Hoffmann's result, it allows to reach low orders of smoothness for  $b$  ( $\alpha > 1/2$ ) and for  $\sigma^2$  (namely  $1/2 < \beta < 3/2$ ). On the other hand, to separate the variance of the noise from the mean function, it is empirically natural to ask that the latter is much smoother than the former, otherwise it is hard to distinguish between them.

The plan of the paper is as follows. Section 2 presents the whole estimation procedure, namely the building of both estimators of  $b$  and  $\sigma^2$  and the assumptions on the functions, the variables and the collections of models. The results in terms of inequality of type (1.8) and of minimax rates on Besov balls are given in Section 3. Section 4 explains our simulation methods and describes the results of intensive simulation experiments. We used in particular models recently studied by Härdle and Tsybakov (1997) and Fan and Yao (1998) but also many others. Lastly, almost all proofs are gathered in Section 5 while section 6 contains some complementary informations about the simulations.

## 2 The estimation procedure

### 2.1 Assumptions on the linear spaces of estimation

We assume that we aim to estimate the functions on a given compact set  $A$ . We consider families of linear subspaces  $S_m$  of  $\mathbb{L}_2(A, dx)$  and we call those families *collections of models*. It is standard to set the following assumptions on the collections  $(S_m^{(i)})_{m \in \mathcal{M}_n^{(i)}}$ ,  $i = 1, 2$ :

- ( $\mathbf{H}_{\Phi_i}$ )
1. Each  $S_m^{(i)}$  is a finite dimensional linear subspace of  $\mathbb{L}_2(A, dx)$  with dimension  $\dim(S_m^{(i)}) = D_m^{(i)}$  and maximal dimension denoted by  $D_n^{(i)}$ .
  2. There exists a constant  $\Phi_i$  such that for any pair  $(m, m') \in (\mathcal{M}_n^{(i)})^2$ , and any  $t \in S_m^{(i)} + S_{m'}^{(i)}$

$$\|t\|_\infty \leq \Phi_i \sqrt{\dim(S_m^{(i)} + S_{m'}^{(i)})} \|t\| \quad (2.9)$$

where  $\|t\| = \int_A t^2(x) dx = \int_A t^2(x) dx$ .

3. There exists a constant  $K$  such that  $D_n^{(i)} \leq K\sqrt{n}/\ln(n)$  in the general case,  $D_n^{(i)} \leq Kn/\ln^2(n)$  for wavelets (family **(W)** below) and for piecewise polynomials (families **(DP)** and **(RP)** below).

( $\mathbf{H}_{(a_i, b_i)}$ ) There exist some nonnegative constants  $a_i, b_i, \Sigma_i, T_i$  such that

$$\sum_{m \in \mathcal{M}_n^{(i)}} (D_m^{(i)})^{-a_i} \leq \Sigma_i < \infty$$

and  $|\mathcal{M}_n^{(i)}| \leq T_i n^{b_i}$ .

**Comments. 1.** Assumption  $(\mathbf{H}_{\Phi_i})$  2. is an assumption of connection between the two norms  $\|\cdot\|_\infty$  and  $\|\cdot\|$ . It implies in particular that for all  $t \in S_m^{(i)}$ ,  $\|t\|_\infty \leq \Phi_i \sqrt{D_m^{(i)}} \|t\|$ . It follows from Barron et al. (1999), equations (3.2) and (3.3), that, for any orthonormal basis  $(\varphi_\lambda)_{\lambda \in \Lambda}$  of  $S_m^{(i)} + S_{m'}^{(i)}$ :

$$\left\| \sum_{\lambda \in \Lambda} \varphi_\lambda^2 \right\|_\infty = \sup_{t \in S_m^{(i)} + S_{m'}^{(i)}, t \neq 0} \frac{\|t\|_\infty}{\|t\|}. \quad (2.10)$$

**2.** Assumption  $(\mathbf{H}_{(a_i, b_i)})$  is a limitation on the number of models which have the same dimension and consequently on the global number of models. It guarantees in particular that we do not consider too many models. Note also that the choice  $b_i = a_i$ ,  $T_i = \Sigma_i$  suits. Indeed, since  $D_m^{(i)} \leq n$ , for any  $m \in \mathcal{M}_n^{(i)}$ ,

$$\Sigma_i \geq \sum_{m \in \mathcal{M}_n^{(i)}} (D_m^{(i)})^{-a_i} \geq \sum_{m \in \mathcal{M}_n^{(i)}} n^{-a_i} = |\mathcal{M}_n^{(i)}| n^{-a_i}$$

which implies that:  $|\mathcal{M}_n^{(i)}| \leq \Sigma_i n^{a_i}$ . In other words the number of models is at most polynomial with respect to  $n$ .

We shall essentially consider in the sequel three kinds of specific families of models  $(S_m^{(i)})_{m \in \mathcal{M}_n^{(i)}}$  satisfying  $(\mathbf{H}_{\Phi_i})$  and  $(\mathbf{H}_{(a_i, b_i)})$ : trigonometric polynomials, wavelets and piecewise polynomials that can be described as follows.

- (T)** Trigonometric polynomials : we consider spaces of dimension  $D_m^{(i)}$  generated by the functions  $\varphi_0(x) = 1$ ,  $\varphi_{2j}(x) = \sqrt{2} \cos(2\pi jx)$ ,  $\varphi_{2j+1}(x) = \sqrt{2} \sin(2\pi jx)$  for  $j = 0, \dots, d_m^{(i)}$ , where  $D_m^{(i)} = 2d_m^{(i)} + 1$  is the dimension of  $S_m^{(i)}$ . Any such  $S_m^{(i)}$  is entirely defined by its dimension. The family of models  $\mathcal{M}_n^{(i)}$  is in that case the set of all possible dimensions such that  $(\mathbf{H}_{\Phi_i})$  2. holds:  $\mathcal{M}_n^{(i)} = \{1, \dots, \sqrt{n}\}$ . Here  $\Phi_i = \sqrt{2}$  in  $(\mathbf{H}_{\Phi_i})$  2.,  $a_i = 1 + \epsilon$ ,  $\forall \epsilon > 0$ , and  $b_i = 1/2$  in  $(\mathbf{H}_{(a_i, b_i)})$ .
- (RP)** Regular piecewise polynomials: we consider the regular partitions  $\mathcal{I}_m$  defined by  $\mathcal{I}_m = \{[j/m, (j+1)/m], j = 0, 1, \dots, m-1\}$ . Given some positive integer  $r$ , we define  $S_m^{(i)}$  to be the space of piecewise polynomials with degree bounded by  $r-1$  on the partition  $\mathcal{I}_m$ . Then  $D_m^{(i)} = rm$ . The maximal value of  $m$ ,  $m(n)$  is the greatest integer such that  $rm \leq n/\ln^2(n)$ , i.e.  $m(n) = \lfloor n/(r \ln^2(n)) \rfloor = |\mathcal{M}_n^{(i)}|$  where  $[z]$  denotes the integer part of  $z$ . Here  $\Phi_i = \sqrt{(r+2)(2r+1)}$  (see Barron et al. (1999) p. 323),  $a_i = 1 + \epsilon$ ,  $\forall \epsilon > 0$  and  $b_i = 1$  suit.
- (DP)** Dyadic piecewise polynomials: we consider now dyadic partitions  $\mathcal{I}_m = \{[j/2^m, (j+1)/2^m], j = 0, \dots, 2^m-1\}$ . Given some positive integer  $r$ , we define  $S_m^{(i)}$  to be the space of piecewise polynomials with degree bounded by  $r-1$  on the partition  $\mathcal{I}_m$ . Then  $D_m^{(i)} = r2^m$ . The maximal value of  $m$ ,  $m(n)$  is the greatest integer such that  $r2^m \leq n/\ln^2(n)$ , i.e.  $m(n) = \lfloor \ln(n/(r \ln^2(n)))/\ln(2) \rfloor = |\mathcal{M}_n^{(i)}|$ . Again  $\Phi_i = \sqrt{(r+2)(2r+1)}$  (see Barron et al. (1999) p. 323), but now any positive  $a_i, b_i$  suit.

(W) Compactly supported wavelets: Let  $\Lambda(j) = \{(j, k), k = 1, \dots, 2^j\}$  and let

$$\{\phi_{J_0, k}, (J_0, k) \in \Lambda(J_0)\} \cup \{\varphi_{j, k}, (j, k) \in \bigcup_{J=J_0}^{+\infty} \Lambda(J)\}$$

be an  $\mathbb{L}_2([0, 1], dx)$ -orthonormal system of compactly supported wavelets of regularity  $r$  built by Cohen et al. (1993); for a precise description, see Donoho and Jonstone (1998). These new functions derive from Daubechies (1992)'s wavelets at the interior of  $[0, 1]$  and are boundary corrected at the “edges”. For any  $J_n > J_0$ , let  $\mathcal{S}_n$  be the space spanned by the  $\phi_{J_0, k}$ 's for  $(J_0, k) \in \Lambda(J_0)$  and by the  $\varphi_{j, k}$ 's for  $(j, k) \in \bigcup_{J=J_0}^{J_n} \Lambda(J)$ . It follows that  $\dim(\mathcal{S}_n) = 2^{J_n} \leq n$  if  $J_n \leq \ln_2(n)$ . For any  $m \in \mathcal{M}_n = \{J_0, \dots, J_n - 1\}$ , we take for  $S_m^{(i)}$  the linear span of the  $\phi_{J_0, k}$ 's for  $(J_0, k) \in \Lambda(J_0)$  and of the  $\varphi_{j, k}$ 's for  $(j, k) \in \bigcup_{J=J_0}^m \Lambda(J)$ . This implies that  $D_m^{(i)} \leq 2^m$ . We know from Barron et al. (1999), p.322, that  $\Phi_i = 2 + \sqrt{2}$  suits and any positive  $a_i, b_i$  suit.

## 2.2 The assumptions on the model

All along the paper, we consider model (1.1) with  $\varepsilon_i$  i.i.d.,  $\mathbb{E}(\varepsilon_1) = \mathbb{E}(\varepsilon_1^3) = 0$  and  $\text{Var}(\varepsilon_1) = 1$ . We assume that the process  $(X_i)$  is strictly stationary. Let us recall that a stationary process  $(X_i)$  is said to be absolutely regular or  $\beta$ -mixing (Kolmogorov and Rozanov, 1960) if

$$\frac{1}{2} \sup \left\{ \sum_{i=1}^I \sum_{j=1}^J |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)| \right\} = \beta_k \rightarrow 0 \text{ when } k \rightarrow +\infty.$$

where the supremum is taken over all finite partitions  $(A_i)_{1 \leq i \leq I}$  and  $(B_j)_{1 \leq j \leq J}$  of the probability space  $\Omega$  respectively  $\mathcal{F}_{-\infty}^0$  and  $\mathcal{F}_k^\infty$  measurable where  $\mathcal{F}_i^k$  is the  $\sigma$ -algebra generated by  $\{X_j, i \leq j \leq k\}$ . The mixing is said to be geometrical if there exist positive  $M$  and  $\theta$  such that  $\beta_k \leq M e^{-\theta k}$ . The mixing is said to be arithmetical if there exist positive  $M$  and  $\theta$  such that  $\beta_k \leq M k^{-\theta}$ .

We work under the following assumptions:

A1  $(X_t)_{t \in \mathbf{Z}}$  is geometrically  $\beta$ -mixing.

A2(p)  $X$ ,  $b(X)$ ,  $\sigma(X)$  and  $\varepsilon$  admit moments until order  $p$ ,  $p \geq 4$ .

A3  $b$  and  $\sigma$  are bounded on compact sets.

A4  $X$  admits a density  $h_X$  such that for any compact set  $A$  in the support of  $h_X$ , there exist  $h_0, h_1$  (depending on  $A$ ) such that

$$\forall x \in A, \quad 0 < h_0 \leq h_X(x) \leq h_1. \quad (2.11)$$

Under A2(p), we denote by  $m_4 = \mathbb{E}[(\varepsilon^2 - 1)^2] (< \infty)$  and by  $\sigma_q^q = \mathbb{E}[|\varepsilon_1|^q]$  for  $q \in (0, p]$ . Note that assumptions A1, A2(p), A3 are fulfilled under standard assumptions given by Ango Nze (1992), Proposition 3, (see also Doukhan (1994) p. 107). More precisely, here is a set of assumptions implying A1-A4:

B1 There exists constants  $C_1 > 0$  and  $C_2 > 0$  such that, for all  $y \in \mathbb{R}$ ,

$$|b(y)| \leq C_1(1 + |y|), \quad |\sigma(y)| \leq C_2(1 + |y|).$$

B2 The function  $\sigma$  satisfies  $\inf_{y \in \mathbb{R}} \sigma(y) > C_3$  for a  $C_3 > 0$ .

B3(p)  $\mathbb{E}(|\varepsilon_1|^p) < +\infty$  for some  $p \geq 4$  and  $\mathbb{E}[C_1 + C_2|\varepsilon_1|]^p < 1$ .

B4 The density  $h_\varepsilon$  of  $\varepsilon$  exists and  $h_\varepsilon$  is continuous on its support.

Those assumptions are quite near of those required by Härdle and Tsybakov (1997). Under B1-B4, the Markov chain  $(X_i)$  given by (1.1) is geometrically ergodic and the stationary law is geometrically absolutely regular; this ensures A1.

Under B3(p), we know (see Duflo (1990), p.178) that for any initial condition  $X_0$  in  $\mathbb{L}^p$  independent of  $\varepsilon$ , the  $X_i$ 's admit moments of the same order as the  $\varepsilon_i$ 's (and thus, so do  $\sigma(X_i)$  and  $b(X_i)$  with B1). Thus B3(p) ensures A2(p).

As a consequence of B1, it is clear that  $b$  and  $\sigma$  are bounded on compact sets, which gives A3. Note that we estimate  $b$  and  $\sigma$  on the compact set  $A$  only, the same for both functions.

Moreover, if  $\mu$  denotes the stationary law of  $X_1$  (which exists under B1-B4), we know with B2 and B4 that  $d\mu(x) = h_X(x)dx$  with:

$$h_X(x) = \int h_\varepsilon \left( \frac{x - b(u)}{\sigma(u)} \right) \frac{1}{\sigma(u)} d\mu(u).$$

Indeed the positivity of  $\sigma$  ensures that the change of variable can be done and the continuity of  $h_\varepsilon$  implies the continuity of  $h_X$ . Thus  $h_X$  is positive on its support and continuous which ensures A4 for any compact set  $A$  in the support of  $h_X$ .

In other words B1-B2-B3(p)-B4 imply A1-A2(p)-A3-A4.

Since the random variables  $X_i$  are geometrically  $\beta$ -mixing, this will allow to apply some results established in Baraud et al. (1999).

**Comments:** **1.** Ango Nze (1998) gives also conditions on autoregressive models to generate arithmetically mixing variables still admitting a stationary ergodic law. Moreover the results of Baraud et al. (1999) also allow to consider arithmetically mixing variables. This implies some robustness of the results with respect to stronger types of dependence. But such results lead to much stronger conditions on the errors and on the size of the collections of models.

**2.** All the given results would hold for model (1.2) with  $u_i$  i.i.d.,  $\mathbb{E}(u_1) = \mathbb{E}(u_1^3) = 0$  and  $\text{var}(u_1) = 1$ , (see for such extensions Baraud et al. (1999)) under the assumptions A3, A4 and

C1  $(Y_t, X_t)_{t \in \mathbf{Z}}$  is geometrically  $\beta$ -mixing.

C2(p)  $Y$ ,  $b(X)$ ,  $\sigma(X)$  and  $\varepsilon$  admit moments until order  $p$ ,  $p \geq 4$ .

**3.** Lastly the real valued random variables  $X_i$  could be replaced by a  $r$ -dimensional random vector  $\vec{X}_i = (X_i^{(1)}, \dots, X_i^{(r)})$  under the same kind of assumptions and the autoregression of order one can in the same way be generalized into an autoregression of order  $r$ . For the extension of assumptions B1-B4 ensuring A1-A4, see Ango Nze (1992) or the application of these results in Härdle et al. (1998). The functions  $b$  and  $\sigma$  remain real valued and the errors  $\varepsilon_i$  as well, which makes most extensions straightforward.

