Forecasting in functional regressive or autoregressive models

Anne Philippe[1] and Marie-Claude Viano[2]

Université de Nantes Université de Lille 1

2008-2009

[1] Laboratoire de mathématiques Jean Leray, 2 rue de la Houssinière 44322 Nantes, France
Anne.Philippe@univ-nantes.fr

[2] Laboratoire Paul Painlevé, Villeneuve d'Ascq, 59655 Cedex, France
viano@math.univ-lille1.fr

Contents

CHAPTER 1

Introduction

This document is devoted to questions related to forecasting time series models, a topic which has now a growing importance in various domains like signal and image processing, agro-industry, econometrics, geophysics and all socio-economics areas where good forecasts can greatly improve the gains and limit the wasting.

1. General model

We are interested in models like

$$X_k = a(X_{k-1}, \dots, X_{k-p}) + b(e_k, \dots, e_{k-q}) + \varepsilon_k \quad \forall k, \tag{1}$$

where the observed variables are $(X_n, \dots, X_1)$ and $(e_{n+1}, \dots, e_1)$ and where the sequence (ε_j) is an unobserved white noise. The goal is to predict the value of X_{n+h} for $h = 1, 2, \dots$ from the observed variables. For convenience, in most cases, we shall take the forecasting horizon $h = 1$. But it should be clear to the reader that this is a real loss of generality. Other values of h shall be treated in exercises.

Notice in (1) the simultaneous presence, in the right hand side, of an autoregressive summand and of a purely regressive one.

- the autoregressive part $a(X_{k-1}, \dots, X_{k-p})$ means that the past values of the time series, up to a lag of length p, affect the value X_{k+h}.
- the regressive part $b(e_k, \dots, e_{k-q})$ summarizes the action of an exogeneous sequence (e_j).

For example imagine that the electricity consumption X_k at time k depends on the consumption at the p instants just before and on the temperature $e_k, \dots, e_{k-q}$ at the moments $k, \dots, k-q$.

2. Optimal predictor

The most usual forecasting method consists in minimizing a quadratic criterion (assuming that the second order moments are finite). Namely

$$\tilde{X}_{n+h} = \text{Argmin}\{(X_{n+h} - Z)^2 | Z \in \mathcal{F}_n\},$$

where $\mathcal{F}_n$ is the σ-algebra generated by $(X_n, \dots, X_1), (e_{n+1}, \dots, e_1)$. With this criterion, $\tilde{X}_{n+h}$ is nothing else than the conditional expectation

$$\tilde{X}_{n+h} = \mathbb{E}(X_{n+h} | \mathcal{F}_n)$$

Consider the case $h = 1$. In all the situations studied below, ε_n is independent of $\mathcal{F}_n$, so that the one step ahead optimal predictor is

$$\tilde{X}_{n+1} = a(X_n, \ldots, X_{n-p+1}) + b(e_{n+1}, \ldots, e_{n-q+1}),$$

and consequently, ε_{n+1} is the forecasting error at horizon $h = 1$. Unfortunately, the functions a and b are generally unknown, so the statistician has to plug in an estimation of the functions. Consequently, the forecasting error includes both the theoretical error ε_n and the estimation error. More precisely, we have to replace $\tilde{X}_{n+1}$ by

$$\hat{X}_{n+1} = \hat{a}(X_n, \ldots, X_{n-p+1}) + \hat{b}(e_{n+1}, \ldots, e_{n-q+1}), \tag{2}$$

implying that

$$\begin{aligned} X_{n+1} - \hat{X}_{n+1} = \varepsilon_n \quad &+ \quad (a(X_n, \ldots, X_{n-p+1}) - \hat{a}(X_n, \ldots, X_{n-p+1})) \\ &+ \quad \Big(b(e_{n+1}, \ldots, e_{n-q+1}) - \hat{b}(e_{n+1}, \ldots, e_{n-q+1})\Big) . \end{aligned}$$

3. Difficulties

The theoretical treatments of the general model (1) are difficult for several reasons.

- The first reason is the fact these two regressions have a functional form: in order to predict X_{n+1}, one has to estimate two functions a and b. Estimating functions is always more tricky than to estimate finite-dimensional parameters.
- The second one is the simultaneous presence of regression and autoregression. Regression is easy to treat, being a relatively well known situation. Autoregression, which induces stochastic dependence between the $X_j's$, is much more difficult to handle, except in the familiar case of linear autoregression.

We shall proceed step by step. Firstly, in section 2 we deal with a linear version of (1). Then we treat in sections 3 and 4 simple regression models, and a simple autoregression one in section 5. In these two cases, we shall take $p = q = 1$ in (1), keeping in mind that the general case can be treated as well, despite a necessary multivariate treatment (see section 6 for example).

CHAPTER 2

Linear models

We begin with the linear version of (1)

$$X_k = a_0 + a_1 X_{k-1} \ldots + a_p X_{k-p} + b_0 e_k + \ldots + b_q e_{k-q} + \varepsilon_k \quad \forall k. \tag{3}$$

1. Assumptions

- *The noise* (ε_n) is a Gaussian zero-mean i.i.d sequence, with variance $\sigma^2 \neq 0$.
- *The exogeneous sequence* (e_n) is i.i.d, Gaussian, with zero mean and $\mathrm{Var}(e_n) = 1$.
- *Independence*: The two sequences (ε_n) and (e_n) are independent.
- *Stationarity*: $a_p \neq 0$ and the polynomial $A(z) = z^p - a_1 z^{p-1} - \ldots a_p$ does not vanish on the domain $|z| \geq 1$.
- *Minimality*: the two polynomials $A(z)$ and $B(z) = b_0 z^q + b_1 z^{q-1} + \ldots + b_q$ have no common root.
- *Stationarity again*: The process (X_n) is the unique stationary solution of (3).

Remark 1. Whiteness assumption of the input noise (ε_n) is rather natural, at least in a first approach. So is the independence of (ε_n) and of (e_n).

Remark 2. The Gaussian assumption is convenient, but could easily be relaxed.

Remark 3. Assuming that (e_n) is i.i.d. is not realistic in most cases (for example when e_n represents the temperature!), and should be relaxed. However, this situation is chosen here because it makes the developments more easy.

Indeed, in this case, equations (3) have a unique Gaussian stationary solution that satisfies the ARMA(p,q) representation

$$Y_k - a_1 Y_{k-1} - \ldots - a_p Y_{k-p} \quad = \quad \eta_k + c_1 \eta_{k-1} + \ldots + c_q \eta_{k-q} \tag{4}$$

with

$$Y_k = X_k - \mathbb{E}(X_k) \quad = \quad X_k - \frac{a_0}{1 - a_1 - \ldots - a_p} =: X_k - m_0. \tag{5}$$

where (η_k) is a zero-mean Gaussian white noise (see exercise 1 for the proof of this result in a simple case). This solution also writes

$$Y_k = u_k + d_1 u_{k-1} + \ldots + d_l u_{k-l} \ldots \tag{6}$$

where

$$u_k = b_0 e_k + \ldots + b_q e_{k-q} + \varepsilon_k, \tag{7}$$

and where the $d_j's$ are the coefficients of the expansion of the autoregressive part

$$(1 - a_1 z - \ldots - a_p z^p)^{-1} = 1 + d_1 z + \ldots + d_l z^l + \ldots$$

REMARK 4. The minimality assumption implies that it is impossible to find for the same model a shorter representation like

$$X_k = c_0 + c_1 X_{k-1} \ldots + c_{p-1} X_{k-p+1} + d_0 e_k + \ldots + d_{q-1} e_{k-q+1} + \varepsilon_k \quad \forall k.$$

2. Parameter estimation

In the ARMA representation (4), we know that the maximum likelihood estimate $\hat{\theta}_n^*$ of the vector parameter $\theta^* = {}^t(a_1, \ldots, a_p, c_1, \ldots, c_q)$ (hereafter, ${}^t v$ is written for the transpose of vector v) is almost surely convergent and that $n^{1/2}(\hat{\theta}_n^* - \theta)$ is asymptotically normally distributed (see for example [4], chapter 8). However, this is not a very useful result, for two reasons. The first one is that the vector we want to estimate is not θ^*, but $\theta =^t (a_1, \ldots, a_p, b_0, \ldots, b_q)$, the link between the two vectors being highly non linear (see exercise 1). The second reason is that in the classical ARMA theory, the input noise (η_n) is unobserved, while in model (3), the exogeneous part $e_k, \ldots, e_{k-q}$ is observed. The only unobserved term being ε_k. Estimator $\hat{\theta}_n^*$ is not fitted to this situation.

For those reasons, it is better to estimate θ by a direct least mean square method

$$\hat{\theta}_n = \text{Argmin} \left\{ \sum_{k=1+p\vee q}^{n} \left(X_k - \alpha_0 - \sum_{j=1}^{p} \alpha_j X_{k-j} - \sum_{j=0}^{q} \beta_j e_{k-j} \right)^2 \right\}, \tag{8}$$

the minimum being taken over $(\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_0, \ldots, \beta_q)$. Denoting

$$\begin{aligned} k_0 &= 1 + p \vee q \\ \phi_k &= {}^t(1, X_{k-1}, \ldots, X_{k-p}, e_k, \ldots, e_{k-q}) \\ M_n &= \sum_{k=k_0}^{n} \phi_k {}^t\phi_k \end{aligned}$$

it is easy to check that, if M_n is invertible, (8) has a unique solution given by

$$\hat{\theta}_n = M_n^{-1} \sum_{k=k_0}^{n} X_k \phi_k. \tag{9}$$

For this estimator the following result holds.

PROPOSITION 1. *As $n \to \infty$, with $\hat{\theta}_n$ defined as in (8),*
(i) $\hat{\theta}_n - \theta = o_{as}(n^{-\alpha})$ for every $\alpha < 1/2$
(ii) $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 M^{-1})$
where $\sigma^2 = \text{Var}(\varepsilon_n)$ and where $M = \mathbb{E}(\phi_k {}^t\phi_k)$, is invertible.

Proof. Recall that $u_n = o_{as}(v_n)$ means that $u_n v_n^{-1} \xrightarrow{a.s.} 0$.

Before proceeding, it is useful to see that the vector sequence $(\phi_k)_{k\geq k_0}$ is Gaussian, stationary and ergodic. Indeed, using (7),

$$\phi_k = \begin{pmatrix} 1 \\ a_0 + u_{k-1} + \sum_{j=1}^{\infty} d_j u_{k-1-j} \\ \vdots \\ a_0 + u_{k-p} + \sum_{j=1}^{\infty} d_j u_{k-p-j} \\ e_k \\ \vdots \\ e_{k-q} \end{pmatrix}$$

so that, introducing the backward shift operator B (defined by $B^m z_n = z_{n-m}$), we can write

$$\phi_k = \begin{pmatrix} 1 \\ a_0 \\ \vdots \\ a_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ B + \sum_{j=1}^{\infty} d_j B^j & 0 \\ \vdots & \vdots \\ B^p + \sum_{j=1}^{\infty} d_j B^{p+j} & 0 \\ 0 & B^0 \\ \vdots & \vdots \\ 0 & B^q \end{pmatrix} \begin{pmatrix} u_k \\ e_k \end{pmatrix}$$

so that the sequence $(\phi_k)_{k\geq k_0}$ is clearly constructed by linear filtering from the Gaussian sequence $(u_k, e_k)_{k\geq 1}$. Finally, taking (7) into account, $(\phi_k)_{k\geq k_0}$ is obtained from ${}^t(e_k, \varepsilon_k)_{k\geq 1}$ by linear filtering. Now we recall two results:

- If a stationary sequence (w_k) has a spectral density, it is also the case for every sequence (w'_k) obtained from w by linear filtering (i.e. $w'_k = \sum_0^{\infty} \gamma_j w_{k-j}$, with $\sum \|\gamma_j\|^2 < \infty$).
- A stationary Gaussian sequence having a spectral density is ergodic. For ergodicity and related properties see for example [5]).

First step: consider the matrix M_n. Since $(\phi_k)_{k\geq k_0}$ is ergodic, the law of large numbers applies, leading to

$$n^{-1} M_n \xrightarrow{a.s.} M = \mathbb{E}(\phi_k^t \phi_k). \tag{10}$$

Now, suppose that $\mathbb{E}(\phi_k^t \phi_k)$ is not invertible. This means that there exists a non zero vector v such that

$$\mathbb{E}({}^t v \phi_k^t \phi_k v) = 0 \quad \forall k$$

Hence, ${}^t v \phi_k =_{a.s} 0$ for all k. This implies that there exists coefficients such that

$$v_0 + v_1 X_{k-1} + \ldots + v_p X_{k-p} + w_0 e_k + \ldots w_q e_{k-q} = 0 \quad \forall k.$$

As e_k is independent of the other variables in the expression above, this implies that $w_0 = 0$, so that

$$v_1 X_{k-1} + \ldots + v_p X_{k-p} + w_1 e_{k-1} \ldots w_q e_{k-q} = -v_0 \quad \forall k.$$

But in turn this contradicts the hypothesis of minimality (see the assumption 4 in subsection 1) . Consequently, M is invertible, so that, almost surely, M_n is invertible for n sufficiently large, and formula (9) is then valid.

Second step: let us prove the almost sure convergence. From (9), and from the fact that $X_k = {}^t\phi_k \theta + \varepsilon_k$, write

$$\begin{aligned} n^\alpha(\hat{\theta}_n - \theta) &= M_n^{-1} \sum_{k=k_0}^{n} X_k \phi_k - \theta \\ &= n^\alpha \left(M_n^{-1} \sum_{k=k_0}^{n} \phi_k ({}^t\phi_k \theta + \varepsilon_k) - \theta \right) \\ &= n^\alpha \left(M_n^{-1} \sum_{k=k_0}^{n} \phi_k \varepsilon_k \right) = \left(n M_n^{-1} \right) \left(\frac{\sum_{k=k_0}^{n} \phi_k \varepsilon_k}{n^{1-\alpha}} \right) \end{aligned} \tag{11}$$

Now, ε_k is independent of all the coordinates of ϕ_k, so that $\mathbb{E}(\phi_k \varepsilon_k | \mathcal{F}_{k-1}) = 0$. In other words, $(\phi_k \varepsilon_k)_{k \geq k_0}$ is a (vector) martingale difference sequence with respect to the sequence $(\mathcal{F}_k)_{k \geq 1}$. Moreover, for $\beta > 1/2$

$$\sum_{k \geq k_0} \frac{\mathbb{E}\left(\|\phi_k \varepsilon_k\|^2 \right)}{k^{2\beta}} = \sigma^2 \mathbb{E}\left(\|\phi_k\|^2 \right) \sum_{k \geq k_0} \frac{1}{k^{2\beta}} < \infty$$

Hence, applying Theorem 3.3.1 in [16],

$$n^{-\beta} \sum_{k=k_0}^{n} \phi_k \varepsilon_k \xrightarrow{a.s.} 0.$$

Finally, as was seen above, $n M_n^{-1} \xrightarrow{a.s.} M^{-1}$. Using (10), the almost sure convergence is proved.

Third step: we prove now that

$$\frac{\sum_{k=k_0}^{n} \phi_k \varepsilon_k}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 M) \tag{12}$$

To prove this result, let T be a fixed integer, and consider the truncated expansion

$$X_k^{(T)} = m_0 + u_k + \sum_{j=1}^{T} d_j u_{k-j},$$

and the corresponding vector

$$\phi_k^{(T)} = {}^t(1, X_{k-1}^{(T)}, \ldots, X_{k-p}^{(T)}, e_k, \ldots, e_{k-q}).$$

It is easy to check that the sequence $(\varepsilon_k \phi_k^{(T)})_{k \geq k_0}$ is $T+p+q$-dependent (that is to say: $\varepsilon_k \phi_k^{(T)}$ and $\varepsilon_{k+h} \phi_{k+h}^{(T)}$ are independent as soon as $h > T+p+q$). Hence the central limit theorem holds. In order to find the covariance matrix of the limiting law, notice that $\mathbb{E}(\varepsilon_k \varepsilon_{k+h} \phi_k {}^t\phi_{k+h}) = 0$ if $h \neq 0$. Hence,

$$\frac{\sum_{k=k_0}^{n} \phi_k \varepsilon_k}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 M_T)$$

where $M_T = \mathbb{E}\left(\phi_k^{(T)\,t} \phi_k^{(T)}\right)$.

Finally, as $T \to \infty$, $M_T \to M$ and $\mathrm{Var}(\phi_k^{(T)} - \phi_k) \to 0$. To prove (11) it remains to apply the following lemma whose proof is left as an exercise.

Lemma 2. *Suppose that,*

$$Z_n = Z_{T,n,1} + Z_{T,n,2} \quad \forall n, T$$

where

- *for fixed T, as $n \to \infty$, $Z_{T,n,1} \xrightarrow{\mathcal{L}} \mathcal{N}(0, V_T)$*
- *$V_T \to V$ as $T \to \infty$*
- *$\mathrm{Var}(Z_{T,n,2}) \to 0$ uniformly with respect to n, as $T \to \infty$*

then $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, V)$ when $n \to \infty$.

□

3. Forecasting

As was seen in the introduction, we choose

$$\hat{X}_{n+1} = {}^t\hat{\theta}_n \phi_{n+1},$$

and, consequently, the forecasting error at horizon 1 is

$$X_{n+1} - \hat{X}_{n+1} = (\theta - {}^t\hat{\theta}_n)\phi_{n+1} + \varepsilon_{n+1},$$

where the two summands in the right hand side are independent. From Proposition 1, for every $\alpha < 1/2$, $n^{-\alpha}(\theta - {}^t\hat{\theta}_n) \xrightarrow{a.s.} 0$. Moreover the distribution of ϕ_n does not depend on n. To summarize,

Proposition 3. *With the same assumptions as in Proposition 1,*

$$X_{n+1} - \hat{X}_{n+1} = \varepsilon_{n+1} + T_n,$$

where ε_{n+1} and T_n are independent and where, as $n \to \infty$, $T_n = o_P(n^{-\alpha})$ for every $\alpha < 1/2$.

4. Exercises

Exercise 1. Prove representation (4) for the simple model

$$X_k - a_1 X_{k-1} = b_0 e_k + b_1 e_{k-1} + \varepsilon_k,$$

and give a hint for the proof in the general case (3). ⋆

Exercise 2. Prove Lemma 2. ⋆

Exercise 3. Give the expression of the optimal forecast $\tilde{X}_{n+2}$ at horizon $h = 2$ in the model of exercise 1. ⋆

Exercise 4. Consider now the linear model

$$X_{k+1} = aX_k + be_{k+1} + \varepsilon_{k+1},$$

and suppose that now that (e_k) is an autoregressive sequence

$$e_{k+1} = ce_k + \eta_{k+1}$$

where (η_k) is a zero mean white noise.

(1) Show that, if $|a| < 1$ and $|c| < 1$, there is a stationary solution (X_k, e_k).

(2) Working with this stationary solution, propose an estimator of the parameters a, b and c.

⋆

CHAPTER 3

Preliminaries on kernel methods in functional regression estimation

Kernel methods are old and popular methods used in all areas where the statistician has to estimate a functional parameter.

As examples,

- let the data $(Z_1, \ldots, Z_n)$ represent a sample from an i.i.d. sequence, the question being to estimate the density of the marginal distribution.
- Or let $[(Y_1, Z_1), \ldots, (Y_n, Z_n)]$ be a sample of an i.i.d. sequence, the problem being then to estimate $\mathbb{E}(Z_1|Y_1 = y)$.

The first example is the problem of density estimation, for which kernel methods were proposed by Parzen in 1962. The second is the problem of regression estimation, for which kernel methods were proposed by Nadaraya and Watson in 1964.

Here we concentrate on regression estimation, and the so-called Nadaraya-Watson estimator.

1. Heuristic approach

1.1. Step 1. In the case of discrete data, when the denominator does not vanish

$$r(y) := \mathbb{E}(Z_1|Y_1 = y) = \frac{\mathbb{E}(Z_1\, \mathbb{I}_{Y_1=y})}{P(Y_1 = y)}. \tag{13}$$

hence, from the sample $(Y_1, Z_1), \ldots, (Y_n, Z_n)$, it is natural estimate $r(y)$ by

$$\hat{r}_n(y) = \frac{n^{-1}\sum_{j=1}^n Z_j\, \mathbb{I}_{Y_j=y}}{n^{-1}\sum_{j=1}^n \mathbb{I}_{Y_j=y}} = \frac{\sum_{j=1}^n Z_j\, \mathbb{I}_{Y_j=y}}{\sum_{j=1}^n \mathbb{I}_{Y_j=y}}$$

which, thanks to the law of large numbers, converges towards the conditional expectation. Now, when the data are not discrete, formula (12) no longer holds, both numerator and denominator generally being zero.

1.2. Step 2. However, the same method could be applied to estimate (if the denominator is non zero)

$$\mathbb{E}\left(Z_1|Y_1 \in [y-h, y+h]\right) = \frac{\mathbb{E}(Z_1\, \mathbb{I}_{Y_1\in[y-h,y+h]})}{P(Y_1 \in [y-h, y+h])} \tag{14}$$

by

$$\frac{\sum_{j=1}^n Z_j \, \mathbb{I}_{Y_j \in [y-h, y+h]}}{\sum_{j=1}^n \mathbb{I}_{Y_j \in [y-h, y+h]}}.$$

1.3. Step 3. As every one knows, if $h \to 0$ in (13), the left hand side tends to $\mathbb{E}(Z_1 | Y_1 = y)$, at least under suitable smoothness assumptions.

From this, it seems natural to replace h by a sequence h_n tending to zero as $n \to \infty$, and take

$$\hat{r}_n(y) = \frac{\sum_{j=1}^n Z_j \, \mathbb{I}_{Y_j \in [y-h_n, y+h_n]}}{\sum_{j=1}^n \mathbb{I}_{Y_j \in [y-h_n, y+h_n]}}, \tag{15}$$

where h_n has to decrease when the sample size increases. This last point has to be developed. From now on, in order to have a well defined estimator, we take $0/0 = 0$.

1.4. Step 4. Fast enough, but not too fast! Writing

$$Z_j = \mathbb{E}(Z_j | Y_j) + (Z_j - \mathbb{E}(Z_j | Y_j)) = r(Y_j) + \eta_j$$

where Y_j and $Z_j - \mathbb{E}(Z_j | Y_j) =: \eta_j$ are uncorrelated, we get

$$\begin{aligned}
\hat{r}_n(y) - r(y) = \frac{\sum_{j=1}^n (Z_j - r(y)) \, \mathbb{I}_{Y_j \in [y-h_n, y+h_n]}}{\sum_{j=1}^n \mathbb{I}_{Y_j \in [y-h_n, y+h_n]}} &= \frac{\sum_{j=1}^n (r(Y_j) - r(y)) \, \mathbb{I}_{Y_j \in [y-h_n, y+h_n]}}{\sum_{j=1}^n \mathbb{I}_{Y_j \in [y-h_n, y+h_n]}} \\
&+ \frac{\sum_{j=1}^n \eta_j \, \mathbb{I}_{Y_j \in [y-h_n, y+h_n]}}{\sum_{j=1}^n \mathbb{I}_{Y_j \in [y-h_n, y+h_n]}} \\
&= A_n + B_n.
\end{aligned}$$

First consider A_n. If r is continuous, it is clear that this term tends to zero if $h_n \to 0$. In fact, the smaller h_n is, the smaller A_n.

Now consider B_n. For the sake of simplicity, suppose that (Y_j, Z_j) is Gaussian. Then for every j, η_j is independent from all the indicators $\mathbb{I}_{Y_l \in [y-h_n, y+h_n]}$, so that

$$B_n = \sum_{j=1}^n \eta_j u_j$$

where, for every j, η_j and u_j are independent and $\mathbb{E}(\eta_j = 0)$. This implies that $\mathbb{E}(B_n) = 0$ and that, with $p_n = P(Y_j \in [y - h_n, y + h_n])$

$$\begin{aligned}
\operatorname{Var}(B_n) &= \mathbb{E}(\operatorname{Var}(B_n | u_1, \ldots, u_n)) = \operatorname{Var}(\eta_1) \mathbb{E}\left(\sum_{j=1}^n u_j^2 \right) \\
&= \operatorname{Var}(\eta_1) \mathbb{E}\left(\frac{1}{\sum_{j=1}^n \mathbb{I}_{Y_j \in [y-h_n, y+h_n]}} \right) \geq \operatorname{Var}(\eta_1) \frac{1}{np_n}. \qquad (16)
\end{aligned}$$

This proves that $np_n \to \infty$ is necessary for $\operatorname{Var}(B_n) \to 0$. From this it is clear that for the convergence of B_n to zero, h_n has to tend to zero not too fast. For example, if the Y_j are uniformly distributed, we have $p_n \sim ch_n$, and then we see that the two conditions are

$$h_n \to 0 \qquad \text{and} \qquad nh_n \to \infty.$$

More generally, we shall see that it is a general feature when a kernel method is used for estimating a function that the same kind of antagonist constraints hold. The consequence for the practitioner is that *the smoothing parameter has to be carefully regulated.*

1.5. Step 5. Choice of the kernel. The estimator in (14) also writes

$$\hat{r}_n(y) = \sum_{j=1}^{n} Z_j \frac{K\left(\frac{Y_j - y}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{Y_j - y}{h_n}\right)}, \tag{17}$$

where $K(x) = \mathbb{I}_{[-1,1]}(x)$. This kernel is often refereed to as the rectangular kernel. Other ones are commonly proposed. What is asked is some smoothness at $x = 0$, symmetry and some integrability conditions. As formula (14) shows, K is defined up to a multiplicative constant.

