INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2009-18

Combining Nonparametric and Optimal Linear Time Series Predictions

S. DABO-NIANG[1]
C. FRANCQ[2]
J.-M. ZAKOIAN[3]

Les documents de travail ne reflètent pas la position de l'INSEE et n'engagent que leurs auteurs.

Working papers do not reflect the position of INSEE but only the views of the authors.

[1] Université Lille III, GREMARS-EQUIPPE, BP 60149, 59653 Villeneuve d'Ascq Cedex, France. Email : sophie.dabo@univ-lille3.fr

[2] Université Lille III, GREMARS-EQUIPPE, BP 60149, 59653 Villeneuve d'Ascq Cedex, France. Email : christian.francq@univ.lille3.fr

[3] CREST and EQUIPPE-GREMARS, 15 boulevard Gabriel Péri, 92245 Malakoff Cedex, France. Email : zakoian@ensae.fr

Combining nonparametric and optimal linear time series predictions

Sophie Dabo-Niang[*], Christian Francq[†] and Jean-Michel Zakoïan[‡]

Abstract: We introduce a semiparametric procedure for more efficient prediction of a strictly stationary process admitting an ARMA representation. The procedure is based on the estimation of the ARMA representation, followed by a nonparametric regression where the ARMA residuals are used as explanatory variables. Compared to standard nonparametric regression methods, the number of explanatory variables can be reduced because our approach exploits the linear dependence of the process. We establish consistency and asymptotic normality results. A Monte Carlo study and an empirical application on stock indices suggest that significant gains can be achieved with our approach.

Résumé: Nous introduisons une procédure semi-paramétrique afin de prévoir plus efficacement un processus strictement stationnaire admettant une représentation ARMA. Cette procédure est fondée sur l'estimation de la représentation ARMA, suivie d'une régression non paramétrique dans laquelle les résidus ARMA sont utilisés comme variables explicatives. Par rapport aux méthodes standard de régression non paramétrique, cette approche permet de réduire le nombre de variables explicatives car elle exploite la dépendance linéaire du processus. On établit des résultats de convergence et de normalité asymptotique. Une étude par simulation et une application sur données d'indices boursiers montrent que des gains d'efficacité significatifs peuvent être obtenus par cette méthode.

[*] Université Lille III, EQUIPPE-GREMARS, BP 60 149, 59653 Villeneuve d'Ascq cedex, France. E-mail: sophie.dabo@univ-lille3.fr

[†] Université Lille III, EQUIPPE-GREMARS, BP 60 149, 59653 Villeneuve d'Ascq cedex, France. E-mail: christian.francq@univ-lille3.fr

[‡] CREST and EQUIPPE-GREMARS, 15 Bd G. Péri, 92245 Malakoff Cedex, France. E-mail: zakoian@ensae.fr

1 Introduction

After three decades of non-linear time series models, the class of ARMA models remains the most widely employed parametric family. Reasons can be found in the generality of the class (when the noise is only assumed to be uncorrelated), the relative ease of implementation and the ability to provide optimal linear predictions. If a stationary processes (W_t) is an ARMA process, its optimal linear prediction

$$EL(W_t \mid \{W_u, u < t\}) = \sum_{i=1}^{\infty} a_i W_{t-i}, \tag{1.1}$$

is obtained from the ARMA model. However, ARMA models also have important drawbacks. Their generality vanishes when strong assumptions (such as independence, martingale difference) are made on the noise. The optimal linear prediction does not coincide, in this case, with the optimal prediction

$$E(W_t \mid \{W_u, u < t\}) = \phi(W_{t-1}, W_{t-2}, \dots). \tag{1.2}$$

Nonlinear models have been introduced to solve the problem, but they may be hard to identify.

In situations where parametric families cannot be adopted with confidence, nonparametric models offer an alternative. Nonparametric kernel regressors seem attractive because they aim at estimating the regression of the observed process W_t on its past values $W_{t-1}, \dots, W_{t-d}$,

$$r(W_{t-1}, \dots, W_{t-d}) = E(W_t \mid W_{t-1}, \dots, W_{t-d}) \tag{1.3}$$

without requiring strong assumptions on the data generating process. The choice of the number d of lags is however crucial for the following reason. When d is chosen too small, the nonparametric predictions are likely to be far from the optimality, even when the number of observations n increases to infinity. On the other hand when d is large, the method is subject to the so-called *curse of dimensionality* (the kernel estimator converges at a rate $n^{2/(4+d)}$ which is low when d is large).

In this work we consider a third approach to the problem of time series prediction, which combines parametric and nonparametric methods. The idea is to utilize ARMA residuals as regressors in the nonparametric approach, to forecast the subsequent behaviour of W_t. More precisely, we consider two approaches. In the first one we use

$$\tilde{r}(W_{t-1}, \dots, W_{t-\ell}, \epsilon_{t-1}, \dots, \epsilon_{t-m}) = E(W_t \mid W_{t-1}, \dots, W_{t-\ell}, \epsilon_{t-1}, \dots, \epsilon_{t-m}) \tag{1.4}$$

as an approximation to the optimal prediction in (1.2), where $\epsilon_t = W_t - EL(W_t \mid \{W_u, u < t\})$ denotes the linear innovation of the stationary process (W_t). The use of a nonlinear regression aims to account for the underlying nonlinear structure of W_t. On the other hand, the use of linear innovations aims to alleviate the effects of the above-mentioned curse of dimensionality. Since the ϵ_t's are not observable, they are replaced by the residuals of an ARMA model. For this method we will show that the rate of convergence approaches $n^{2/(4+\ell+m)}$, which seems advantageous compared to the traditional nonparametric regression when $\ell + m < d$. In the second approach we use the decomposition of the optimal prediction in (1.2) as the sum of the optimal linear prediction and the optimal (nonlinear) prediction of the linear innovation process:

$$E(W_t \mid \{W_u, u < t\}) = EL(W_t \mid \{W_u, u < t\}) + E(\epsilon_t \mid \{\epsilon_u, u < t\}). \tag{1.5}$$

The idea is to estimate the first term parametrically, and the second term nonparametrically, in the right-hand side of (1.5). Again the innovations are replaced by residuals to obtain a feasible predictor. Under slightly different assumptions than in the first method, we will establish the consistency and asymptotic normality of the proposed estimator.

Our main motivation for using residuals in nonparametric estimators is parsimony. It is well-known that, in view of the parsimony principle, the class of ARMA models is preferable to the class of AR models (although both classes are dense in the set of the stationary processes). This is precisely the idea which is behind the approaches based on (1.4) and (1.5). If the same asymptotic precision is achieved with a first regression on a large number of past values and a second regression on only few past values and past linear innovations, it is reasonable to think that, in view of the curse of dimensionality, the second regression will do a better job in finite samples.

The essential difficulty in the derivation of the asymptotic results is that variables depending on a first-step estimator are included in the regressors and, for the method based on (1.5), are also included in the regressand. To cope with this problem, the idea is to interpret the ARMA residuals as noisy innovations, that is innovations that are corrupted by the effect of the parameters estimation. We therefore need to establish general asymptotic results for nonparametric regression based on "noisy time series". Specifically, we consider the case where the time t observation is the sum of the realization of an underlying stochastic process and a disturbance, which is allowed to depend on the sample size n.

The intuition behind this semiparametric method is simple and has obvious connections with numerous methods already proposed in the literature, in particular *i)* the combination of forecasts from individual models (see Timmermann (2006) and the references therein), *ii)* the pre-whitening methods, like the one proposed by Carroll, Linton, Mammen and Xiao (2002) for a regression model with autocorrelated errors, *iii)* the adaptive estimation methods, like the one proposed by Xu and Phillips (2008) for the inference of AR models with heteroscedasticity of unknown form (see also Phillips and Xu (2005), *iv)* the convex combination of parametric and nonparametric predictions proposed by Einsporn and Birch (1993) and Burman and Chaudhuri (1994) for possibly misspecified regression models (see also Fan and Ullah (1999)), *v)* the Model-Robust Regression method proposed by Mays, Birch, and Einsporn (2000), *vi)* the nonparametric correction factor proposed by Glad (1998). The methods mentioned in *v)* and *vi)* have been developed to protect parametric regressions against a model misspecification, which is not the concern of the present paper, but these two methods are close in spirit to the semiparametric method we consider in the present paper because they combine (additively in Mays et al. (2000) and multiplicatively in Glad (1998)) a parametric fit of the raw data and a nonparametric fit of the parametric residuals. In this sense, the method based on (1.5) can be viewed as a semiparametric method of type iv)-vi) in which the parametric fit is the optimal linear predictor.

Our approach is also related to papers dealing with nonparametric regression, or density estimation, in the presence of measurement error. Recent references on this topic are Carroll, Maca and Ruppert (1999) and Schennach (2004), among others. In contrast to those articles, the measurement error in our framework, that is the difference between the innovation and the residual is not independent from the latter. More importantly, it also depends on the sample size. Finally, it does not only concern the regressors but also the regressand.

The paper is organized as follows. In Section 2 we consider nonparametric density estimation and nonparametric regression for noisy data. Consistency and asymptotic normality are established under mixing assumptions on the unobserved stationary process, and a control of the size of the noise in the data. Section 3 uses these results for Kernel estimators based on ARMA residuals. Hybrid predictors, combining parametric and nonparametric techniques, are studied. In Section 4, their finite sample properties are investigated by means of simulations. An application to stock return data is also presented.

Concluding remarks are given in Section 5.

The symbol $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution. For any function $f : \mathbb{R}^d \to \mathbb{R}$, let $D_{i_1 \dots i_k} f(x) = (\partial^k f / \partial x_{i_1} \dots \partial x_{i_k})(x)$. The notation $o_P(1)$ is used for a sequence of random vectors that converge to zero in probability. The notation $R_n = O_P(S_n)$ means that $R_n = S_n T_n$ for a sequence T_n which is bounded in probability.

2 Kernel estimators applied to noisy data

In this section we study the behavior of the kernel density and regression estimators when they are computed on noisy data. This will be applied to ARMA residuals, considered as noisy innovations. The section may however have its own interest.

Consider a strictly stationary process $Z = (Z_t)_{t \in \mathbb{Z}}$ where $Z_t = (Y_t', X_t')'$, with $Y_t \in \mathbb{R}^{d_0}$ and $X_t \in \mathbb{R}^d$. Let $g : \mathbb{R}^{d_0} \to \mathbb{R}$ be a measurable function. Our goal is to estimate the regression

$$r(x) = E\ (g(Y_t) \mid X_t = x)$$

which is assumed to exist. We do not observe the process (Z_t) but, instead, we have n consecutive noisy observations of the form

$$\tilde{Z}_{1,n}, \dots, \tilde{Z}_{n,n} \quad \text{with} \quad \tilde{Z}_{t,n}' = (\tilde{Y}_{t,n}', \tilde{X}_{t,n}') := (Y_t' + V_{t,n}', X_t' + U_{t,n}')$$

where the $V_{t,n}$ and $U_{t,n}$ are disturbance terms. Observe that the "noise" is present both in the regressand and the regressors.

Let $f = f_X$, f_Y and f_Z be the densities of X_t, Y_t and Z_t. Since the seminal paper by Rosenblatt (1956), kernel estimators have been extensively employed for nonparametrically estimating f and r. Given a kernel $K : \mathbb{R}^d \to \mathbb{R}$ and a sequence of scalar bandwidths $(b_n) > 0$, the kernel density estimator of $f(x)$ is defined by

$$\tilde{f}(x) = \frac{1}{n b_n^d} \sum_{t=1}^{n} K\left(\frac{x - \tilde{X}_{t,n}}{b_n}\right). \tag{2.1}$$

When $\tilde{f}(x) \neq 0$, the regression $r(x)$ can be estimated by the Nadaraya-Watson estimator

$$\tilde{r}(x) = \frac{\tilde{\varphi}(x)}{\tilde{f}(x)}, \qquad \tilde{\varphi}(x) = \frac{1}{n b_n^d} \sum_{t=1}^{n} g(\tilde{Y}_{t,n}) K\left(\frac{x - \tilde{X}_{t,n}}{b_n}\right). \tag{2.2}$$

It will be convenient to consider the pseudo-estimators

$$\hat{f}(x) = \frac{1}{n b_n^d} \sum_{t=1}^{n} K\left(\frac{x - X_t}{b_n}\right), \qquad \hat{r}(x) = \frac{\hat{\varphi}(x)}{\hat{f}(x)}, \qquad \hat{\varphi}(x) = \frac{1}{n b_n^d} \sum_{t=1}^{n} g(Y_t) K\left(\frac{x - X_t}{b_n}\right) \tag{2.3}$$

based on the non-observable variables $Z_1, \dots, Z_n$.