## 2.3 First step of the estimation procedure

To estimate  $b$  on a given compact set  $A$ , we consider the contrast

$$\gamma_n^{(1)}(t) = \frac{1}{n} \sum_{i=1}^n [X_{i+1} - t(X_i)]^2 \quad (2.12)$$

based on the observations  $X_1, X_2, \dots, X_{n+1}$ . We consider a collection of linear subspaces of  $\mathbb{L}_2(A, dx)$ ,  $(S_m^{(1)})_{m \in \mathcal{M}_n^{(1)}}$  of dimension  $D_m^{(1)}$ , as described in section 2.1 and satisfying Assumption  $(\mathbf{H}_{\Phi_1})$ . Baraud et al. (1999) proved non asymptotic risk bounds for the estimate  $\hat{b}_{\hat{m}_1}$  defined as follows, when the variance is a known constant denoted by  $\sigma_2$ . Let

$$\hat{b}_m \text{ be a minimizer of } \gamma_n^{(1)}(t), \text{ over } t \in S_m^{(1)}.$$

The  $\hat{b}_m$ 's define a collection of estimators of  $b$ . Then choose:

$$\hat{m}_1 = \arg \min_{m \in \mathcal{M}_n^{(1)}} \left( \gamma_n^{(1)}(\hat{b}_m) + \text{pen}(m) \right) \quad \text{where } \text{pen}(m) = \kappa \sigma_2^2 \frac{D_m^{(1)}}{n},$$

and  $\kappa$  is a universal constant. Baraud et al. (1999)'s results extend straightforwardly to a known varying variance function by considering the estimate  $\hat{b}_{\hat{m}_1}$  with:

$$\hat{m}_1 = \arg \min_{m \in \mathcal{M}_n} \left( \gamma_n^{(1)}(\hat{b}_m) + \text{pen}_{th}^{(1)}(m) \right) \quad \text{where } \text{pen}_{th}^{(1)}(m) = \kappa \Phi_1^2 \|\sigma\|_\mu^2 \frac{D_m^{(1)}}{n},$$

where  $\mu$  is the stationary law of the  $X_i$ 's and  $\kappa$  a universal constant. Then  $\hat{b}_{\hat{m}_1}$  has the same properties as in the case of a known constant variance.

As  $\|\sigma\|_\mu^2$  is unknown, we complete the procedure by replacing this quantity by an estimate. Let

$$\hat{r}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_{i+1} - \hat{b}_{m_n}(X_i))^2 \quad (2.13)$$

where  $S_{m_n}^{(1)}$  is a space of the family for a given  $m_n \in \mathcal{M}_n^{(1)}$  with dimension  $D_{m_n}^{(1)}$  to be chosen in Theorem 3.1 (see also the comments herewith).

Then we define the final estimate as

$$\tilde{b} := \hat{b}_{\hat{m}_1} \text{ where } \hat{m}_1 = \arg \min_{m \in \mathcal{M}_n^{(1)}} \left( \gamma_n^{(1)}(\hat{b}_m) + \text{pen}^{(1)}(m) \right) \quad (2.14)$$

with

$$\text{pen}^{(1)}(m) = \kappa \Phi_1^2 \hat{r}_n^2 \frac{D_m^{(1)}}{n} \quad \text{and } \hat{r}_n^2 \text{ given by (2.13).} \quad (2.15)$$

**Comment:** It is now well-known that it is safer to take for  $\kappa$  too great than too small values. An empirical calibration study, similar to the one extensively done for density estimation by Birgé and Rozenholc (2000), can be lead in order to compute  $\kappa$ . When the collection of models is chosen,  $\Phi_1$  is known but it is probably a computational artefact rather than a structural constant of the penalty. Indeed, in an independent fixed design framework with constant volatility  $\sigma_2^2$ , the optimal penalty is found by Baraud (2000) to be  $2\sigma_2^2 D_m^{(1)}/n$ .

## 2.4 Second step of the estimation procedure

We consider now the following procedure. Let  $S_m^{(2)}$ ,  $m \in \mathcal{M}_n^{(2)}$ , be a collection of linear subspaces of  $\mathbb{L}_2(A, dx)$ , of dimension  $D_m^{(2)}$ , as described in section 2.1 and satisfying assumption  $(\mathbf{H}_{\Phi_2})$ . Let

$$\gamma_n^{(2)}(t) = \frac{1}{n} \sum_{i=1}^n \left[ X_{i+1}^2 - \tilde{b}^2(X_i) - t(X_i) \right]^2 \quad (2.16)$$

and define  $\widehat{\sigma}_m^{(2)}$  as a minimizer of  $\gamma_n^{(2)}(t)$  over  $t \in S_m^{(2)}$ . Then our estimate is

$$\widetilde{\sigma}^2 = \widehat{\sigma}_{\hat{m}_2}^{(2)} \text{ with } \hat{m}_2 = \arg \min_{m \in \mathcal{M}_n^{(2)}} \left( \gamma_n^{(2)}(\widehat{\sigma}_m^{(2)}) + \text{pen}^{(2)}(m) \right) \quad (2.17)$$

where

$$\text{pen}^{(2)}(m) = \kappa \Phi_2^2 \hat{s}_n^2 \frac{D_m^{(2)}}{n}. \quad (2.18)$$

and

$$\hat{s}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_{i+1}^2 - \hat{g}_{m_n}(X_i))^2 \text{ and } \hat{g}_{m_n} = \arg \min_{t \in S_{m_n}^{(2)}} \frac{1}{n} \sum_{i=1}^n [X_{i+1}^2 - t(X_i)]^2 \quad (2.19)$$

on some well chosen  $S_m^{(2)} = S_{m_n}^{(2)}$ . The theoretical value of the penalty that  $\text{pen}^{(2)}$  estimates is:

$$\text{pen}_{th}^{(2)} = \kappa \Phi_2^2 (m_4 \mathbb{E}_\mu(\sigma^4) + 4 \mathbb{E}_\mu(b^2 \sigma^2)) \frac{D_m^{(2)}}{n}.$$

**Comment.** The choice  $(1/n) \sum_{i=1}^n [(X_{i+1} - \tilde{b}(X_i))^2 - t(X_i)]^2$  for the contrast is more standard and is the one empirically used. Only technical reasons lead to our slightly different choice.

## 3 The theoretical results

### 3.1 Estimation of the mean

Recall that the empirical euclidian norm is  $\|u\|_n^2 = (1/n) \sum_{i=1}^n u^2(X_i)$  and that  $A$  is the given compact set on which we aim to estimate the functions. We denote by  $b_m$  the  $\mathbb{L}_2(A, dx)$ -orthogonal projection of  $b$  on  $S_m$ . We have the following result:

**Theorem 3.1** *Let  $X_1, \dots, X_n$  be a stationary sequence drawn from model (1.1) and consider a collection of models satisfying  $(\mathbf{H}_{\Phi_1})$  and  $(\mathbf{H}_{(a_1, b_1)})$ . Assume that A1, A2(p), A3, A4 are fulfilled with*

$$p \geq 8, \quad p \geq 2(1 + a_1) \quad \text{and} \quad p > 6 + 4b_1, \quad (3.20)$$

*then  $\tilde{b}$ , defined by (2.14)-(2.15), with  $\hat{r}_n^2$  defined by (2.13) and such that*

$$\dim(S_{m_n}^{(1)}) = D_{m_n}^{(1)} \leq n^{1/2-2/p}, \quad (3.21)$$

*satisfies*

$$\mathbb{E}[\|b\mathbf{I}_A - \tilde{b}\|_n^2] \leq C \inf_{m \in \mathcal{M}_n^{(1)}} \left( \|b\mathbf{I}_A - b_m\|_\mu^2 + \frac{D_m^{(1)}}{n} (\|b\|_\mu^2 + \|\sigma\|_\mu^2) \right) + \frac{R}{n}$$

where  $C$  is a universal constant and  $R$  is a constant depending on  $\sigma_p$ ,  $\Phi_1$ ,  $\|b\mathbf{I}_A\|_\infty$ ,  $\|\sigma\mathbf{I}_A\|_\infty$ ,  $\Sigma_1$ ,  $C_1$ .

**Comments: 1.** The estimate performs almost as well as the best estimator that could be chosen among the collection. We insist on the fact that the procedure automatically selects a model very close to the (unobservable) best model (called oracle) in the collection i.e. the most adequate dimension for the space of approximation.

**2.** For the families **(W)** and **(DP)**, any  $a_1, b_1 > 0$  suit so that condition (3.20) reduces to  $p \geq 8$ . For the family **(T)**, as  $a_1 = 1 + \varepsilon$  for any  $\varepsilon > 0$  and  $b_1 = 1/2$ , condition (3.20) becomes  $p > 8$ . For family **(RP)**, if we want to consider the maximal number of possible models, we find  $p \geq 10$ .

**3.** In the general case, under condition (3.20), the constraint (3.21) is fulfilled as soon as  $D_{m_n}^{(1)} \leq n^{1/4}$ . For Gaussian errors, we can take  $p \rightarrow +\infty$  and we find  $D_{m_n}^{(1)} \leq \sqrt{n}$ .

This kind of result is known to lead to results of adaptation to unknown smoothness. For further applications, see Barron et al. (1999). We first recall that a function  $f$  belongs to the Besov space  $\mathcal{B}_{\alpha,l,\infty}([0,1])$  if it satisfies

$$|f|_{\alpha,l} = \sup_{y>0} y^{-\alpha} w_d(f, y)_l < +\infty, \quad d = [\alpha] + 1,$$

where  $w_d(f, y)_l$  denotes the modulus of smoothness. For a precise definition of those notions, we refer to DeVore and Lorentz (1993) Chapter 2, Section 7, where it also proved that  $\mathcal{B}_{\alpha,l,\infty}([0,1]) \subset \mathcal{B}_{\alpha,2,\infty}([0,1])$  for  $l \geq 2$ . This justifies that we now restrict our attention to  $\mathcal{B}_{\alpha,2,\infty}(A)$ .

**Proposition 3.1** *Assume that the Assumptions of Theorem 3.1 hold and consider the families **(RP)**, **(DP)**, **(W)** or **(T)**. Let  $\alpha$  be a real number greater than  $1/2$ ,  $\alpha \leq r$  for **(RP)**, **(DP)** and **(W)** and assume that  $b$  belongs to some Besov space  $\mathcal{B}_{\alpha,2,\infty}(A)$ . Then*

$$\left( \sup_{b \in \mathcal{B}_{\alpha,2,\infty}(L)} \mathbb{E} \|b\mathbf{I}_A - \tilde{b}\|_n^2 \right)^{\frac{1}{2}} \leq C(\alpha, L) n^{-\frac{\alpha}{2\alpha+1}} \quad (3.22)$$

where  $\mathcal{B}_{\alpha,2,\infty}(L) = \{t \in \mathcal{B}_{\alpha,2,\infty}(A), |t|_{\alpha,2} \leq L\}$ .

*Proof.* The result is straightforward with Lemma 12 in Barron et al. (1999) which imply that  $\|b\mathbf{I}_A - b_m\|$  is of order  $(D_m^{(1)})^{-\alpha}$  on the specified collections of models. Moreover the norm  $\mu$  on the compact  $A$  is bounded by  $h_1$  times the Lebesgue-norm as mentioned in (2.11).  $\square$

**Remark.** Since the optimal choice is  $D_{m^*}^{(1)} = n^{1/(2\alpha+1)}$ , it satisfies in particular  $D_{m^*}^{(1)} \leq \sqrt{n}$ ,  $\forall \alpha > 1/2$ . This allows to reach the optimal rate even with the family **(T)**, restricted to dimensions less than  $\sqrt{n}$ .

## 3.2 Estimation of $\sigma^2$

The result for the variance function can be given in two steps.  $\sigma_m^2$  denotes the  $\mathbb{L}_2$  projection of  $\sigma^2$  on  $S_m^{(2)}$ .

**Theorem 3.2** *Let  $X_1, \dots, X_n$  be a stationary sequence drawn from model (1.1) and consider a collection of models satisfying  $(\mathbf{H}_{\Phi_2})$  and  $(\mathbf{H}_{(a_2, b_2)})$ . Assume that A1, A2(p), A3, A4 are fulfilled with*

$$p \geq 16, \quad p \geq 4(1 + a_2) \quad \text{and} \quad p > 8b_2 + 12, \quad (3.23)$$

*then  $\tilde{\sigma}^2$ , defined by (2.17)-(2.18), with  $\hat{s}_n^2$  defined by (2.19) and such that*

$$\dim(S_{m_n}^{(2)}) = D_{m_n}^{(2)} \leq n^{1/2-4/p}, \quad (3.24)$$

*satisfies*

$$\begin{aligned} \mathbb{E}[\|\sigma^2 \mathbf{I}_A - \tilde{\sigma}^2\|_n^2] &\leq C \inf_{m \in \mathcal{M}_n^{(2)}} \left( \|\sigma^2 \mathbf{I}_A - \sigma_m^2\|_\mu^2 + S^2 \Phi_2^2 \frac{D_m^{(2)}}{n} \right) \\ &\quad + \frac{R}{n} + C' \mathbb{E}[\|b^2 \mathbf{I}_A - \tilde{b}^2\|_n^2] \end{aligned} \quad (3.25)$$

*where  $C$  and  $C'$  are universal constants,  $S^2 = \|\sigma^2 + b^2\|_\mu^2 + m_4 \|\sigma^2\|_\mu^2 + 4\|b\sigma\|_\mu^2$  and  $R$  is a constant depending on  $\sigma_{16}$ ,  $\Phi_2$ ,  $\|b \mathbf{I}_A\|_\infty$ ,  $\|\sigma \mathbf{I}_A\|_\infty$ .*

Condition (3.23) reduces to  $p \geq 16$  for families **(DP)** or **(W)**, to  $p \geq 20$  for family **(RP)** and to  $p > 16$  for the family **(T)**.

**Remark.** Note that considering the contrast

$$\tilde{\gamma}_n(t) = \frac{1}{n} \sum_{i=1}^n [X_{i+1}^2 - t(X_i)]^2$$

leads to an estimate  $\tilde{f}$  of  $f = b^2 + \sigma^2$ . In particular, it is possible to provide in an analogous way a bound for  $\|f - \tilde{f}\|_n^2$ . This gives the rate  $n^{-\gamma/(2\gamma+1)}$  where  $\gamma = \min(\alpha, \beta)$  if  $b$  belongs to a Besov space  $\mathcal{B}_{\alpha,2,\infty}$  and  $\sigma^2$  to a Besov space  $\mathcal{B}_{\beta,2,\infty}$ . But it does not allow to separate the smoothness  $\alpha$  and  $\beta$  of  $b$  and  $\sigma$ , without avoiding the loss in the rate when coming back to the evaluation of the rate of convergence of the estimator of  $\sigma^2$  given by  $\tilde{f} - (\tilde{b})^2$ .

The interest of (3.25) is to illustrate the dependency in the first step estimator, and to show where some loss in the rates of convergence can happen. Indeed as

$$\begin{aligned} \|b^2 \mathbf{I}_A - \tilde{b}^2\|_n^2 &= \|(b \mathbf{I}_A - \tilde{b})(b \mathbf{I}_A + \tilde{b})\|_n^2 = \|(b \mathbf{I}_A - \tilde{b})(2b \mathbf{I}_A - (b \mathbf{I}_A - \tilde{b}))\|_n^2 \\ &\leq 4\|b \mathbf{I}_A\|_\infty \|b \mathbf{I}_A - \tilde{b}\|_n^2 + 2\|(b \mathbf{I}_A - \tilde{b})^2\|_n^2 \\ &\leq 4\|b \mathbf{I}_A\|_\infty \|b \mathbf{I}_A - \tilde{b}\|_n^2 + 2n\|b \mathbf{I}_A - \tilde{b}\|_n^4, \end{aligned}$$

so that we can find as another extension of Theorem 3.1 and as a tool for completing (3.25) in Theorem 3.2, the following bound:

**Proposition 3.2** *Under the Assumptions of Theorem 3.1 and if*

$$p \geq 2(a_1 + 2) \quad \text{and} \quad p > 4b_1 + 10, \quad (3.26)$$

we have:

$$\begin{aligned} \mathbb{E}(\|b\mathbf{I}_A - \tilde{b}\|_n^4) &\leq C \inf_{m \in \mathcal{M}_n^{(1)}} \left( \|b\mathbf{I}_A - b_m\|^4 + \frac{\|b\mathbf{I}_A - b_m\|_8^2}{n} + \frac{(D_m^{(1)})^2}{n^2} \right) \\ &\quad + \frac{R'}{n^2}. \end{aligned} \quad (3.27)$$

where  $C$  depends now on  $C$ ,  $h_1$ ,  $M$ ,  $\theta$  and  $\|b\|_\mu^2 + \|\sigma\|_\mu^2$  and  $\|f\|_8^8 = \int |f(x)|^8 dx$ .