As examples (up to multiplicative constants):

- Triangular kernel:
$$K(x) = (1 - |x|)\, \mathbb{I}_{[-1,1]}(x)$$
- Epanechnikov kernel
$$K(x) = (1 - x^2)\, \mathbb{I}_{[-1,1]}(x)$$
- Biweight kernel
$$K(x) = (1 - x^2)^2\, \mathbb{I}_{[-1,1]}(x)$$
- Gaussian kernel
$$K(x) = e^{-\frac{x^2}{2}}$$

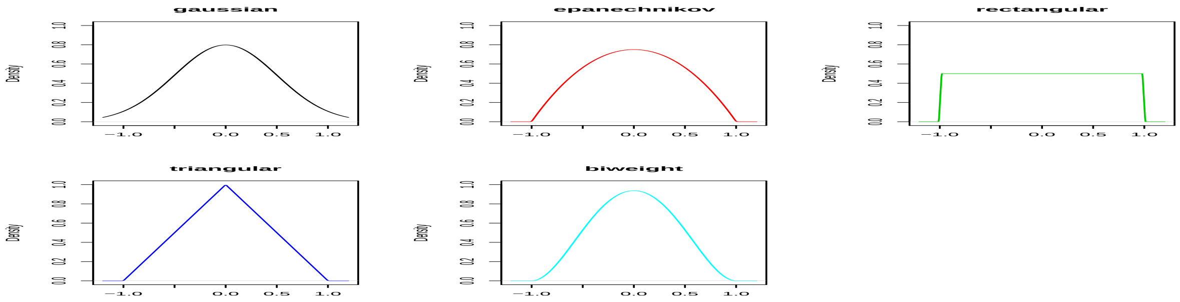

Figure 1. Graph of the kernels: [Top] Gaussian, Epanechnikov and Rectangular. [Bottom] Triangular and Biweight.

Except the rectangular kernel, they all are everywhere continuous. This is the reason why the rectangular kernel is rarely used.

Notice also that all these kernels are non negative. In chapter 4 we use non positive kernels in order to get better rates (see exercise 10).

2. Naive interpretation

The estimator, in (16), writes also

$$\hat{r}_n(y) = \sum_{j=1}^{n} Z_j W_j,$$

so it is clear that the estimator is a weighted sum of the Z_j's. The weights W_j are random positive variables and

$$\sum_{j=1}^{n} W_j = 1.$$

Now, for all the kernels proposed above, the weight $W_j = \frac{K\left(\frac{Y_j - y}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{Y_j - y}{h_n}\right)}$ indicates whether Y_j is close or not to y. The closer Y_j and y are, the larger is the weight. For the rectangular kernel, the weights simply are 0 (if the distance is too large), or 1.

To summarize, the estimator of $\mathbb{E}(Z_1|Y_1 = y)$ is a weighted sum of the Z_j's, with weights calculated according the distance between the Y_j's and y.

3. Exercises

Exercise 5. Prove the last inequality in formula (15). ⋆

Exercise 6. How the above method can be used to predict Z_n from the observation of Y_n and of the (Y_j, Z_j)'s for $j \leq n-1$?

Could you give a naive interpretation of this predictor? ⋆

Exercise 7. *Estimation of a distribution density.*

Let $X_1, \cdots, X_n$ n be i.i.d. variables having a density f. Let K be a kernel such that

- $\int K(u)du = 1$
- $\int K^2(u)du < \infty$,
- $\int |uK(u)|du < \infty$

Consider the estimator of f given by

$$\hat{f}_n(x) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{h_n} K\left(\frac{x - X_k}{h_n}\right)$$

Suppose that f is C^1 and that f and f' are bounded.

1) Prove that

- $h_n \to 0$ when $n \to +\infty$ then, for all x, $\mathbb{E}\hat{f}_n(x) \to f(x)$

- and that if $nh_n \to +\infty$ then for all x,

$$\operatorname{Var}\hat{f}_n(x) = O\left(\frac{1}{nh_n}\right)$$

2) Prove that for all x,

$$\mathbb{E}\left|\hat{f}_n(x) - f(x)\right|^2 \leq c_1 h_n^2 + \frac{c_2}{nh_n}$$

and conclude that, if $h_n \sim n^\alpha$, there is a value of α for which the rate of convergence of the quadratic risk is optimal.

This result shall be improved in the next chapter (see exercise 10). ⋆

Exercise 8. In exercise 7, take the rectangular kernel, and compare the obtained estimator with the familiar histogram. ⋆

Exercise 9. Discuss the results given by Figure 2 and Figure 3 ? What is the sensitivity of the kernel estimate to the choice of the kernels and of the bandwidths?

Explain why you could have guessed your conclusions from the results of this chapter (and of the following ones!).

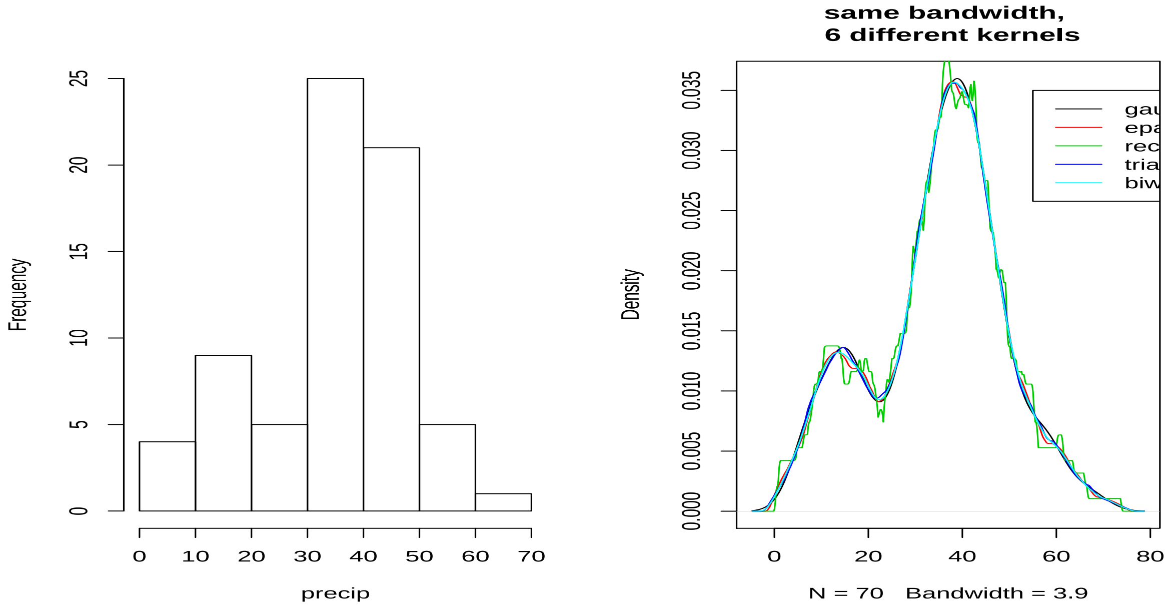

Figure 2. [Left] Histogram of the average amount of precipitation (rainfall) in inches for each of 70 United States, [Right] Kernel density estimates with 5 different kernels

⋆

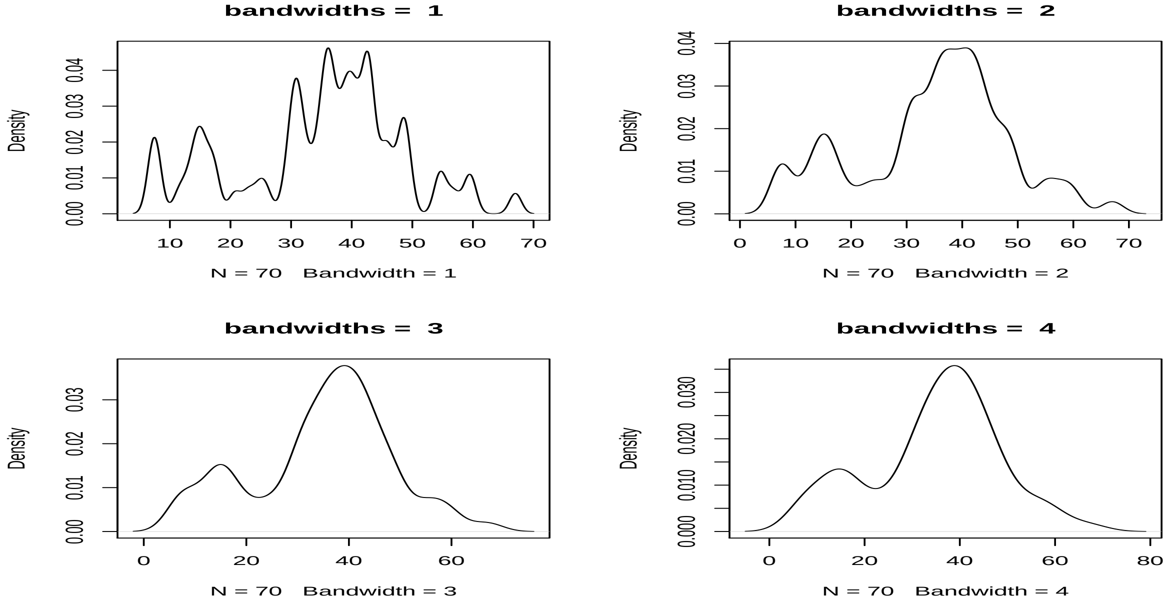

FIGURE 3. Same data set as Figure 2. Kernel density estimates with Gaussian kernel and 4 different bandwidths.

CHAPTER 4

More on functional regression estimation

1. Introduction

We consider here model (1) where only the purely regressive part is present, and where $q = 0$. Namely

$$X_k = b(e_k) + \varepsilon_k, \tag{18}$$

and we suppose that the noise $(\varepsilon_n)_{n\geq 1}$ and the exogeneous sequence $(e_n)_{n\geq 1}$ are two independent i.i.d. sequences. Recall that the question is to predict X_{n+1} from the observation of $e_{n+1}, \dots, e_1$.

1.1. The estimator. Notice first that, under the above hypotheses, the sequence $(e_k, X_k)_{k\geq 1}$ is i.i.d. Hence, the situation is exactly the same as in the preceding chapter. Here, $\mathbb{E}(X_n | e_n = e) = b(e)$, and the function $b(e)$ is estimated by

$$\hat{b}_n(e) = \frac{\sum_{j=1}^n X_j K\left(\frac{e_j - e}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{e_j - e}{h_n}\right)} \tag{19}$$

and the predictor is

$$\hat{X}_{n+1} = \hat{b}_n(e_{n+1}) \tag{20}$$

The aim is to complete the heuristic results of chapter 3. Two types of convergence shall be investigated. Section 2 is devoted to uniform almost sure convergence

$$\sup_e |\hat{b}_n(e) - b(e)| \xrightarrow{a.s.} 0, \tag{21}$$

and section 3 to the integrated quadratic error

$$\mathbb{E}\left(\int (\hat{b}_n(e) - b(e))^2 w(e) de\right).$$

In both cases, rates of convergence are given.

1.2. Assumptions. Among the following hypotheses, some are only technical (such as boundedness of b and of the noise) and could easily be released. They are chosen to shorten some proofs. Other ones (like smoothness of b) are more fundamental, as can be seen from some simulations.

- The noise and the variables e_j are two i.i.d. independent sequences

- there exists a deterministic constant m such that

$$|e_n| \leq m \quad \text{and} \quad |\varepsilon_n| \leq m \quad \forall n$$

- the exogeneous variable e_1 has a density f, strictly positive on $[-m, m]$ and C^2.
- b is C^2
- On the kernel: K is bounded, compactedly supported and

$$\int K(u)du = 1 \tag{22}$$

$$\int uK(u)du = 0, \tag{23}$$

$$\int u^2 K(u)du \neq 0. \tag{24}$$

Suppose also that there exists $\beta > 0$ and a constant γ such that

$$|K(e_1) - K(e_2)| \leq \gamma |e_1 - e_2|^\beta \quad \text{if } -m \leq e_1, e_2 \leq m. \tag{25}$$

Notice that, from the boundedness hypotheses,

$$|X_n| \leq \sup_{-m \leq e \leq m} |b(e)| + m \quad \forall n. \tag{26}$$

Notice also that the noise can't be Gaussian.

2. Uniform almost sure convergence

Theorem 4. *We consider the estimator $\hat{b}_n$ defined in (18), with a kernel satisfying assumptions (22), (23) and (24). Under the hypotheses above, and if*

$$h_n \to 0 \quad \textit{and} \quad \frac{nh_n}{\ln n} \to \infty$$

then,

$$\sup_e |\hat{b}_n(e) - b(e)| = O_{as}(h_n^2) + O_{as}\left(\sqrt{\frac{\ln n}{nh_n}}\right)$$

Let u_n and v_n be random sequences. Recall that $v_n = O_{as}(u_n)$ means that $|v_n/u_n|$ is almost surely bounded. Of course the bound may be a random variable.

Let us begin with the proof of the theorem. Then we shall give some remarks. We choose a proof largely inspired by [9]. First rewrite the estimator as:

$$\hat{b}_n(e) = \frac{\sum_{j=1}^n X_j K\left(\frac{e_j - e}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{e_j - e}{h_n}\right)} = \frac{\frac{\sum_{j=1}^n X_j K\left(\frac{e_j - e}{h_n}\right)}{nh_n}}{\frac{\sum_{j=1}^n K\left(\frac{e_j - e}{h_n}\right)}{nh_n}} =: \frac{\hat{g}_n(e)}{\hat{f}_n(e)} \tag{27}$$

It should be clear (see chapter 3 and exercise 7) that $\hat{f}_n(e)$ estimates the density $f(e)$ and that $\hat{g}_n(e)$ estimates $g(e) := b(e)f(e)$. Now, decompose the estimation error in

$$\hat{b}_n(e) - b(e) = \frac{\hat{g}_n(e)}{\hat{f}_n(e)} - \frac{g(e)}{f(e)} \quad = \quad \frac{\hat{g}_n(e) - g(e)}{\hat{f}_n(e)} + (f(e) - \hat{f}_n(e))\frac{b(e)}{\hat{f}_n(e)}$$

implying that

$$\sup_e |\hat{b}_n(e) - b(e)| \quad \leq \quad \frac{\sup_e |\hat{g}_n(e) - g(e)|}{\inf_e |\hat{f}_n(e)|} + \|b\|_\infty \frac{\sup_e |f(e) - \hat{f}_n(e)|}{\inf_e |\hat{f}_n(e)|}$$

and we treat separately the two numerators and the denominator in the following subsections.

2.0.1. *Rate of convergence of* $\sup_e |\hat{g}_n(e) - g(e)|$. We are going to prove that

Lemma 5. *With the hypotheses of Theorem 4,*

$$\sup_e |\hat{g}_n(e) - g(e)| = O_{as}\left(\sqrt{\frac{\ln n}{nh_n}}\right) + O(h_n^2) \tag{28}$$

Proof. Since

$$\hat{g}_n(e) - g(e) = \hat{g}_n(e) - \mathbb{E}(\hat{g}_n(e)) + \mathbb{E}(\hat{g}_n(e)) - g(e),$$

we shall give a bound for $\sup_e |\hat{g}_n(e) - \mathbb{E}(\hat{g}_n(e))|$, and for $\sup_e |\mathbb{E}(\hat{g}_n(e)) - g(e)|$.

- We start with $\hat{g}_n(e) - \mathbb{E}(\hat{g}_n(e))$ for fixed e. Define the i.i.d variables U_j by

$$U_j = \frac{1}{h_n}\left(X_j K\left(\frac{e_j - e}{h_n}\right) - \mathbb{E}\left(X_j K\left(\frac{e_j - e}{h_n}\right)\right)\right) \qquad j = 1, \ldots, n \tag{29}$$

From (25) and since K is bounded, it is clear that there is a constant C such that

$$|U_j| \leq C_1/h_n.$$

Secondly,

$$\mathbb{E}(U_j^2)| \leq C_2/h_n,$$

because

$$\begin{aligned}
\mathbb{E}(U_j^2) &= \frac{1}{h_n^2}\mathrm{Var}\left(X_j K\left(\frac{e_j - e}{h_n}\right)\right) \leq \frac{1}{h_n^2}\mathbb{E}\left(X_j K\left(\frac{e_j - e}{h_n}\right)\right)^2 \\
&= \frac{1}{h_n^2}\mathbb{E}\left(K^2\left(\frac{e_j - e}{h_n}\right)\mathbb{E}\left(X_j^2|e_j\right)\right) = \frac{1}{h_n^2}\mathbb{E}\left(K^2\left(\frac{e_j - e}{h_n}\right)(\sigma^2 + b^2(e_j))\right) \\
&= \frac{1}{h_n^2}\int(\sigma^2 + b^2(u))K^2\left(\frac{u - e}{h_n}\right) f(u)du \\
&= \frac{1}{h_n}\int(\sigma^2 + b^2(vh_n + e))K^2(v)f(vh_n + e)dv \leq \frac{C_2}{h_n},
\end{aligned}$$

where the last integral is obtained via the change of variables $v = (u-e)/h_n$, and the last bound from boundedness assumptions on K, b and f.

Then it is possible to apply the following key exponential inequality of Hoeffding

LEMMA 6. *Let $U_1, \dots, U_n$ be i.i.d. variables such that*

$$\mathbb{E}(U_j) = 0 \quad \text{and} \quad |U_j| \leq d.$$

Then, for every $\varepsilon \in]0, \delta^2/d[$,

$$P\left(\left|\frac{\sum_{j=1}^n U_j}{n}\right| > \varepsilon\right) \leq 2e^{-\frac{n\varepsilon^2}{4\delta^2}}$$

where δ^2 is any real number such that $\mathbb{E}(U_i^2) \leq \delta^2$

Applying this lemma to the variables U_j defined in (28), with $d = \delta^2 = \frac{C}{h_n}$ gives, for $0 < \varepsilon < 1$,

$$P\left(|\hat{g}_n(e) - \mathbb{E}(\hat{g}_n(e))| > \varepsilon)\right) = P\left(\left|\frac{\sum_{j=1}^n U_j}{n}\right| > \varepsilon\right) \leq 2e^{-\frac{nh_n\varepsilon^2}{4C}} \tag{30}$$

- The result (29) concerns a fixed value of e. We have now to consider the supremum over e.

  The method is simple. Cover $[-m, m]$ by J_n intervals of length $2m/J_n$, respectively centered in $e_1, \dots, e_{J_n}$. For any function ϕ, write

$$\phi(e) = \phi(e_{j(e)}) + \phi(e) - \phi(e_{j(e)})$$

where $e_{j(e)}$ is the nearest neighbour of e among $e_1, \dots, e_{J_n}$. So,

$$\sup_{-m\leq e\leq m} |\phi(e)| \leq \max_{j=1,\dots,J_n} |\phi(e_j)| + \sup_{-m\leq e\leq m} |\phi(e) - \phi(e_{j(e)})|,$$

which in turn implies that

$$\sup_{-m\leq e\leq m} |\phi(e)| \geq \varepsilon \Longrightarrow \{\max_{j=1,\dots,J_n} |\phi(e_j)| \geq \varepsilon/2 \quad \text{or} \quad \sup_{-m\leq e\leq m} |\phi(e) - \phi(e_{j(e)})| \geq \varepsilon/2\}$$

Let's apply this to $\phi(e) = \hat{g}_n(e) - \mathbb{E}(\hat{g}_n(e))$. We have, using inequality (6)

$$\begin{aligned} P\left(\max_{j=1,\dots,J_n} |\hat{g}_n(e_j) - \mathbb{E}(\hat{g}_n(e_j))| \geq \varepsilon/2\right) &\leq \sum_{j=1}^{J_n} P\left(|\hat{g}_n(e_j) - \mathbb{E}(\hat{g}_n(e_j))| \geq \varepsilon/2\right) \\ &\leq 2J_n e^{\frac{-nh_n\varepsilon^2}{C_1}}. \end{aligned}$$

Then, noticing that for every e, $|e - e_{j(e)}| \leq m/J_n$ and using the Lipschitz property of the kernel (see (24)),

$$\sup_{-m\leq e\leq m} \left|\hat{g}_n(e) - \mathbb{E}(\hat{g}_n(e)) - \hat{g}_n(e_{j(e)}) + \mathbb{E}(\hat{g}_n(e_{j(e)}))\right| \leq \frac{C_2}{J_n^\beta h_n^{1+\beta}} \tag{31}$$

Now, chose J_n such that

$$\sqrt{\frac{nh_n}{\ln n}}\frac{1}{h_n^{1+\beta}} = o(J_n^\beta) \tag{32}$$

For such a choice, the first member of (30) is smaller than $\varepsilon_0\sqrt{\frac{\ln n}{nh_n}}$, at least for n large enough.

So, for n large enough,

$$\begin{aligned} & P\left(\sup_{-m\leq e\leq m}|\hat{g}_n(e)-\mathbb{E}(\hat{g}_n(e))| > \varepsilon_0\sqrt{\frac{\ln n}{nh_n}}\right) \leq P\left(\max_{j=1,\ldots,J_n}|\hat{g}_n(e_j)-\mathbb{E}(\hat{g}_n(e_j))| > \varepsilon_0\sqrt{\frac{\ln n}{nh_n}}\right) \\ + \quad & P\left(\sup_{-m\leq e\leq m}\left|\hat{g}_n(e)-\mathbb{E}(\hat{g}_n(e))-\hat{g}_n(e_{j(e)})+\mathbb{E}(\hat{g}_n(e_{j(e)}))\right| > \varepsilon_0\sqrt{\frac{\ln n}{nh_n}}\right) \\ = \quad & P\left(\max_{j=1,\ldots,J_n}|\hat{g}_n(e_j)-\mathbb{E}(\hat{g}_n(e_j))| > \varepsilon_0\sqrt{\frac{\ln n}{nh_n}}\right) \leq 2J_n e^{-\frac{\varepsilon_0^2\ln n}{C_1}} = 2J_n n^{-\frac{\varepsilon_0^2}{C_1}} \end{aligned}$$

To finish with, take $J_n = n^\beta$, and ε_0 large enough in order to obtain $\sum n^{\beta-\frac{\varepsilon_0^2}{C_1}} < \infty$, implying, via Borel Cantelli lemma, that almost surely,

$$\sup_{-m\leq e\leq m}|\hat{g}_n(e)-\mathbb{E}(\hat{g}_n(e))| \leq \varepsilon_0\sqrt{\frac{\ln n}{nh_n}}$$

holds for n large enough. This proves that

$$\sup_{-m\leq e\leq m}|\hat{g}_n(e)-\mathbb{E}(\hat{g}_n(e))| = O_{as}\left(\sqrt{\frac{\ln n}{nh_n}}\right) \tag{33}$$

which is the first part in the right hand member of (27).

- We turn now to $\mathbb{E}(\hat{g}_n(e)) - g(e)$, the so-called bias term.

  From the definition of $\hat{g}_n$ and from stationarity,

$$\mathbb{E}(\hat{g}_n(e)) - g(e) = \frac{1}{h_n}\mathbb{E}\left[X_1 K\left(\frac{e_1-e}{h_n}\right)\right] - b(e)f(e).$$

Then replacing X_1 by its conditional expectation $\mathbb{E}(X_1|e_1) = b(e_1)$,

$$\begin{aligned} \mathbb{E}(\hat{g}_n(e)) - g(e) &= \frac{1}{h_n}\mathbb{E}\left[b(e_1)K\left(\frac{e_1-e}{h_n}\right)\right] - b(e)f(e) \\ &= \frac{1}{h_n}\int b(u)K\left(\frac{u-e}{h_n}\right)f(u)du - b(e)f(e) \\ &= \int (b(vh_n+e)f(vh_n+e) - b(e)f(e))K(v)dv \end{aligned}$$

where the last line comes via the change of variable $(u-e)/h_n = v$ and from (21). Now, since b and f are C^2, so is the product bf and

$$b(vh_n+e)f(vh_n+e) = b(e)f(e) + vh_n[bf]'(e) + (vh_n)^2\psi_n(v,e)$$

where $\psi_n(v,e)$ is uniformly bounded with respect to n, e and v, because the second derivative of bf is continuous and the domain of the variables is compact.