The main asymptotic properties of the latter estimators, when $Z_1, \dots, Z_n$ are observed, are available in the statistical literature (see e.g. the monographs by Prakasa Rao (1983), Fan and Yao (2003)). We start by examining the consistency properties of the density and regression estimators when applied to noisy data.

2.1 Consistency

Establishing consistency requires some technical assumptions which we are listing here. Let $\|\cdot\|$ denote any norm on $\mathbb{R}^d$ or $\mathbb{R}^{d_0}$.

A1: K is a density with respect to the Lebesgue measure, $\int_{\mathbb{R}^d} uK(u)du = 0$, $\int_{\mathbb{R}^d} \|u\|^2 K(u)du < \infty$ and $\lim_{\|u\|\to\infty} \|u\|^d K(u) = 0$. In addition K satisfies the Lipschitz condition $|K(u) - K(v)| < C\|u - v\|$ for some constant C.

A2: The strong mixing coefficients of the process Z, defined by

$$\alpha_Z(k) = \sup_{A\in\sigma(Z_u, u\le t),\, B\in\sigma(Z_u, u\ge t+k)} |P(A\cap B) - P(A)P(B)|,$$

are such that

$$\sum_{h=0}^{\infty} \{\alpha_Z(h)\}^{\frac{\nu}{2+\nu}} < \infty \quad \text{for some } \nu > 0.$$

A3: The vector $x \in \mathbb{R}^d$ is such that $f(x) > 0$. The functions f_Z, f and $\varphi := rf$ are twice derivable with continuous and bounded second order derivatives. We have $E|g(Y_t)|^{2+\nu} < \infty$, and $\sup_u \int |g(y)|^{2+\nu} f_Z(y, u)dy < \infty$, where ν is defined in **A2**. There exists some constant C such that $|g(y') - g(y)| < C\|y' - y\|$ for all $y, y' \in \mathbb{R}^{d_0}$.

A4: $b_n \to 0$ and $nb_n^{d\left(1+\frac{\nu}{2+\nu}\right)} \to \infty$ as $n \to \infty$, for the constant $\nu > 0$ involved in **A2**–**A3**.

An assumption similar to $\sup_x \int |g(y)|^{2+\nu} f_Z(y, x)dy < \infty$ is also made in Mack and Silverman (1982). Note that Pham (1986) and Carrasco and Chen (2002) have shown that, for a wide class of processes, the mixing conditions made in Assumption **A2** are satisfied.

On the disturbance terms, we make the following assumption.

B1: There exists $\tau < 1$ such that

$$\sum_{t=1}^{n} \|U_{t,n}\| + \sum_{t=1}^{n} \|V_{t,n}\| + \sum_{t=1}^{n} (\|Y_t\| + \|V_{t,n}\|) \|U_{t,n}\| = O_P(n^\tau).$$

We will see that this situation arises with $\tau = 1/2$ when the kernel estimators are applied to residuals of parametric models.

Theorem 2.1 *Under Assumptions* **A1**–**A4** *and* **B1**, *the kernel density and regression estimators based on the noisy observations satisfy*

$$\tilde{f}(x) = f(x) + o_P(1) \qquad \text{and} \qquad \tilde{r}(x) = r(x) + o_P(1), \tag{2.4}$$

whenever $nb_n^{(1+d)/(1-\tau)} \to \infty$. *When* $b_n = cn^{-1/\{d+4+d\nu/(2+\nu)\}}$, $c > 0$, *and* $\tau < \{1 + d\nu/(2+\nu)\} / \{d+4+d\nu/(2+\nu)\}$,

$$\begin{aligned} \tilde{f}(x) - f(x) &= O_P\left(n^{-2/\{d+4+d\nu/(2+\nu)\}}\right), & (2.5)\\ \tilde{r}(x) - r(x) &= O_P\left(n^{-2/\{d+4+d\nu/(2+\nu)\}}\right). & (2.6)\end{aligned}$$

In the proof below we use the fact that under the assumptions of this proposition, except **B1**, (2.4), (2.5), and (2.6) hold for the pseudo-estimators obtained by replacing $\tilde{f}(x)$ and $\tilde{r}(x)$ by $\hat{f}(x)$ and $\hat{r}(x)$. Such a result is standard and can be obtained under many other assumptions (see *e.g.* Bosq (1996) or Härdle (1990)). This proposition thus shows that the asymptotic behavior of the kernel estimators is not affected by the presence of small disturbances. Note also that for an exponential mixing rate $\alpha_Z(h) = O(\rho^h)$ with $\rho \in (0,1)$, the constant $\nu > 0$ can be chosen arbitrarily small, so that the rate of convergence (2.5)–(2.6) is arbitrarily close to the optimal rate $O_P\left(n^{-2/(d+4)}\right)$ of the case of identically and independently distributed (iid) variables (Z_t).

Proof. In this proof and the subsequent ones, C denotes a generic positive constant whose exact value is unimportant and may vary from line to line.

Under **A1**–**A4** we have[1]

$$\hat{f}(x) = f(x) + o_P(1) \quad \text{and} \quad \hat{r}(x) = r(x) + o_P(1) \tag{2.7}$$

and for $b_n = cn^{-1/\{d+4+d\nu/(2+\nu)\}}$, $c > 0$

$$\max\{\hat{f}(x) - f(x), \hat{r}(x) - r(x)\} = O_P\left(n^{-2/\{d+4+d\nu/(2+\nu)\}}\right). \tag{2.8}$$

[1] For the convenience of the reader, and also because we have not been able to find a reference establishing these results with exactly the same assumptions, a complete proof of (2.7)–(2.8) is available from the authors.

Now, by **A1** and **B1** we have

$$\begin{aligned}\left|\tilde{f}(x)-\hat{f}(x)\right| &\leq \frac{1}{nb_n^d}\sum_{t=1}^{n}\left|K\left(\frac{x-X_t-U_{t,n}}{b_n}\right)-K\left(\frac{x-X_t}{b_n}\right)\right| \\ &\leq \frac{C}{nb_n^d}\sum_{t=1}^{n}\left\|\frac{U_{t,n}}{b_n}\right\| = O_P\left(n^{\tau-1}b_n^{-d-1}\right). \qquad (2.9)\end{aligned}$$

The right-hand side of the last equality tends to zero when $nb_n^{(1+d)/(1-\tau)} \to \infty$. The first consistency in (2.4) then follows from (2.7) and the triangular inequality

$$\left|\tilde{f}(x)-f(x)\right| \leq \left|\tilde{f}(x)-\hat{f}(x)\right| + \left|\hat{f}(x)-f(x)\right|.$$

In view of this inequality and (2.8), the optimal rate is reached since $n^{\tau-1}b_n^{-d-1} = o(n^{-2/\{d+4+d\nu/(2+\nu)\}})$ for $b_n = cn^{-1/\{d+4+d\nu/(2+\nu)\}}$, which entails (2.5).

We now consider the consistency of the regressor $\tilde{r}(x) = \tilde{\varphi}(x)/\tilde{f}(x)$. First note that (2.7) implies

$$|\hat{\varphi}(x)-\varphi(x)| = o_P(1). \qquad (2.10)$$

We have

$$\begin{aligned}\tilde{\varphi}(x)-\hat{\varphi}(x) &= \frac{1}{nb_n^d}\sum_{t=1}^{n}\{g(Y_t+V_{t,n})-g(Y_t)\}K\left(\frac{x-X_t}{b_n}\right) \\ &+ \frac{1}{nb_n^d}\sum_{t=1}^{n}g(Y_t+V_{t,n})\left\{K\left(\frac{x-X_t-U_{t,n}}{b_n}\right)-K\left(\frac{x-X_t}{b_n}\right)\right\}. \qquad (2.11)\end{aligned}$$

The assumptions made in **A1** entail that $\overline{K} := \sup_u |K(u)| < \infty$. By the Lipschitz condition in **A3** we have $|g(x)| = |g(x)-g(0)+g(0)| \leq C(\|x\|+1)$. Using **A3**, we then obtain

$$\begin{aligned}|\hat{\varphi}(x)-\tilde{\varphi}(x)| &\leq \frac{C}{nb_n^d}\sum_{t=1}^{n}K\left(\frac{x-X_t}{b_n}\right)\|V_{t,n}\| + \frac{C}{nb_n^{d+1}}\sum_{t=1}^{n}|g(Y_t+V_{t,n})|\,\|U_{t,n}\| \\ &\leq \frac{C\overline{K}}{nb_n^d}\sum_{t=1}^{n}\|V_{t,n}\| + \frac{C^2}{nb_n^{d+1}}\sum_{t=1}^{n}\left(\|Y_t\|+\|V_{t,n}\|+1\right)\|U_{t,n}\|.\end{aligned}$$

Thus, in view of **B1**,

$$|\hat{\varphi}(x)-\tilde{\varphi}(x)| = O_P\left(n^{\tau-1}b_n^{-d-1}\right). \qquad (2.12)$$

The consistency of the numerator of $\tilde{r}(x) = \tilde{\varphi}(x)/\tilde{f}(x)$ follows from (2.10) and (2.12). The consistency of the denominator has already been established. The second equality of (2.4) then follows by Slutsky's lemma.

From (2.8), (2.12) and the consistency of $\hat{f}(x)$, we obtain

$$\begin{aligned}\tilde{\varphi}(x)-\varphi(x) &= \tilde{\varphi}(x)-\hat{\varphi}(x)+\{\hat{r}(x)-r(x)\}\hat{f}(x)+r(x)\left\{\hat{f}(x)-f(x)\right\}\\ &= O_P\left(n^{-2/\{d+4+d\nu/(2+\nu)\}}\right)\end{aligned}$$

under the assumptions of the theorem on b_n and τ. Under the same conditions,

$$\begin{aligned}\tilde{r}(x)-r(x) &= \frac{\tilde{\varphi}(x)-\varphi(x)}{\tilde{f}(x)}+\varphi(x)\frac{f(x)-\tilde{f}(x)}{f(x)\tilde{f}(x)}\\ &= O_P\left(n^{-2/\{d+4+d\nu/(2+\nu)\}}\right)\end{aligned}$$

which completes the proof. □

2.2 Asymptotic normality of the regression estimator

The previous assumptions are strengthened as follows.

A1': Assumption **A1** holds and the Kernel function K is three times differentiable with bounded derivatives, and the first two derivatives are integrable.

A2':

$$\sum_{h=0}^{\infty} h^{\varrho}\{\alpha_Z(h)\}^{\frac{\nu}{2+\nu}}<\infty, \quad \text{where } \nu>0 \text{ and } \varrho>\tfrac{\nu}{2+\nu}.$$

A3': Assumption **A3** holds with $f_Z(y,x)\leq Cf_Y(y)$, for some positive constant C, for all $x\in\mathbb{R}^d, y\in\mathbb{R}^{d_0}$. For all $h>0$, the vector $(Z_t',Z_{t-h}')=(Y_t',X_t',Y_{t-h}',X_{t-h}')$ admits a continuous density f_{Z_h,Z_0} such that $f_{Z_h,Z_0}(y,x,\tilde{y},\tilde{x})\leq Cf_{Y_h,Y_0}(y,\tilde{y})$, for some positive constant C, for all $x,\tilde{x}\in\mathbb{R}^d, y,\tilde{y}\in\mathbb{R}^{d_0}$, where f_{Y_h,Y_0} denotes the density of (Y_t',Y_{t-h}').

When $f_Z(y,x)\leq Cf_Y(y)$, the function $f_Z(y,x)$ is said to be uniformly in the order of $f_Y(y)$. Note that this assumption, together with $E|g(Y_t)|^{2+\nu}<\infty$, entails the condition $\sup_u\int|g(y)|^{2+\nu}f_Z(y,u)dy<\infty$ in **A3**. When Z_t is gaussian and $Y_t\in\mathbb{R}$, one can take

$$C=(2\pi)^{-d_0/2}\left|\text{Var}(Y_t)-\text{Cov}(Y_t,X_t)\text{Var}(X_t)^{-1}\text{Cov}(X_t,Y_t)\right|^{-1/2}.$$

A4': For the constants ϱ and ν involved in **A2'** and **A3** we have, as $n\to\infty$,

$$nb_n^{4+d}\to 0,\quad nb_n^{\frac{\nu(4+d)}{\varrho(2+\nu)}}\to\infty,\quad nb_n^{d\left(1+\frac{\nu}{2+\nu}\right)}\to\infty.$$

For a real random variable X and a constant $s>0$, let $\|X\|_s = \left\{\int |x|^s dP(x)\right\}^{1/s}$. For a random vector $X=(X_1,\dots,X_k)'$, let $\|X\|_s = \left\|\sum_{i=1}^k |X_i|\right\|_s$.