Therefore if  $b$  belongs to some Besov space  $\mathcal{B}_{\alpha,2,\infty}$  for  $\alpha > 1/2$ , then  $\|b\mathbf{I}_A - b_m\|_8$  is of order  $(D_m^{(1)})^{-(\alpha-(1/2-1/8))}$ ,  $\|b\mathbf{I}_A - b_m\|$  is of order  $(D_m^{(1)})^{-\alpha}$ . Therefore, choosing  $D_m^{(1)}$  of order  $n^{1/(2\alpha+1)}$  ensures that the infimum in (3.27) is less than  $Cn^{-4\alpha/(2\alpha+1)}[1+n^{-(2\alpha-1/2)/(2\alpha+1)}]$  and therefore less than  $2Cn^{-4\alpha/(2\alpha+1)}$ ,  $\forall \alpha > 1/2$ . The rate corresponding to the term depending on  $b$  via  $n\mathbb{E}(\|b\mathbf{I}_A - \tilde{b}\|_n^4)$  is

$$n \times n^{-4\alpha/(2\alpha+1)} = n^{-(2\alpha-1)/(2\alpha+1)}.$$

Next if  $\sigma^2$  is in some  $\mathcal{B}_{\beta,2,\infty}$ , then the first term of the right-hand-side of (3.25) is of order

$$n^{-\frac{2\beta}{2\beta+1}}.$$

Thus it is easy to see that the minimax rate is obtained for  $\sigma^2$  if

$$\alpha \geq 2\beta + 1/2,$$

i.e. it requires the regularity of  $b$  to be significantly greater than that of  $\sigma^2$ . Moreover for the part  $\mathbb{E}(\|b\mathbf{I}_A - \tilde{b}\|_n^2)$  which has rate  $n^{-2\alpha/(2\alpha+1)}$ , it is negligible with respect to  $n^{-2\beta/(2\beta+1)}$  as soon as  $\alpha\beta$ . Therefore, we proved the following result:

**Proposition 3.3** *Assume that the Assumptions of Theorems 3.1, 3.2 and (3.26) hold and consider the collections of models (DP), (RP), (W) or (T). Let  $\alpha$  and  $\beta$  be real numbers greater than  $1/2$  and less than  $r$  for families (RP), (DP) or (W) and assume that  $b$  belongs to some Besov space  $\mathcal{B}_{\alpha,2,\infty}(A)$  and that  $\sigma^2$  belongs to some Besov space  $\mathcal{B}_{\beta,2,\infty}(A)$  with  $\alpha \geq 2\beta + 1/2$ . Then*

$$\sup_{b \in \mathcal{B}_{\alpha,2,\infty}(R_1,R_2), \sigma^2 \in \mathcal{B}_{\beta,2,\infty}(L)} \mathbb{E}\|\sigma^2\mathbf{I}_A - \widetilde{\sigma^2}\|_n^2 \leq C(\alpha, L, R_1, R_2)n^{-\frac{2\beta}{2\beta+1}} \quad (3.28)$$

where  $\mathcal{B}_{\alpha,2,\infty}(R_1, R_2) = \{t \in \mathcal{B}_{\alpha,2,\infty}(A), |t|_{\alpha,2} \leq R_1, |t|_\infty \leq R_2\}$  and  $\mathcal{B}_{\beta,2,\infty}(L) = \{t \in \mathcal{B}_{\beta,2,\infty}(A), |t|_{\beta,2} \leq L\}$ .

**Comments: 1.** If the condition  $\alpha \geq 2\beta + 1/2$  is not fulfilled, the rate becomes  $n^{-(2\alpha-1/2)/(2\alpha+1)}$  and is clearly suboptimal.

**2.** It has already been mentioned in the introduction that Neumann (1994) reaches the optimal rate for the estimation of  $\sigma^2$  under the simpler condition  $\alpha > 1$ ; but he works with a fixed design regressive model under moment condition of any order for  $\varepsilon$ . It is also worth comparing this result with Hoffmann (1999) who deals with a more general risk  $\mathbb{L}^{p'}$  and with functions belonging to more general Besov spaces  $\mathcal{B}_{s,p,q}$ , with  $s = \alpha$  or  $s = \beta$ . Taking  $p' = p = 2$  and  $q = \infty$  in his main result for comparison shows that his conditions reduce simply to  $\alpha > 3/2$  (even when estimating  $b$  alone) and  $\beta > 3/2$ . Moreover, he requires the finiteness of exponential moments of the noise and reaches the optimal rate up to  $\ln(n)$  factors. Therefore, the result given in Proposition 3.1 is always better, and the result given by Proposition 3.3 is better if  $\alpha \geq 2\beta + 1/2$ .

## 4 Simulation results

We generate samples using several regressive and autoregressive models. All models are denoted by  $Y_i = b(X_i) + \sigma(X_i)\varepsilon_{i+1}$ , with possibly  $Y_i = X_{i+1}$  in the autoregressive case. For all paths, to make sure that the process has reached stationarity in the autoregressive case, we forget the 200 first data. For each model, we consider  $S = 400$  samples with length  $n = 500$  which provides paths denoted by  $(Y_i^{(s)}, X_i^{(s)})_{1 \leq i \leq n}$  for  $s = 1, \dots, S$ . We consider various couples of regression or autoregression functions  $(b, \sigma)$ . The couples of functions are gathered in Table 1. The values of the parameters in the regressive and autoregressive cases are given in the appendix (section 6).

| Model            | drift and volatility                                                                                                 |
|------------------|----------------------------------------------------------------------------------------------------------------------|
| M1               | $b(x) = 0.4x + 1, \sigma(x) \equiv \sigma$                                                                           |
| M2               | $b(x) = (0.5 + 0.25x) \exp(0.5 - 0.25x), \sigma(x) \equiv \sigma$                                                    |
| M3               | $b(x) = 0.5(x + 2 \exp(-16x^2)), \sigma(x) \equiv \sigma$                                                            |
| M4               | $b(x) = \sin(2x) + 2 \exp(-16x^2), \sigma(x) \equiv \sigma$                                                          |
| M5, M6 & M7      | $b(x) = \sin(2\omega\pi x + \pi/3), \sigma(x) \equiv \sigma$                                                         |
| M8               | $b(x) = \sin(2\pi x + \pi/3), \sigma(x) = \sigma\sqrt{ x }$                                                          |
| M9               | $b(x) = \sin(2\pi x + \pi/3)$<br>$\sigma(x) = \sigma(0.31 + 0.7 \exp(-5x^2))$                                        |
| M10              | $b(x) = a(x + 2 \exp(-16x^2))$<br>$\sigma(x) = \sigma(0.2 + 0.4 \exp(-2x^2))$                                        |
| M11 <sup>1</sup> | $b(x) = 1/(1 + \exp(-x))$<br>$\sigma(x) = \sigma(\varphi(x + 1.2) + 1.5\varphi(x - 1.2))$                            |
| M12 & M13        | $b(x) = ax$<br>$\sigma(x) = 0.05 + 1/(1 + \beta x^2)$                                                                |
| M14 & M15        | $b(x) = ax, \sigma(x) = 0.05 + \pi/2 + \arctan(\beta x)$                                                             |
| M16              | $b(x) = \begin{cases} a x  & \text{if } x < x_0, \\ a(x - 2x_0) & \text{else.} \end{cases} \quad \sigma(x) = \sigma$ |

Table 1. Couples of functions used to generate the models

Note that the regressive case corresponds to the independent framework and the autoregressive case corresponds to the dependent context. Moreover models M1 to M7 correspond to (auto-)regressive models with constant volatility. Model M8 studies the problem of possible nullity of the variance function, together with some regularity problems in the volatility function. The models M10 and M11 are the one studied by Fan and Yao (1998) and Härdle and Tsybakov (1997) respectively. Lastly, model M16 studies the effect of a discontinuity in the mean function.

In the regressive case, the parameters are chosen to give some fixed level of the signal to noise ratio, denoted in all the following by  $s2n$ . Since in the regressive case, the  $X_i$ 's are taken uniform on  $[-2, 2]$ , we have  $s2n(\text{reg}) = \int_{-2}^2 b^2(x)dx / \int_{-2}^2 \sigma^2(x)dx$ . In the autoregressive case, the choice of the parameters is done both to ensure the stability of the models and to provide some given signal to noise ratio. Since the law of the  $X_i$ 's is unknown in this case, we compute for a given long sample  $s2n(\text{autoreg}) = \sum_{i=1}^n b^2(X_i) / \sum_{i=1}^n \sigma^2(X_i)$  and choose the coefficients giving the desired value of  $s2n(\text{autoreg})$  in this particular case.

---

<sup>1</sup> $\varphi$  is the normal density

The results for models M12, M14, M16 are not reported in that context because the adjustment of most s2n ratios generate unstable models.

The estimation procedure is done using for both  $b$  and  $\sigma^2$  the collection of models **(RP)** with degree  $r \leq 5$ . We have implemented several procedures: six procedures working with piecewise polynomials of given (fixed) degree from  $r = 0$  to  $r = 5$ , and a seventh procedure that chooses among those six global degrees the best one in terms of a penalized contrast. The interest of fixed degree estimation is that we can compute oracles which provide a benchmark to evaluate the performances of our estimates. More precisely, for each model, each degree  $r$ , each given dimension  $D = 1, \dots, D_{max} = \lceil n/((r+1)\ln(n)) \rceil$ , we compute

$$L^2(b, r, D) = \frac{1}{S} \sum_{s=1}^S \left( \frac{1}{n} \sum_{i=1}^n [b(X_i^{(s)}) - \hat{b}_D^{(s)}(X_i^{(s)})]^2 \right),$$

where  $\hat{b}_D^{(s)}$  is a mean square estimator based on the sample  $(Y_i^{(s)}, X_i^{(s)})_{1 \leq i \leq n}$ , as an estimation of  $\mathbb{E}[\|b - \hat{b}_D\|_n^2]$ . Then we know

$$L_{opt}^2(b, r) = \min_{1 \leq D \leq D_{max}} L^2(b, r, D) \text{ and } D_{opt} = \arg \min_{1 \leq D \leq D_{max}} L^2(b, r, D).$$

The oracle is then given by

$$L_{oracle}^2(b) = \min_{0 \leq r \leq 5} L_{opt}^2(b, r).$$

We define and compute analogously the oracles for  $\sigma^2$ ,  $L_{opt}^2(\sigma^2, r)$  by using  $([Y_i^{(s)} - b(X_i^{(s)})]^2, X_i^{(s)})$  as new data set and keep  $L_{oracle}^2(\sigma^2) = \min_{0 \leq r \leq 5} L_{opt}^2(\sigma^2, r)$ . Note that the oracles for  $\sigma^2$  are computed with assuming that  $b$  is known. The oracles gives the best reachable performance, and are in practice unknown since the choice is performed with respect to the true function. The computation of the oracles represents the (very) long part (in time) of the numerical procedure.

|     | s2n=1 |          | s2n=3 |          | s2n=7 |          | s2n=10 |          |
|-----|-------|----------|-------|----------|-------|----------|--------|----------|
|     | $b$   | $\sigma$ | $b$   | $\sigma$ | $b$   | $\sigma$ | $b$    | $\sigma$ |
| M1  | 2.6   | 0.7      | 1.7   | 0.7      | 1.2   | 0.7      | 1.1    | 0.7      |
| M2  | 1.5   | 0.7      | 2.1   | 0.7      | 1.4   | 0.7      | 1.5    | 0.7      |
| M3  | 1.4   | 0.7      | 1.8   | 0.7      | 1.4   | 0.8      | 1.7    | 0.8      |
| M4  | 1.6   | 0.7      | 1.4   | 0.8      | 1.7   | 0.8      | 1.4    | 0.8      |
| M5  | 1.7   | 0.7      | 1.1   | 0.7      | 1.4   | 0.7      | 1.4    | 0.7      |
| M6  | 1.3   | 0.7      | 1.3   | 0.7      | 1.5   | 0.8      | 1.5    | 0.3      |
| M7  | 1.4   | 0.9      | 1.3   | 0.9      | 1.3   | 1.0      | 1.3    | 1.0      |
| M8  | 1.3   | 1.4      | 1.2   | 1.3      | 1.6   | 1.4      | 1.2    | 1.5      |
| M9  | 1.6   | 1.2      | 1.4   | 1.2      | 1.5   | 1.2      | 1.6    | 1.2      |
| M10 | 1.8   | 1.4      | 2.1   | 1.4      | 1.5   | 1.5      | 1.6    | 1.4      |
| M11 | 2.7   | 2.1      | 1.6   | 2.0      | 1.2   | 2.1      | 1.1    | 2.1      |
| M12 | 1.7   | 1.0      | 1.2   | 1.0      | 1.1   | 1.0      | 1.1    | 1.0      |
| M13 | 2.2   | 1.1      | 1.8   | 1.1      | 1.6   | 1.1      | 1.6    | 1.1      |
| M14 | 1.9   | 1.7      | 1.3   | 1.7      | 1.1   | 1.7      | 1.1    | 1.7      |
| M15 | 1.9   | 1.2      | 1.3   | 1.2      | 1.1   | 1.2      | 1.1    | 1.2      |
| M16 | 1.2   | 0.9      | 0.8   | 1.8      | 0.3   | 4.2      | 0.3    | 12       |

Table 2. Ratio to the oracle of the  $L_2$  risk for the first step estimator in the regressive case for gaussian errors

Let  $S_{r,D}$  be the space of piecewise polynomials of degree  $r$  on the partition  $[(d-1)/D, d/D[, d = 1, \dots, D$ . Let  $X, Y, sX$  be vectors of  $\mathbb{R}^n$  with coordinates  $X_i, Y_i, sX_i$ , where  $sX_i$  will be defined later, and let  $t$  be a function in some  $S_{r,D}$ . We define here the contrast and the penalty function as

$$g_n(X, Y, sX; t) = \frac{1}{n} \sum_{i=1}^n \left( \frac{Y_i - t(X_i)}{sX_i} \right)^2 \quad \text{and} \quad \text{pen}(D) = D + \ln^2(D).$$

Then we consider the following general procedures:  $\mathcal{E}_r(X, Y, sX, fX)$  for  $r = 0, \dots, 5$  and  $\mathcal{E}(X, Y, sX, fX)$ , with input the  $\mathbb{R}^n$  vectors  $X, Y, sX$  previously described and  $fX$  with coordinates  $f(X_i)$  for some given function  $f$ . The procedure  $\mathcal{E}_r(X, Y, sX, fX)$  proceeds as follows.

- For  $D = 1, \dots, D_{max} = \lceil n/((r+1)\ln(n)) \rceil$ , compute  $\hat{f}_{r,D}$  (in fact  $\hat{f}_{r,D}(X_i), i = 1, \dots, n$ ) the piecewise polynomial of  $S_{r,D}$  minimizing  $g_n(X, Y, sX; t)$  over all  $t$  in  $S_{r,D}$ .
- Compute  $\hat{D}_r = \arg \min_{1 \leq D \leq D_{max}} [g_n(X, Y, sX; \hat{f}_{r,D}) + 2\hat{\sigma}_r^2 \text{pen}(D(r+1))]$  where

$$\hat{\sigma}_r^2 = \begin{cases} g_n(X, Y, sX, \hat{f}_{r, [\min(\sqrt{n}, n/((r+1)\ln(n))])}) & \text{if } sX_i = 1, \quad \forall i = 1, \dots, n \\ 1 & \text{else} \end{cases}.$$

- Keep  $\hat{\sigma}_r^2, (\hat{f}_{r, \hat{D}_r}(X_1), \dots, \hat{f}_{r, \hat{D}_r}(X_n))$  and  $\|f - \hat{f}_{r, \hat{D}_r}\|_n^2 = (1/n) \sum_{i=1}^n (f(X_i) - \hat{f}_{r, \hat{D}_r}(X_i))^2$ .

The procedure  $\mathcal{E}(X, Y, sX, fX)$  follows then and selects

$$\hat{r} = \arg \min_{0 \leq r \leq 5} \left\{ g_n(X, Y, sX, \hat{f}_{r, \hat{D}_r}) + 2\hat{\sigma}_r^2 \text{pen}((r+1)\hat{D}_r) \right\}.$$

The output is therefore  $\tilde{f} = \hat{f}_{\hat{r}, \hat{D}_{\hat{r}}}$  and the associated error  $\|f - \tilde{f}\|_n^2$ .

|     | s2n=1 |          | s2n=3 |          | s2n=7 |          | s2n=10 |          |
|-----|-------|----------|-------|----------|-------|----------|--------|----------|
|     | $b$   | $\sigma$ | $b$   | $\sigma$ | $b$   | $\sigma$ | $b$    | $\sigma$ |
| M1  | 1.4   | 0.8      | 1.4   | 0.8      | 1.4   | 0.8      | 1.4    | 0.8      |
| M2  | 2.2   | 1.2      | 2.1   | 0.8      | 2.1   | 0.8      | 2.1    | 0.8      |
| M3  | 1.2   | 0.8      | 1.5   | 1.0      | 1.4   | 0.9      | 1.2    | 0.8      |
| M4  | 1.1   | 1.3      | 1.2   | 1.2      | 1.7   | 126      | 0.6    | 0.7      |
| M5  | 1.3   | 0.9      | 1.4   | 1.2      | 1.5   | 1.1      | 1.4    | 0.8      |
| M6  | 1.3   | 1.3      | 1.2   | 1.0      | 1.2   | 1.1      | 1.1    | 1.2      |
| M7  | 1.6   | 2.6      | 1.4   | 2.9      | 1.4   | 1.1      | 1.3    | 1.4      |
| M8  | 1.1   | 4.1      | 1.2   | 2.3      | 1.3   | 2.3      | 1.1    | 1.6      |
| M9  | 1.5   | 1.6      | 1.1   | 1.7      | 1.3   | 1.7      | 1.5    | 1.6      |
| M10 | 1.5   | 1.6      | 1.6   | 1.3      | 1.5   | 1.3      | 1.2    | 2.2      |
| M11 | 1.5   | 2.2      | 1.6   | 1.7      | 1.4   | 1.2      | 1.4    | 1.2      |

Table 3. Ratio to the oracle of the  $L_2$  risk for the first step estimator in the autoregressive case for gaussian errors.