Finally, remembering (22)

$$\sup_e |\mathbb{E}(\hat{g}_n(e)) - g(e)| \quad = \quad h_n^2 \sup_e \left|\int \psi_n(v,e)v^2K(v)dv\right| \leq Ch_n^2 \int v^2K(v)dv.$$

□

2.0.2. *Rate of convergence of* $\sup_{-m\leq e\leq m}\left|\hat{f}_n(e) - \mathbb{E}(\hat{f}_n(e))\right|$.

Since $\hat{f}_n(e)$ has the same form as $\hat{g}_n(e)$ (simply replace X_i by 1), it is not so difficult to understand that the same sort of technical proof as for (27) above leads to the following result, whose proof is left to the reader.

LEMMA 7. *Under the hypotheses of theorem 4, as* $n \to \infty$

$$\sup_e |\hat{f}_n(e) - f(e)| = O_{as}\left(\sqrt{\frac{\ln n}{nh_n}}\right) + O(h_n^2) \tag{34}$$

2.0.3. *A lower bound for* $\inf_e |\hat{f}_n(e)|$.

Being C^2 and strictly positive on $[-m,m]$, f has a non zero lower bound

$$\inf_e f(e) = i > 0.$$

Then, writing $f(e) = \hat{f}_n(e) + f(e) - \hat{f}_n(e)$ gives, for all e

$$i \leq f(e) = |f(e)| \leq |\hat{f}_n(e)| + \sup_e |\hat{f}_n(e) - f(e)|$$

and consequently from (33),

$$\inf_e |\hat{f}_n(e)| \geq i - O_{as}\left(\sqrt{\frac{\ln n}{nh_n}}\right) - O(h_n^2)$$

proving that almost surely $\inf_e |\hat{f}_n(e)| \geq i/2$ for n large enough.

Collecting the results of the three subsections concludes the proof of Theorem 4.

REMARK 5. Forgetting the technical details, the reader can notice that two types of rates are obtained all along this proof

- rates like h_n^2 arise from bias terms $\mathbb{E}(\hat{g}_n) - g$ or $\mathbb{E}(\hat{f}_n) - f$
- rates like $\sqrt{\frac{\ln n}{nh_n}}$ arise from $\hat{g}_n - \mathbb{E}(\hat{g}_n)$ or $\hat{f}_n - \mathbb{E}(\hat{f}_n)$, dispersions of the estimators from their expectations.

It is interesting to note again that (see also chapter 3, section1.4) the smoothing parameter h_n plays antagonistic roles in the bias and in the dispersion. **Large h_n increases the bias and decreases the dispersion.**

2.1. Optimal rate. Suppose that $h_n \sim c\left(\frac{n}{lnn}\right)^\beta$ for some negative β. Then, the best rate of convergence to zero of the bound

$$O_{as}(h_n^2) + O_{as}\left(\sqrt{\frac{\ln n}{nh_n}}\right) = O_{as}\left(\frac{ln}{n}\right)^{-2\beta} + O_{as}\left(\frac{\ln n}{n}\right)^{(1+\beta)/2}$$

is obtained for $-2\beta = (\beta+1)/2$, that is for $\beta = -1/5$. This is summarized in the next corollary

Corollary 8. *With the hypotheses of Theorem 4, if*

$$h_n \sim c\left(\frac{\ln n}{n}\right)^{1/5},$$

then

$$\sup_e |\hat{b}_n(e) - b(e)| = O_{as}\left(\frac{\ln n}{n}\right)^{2/5},$$

which happens to be optimal for the uniform convergence in this functional situation and when the kernel is positive (see [10]).

Remark 6. Now let us compare with the results obtained in the linear case (chapter 2). In Proposition 1, the rate of convergence of the coefficient's estimator is $1/n^\alpha$ for all $\alpha < 1/2$. So, roughly speaking, in the linear case the rate is $n^{-1/2}$ while in the non linear case it is $n^{-2/5}$. Comparing $1/2$ and $2/5$ gives a good idea of the price to pay when passing from a parametric to a non parametric estimation.

3. Integrated quadratic error

It is also interesting to consider the integrated quadratic error

$$\mathbb{E}\left(\int (\hat{b}_n(e) - b(e))^2 w(e) de\right), \tag{35}$$

where w is a positive function (for example it can be the density f). We just give the result:

Proposition 9. *Under the assumptions of Theorem 4, if the weight w is bounded and compactedly supported*

$$\mathbb{E}\left(\int (\hat{b}_n(e) - b(e))^2 w(e) de\right) = O(h_n^4) + O\left(\frac{1}{nh_n}\right)$$

Remark 7. Compared to Theorem 4, there is no logarithmic factor in the second term. The result is better than what is obtained by directly replacing $(\hat{b}_n(e) - b(e))^2$ by $\sup_e |\hat{b}_n(e) - b(e)|^2$ in the integral, and using the bound in Theorem 4.

REMARK 8. It is worth noticing that, if $h_n \sim n^{\beta}$, the optimal value of β is $-1/5$, the optimal rate of the right hand side is $n^{-4/5}$. Hence, up to a logarithmic factor, we obtain the same optimal rate of the error as in the preceding remark.

4. Illustration

We illustrate the properties of the estimate (18) on different simulated data sets.

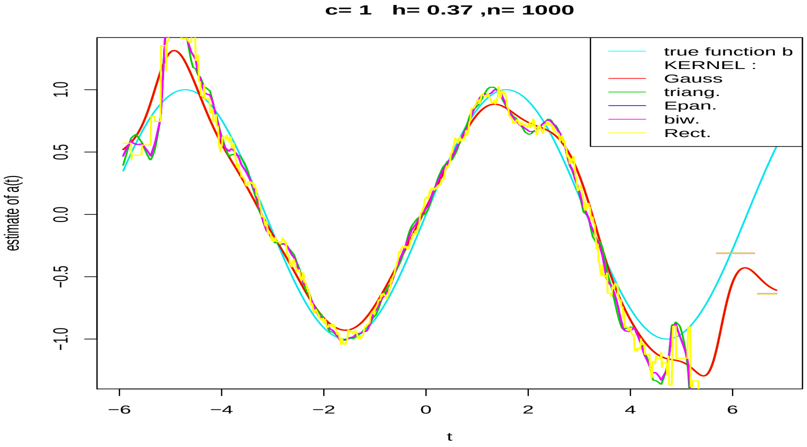

FIGURE 1. The model is defined by $b(e) = \sin(e)$, (e_n) are iid from a Gaussian $\mathcal{N}(0,4)$ and a Gaussian noise $\mathcal{N}(0,1)$. The sample size is $n = 1000$ and the bandwidth $h_n = 0.37$

As shown Fig 1, the choice of the kernel has few effects on the convergence properties of the estimate of b, except the rectancular kernel which provide a less regular estimate.

Hereafter

We only consider the case of the Gaussian kernel and we evaluate the effects of the bandwidth h_n. According to the theoretical result we take h_n of the form $C(\log(n)/n)^{1/5}$ for different values of C.

4.1. Presentation.

The following pictures provide

- The set of point (e_i, X_i) and the histogram of both series (X_i) and (e_i)
- The kernel estimate for the sample size $n = 500,\ 5000$ and the constant $C = 0.1,\ 0.5,\ 1,\ 2$.
- Figures 2, 3 and 4 : the model is defined by $b(e) = \sin(e)$
  - Fig. 2 and Fig.3 : the random variables (e_n) are iid from the Gaussian $\mathcal{N}(0,2)$ and the noise is Gaussian with variance equal to 1 (Fig. 2) and 4 (Fig. 3)
  - Fig. 4: the random variables (e_n) are iid from the uniform distribution on $[-2,2]$ and the noise is Gaussian $\mathcal{N}(0,1)$

- Figure 5 : the model is defined by $b(e) = 2 * \text{sign}(e)$, (e_n) are iid from a uniform distribution on $(-2\pi, 2\pi)$ and a Gaussian noise $\mathcal{N}(0, 9)$.
- Figure 6 : the model is defined by $b(e) = -2e\mathbb{I}_{[0,1]}(e) + 2e\mathbb{I}_{[-1,0]}(e)$, (e_n) are iid from a uniform distribution on $(-2\pi, 2\pi)$ and a Gaussian noise $\mathcal{N}(0, 1)$.

4.2. Comments. The main features to be noticed as illustrating the theory are the following:

4.2.1. *Influence of h_n.* Too small values of the smoothing parameter lead to small bias and large variance, while too large values lead to *oversmoothing*, that is small variance and bad bias.

4.2.2. *Influence of the constant.* In all the examples the chosen rate is the optimal rate $(\ln n/n)^{2/5}$, multiplied by a constant c. In view of the preceeding comment, for a fixed n, the value of c is important.

4.2.3. *Influence of the law of X_n.* The histogram of the values X_j is depicted on the top graphic in each page. Since there are less observations on the tails of the histogram, the function b is badly estimated in these zones. Keeping this in mind, compare Figures 1 and 2 with the other ones.

4.2.4. *Smoothness of b.* See exercise 14 below.

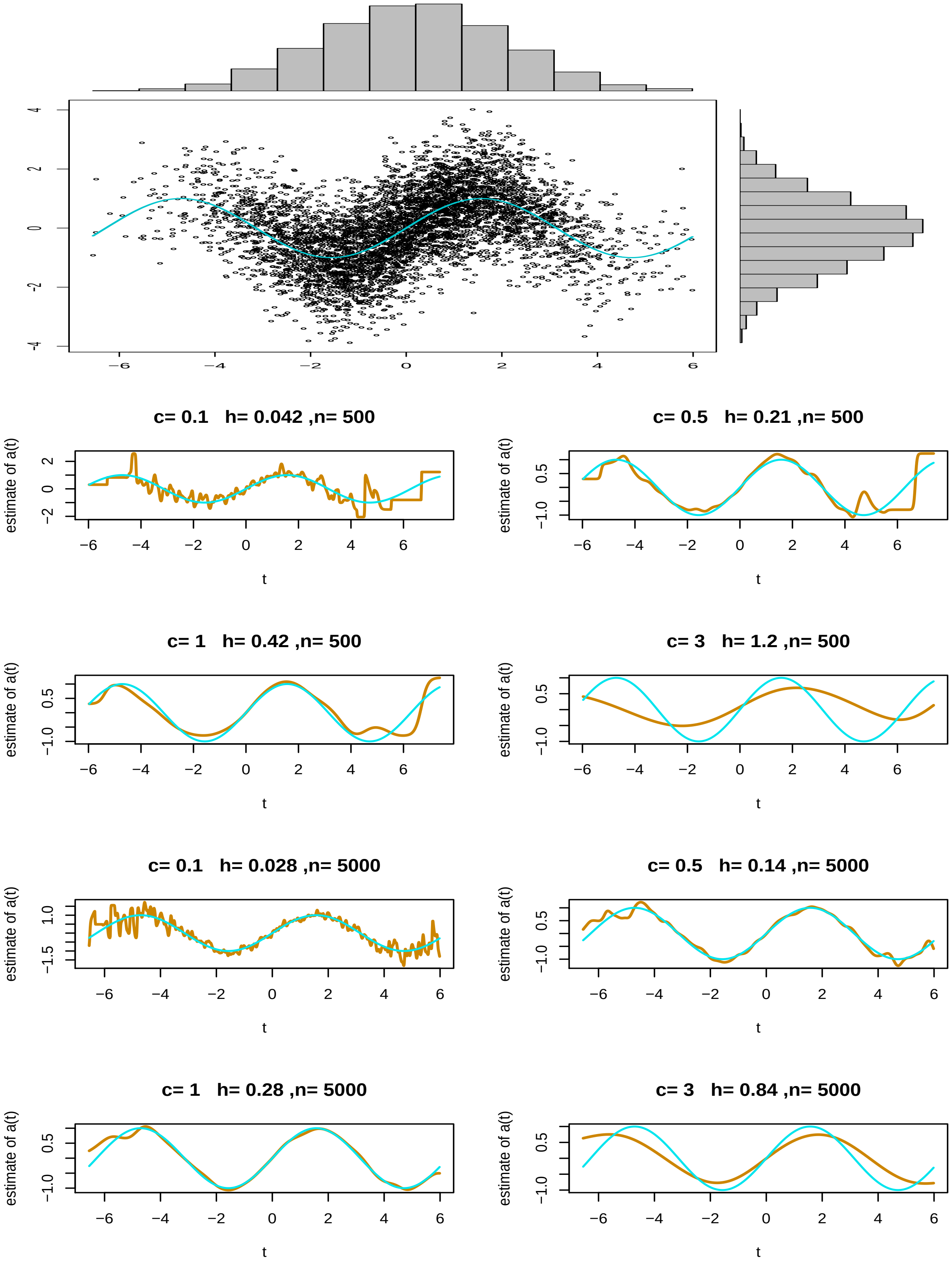

FIGURE 2. The model is defined by $b(e) = \sin(e)$, (e_n) are iid from a Gaussian $\mathcal{N}(0,4)$ and a Gaussian noise $\mathcal{N}(0,1)$.

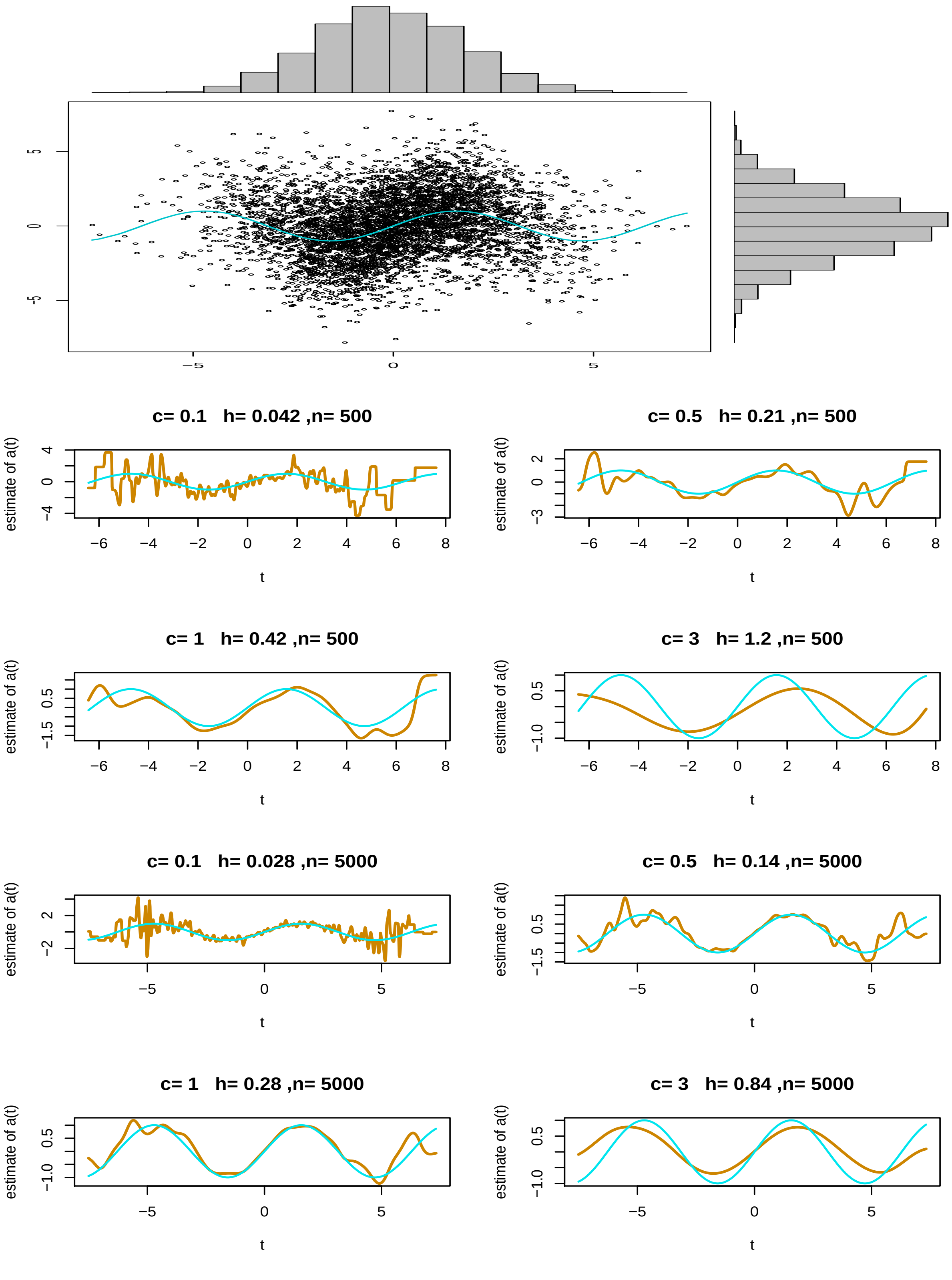

FIGURE 3. The model is defined by $b(e) = \sin(e)$, (e_n) are iid from a Gaussian $\mathcal{N}(0,4)$ and a Gaussian noise $\mathcal{N}(0,4)$.

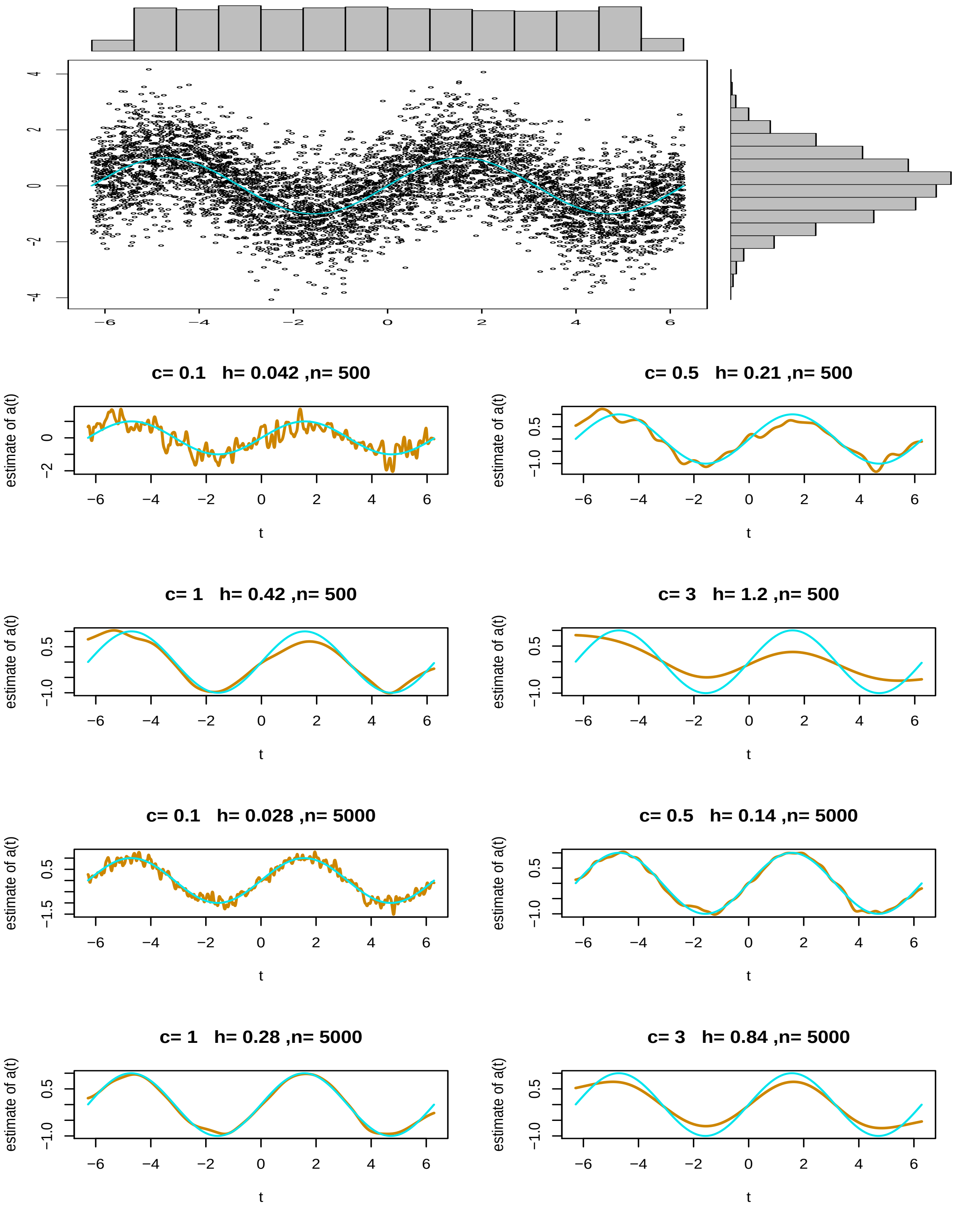

FIGURE 4. The model is defined by $b(e) = \sin(e)$, (e_n) are iid from a uniform distribution on $(-2\pi, 2\pi)$ and a Gaussian noise $\mathcal{N}(0, 1)$.

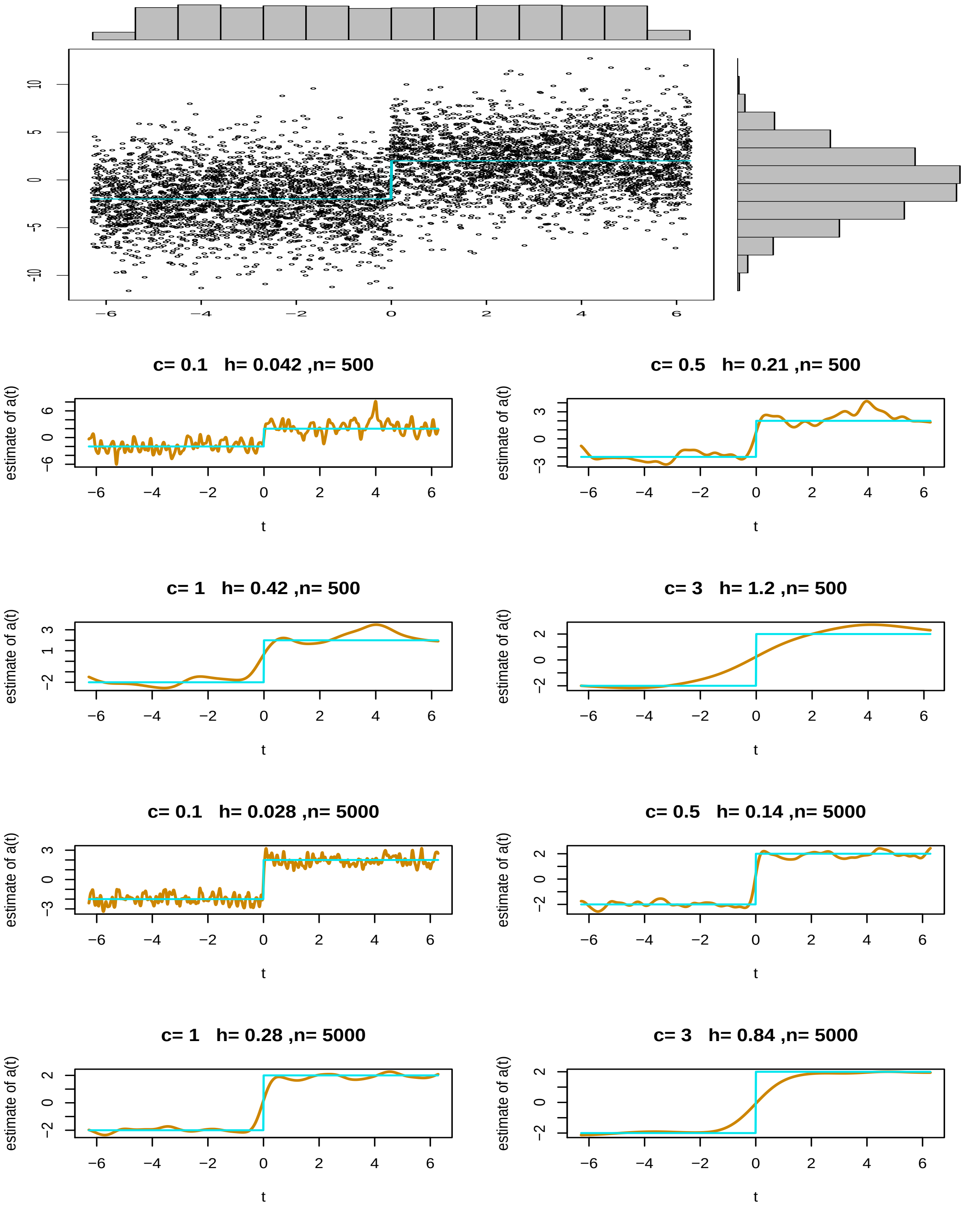

FIGURE 5. The model is defined by $b(e) = 2 * \text{sign}(e)$, (e_n) are iid from a uniform distribution on $(-2\pi, 2\pi)$ and a Gaussian noise $\mathcal{N}(0, 9)$.