B2: There exist nonnegative numbers τ_0 and τ_1, positive constants ζ_1, ζ_2, γ_1, γ_2, γ_3 with $\frac{1}{\zeta_1}+\frac{1}{\zeta_2}=1$ and $\frac{1}{\gamma_1}+\frac{1}{\gamma_2}+\frac{1}{\gamma_3}=1$, sequences of positive random variables (μ_t), (ρ_t), (h_n), (k_n), $(u_{t,n})$ and $(v_{t,n})$, a constant $C>0$ and a constant $\rho\in(0,1)$, a sequence of integers (a_n), such that

$$a_n \to \infty \quad \text{and} \quad a_n = o(\sqrt{nb_n^d}) \quad \text{as} \quad n\to\infty,$$

and

$$\|V_{t,n}\| \le \rho_t + k_n v_{t,n}, \quad \|U_{t,n}\| \quad \le \quad \mu_t + h_n u_{t,n}, \tag{2.13}$$

where

$$k_n = O_P(n^{-\tau_0}), \quad h_n = O_P(n^{-\tau_1}), \quad \max\{\rho_t,\mu_t\} \le C\rho^t \text{ a.s.},$$
$$\max\left\{\|Y_t\|_{\zeta_1}, \|v_{t,n}\|_{\zeta_1}, \left\|u_{t,n}^3\right\|_{\zeta_2}, \|v_{t,n}\|_{\gamma_1}, \|Y_t\|_{\gamma_1}, \left\|u_{t,n}^2\right\|_{\gamma_2}\right\} \le C.$$

In the following result, we establish the asymptotic distribution of $\hat{r}(x)$ under two sets of assumptions. The first one is simpler, but it cannot be used for the application developed in Section 3 since, as we shall see, the real number involved in **B1** is $\tau = 1/2$ when the noisy data consist of ARMA residuals.

Theorem 2.2 *Assume* **A1'-A4'** *and either*

1. **B1** *with* $n^{2\tau-1}b_n^{-(2+d)} \to 0$,

or

2. **B2** *with, for* $k=1,2,$

$$\begin{array}{llll}
1) & n^{-\tau_0+\frac{1}{2}} b_n^{\frac{d}{2}(\frac{1}{\zeta_2}-\frac{1}{\zeta_1})} \to 0, & 2) & n\rho^{2a_n} b_n^{-(2k+d)+\frac{2d}{\zeta_2}} \to 0,\\
3) & n^{-1}\rho^{2a_n} b_n^{-(2k+d)} \to 0, & 4) & n^{1-2k\tau_1} b_n^{-(2k+d)+\frac{2d}{\gamma_3}} \to 0,\\
5) & n^{1-2\tau_0}\rho^{2a_n} b_n^{-(2k+d)+\frac{2d}{\zeta_2}} \to 0, & 6) & n^{1-2(\tau_0+k\tau_1)} b_n^{-(2k+d)+\frac{2d}{\gamma_3}} \to 0,\\
7) & n\rho^{2a_n} b_n^{-(6+d)} \to 0, & 8) & n^{1-6\tau_1} b_n^{-(6+d)} \to 0,\\
9) & n^{1-2\tau_0}\rho^{2a_n} b_n^{-(6+d)} \to 0, & 10) & n^{1-2(\tau_0+3\tau_1)} b_n^{-(6+d)} \to 0.
\end{array} \tag{2.14}$$

Then, letting $F_2(x) = \varphi_2(x) - f(x)r^2(x)$,

$$\sqrt{nb_n^d}\,\{\tilde{r}(x) - r(x)\} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{F_2(x)}{f^2(x)}\int_{\mathbb{R}^d} K^2(u)du\right). \tag{2.15}$$

It should be noted that, under **A1'-A4'**, the asymptotic distribution of $\hat{r}(x)$ is exactly the same as in (2.15). Hence, the asymptotic behavior of regression estimator is not affected by the presence of a "small" noise in the data.

Before proving this result, we establish the following technical lemma.

Lemma 2.1 *Let (X_t) be as in Section 2 and let the density f of X_t satisfy* **A3**. *Let $\gamma > 1$ and $H : \mathbb{R}^d \to \mathbb{R}^+$ such that $\int_{\mathbb{R}^d} H^\gamma(t)\,dt < \infty$, and let (b_n) be a sequence of positive numbers such that $b_n \to 0$ as $n \to \infty$. Then we have*

$$\left\| H\left(\frac{x - X_t}{b_n}\right) \right\|_\gamma = O\left(b_n^{\frac{d}{\gamma}}\right).$$

Proof. We have, by the change of variable formula

$$\left\| H\left(\frac{x - X_t}{b_n}\right) \right\|_\gamma = \left\{ \int_{\mathbb{R}^d} H^\gamma\left(\frac{x-u}{b_n}\right) f(u)du \right\}^{\frac{1}{\gamma}} = \left\{ b_n^d \int_{\mathbb{R}^d} H^\gamma(t) f(x - b_n t)dt \right\}^{\frac{1}{\gamma}}.$$

Note that f is bounded under **A3**. When $n \to \infty$ the latter integral converges to $f(x) \int_{\mathbb{R}^d} H^\gamma(t)\,dt$ by the dominated convergence theorem. The conclusion follows. □

Proof of Theorem 2.2. Asymptotic normality of regression estimators under strong mixing assumptions were first established by Robinson (1983). Under **A1** and **A2'-A4'** we have[2]

$$\sqrt{nb_n^d}\,\{\hat{r}(x) - r(x)\} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{F_2(x)}{f^2(x)} \int_{\mathbb{R}^d} K^2(u)du\right). \tag{2.16}$$

We have

$$\hat{r}(x) - \tilde{r}(x) = \frac{\hat{\varphi}(x) - \tilde{\varphi}(x)}{\hat{f}(x)} - \tilde{\varphi}(x)\frac{\hat{f}(x) - \tilde{f}(x)}{\hat{f}(x)\tilde{f}(x)}. \tag{2.17}$$

Under **B1**, by (2.9), (2.12) and (2.17),

$$\sqrt{nb_n^d}(\hat{r}(x) - \tilde{r}(x)) = O_P\left(n^{\tau-1/2} b_n^{-d/2-1}\right) = o_P(1). \tag{2.18}$$

Thus (2.15) follows, in view of (2.16), under the first set of assumptions.

Now suppose that **B2** holds. We will show that $\sqrt{nb_n^d}(\hat{r}(x) - \tilde{r}(x))$ converges to zero in probability, which, by (2.16), will be sufficient to prove (2.15). In view of (2.17) it suffices to prove that

$$\sqrt{nb_n^d}\,|\hat{\varphi}(x) - \tilde{\varphi}(x)| = o_P(1) \quad \text{and} \quad \sqrt{nb_n^d}\left|\hat{f}(x) - \tilde{f}(x)\right| = o_P(1). \tag{2.19}$$

[2] A proof is available from the authors.

In view of (2.11) and using **A3** we have

$$\begin{aligned}\sqrt{nb_n^d}\,|\hat{\varphi}(x)-\tilde{\varphi}(x)| &\leq \frac{C}{\sqrt{nb_n^d}}\sum_{t=1}^n K\left(\frac{x-X_t}{b_n}\right)\|V_{t,n}\| \\ &+\frac{C}{\sqrt{nb_n^d}}\sum_{t=1}^n |g(Y_t+V_{t,n})|\left|K\left(\frac{x-X_t-U_{t,n}}{b_n}\right)-K\left(\frac{x-X_t}{b_n}\right)\right| \\ :=\ & S_1+S_2. \end{aligned} \tag{2.20}$$

With the notation $\overline{K}=\sup_u |K(u)|$ introduced in the proof of Theorem 2.1, we have

$$\begin{aligned} S_1 &\leq \frac{C}{\sqrt{nb_n^d}}\sum_{t=1}^n K\left(\frac{x-X_t}{b_n}\right)(\rho_t+k_n v_{t,n}) \\ &\leq \frac{C\overline{K}}{\sqrt{nb_n^d}}\sum_{t=1}^n \rho_t+\frac{Ck_n}{\sqrt{nb_n^d}}\sum_{t=1}^n K\left(\frac{x-X_t}{b_n}\right)v_{t,n}. \end{aligned} \tag{2.21}$$

The first term in the right-hand side of the last inequality converges to 0 in probability by **B2** and **A4'** (implying $nb_n^d\to\infty$). Moreover, using successively the Hölder inequality, Lemma 2.1 and **B2** we find

$$\begin{aligned} & E\left\{\frac{1}{\sqrt{nb_n^d}}\sum_{t=1}^n K\left(\frac{x-X_t}{b_n}\right)v_{t,n}\right\} \\ \leq\ & \frac{C}{\sqrt{nb_n^d}}\sum_{t=1}^n \left\|K\left(\frac{x-X_t}{b_n}\right)\right\|_{\zeta_2}\|v_{t,n}\|_{\zeta_1} \\ \leq\ & \frac{Cb_n^{d/\zeta_2}}{\sqrt{nb_n^d}}\sum_{t=1}^n \|v_{t,n}\|_{\zeta_1}=O\left(n^{1/2}b_n^{\frac{d}{2}(\frac{1}{\zeta_2}-\frac{1}{\zeta_1})}\right). \end{aligned}$$

From **B2** and 1) in (2.14) we deduce that the second term in (2.21) converges to 0 in probability. It follows that

$$S_1=o_P(1). \tag{2.22}$$

To handle S_2 we will make a third-order Taylor expansion of the Kernel function around $(x-X_t)/b_n$. Write $U_{t,n}=(U_{1,t,n},\ldots,U_{d,t,n})'$. We have

$$\begin{aligned} K\left(\frac{x-X_t-U_{t,n}}{b_n}\right) &= K\left(\frac{x-X_t}{b_n}\right)-\frac{1}{b_n}\sum_{i=1}^d U_{i,t,n}D_iK\left(\frac{x-X_t}{b_n}\right) \\ &+\frac{1}{2b_n^2}\sum_{i,j=1}^d U_{i,t,n}U_{j,t,n}D_{ij}K\left(\frac{x-X_t}{b_n}\right) \\ &-\frac{1}{6b_n^3}\sum_{i,j,k=1}^d U_{i,t,n}U_{j,t,n}U_{k,t,n}D_{ijk}K\left(\frac{x-X_{t,n}}{b_n}\right), \end{aligned}$$

where $X_{t,n}$ is between X_t and $X_t + U_{t,n}$. Thus, using the vector norm $\|a\| = \sum |a_i|$,

$$\left| K\left(\frac{x - X_t - U_{t,n}}{b_n}\right) - K\left(\frac{x - X_t}{b_n}\right)\right| \leq \frac{1}{b_n}\|U_{t,n}\| \sum_{i=1}^{d} \left| D_i K\left(\frac{x - X_t}{b_n}\right)\right|$$

$$+\frac{1}{2b_n^2}\|U_{t,n}\|^2 \sum_{i,j=1}^{d} \left| D_{ij} K\left(\frac{x - X_t}{b_n}\right)\right| + \frac{1}{6b_n^3}\|U_{t,n}\|^3 \sum_{i,j,k=1}^{d} \left| D_{ijk} K\left(\frac{x - X_{t,n}}{b_n}\right)\right|.$$