It follows that, as an output of the procedure  $\mathcal{E}_r(X^{(s)}, Y^{(s)}, \mathbf{1}, bX^{(s)})$ , where  $(X^{(s)}, Y^{(s)})$  is the  $s$ th sample drawn from a given regressive model, we obtain  $\tilde{b}_r^{(s)}$ , and  $\tilde{b}^{(s)}$  as an output of  $\mathcal{E}(X^{(s)}, Y^{(s)}, \mathbf{1}, bX^{(s)})$ . We compute  $\|b - \tilde{b}^{(s)}\|_n^2$  for each sample. This allows to compute the mean  $\mathbb{L}^2$ -empirical error:

$$L_{emp}^2(b, \tilde{b}) = \mathbb{E}^{(S)} \left[ \|b - \tilde{b}_r\|_n^2 \right] = \frac{1}{S} \sum_{s=1}^S \left( \frac{1}{n} \sum_{i=1}^n [b(X_i^{(s)}) - \tilde{b}_r^{(s)}(X_i^{(s)})]^2 \right).$$

Therefore, we are interested in the ratios  $L_{emp}^2(b, \tilde{b})/L_{oracle}^2(b)$ . They are generally greater than one. The nearer of one, the better our method.

We also computed the output of  $\mathcal{E}(X^{(s)}, Y^{(s)}, \sigma X^{(s)}, bX^{(s)})$ , where  $\sigma X^{(s)}$  has coordinates  $\sigma(X_i^{(s)})$  for  $i = 1, \dots, n$ , which delivers an estimator denoted by  $\tilde{b}_\sigma^{(s)}$  of  $b$  if  $\sigma$  were known. We compared the associated ratio  $L_{emp}^2(b, \tilde{b}_\sigma)/L_{oracle}^2(b)$  to the previous one.

Analogously,  $\mathcal{E}_r(X^{(s)}, [Y^{(s)} - \tilde{b}^{(s)}(X^{(s)})]^2, \mathbf{1}, \sigma^2 X^{(s)})$  where  $[Y^{(s)} - \tilde{b}^{(s)}(X^{(s)})]^2$  has coordinates  $[Y_i^{(s)} - \tilde{b}^{(s)}(X_i^{(s)})]^2$  for  $i = 1, \dots, n$ , gives estimators  $\widetilde{\sigma^2_r}^{(s)}$ , and  $\mathcal{E}(X^{(s)}, [Y^{(s)} - \tilde{b}^{(s)}(X^{(s)})]^2, \mathbf{1}, \sigma^2 X^{(s)})$  gives  $\widetilde{\sigma^2}^{(s)}$ . When  $s$  is varying we compute the error:

$$L_{emp}^2(\sigma^2, \widetilde{\sigma^2}^{(s)}) = \frac{1}{S} \sum_{s=1}^S \left\{ \frac{1}{n} \sum_{i=1}^n \left[ \sigma^2(X_i^{(s)}) - \widetilde{\sigma^2}^{(s)}(X_i^{(s)}) \right]^2 \right\}.$$

It can be compared with the estimate, denoted by  $\widetilde{\sigma^2_b}^{(s)}$  of  $\sigma^2$  if  $b$  were known by using  $\mathcal{E}(X^{(s)}, [Y^{(s)} - b(X^{(s)})]^2, \mathbf{1}, \sigma^2 X^{(s)})$ .

Moreover, we studied a second stage of the procedure by computing  $\tilde{b}^{(s)}$  as the output of  $\mathcal{E}_r(X^{(s)}, Y^{(s)}, \widetilde{\sigma^2}^{(s)}, bX^{(s)})$  and  $\widetilde{\sigma^2}^{(s)}$  as the output of  $\mathcal{E}(X^{(s)}, [Y^{(s)} - \tilde{b}^{(s)}(X^{(s)})]^2, \mathbf{1}, \sigma^2 X^{(s)})$ . But this procedure happened to be very unstable in spite of several attempts to stabilize it.

We need to make two remarks about our procedure. First, when we have to divide by some estimate of the variance, when computing  $\tilde{b}^{(s)}$  for instance, we divide in fact by the supremum of the value of interest and the 2%-quantile of all the positive estimated values. Secondly, there may be some restrictions on the values of the degrees when too few observations lie in one bin. In the regressive case, when working with global degree  $r$ , we take in fact locally on the subinterval  $[d - 1/D, d/D]$ , the degree

$$\min \left( r, \left| \left\{ X_i \in \left[ \frac{d-1}{D}, \frac{d}{D} \right] \right\} \right| - 1 \right).$$

In the autoregressive case, we take  $\min(r, \hat{R}_d - 1)$  where  $\hat{R}_d = \text{rank}(V(d, D))$  and  $V(d, D) = (X_{i_p}^{q-1} / \tilde{\sigma}(X_{i_p}))_{1 \leq p \leq k, 1 \leq q \leq r+1}$ , for  $i_1, \dots, i_k$  the indexes of the  $X_i$ 's in  $[(d-1)/D, d/D]$ . This is required for the inversion of the local linear system associated to the local computation of the estimator.

Our results are gathered in the Tables 2 and 3 in the case of gaussian errors. Table 4 and 5 give the results for uniform errors and since they are of the same type, they are deferred to the appendix in section 6. All tables give the error ratios  $L_{emp}^2/L_{oracle}^2$  for  $b$  and  $\sigma$  as in the models given in Table 1 and for different values of the signal-to-noise ratio  $s2n$ . We can give several comments about these tables and other unreported results.

1. We can see that most ratios are near of 1, and almost all are less than 2, which means that our estimates perform very well.
2. We give the results of the first step estimator for  $b$  and  $\sigma$  because the second step is often unstable.
3. The results for  $b$  are most of the time better as those obtained by working with known  $\sigma$  and the knowledge of  $b$  does not improve significantly the estimation of  $\sigma$ .
4. We must also emphasize that the last step of the procedure which chooses the degree performs quite well and gives empirical errors of the same order as the error associated to the degree implying the lowest error.

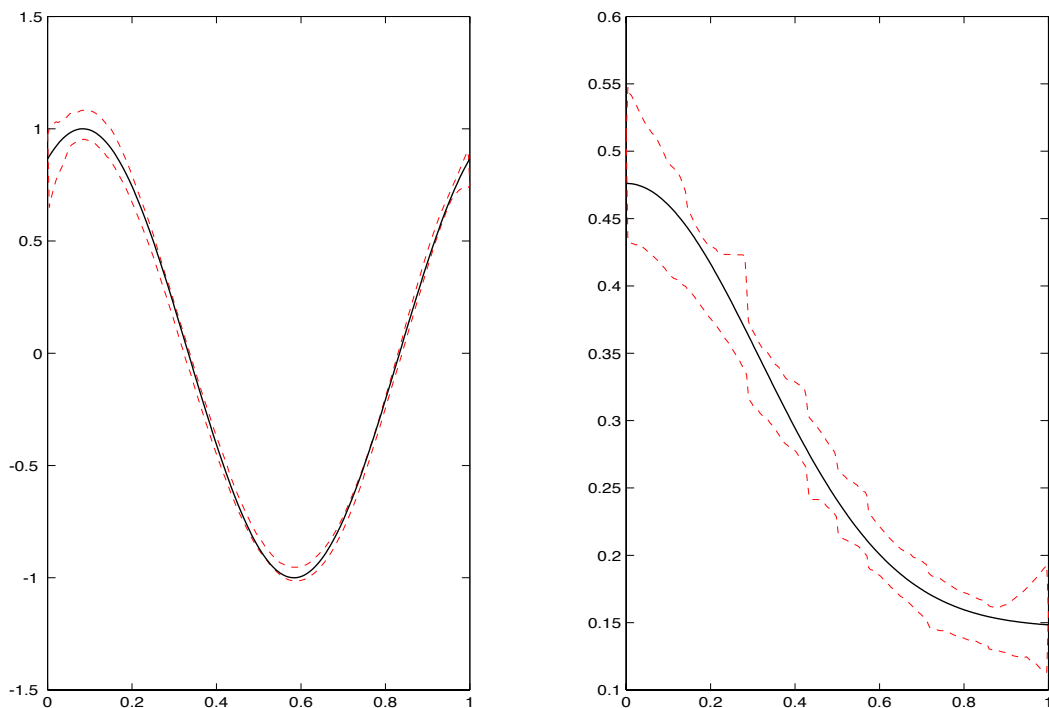


Figure 1: True functions (thick curves)  $b$  and  $\sigma^2$  in the autoregressive model M9 with value of the parameters corresponding to  $s2n=3$  and gaussian errors. Tenth and nineteenth percentiles (dotted curves) for  $S = 100$  samples with length  $n = 500$  of the estimation of  $b$  (right curves) and  $\sigma^2$  (left curves) given by the algorithm.

5. When the  $s2n$  ratio become higher, the results do not improve significantly because the oracle decreases considerably in the same time.

In order to give a visual illustration of the results, we give confidence intervals (tenth and nineteenth percentiles) for curve estimation of  $b$  and  $\sigma^2$  in 3 cases: Model M9 in the gaussian autoregressive case for  $s2n=3$  (Figure 1), Model M10 for  $s2n=7$  in the gaussian regressive (Figure 2) and autoregressive (Figure 3) cases. We generated here  $S = 100$  samples with length  $n = 500$ . It appears clearly that the estimation of the mean function  $b$  is always very good, whereas the estimation of  $\sigma^2$  is generally better in the regressive context than in the autoregressive one.

## 5 Proofs

### 5.1 Proof of Theorem 3.1

For the sake of simplicity, we omit the superscript (1) for the spaces and the dimensions and write  $S_m$  for  $S_m^{(1)}$ ,  $D_m$  for  $D_m^{(1)}$ ,  $\mathcal{M}_n$  for  $\mathcal{M}_n^{(1)}$ . There is no ambiguity all along this proof.

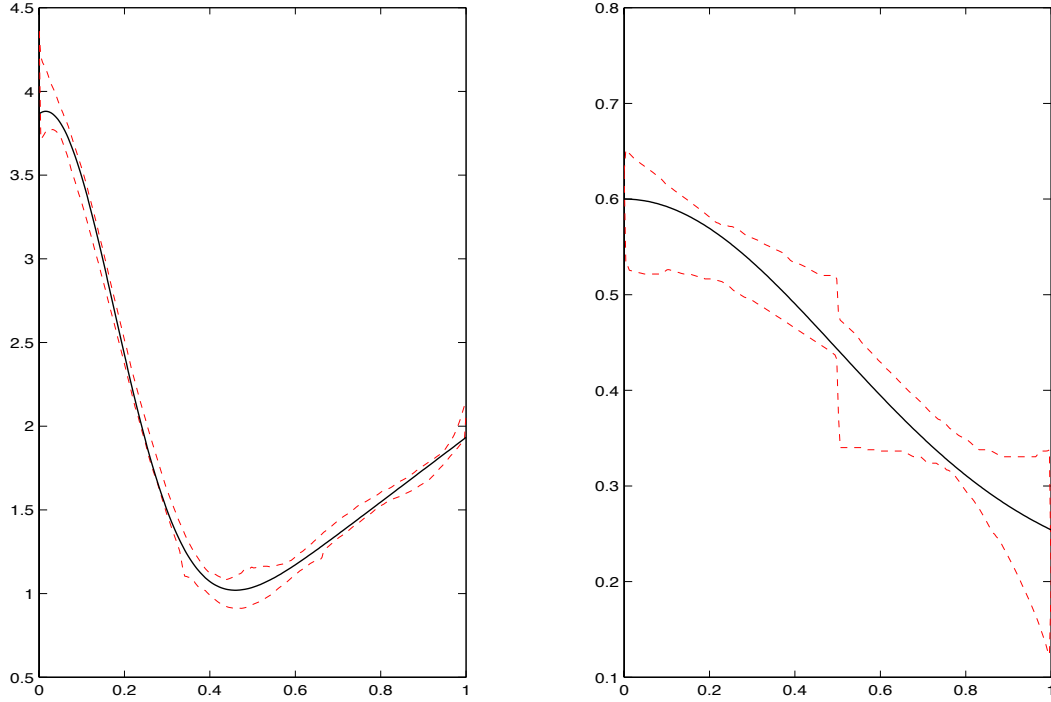


Figure 2: True functions (thick curves)  $b$  and  $\sigma^2$  in the regressive model M10 with value of the parameters corresponding to  $s2n=7$  and gaussian errors. Tenth and nineteenth percentiles (dotted curves) for  $S = 100$  samples with length  $n = 500$  of the estimation of  $b$  (right curves) and  $\sigma^2$  (left curves) given by the algorithm.

We follow the line of the proof of Theorem 1 in Baraud et al. (1999) and we use the same notations. We only recall that  $\Omega^*$  is the event

$$\Omega^* = \{(\varepsilon_{i+1}, X_i) = (\varepsilon_{i+1}^*, X_i^*), i = 1, \dots, n\}$$

where the variables  $(\varepsilon_{i+1}^*, X_i^*)$  are associated to the  $(\varepsilon_{i+1}, X_i)$  as in Claim 2 of Baraud et al. (1999) recalled below:

**Claim 2:** Let  $q_n, q_{n,1}$  be integers such that  $0 \leq q_{n,1} \leq q_n/2, q_n \geq 1$ . Set  $u_i = (\varepsilon_i, X_i)$ ,  $i = 1, \dots, n$ , then there exist random variables  $u_i^* = (\varepsilon_i^*, X_i^*)$ ,  $i = 1, \dots, n$  satisfying the following properties:

- For  $\ell = 1, \dots, \ell_n = \lfloor n/q_n \rfloor$ , the random vectors

$$\vec{U}_{\ell,1} = (u_{(\ell-1)q_n+1}, \dots, u_{(\ell-1)q_n+q_{n,1}})' \text{ and } \vec{U}_{\ell,1}^* = (u_{(\ell-1)q_n+1}^*, \dots, u_{(\ell-1)q_n+q_{n,1}}^*)'$$

have the same distribution, and so have the random vectors

$$\vec{U}_{\ell,2} = (u_{(\ell-1)q_n+q_{n,1}+1}, \dots, u_{\ell q_n})' \text{ and } \vec{U}_{\ell,2}^* = (u_{(\ell-1)q_n+q_{n,1}+1}^*, \dots, u_{\ell q_n}^*)'.$$

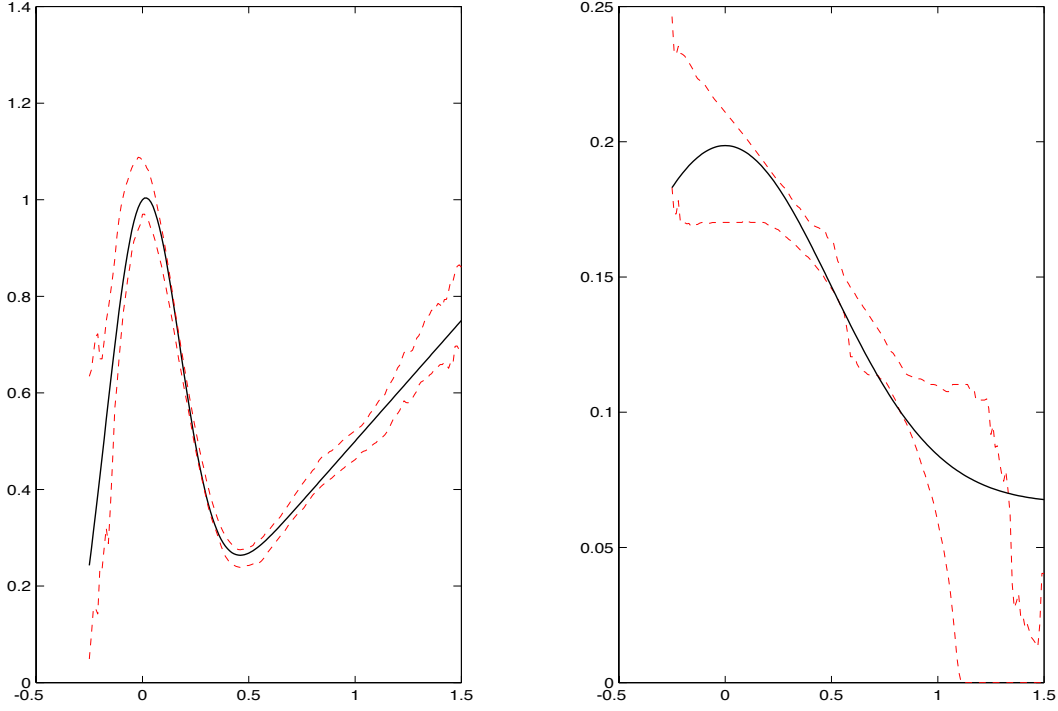


Figure 3: True functions (thick curves)  $b$  and  $\sigma^2$  in the autoregressive model M10 with value of the parameters corresponding to  $s2n=7$  and gaussian errors. Tenth and nineteenth percentiles (dotted curves) for  $S = 100$  samples with length  $n = 500$  of the estimation of  $b$  (right curves) and  $\sigma^2$  (left curves) given by the algorithm.