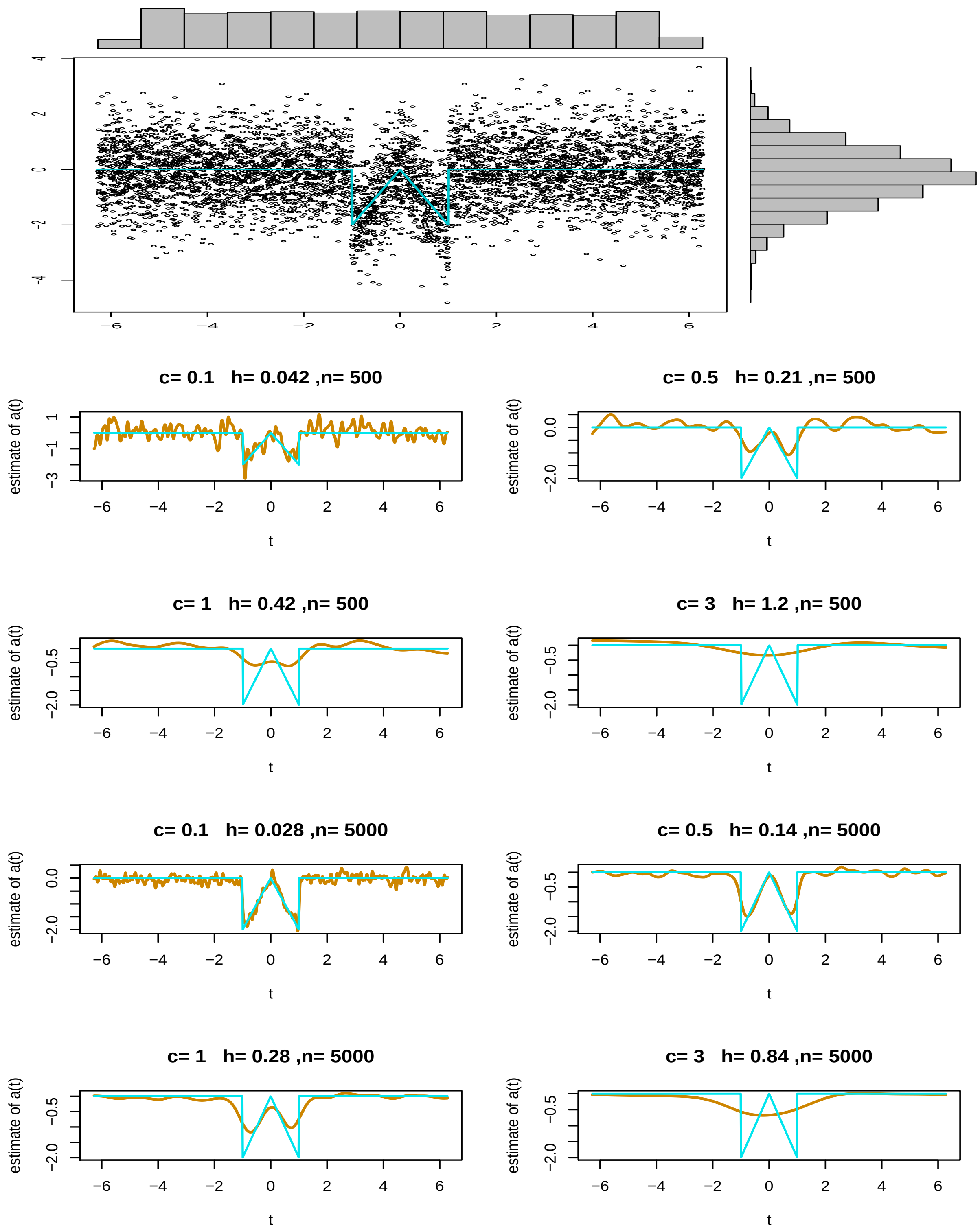

FIGURE 6. The model is defined by $b(e) = -2e\mathbb{I}_{[0,1]}(e) + 2e\mathbb{I}_{[-1,0]}(e)$, (e_n) are iid from a uniform distribution on $(-2\pi, 2\pi)$ and a Gaussian noise $\mathcal{N}(0,1)$.

5. Forecasting

Recall that the problem consists in predicting X_{n+1} from the observed values $(X_n, \ldots, X_1, e_{n+1}, \ldots, e_1)$. In the model (17), taking into account the fact that e_{n+1} is independent of $(X_n, \ldots, X_1, e_n, \ldots, e_1)$,

$$\mathbb{E}(X_{n+1}|X_n, \ldots, X_1, e_{n+1}, \ldots, e_1) = \mathbb{E}(X_{n+1}|e_{n+1}) = b(e_{n+1}).$$

So, the optimal predictor is $b(e_n)$. As in general the function b is unknown, we replace it by the estimator (18), and take

$$\hat{X}_{n+1} = \hat{b}_n(e_{n+1}).$$

The forecasting error is

$$X_{n+1} - \hat{X}_{n+1} = \varepsilon_{n+1} + b(e_{n+1}) - \hat{b}_n(e_{n+1}).$$

5.1. Theoretical forecasting error.

From the uniform convergence result of Theorem 4:

Proposition 10. *With the same assumptions as in Theorem 4,*

$$X_{n+1} - \hat{X}_{n+1} = \varepsilon_{n+1} + T_n,$$

where ε_{n+1} and T_n are independent and where, as $n \to \infty$,

$$T_n = O_{as}(h_n^2) + O_{as}\left(\sqrt{\frac{\ln n}{nh_n}}\right).$$

5.2. How to build the forecasting interval?

Proposition10 implies that the distribution of forecasting error converges to the law of the noise. If the statistician knows this law, he can, neglecting the estimation error T_n, take as forecasting interval

$$[\hat{X}_{n+1} + Q_\alpha \ , \ \hat{X}_{n+1} + Q_{1-\alpha}]$$

where Q_α and $Q_{1-\alpha}$ are the two quantiles of order α and $1-\alpha$ of the law of ε_1.

Unfortunately, the distribution of ε_1 is generally unknown and the quantiles are to be estimated. The following consequence of Corollary 8 and of Proposition 10 gives a method

Corollary 11. *Under assumptions of Proposition 10, denoting by F_ε the marginal distribution function of ε,*

$$\sup_u \left| \frac{\sum_{j=1}^n \mathbb{I}_{]-\infty,u]}(X_j - \hat{b}_n(e_j))}{n} - F_\varepsilon(u) \right| \xrightarrow{a.s.} 0.$$

Proof. For every fixed j and u, from Proposition 10, $X_{j+1} - \hat{b}_n(e_{j+1}) \xrightarrow{a.s.} 0$ as $n \to \infty$. Since the noise has a marginal density, $P(\varepsilon_j = u) = 0$. Hence, $\mathbb{I}_{]-\infty,u]}(X_j - \hat{b}_n(e_j)) - \mathbb{I}_{]-\infty,u]}(\varepsilon_j) \xrightarrow{a.s.} 0$, which in turn implies that

$$\frac{\sum_{j=1}^n \mathbb{I}_{]-\infty,u]}(X_j - \hat{b}_n(e_j)) - \mathbb{I}_{]-\infty,u]}(\varepsilon_j)}{n} \xrightarrow{a.s.} 0.$$

Then, by the law of large numbers applied to the noise,

$$\frac{\sum_{j=1}^{n} \mathbb{I}_{]-\infty,u]}(\varepsilon_j)}{n} \xrightarrow{a.s.} F_\varepsilon(u),$$

leading to

$$\frac{\sum_{j=1}^{n} \mathbb{I}_{]-\infty,u]}(X_j - \hat{b}_n(e_j))}{n} \xrightarrow{a.s.} F_\varepsilon(u).$$

The uniform convergence is a consequence of the fact that we deal with distribution functions. □

This corollary means that the statistician can treat the sample of prediction errors as a sample of estimated ε_j and use it to estimate the law of the noise. As this law is also the limit law of the forecasting error, the estimated quantiles $\hat{Q}_{n,\alpha}$ and $\hat{Q}_{n,1-\alpha}$ can be used to build a forecast interval of asymptotic level α

$$[\hat{X}_{n+1} + \hat{Q}_{n,\alpha}\ ,\ \hat{X}_{n+1} + \hat{Q}_{n,1-\alpha}].$$

6. Increasing the memory

We now consider models of the form

$$X_k = b(e_k, \dots, e_{k-q+1}) + \varepsilon_k.$$

Now we have to estimate a function of q variables $b(e^{(1)}, \dots, e^{(q)})$. The more natural idea is to replace in (18) the index measuring the distance between e_j and e by the distance between the two vectors

$$\underline{e}_{j-q+1}^{j} := {}^t(e_j, \dots, e_{j-q+1}) \quad \text{and} \quad \underline{e} := {}^t(e^{(1)}, \dots, e^{(q)}),$$

and estimate $b(e_1, \dots, e_q) = b(\underline{e})$ by

$$\hat{b}_n(\underline{e}) = \frac{\sum_{j=1}^{n} X_j K\left(\left\| \frac{\underline{e}_{j-q+1}^{j} - \underline{e}}{h_n} \right\|_2\right)}{\sum_{j=1}^{n} K\left(\left\| \frac{\underline{e}_{j-q+1}^{j} - \underline{e}}{h_n} \right\|_2\right)}.$$

Remark 9. Recall the naive interpretation of the preceding chapter (section2). The estimator is a weighted sum of the observations, each X_j having a small or large weight according to the distance of its immediate past of length q from the fixed block $(e_1, \dots, e_q)$.

In this situation, and with the same hypotheses as in the previous sections (some of them have to be adapted because now b is a function of several variables) if the smoothing parameter has the form $h_n \sim L_1(n) n^{-1/(q+4)}$ where L_1 is a logarithmic function, then

$$\sup_{\underline{e}} |\hat{b}_n(\underline{e}) - b(\underline{e})| = O_{as}\left(\frac{L_2(n)}{n^{2/(q+4)}}\right) \tag{36}$$

where L_2 is another logarithmic function. For the proof, for details on the hypotheses and on the functions L_1 and L_2 see [2].

Remark 10. For $q = 1$ we get back to the previous sections. As q increases, $2/(q+4)$ decreases and, since the bound in (35) is optimal, the rate of convergence really decreases. As a result, **the quality of estimation is rapidly deteriorating for dimensions** $q > 1$.

One of the methods aiming to remedy this so-called "curse of dimensionality" consists in adopting additive models such as

$$X_k = \sum_{j=1}^{q} b_j(e_{k-j+1}) + \varepsilon_k,$$

models for which we have to estimate q functions of one variable instead of one function of q variables (see[11]).

7. Exercises

Exercise 10. Suppose that b and f are C^k (for some $k > 2$) and that

$$\begin{aligned} \int u^j K(u)du &= 0 \quad j = 1, \dots, k-1 \\ \int u^k K(u)du &\neq 0, \end{aligned}$$

(which implies of course that K can take negative values). Prove that

$$\mathbb{E}(\hat{g}_n(e)) - g(e) = O(h_n^k),$$

and give the best rate of convergence of the estimator $\hat{b}_n$ when $h_n \sim cn^\beta$. ⋆

Exercise 11. Find symmetric, bounded and compactly supported kernels satisfying assumptions of exercise above. ⋆

Exercise 12. Use the idea of exercise 10 to improve the result of exercise 7. Compare the rates to what obtains Proposition 9. Could you give one reason for preferring positive kernels? ⋆

Exercise 13. Try to prove (at least give the main lines) the result of section 6 for the model

$$X_k = b(e_k, e_{k-1}) + \varepsilon_k$$

⋆

Exercise 14. Comment Figures 4 and 5 where the function b does not satisfy hypotheses of Theorem 4.

⋆

Exercise 15. For the models of Figures 2 to 5, give the density of X_n and comment the histograms depicted on the top of each corresponding page.

⋆

Exercise 16. Consider the additive model

$$X_k = b_1(e_k) + b_2(e_{k-1}) + \varepsilon_k,$$

with i.i.d noise and i.i.d (e_k).

(1) Notice that you have to suppose that either $\mathbb{E}(b_1(e_k)) = 0$ or $\mathbb{E}(b_2(e_k)) = 0$ for the model to be identifiable. Why?
(2) Suppose that $\mathbb{E}(b_2(e_{k-1})) = 0$. Give the expression of $\mathbb{E}(X_k|e_k)$.
(3) Use this result to propose a method to estimate $b_1(x)$.
(4) And now, use the same idea to build an estimator of $b_2(x)$.
(5) What do you think of your estimators (try to give the main lines of a proof).
(6) What happens if the e_k are not independent?

⋆

CHAPTER 5

Functional autoregression models

1. Introduction

In this chapter we turn to functional autoregressive models

$$X_k = a(X_{k-1}, \dots, X_{k-p}) + \varepsilon_k \tag{37}$$

that is models (1) where the exogeneous part is missing. The problem remains the same as previously: find a good forecasting method for X_{n+1} based on the passed values $X_n, \dots, X_1$. In fact, for the sake of simplicity, we shall suppose that $p = 1$. In other words, we deal with the model

$$X_k = a(X_{k-1}) + \varepsilon_k, \qquad k \geq 2 \tag{38}$$

where (ε_k) is an i.i.d. sequence.

Suppose for the moment that X_1 is independent from the noise (ε_k). It should be clear that the sequence (X_k) is a Markov process, and that

$$\mathbb{E}(X_k | X_{k-1}, \dots, X_1) = \mathbb{E}(X_k | X_{k-1}) = a(X_{k-1}),$$

implying that the optimal forecast consists in taking

$$\tilde{X}_{n+1} = a(X_n).$$

Then, why not estimate a by a kernel method analogously to what was done in (18), and take

$$\hat{a}_n(x) = \frac{\sum_{j=1}^{n-1} X_{j+1} K\left(\frac{X_j - x}{h_n}\right)}{\sum_{j=1}^{n-1} K\left(\frac{X_j - x}{h_n}\right)} \tag{39}$$

and then plug in the value of X_n to obtain

$$\hat{X}_{n+1} = \hat{a}_n(X_n).$$

1.1. Heuristic interpretation. The same naive interpretation as for the pure autoregression can be developed. For each X_{j+1}, the estimator calculates a weight measuring the vicinity of the observation X_j just before from the fixed value x. Then the estimator is the weighted sum of the X_j's.

1.2. Theoretical difficulties. There is an important difference between the present chapter and chapters 3 and 4.

Formally, the problem is the same in all the cases: estimate $\mathbb{E}(Z_{n+1}|Y_{n+1}=y)$, using the available observations. In the two preceding chapters, the (Z_j, Y_j)'s are i.i.d. For example in the pure regression situation, we have $Z_j = Y_j$ and $X_j = e_j$ and the (X_j, e_j)'s are independent.

Here, $Z_j = X_{j+1}$ and $Y_j = X_j$, and the (X_{j+1}, X_j)'s are certainly not independent.

So it should be evident that some knowledge on the dependence between the X_j's is necessary for studying the properties of the estimator (38).

- When $a(x) = a_1 x + a_2$, you recognize the usual linear AR_1 (non centered) model,
$$X_k = a_1 X_{k-1} + a_2 + \varepsilon_k,$$
about which nearly every thing is known. In particular, it is well known that the linear equations above admit a strictly stationary solution iff $|a_1| < 1$.
- In the other cases, we give in the following section some results on the existence of a stationary solution and on its dependence structure.

2. Weak dependence of non-linear autoregressions

Without giving any proof, we refer here to several papers or books, where details and proofs can be found. For example: [6] is devoted to mixing properties, [7] treats precisely markov processes like (37) and [10] and [17] include reviews on the question of weak dependence of sequences and particularly of Markov sequences.

The main result is that, modulo ad hoc assumptions on the function a and on the noise sequence, (37) has a stationary solution, and that, for this solution, the X_j's are not dependent enough to modify the results of the preceding chapter.

The most important notion to quantify weakness of dependencies is the notion of strong-mixing.

Given a sequence $(U_n)_n$ of stationary random variables (or random vectors), denote by $\mathcal{U}_l^k$ the sigma-algebra generated by $U_l, \ldots, U_k$

Definition 1. *The strong-mixing coefficients α_n of the sequence $(U_k)_{k\geq 1}$ are defined by*

$$\alpha_n = \sup_k \sup_{A\in\mathcal{U}_0^k, B\in\mathcal{U}_{k+n}^\infty} |P(A\cap B) - P(A)P(B)|$$

Definition 2. *The sequence $(U_n)_n$ is strong mixing if*

$$\alpha_n \to 0 \quad \textit{when} \quad n\to\infty.$$

The sequence is geometrically strong mixing if the convergence to zero is as fast as the convergence of a geometric sequence, meaning that there exists $\tau \in]0,1[$ such that

$$\alpha_n \leq c\tau^n \quad \textit{for} \quad n \geq n_0$$

Remark 11. Clearly, for an i.i.d. sequence, $\alpha_n = 0$ for every $n \geq 1$.

Remark 12. Roughly speaking, in a strong mixing sequence $(U_n)_{n\geq 1}$, the dependence between U_j and U_k disappears when $|j-k|$ increases.

Remark 13. If $(U_n)_{n\in\mathbb{Z}}$ is stationary, the dependence between X_k and X_{k+n} only depends on n, so that α_n can be redefined by

$$\alpha_n = \sup_{A\in\mathcal{U}_{-\infty}^0, B\in\mathcal{U}_n^\infty} |P(A\cap B) - P(A)P(B)|$$

Remark 14. If $(U_n)_n$ is a stationary Markov sequence,

$$\alpha_n = \sup_{A\in\mathcal{U}_0, B\in\mathcal{U}_n} |P(A\cap B) - P(A)P(B)|$$

where $\mathcal{U}_k = \mathcal{U}_k^k$ is the sigma algebra generated by X_k.

Concerning model (37) we shall use the following result (see for example [6], or [7], or [17])

Theorem 12. *If $(\varepsilon_k)_{k\geq 1}$ is an i.i.d. sequence having a strictly positive marginal density, and if the function a is bounded then Markov model (37) has a strictly stationary solution $(X_k)_{k\geq 1}$, and this solution is geometrically strongly mixing.*

3. Properties of strong mixing sequences, and their consequences

3.1. Invariance. The mixing property is invariant by simple transformations. For example

Lemma 13. *If $(U_n)_{n\geq 1}$ is strong mixing, so is the sequence $(V_n = \phi(U_{n-k_1}, \dots, U_{n+k_2}))_{n\geq 1}$, where k_1 and k_2 are fixed integers and ϕ any $\mathbb{R}^p$-valued function. The rate of convergence to 0 of the mixing coefficient is the same for the two sequences.*

For example, it is easy to deduce from this Lemma that, under the assumption of Theorem 12, the sequence $(X_{k+1}, X_k)_{k\geq 1}$ is geometrically strongly mixing.

3.2. Exponential inequality. As mixing is a kind of weak dependence, it is not surprising that most classical results for i.i.d. sequences still hold with minor changes for mixing ones under a suitable rate of convergence of the mixing sequence.

As an example, take Lemma 6, which plays a key role in the proof of Theorem 4. This lemma is stated for i.i.d. sequences. There are many analogous results for mixing sequences. The following one is well fitted to our problem. See [15] for the proof.

Lemma 14. *Let V_j be a geometrically strong mixing sequence of centered bounded random variables. For any $a > 1$, $r > 1$ and $\varepsilon > 0$,*

$$P\left(\left|\sum_{j=1}^{n} V_j\right| > 4\varepsilon\right) \leq 4\left(1+\frac{\varepsilon^2}{rs_n^2}\right)^{-r/2} + 2c\frac{n}{r}\left(\frac{2r}{\varepsilon}\right)^{a}$$

where $s_n^2 = \sum_{1\leq j,k\leq n} |\mathrm{Cov}(V_j, V_k)|$

3.3. Covariances. In order to use this inequality, we shall need to evaluate s_n^2. The key result to do that concerns the link between the covariance sequence and the sequence of mixing coefficients (see [6] for other results of the same type).

Lemma 15. *Let $(V_n)_{n\geq 1}$ be a stationary sequence, and (α_n) its sequence of mixing coefficients defined in 1. Suppose that there exists a constant m such that $|V_j| \leq m$ for all j. Then*

$$|\mathrm{Cov}(V_j, V_k)| \leq 4m^2\alpha_{j-k} \quad \forall j, k$$

This inequality can be used for example to prove that, if $\alpha_n \to 0$ fastly enough, $s_n^2 \sim n$ as $n \to \infty$, that is to say that its asymptotic behaviour is (up to a multiplicative constant) the same as if the variables were i.i.d. (see exercise 18 for details).

4. Estimation of a

We proceed exactly as in chapter 4, only changing (e_k, X_k) for (X_k, X_{k+1}), as mentioned in the introduction. So, we estimate $a(x)$ by $\hat{a}_n(x)$ defined in (38).

4.1. Assumptions.

- The noise is i.i.d. and there exists a deterministic constant m such that

$$|\varepsilon_n| \leq m \quad \forall n$$

- a is bounded and C^2
- The marginal distribution of the stationary solution X_n has a density ϕ, strictly positive on $[-m, m]$ and C^2.
- For every j, k, the distribution of (X_j, X_k) has a bounded density $\phi_{j,k}$
- On the kernel: K is bounded, compactedly supported and satisfies the conditions (21),(22),(23) and (24)

Remark 15. From these hypotheses

$$|X_k| \leq \|a\|_\infty + m \quad \forall k$$

4.2. Convergence result.

The result of Theorem 4 becomes now:

THEOREM 16. *Under assumptions above,*

$$\sup_x |\hat{a}_n(x) - a(x)| = O_{as}\left(\sqrt{\frac{\ln n}{nh_n}}\right) + O(h_n^2).$$

PROOF. The proof follows the same lines as that of Theorem 4, modulo the change indicated above. We rewrite the estimator:

$$\hat{a}_n(x) = \frac{\sum_{j=1}^{n-1} X_{j+1} K\left(\frac{X_j - x}{h_n}\right)}{\sum_{j=1}^{n-1} K\left(\frac{X_j - x}{h_n}\right)} = \frac{\frac{\sum_{j=1}^{n-1} X_{j+1} K\left(\frac{X_j - x}{h_n}\right)}{nh_n}}{\frac{\sum_{j=1}^{n-1} K\left(\frac{X_j - x}{h_n}\right)}{nh_n}} =: \frac{\hat{\psi}_n(x)}{\hat{\phi}_n(x)}, \tag{40}$$

where $\hat{\phi}_n(e)$ estimates the marginal density $\phi(x)$ of X_j and where $\hat{\psi}_n(x)$ estimates

$$\psi(x) := \mathbb{E}(X_2 \, \mathbb{I}_{X_1 = x}) = a(x)\phi(x).$$

So, the estimation error is splitted into

$$\hat{a}_n(x) - a(x) = \frac{\hat{\psi}_n(x)}{\hat{\phi}_n(x)} - \frac{\psi(x)}{\phi(x)} \quad = \quad \frac{\hat{\psi}_n(x) - \psi(x)}{\hat{\phi}_n(x)} + (\phi(x) - \hat{\phi}_n(x))\frac{a(x)}{\hat{\phi}_n(x)}$$

implying that

$$\sup_x |\hat{a}_n(x) - a(x)| \quad \leq \quad \frac{\sup_x |\hat{\psi}_n(x) - \psi(x)|}{\inf_x |\hat{\phi}_n(x)|} + \|a\|_\infty \frac{\sup_x |\phi(x) - \hat{\phi}_n(x)|}{\inf_x |\hat{\phi}_n(x)|}.$$

From this point, the only modifications from the proof of theorem 4 concern inequalities (29) and (32). The basic Lemma 6 is now replaced by Lemma 14 which we apply to the variables

$$V_j := X_{j+1} K\left(\frac{X_j - x}{h_n}\right) - \mathbb{E}\left(X_{j+1} K\left(\frac{X_j - x}{h_n}\right)\right). \tag{41}$$

These variables are bounded by a constant C (see what concerns variables U_j in the proof of Lemma 27). Moreover, applying Theorem 12 and Lemma 13 shows that the sequence $(V_j)_{j \geq 1}$ is geometrically strong mixing.