Hence, with the elementary inequality $(|a| + |b|)^k \leq 2^k(|a|^k + |b|^k)$,

$$\begin{aligned} S_2 \quad &\leq \quad \frac{C}{\sqrt{nb_n^d}} \sum_{t=a_n}^{n} \frac{1}{b_n} \left| g(Y_t + V_{t,n}) \right| (\mu_t + h_n u_{t,n}) \sum_{i=1}^{d} \left| D_i K\left(\frac{x - X_t}{b_n}\right)\right| \\ &\quad + \frac{C}{\sqrt{nb_n^d}} \sum_{t=a_n}^{n} \frac{1}{b_n^2} \left| g(Y_t + V_{t,n}) \right| (\mu_t^2 + h_n^2 u_{t,n}^2) \sum_{i,j=1}^{d} \left| D_{ij} K\left(\frac{x - X_t}{b_n}\right)\right| \\ &\quad + \frac{C}{\sqrt{nb_n^d}} \sum_{t=a_n}^{n} \frac{1}{b_n^3} \left| g(Y_t + V_{t,n}) \right| (\mu_t^3 + h_n^3 u_{t,n}^3) \sum_{i,j,k=1}^{d} \left| D_{ijk} K\left(\frac{x - X_{t,n}}{b_n}\right)\right| \\ &\quad + \frac{C}{\sqrt{nb_n^d}} \sum_{t=1}^{a_n} (1 + \|Y_t\| + \|V_{t,n}\|) \\ := \quad & S_{21} + S_{22} + S_{23} + S_{24}, \end{aligned} \tag{2.23}$$

where the last term follows from the fact that the kernel function is bounded and from the Lipschitz condition on g. We have

$$\begin{aligned} S_{21} \quad &\leq \quad \frac{C}{b_n\sqrt{nb_n^d}} \sum_{t=a_n}^{n} (1 + \|Y_t\| + \rho_t)\mu_t \sum_{i=1}^{d} \left| D_i K\left(\frac{x - X_t}{b_n}\right)\right| \\ &\quad + \frac{Ch_n}{b_n\sqrt{nb_n^d}} \sum_{t=a_n}^{n} (1 + \|Y_t\| + \rho_t)u_{t,n} \sum_{i=1}^{d} \left| D_i K\left(\frac{x - X_t}{b_n}\right)\right| \\ &\quad + \frac{Ck_n}{b_n\sqrt{nb_n^d}} \sum_{t=a_n}^{n} v_{t,n}\mu_t \sum_{i=1}^{d} \left| D_i K\left(\frac{x - X_t}{b_n}\right)\right| \\ &\quad + \frac{Ch_nk_n}{b_n\sqrt{nb_n^d}} \sum_{t=a_n}^{n} v_{t,n}u_{t,n} \sum_{i=1}^{d} \left| D_i K\left(\frac{x - X_t}{b_n}\right)\right| \\ := \quad & S_{211} + h_n S_{212} + k_n S_{213} + h_n k_n S_{214}. \end{aligned} \tag{2.24}$$

Because the derivatives of K are bounded and ρ_t and μ_t are $O(\rho^t)$ with probability one, we have

$$S_{211} \leq \frac{C\rho^{a_n}}{b_n\sqrt{nb_n^d}} \sum_{t=a_n}^{n} \|Y_t\| \sum_{i=1}^{d} \left| D_i K\left(\frac{x - X_t}{b_n}\right)\right| + O_P\left(\frac{\rho^{a_n}}{b_n\sqrt{nb_n^d}}\right). \tag{2.25}$$

Denote by S_{211}^* the first term of the right-hand side of this equality. It is easy to check that **A1'** entails $\int |D_i K(t)|^\gamma \, dt < \infty$ for any power $\gamma \geq 1$. Using Lemma 2.1 and by the

Hölder inequality, we have

$$\begin{aligned} E(S_{211}^{*}) &\leq \frac{C\rho^{a_n}}{b_n\sqrt{nb_n^d}}\sum_{t=a_n}^{n}\|Y_t\|_{\zeta_1}\sum_{i=1}^{d}\left\|D_iK\left(\frac{x-X_t}{b_n}\right)\right\|_{\zeta_2} \\ &\leq \frac{C\rho^{a_n}b_n^{\frac{d}{\zeta_2}}}{b_n\sqrt{nb_n^d}}\sum_{t=a_n}^{n}\|Y_t\|_{\zeta_1} \leq \frac{C\rho^{a_n}nb_n^{\frac{d}{\zeta_2}}}{b_n\sqrt{nb_n^d}} = o(1), \end{aligned}$$

where the last equality follows from 2) in (2.14) with $k=1$. Thus, noting that $S_{211}^{*}\geq 0$, we have $S_{211}^{*}=o_P(1)$. Using 3) with $k=1$ and (2.25) we then obtain $S_{211}=o_P(1)$. By Lemma 2.1 and the Hölder inequality, we have

$$\begin{aligned} E(S_{212}) &\leq \frac{1}{b_n\sqrt{nb_n^d}}\sum_{t=a_n}^{n}(1+\|Y_t\|_{\gamma_1}+\|\rho_t\|_{\gamma_1})\|u_{t,n}\|_{\gamma_2}\sum_{i=1}^{d}\left\|D_iK\left(\frac{x-X_t}{b_n}\right)\right\|_{\gamma_3} \\ &\leq \frac{Cb_n^{-1-\frac{d}{2}+\frac{d}{\gamma_3}}}{\sqrt{n}}\sum_{t=a_n}^{n}(1+\|Y_t\|_{\gamma_1}+C\rho^t)\|u_{t,n}\|_{\gamma_2} = O\left(b_n^{-1-\frac{d}{2}+\frac{d}{\gamma_3}}\sqrt{n}\right). \end{aligned}$$

Thus, 4) with $k=1$ in (2.14) entails $h_nS_{212}=o_P(1)$. For the same reasons, 5) with $k=1$ entails $k_nS_{213}=o_P(1)$, and 6) with $k=1$ entails $h_nk_nS_{214}=o_P(1)$. Thus, in view of (2.24), we have shown that $S_{21}=o_P(1)$. By exactly the same arguments, 2)-6) with $k=2$ entail $S_{22}=o_P(1)$. For S_{23} we use the boundedness of the third derivatives (**A1'**) and conclude similarly using the convergence 7)-10) in (2.14). Finally,

$$S_{24}\leq \frac{C}{\sqrt{nb_n^d}}\sum_{t=1}^{a_n}(1+\|Y_t\|+\rho_t)+\frac{Ck_n}{\sqrt{nb_n^d}}\sum_{t=1}^{a_n}(1+\|Y_t\|+v_{t,n}) := S_{241}+k_nS_{242}.$$

We have

$$\|S_{24i}\|_1 \leq \frac{Ca_n}{\sqrt{nb_n^d}} = o(1), \quad i=1,2$$

which proves that $S_{24i}=o_P(1)$ and thus that $S_{24}=o_P(1)$. Therefore we have

$$S_2 = o_P(1). \tag{2.26}$$

In view of (2.22) and (2.20) this proves that the first equality in (2.19) holds. The second equality follows along the same lines and the proof of Theorem 2.2 is complete. □

3 Kernel estimators applied to ARMA residuals

Many non linear processes admit ARMA representations (see *e.g.* Romano and Thombs (1996), Francq, Roy and Zakoïan (2005) and the references therein). This is not very surprising because, from the Wold theorem, any purely non deterministic second-order stationary process (X_t) has an infinite MA representation, which can be closely approximated by finite order ARMA models. The noise in these ARMA representations is only the linear innovation of (X_t) and is not an iid sequence (otherwise (X_t) would be a linear process). These representations are referred to as weak ARMA representations, by opposition to the standard strong ARMA representations where the noise is supposed to be iid.

In this section we will show how the weak ARMA residuals can be used in nonparametric predictors. Density estimators based on residuals of time series models have already been studied by Robinson (1983) and Liebscher (2001), among others. Such residual-based estimators can be used to obtain adaptive estimators (see Drost Klaassen and Werker, 1997) and to obtain $\sqrt{n}$-consistent plug-in estimators for functionals of a density (see Schick and Wefelmeyer, 2004). Our framework here is quite different, since we study the asymptotic behavior of kernel estimators of autoregressions when lagged values of ARMA residuals are taken as explanatory variables (in Section 3.2) and when the ARMA residuals constitute the dependent variable (in Section 3.3).

3.1 Weak ARMA residuals

We now introduce the assumptions we need. Let $W = (W_t)_{t\in\mathbb{Z}}$ be a real second-order stationary ARMA(p,q) process such that

$$W_t + \sum_{i=1}^{p} \phi_i W_{t-i} = \epsilon_t + \sum_{i=1}^{q} \psi_i \epsilon_{t-i}, \quad \forall t \in \mathbb{Z}. \tag{3.1}$$

The parameter $\theta_0 = (\phi_1, \ldots, \phi_p, \psi_1, \ldots, \psi_q)'$ is unknown and we observe a sample $W_1, W_2, \ldots, W_n$ of W. For any $\theta = (\theta_1, \ldots, \theta_{p+q})' \in \mathbb{R}^{p+q}$, define two polynomials by $\Phi_\theta(z) = 1 + \theta_1 z + \cdots + \theta_p z^p$ and $\Psi_\theta(z) = 1 + \theta_{p+1} z + \cdots + \theta_{p+q} z^q$. For any $\delta > 0$, let the compact set

$$\Theta_\delta = \{\theta \in \mathbb{R}^{p+q};\ \text{ the zeros of } \Phi_\theta(z) \text{ and } \Psi_\theta(z) \text{ have moduli} \geq 1 + \delta\}.$$

We make the following assumptions.

C1. $\epsilon = (\epsilon_t)$ is a strictly stationary sequence of uncorrelated random variables with zero mean and variance $\sigma^2 > 0$.

C2. θ_0 belongs to the interior of Θ_δ, and the polynomials $\Phi_{\theta_0}(z)$ and $\Psi_{\theta_0}(z)$ have no zero in common.

C3. ϕ_p and ψ_q are not both equal to zero (by convention $\phi_0 = \psi_0 = 1$).

The sequence of the $\alpha-$mixing (strong mixing) coefficients of some process $(U) = (U_t)_{t\in\mathbb{Z}}$ is denoted by $\{\alpha_U(h)\}_{h\geq 0}$. We will consider, alternatively, the following mixing assumptions.

C4. The process (W, ϵ) satisfies Model (3.1), the moment condition $E|W_t|^{4+2\nu} < \infty$ (or equivalently $E|\epsilon_t|^{4+2\nu} < \infty$) and the mixing condition

$$\sum_{h=0}^{\infty} \{\alpha_{W,\epsilon}(h)\}^{\frac{\nu}{2+\nu}} < \infty \quad \text{for some } \nu > 0.$$

C5. The process (W, ϵ) satisfies Model (3.1), with $E|W_t|^{4+2\nu} < \infty$ and

$$\sum_{h=0}^{\infty} \{\alpha_{\epsilon}(h)\}^{\frac{\nu}{2+\nu}} < \infty \quad \text{for some } \nu > 0.$$

For all $\theta \in \Theta$, let $\epsilon_t(\theta) = \Psi_\theta^{-1}(B)\Phi_\theta(B)W_t$ and $e_t(\theta) = \Psi_\theta^{-1}(B)\Phi_\theta(B)(W_t 1_{1\leq t\leq n})$, where B denotes the backshift operator. Note that $\epsilon_t(\theta)$ is not computable from the sample, but it is introduced for theoretical purpose. Let $\hat{\theta}_n$ be a Least Squares Estimator (LSE) satisfying, almost surely,

$$Q_n(\hat{\theta}_n) = \min_{\theta\in\Theta_\delta} Q_n(\theta) \quad \text{where} \quad Q_n(\theta) = \frac{1}{2n}\sum_{t=1}^{n} e_t^2(\theta). \tag{3.2}$$

The ARMA(p, q) residuals are then defined by $\hat{\epsilon}_t = e_t(\hat{\theta}_n)$ for $t = 1, \ldots, n$. Francq and Zakoïan (1998) have shown that the LSE is strongly consistent and asymptotically normal under **C1-C4**. We will need the additional technical lemma, giving the behavior of the weak ARMA residuals.

Lemma 3.1 *If* **C1-C3** *and either* **C4** *or* **C5** *hold then*

$$\hat{\epsilon}_t = \epsilon_t + s_t + O_P(n^{-1/2}) \quad \textit{with} \quad |s_t| \leq C\rho^t, \tag{3.3}$$

where the constants $\rho \in (0, 1)$ and C only depend on the initial values W_0, ..., W_{1-p}, ϵ_0, ..., ϵ_{1-q}. Moreover $\sum_{t=1}^{n} |\epsilon_t - \hat{\epsilon}_t| = O_P(n^{1/2})$.