- For  $\ell = 1, \dots, \ell_n$ ,

$$\mathbb{P} \left[ \vec{U}_{\ell,1} \neq \vec{U}_{\ell,1}^* \right] \leq \beta_{(q_n - q_{n,1})} \text{ and } \mathbb{P} \left[ \vec{U}_{\ell,2} \neq \vec{U}_{\ell,2}^* \right] \leq \beta_{q_{n,1}}. \quad (5.29)$$

- For each  $\delta \in \{1, 2\}$ , the random vectors  $\vec{U}_{1,\delta}^*, \dots, \vec{U}_{\ell_n,\delta}^*$  are independent.

The variables  $u_i^*$  are built using Berbee's coupling Lemma as in Viennet (1997). For sake of simplicity, we assume that  $n = q_n \ell_n$ . For  $\rho \geq 1$  we also recall that  $\Omega_\rho$  is the event:

$$\Omega_\rho = \left\{ \|t\|_\mu^2 \leq \rho \|t\|_n^2, \forall t \in \bigcup_{m,m' \in \mathcal{M}_n} S_m + S_{m'} \right\},$$

that is  $\Omega_\rho$  is the event where the norms  $\|\cdot\|$  and  $\|\cdot\|_n$  can be compared. Lastly, we add, for some  $\tau \in ]0, 1[$ , the definition of the following event:

$$\Omega_\tau = \left\{ (1 - \tau) \|\sigma\|_\mu^2 \leq \hat{r}_n^2 \leq 2(1 + \tau) (\|b\|_\mu^2 + \|\sigma\|_\mu^2) \right\}, \quad (5.30)$$

where  $\hat{r}_n$  is defined by (2.13). We denote by  $\Omega_{\tau,\rho}^* := \Omega_\tau \cap \Omega_\rho \cap \Omega^*$ , by  $B(m', \mu) = \{t \in S_m + S_{m'}, \|t\|_\mu \leq 1\}$ , and by  $D(m') = \dim(S_m + S_{m'})$ . Since  $m$  is fixed, we do not mention

the dependence on  $m$  of the previous terms. Then we write the decomposition

$$\gamma_n(t) - \gamma_n(s) = \|b_A - t\|_n^2 - \|b_A - s\|_n^2 + 2\langle s - t, \sigma\varepsilon \rangle_n$$

where  $\langle t, \sigma\varepsilon \rangle_n = (1/n) \sum_{i=1}^n t(X_i) \sigma(X_i) \varepsilon_{i+1}$ . Moreover the definition of  $\tilde{b}$  implies that  $\forall m \in \mathcal{M}_n$

$$\gamma_n(\tilde{b}) + \text{pen}^{(1)}(\hat{m}) \leq \gamma_n(b_m) + \text{pen}^{(1)}(m)$$

with  $\text{pen}^{(1)}$  defined by (2.15). Therefore, using that  $2ab \leq xa^2 + x^{-1}b^2$  and  $(a+b)^2 \leq (1+y)a^2 + (1+y^{-1})b^2$  for all positive  $a, b, x$  and  $y$ , we find, analogously to (38) in Baraud et al. (1999), Claim 3, that on  $\Omega_{\tau, \rho}^*$

$$\begin{aligned} & \left(1 - \rho \frac{1+y}{x}\right) \|b\mathbf{I}_A - \tilde{b}\|_n^2 \\ & \leq \left(1 + \rho \frac{1+y^{-1}}{x}\right) \|b\mathbf{I}_A - b_m\|_n^2 + \text{pen}^{(1)}(m) \\ & \quad + \frac{x}{n^2} \left( \sup_{t \in B(\hat{m}, \mu)} \sum_{i=1}^n \varepsilon_{i+1}^* \sigma(X_i^*) t(X_i^*) \right)^2 - \text{pen}^{(1)}(\hat{m}) \\ & \leq \left(1 + \rho \frac{1+y^{-1}}{x}\right) \|b\mathbf{I}_A - b_m\|_n^2 + 2(1+\tau)\kappa(\|b\|_\mu^2 + \|\sigma\|_\mu^2) \frac{D_m}{n} \Phi_1^2 \\ & \quad + \frac{x}{n^2} \left[ \left( \sup_{t \in B(\hat{m}, \mu)} \sum_{i=1}^n \varepsilon_{i+1}^* \sigma(X_i^*) t(X_i^*) \right)^2 - x^2 n D(\hat{m}) \Phi_1^2 \|\sigma\|_\mu^2 \right]_+ \\ & \quad + \frac{x^3 D(\hat{m})}{n} \Phi_1^2 \|\sigma\|_\mu^2 - \frac{\kappa(1-\tau)\|\sigma\|_\mu^2 D_{\hat{m}}}{n} \Phi_1^2 \end{aligned}$$

The real numbers  $x$  and  $y$  are positive constant numbers to be chosen. Then setting

$$\tilde{W}_n(m') = \left[ \left( \sup_{t \in B(m', \mu)} \sum_{i=1}^n \varepsilon_{i+1}^* \sigma(X_i^*) t(X_i^*) \right)^2 - x^2 n D(m') \Phi_1^2 \|\sigma\|_\mu^2 \right]_+$$

we find that if  $x, \rho$  are numbers satisfying  $x > \rho > 1$ ,  $y = (x - \rho)/(x + \rho)$ ,  $\tau > 0$ , and if  $\kappa = x^3/(1 - \tau)$  in  $\text{pen}^{(1)}$ , then:

$$\begin{aligned} \|b\mathbf{I}_A - \tilde{b}\|_n^2 \mathbf{I}_{\Omega_{\tau, \rho}^*} & \leq K_1(x, \rho) \left[ \|b\mathbf{I}_A - b_m\|_n^2 + 4 \frac{(\|b\|_\mu^2 + \|\sigma\|_\mu^2) D_m}{1 - \tau} \frac{\Phi_1^2}{n} \right] \\ & \quad + \frac{x}{n^2} \tilde{W}_n(\hat{m}). \end{aligned} \tag{5.31}$$

with  $K_1(x, \rho) = (x + \rho)^2/(x - \rho)^2$ .

In Baraud et al. (1999),  $\tilde{W}_n(m')$  for  $\sigma \equiv \sigma_2$  where  $\sigma_2$  is a constant is denoted by  $W_n(m')$  and Proposition 6 in Baraud et al. (1999) states that, under assumptions analogous to the assumptions of Theorem 3.1, for any  $\bar{p} < p/2$ ,

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}(W_n(m')^{\bar{p}}) \leq K_2 n^{\bar{p}} \left[ \sum_{m' \in \mathcal{M}_n} D_{m'}^{-p/2 + \bar{p}} + \frac{q_n^p |\mathcal{M}_n|}{n^{p(p-2)/[4(p-1)] - \bar{p}}} \right] \tag{5.32}$$

where  $K_2 = C(p, \bar{p}, x)(\Phi_1 h_0^{-1/2})^p \sigma_p^{2\bar{p}}$ ,  $\sigma_p^p = \mathbb{E}(|\varepsilon_1|^p)$  and  $C(p, \bar{p}, x)$  is a constant depending on  $p$ ,  $\bar{p}$  and  $x$ .

This result has a straightforward extension to non constant variance function with only a  $\|\sigma_A\|_\infty$  replacing  $\sigma_2$  and  $(\|\sigma_A\|_\infty \sigma_p)^{2\bar{p}}$  replacing  $\sigma_p^{2\bar{p}}$ . The only point to check is indeed the bound for the analogous of  $\mathbb{E} \left[ \sup \sum_{\ell=1}^{\ell_n} G_\ell(t) \right]$  with now

$$\tilde{G}_\ell(t) = \sum_{i \in \mathcal{I}_\ell^{(1)}} \varepsilon_{i+1}^* \sigma_A(X_i^*) t(X_i^*).$$

We still find that for  $t = \sum_{j=1}^{D(m')} a_j \varphi_j$  where  $(\varphi_j)_{1 \leq j \leq D(m')}$  is a  $\mu$ -orthonormal basis of  $S_m + S_{m'}$ ,

$$\mathbb{E} \left[ \sup_{t \in B(m', \mu)} \sum_{\ell=1}^{\ell_n} \tilde{G}_\ell(t) \right] \leq \left[ \sum_{j=1}^{D(m')} \mathbb{E} \left( \sum_{\ell=1}^{\ell_n} \tilde{G}_\ell(\varphi_j) \right)^2 \right]^{1/2} = \left[ \sum_{j=1}^{D(m')} \sum_{\ell=1}^{\ell_n} \mathbb{E} \left( \tilde{G}_\ell^2(\varphi_j) \right) \right]^{1/2}$$

since the blocks are independent and centered. The difference is that here

$$\mathbb{E}(\tilde{G}_\ell^2(\varphi_j)) = \sum_{i \in \mathcal{I}_\ell^{(1)}} \mathbb{E}(\varepsilon_{i+1}^2) \mathbb{E}(\sigma_A^2(X_i) \varphi_j^2(X_i)) = q_{n,1} \|\sigma_A \varphi_j\|_\mu^2.$$

Then using consequence (2.10) of assumption **(H<sub>Φ<sub>1</sub></sub>) 2.**, we have  $\|\sum_j \varphi_j^2\|_\infty \leq \Phi_1^2 D(m')$ , and therefore

$$\mathbb{E} \left[ \sup_{t \in B(m', \mu)} \sum_{\ell=1}^{\ell_n} \tilde{G}_\ell(t) \right] \leq \Phi_1 \|\sigma\|_\mu \sqrt{\ell_n q_{n,1} D(m')}.$$

This gives the announced extension of Proposition 6 of Baraud et al. (1999), namely, for  $\bar{p} = 1$ :

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E} \left[ \frac{\tilde{W}_n(m')}{n^2} \right] \leq K_3 n^{-1} \left[ \sum_{m' \in \mathcal{M}_n} D_{m'}^{-p/2+1} + \frac{q_n^p |\mathcal{M}_n|}{n^{p(p-2)/[4(p-1)]-1}} \right], \quad (5.33)$$

where  $K_3 = C(x, p) \|\sigma_A\|_\infty^2 \sigma_p^2 (\Phi_1^2 h_0^{-1})^{p/2}$ . Thus, in view of **(H<sub>(a<sub>1</sub>, b<sub>1</sub>)</sub>)**, the last bracketed term in (5.33) is uniformly bounded if

$$-p/2 + 1 \leq -a_1 \text{ and } b_1 + 1 - \{p(p-2)/[4(p-1)]\} < 0. \quad (5.34)$$

Since  $p(p-2)/[4(p-1)] = (p-2)/4 + (p-2)/[4(p-1)] \geq (p-2)/4$ , (5.34) is fulfilled under (3.20).

Since  $\mathbb{E} \left[ \tilde{W}_n(\hat{m})/n^2 \right] \leq \sum_{m' \in \mathcal{M}_n} \mathbb{E} \left[ \tilde{W}_n(m')/n^2 \right]$ , (5.31), (5.33) and (3.20) imply that

$$\begin{aligned} \mathbb{E}(\|b \mathbf{I}_A - \tilde{b}\|_{\Omega_{\tau, \rho}^*}^2) &\leq K_1(x, \rho) \left[ \|b \mathbf{I}_A - b_m\|_n^2 + 4 \frac{(\|b\|_\mu^2 + \|\sigma\|_\mu^2) D_m}{1 - \tau} \frac{\Phi_1^2}{n} \right] \\ &\quad + \frac{K_4}{n}. \end{aligned} \quad (5.35)$$

where  $K_4 = K_4(x, \rho, \Phi_1, h_0, \Sigma_1, T_1)$ . It remains to bound the expectation on the complementary of  $\Omega_{\tau, \rho}^*$ . Let  $C$  denote a constant that may change from line to line. It follows

from Claim 5 in Baraud et al. (1999) that  $\mathbb{P}((\Omega_\rho \cap \Omega^*)^c) \leq C/n^2$  under geometrical mixing condition and  $(\mathbf{H}_{\Phi_1})$  3. and that

$$\mathbb{E} \left[ \|b\mathbf{I}_A - \tilde{b}\|_n^2 \mathbf{I}_{\Omega_{\tau,\rho}^{*c}} \right] \leq C(h_0, h_1, \Phi_1, \rho)n^{-1}$$

as soon as

$$\mathbb{P}((\Omega_{\tau,\rho}^*)^c) \leq \frac{C}{n^2}. \quad (5.36)$$

Since  $\mathbb{P}((\Omega_{\tau,\rho}^*)^c) \leq \mathbb{P}(\Omega_\tau^c) + \mathbb{P}((\Omega_\rho^*)^c)$  with obvious notations, and since

$$\mathbb{P}(\Omega_\tau^c) = \mathbb{P}(\Omega_\tau^c \cap \Omega_\rho \cap \Omega^*) + \mathbb{P}((\Omega_\rho \cap \Omega^*)^c),$$

we find that (5.36) holds if we have

$$\mathbb{P}(\Omega_\tau^c \cap \Omega_\rho \cap \Omega^*) \leq C/n^2.$$

This is ensured by the following Lemma:

**Lemma 5.1** *Under the Assumptions of Theorem 3.1 and if*

$$D_{m_n}^{(1)} \leq n^{1/2-k/p} \text{ for } k = 2 \text{ or } k = 4, \quad (5.37)$$

*and  $p \geq 8$  if  $k = 2$  and  $p \geq 16$  if  $k = 4$  (so that  $D_{m_n}^{(1)}$  can always be taken of order  $n^{1/4}$ ), then*

$$\mathbb{P}(\Omega_\tau^c \cap \Omega_\rho \cap \Omega^*) \leq Cn^{-k}$$

*where  $C$  is a constant depending in particular on  $\Phi_1, p, \rho, \|\sigma\|_\mu, \|\sigma_A\|_\infty, \|b_A\|_\infty$ .*

Recall that  $p$  denotes the order of the moment of the  $\varepsilon_i$ 's in model (1.1) and that  $D_{m_n}^{(1)}$  is defined by (2.13) and the line following. This ends the proof of Theorem 3.1.  $\square$ .

The limit choices  $\tau \rightarrow 0$  and  $x \rightarrow 0$  give  $\kappa \rightarrow 1$  but imply that the multiplicative constant tends to infinity. The choice  $\tau = 1/2, x = 2$  and  $\rho = 1$  gives  $\kappa = 4$  and reasonable orders for the multiplicative constants. The value of  $\kappa$  must be investigated by simulation experiments.

*Proof of Lemma 5.1.* Most elements of this proof (in the case of a simpler model) can be found in a first draft of Baraud et al. (1999) but it disappeared of the final version.