Then we apply Lemma 14. Firstly we need an estimation of $s_n^2 = \sum_{1 \leq j,k \leq n} |\mathrm{Cov}(V_j, V_k)|$.

LEMMA 17. *If, as $n \to \infty$, $h_n \sim cn^{\beta_1}(\ln n)^{\beta_2}$ then*

$$s_n^2 = O(nh_n) \tag{42}$$

Let us prove the lemma. With the same kind of proofs as for the variables U_j (see again proof of Lemma 27) we obtain

$$\mathrm{Var}(V_j) \leq Ch_n \quad \forall j, \tag{43}$$

$$
\begin{aligned}
|\mathrm{Cov}(V_j, V_k)| &\leq \int K\left(\frac{u-x}{h_n}\right) K\left(\frac{v-x}{h_n}\right) \phi_{j,k}(u,v)dudv \\
&+ \left(\int K\left(\frac{u-x}{h_n}\right) \phi(u)du\right)^2 \\
&= h_n^2 \int K(u)K(v)\phi_{j,k}(h_n u - x, h_n v - x)dudv \\
&+ h_n^2 \left(\int K(u)\phi(h_n u - x)du\right)^2 = O(h_n^2) \quad \forall j \neq k
\end{aligned}
\tag{44}
$$

and, from Lemma 15

$$
|\mathrm{Cov}(V_j, V_k)| \leq 4C^2 \alpha_{j-k} \sim C_1 \tau^{|j-k|}
\tag{45}
$$

We use inequality (44) for large values of $|j-k|$, inequality (42) for the variances and inequality (43) otherwise. For a sequence δ_n to be precised,

$$
\begin{aligned}
s_n^2 &= \sum_{|j-k| \leq \delta_n} |\mathrm{Cov}(V_j, V_k)| + \sum_{|j-k| > \delta_n} |\mathrm{Cov}(V_j, V_k)| \\
&= n\mathrm{Var}(X_1) + \sum_{1 < |j-k| \leq \delta_n} |\mathrm{Cov}(V_j, V_k)| + \sum_{|j-k| > \delta_n} |\mathrm{Cov}(V_j, V_k)| \\
&\leq C_2(nh_n + n\delta_n h_n^2 + n^2 \alpha_{\delta_n}).
\end{aligned}
$$

Then we take $\delta_n = 1/(h_n \ln n)$ and obtain

$$
s_n^2 = O(nh_n + n^2 \tau^{1/(h_n \ln n)})
$$

Taking $h_n \sim cn^{\beta_1}(\ln n)^{\beta_2}$ and using the fact that $\tau^x = o(x^{-k})$ for every $k > 0$, it is easy to see that the second term is negligible compared with the first one, and the lemma is proved.

Now, from Lemma 14, together with the bound (41), we deduce for any $a > 1$, $r > 1$ and $\varepsilon > 0$,

$$
P\left(\left|\sum_{j=1}^{n} V_j\right| > 4\varepsilon\right) \leq 4\left(1 + \frac{C_3 \varepsilon^2}{rnh_n}\right)^{-r/2} + 2c\frac{n}{r}\left(\frac{2r}{\varepsilon}\right)^a,
$$

leading to

$$\begin{aligned} P\Big(|\hat{\psi}_n(x)-\psi(x)|\geq \varepsilon_0\sqrt{\frac{\ln n}{nh_n}}\Big) &= P\left(\frac{|\sum_{j=1}^n V_j|}{nh_n}\geq \varepsilon_0\sqrt{\frac{\ln n}{nh_n}}\right) = \\ = \ & P\left(\left|\sum_{j=1}^n V_j\right|\geq \varepsilon_0\sqrt{nh_n\ln n}\right)\leq 4\left(1+\frac{C_3\varepsilon_0^2 nh_n\ln n}{16rnh_n}\right)^{-r/2}+ \\ + \ & 2c\frac{n}{r}\left(\frac{2r}{\varepsilon_0\sqrt{nh_n\ln n}}\right)^a \leq 4e^{-C_4\frac{r}{2}\frac{\varepsilon_0^2\ln n}{16r}}+2c\frac{n}{r}\left(\frac{2r}{\varepsilon_0\sqrt{nh_n\ln n}}\right)^a \\ = \ & 4e^{-\frac{C_4\varepsilon_0^2\ln n}{32}}+2c\frac{n}{r}\left(\frac{2r}{\varepsilon_0\sqrt{nh_n\ln n}}\right)^a = 4n^{-\frac{C_4\varepsilon_0^2}{16}}+2c\frac{n}{r}\left(\frac{2r}{\varepsilon_0\sqrt{nh_n\ln n}}\right)^a. \end{aligned}$$

Then, take $r=n^\beta$. Remembering that $h_n\sim cn^{\beta_1}\ln n^{\beta_2}$ gives

$$P\Big(|\hat{\psi}_n(x)-\psi(x)|\geq \varepsilon_0\sqrt{\frac{\ln n}{nh_n}}\Big)\leq 4n^{-\frac{C_4\varepsilon_0^2}{16}}+\frac{2^{a+1}c}{\varepsilon_0^a}\frac{n^{1/2+b(1-a)-\beta_1/2}}{(\ln n)^{(1+\beta_1)/2}}.$$

Then, it remains to chose ε_0 large enough to have $C_4\varepsilon_0^2>16$, and a and b large enough to have $1/2+b(1-a)-\beta_1/2<-1$. Then the series $\sum_n P\Big(|\hat{\psi}_n(x)-\psi(x)|\geq \varepsilon_0\sqrt{\frac{\ln n}{nh_n}}\Big)$ converges, which implies that

$$\hat{\psi}_n(x)-\psi(x)=O_{as}\left(\sqrt{\frac{\ln n}{nh_n}}\right).$$

The rest of the proof goes similarly as for theorem 4 and is omitted. □

Remark 16. Notice that the rate of convergence is the same as in the pure regression problem. The reason is, as was already pointed out in the introduction, the weakness of dependence between the X_j's.

4.3. Optimal rate. From Proposition 16, with smoothing parameter $h_n\sim cn^{\beta_1}\ln n^{\beta_2}$,

$$\sup_x|\hat{a}_n(x)-a(x)|=O_{as}\left(n^{(1-\beta_1)/2}\ln n^{(1+\beta_2)/2)}\right)+O(n^{2\beta_1}\ln n^{2\beta_2}).$$

The optimal rate is obtained for $\beta_2=-\beta_1=1/5$. Hence

Corollary 18. *For smoothing parameters having the form $h_n\sim cn^{\beta_1}\ln n^{\beta_2}$, the optimal rate of convergence, obtained for*

$$h_n\sim c\left(\frac{\ln n}{n}\right)^{1/5},$$

is

$$\sup_x|\hat{a}_n(x)-a(x)|=O_{as}\left(\frac{\ln n}{n}\right)^{2/5}$$

5. Illustration

We illustrate the properties of the estimate (38) on different simulated data sets.

We only consider the case of the Gaussian kernel and we evaluate the effects of the bandwidth h_n. According to the theoretical result we take h_n of the form $C(\log(n)/n)^{1/5}$ for different values of C.

5.1. Presentation. The following pictures provide

- The time series (X_i) with its auto correlations function and the set of points (X_i, X_{i+1})
- The kernel estimate for the sample size $n = 500,\ 5000$ and the constant $C = 0.1,\ 0.5,\ 1,\ 2$.
- Figures 4, 1 and 2 : the model is defined by $a(x) = \sin(x)$
  - Fig. 4 :
  - Fig. 1 :
  - Fig. 2
- Figure 3 : the model is defined by $a(x) = 1/(1+x^2)$ and a Gaussian noise $\mathcal{N}(0,1)$
- Figure 5 : the model is defined by $a(x) = 2 * \mathrm{sign}(x)$ and a Gaussian noise $\mathcal{N}(0,1)$
- Figure 6 : the model is defined by $a(x) = -2x\mathbb{I}_{[0,1]}(x) + 2x\mathbb{I}_{[-1,0]}(x)$ and a Gaussian noise $\mathcal{N}(0,1)$

5.2. Comments. The same comments as in Chapter 4 can be given. We leave them to the reader. It may be interesting to look at the empirical autocorrelations given on the first line of each page, and to think of the ARMA (linear) models which could be adapted to the data.

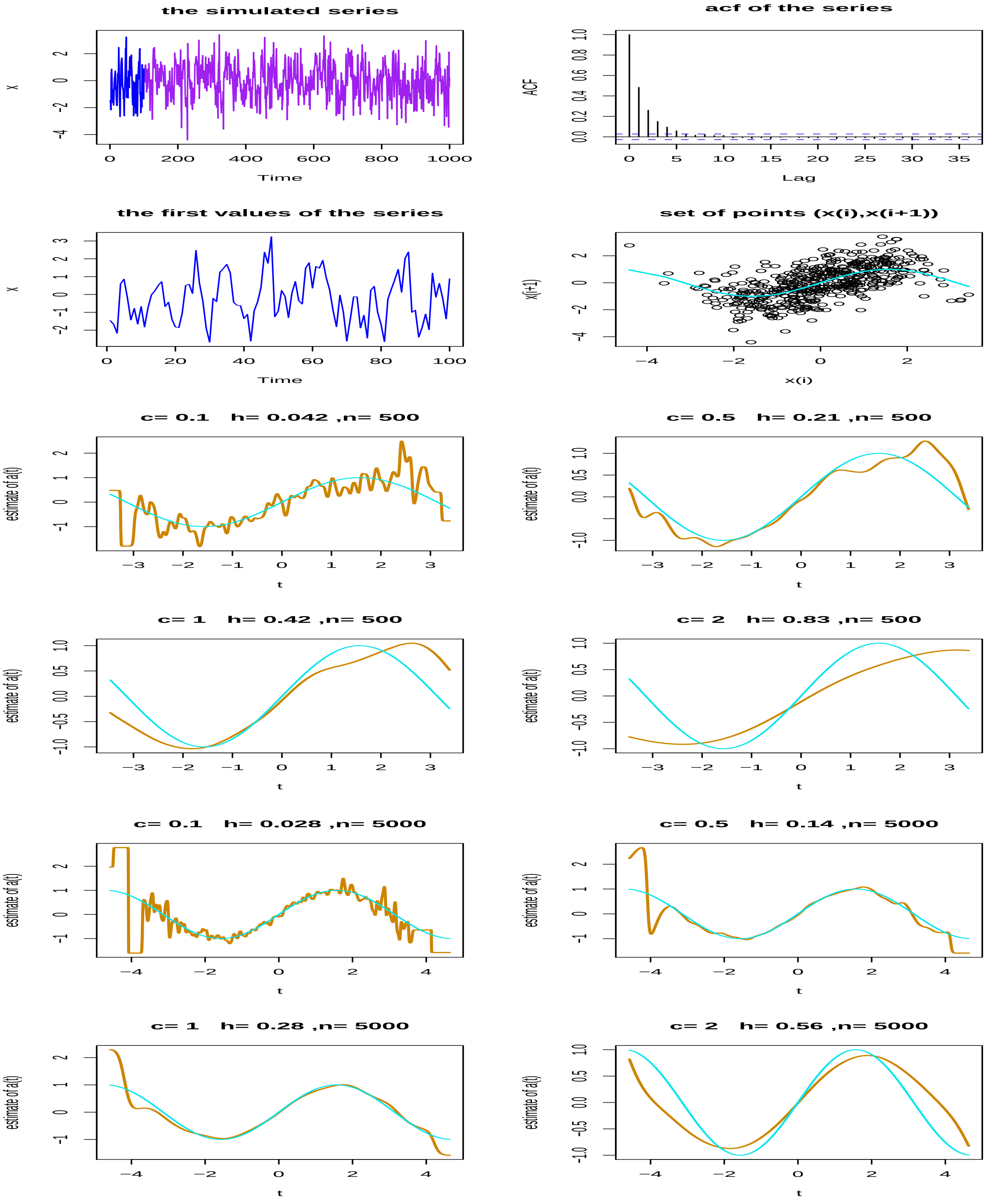

Figure 1. The model is defined by $a(x) = \sin(x)$ and a Gaussian noise $\mathcal{N}(0, 1)$.

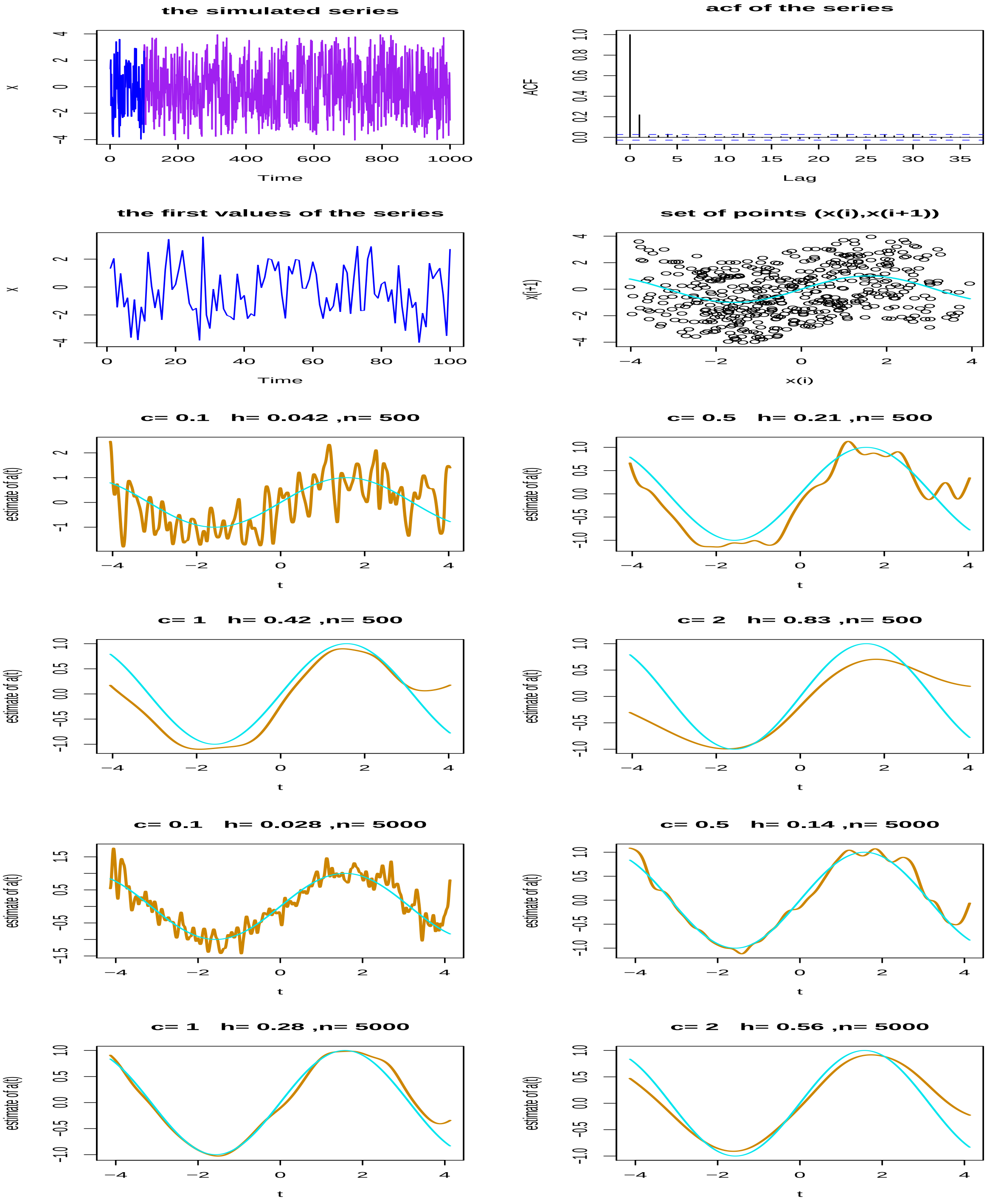

Figure 2. The model is defined by $a(x) = \sin(x)$ and a uniform noise on $(-\pi, \pi)$.

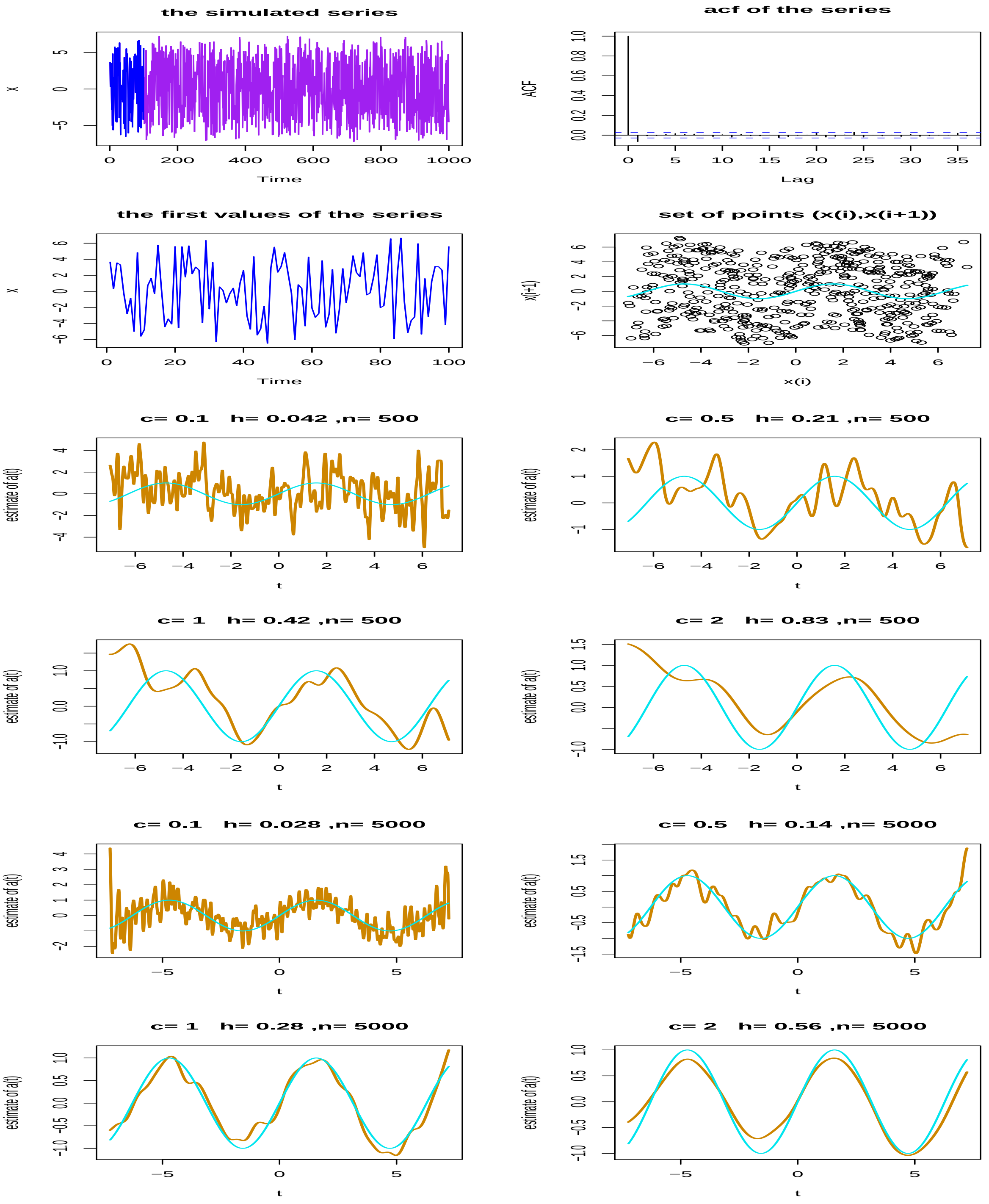

FIGURE 3. The model is defined by $a(x) = \sin(x)$ and a uniform noise on $(-2\pi, 2\pi)$.

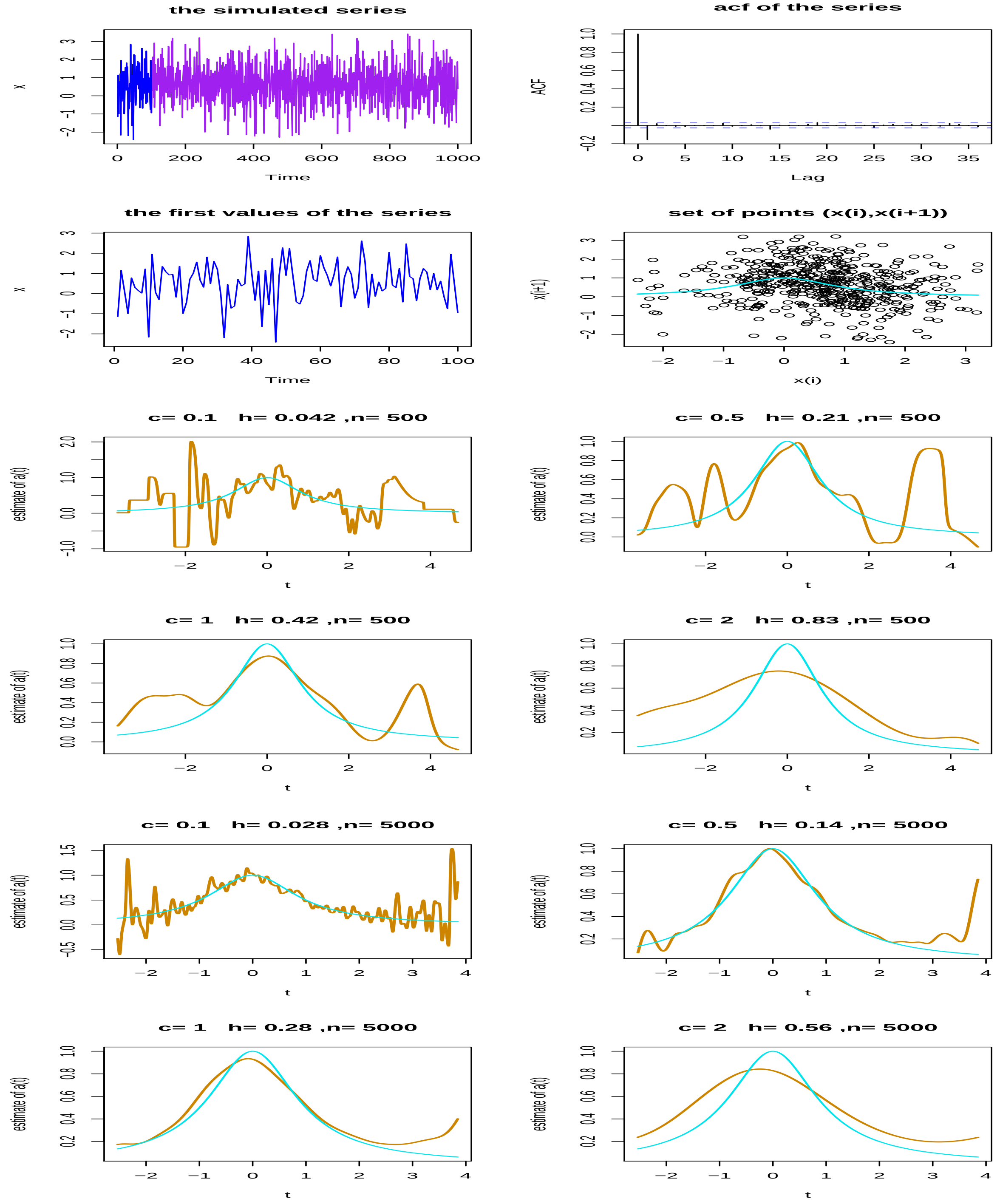

Figure 4. The model is defined by $a(x) = 1/(1+x^2)$ and a Gaussian noise $\mathcal{N}(0,1)$.