Proof. Write $s_t = s_t(\theta_0)$, where $s_t(\theta) = e_t(\theta) - \epsilon_t(\theta)$. A Taylor expansion of $e_t(\cdot)$ around θ_0 yields

$$\hat{\epsilon}_t = e_t(\hat{\theta}_n) = s_t + \epsilon_t(\theta_0) + (\hat{\theta}_n - \theta_0)'\frac{\partial e_t}{\partial\theta}(\theta^*),$$

where θ^* is between $\hat{\theta}_n$ and θ_0. It is shown in Francq and Zakoïan (1998, Lemma 1 and Theorem 2, and 2000, Lemmas A.1 and A.2) that $\sup_{\theta\in\Theta}|s_t(\theta)| \leq C\rho^t$, where C is a measurable function of the initial values, that $\frac{\partial e_t}{\partial \theta_j}(\theta) = \sum_{i=1}^{t-1} c_{i,j}(\theta)W_{t-i}$ with $c_i := \max_{j\in\{1,\dots,p+q\}}\sup_{\theta\in\Theta}\|c_{i,j}(\theta)\| = O(\rho^i)$, and that $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_P(1)$ under **C4**. The same results can be obtained following the same lines of proof under **C5**. The proof of (3.3) follows. Now

$$\sum_{t=1}^{n}|\epsilon_t - \hat{\epsilon}_t| \leq \sum_{t=1}^{n}|s_t| + \left\|\hat{\theta}_n - \theta_0\right\|\sum_{t=1}^{n}\sum_{i\geq 1} c_i\,|W_{t-i}|\,, \tag{3.4}$$

and the conclusion follows from the $\sqrt{n}$-consistency of $\hat{\theta}_n$ and the fact that $E\sum_{t=1}^{n}\sum_{i\geq 1} c_i\,|W_{t-i}| \leq Cn\sum_{i\geq 1}\rho^i = O(n)$. □

3.2 Nonparametric prediction based on past observables and ARMA residuals

Because we will now consider different regressions, we need to modify the notation of Section 2 for the regression functions $r(x)$ and $\tilde{r}(x)$. Let ℓ and m be two integers such that $d := \ell + m \neq 0$. Recall that the $\hat{\epsilon}_t$ are the residuals of the LSE of the weak ARMA(p,q) model (3.1). With some abuse of notation and obvious conventions, write $\tilde{r}\{W_t \mid (W_{t-1},\dots,W_{t-\ell},\hat{\epsilon}_{t-1},\dots,\hat{\epsilon}_{t-m}) = x\}$ for the kernel estimator of the regression of W_t on $W_{t-1},\dots,W_{t-\ell},\hat{\epsilon}_{t-1},\dots,\hat{\epsilon}_{t-m}$ evaluated at $x = (x_1,\dots,x_d)'$. For any real sequence (X_t) and $k < \ell$, let $\mathbf{X}_{t-k}^{t-\ell} = (X_{t-k}, X_{t-k-1},\dots,X_{t-\ell})$.

Our first main result in this section is the following, showing that under mild regularity conditions, a kernel regression on ARMA residuals is equivalent to a (theoretical) regression on (non observed) linear innovations.

Theorem 3.1 *Assume that* **A1** *and* **C1**–**C4** *hold true and that for all integers d_1 and d_2, and all indices $t_1,\dots,t_{d_1},t'_1,\dots,t'_{d_1}$ of $\mathbb{Z}$, the vector $(W_{t_1},\dots,W_{t_{d_1}},\epsilon_{t'_1},\dots,\epsilon_{t'_{d_2}})$ has a strictly positive density f_Z which is uniformly in the order of each of its marginal densities. Assume also that the functions $f_Z(z_1,\dots,z_{d_1+d_2})$ and $(z_2,\dots,z_{d_1+d_2}) \mapsto \int z_1 f_Z(z_1,\dots,z_{d_1+d_2})dz_1$ are twice derivable, with continuous and bounded second order derivatives. Then*

(i) If $b_n \to 0$ and $nb_n^{2(1+d)} \to \infty$ as $n \to \infty$,

$$\tilde{r}\left\{W_t \mid (\mathbf{W}_{t-1}^{t-\ell}, \hat{\boldsymbol{\epsilon}}_{t-1}^{t-m}) = x\right\} \to E\left\{W_t \mid (\mathbf{W}_{t-1}^{t-\ell}, \boldsymbol{\epsilon}_{t-1}^{t-m}) = x\right\}, \quad \textit{in probability.}$$

(ii) If, in addition, **A1'**, **A2'** *and* **A4'** *hold with* $Z = (W, \epsilon)$, *if* $E|W_t|^s < \infty$ *with* $s \geq 4 + \nu$ *and* $s > 6d/(d-2)$, $d > 2$, *and if* $n^2 b_n^{6+d} \to \infty$ *hold,*

$$\begin{aligned}
&\sqrt{nb_n^d}\left[\tilde{r}\left\{W_t \mid (\mathbf{W}_{t-1}^{t-\ell}, \hat{\boldsymbol{\epsilon}}_{t-1}^{t-m}) = x\right\} - E\left\{W_t \mid (\mathbf{W}_{t-1}^{t-\ell}, \boldsymbol{\epsilon}_{t-1}^{t-m}) = x\right\}\right] \\
&\xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\operatorname{Var}\left\{W_t \mid (\mathbf{W}_{t-1}^{t-\ell}, \boldsymbol{\epsilon}_{t-1}^{t-m}) = x\right\}}{f(x)} \int_{\mathbb{R}^d} K^2(u)du\right)
\end{aligned}$$

where f *is the density of* $(\mathbf{W}_{t-1}^{t-\ell}, \boldsymbol{\epsilon}_{t-1}^{t-m})$.

Concerning the assumptions, the following remark can be made. For a purely non deterministic strong ARMA process with innovations admitting a positive density over $\mathbb{R}$, the vector $(W_t, \dots, W_{t-\ell}, \epsilon_{t-1}, \dots, \epsilon_{t-m})$ has also a positive density. This a not necessarily the case for a general weak ARMA model (3.1) because the distribution of W is not entirely defined by the distribution of ϵ and the ARMA coefficients.

Proof. To use the results of Section 2 we set

$$Z_t = (\mathbf{W}_t^{t-\ell}, \boldsymbol{\epsilon}_{t-1}^{t-m})', \quad \tilde{Z}_{t,n} = (\mathbf{W}_t^{t-\ell}, \hat{\boldsymbol{\epsilon}}_{t-1}^{t-m})', \quad U_{t,n} = (0, \dots, 0, \boldsymbol{\epsilon}_{t-1}^{t-m} - \hat{\boldsymbol{\epsilon}}_{t-1}^{t-m})',$$

$V_{t,n} = 0$, and $Y_t = W_t$. Using (3.4), we obtain

$$\begin{aligned}
&\sum_{t=1}^{n} \{\|U_{t,n}\| + \|V_{t,n}\| + (\|Y_t\| + \|V_{t,n}\|)\|U_{t,n}\|\} \\
\leq\ & \sum_{j=1}^{m}\sum_{t=1}^{n}\left\{C\rho^t + \left\|\hat{\theta}_n - \theta_0\right\| \sum_{i\geq 1} c_i |W_{t-j-i}|\right\}(1 + |W_t|) \\
=\ & O_P(n^{1/2}).
\end{aligned}$$

Thus **B1** holds with $\tau = 1/2$. Because $\alpha_Z(h) \leq \alpha_{W,\epsilon}(h - \max\{\ell, m\})$, Assumption **A2** is satisfied. Assumption **A3** is satisfied because of conditions assumed in the proposition. Since $d(1 + \nu/(2+\nu)) < 2(1+d)$ for all $\nu > 0$, **A4** is satisfied. Then, the convergence of probability follows from (2.4) of Theorem 2.1.

Now we turn to the asymptotic normality. In view of Theorem 2.2, it suffices to verify that assumptions **A3'**, **B2** and (2.14) are satisfied. Assumption **A3'** is directly implied by the conditions made in the proposition. Let us turn to **B2**. Obviously one can take $\rho_t = k_n = v_{t,n} = 0$ in (2.13). Thus τ_0 can be chosen arbitrarily large. To derive the second inequality of (2.13), let us write, as in the proof of Lemma 3.1,

$$\hat{\epsilon}_t - \epsilon_t = s_t + (\hat{\theta}_n - \theta_0)' \frac{\partial e_t}{\partial \theta}(\theta^*) = s_t - \frac{\partial Q_n}{\partial \theta'}(\theta_0) J_n^{-1} \frac{\partial e_t}{\partial \theta}(\theta^*)$$

where the matrix $J_n = \left[\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta^*_{n,ij})\right]$ is non-singular for sufficiently large n, the $\theta^*_{n,ij}$ and θ^* being between θ_0 and $\hat{\theta}_n$. Thus, using the multiplicativity of the norm $\|A\| = \sum |a_{ij}|$, we obtain

$$|\hat{\epsilon}_t - \epsilon_t| \quad \leq \quad |s_t| + \left\|\frac{\partial Q_n}{\partial\theta'}(\theta_0)\right\| \|J_n^{-1}\| \left\|\frac{\partial e_t}{\partial\theta}(\theta^*)\right\|,$$

and thus

$$\|U_{t,n}\| \quad \leq \quad \sum_{j=1}^{m}|s_{t-j}| + \left\|J_n^{-1}\right\| \left\|\frac{\partial Q_n}{\partial\theta'}(\theta_0)\right\| \sum_{j=1}^{m}\left\|\frac{\partial e_{t-j}}{\partial\theta}(\theta^*)\right\| = \mu_t + h_n u_{t,n}$$

with $\mu_t = \sum_{j=1}^m |s_{t-j}|, h_n = \left\|J_n^{-1}\right\| \left\|\frac{\partial Q_n}{\partial\theta}(\theta_0)\right\|$ and $u_{t,n} = \sum_{j=1}^m \left\|\frac{\partial e_{t-j}}{\partial\theta}(\theta^*)\right\|$. It is shown in Francq and Zakoïan (1998) that $\sqrt{n}\frac{\partial Q_n}{\partial\theta}(\theta_0)$ converges in distribution to a non degenerated gaussian distribution and that J_n converges almost surely to a non-singular matrix J. Therefore $h_n = O_P(n^{-1/2})$, and one can take $\tau_1 = 1/2$. Using arguments given in the proof of Lemma 3.1 and the Minkowski inequality we obtain

$$\left\|\frac{\partial e_t}{\partial\theta}(\theta^*)\right\|_\zeta \leq \sum_{i=1}^{\infty}\sum_{j=1}^{p+q}\sup_{\theta\in\Theta}|c_{i,j}(\theta)| \, \|W_{t-i}\|_\zeta \leq C,$$

for all ζ such that $E|W_t|^\zeta < \infty$. We deduce that the moment conditions of **B2** are satisfied if W_t admits moments of order ζ_1, $3\zeta_2$, γ_1 and $2\gamma_2$. Since $\zeta_1 = 3\zeta_2$ with $\zeta_1^{-1} + \zeta_2^{-1} = 1$ if and only if $\zeta_1 = 4$, a moment of order 4, at least, is required for W_t.

Because τ_0 can be chosen arbitrarily large, the conditions 1), 6), 9) and 10) in (2.14) are always satisfied.

With $a_n = [(\log n)^2]$ (the integer part of $(\log n)^2$) we have $n^k\rho^{a_n} \to 0$ as $n \to \infty$ for all k. Thus 2), 3), 5) and 7) in (2.14) are also always satisfied. Because $nb_n^2 \to \infty$ it is easy to see that Condition 4) with $k = 1$ entails Condition 4) with $k = 2$. Condition 4) with $k = 1$ is satisfied when $\gamma_3 < 2d/(2+d)$. Since γ_3 must also be strictly greater than 1, this is only possible when $d > 2$. Taking $\gamma_1 = 2\gamma_2$ the required moment condition is $E|W_t|^{3\gamma_3/(\gamma_3-1)} < \infty$. Note that the minimum of the function $\gamma_3 \mapsto 3\gamma_3/(\gamma_3 - 1)$ is $6d/(d-2)$ on $(1, 2d/(2+d))$. It is thus possible to find a suitable γ_3 when W_t admits moments of order greater than $6d/(d-2)$. Condition 8) is satisfied when $n^2b_n^{6+d} \to \infty$, which completes the proof. □

3.3 Linear prediction plus nonlinear prediction of ARMA residuals

An alternative estimator can be constructed as follows. Note that under **C2**, (ϵ_t) is the linear innovation process of (W_t), that is

$$W_t = EL(W_t \mid \{W_u, u < t\}) + \epsilon_t$$

and that the σ-fields generated by $\{W_u, u < t\}$ and $\{\epsilon_u, u < t\}$ coincide. It follows that (1.5) holds true. Let $\hat{W}_t^L$ denote the linear prediction of W_t based on the estimated ARMA model,

$$\hat{W}_t^L = -\sum_{i=1}^{t-1} \hat{\pi}_i W_{t-i}, \quad \text{where} \quad \sum_{i=0}^{\infty} \hat{\pi}_i z^i = \Psi_{\hat{\theta}_n}^{-1}(z)\Phi_{\hat{\theta}_n}(z), \quad |z| \leq 1. \tag{3.5}$$

Write $\tilde{r}(x) = \tilde{r}\{\hat{\epsilon}_t \mid (\hat{\epsilon}_{t-1}, \dots, \hat{\epsilon}_{t-d}) = x\}$ for the kernel estimator of the regression of $\hat{\epsilon}_t$ on $\hat{\epsilon}_{t-1}, \dots, \hat{\epsilon}_{t-d}$ evaluated at $x = (x_1, \dots, x_d)$. The use of

$$\tilde{W}_t = \hat{W}_t^L + \tilde{r}(\hat{\epsilon}_{t-1}, \dots, \hat{\epsilon}_{t-d})$$

as an estimator of $E(W_t \mid \{W_u, u < t\})$ is legitimated by the following result.