For the sake of simplicity, we work on a space  $S_m$  with dimension  $D_m$  (instead of  $D_{m_n}^{(1)}$ ). We shall denote by  ${}^tX$  the transpose of a vector  $X$ . Let  $R = {}^t(X_2, \dots, X_{n+1})$  and let  $\varepsilon = {}^t(\varepsilon_2, \dots, \varepsilon_{n+1})$ . All along this section we abusively denote the same way a function  $g$  mapping  $\mathbb{R}$  into  $\mathbb{R}$  and the vector of  $\mathbb{R}^n$   ${}^t(g(X_1), \dots, g(X_n))$ .  $\mathbb{R}^n$  is provided with the inner product  $\langle u, v \rangle = \sum_{i=1}^n u_i v_i$ , we denote the corresponding norm by  $\|\cdot\|$  and by  $\|\cdot\|_n$  the empirical norm:  $\|u\|_n^2 = (1/n) \sum_{i=1}^n u_i^2$ . From now on  $\{\varphi_\lambda, \lambda \in \Lambda_m\}$  denotes an orthonormal basis of  $S_m$  relatively to  $\mu$  and  $\Phi_m(X)$  is the  $D_m \times D_m$  normalized Gram matrix defined by

$$\Phi_m(X) = \left( \frac{1}{n} \sum_{i=1}^n \varphi_\lambda(X_i) \varphi_{\lambda'}(X_i) \right)_{\lambda, \lambda' \in \Lambda_m}.$$

It follows from the definition of  $\hat{b}_m$  that

$$\hat{b}_m = V_m [n\Phi_m(X)]^{-1} {}^tV_m R = \frac{1}{n} V_m \Phi_m^{-1}(X) {}^tV_m R$$

where  $V_m$  denotes the  $n \times D_m$  matrix satisfying  $(V_m)_{(i,\lambda)} = \varphi_\lambda(X_i)$  for  $i = 1, \dots, n$  and  $\lambda \in \Lambda_m$ . We denote by  $\Pi_m(X)$  the projection matrix  $V_m \Phi_m^{-1}(X) {}^t V_m$ . Note that

$$\Pi_m(X) {}^t \Pi_m(X) = V_m \Phi_m^{-1}(X) ({}^t V_m V_m) \Phi_m^{-1}(X) {}^t V_m = n V_m \Phi_m^{-1}(X) {}^t V_m = n \Pi_m(X).$$

Since  $\hat{r}_n^2 = \|R - \hat{b}_m\|_n^2$  for  $R = b + \sigma\varepsilon$ , we have

$$\begin{aligned} \hat{r}_n^2 &= \|R - \Pi_m(X)R\|_n^2 \\ &= \|b - \Pi_m(X)b + \sigma\varepsilon - \Pi_m(X)\sigma\varepsilon\|_n^2 \\ &= \frac{1}{n} [\|b - \Pi_m(X)b\|^2 + \|\sigma\varepsilon\|^2 - \|\Pi_m(X)\sigma\varepsilon\|^2 + 2\langle b - \Pi_m(X)b, \sigma\varepsilon \rangle]. \end{aligned}$$

We define the measure  $\mathbb{P}^{*,\rho}$  by  $\mathbb{P}^{*,\rho}(B) = \mathbb{P}(B \cap \Omega^* \cap \Omega_\rho)$ , and we take  $\tau = 4\eta$ .

$$\begin{aligned} \mathbb{P}^{*,\rho}(\hat{r}_n^2 \leq (1 - 4\eta)\|\sigma\|_\mu^2) &\leq \mathbb{P}^{*,\rho}(\|\sigma\varepsilon\|_n^2 \leq (1 - \eta)\|\sigma\|_\mu^2) \\ &\quad + \mathbb{P}(\|\Pi_m(X)\sigma_A\varepsilon\|_n^2 \geq \eta\|\sigma\|_\mu^2) \\ &\quad + \mathbb{P}^{*,\rho}(2|\langle b - \Pi_m(X)b, \sigma\varepsilon \rangle| \geq 2n\eta\|\sigma\|_\mu^2). \end{aligned}$$

We denote by  $\sigma_A = \sigma \mathbf{I}_A$ . Note that  $\sigma$  can be replaced by  $\sigma_A$  each time it is multiplied by  $A$ -supported functions. The same holds for  $b$  and  $b_A = b \mathbf{I}_A$ .

Let us bound first  $\mathbb{P}^{*,\rho}(\|\Pi_m(X)\sigma_A\varepsilon\|_n^2 \geq 2\eta\|\sigma\|_\mu^2)$ .

$$\begin{aligned} \|\Pi_m(X)\sigma_A\varepsilon\|_n^2 &= \frac{1}{n^2} {}^t(\sigma_A\varepsilon)\Pi_m(X)\Pi_m(X)(\sigma_A\varepsilon) = \frac{1}{n} {}^t({}^t V_m \sigma_A \varepsilon) \Phi_m^{-1}(X) ({}^t V_m \sigma_A \varepsilon) \\ &= \frac{1}{n} \langle {}^t V_m \sigma_A \varepsilon, \Phi_m^{-1}(X) ({}^t V_m \sigma_A \varepsilon) \rangle \\ &\leq \frac{\rho(\Phi_m^{-1}(X))}{n} \|{}^t V_m \sigma_A \varepsilon\|^2, \end{aligned}$$

where  $\rho(M)$  denotes the spectral radius of the matrix  $M$ . We know from Baraud (2000), Lemma 3.1 p. 475, that

$$\rho(\Phi_m^{-1}(X)) = \sup_{t \in S_m / \{0\}} \frac{\|t\|_\mu^2}{\|t\|_n^2}.$$

Therefore on  $\Omega_\rho$ , we have  $\rho(\Phi_m^{-1}(X)) \leq \rho$ . This implies that on  $\Omega^* \cap \Omega_\rho$ ,

$$\|\Pi_m(X)\sigma_A\varepsilon\|_n^2 \leq \frac{\rho}{n^2} \|{}^t V_m \sigma_A \varepsilon\|^2 = \rho \sum_{\lambda \in \Lambda_m} \left( \frac{1}{n} \sum_{i=1}^n \varepsilon_{i+1} \varphi_\lambda(X_i) \sigma_A(X_i) \right)^2.$$

Therefore

$$\begin{aligned} &\mathbb{P}^{*,\rho}(\|\Pi_m(X)\sigma\varepsilon\|_n^2 \geq \eta\|\sigma\|_\mu^2) \\ &\leq \mathbb{P}^{*,\rho} \left( \sum_{\lambda \in \Lambda_m} \left( \frac{1}{n} \sum_{i=1}^n \varepsilon_{i+1} \varphi_\lambda(X_i) \sigma_A(X_i) \right)^2 \geq \frac{\eta}{\rho} \|\sigma\|_\mu^2 \right) \\ &\leq \left( \frac{\rho}{\eta\|\sigma\|_\mu^2} \right)^{\frac{p}{2}} \mathbb{E} \left( \sum_{\lambda \in \Lambda_m} \left( \frac{1}{n} \sum_{i=1}^n \varepsilon_{i+1}^* \varphi_\lambda(X_i^*) \sigma_A(X_i^*) \right)^2 \right)^{\frac{p}{2}} \\ &\leq \left( \frac{\rho}{\eta} \right)^{\frac{p}{2}} \frac{D_m^{p/2-1}}{n^p \|\sigma\|_\mu^p} \sum_{\lambda \in \Lambda_m} \mathbb{E} \left| \sum_{i=1}^n \varepsilon_{i+1}^* \varphi_\lambda(X_i^*) \sigma_A(X_i^*) \right|^p. \end{aligned}$$

This term is handled by using a Rosenthal moment inequality (see Petrov (1995) or a recall in Baraud (2000), Theorem 8.1) applied to centered and block-independent variables admitting moments of order  $p$  ( $\mathcal{I}_\ell^{(1,2)}$  is set for successively  $\mathcal{I}_\ell^{(1)}$  and  $\mathcal{I}_\ell^{(2)}$ ): there exists a constant  $c(p)$  such that

$$\mathbb{E} \left| \sum_{i=1}^n Z_i \right|^p \leq 2c(p) \left\{ \sum_{\ell=1}^{\ell_n} \mathbb{E} \left| \sum_{i \in \mathcal{I}_\ell^{(1,2)}} Z_i \right|^p + \left[ \sum_{\ell=1}^{\ell_n} \mathbb{E} \left( \sum_{i \in \mathcal{I}_\ell^{(1,2)}} Z_i \right)^2 \right]^{p/2} \right\} \quad (5.38)$$

where  $Z_i = \varepsilon_{i+1}^* \varphi_\lambda(X_i^*) \sigma_A(X_i^*)$ . Next we bound both terms separately.

$$\begin{aligned} \mathbb{E} \sum_{\ell=1}^{\ell_n} \left| \sum_{i \in \mathcal{I}_\ell^{(1,2)}} Z_i \right|^p &\leq \sum_{\ell=1}^{\ell_n} \mathbb{E} \left( \sum_{i \in \mathcal{I}_\ell^{(1,2)}} |\varepsilon_{i+1} \varphi_\lambda(X_i) \sigma_A(X_i)| \right)^p \\ &\leq (\Phi \sqrt{D_m})^p \|\sigma_A\|_\infty^p q_n^{p-1} \sum_{\ell=1}^{\ell_n} \mathbb{E} \left( \sum_{i \in \mathcal{I}_\ell^{(1,2)}} |\varepsilon_{i+1}|^p \right) \\ &\leq (2\Phi \|\sigma_A\|_\infty \sigma_p)^p q_n^{p-1} n D_m^{p/2} \end{aligned}$$

using that  $n = \ell_n q_n$  and

$$\left( \sum_{\ell=1}^{\ell_n} \mathbb{E} \left( \sum_{i \in \mathcal{I}_\ell^{(1,2)}} Z_i \right)^2 \right)^{p/2} \leq \|\sigma_A\|_\infty^p \sigma_p^2 \left[ \sum_{i=1}^n \mathbb{E} \varphi_\lambda^2(X_i) \right]^{p/2} \leq (\|\sigma_A\|_\infty \sigma_2)^p n^{p/2}.$$

Therefore

$$\mathbb{P}^{*,\rho} \left( \frac{1}{n^2} \sum_{\lambda \in \Lambda_m} \left( \sum_{i=1}^n \varepsilon_{i+1} \varphi_\lambda(X_i) \sigma(X_i) \right)^2 \geq \frac{\eta}{\rho} \|\sigma\|_\mu^2 \right) \leq C \left( D_m^p n^{1-p} q_n^{p-1} + \frac{D_m^{p/2}}{n^{p/2}} \right) \quad (5.39)$$

where  $C = C(p, \rho, \tau, \Phi_1, \sigma_p, \|\sigma_A\|_\infty)$ . Therefore, under (5.37), and for  $p \geq 8$ , we have

$$\mathbb{P}^{*,\rho}(\|\Pi_m(X) \varepsilon\|_n^2 \geq \eta \|\sigma\|_\mu^2) \leq C(\tau, p, \Phi_1, \rho) n^{-k}.$$

A bound for  $\mathbb{P}(\|\sigma \varepsilon\|_n^2 \leq (1-\eta) \|\sigma\|_\mu^2)$  is obtained by applying a Rosenthal type inequality as well.

$$\begin{aligned} &\mathbb{P}^{*,\rho}(\|\sigma \varepsilon\|_n^2 \leq (1-\eta) \|\sigma\|_\mu^2) \\ &= \mathbb{P}^{*,\rho} \left( \frac{1}{n} \sum_{i=1}^n \sigma^2(X_i) (\varepsilon_{i+1}^2 - 1) + \frac{1}{n} \sum_{i=1}^n [\sigma^2(X_i) - \mathbb{E}(\sigma^2(X_i))] \leq -\eta \|\sigma\|_\mu^2 \right) \\ &\leq \mathbb{P}^{*,\rho} \left( \frac{1}{n} \left| \sum_{i=1}^n \sigma^2(X_i) (\varepsilon_{i+1}^2 - 1) \right| \geq \frac{\eta}{2} \|\sigma\|_\mu^2 \right) \\ &\quad + \mathbb{P}^{*,\rho} \left( \frac{1}{n} \left| \sum_{i=1}^n [\sigma^2(X_i) - \mathbb{E}(\sigma^2(X_i))] \right| \geq \frac{\eta}{2} \|\sigma\|_\mu^2 \right) \end{aligned}$$

$$\begin{aligned}
&\leq \frac{\mathbb{E} \left| \sum_{i=1}^n \sigma^2(X_i^*) (\varepsilon_{i+1}^* - 1) \right|^{p/2}}{n^{p/2} (\frac{\eta}{2} \|\sigma\|_\mu^2)^{p/2}} + \frac{\mathbb{E} \left| \sum_{i=1}^n [\sigma^2(X_i^*) - \mathbb{E}(\sigma^2(X_i^*))] \right|^{p/2}}{n^{p/2} (\frac{\eta}{2} \|\sigma\|_\mu^2)^{p/2}} \\
&\leq \frac{2c(p) \mathbb{E}_\mu(|\sigma|^p)}{n^{p/2} (\frac{\eta}{2} \|\sigma\|_\mu^2)^{p/2}} \left( n^{p/2} q_n^{p/2-1} \mathbb{E} |\varepsilon_1^2 - 1|^{p/2} + n^{p/4} m_4^{p/4} + 2^{p/2} n q_n^{p/2-1} + 2^{p/2} n^{p/4} \right).
\end{aligned}$$

Since  $p > 4$ ,  $n^{1-p/2} < n^{-p/4}$  and thus the order is  $n^{-p/4}$  which is less than  $n^{-2}$  if  $p \geq 8$  and less than  $n^{-4}$  if  $p \geq 16$ .

To bound the last term  $\mathbb{P}^{*,\rho}(2|\langle b - \Pi_m(X)b, \sigma\varepsilon \rangle_n| \geq 2\eta\|\sigma\|_\mu^2)$ , we consider the two terms

$$\mathbb{P}^{*,\rho}(|\langle b, \sigma\varepsilon \rangle_n| \geq \eta\|\sigma\|_\mu^2/2) \text{ and } \mathbb{P}^{*,\rho}(|\langle \Pi_m(X)b, \sigma\varepsilon \rangle_n| \geq \eta\|\sigma\|_\mu^2/2).$$

Again, we clearly have

$$\mathbb{P}(2|\langle b, \sigma\varepsilon \rangle_n| \geq \eta\|\sigma\|_\mu^2/2) \leq 2^{p/2} \frac{\mathbb{E} \left| \sum_{i=1}^n b(X_i^*) \sigma(X_i^*) \varepsilon_{i+1}^* \right|^{p/2}}{n^{p/2} (\eta\|\sigma\|_\mu^2)^{p/2}}.$$

The moment of order  $p/2$  is bounded by applying again the moment inequality (5.38) to the blocks of  $b(X_i^*) \sigma(X_i^*) \varepsilon_{i+1}^*$ :

$$\mathbb{E} \left| \sum_{i=1}^n b(X_i^*) \sigma(X_i^*) \varepsilon_{i+1}^* \right|^p \leq c(p) \left[ (2\mathbb{E}_\mu(|b\sigma|^{p/2}) \sigma_p^p n q_n^{p/2-1} + n^{p/4} \sigma_2^p [\mathbb{E}_\mu(\sigma^2 b^2)]^{p/4} \right].$$

Thus as previously and since  $q_n$  is of order  $\ln(n)$ ,  $p > 4$ , the order is  $n^{-p/4}$  so that:

$$\mathbb{P}^{*,\rho} \left( |\langle b, \sigma\varepsilon \rangle_n| \geq \frac{\eta\|\sigma\|_\mu^2}{2} \right) \leq C(p, \tau, \mathbb{E}_\mu(|b|^p), \mathbb{E}_\mu(|\sigma|^p), \sigma_p) n^{-k}.$$

A2(p) ensures that  $\mathbb{E}_\mu(|b|^p), \mathbb{E}_\mu(|\sigma|^p)$  are finite.