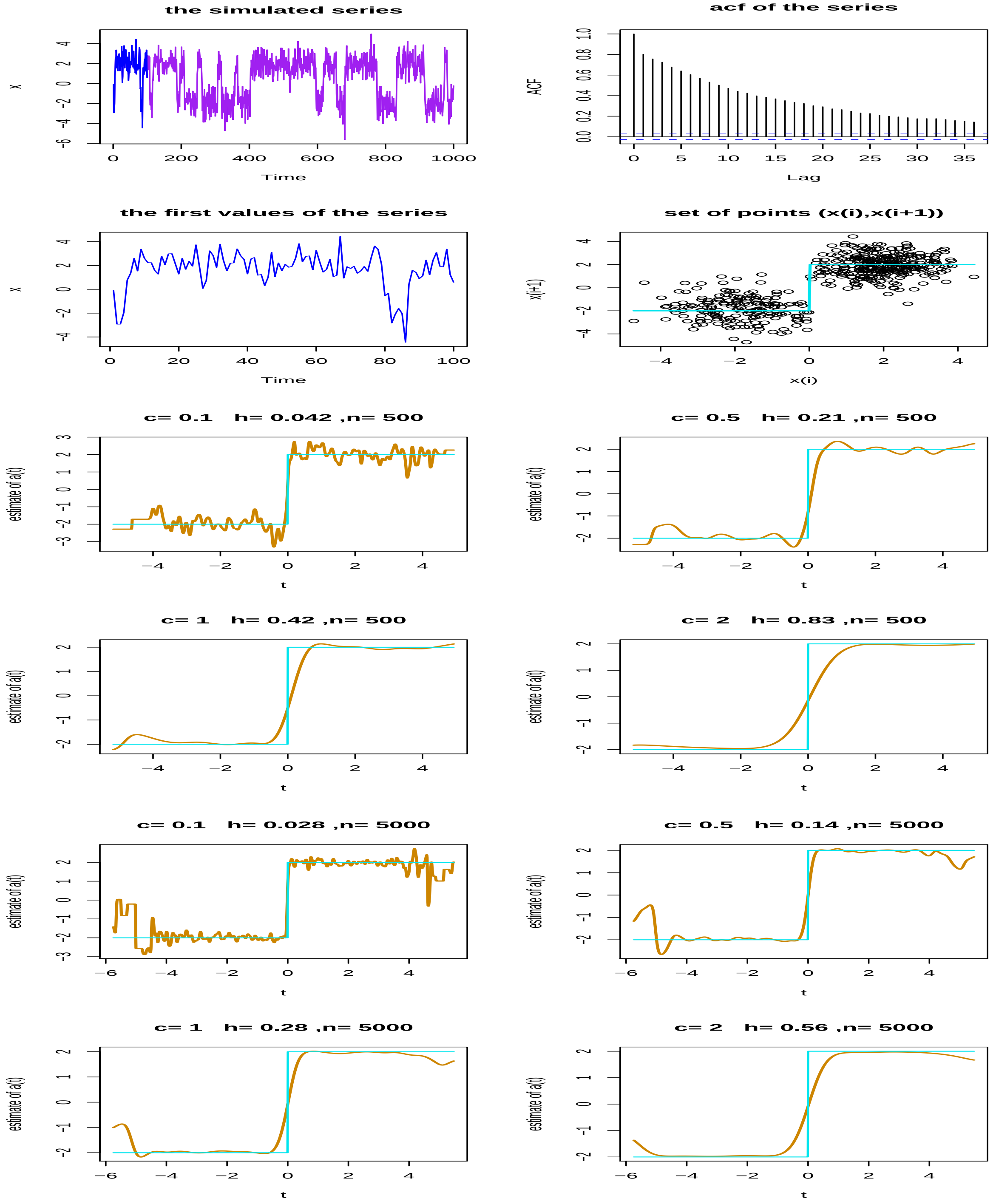

FIGURE 5. The model is defined by $a(x) = 2\text{sign}(x)$ and a Gaussian noise $\mathcal{N}(0,1)$.

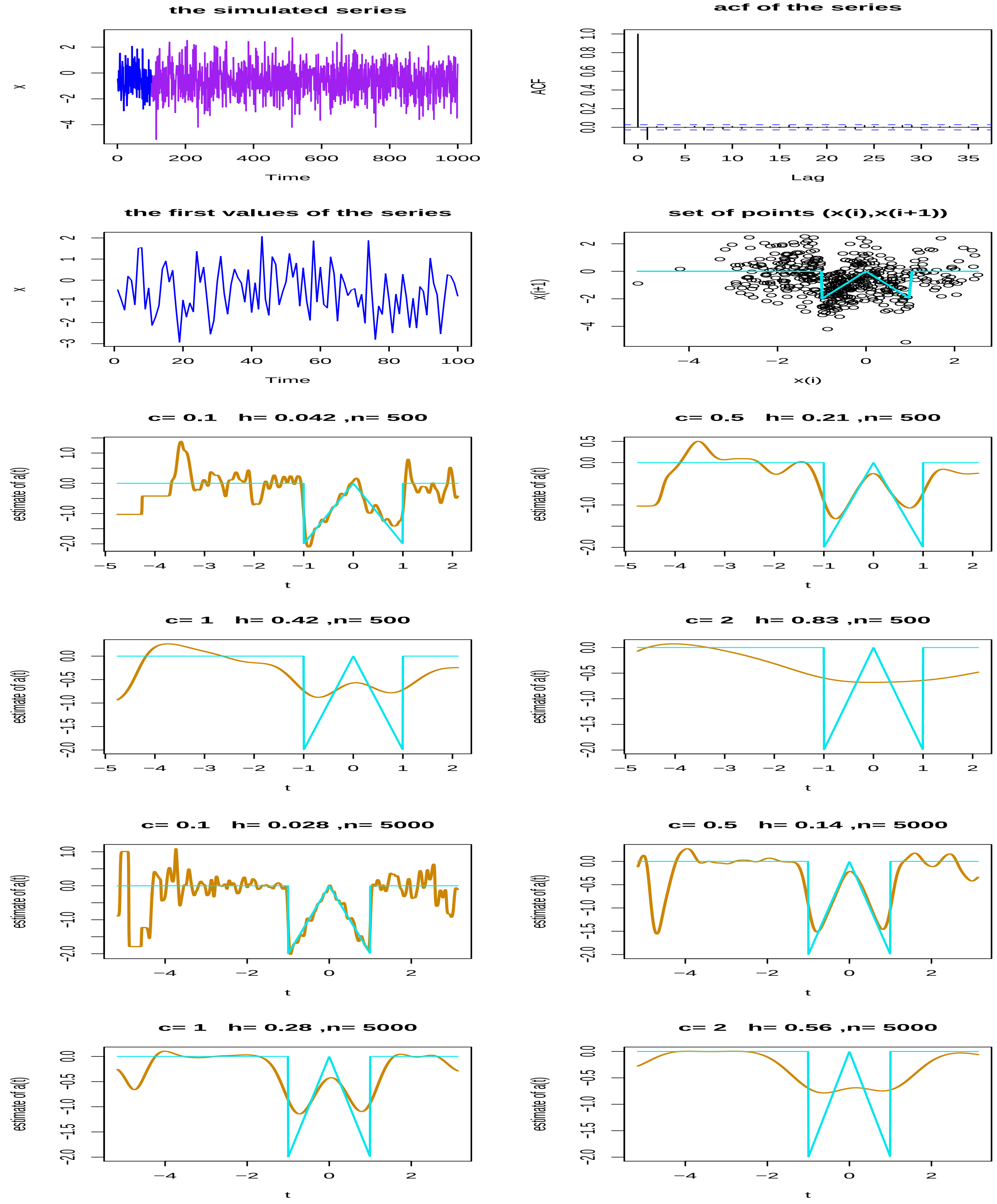

FIGURE 6. The model is defined by $a(x) = -2x\mathbb{I}_{[0,1]}(x) + 2x\mathbb{I}_{[-1,0]}(x)$ and a Gaussian noise $\mathcal{N}(0,1)$.

6. Forecasting

In order to predict X_{n+1} from the observed values $X_n, \ldots, X_1$, take

$$\hat{X}_{n+1} = \hat{a}_n(X_n),$$

with $\hat{a}_n()$ is defined in (38). From the uniform result given in Proposition 16, if h_n is chosen as in Corollary 18,

$$X_{n+1} - \hat{X}_{n+1} = \varepsilon_{n+1} + O_{as}\left(\frac{\ln n}{n}\right)^{2/5},$$

and the construction of a forecasting interval of level α is exactly the same as in the pure regression situation.

7. Increasing the memory

7.1. Theoretical results. The reason for taking $p = 1$ in the autoregressive model was only simplicity of the proofs. Modulo a few changes, the results also hold for the more general model

$$X_k = a(X_{k-1}, \ldots, X_{k-p}) + \varepsilon_k.$$

To build the estimator, it is natural to insert in (38) an index measuring the distance between the two vectors

$$X_{j-p+1}^{j} := {}^t(X_j, \ldots, X_{j-p+1}) \quad \text{and} \quad \underline{x} := {}^t(x^{(1)}, \ldots, x^{(p)}),$$

and estimate $a(x_1, \ldots, x_p) := a(\underline{x})$ by

$$\hat{a}_n(\underline{x}) = \frac{\sum_{j=1}^{n-1} X_{j+1} K\left(\left\|\frac{X_{j-p+1}^{j} - \underline{x}}{h_n}\right\|_2\right)}{\sum_{j=1}^{n-1} K\left(\left\|\frac{X_{j-p+1}^{j} - \underline{x}}{h_n}\right\|_2\right)}. \tag{46}$$

This general case is treated for example in [2] (Theorem 3.2), where it is proved that the optimal rate of convergence is

$$\hat{a}_n(\underline{x}) - a(\underline{x}) = O_{as}\left(\frac{L(n)}{n^{2/(p+4)}}\right). \tag{47}$$

The comments are the same as for the pure regressive case: the rate is a decreasing function of p. A practical consequence is that **the method behaves rather badly for autoregressions of order larger than** $p = 1$.

This is the reason why additive models

$$X_k = a_1(X_{k-1}) + \ldots + a_p(X_{k-p}) + \varepsilon_k$$

have been successfully introduced (see [11] and exercise 22 below).

7.2. Some illustrations for $p = 2$. We illustrate the properties of estimate (45) when the order of the model is equal to 2. We consider a bandwidth h_n of the form $Cn^{-1/6}$ for different values of C.

7.3. Presentation. The following pictures provide

- the set of points $(X_{i+\ell}, X_{i+k})$ with $\ell, k = 0, 1, 2$
- The kernel estimate for the sample size $n = 5000$, 10000 and 50000, and the constant $C = 0.5, 1, 3$.

7.4. Comments. The estimations are to be compared with the function $a(x, y)$ on the graphic on the right of the last line of each page. Some conclusions are evident

- The model $\sin(x) + \sin(y)$ is easier to estimate than the model $\sin(x) + 1/(1 + y^2)$ (compare for example the results for n=10000 for both models).
- As before, for a fixed sample size n, the value h_n of the smoothing parameter only depends on the constant c, and the antagonism *bias-variance* which gives chaotic estimates for small values of the constant c and oversmoothing (correct average shape, but missing contrasts) for large values is well visible.

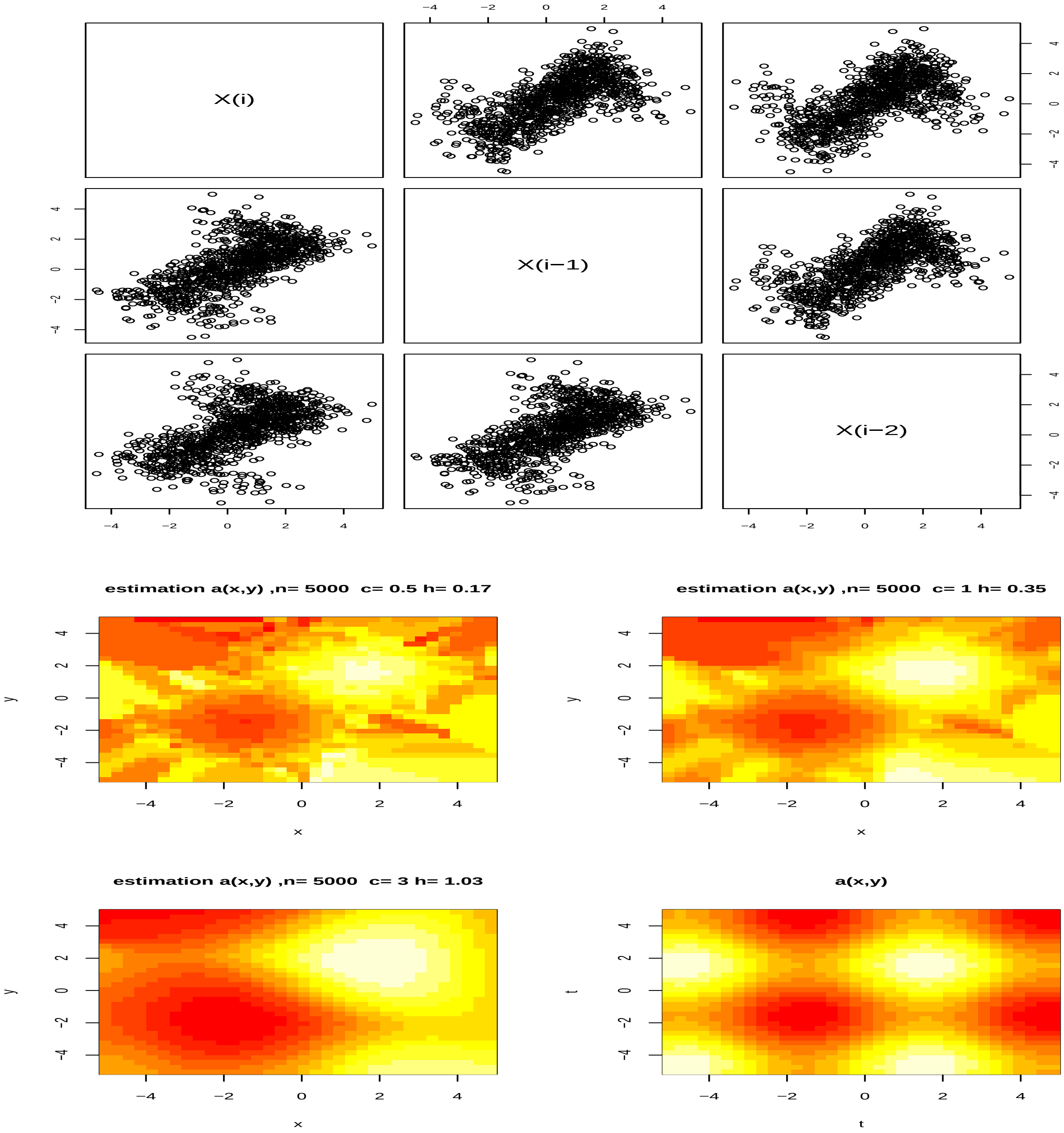

FIGURE 7. The model is defined by $a(x, y) = sin(x)+sin(y)$, and a Gaussian noise $\mathcal{N}(0, 1)$.

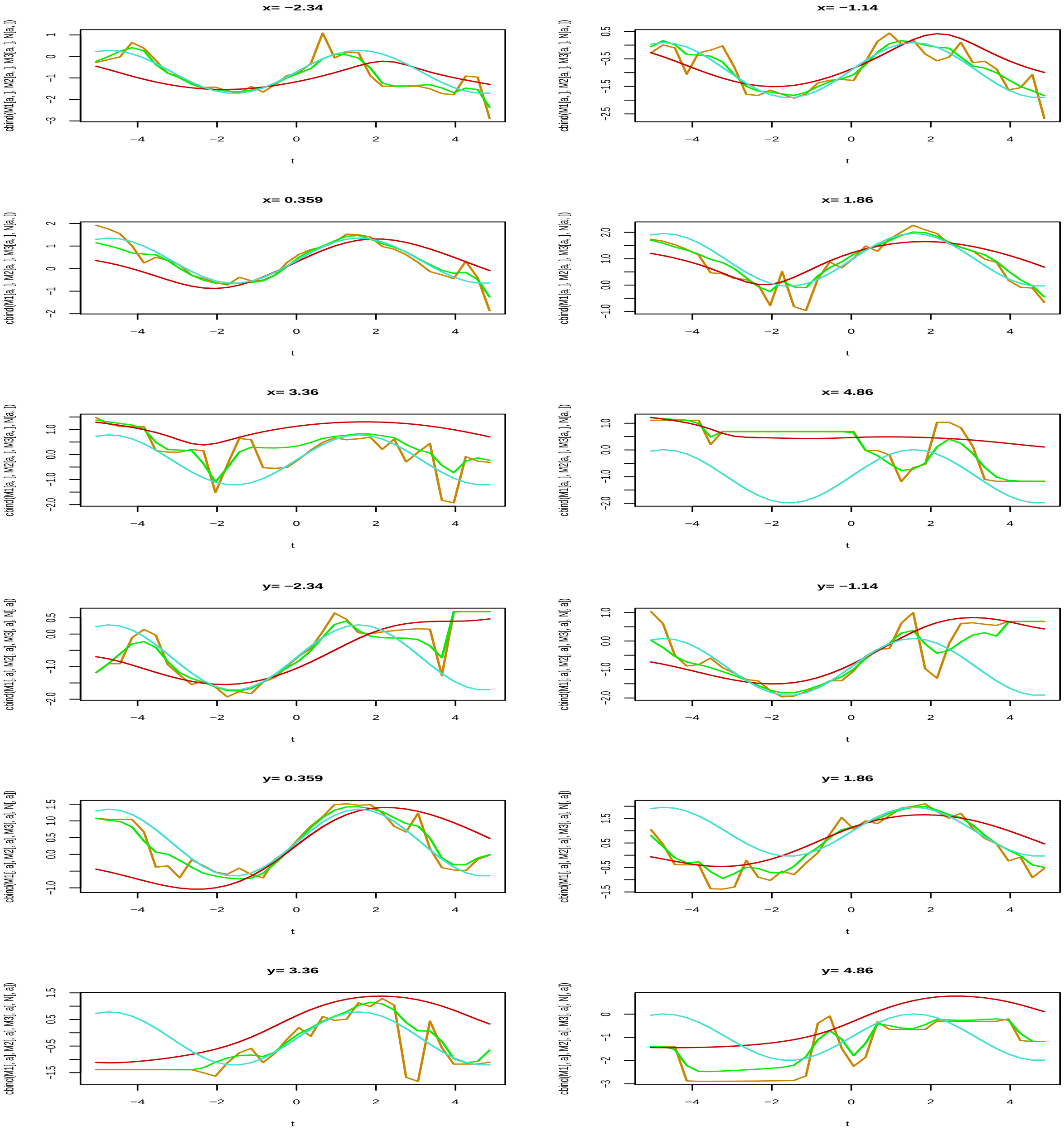

Figure 8. The same model as Figure 7

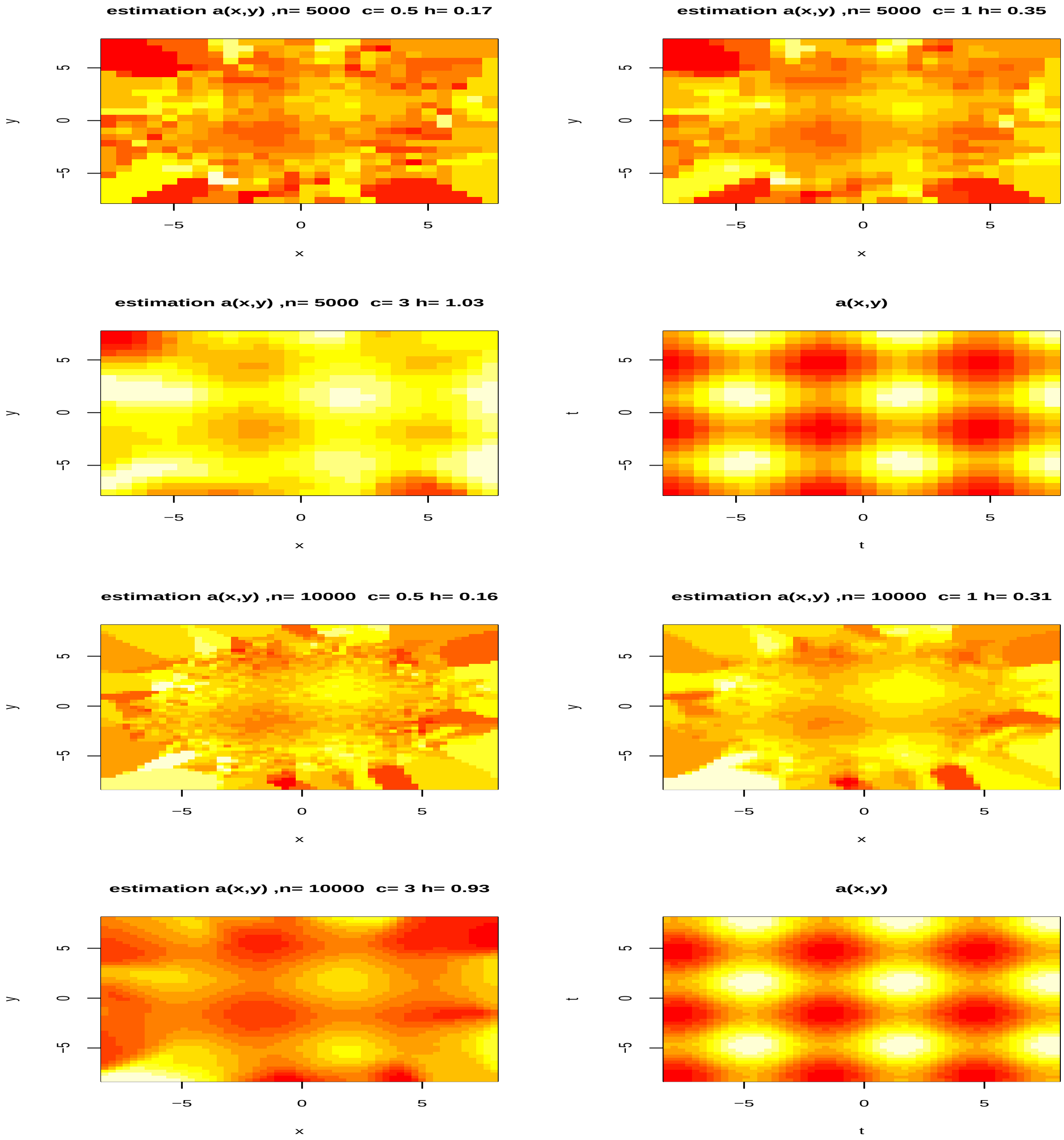

FIGURE 9. The model is defined by $a(x, y) = sin(x)+sin(y)$, and a Gaussian noise $\mathcal{N}(0, 4)$.

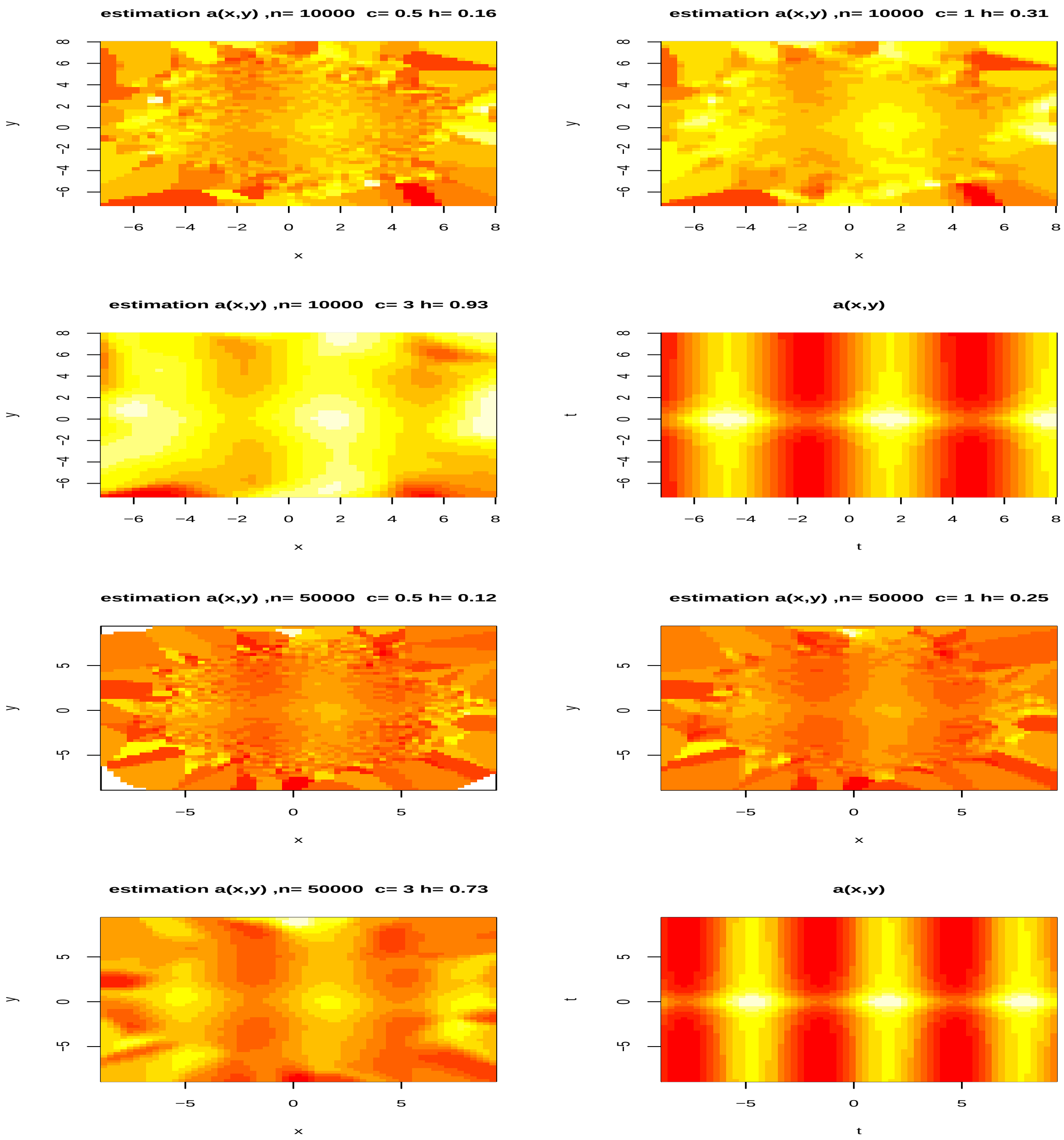

FIGURE 10. The model is defined by $a(x,y) = sin(x) + 1/(1+y^2)$, and a Gaussian noise $\mathcal{N}(0,4)$.