Theorem 3.2 *Let* **A1**, **C1**–**C3** *and* **C5** *hold true. For all integer d_1, the vector $(\epsilon_t, \dots, \epsilon_{t-d_1})$ is assumed to have a strictly positive density f_ϵ which is uniformly in the order of each of its marginal densities, and the functions $f_\epsilon(x_1, \dots, x_{d_1+1})$ and $(x_2, \dots, x_{d_1+1}) \mapsto \int x_1 f_\epsilon(x_1, \dots, x_{d_1+1}) dx_1$ are supposed to be twice derivable, with continuous and bounded second order derivatives. Then*

(i) If $b_n \to 0$ and $nb_n^{2(1+d)} \to \infty$,

$$\tilde{r}\left\{\hat{\epsilon}_t \mid \hat{\boldsymbol{\epsilon}}_{t-1}^{t-d} = x\right\} \to E\left\{\epsilon_t \mid \boldsymbol{\epsilon}_{t-1}^{t-d} = x\right\}, \quad \textit{in probability, as } n \to \infty.$$

(ii) If, in addition, **A1'**, **A2'** *and* **A4'** *hold with $Z = \epsilon$, if $d > 2$ and $E|W_t|^s < \infty$ with $s \geq 4 + \nu$ and $s > 6d/(d-2)$, if $n^2 b_n^{6+d} \to \infty$ hold,*

$$\sqrt{nb_n^d}\left[\tilde{r}\left\{\hat{\epsilon}_t \mid \hat{\boldsymbol{\epsilon}}_{t-1}^{t-d} = x\right\} - E\left\{\epsilon_t \mid \boldsymbol{\epsilon}_{t-1}^{t-d} = x\right\}\right] \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\operatorname{Var}\left\{\epsilon_t \mid \boldsymbol{\epsilon}_{t-1}^{t-d} = x\right\}}{f(x)} \int_{\mathbb{R}^d} K^2(u)du\right)$$

where f is the density of $\boldsymbol{\epsilon}_{t-1}^{t-d}$.

Note that the assumptions are slightly weaker than those of Theorem 3.1, since no assumption is made on the density or the mixing coefficients of (W, ϵ). Note also that when (ϵ_t) is a strong noise, f_ϵ is uniformly in the order of its marginal densities if and only if ϵ_t admits a bounded density.

Proof. We now set $Z_t = \boldsymbol{\epsilon}_t^{t-d}$, $Y_t = \epsilon_t$, $\tilde{Z}_{t,n} = \hat{\boldsymbol{\epsilon}}_t^{t-d}$, $V_{t,n} = \epsilon_t - \hat{\epsilon}_t$, $U_{t,n} = \boldsymbol{\epsilon}_{t-1}^{t-d} - \hat{\boldsymbol{\epsilon}}_{t-1}^{t-d}$. Thus

$$\begin{aligned}
&\sum_{t=1}^{n} \left\{ \|U_{t,n}\| + \|V_{t,n}\| + (\|Y_t\| + \|V_{t,n}\|) \|U_{t,n}\| \right\} \\
\leq\ & C \sum_{t=1}^{n} \rho^t + \left\| \hat{\theta}_n - \theta_0 \right\| \sum_{t=1}^{n} \sum_{i \geq 1} c_i |W_{t-i}| + \sum_{j=1}^{d} \sum_{t=1}^{n} \left\{ C\rho^t + \left\| \hat{\theta}_n - \theta_0 \right\| \sum_{i \geq 1} c_i |W_{t-j-i}| \right\} \\
& \times \left\{ 1 + |\epsilon_t| + C\rho^t + \left\| \hat{\theta}_n - \theta_0 \right\| \sum_{i \geq 1} c_i |W_{t-i}| \right\} \\
=\ & O_P(n^{1/2}),
\end{aligned}$$

by arguments already used to prove (3.4). In particular we used the fact that $E|\epsilon_t||W_{t'}|$ are $E|W_t||W_{t'}|$ finite for any t, t'. We also argue that the LSE is $\sqrt{n}$-consistent under **C5** (see Francq and Zakoïan, 1998). The consistency follows as in the proof of Theorem 3.1.

To show the asymptotic normality, an adaptation of the proof of Theorem 3.1 is needed. One can take $\|V_{t,n}\| \leq \rho_t + k_n v_{t,n}$ with $\rho_t = |s_t|$, $k_n = \left\| J_n^{-1} \right\| \left\| \frac{\partial Q_n}{\partial \theta}(\theta_0) \right\|$ and $v_{t,n} = \left\| \frac{\partial e_t}{\partial \theta}(\theta^*) \right\|$, and we still have $\|U_{t,n}\| \leq \mu_t + h_n u_{t,n}$ with $\mu_t = \sum_{j=1}^{d} |s_{t-j}|$, $h_n = k_n$ and $u_{t,n} = \left\| \sum_{j=1}^{d} \frac{\partial e_{t-j}}{\partial \theta}(\theta^*) \right\|$. The arguments given in the proof of Theorem 3.1 then show that one can take $\tau_0 = \tau_1 = 1/2$ in **B2**, and that the moment conditions of **B2** are satisfied if W_t admits moments of orders ζ_1, $3\zeta_2$, γ_1 and $2\gamma_2$.

Convergence 1) in (2.14) only requires the condition $\zeta_1 > 2$, which can be satisfied when W_t admits a moment of order greater than 4. By already given arguments, conditions 2), 3), 5), 7) and 9) in (2.14) are always satisfied. Moreover, Condition 4) is satisfied when $\gamma_3 < 2d/(2+d)$, which requires $d > 2$ and the moment condition $E|W_t|^s < \infty$ with $s > 6d/(d-2)$. Noting that Condition 6) is entailed by Condition 4, and Condition 10) by Condition 8), we conclude as in the proof of Theorem 3.1. □

3.4 Implementation

For simplicity, assume that $W_{1-d}, \ldots, W_n$ is observed, and consider the one-step ahead prediction of W_{n+1}. Three predictors of W_{n+1} can be investigated:

1) the purely nonparametric estimator

$$\hat{W}_{n+1}^{NP} = \hat{r}\left(W_n, \ldots, W_{n-d+1}\right),$$

where $\hat{r}$ is defined by (2.3), replacing X_t by $(W_{t-1}, \ldots, W_{t-d})$, Y_t by W_t and g by the identity function;

2) the purely parametric estimator $\hat{W}_{n+1}^L$ defined by (3.5);

3) the mixed predictor of Section 3.3

$$\hat{W}_{n+1}^M = \hat{W}_{n+1}^L + \tilde{r}\left(\hat{\epsilon}_n, \ldots, \hat{\epsilon}_{n-d+1}\right),$$

where the $\hat{\epsilon}_t$'s are the ARMA residuals and $\tilde{r}$ is defined by (2.2), replacing $\tilde{X}_{t,n}$ by $(\hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-d})$, $\tilde{Y}_{t,n}$ by $\hat{\epsilon}_t$ and g by the identity function.

The mixed predictor of Section 3.2 could be implemented as well, but for the numerical illustrations we have chosen to concentrate on that of Section 3.3.

3.5 Testing the nullity of the autoregression function of the linear innovation process

Note that if the observed process (W_t) is a strong ARMA model, or more generally if the linear innovation process (ϵ_t) of (W_t) is a sequence of martingale differences, then the following null hypothesis holds :

$$H_0: \quad r\left(\epsilon_{t-1}, \ldots, \epsilon_{t-d}\right) := E\left(\epsilon_t \mid \epsilon_{t-1}, \ldots, \epsilon_{t-d}\right) \equiv 0.$$

It is important to test for H_0 because, if H_0 holds then the mixed predictor defined in 3) of Section 3.4 has no chance to improve the purely linear predictor 2). Conversely, when H_0 is rejected the linear model is not optimal in terms of prediction mean squared error (MSE), and it is worth considering the alternative predictors 1) and 3).

The problem of testing a particular specification of a regression against a nonparametric alternative has been intensively studied in the literature (for recent references see Gao and Tong (2002), Hall and Yatchewa (2005), and the references therein). For iid observations, Härdle and Mammen (1993) proposed a goodness-of-fit test based on a distance between a Nadaraya-Watson estimator and the specification of the regression under the null. Kreiss, Neumann and Yao (2008) (denoted by KNY hereafter) extended the test to a time series context. If the linear innovations were observed, the test statistic of KNY would be in our framework

$$S_n = \int_{\mathbb{R}^d} \left\{ \frac{1}{n b_n^d} \sum_{t=1}^{n} \epsilon_t K\left(\frac{x - (\epsilon_{t-1}, \ldots, \epsilon_{t-d})}{b_n} \right) \right\}^2 \omega(x) dx$$

where $\omega(\cdot)$ is a weight function, which is $\omega(\cdot) \equiv 1$ in the forthcoming applications. Because the ϵ_t are not observed, it is natural to replace S_n by the statistic

$$\tilde{S}_n = \int_{\mathbb{R}^d} \left\{ \frac{1}{nb_n^d} \sum_{t=1}^{n} \hat{\epsilon}_t K \left(\frac{x - (\hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-d})}{b_n} \right) \right\}^2 \omega(x) dx \tag{3.6}$$

where the $\hat{\epsilon}_t$'s are the ARMA residuals obtained in the purely parametric prediction step. It is clear that the kernel and bandwidth involved in (3.6) are not necessarily the same as those involved in the mixed and purely nonparametric predictors (but in our applications we employed the same parameters). Note that when $\omega(\cdot) \equiv 1$[3] and the kernel K is the gaussian density, we simply have

$$\tilde{S}_n = \frac{1}{2^d \pi^{d/2} n^2 b_n^d} \sum_{t=1}^{n} \sum_{s=1}^{n} \hat{\epsilon}_t \hat{\epsilon}_s \prod_{i=1}^{d} \exp \left(-\frac{(\hat{\epsilon}_{t-i} - \hat{\epsilon}_{s-i})^2}{4b_n^2} \right).$$

KNY showed that, under a set of regularity conditions, the asymptotic distribution of $nb_n^{d/2}(S_n - ES_n)$ is gaussian under the null. The approximation of the finite-sample distribution of S_n by a normal distribution is however too crude in practice, and Härdle and Mammen (1993) and KNY implement the wild bootstrap to determine the critical values of their tests. We employed exactly the same resampling scheme as in KNY to obtain the critical value t_α^* of a test of critical region $\{\tilde{S}_n > t_\alpha^*\}$. More precisely, conditionally on $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$, the critical value t_α^* is defined as the $(1-\alpha)$ quantile of the distribution of the bootstrap statistic

$$S_n^* = \int_{\mathbb{R}^d} \left\{ \frac{1}{nb_n} \sum_{t=1}^{n} \xi_t \hat{\epsilon}_t K \left(\frac{x - (\hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-d})}{b_n} \right) \right\}^2 \omega(x) dx,$$

where the ξ_t's are iid $\mathcal{N}(0,1)$, and are independent of the $\hat{\epsilon}_t$'s.

4 Numerical illustrations

We first investigate the performance of the three procedures presented in Section 3.4 on simulated data. Then we present an illustration to the prediction of the volatility of stock market returns.