In the same way as previously,

$$\frac{1}{n} |\langle \Pi_m(X)b, \sigma\varepsilon \rangle| = \frac{1}{n^2} | {}^t b_A V_m \Phi_m(X)^{-1} {}^t V_m \sigma_A \varepsilon | \leq \frac{\rho}{n^2} \| {}^t V_m b_A \| \| {}^t V_m \sigma_A \varepsilon \|.$$

Since

$$\begin{aligned}
\| {}^t V_m b \|^2 &= \sum_{\lambda \in \Lambda_m} \left( \sum_{i=1}^n \varphi_\lambda(X_i) b_A(X_i) \right)^2 \leq n \|b_A\|_\infty^2 \sum_{\lambda \in \Lambda_m} \sum_{i=1}^n \varphi_\lambda^2(X_i) \\
&\leq n^2 \|b_A\|_\infty^2 \left\| \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2 \right\|_\infty^2 \leq n^2 \|b_A\|_\infty^2 \Phi_1^2 D_m.
\end{aligned}$$

This implies

$$\begin{aligned}
&\mathbb{P}^{*,\rho} \left( \frac{1}{n} |\langle \Pi_m(X)b, \sigma\varepsilon \rangle| \geq \eta\|\sigma\|_\mu^2/2 \right) \\
&\leq \mathbb{P}^{*,\rho} \left( \frac{1}{n^2} \| {}^t V_m \sigma_A \varepsilon \| \geq \frac{\eta^2 \|\sigma\|_\mu^2}{4 \|b_A\|_\infty^2 \Phi_1^2 D_m} \right) \\
&= \mathbb{P}^{*,\rho} \left( \sum_{\lambda \in \Lambda_m} \left( \frac{1}{n} \sum_{i=1}^n \varphi_\lambda(X_i) \sigma_A(X_i) \varepsilon_{i+1} \right)^2 \geq \frac{\eta^2 \|\sigma\|_\mu^2}{4 \|b_A\|_\infty^2 \Phi_1^2 D_m} \right)
\end{aligned}$$

This term is nearly the same as (5.39) except that there is a loss of  $D_m^{p/2}$  in the final order. Ignoring the constants, we find for this probability an order  $D_m^{3p/2}n^{1-p}q_n^{p-1} + D_m^p n^{-p/2}$ . The final order is  $D_m^p n^{-p/2}$  and is less than  $n^{-k}$  as soon as  $D_m \leq n^{1/2-k/p}$ . Lastly, we have to bound  $\mathbb{P}(\hat{r}_n^2 \geq 2(1+\tau)(\|b\|_\mu^2 + \|\sigma\|_\mu^2))$ , but

$$\begin{aligned}\hat{r}_n^2 &\leq 2\|b - \Pi_m(X)b\|_n^2 + 2\|\sigma\varepsilon\|_n^2 \leq 2\|b\mathbf{1}_{A^c}\|_n^2 + 2\|b_A - \Pi_m(X)b_A\|_n^2 + 2\|\sigma\varepsilon\|_n^2 \\ &\leq 2\|b_{A^c}\|_n^2 + 2\|b_A\|_n^2 + 2\|\sigma\varepsilon\|_n^2 = 2\|b\|_n^2 + 2\|\sigma\varepsilon\|_n^2\end{aligned}$$

so that

$$\begin{aligned}\mathbb{P}(\hat{r}_n^2 \geq 2(1+\tau)(\|b\|_\mu^2 + \|\sigma\|_\mu^2)) &\leq \mathbb{P}(\|\sigma\varepsilon\|_n^2 \geq (1+\tau)\|\sigma\|_\mu^2) \\ &\quad + \mathbb{P}(\|b\|_n^2 \geq (1+\tau)\|b\|_\mu^2)\end{aligned}$$

and the first right-hand-side term has already been studied; the second one gives the same order with a Rosenthal inequality again. This completes the proof of Lemma 5.1.  $\square$

## 5.2 Proof of Theorem 3.2

We have

$$\begin{aligned}\gamma_n^{(2)}(t) - \gamma_n^{(2)}(s) &= \|t - \sigma^2\|_n^2 - \|s - \sigma^2\|_n^2 + \frac{4}{n} \sum_{i=1}^n b(X_i)\sigma(X_i)(s-t)(X_i)\varepsilon_{i+1} \\ &\quad + \frac{2}{n} \sum_{i=1}^n \sigma^2(X_i)(s-t)(X_i)(\varepsilon_{i+1}^2 - 1) \\ &\quad + \frac{2}{n} \sum_{i=1}^n (b^2 - \tilde{b}^2)(X_i)(s-t)(X_i)\end{aligned}$$

Since all functions  $s, t$  are  $A$ -supported, we can replace  $b$  and  $\sigma$  by  $b\mathbf{1}_A = b_A$  and  $\sigma\mathbf{1}_A = \sigma_A$  everywhere. Moreover, for any  $\theta > 0$ ,

$$\frac{2}{n} \sum_{i=1}^n (b_A^2 - \tilde{b}^2)(X_i)(s-t)(X_i) \leq \theta \|b_A^2 - \tilde{b}^2\|_n^2 + \frac{2}{\theta} (\|\sigma_A^2 - t\|_n^2 + \|\sigma_A^2 - s\|_n^2).$$

Next, as  $\gamma_n(\widetilde{\sigma^2}) - \gamma_n(\sigma_m^2) \leq \text{pen}(m) - \text{pen}(\hat{m})$ , we have, taking  $\theta = 16$ ,

$$\begin{aligned}\frac{7}{8} \|\widetilde{\sigma^2} - \sigma_A^2\|_n^2 &\leq \frac{9}{8} \|\sigma_m^2 - \sigma_A^2\|_n^2 + \frac{4}{n} \sum_{i=1}^n b_A(X_i)\sigma_A(X_i)(\widetilde{\sigma^2} - \sigma_m^2)(X_i)\varepsilon_{i+1} \\ &\quad + \frac{2}{n} \sum_{i=1}^n \sigma_A^2(X_i)(\widetilde{\sigma^2} - \sigma_m^2)(X_i)(\varepsilon_{i+1}^2 - 1) + 16\|b_A^2 - \tilde{b}^2\|_n^2 \\ &\quad + \text{pen}^{(2)}(m) - \text{pen}^{(2)}(\hat{m}).\end{aligned}$$

The terms to control are

$$\sup_{t \in B_2(m', \mu)} \frac{1}{n} \sum_{i=1}^n b_A(X_i)\sigma_A(X_i)t(X_i)\varepsilon_{i+1} \quad \text{and} \quad \sup_{t \in B_2(m', \mu)} \frac{1}{n} \sum_{i=1}^n \sigma_A^2(X_i)t(X_i)u_{i+1}$$

where  $u_i = \varepsilon_i^2 - 1$  are i.i.d. centered variables, with  $u_i$  independent of  $X_{i-1}$  and  $B_2(m', \mu) = \{t \in S_m^{(2)} + S_{m'}^{(2)}, \|t\|_\mu \leq 1\}$ ,  $D_2(m') = \dim(S_m^{(2)} + S_{m'}^{(2)})$ . They both are of the type of  $\tilde{W}$  previously studied. We set:

$$\begin{aligned}\tilde{W}_n^{(1)}(m') &= \left[ \left( \sup_{t \in B_2(m', \mu)} \sum_{i=1}^n \varepsilon_{i+1}^* b_A(X_i^*) \sigma_A(X_i^*) t(X_i^*) \right)^2 - x^2 n D_2(m') \Phi_2^2 \|b\sigma\|_\mu^2 \right]_+ \\ \tilde{W}_n^{(2)}(m') &= \left[ \left( \sup_{t \in B_2(m', \mu)} \sum_{i=1}^n u_{i+1}^* \sigma_A^2(X_i^*) t(X_i^*) \right)^2 - x^2 n D_2(m') \Phi_2^2 m_4 \|\sigma^2\|_\mu^2 \right]_+\end{aligned}$$

and we consider the new  $\tilde{\Omega}_{\tau, \rho}^* := \tilde{\Omega}_\tau \cap \Omega_\rho \cap \Omega^*$  where now

$$\tilde{\Omega}_\tau = \{(1 - \tau)s^2 \leq \hat{s}_n^2 \leq 2(1 + \tau) [\mathbb{E}_\mu[(b^2 + \sigma^2)^2] + s^2]\}, \quad (5.40)$$

with

$$s^2 = 4\mathbb{E}_\mu(b^2 \sigma^2) + m_4 \mathbb{E}_\mu(\sigma^4).$$

We recall that  $S^2 = s^2 + \mathbb{E}_\mu[(b^2 + \sigma^2)^2]$ . Therefore, we find:

$$\begin{aligned}\frac{7}{8} \|\widetilde{\sigma^2} - \sigma_A^2\|_n^2 \mathbf{1}_{\tilde{\Omega}_{\tau, \rho}^*} &\leq \frac{9}{8} \|\sigma_m^2 - \sigma_A^2\|_n^2 + \text{pen}^{(2)}(m) + 8\Phi_2^2 x^2 s^2 \frac{D_m^{(2)}}{n} \\ &\quad + 32\tilde{W}_n^{(1)}(\hat{m}_2) + 8\tilde{W}_n^{(2)}(\hat{m}_2) + \frac{\rho}{4} \|\widetilde{\sigma^2} - \sigma_m^2\|_n^2 \\ &\quad + 16\|b_A^2 - \tilde{b}^2\|_n^2 + 8\Phi_2^2 x^2 s^2 \frac{D_{\hat{m}_2}}{n} - \text{pen}^{(2)}(\hat{m}_2).\end{aligned} \quad (5.41)$$

For simplicity, we choose  $\rho = 3/2$  and this yields

$$\begin{aligned}\frac{1}{8} \|\widetilde{\sigma^2} - \sigma_A^2\|_n^2 \mathbf{1}_{\tilde{\Omega}_{\tau, \rho}^*} &\leq \frac{15}{8} \|\sigma_m^2 - \sigma_A^2\|_n^2 + 8 \frac{3 + \tau}{1 - \tau} \Phi_2^2 x^2 S^2 \frac{D_m^{(2)}}{n} \\ &\quad + 32 \sum_{m' \in \mathcal{M}_n} \tilde{W}_n^{(1)}(m') + 8 \sum_{m' \in \mathcal{M}_n} \tilde{W}_n^{(2)}(m') + 16\|b_A^2 - \tilde{b}^2\|_n^2\end{aligned}$$

provided that  $\kappa$  in  $\text{pen}^{(2)}$  is chosen in such a way that the last term in (5.41) is nonpositive i.e.  $\kappa = 8x^2/(1 - \tau)$ .

The bound for  $\sum_{m' \in \mathcal{M}_n^{(2)}} \mathbb{E}[\tilde{W}_n^{(1)}(m')]$  is the same as the one given in (5.33) with only  $\|\sigma_A\|_\infty$  replaced by  $\|b_A \sigma_A\|_\infty$  and the same conditions on  $p$ . To bound  $\sum_{m' \in \mathcal{M}_n} \mathbb{E}[\tilde{W}_n^{(2)}(m')]$  we must take into account that the  $u_i$ 's admit moments of order  $p/2$  only, thus (5.33) holds with  $\|\sigma_A\|_\infty$  replaced by  $\|\sigma_A\|_\infty^2$  and  $p$  replaced by  $p/2$ . Then the conditions required now to bound the last term (see (5.33) with  $p$  replaced by  $p/2$ ) are  $-p/4 + 1 \leq -a_2$  and  $b_2 + 1 - \{(p/2)(p/2 - 2)/[4(p/2 - 1)]\} < 0$ ; those conditions are fulfilled under (3.23). Therefore, the end being the same as in the proof of Theorem 3.1, the result follows from the following Lemma:

**Lemma 5.2** *Under the assumptions of Theorem 3.2 and if*

$$D_{m_n}^{(2)} \leq n^{1/2-4/p} \quad (5.42)$$

*and  $p \geq 16$  (so that  $D_m^{(2)}$  can be of order  $n^{1/4}$ ), then*

$$\mathbb{P}\left(\tilde{\Omega}_\tau^c \cap \Omega_\rho \cap \Omega^*\right) \leq Cn^{-2}$$

*where  $C$  is a constant depending in particular on  $\Phi_2, p, \rho, \|\sigma\|_\mu, \|\sigma_A\|_\infty, \|b_A\|_\infty$ .*

This ends the proof of Theorem 3.2.  $\square$

*Proof of Lemma 5.2.* We follow the line and the notations of the proof of Lemma 5.1 and write, if  $X^2$  has coordinates  $(X_{i+1}^2)$ ,  $i = 1, \dots, n$ :

$$\begin{aligned}\hat{s}_n^2 &= \|X^2 - \Pi_m(X)X^2\|_n^2 \\ &= \|(b^2 + \sigma^2) - \Pi_m(b^2 + \sigma^2)\|_n^2 + \|2b\sigma\varepsilon + \sigma^2(\varepsilon^2 - 1)\|_n^2 \\ &\quad - \|\Pi_m(X)(2b\sigma\varepsilon + \sigma^2(\varepsilon^2 - 1))\|_n^2 \\ &\quad + \frac{2}{n} < (b^2 + \sigma^2) - \Pi_m(b^2 + \sigma^2), 2b\sigma\varepsilon + \sigma^2(\varepsilon^2 - 1) >.\end{aligned}$$

Then all terms can be treated as previously. For instance

$$\begin{aligned}& \mathbb{P}^{*,\rho} (\|2b\sigma\varepsilon + \sigma^2(\varepsilon^2 - 1)\|_n^2 \leq (1 - \eta)(4\|b\sigma\|_\mu^2 + m_4\|\sigma^2\|_\mu^2)) \\ & \leq \mathbb{P} \left( \frac{4}{n} \left| \sum_{i=1}^n b(X_i^*)\sigma^3(X_i^*)(\varepsilon_{i+1}^{*2} - 1)\varepsilon_{i+1}^* \right| \geq (\eta/5)(4\|b\sigma\|_\mu^2 + m_4\|\sigma^2\|_\mu^2) \right) \\ & \quad + \mathbb{P} \left( \frac{4}{n} \left| \sum_{i=1}^n b^2(X_i^*)\sigma^2(X_i^*)(\varepsilon_{i+1}^{*2} - 1) \right| \geq (\eta/5)(4\|b\sigma\|_\mu^2 + m_4\|\sigma^2\|_\mu^2) \right) \\ & \quad + \mathbb{P} \left( \frac{4}{n} \left| \sum_{i=1}^n (b^2(X_i^*)\sigma^2(X_i^*) - \mathbb{E}_\mu(b^2\sigma^2)) \right| \geq (\eta/5)(4\|b\sigma\|_\mu^2 + m_4\|\sigma^2\|_\mu^2) \right) \\ & \quad + \mathbb{P} \left( \frac{1}{n} \left| \sum_{i=1}^n \sigma^4(X_i^*)[(\varepsilon_{i+1}^{*2} - 1)^2 - m_4] \right| \geq (\eta/5)(4\|b\sigma\|_\mu^2 + m_4\|\sigma^2\|_\mu^2) \right) \\ & \quad + \mathbb{P} \left( \frac{m_4}{n} \left| \sum_{i=1}^n [\sigma^4(X_i^*) - \mathbb{E}_\mu(\sigma^4)] \right| \geq (\eta/5)(4\|b\sigma\|_\mu^2 + m_4\|\sigma^2\|_\mu^2) \right)\end{aligned}$$

and all terms can be treated thanks to a Rosenthal inequality of order  $p/4$ . This implies an order  $n^{-p/8}$ , less than  $n^{-2}$  for  $p \geq 16$  as assumed in (3.23).

Analogously the term  $\|\Pi_m(X)(2b\sigma\varepsilon + \sigma^2(\varepsilon^2 - 1))\|_n^2$  is found of order

$$(D_m^{(2)})^{p/2} n^{1-p/2} q_n^{p/2-1} + (D_m^{(2)})^{p/4} n^{-p/4}$$

and the scalar product term of order

$$(D_m^{(2)})^{3p/4} n^{1-p/2} q_n^{p/2-1} + (D_m^{(2)})^{p/2} n^{-p/4}.$$

They are less than  $n^{-2}$  if  $(D_m^{(2)})^{p/2} \leq n^{p/4-2}$  and  $p > 8$  which explains condition (5.42).  $\square$

### 5.3 Proof of Proposition 3.2

We start from (5.31) which only requires to be squared:

$$\begin{aligned}\|b_A - \tilde{b}\|_n^4 \mathbf{1}_{\Omega_{\tau,\rho}^*} &\leq C'_1(x, \tau, \rho) \left[ \|b_A - b_m\|_n^4 + (\|b\|_\mu^4 + \|\sigma\|_\mu^4) \frac{(D_m^{(1)})^2}{n^2} \Phi_1^4 \right] \\ &\quad + \frac{2x^2}{n^4} \tilde{W}_n^2(\hat{m}).\end{aligned}\tag{5.43}$$

Choosing  $\bar{p} = 2$  in (5.32) and using the extended Proposition 6 of Baraud et al. (1999), we can replace (5.33) by

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E} \left[ \frac{(\tilde{W}_n(m'))^2}{n^4} \right] \leq K n^{-2} \left[ \sum_{m' \in \mathcal{M}_n^{(1)}} (D_{m'}^{(1)})^{-p/2+2} + \frac{q_n^p |\mathcal{M}_n^{(1)}|}{n^{p(p-2)/[4(p-1)]-2}} \right].$$

where  $K = C(x, p)(\Phi_1^2 h_0)^{p/2} \|\sigma_A\|_\infty^2 \sigma_p^2$ . The last bracketed term is bounded if  $-p/2 + 2 \leq -a_1$  and  $b_1 + 2 - \{p(p-2)/[4(p-1)]\} < 0$ . This gives the conditions  $p \geq 2(2 + a_1)$  and  $p > 4b_1 + 10$ ; these conditions are fulfilled under (3.26). Therefore under (3.26),  $\sum_{m' \in \mathcal{M}_n} \mathbb{E}(\tilde{W}_n^2(m')/n^4)$  is of order  $1/n^2$ .