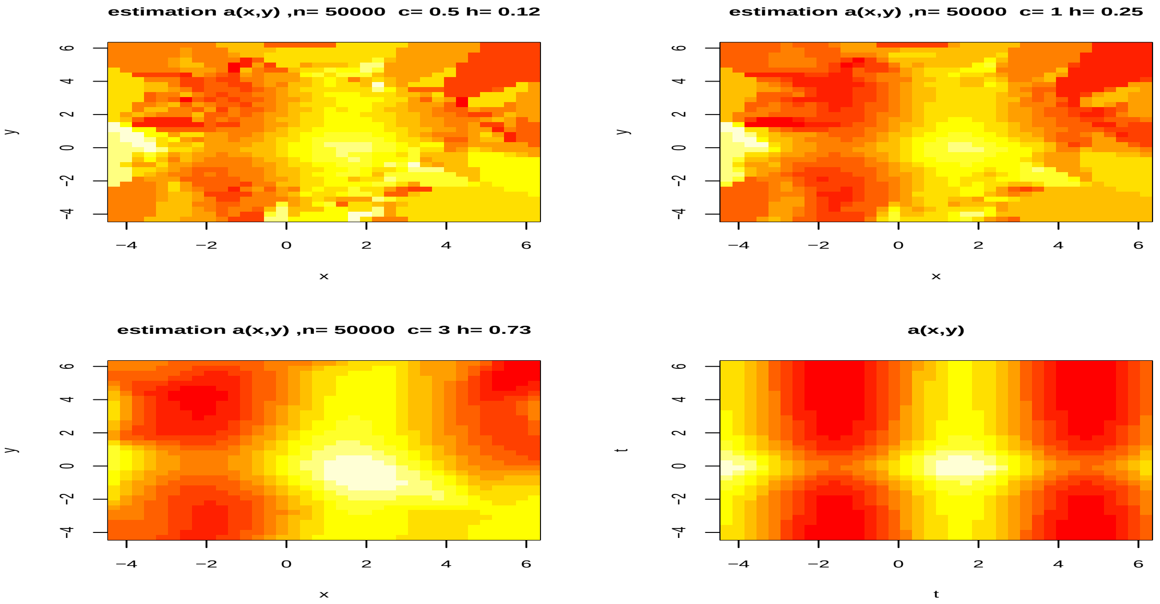

FIGURE 11. The model is defined by $a(x, y) = sin(x) + 1/(1 + y^2)$, and a Gaussian noise $\mathcal{N}(0, 1)$.

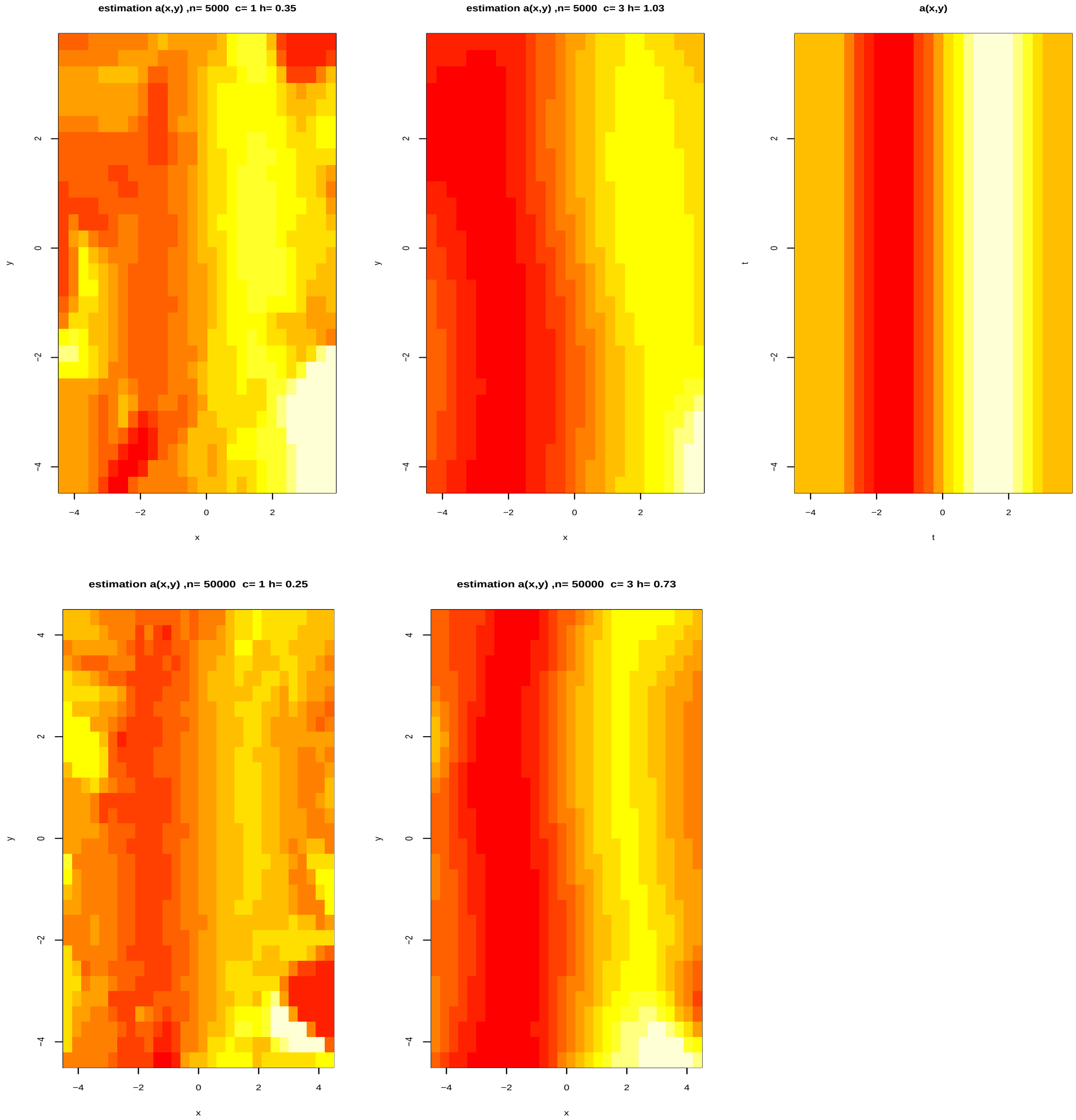

FIGURE 12. The model is defined by $a(x, y) = sin(x)\,\mathbb{I}_{[-4,4]}(y)$, and a Gaussian noise $\mathcal{N}(0, 1)$.

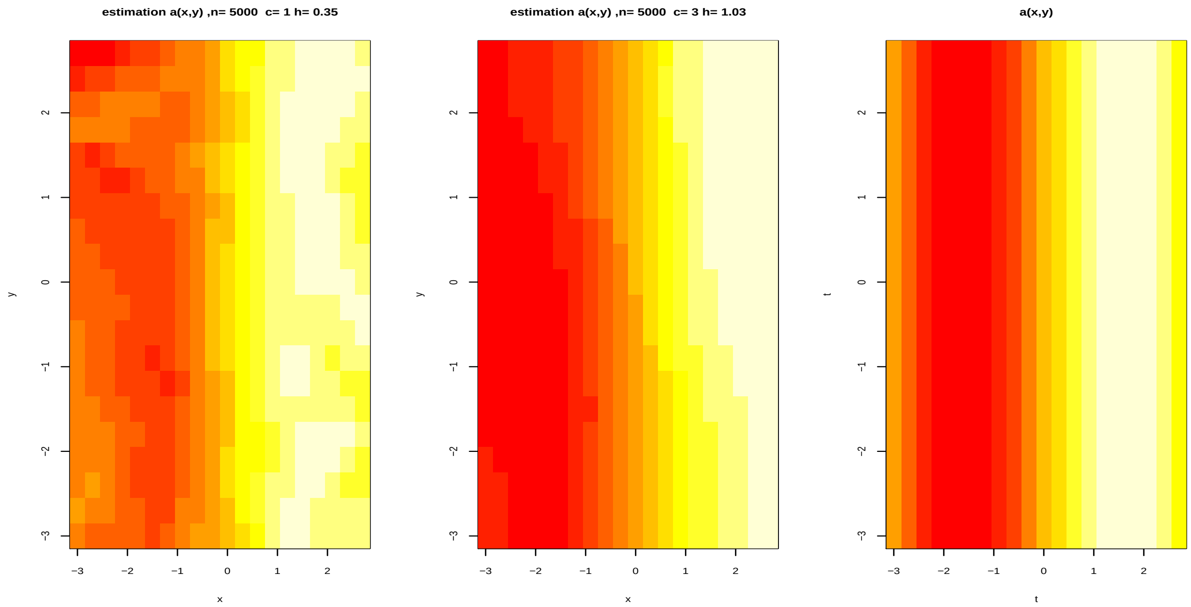

FIGURE 13. The model is defined by $a(x, y) = sin(x)\,\mathbb{I}_{[-4,4]}(y)$, and a uniform noise on $(-2, 2)$.

8. Exercises

Exercise 17. Prove Lemma 13. ⋆

Exercise 18. Let (U_j) be a stationary bounded strong mixing sequence such that, as $n \to \infty$, $\alpha_n \sim c\tau^n$ where $0 < \tau < 1$.

(1) Prove that the series $\sum_k \mathrm{Cov}(X_1, X_k)$ is absolutely convergent

(2) Deduce that, as $n \to \infty$,

$$\frac{1}{n}\mathrm{Var}\left(\sum_{j=1}^{n} U_j\right)$$

is convergent. Give an expression of the limit and compare with the i.i.d. case.

(3) Deduce also that $s_n^2 \sim cn$ as $n \to \infty$.

(4) Determine the values of parameter β for which the previous results still hold for an arithmetic strong mixing sequence ($\alpha_n \sim cn^{-\beta}$ as $n \to \infty$).

⋆

Exercise 19. Prove the optimality of $\beta_2 = -\beta_1 = 1/5$ in Corollary 18. ⋆

Exercise 20. Give comments for the acf graphs on the top of Figures 1 to 6 of the present chapter. ⋆

Exercise 21. Give comments for the graphs depicting the set of (X_l, X_{l+1}) in the same Figures. ⋆

Exercise 22. Do you see why the method you proposed in exercise 16 could give bad result for the estimation of the functions $a_1()$ and $a_2()$ in the model

$$X_k = a_1(X_{k-1}) + a_2(X_{k-2}) + \varepsilon_k?$$

An iterative method which seems to goodly perform consists in an iterative scheme (see [8] where chapter 8 for is devoted to the so-called *backfitting*). The main lines are the following.

- First step. Chose a preliminary estimate $\hat{a}_2^{(0)}()$ of $a_2()$. For example chose a constant, which could be the mean of the observations.
- Second step. Use this estimate to estimate $X_j - a_2(X_{k-2})$ by $X_j - \hat{a}_2^{(0)}(X_{k-2})$ and then calculate a first estimate of $a_1()$ by

$$\hat{a}_1^{(1)}(x) = \frac{\sum_{j=2}^{n-1}(X_j - \hat{a}_2^{(0)}(X_{j-2}))K\left(\frac{X_{j-1}-x}{h_n}\right)}{\sum_{j=2}^{n-1} K\left(\frac{X_{j-1}-x}{h_n}\right)}$$

- Third step. Re-estimate a_2 by

$$\hat{a}_2^{(1)}(x) = \frac{\sum_{j=2}^{n-1}(X_j - \hat{a}_1^{(1)}(X_{j-1}))K\left(\frac{X_{j-2}-x}{h_n}\right)}{\sum_{j=2}^{n-1} K\left(\frac{X_{j-2}-x}{h_n}\right)}$$

- Following steps. Re-estimate a_1 as in step 2 from $\hat{a}_2^{(1)}$, and so on. Stop when the estimates stabilize.

Do you understand why this iterative procedure could work? ⋆

CHAPTER 6

Mixed models

1. Introduction

We now turn to general models presenting an additive contribution of a functional autoregression $a(X_k)$ and of a pure regression on exogeneous variables $b(e_k)$

$$X_k = a(X_{k-1}) + b(e_k) + \varepsilon_k.$$

We suppose that $(e_k)_k$ is i.i.d. This is really an unrealistic hypothesis, presented here for easiness of theory. More, the method presented below is not adapted if the independence is lost. In this case, methods of type *backfitting* as presented in exercise 22 are surely more performing (see [8]).

1.1. Basic remark.

Firstly, it is clear that

$$\mathbb{E}(X_{n+1}|X_n, e_{n+1}, \ldots, X_1, e_2) = a(X_n) + b(e_{n+1}),$$

implying that the optimal predictor based on $(X_n, e_{n+1}, \ldots, X_1, e_2)$ is

$$\hat{X}_{n+1} = a(X_n) + b(e_{n+1}).$$

Then, a and b being unknown to the statistician, they have to be estimated. Notice that a direct kernel method based on a 2-dimensional kernel measuring the distance between (X_j, e_{j+1}) and (x, e) can not be used, because, apart its well known bad performances, it would result in a non additive function of (x, e).

Suppose that $\mathbb{E}(e_k) = 0$, then

$$\begin{array}{rcl} \mathbb{E}(X_k|X_{k-1}, \ldots, X_1) & = & a(X_{k-1}) \\ \text{and} & & \\ b(e_k) & = & \mathbb{E}(X_k - a(X_{k-1})|e_k, \ldots, e_1) \end{array}$$

1.2. Estimation.

This is a good reason to propose the following estimates:

(48) $$\hat{a}_n(x) = \frac{\sum_{j=1}^{n-1} X_{j+1} K\left(\frac{X_j - x}{h_n}\right)}{\sum_{j=1}^{n-1} K\left(\frac{X_j - x}{h_n}\right)}$$

(49) $$\hat{b}_n(e) = \frac{\sum_{j=1}^{n-1}(X_{j+1} - \hat{a}_n(X_j))K\left(\frac{e_{j+1}-e}{h_n}\right)}{\sum_{j=1}^{n-1} K\left(\frac{e_{j+1}-e}{h_n}\right)}.$$

2. Assumptions and first consequences

We first gather all the assumptions made in Chapters 4 and 5. They will not be repeated here.

For the sake of identifiability, we suppose also (see the introduction of the present chapter)

(50) $$\mathbb{E}(b(e_1)) = 0.$$

This means that (as in the pure autoregressive scheme) a possibly non zero mean for the stationary solution (X_n) is such that

$$\mathbb{E}(X_1) = \mathbb{E}(a(X_1)).$$

Notice that since the noise and the exogeneous sequence are independent and both i.i.d, the model is formally

$$X_{k+1} = a(X_k) + \eta_{k+1}$$

where $(\eta_k)_k$ is a zero mean white bounded noise having a positive marginal density, obtained by convolution of the density of $b(e_k)$ with the density of ε_k.

Then, since a is bounded there exists a stationary solution to this functional autoregressive model, and this solution has the same mixing properties as in Chapter 5.

3. Convergence results

Two basic facts explain why Proposition 19 below holds. The details are left to the reader

- Denoting $U_{j+1} = X_{j+1} - a(X_j)$, the model $U_{k+1} = b(e_{k+1}) + \varepsilon_{k+1}$ is a pure regression satisfying all hypotheses of Theorem 4. Hence, with

$$\tilde{b}_n(e) = \frac{\sum_{j=1}^{n-1}(X_{j+1} - a(X_j))K\left(\frac{e_{j+1}-e}{h_n}\right)}{\sum_{j=1}^{n-1} K\left(\frac{e_{j+1}-e}{h_n}\right)}$$

if $h_n \sim c\left(\frac{\ln n}{n}\right)^{-\beta}$

$$\sup_e |\tilde{b}_n(e) - b(e)| = O_{as}\left(\frac{\ln n}{n}\right)^{-2\beta} + O_{as}\left(\frac{\ln n}{n}\right)^{(1+\beta)/2}$$

- The autoregression $X_{k+1} = a(X_k) + \eta_{k+1}$ satisfies all the assumptions of Theorem 16 and, for $\hat{a}_n(x)$ defined in (47), if $h_n \sim n^{\beta_1} h_n^{\beta_2}$,

$$\sup_x |\hat{a}_n(x) - a(x)| = O_{as}\left(n^{(1-\beta_1)/2} \ln n^{(1+\beta_2)/2}\right) + O_{as}(n^{2\beta_1} \ln n^{2\beta_2}).$$

It remains to mix these two results an to introduce estimator $\hat{b}_n$ defined in (48).

Proposition 19. *With the assumptions of the introduction,*
(i) if $h_n \sim c\left(\frac{\ln n}{n}\right)^{-\beta}$

$$\sup_x |\hat{a}_n(x) - a(x)| + \sup_e |\tilde{b}_n(e) - b(e)| = O_{as}\left(\frac{\ln n}{n}\right)^{-2\beta} + O_{as}\left(\frac{\ln n}{n}\right)^{(1+\beta)/2},$$

(ii) The optimal rate, reached for $h_n \sim c\left(\frac{\ln n}{n}\right)^{1/5}$ *is*

$$\sup_x |\hat{a}_n(x) - a(x)| + \sup_e |\tilde{b}_n(e) - b(e)| = O_{as}\left(\frac{\ln n}{n}\right)^{2/5}.$$

Remark 17. While a two variable model such as

$$X_{n+1} = A(X_n, e_{n+1}) + \varepsilon_n$$

would have given a rate of convergence $n^{-1/3}$, up to a multiplicative logarithmic factor (there is no such result in this document, but see 35 and 46 to understand why this claim should be correct), the present additive model permits the same rate $n^{-2/5}$ as in the case of one variable. This is the advantage of using additive models.

4. Illustration

We only consider the case of the Gaussian kernel and we evaluate the effects of the bandwidth h_n. According to the theoretical result we take h_n of the form $C(\log(n)/n)^{1/5}$ for different values of C.

The following pictures provide

- The sets of points (X_i, X_{i+1}) and (e_i, X_i)
- The kernel estimates of a and b for the constant $C = 0.1,\ 0.5,\ 1,\ 2.$ and the sample size 500 and 5000

In the three examples, the random variables (e_n) are iid from a uniform distribution on $(-1, 1)$ and the noise is a Gaussian noise $\mathcal{N}(0, 1)$

- Figures 1, 2 : the model is defined by $a(x) = sin(x)$, $b(e) = e\mathbb{I}_{[-1,1]}(e)$.

- Figures 3, 4 : the model is defined by $a(x) = 1/(x^2 + 1)$, $b(e) = e^3 \mathbb{I}_{[-1,1]}(e)$.
- Figures 5, 6 :The model is defined by $a(x) = \mathrm{sign}(x)\mathbb{I}_{[0,1]}(|x|)$, $b(e) = e^3 \mathbb{I}_{[-1,1]}(e)$.

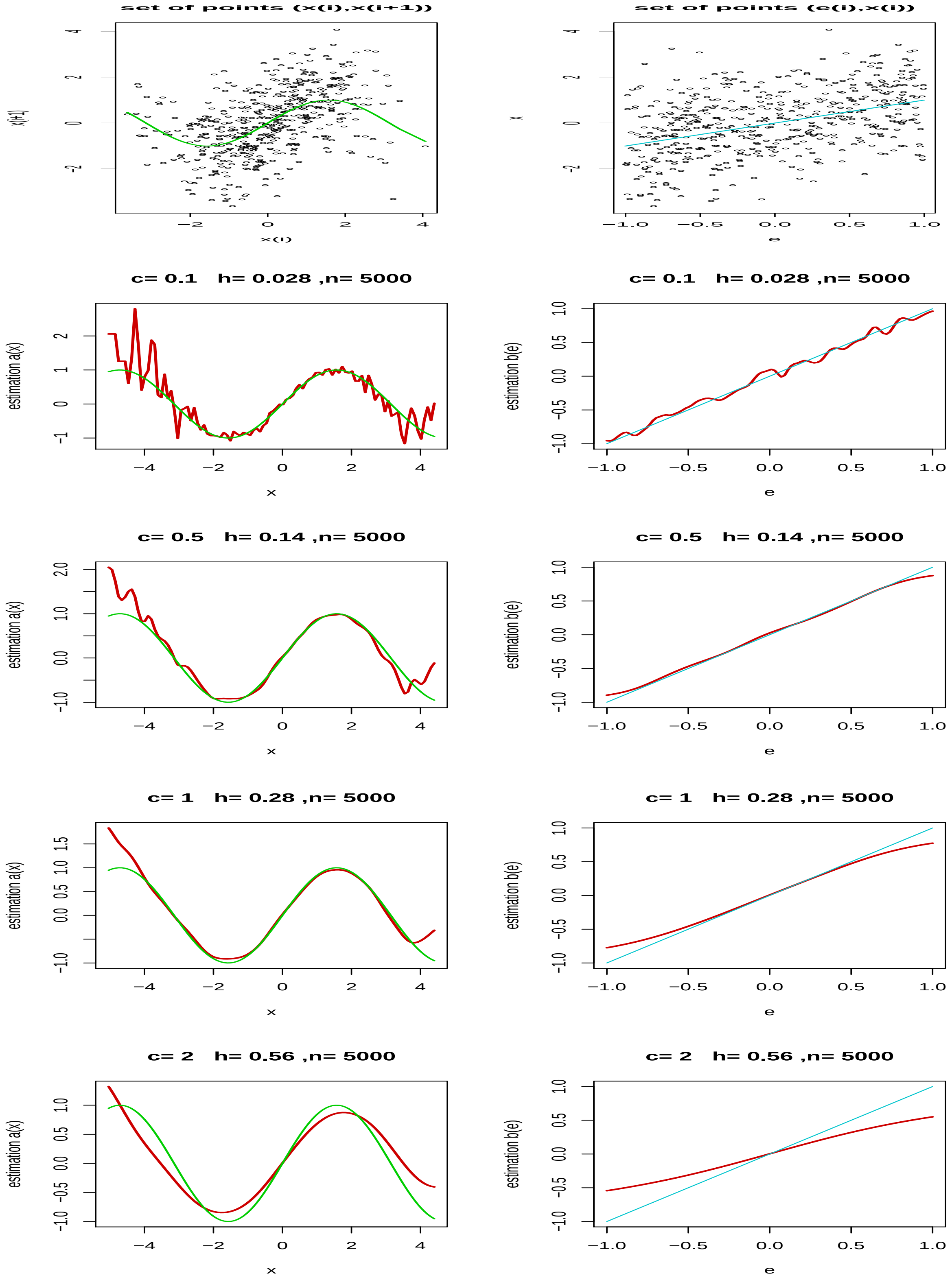

Figure 1. The model is defined by $a(x) = sin(x)$, $b(e) = e\mathbb{I}_{[-1,1]}(e)$, (e_n) are iid from a uniform distribution on $(-1, 1)$ and a Gaussian noise $\mathcal{N}(0, 1)$. The sample size is $n = 5000$