4.1 Monte Carlo experiments

We propose an illustrative example based on a chaotic process (see May (1976)). Let

$$\epsilon_t = u_t - \frac{1}{2} + \eta_t, \quad u_t = 4u_{t-1}(1 - u_{t-1}), \quad t \geq 1 \tag{4.1}$$

[3] With constant weights, the test enjoys the property of scale invariance.

where u_0 has the arc-sinus density $f(x) = \pi^{-1}\{x(1-x)\}^{-1/2}$ on the interval $[0,1]$, $(\eta_t)_{t\geq 1}$ is an iid sequence, independent of u_0, with mean 0 and finite variance σ_η^2. Since f is the invariant density of (u_t), this process is stationary. We have $E\epsilon_t = 0$ and, since u_t and $1-u_t$ have the same law,

$$\begin{aligned}\mathrm{Cov}(\epsilon_t, \epsilon_{t-1}) &= \mathrm{Cov}(u_t, u_{t-1}) \\ &= \mathrm{Cov}\{4u_{t-1}(1-u_{t-1}), u_{t-1}\} = \mathrm{Cov}(u_t, 1-u_{t-1}) = 0.\end{aligned}$$

The same symmetry argument shows that $\mathrm{Cov}(\epsilon_t, \epsilon_{t-h}) = 0$ for all $h \neq 0$. Therefore (ϵ_t) is a white noise. Consequently, given $\{\epsilon_u,\ u \leq t\}$, the best linear predictor of ϵ_{t+h} is equal to zero, for any horizon h. However, in general, the best (nonlinear) predictor is quite different. For illustrative purpose, Figure 1 displays the scatter plot of the pairs $(\epsilon_{t-1}, \epsilon_t)$, for $t = 1, \dots, 1\,000$, obtained by simulation, and the nonlinear regression obtained by tedious computation, in the case where η_t is uniformly distributed over $[-0.6, -0.4]$. This example illustrates the, possibly dramatic, differences between linear and nonlinear predictions of a given weak ARMA process. One can interpret the ratio

$$\tau := \frac{\mathrm{Var}\, u_t}{\mathrm{Var}\, \epsilon_t} = \frac{\mathrm{Var}\, u_t}{\mathrm{Var}\, u_t + \sigma_\eta^2} = \frac{\frac{1}{8}}{\frac{1}{8} + \sigma_\eta^2}$$

as the proportion of the deterministic part in the noise.

To illustrate our method, we will therefore simulate for different values of θ the MA(1) process $W_t = \epsilon_t - \theta\epsilon_{t-1}$, where the noise (ϵ_t) is given by (4.1). We will compare three predictors: the purely nonparametric predictor defined in 1) of Section 3.4 with $d = 1$, the MA(1) predictor, and the mixed predictor defined in 3) Section 3.4. For the implementation of the nonparametric predictors we used the function `sm.autoregression()` contained in the package **sm** of the statistical software R (see http://cran.r-project.org/). We simulated $N = 50$ independent replications of a simulation of length $n+m$ of the MA(1) process W_t. For each replication, the first $n = 500$ simulated values served to adjust the 3 predictors, and the last $m = 100$ values were used to compare the actual simulated values and their one-step ahead predictions. Figure 2 compares the distributions of the $Nm = 5000$ prediction errors obtained with the 3 predictors. As expected, the parametric estimator is slightly superior to the purely nonparametric one when $|\theta|$ is large and the deterministic proportion τ is not too important (lower panel), whereas the purely nonparametric predictor is often more accurate than the MA(1) predictor when τ is high (upper panels). The distribution of the purely nonparametric predictor presents however more extreme values. In all situations considered in Figures 2, the mixed predictor is always more accurate and seems to cumulate the advantages of the linear and nonparametric predictors. To summarize, the predictor can be ranked, in decreasing order of efficiency, as

$$\hat{W}^M \succ \hat{W}^L \succ \hat{W}^{NP} \qquad \text{when } |\theta| \sim 1 \text{ and } \tau \ll 1$$

and $\hat{W}^M \succ \hat{W}^{NP} \succ \hat{W}^L$ otherwise. The simulations clearly show the superiority of our approach on this model.

4.2 Volatility prediction for stock market returns

Our application concerns daily returns of the following world stock market indices : BEL 20 (Brussels), CAC 40, DAX, FTSE, HSI (Hong Kong), Nikkei, NSE (India), SMI (Swiss) , IGBM (Madrid), SP500, SP TSX (Toronto) and SSE (Shanghai), from January 2, 1991 to July 3, 2009 (except for the indices for which such historical data do not exist). The number of observations varies from $n = 1130$ for the BEL 20 index (which begins on February 11, 2005) to $n = 4591$ for the HSI index. Standard models for such financial series are weak white noises of the form $r_t = \sigma_t \eta_t$ where r_t is the log-return, η_t is an iid noise, with variance equal to 1, and σ_t^2 is the so-called volatility. For the GARCH-type models, σ_t is a measurable function of $\{r_s, s < t\}$.

In the sequel, we compare the prediction of the volatility (that is the prediction of the squared returns) obtained with the parametric, nonparametric, and mixed methods of Section 3.4. [4]

Denote by W_t, $t = 1, \dots, n$ the sequence of the squared returns. In a first set of experiments, for the purely linear predictor, as well as for the parametric component of the mixed predictor, a MA(1) model has been chosen. Results not reported here, based on the test of Section 3.5, show that the assumption that the best predictor is provided by the MA(1) model is clearly rejected on the data.

To compare the effective predictions of the different methods, a part of the observations is used to fit the predictors and a part is reserved for forecasting exercises. Consider the case when n is an even number of the form $n = 2k$ (the case when n is odd is handled

[4] For the purely nonparametric predictor, and also for the nonparametric component of the mixed predictor, we used the default implementation of the R function `sm.autoregression()`.

similarly). For $t = 1, \dots, k$, we used $W_t, \dots, W_{t+k-1}$ to define predictors of W_{t+k} with the methods 1)-3) of Section 3.4. Table 1 indicates that the root mean squared error (RMSE) of prediction is lower with the mixed method for all asset.

In a second set of experiments, we use an ARMA(1,1) as parametric predictor of W_t, both for the purely parametric and mixed methods. Note that an ARMA(1,1) for the squared returns is obtained when the returns follow a GARCH(1,1), that is the most widely-used model for such financial series.

The first four columns of Table 2 give the p-values of the KNY test that the regression of W_t on $W_{t-1}, \dots, W_{t-d}$ is constant. This assumption is clearly rejected. The next four columns give the p-values of the KNY-type test of the null hypothesis that the best predictions of the ARMA(1,1) residuals by d past values are identically equal to zero. One can see that the assumption that the ARMA(1,1) model is optimal is rejected for five series at least for one d. Table 3 gives the prediction RMSE of the different methods. For the series for which the tests of Table 2 do not reject the assumption that the best predictor is ARMA(1,1), the purely linear prediction is indeed the best, in general. The mixed predictor can however improve the purely linear predictor when the tests of Section 3.5 reject the assumption that the ARMA(1,1) predictor is optimal. Finally, note that the purely non parametric method is always far from the optimal method.

5 Conclusion

The basic idea behind the semiparametric method studied in this paper is to improve linear parametric predictions by predicting non parametrically what is not linearly predictable. We considered two approaches using ARMA models to capture the linear part of the process, and nonparametric regressions involving ARMA residuals to capture the nonlinear part. In order to avoid the curse of dimensionality inherent to nonparametric estimation, a small number of regressors seems reasonable, whereas the orders p and q of the parametric model are allowed to be relatively large. Compared to a purely nonparametric regressor of the form $E\left(X_t \mid X_{t-1}, \dots, X_{t-d}\right)$, such a mixed method presents the advantage of being able to take into account mid-term linear dynamics. This mixed method could thus be worth considering for time series whose dynamics can not be well taken into account by a very small number of lagged values of the observed process (as it is the case for MA and mixed ARMA models or for seasonal models) and, at the same time, exhibit nonlinearities. This method could be also of interest to distinguish or compare the purely linear and the purely nonlinear dynamics.

In this paper regularity conditions were given for consistency and asymptotic normality of the residual-based nonparametric regressor. We established intermediate results which are also applicable in a more general context of triangular arrays of noisy observations. We presented simulation experiments, and an illustration to the volatility prediction of 12 stock market indices, in which the mixed method outperforms both the linear predictor and the purely nonparametric predictor.

The R code used for the numerical illustrations is available on the web pages of the authors.

References

Bosq, D. (1996) *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction.* Lecture Notes in Statist 110, Springer Verlag.

Burman, P. and P. Chaudhuri (1994) A hybrid approach to parametric and nonparametric regression *Technical Report No. 243, Division of Statistics, University of California Davis.*

Carrasco, M. and X. Chen (2002) Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* **18**, 17–39.

Carroll, R.J., Linton, O., Mammen, E. and Z. Xiao, (2002) More Efficient Kernel Estimation in Nonparametric Regression with Autocorrelated Errors, *STICERD - Econometrics Paper Series /2002/435, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.*

Carroll, R.J., Maca, J.D., and D. Ruppert (1999) Nonparametric regression in the presence of measurement error. *Biometrika* 86, 541–554.

Drost, F.C., Klaassen, C.A.J., and B.J.M. Werker (1997) Adaptive estimation in time-series models. *Annals of Statistics* **25**, 786–817.

Einsporn, R.L., and J.B. Birch (1993) Model robust regression: using nonparametric regression to improve parametric regression analyses. *Technical Report 93-5, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA.*

Fan, Y. and A. Ullah (1999) Asymptotic Normality of a Combined Regression Estimator. *Journal of Multivariate Analysis* **71**, 191–240

Fan, J. and Q. Yao (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer.

Francq, C., Roy, R. and J-M. Zakoïan (2005) Goodness-of-fit Tests for ARMA Models with Uncorrelated Errors. *Journal of the American Statististical Association* **100**, 532–544.

Francq, C. and J-M. Zakoïan (1998) Estimating Linear Representations of Nonlinear Processes. *Journal of Statistical Planning and Inference* **68**, 145–165.

Francq, C. and J-M. Zakoïan (2000) Covariance matrix estimation for estimators of mixing Wold's ARMA. *Journal of Statistical Planning and Inference* **83**, 369–394.

Glad, I. (1998) Parametrically guided non-parametric regression *Scandinavian Journal of Statistics* **25**, 649–668.

Hall, P. and A. Yatchewa (2005) Unified approach to testing functional hypotheses in semiparametric contexts. *Journal of Econometrics* **127**, 225–252.

Gao, J. and H. Tong (2002) Model Specification Tests in Nonparametric Stochastic Regression Models. *Journal of Multivariate Analysis* **83**, 324–359 (2002)

Härdle, W. (1990) *Applied Nonparametric Regression.* Cambridge University Press: Cambridge.

Härdle, W. and E. Mammen (1993) Comparing nonparametric versus parametric regression fits. *Annals of Statistics* **21**, 1926-1947.

Kreiss, J.P., Neumann, M. H and Q.W. Yao (2008) Bootstrap tests for simple structures in nonparametric time series regression. *Statistics and its interface* **1** 367–380.

Liebscher, E. (2001) Estimation of the density and the regression function under mixing conditions. *Statistics* & *Decisions* **19**, 9–26.

Mack, Y.P. and B.W. Silverman (1982) Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitsth. verw. Geb.* **61**, 405–415.

May, R.M. (1976) Simple mathematical models with very complicated dynamics. *Nature* **261**, 459–467.

Mays, J., Birch, J.B. and R. Einsporn (2000) An overview of model-robust regression. *Journal of Statistical Computation and Simulation* **66**, 79–100.

Pham, D.T. (1986) The mixing property of bilinear and generalised random coefficients autoregressive models. *Stochastic Processes and their Applications* **23**, 291–300.

Phillips, P.C.B. and K.-L. Xu (2005) Inference in Autoregression under Heteroskedasticity. *Journal of Time Series Analysis* **27**, 289–308.

Prakasa Rao, B.L.S. (1983) *Nonparametric Functional Estimation.* Academic Press, New-York.

Robinson, P.M. (1983) Nonparametric estimators for time series. *Journal of Time Series Analysis* **4**, 185–207.

Romano, J.P. and L.A. Thombs (1996) Inference for Autocorrelations under Weak Assumptions. *Journal of the American Statististical Association* **91**, 590–600.

Rosenblatt, M. (1956) Remarks on Some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics* **27** 832–835.

Schennach, S.M. (2004) Nonparametric regression in the presence of measurement error. *Econometric Theory* 20, 1046–1093.

Schick, A. and W. Wefelmeyer (2004) Root n consistent and optimal density estimators for moving average processes. *Scandinavian Journal of Statistics* **31**, 63–78.