Taking the expectation of (5.43) gives

$$\mathbb{E} \left[ \|b_A - \tilde{b}\|_n^4 \mathbf{I}_{\Omega_{\tau, \rho}^*} \right] \leq C(x, \tau, \rho) \left[ \mathbb{E}(\|b_A - b_m\|_n^4) + (\|b\|_\mu^4 + \|\sigma\|_\mu^4) \Phi_1^4 \frac{D_m}{n^2} \right] + \frac{K'}{n^2}$$

where  $K'$  depends on  $x, \rho, h_0, \|\sigma_A\|_\infty, \Phi_1, \Sigma_1, T_1$ . We write that

$$\mathbb{E}(\|b_A - b_m\|_n^4) = \|b_A - b_m\|_\mu^4 + \frac{1}{n^2} \text{Var} \left[ \sum_{i=1}^n (b_A - b_m)^2(X_i) \right].$$

From Theorem 2.1 in Viennet (1997), we know that there exists a function  $B$  satisfying  $\mathbb{E}[B^k(X_1)] \leq k \sum_{l \geq 0} (l+1)^{k-1} \beta_l$  and such that

$$\text{Var} \left( \sum_{i=1}^n h(X_i) \right) \leq 2n \int B(x) h^2(x) d\mu(x)$$

for a sequence  $(X_i)$  stationary with stationary law  $\mu$  and absolutely regular with  $\beta$ -mixing coefficients  $\beta_l$ . Therefore

$$\begin{aligned} \text{Var} \left( \sum_{i=1}^n (b_A - b_m)^2(X_i) \right) &\leq 2n [\mathbb{E}(B^2(X_1)) \mathbb{E}((b_A - b_m)^8(X_1))]^{1/2} \\ &\leq 2n h_1^{1/2} \|b_A - b_m\|_8^4 \sqrt{2 \sum_{l=0}^{+\infty} (l+1) M e^{-\theta l}} \\ &\leq C(M, \theta, h_1) n \|b_A - b_m\|_8^4 \end{aligned}$$

This yields

$$\mathbb{E}(\|b_A - b_m\|_n^4) \leq \|b_A - b_m\|_\mu^4 + \frac{C(M, \theta, h_1)}{n} \|b_A - b_m\|_8^4$$

which is the first part of the right hand side of (3.27).

The last thing to check is the order of the expectation of  $\|b_A - b_m\|_n^4$  on the complementary of  $\Omega_{\tau, \rho}^*$ . Since

$$\|b_A - \tilde{b}\|_n^2 = \|b_A - \Pi_{\hat{m}}(X) b_A\|_n^2 + \|\Pi_{\hat{m}}(X) \sigma \varepsilon\|_n^2 \leq \|b_A\|_n^2 + \|\sigma \varepsilon\|_n^2$$

we have

$$\mathbb{E} \left[ \|b_A - \tilde{b}\|_n^4 \mathbf{I}_{(\Omega_{\tau, \rho}^*)^c} \right] \leq 2 \left\{ \|b_A\|_\infty^4 \mathbb{P}[(\Omega_{\tau, \rho}^*)^c] + \mathbb{E} \left[ \frac{1}{n^2} \left( \sum_{i=1}^n \sigma^2(X_i) \varepsilon_{i+1}^2 \right)^2 \mathbf{I}_{(\Omega_{\tau, \rho}^*)^c} \right] \right\}$$

$$\begin{aligned}
&\leq 2 \left[ \|b_A\|_\infty^4 \mathbb{P}[(\Omega_{\tau,\rho}^*)^c] + \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left( \sigma^4(X_i) \varepsilon_{i+1}^4 \mathbf{1}_{(\Omega_{\tau,\rho}^*)^c} \right) \right] \\
&\leq 2 \left( \|b_A\|_\infty + \sqrt{\mathbb{E}(\varepsilon_1^8) \mathbb{E}_\mu(\sigma^8)} \right) \sqrt{\mathbb{P}[(\Omega_{\tau,\rho}^*)^c]}.
\end{aligned}$$

Thus, we need  $\sqrt{\mathbb{P}[(\Omega_{\tau,\rho}^*)^c]}$  to be order  $n^{-2}$  i.e. that  $\mathbb{P}[(\Omega_{\tau,\rho}^*)^c] \leq C/n^4$  which is ensured by Lemma 5.1 (take  $k = 4$ ) under the Assumptions of Theorem 3.2.  $\square$

## 6 Appendix: Value of the parameters associated to the models and further numerical results.

The functions are given in Table 1 and the parameters are given first for the regressive case as a function of  $s2n$ , then in the autoregressive case, in the order corresponding to  $s2n$  values 1, 3, 7, 10, first for gaussian errors and next for uniform errors.

- M1  $\sigma = 1.1015/s2n$ ;  $\sigma = (1.85, 0.562, 0.239, 0.167)$ ,  $\sigma = (1.85, 0.562, 0.2385, 0.1669)$ ,
- M2  $\sigma = 0.7728/s2n$ ;  $\sigma = (0.917, 0.318, 0.137, 0.096)$ ,  $\sigma = (0.917, 0.31, 0.137, 0.096)$ ,
- M3  $\sigma = 0.6416/s2n$ ;  $\sigma = (0.555, 0.161, 0.0549, 0.0371)$ ,  $\sigma = (0.555, 0.169, 0.055, 0.0371)$ ,
- M4  $\sigma = 0.8669/s2n$ ;  $\sigma = (0.928, 0.305, 0.132, 0.0935)$ ,  $\sigma = (0.93, 0.302, 0.1316, 0.0935)$ ,
- M5  $\sigma = 1/(\sqrt{2}s2n)$ ;  $\sigma = (0.7071, 0.2447, 0.1184, 0.0877)$ ,  $\sigma = (0.7071, 0.2405, 0.1175, 0.0881)$ ,
- M6  $\sigma = 1/(\sqrt{2}s2n)$ ;  $\sigma = (0.7071, 0.2357, 0.102, 0.0718)$ ,  $\sigma = (0.7071, 0.2341, 0.1019, 0.0719)$ ,
- M7  $\sigma = 1/(\sqrt{2}s2n)$ ;  $\sigma = (0.7071, 0.2357, 0.101, 0.071)$ ,  $\sigma = (0.7071, 0.2357, 0.1006, 0.0707)$ ,
- M8  $\sigma = 1/(\sqrt{2}s2n)$ ;  $\sigma = (0.777, 0.291, 0.1269, 0.0889)$ ,  $\sigma = (0.7813, 0.2898, 0.1268, 0.0891)$ ,
- M9  $\sigma = \sqrt{2}/s2n$ ;  $\sigma = (1.266, 0.423, 0.188, 0.133)$ ,  $\sigma = (1.266, 0.423, 0.1877, 0.1329)$ ,
- M10  $a = 0.2762s2n$ ,  $\sigma = 1$ ;  $a = 0.5$ ,  $\sigma = (1.255, 0.331, 0.108, 0.0728)$ ,  $a = 0.5$ ,  $\sigma = (1.27, 0.352, 0.1077, 0.0727)$ ,
- M11  $\sigma = 1.1139/s2n$ ;  $\sigma = (1.19, 0.376, 0.1603, 0.1122)$ ,  $\sigma = (1.19, 0.376, 0.1603, 0.1122)$ ,
- M12  $\beta = 2$ ,  $a = s2n/1.5235$ ;  $\beta = 1$ ,  $a = (0.707, 0.9485, 0.9895, 0.9947)$ ,  $\beta = 1$ ,  $a = (0.707, 0.9485, 0.9896, 0.9947)$ ,
- M13  $\beta = 5$ ,  $a = s2n/2.5347$ ; no autoregressive counterpart.

|     | s2n=1 |          | s2n=3 |          | s2n=7 |          | s2n=10 |          |
|-----|-------|----------|-------|----------|-------|----------|--------|----------|
|     | $b$   | $\sigma$ | $b$   | $\sigma$ | $b$   | $\sigma$ | $b$    | $\sigma$ |
| M1  | 1.4   | 1.0      | 1.4   | 1.0      | 1.4   | 1.0      | 1.4    | 1.0      |
| M2  | 1.9   | 1.0      | 2.3   | 1.1      | 2.3   | 1.1      | 2.3    | 1.1      |
| M3  | 1.5   | 1.4      | 1.6   | 1.3      | 1.2   | 1.2      | 1.3    | 1.3      |
| M4  | 1.3   | 1.7      | 1.2   | 2.1      | 1.2   | 1.2      | 1.2    | 1.0      |
| M5  | 1.2   | 1.6      | 1.3   | 1.4      | 1.5   | 1.2      | 1.2    | 1.5      |
| M6  | 1.3   | 2.2      | 1.3   | 1.7      | 1.1   | 2.1      | 1.3    | 2.4      |
| M7  | 1.6   | 3.1      | 1.3   | 2.7      | 1.3   | 2.5      | 1.3    | 2.4      |
| M8  | 1.3   | 8.5      | 1.1   | 2.5      | 1.3   | 1.9      | 1.1    | 2.3      |
| M9  | 1.5   | 1.6      | 1.1   | 1.7      | 1.3   | 1.5      | 1.3    | 1.6      |
| M10 | 1.5   | 1.9      | 1.9   | 1.9      | 1.1   | 1.1      | 1.2    | 1.3      |
| M11 | 1.8   | 2.6      | 1.5   | 2.3      | 1.6   | 1.3      | 1.6    | 1.1      |

Table 4. Ratio to the oracle of the  $L_2$  risk for the first step estimator in the autoregressive case for uniform errors.

M14  $\beta = 4$ ,  $a = s2n/0.6007$ ;  $\beta = 0.25$ ,  $a = (0.708, 0.951, 0.9919, 0.9961)$ ,  $\beta = 0.25$ ,  $a = (0.707, 0.95, 0.9918, 0.9961)$ ,

M15  $\beta = 2$ ,  $a = s2n/0.5675$ ; no autoregressive counterpart

M16  $x_0 = 1/\sqrt{2}$ ,  $\sigma = 0.8616$ ,  $a = s2n$ ;  $a = 1.04$ ,  $x_0 = \sqrt{2}$ ,  $\sigma = (1.044, 0.314, 0.162, 0.125)$ ;  $a = 1.04$ ,  $x_0 = \sqrt{2}$ ,  $\sigma = (1.03, 0.314, 0.162, 0.1254)$ .

|     | s2n=1 |          | s2n=3 |          | s2n=7 |          | s2n=10 |          |
|-----|-------|----------|-------|----------|-------|----------|--------|----------|
|     | $b$   | $\sigma$ | $b$   | $\sigma$ | $b$   | $\sigma$ | $b$    | $\sigma$ |
| M1  | 2.4   | 1.1      | 1.6   | 1.1      | 1.4   | 1.1      | 1.4    | 1.1      |
| M2  | 1.7   | 1.0      | 2.1   | 1.1      | 1.5   | 1.1      | 1.5    | 1.1      |
| M3  | 1.5   | 1.3      | 1.6   | 1.3      | 1.3   | 1.3      | 1.6    | 1.3      |
| M4  | 1.8   | 1.3      | 1.4   | 1.3      | 1.6   | 1.3      | 1.3    | 1.4      |
| M5  | 1.9   | 1.2      | 1.3   | 1.2      | 1.6   | 1.2      | 1.5    | 1.2      |
| M6  | 1.3   | 1.3      | 1.3   | 1.1      | 1.5   | 1.3      | 1.5    | 1.3      |
| M7  | 1.3   | 1.4      | 1.3   | 1.5      | 1.2   | 1.6      | 1.3    | 1.8      |
| M8  | 1.4   | 1.3      | 1.2   | 1.3      | 1.5   | 1.5      | 1.2    | 1.4      |
| M9  | 1.6   | 1.2      | 1.4   | 1.2      | 1.5   | 1.2      | 1.6    | 1.2      |
| M10 | 1.6   | 1.3      | 2.4   | 1.3      | 1.3   | 1.3      | 1.5    | 1.3      |
| M11 | 2.4   | 1.3      | 1.6   | 1.3      | 1.3   | 1.3      | 1.3    | 1.3      |
| M12 | 2.0   | 1.6      | 1.4   | 1.5      | 1.4   | 1.5      | 1.4    | 1.5      |
| M13 | 2.7   | 1.5      | 2.2   | 1.5      | 2.2   | 1.5      | 2.2    | 1.5      |
| M14 | 1.9   | 1.5      | 1.4   | 1.5      | 1.3   | 1.4      | 1.3    | 1.4      |
| M15 | 2.0   | 1.1      | 1.5   | 1.1      | 1.4   | 1.1      | 1.4    | 1.1      |
| M16 | 1.3   | 1.5      | 0.8   | 3.6      | 0.4   | 11       | 0.3    | 31       |

Table 5. Ratio to the oracle of the  $L_2$  risk for the first step estimator in the regressive case for uniform errors.

## References

- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. *Proceedings 2nd International Symposium on Information Theory*, P.N. Petrov and F. Csaki (Eds.), Akademia Kiado, Budapest, 267-281.
- Ango Nze, P., 1992. Critères d'ergodicité de quelques modèles à représentation markovienne. *C. R. Acad. Sci. Paris Ser. I Math.* 315, 1301-1304.

- Ango Nze, P., 1998. Critères d'ergodicité géométrique ou arithmétique de modèles linéaires perturbés à représentation markovienne. C. R. Acad. Sci. Paris Ser. I Math. 326, 371-376.
- Baraud, Y., 2000. Model selection for regression on a fixed design. Probab. Theory Relat. Fields 117, 467-493.
- Baraud, Y, Comte, F. and Viennet, G., 1999. Adaptive estimation in an autoregression or a  $\beta$ -mixing regression. To appear in the Ann. Statist., Preprint 566 University Paris 6, [www.proba.jussieu.fr/mathdoc/preprints](http://www.proba.jussieu.fr/mathdoc/preprints).
- Barron, A., Birgé, L and Massart, P., 1999. Risks bounds for model selection via penalization. Probab. Theory Relat. Fields 113, 301-413.
- Birgé, L. and Massart, P., 1997. From model selection to adaptive estimation. In Festschrift for Lucien Lecam: Research Papers in Probability and Statistics (D. Pollard, E. Torgensen and G. Yangs, eds), 55-87, Springer Verlag, New-York.
- Birgé, L. and Rozenholc, Y., 2000. How many bins must be put in a regular histogram. *Working paper*.
- Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econometrics 31, 307-327.
- Cohen, A. Daubechies, I. and Vial, P., 1993. Wavelet and fast wavelet transform on an interval. Appl. Comp. Harmon. Anal. 1, 54-81.
- Daubechies, I., 1992. Ten lectures on wavelets, SIAM: Philadelphia.
- DeVore, R.A. and Lorentz, C.G., 1993. Constructive Approximation. Springer-Verlag.
- Doukhan, P., 1995. Mixing properties and examples. Springer-Verlag.
- Donoho, D.L. and Johnstone, I.M., 1998. Minimax estimation via wavelet shrinkage. Ann. Statist. 26, 879-921.
- Duflo, M., 1990. Méthodes récursives aléatoires. Masson, Paris.
- Engle, R. F., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50, 987-1007.
- Fan, J. and Yao, Q., 1998. Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85, 645-660.
- Gouriéroux, C. and Monfort, A., 1992. Qualitative threshold ARCH models. J. Econometrics, 52, 159-199.
- Hall, P. and Carroll, R.J., 1989. Variance function estimation in regression: the effect estimation of the mean. J.R. Statist. Soc., B 51, 3-14.
- Härdle, W. and Tsybakov, A., 1997. Local polynomial estimators of the volatility function in nonparametric regression. J. Econometrics 81, 223-242.

- Härdle, W., Tsybakov, A. and Yang, L., 1998. Nonparametric vector autoregression. *J. Statist. Plan. Inf.* 68, 221-245.
- Hoffmann, M., 1999. On nonparametric estimation in nonlinear AR(1)-models. *Statist. Probab. Lett.* 44, 29-45.
- Kolmogorov, A.R. and Rozanov, Y.A., 1960. On the strong mixing conditions for stationary gaussian sequences. *Theor. Probab. Appl.* 5, 204-207.
- Li, K.C., 1987. Asymptotic optimality for  $C_p$ ,  $C_l$  cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* 15, 958-975.
- Lütkepohl, H., 1992. *Introduction to Multiple Time Series Analysis*. Springer, Heidelberg.
- Mallows, C.L., 1973. Some comments on  $C_p$ . *Technometrics* 15, 661-675.
- Mc Keague, I.W. and Zhang, M.J., 1994. Identification of nonlinear time series from first order cumulative characteristics. *Ann. Statist.* 22, 495-514.
- Müller, H.G. and Stadtmüller, U., 1987. Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* 15, 610-625.
- Neumann, M.H., 1994. Fully data-driven nonparametric variance estimators. *Statistics* 25, 189-212.
- Petrov, V.V., 1995. *Limit theorems of probability theory. Sequences of independent random variables*. Oxford Science Publications 4. The Clarendon Press, Oxford University Press, New York.
- Polyak, B.T. and Tsybakov, A., 1992. A family of asymptotically optimal methods for choosing the order of a projective regression estimate. *Theory Probab. Appl.* 37, 471-481.
- Shibata, R., 1976. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* 63, 117-126.
- Viennet, G., 1997. Inequalities for absolutely regular processes: application to density estimation. *Probab. Theory Relat. Fields* 107, 467-492.