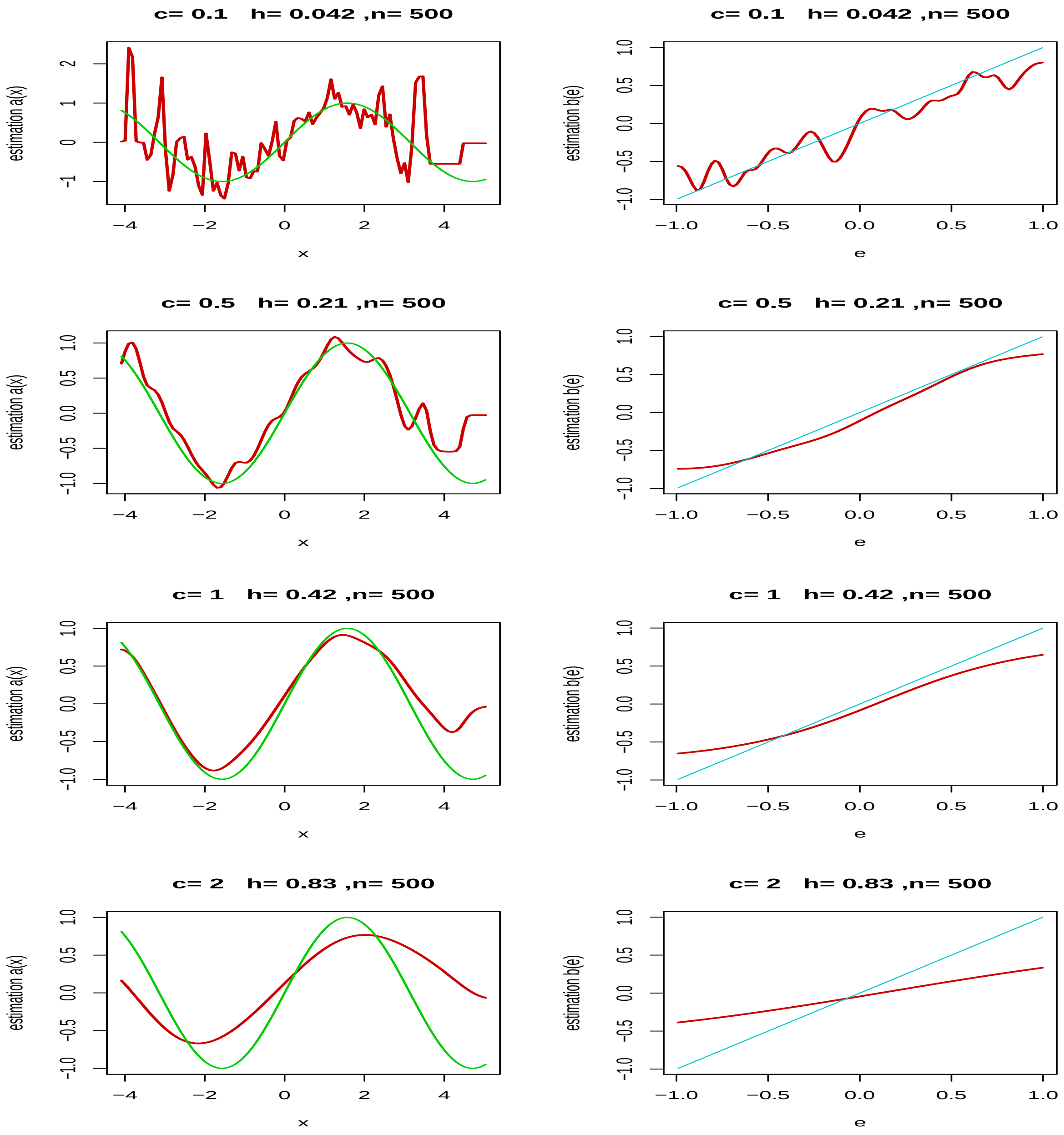

FIGURE 2. The same model as Fig 1 $n = 500$

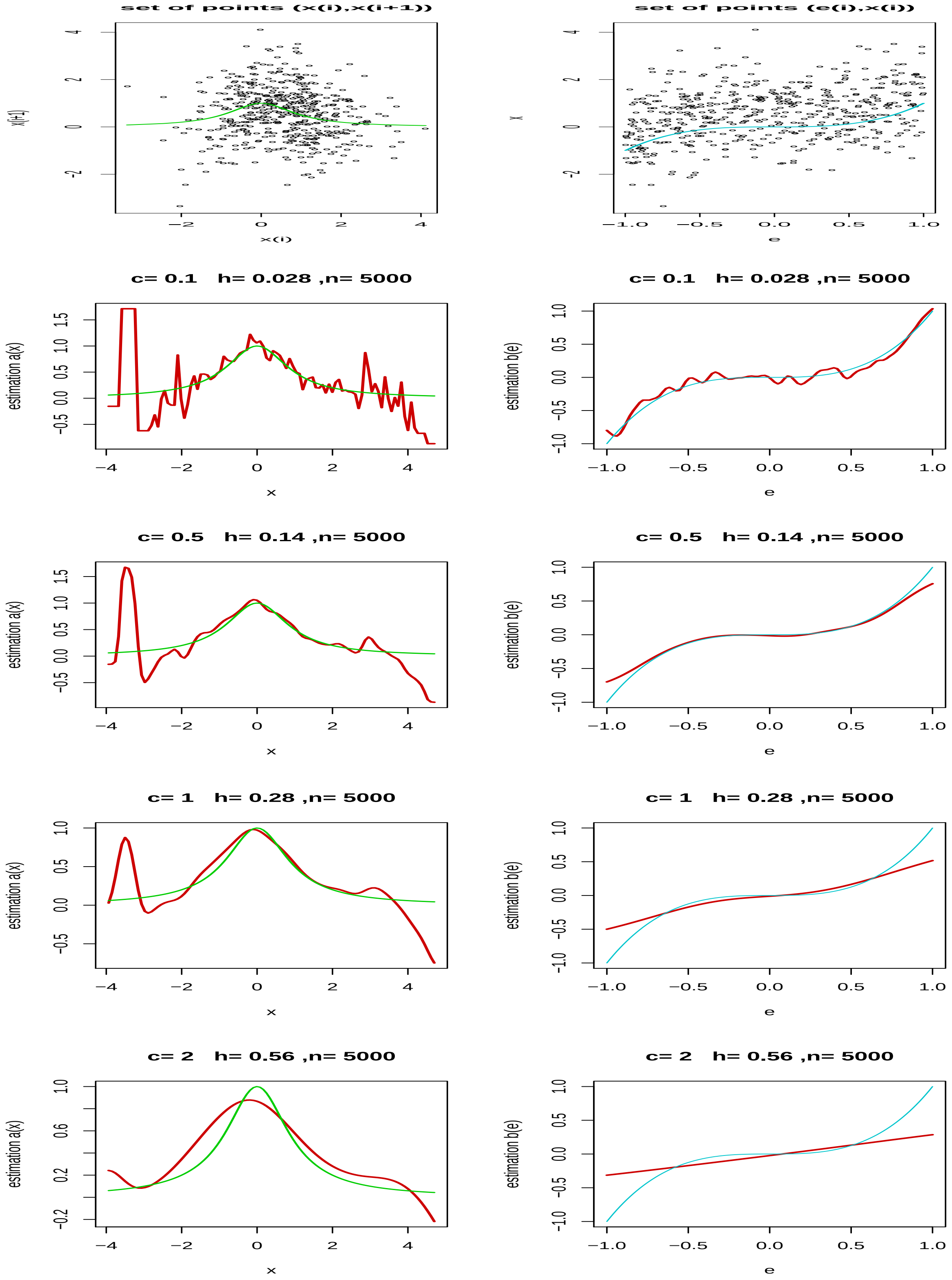

FIGURE 3. The model is defined by $a(x) = 1/(x^2+1)$, $b(e) = e^3\mathbb{I}_{[-1,1]}(e)$, (e_n) are iid from a uniform distribution on $(-1,1)$ and a Gaussian noise $\mathcal{N}(0,1)$. The sample size is $n = 5000$

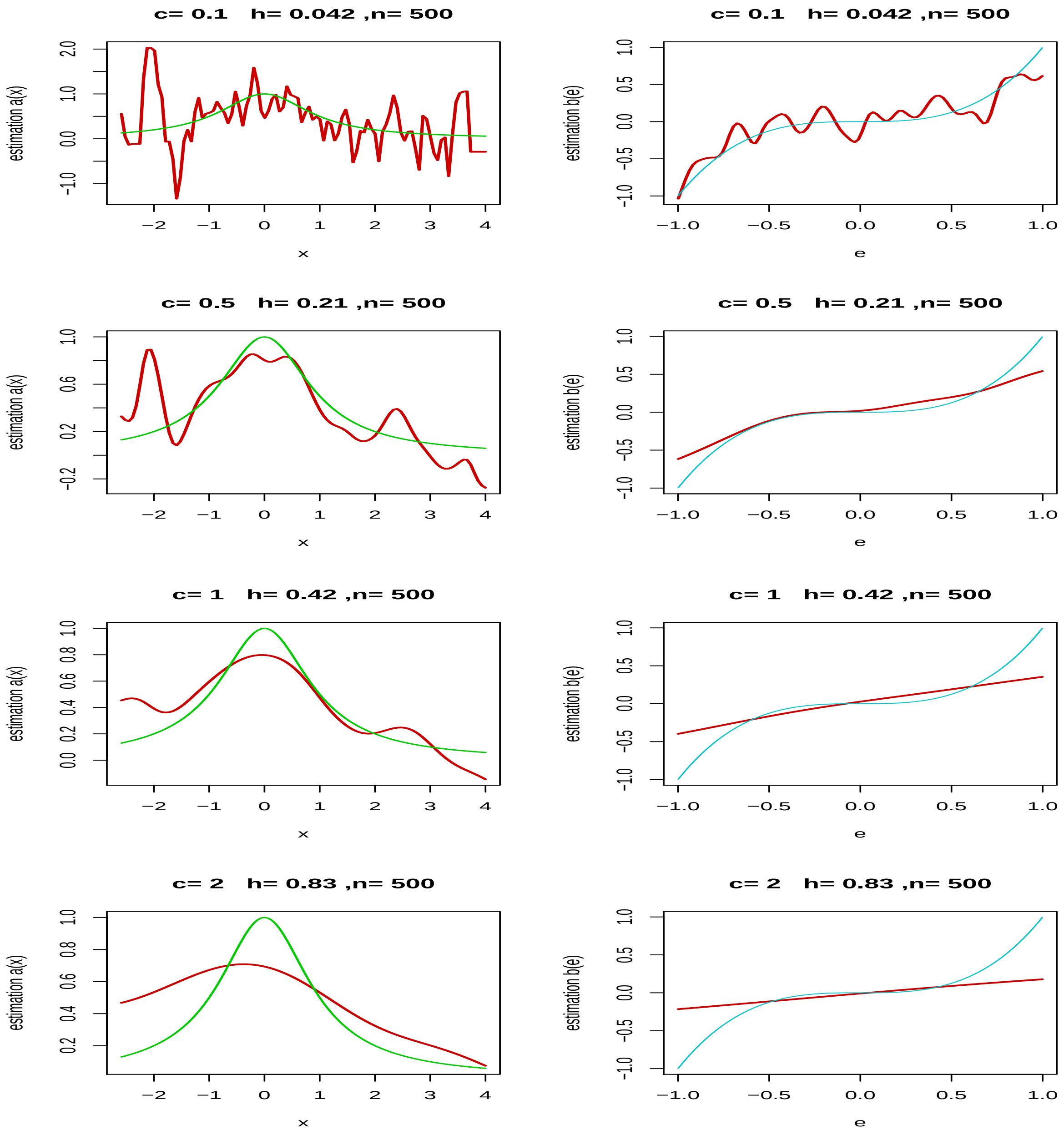

Figure 4. The same model as Fig. 3. $n = 500$

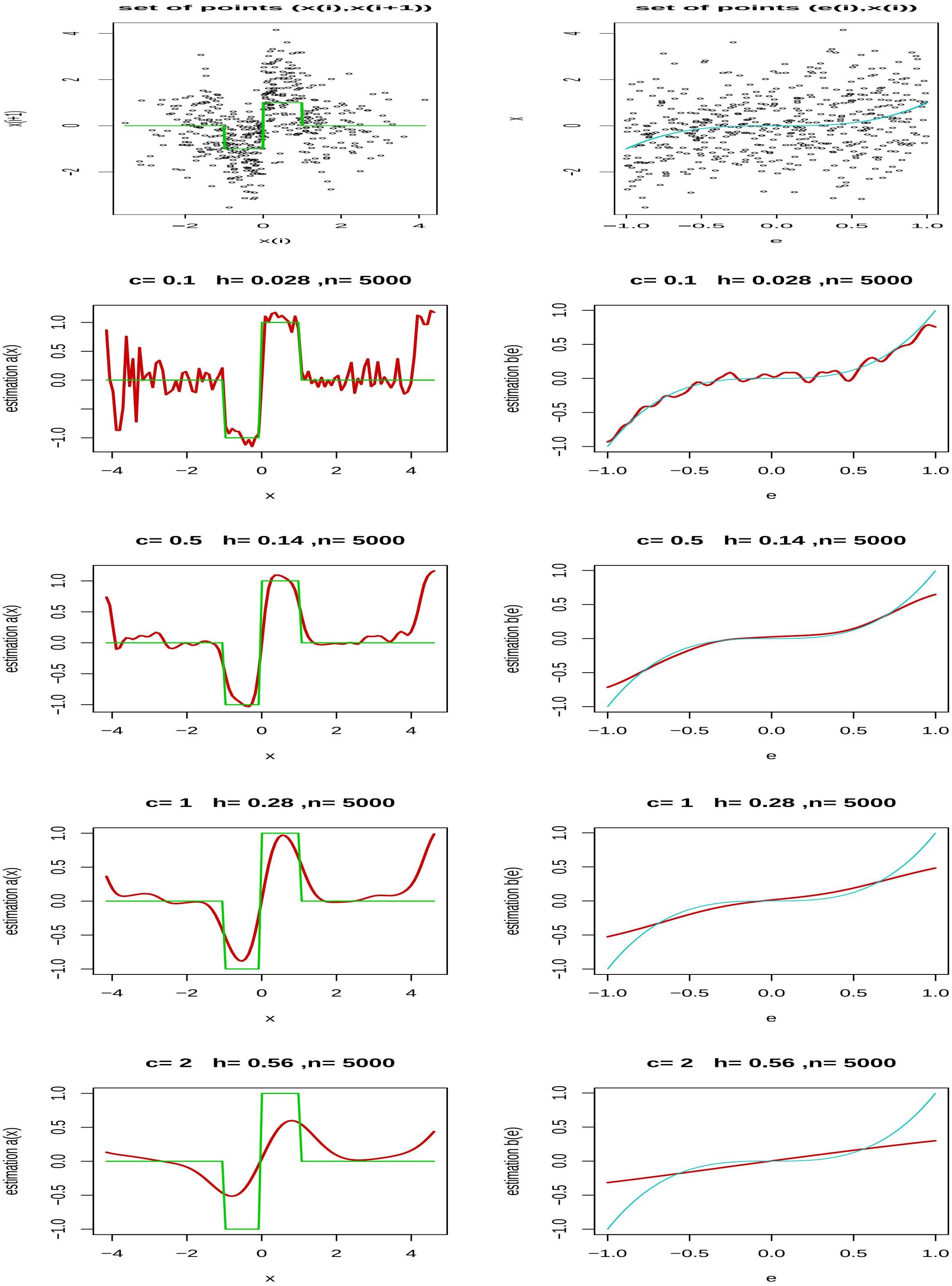

FIGURE 5. The model is defined by $a(x) = \mathrm{sign}(x)\mathbb{I}_{[0,1]}(|x|)$, $b(e) = e^3\mathbb{I}_{[-1,1]}(e)$, (e_n) are iid from a uniform distribution on $(-1,1)$ and a Gaussian noise $\mathcal{N}(0,1)$. The sample size is $n = 5000$

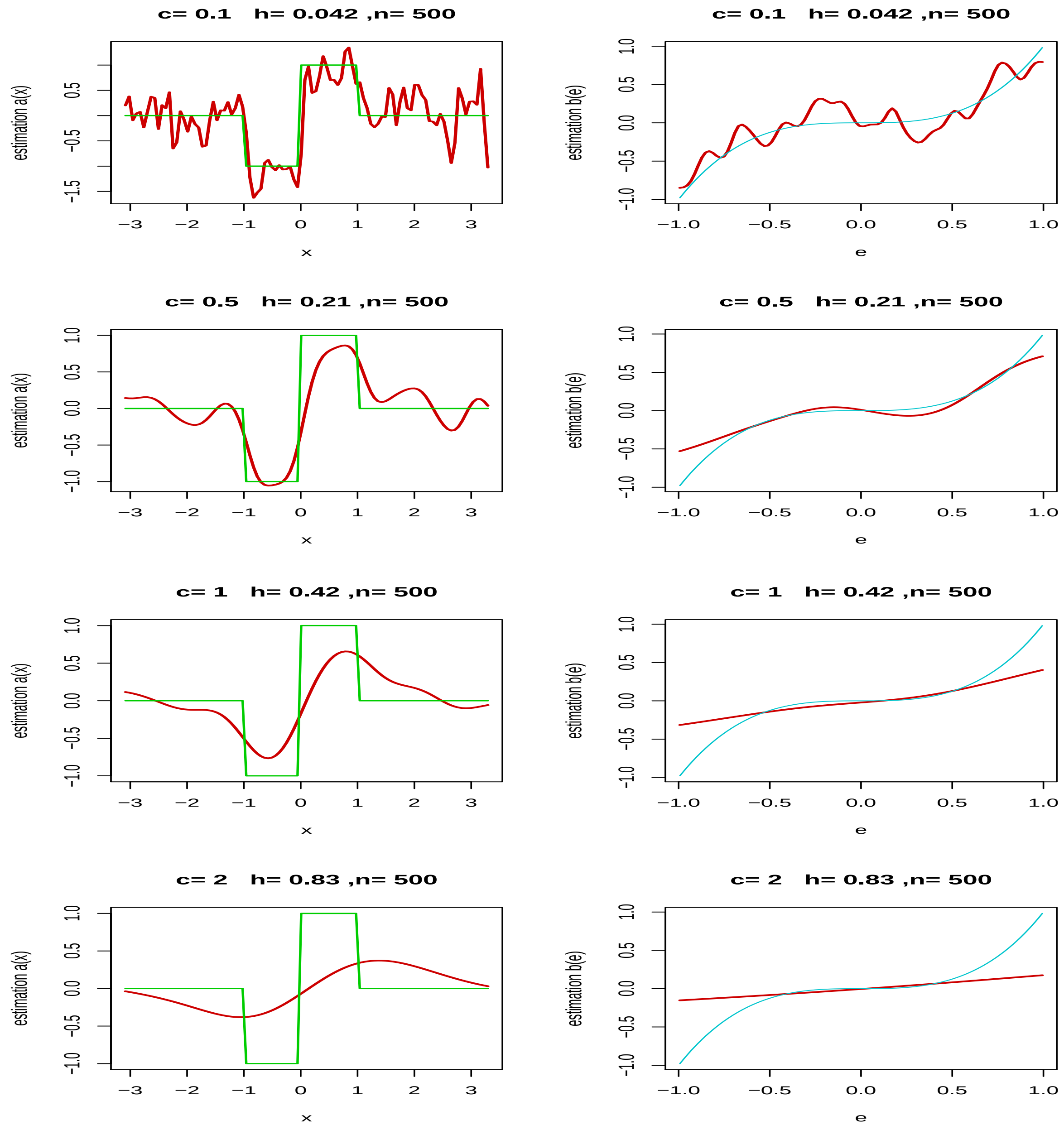

FIGURE 6. The same Fig. 5. $n = 500$

5. Forecasting

The one step ahead predictor is

$$\hat{X}_{n+1} = \hat{a}_n(X_n) + \hat{b}_n(e_{n+1})$$

and the same remarks as in Chapters 4 and 5 can be repeated here.

6. Other methods

The model and the method presented in this chapter are too particular to be interesting for practical purposes.

We mention here two widely used procedures (see [8], [10] and [11] among others).

6.1. Backfitting methods. They are introduced in exercise 22 above. For the moment, theoretical developments are missing in the literature. But these methods are easy to implement and seem rather successful.

6.2. Projection. These methods consist in changing the non-parametric problem of estimating functions in a parametric problem.

- Expand a and b on bases of functions (G_k) and (L_k) (possibly the same)

$$\begin{aligned} a(x) &= \sum_{j\geq 1} \lambda_j G_j(x) \\ b(e) &= \sum_{j\geq 1} \nu_j L_j(e) \end{aligned}$$

- Choose two truncation values I_n and J_n
- Use (for example) a LMS method to estimate the parameters $\lambda_1, \ldots, \lambda_{I_n}$ and $\nu_1, \ldots, \nu_{J_n}$ in the model

$$X_{n+1} = \sum_{j=1}^{I_n} \lambda_j G_j(X_n) + \sum_{j=1}^{J_n} \nu_j L_j(e_{n+1}) + R_n + \varepsilon_{n+1}$$

  where R_n is the remainder of the two expansions.

See [3] where this method is used to treat the case where the exogeneous variables are not i.i.d.

7. Exercises

Exercise 23. Give the details of the proof of Proposition 19. ⋆

Exercise 24. Consider the model of Figures 1 and 2 and suppose that the statistician knows that the second function is linear. In this case he shall estimate the function $a()$ and the parameter b in the model

$$X_{k+1} = a(X_k) + be_{k+1} + \varepsilon_{k+1}.$$

Try to propose a method to perform this estimation. ⋆

Exercise 25. Consider the partially linear model

$$X_{k+1} = aX_k + b(e_{k+1}) + \varepsilon_{k+1}.$$

with the same hypotheses as in the previous chapters.

(1) Prove that $a = \mathrm{Cov}(X_n, X_{n+1})$
(2) Deduce a convergent estimator of a.
(3) Use this estimator to construct a kernel estimator of $b(e)$.

⋆

Exercise 26. Try to use the same kind of method to estimate a_1, a_2 and the function $b()$ in the model

$$X_{n+1} = a_1X_n + a_2X_{n-1} + b(e_{n+1}) + \varepsilon_{n+1}.$$

⋆

Bibliography

[1] Ango Nze P., Doukhan P. (2004) Weak dependence: models and applications to econometrics. Econometric Theory. 20. 995–1045.

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

[3] Bosq D., Shen J. (1998) Estimation of an autoregressive semiparametric model with exogenous variables. Journal of Stat. Plann. and Inf. 68. 105–127.

[4] Brockwell P. J., Davis R. A. (1991) Time Series: Theory and Methods. Springer Verlag. New York.

[5] Cramer H., Leadbetter R. (1967). Stationary and related stochastic processes. Wiley.

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

[7] Doukhan P., Ghindes M. (1980) Etude du processus $X_{n+1} = f(X_n) + \varepsilon_n$. C.R.Acad. Sci. Paris. Ser.A. 290. 921–923.

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

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

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

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

[12] Mokkadem A. (1987) Sur un modèle autorégressif non linéaire, ergodicité et ergodicité géométrique. J. of Time Ser. Anal. 8-2; 195–204.

[13] Nadaraya E. (1964) On estimating regressions. Theory Prob. Appl. 10. 186–196.

[14] Parzen E. (1962). On estimation of a probability density function and mode, Ann. Math. Stat. 33, pp. 1065-1076.

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

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

[17] Tjøstheim D. (1994) Non-linear time series: a selective review. Scand. J. of Stat. vol. 21. 97–130.

[18] Tsybakov A. (2004) *Introduction à l'estimation non paramétrique.* Springer.

[19] Tukey J. W. (1961) Curve as parameters, and touch estimation. Proceedings ot the 4th Symposium on Mathematices, Statistics and Probability, 681-694, Berkeley, CA, USA.

[20] Vieu Ph. (1993) Bandwidth selection for kernel regression: a survey. *Computer Intensive methods in statistics.* Eds Härdle W. & Simar, 134–149. Physica Verlag. Heidleberg. Germany.

[21] Watson G. S. (1964) Smooth regression analysis. Sankhya Ser A. 26. 359–372.

[22] William D. (1997). *Probability with martingales.* Camb. Math. Textbooks.