Timmermann, A. (2006) Forecast Combination. *Handbook of Economic Forecasting*, Vol.1, ed. Elliott G., Granger C. and Timmermann A., Amsterdam: North-Holland, 135–194.

Xu, K.-L. and P.C.B. Phillips (2008) Adaptive Estimation of Autoregressive Models with Time-Varying Variances. *Journal of Econometrics* **142**, 265–280.

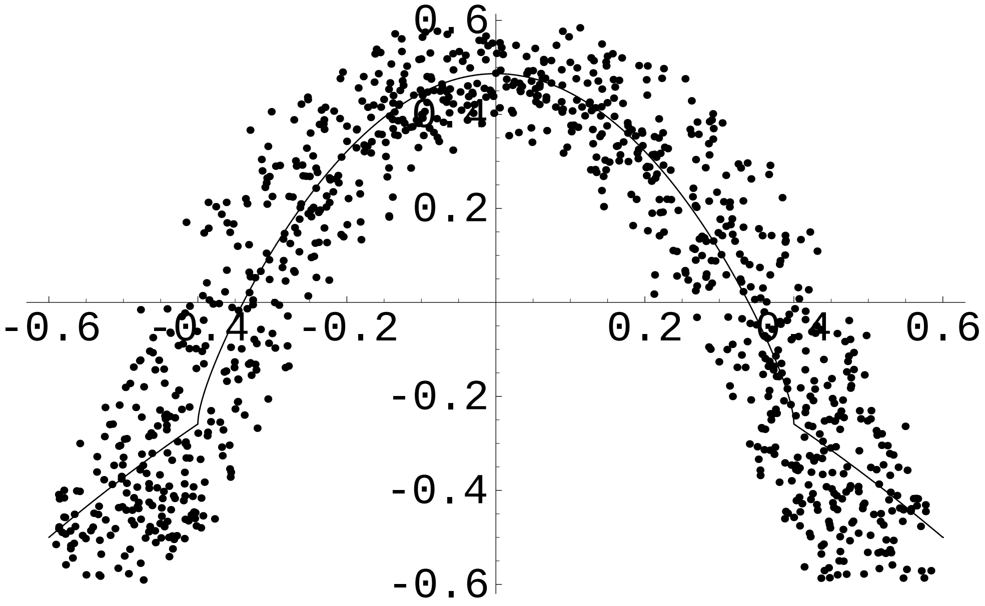

Figure 1: Scatter plot of 1,000 pairs $(\epsilon_t, \epsilon_{t-1})$, simulated from (4.1) with η_t uniformly distributed over $[-0.6, -0.4]$. The full line is the theoretical (nonlinear) regression of ϵ_t on ϵ_{t-1}.

Table 1: RMSE of the MA(1), nonparametric (NP) and mixed (Mixed) predictions. For each series, the smallest RMSE is underlined.

	$MA(1)$	NP				Mixed			
		$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 1$	$d = 2$	$d = 3$	$d = 4$
BEL	8.365	9.165	8.901	8.592	8.856	8.446	8.274	8.387	8.435
CAC	6.739	6.764	6.767	6.746	6.635	6.720	6.631	6.617	6.604
DAX	7.131	7.133	7.077	6.973	6.949	7.044	6.948	6.886	6.871
FTSE	5.229	5.246	5.234	5.230	5.236	5.174	5.074	5.146	5.149
HSI	9.100	9.359	9.244	9.177	9.189	8.879	8.861	8.775	8.786
Nikkei	7.971	7.973	7.975	7.961	7.975	7.904	7.888	7.902	7.910
NSE	13.790	13.693	13.545	13.654	13.540	13.705	13.532	13.587	13.557
SMI	4.845	4.974	5.025	5.017	4.948	4.794	4.776	4.764	4.742
IGBM	7.694	7.777	8.392	8.242	7.792	7.556	7.636	7.631	7.663
SP500	6.232	6.244	6.292	6.349	6.261	6.141	6.142	6.158	6.195
SPTSX	6.634	6.963	7.061	7.214	7.232	6.620	6.666	6.630	6.607
SSE	8.476	8.625	8.628	8.563	8.596	8.565	8.534	8.481	8.435

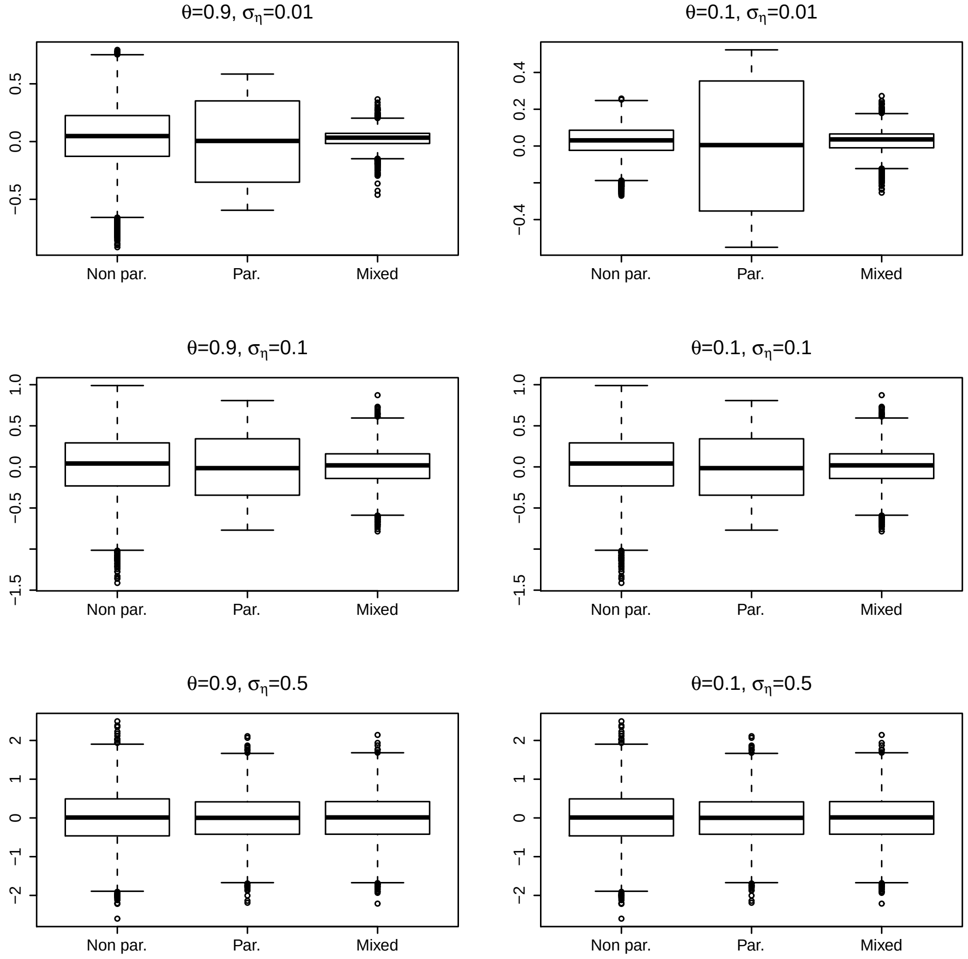

Figure 2: Comparison of the prediction errors of the purely nonparametric, the purely parametric, and the mixed predictors. The simulated process is the MA(1) $W_t = \epsilon_{t-1} + \theta\epsilon_{t-1}$, with $\theta = 0.9$ or $\theta = 0.1$, where (ϵ_t) is the weak white noise (4.1) in which $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$ with $\sigma_\eta = 0.01$ (deterministic proportion $\tau = 99.9\%$), $\sigma_\eta = 0.1$ ($\tau = 92.5\%$) and $\sigma_\eta = 0.5$ ($\tau = 33.3\%$).

Table 2: Test that the optimal predictor is a constant and test that the optimal predictor is ARMA(1,1): p-value of the tests of null hypotheses $H_0^W(d) : E(W_t \mid W_{t-1}, \dots, W_{t-d}) \equiv EW_t$ and $H_0^\epsilon(d) : E(\epsilon_t \mid \epsilon_{t-1} \dots, \epsilon_{t-d}) \equiv 0$, where the ϵ_t's are approximated by ARMA(1,1) residuals. In the last four columns, the p-values are underlined when they are less than 5%.

	$H_0^W(d)$				$H_0^\epsilon(d)$			
	$d=1$	$d=2$	$d=3$	$d=4$	$d=1$	$d=2$	$d=3$	$d=4$
BEL	0	0	0	0	0.814	0.320	0.342	0.304
CAC	0.006	0	0	0	0.818	0.518	0.318	0.428
DAX	0.026	0	0	0	0.128	0.232	0.272	0.238
FTSE	0	0	0	0	$\underline{0.008}$	0.062	0.242	0.274
HSI	0	0	0	0	0.062	$\underline{0.022}$	$\underline{0.004}$	$\underline{0.012}$
Nikkei	0.004	0	0	0	0.160	$\underline{0.030}$	0.090	0.116
NSE	0.016	0.002	0	0	0.814	0.338	0.194	$\underline{0.046}$
SMI	0	0	0	0	0.762	0.292	0.268	0.384
IGBM	0.004	0	0	0	0.642	0.558	0.286	0.188
SP500	0.016	0	0	0	0.380	0.174	0.228	0.246
SPTSX	0	0	0	0	0.410	0.358	0.310	0.286
SSE	0.178	0.006	0	0	0.226	0.164	$\underline{0.026}$	0.052

Table 3: RMSE of the ARMA(1,1), nonparametric (NP) and mixed (Mixed) predictions. For each series, the smallest RMSE is underlined.

	$ARMA(1,1)$	NP				Mixed			
		$d=1$	$d=2$	$d=3$	$d=4$	$d=1$	$d=2$	$d=3$	$d=4$
BEL	7.632	9.165	8.901	8.592	8.856	7.742	8.087	7.694	7.647
CAC	6.177	6.764	6.767	6.746	6.635	6.205	6.275	6.167	6.209
DAX	6.512	7.133	7.077	6.973	6.949	6.516	6.586	6.598	6.540
FTSE	4.753	5.246	5.234	5.230	5.236	4.769	4.804	4.774	4.752
HSI	8.733	9.359	9.244	9.177	9.189	8.702	8.737	8.725	8.687
Nikkei	7.338	7.973	7.975	7.961	7.975	7.311	7.318	7.352	7.347
NSE	13.505	13.693	13.545	13.654	13.540	13.497	13.453	13.434	13.422
SMI	4.480	4.974	5.025	5.017	4.948	4.502	4.533	4.513	4.489
IGBM	7.322	7.777	8.392	8.242	7.792	7.308	7.538	7.432	7.360
SP500	5.613	6.244	6.292	6.349	6.261	5.612	5.692	5.637	5.617
SPTSX	6.243	6.963	7.061	7.214	7.232	6.268	6.302	6.322	6.279
SSE	8.234	8.625	8.628	8.563	8.435	8.303	8.270	8.596	8.293

Combining nonparametric and optimal linear time series predictions: complementary results

A Proof of (2.7) and (2.8).

The dominated convergence theorem entails that, under **A1** and **A3**,

$$\begin{aligned} E\,\hat{f}(x) - f(x) &= \frac{1}{b_n^d}\int_{\mathbb{R}^d} K\left(\frac{x-y}{b_n}\right) f(y)dy - f(x) \\ &= \int_{\mathbb{R}^d} K(t)\left\{f(x-b_nt)-f(x)\right\}dt \to 0 \qquad \text{(A.1)} \end{aligned}$$

when $b_n \to 0$.

By stationarity, we have

$$\operatorname{Var}\hat{f}(x) = \frac{1}{nb_n^{2d}}\sum_{h=-n+1}^{n-1}\left(\frac{n-|h|}{n}\right)\operatorname{Cov}\left\{K\left(\frac{x-X_t}{b_n}\right), K\left(\frac{x-X_{t-|h|}}{b_n}\right)\right\}.$$

Davydov's inequality (1968) entails

$$\begin{aligned} &\left|\operatorname{Cov}\left\{K\left(\frac{x-X_t}{b_n}\right), K\left(\frac{x-X_{t-|h|}}{b_n}\right)\right\}\right| \\ \leq\ & C\left\|K\left(\frac{x-X_t}{b_n}\right)\right\|_{2+\nu}^2 \left\{\alpha_X(|h|)\right\}^{\frac{\nu}{2+\nu}}. \end{aligned}$$

Note that, in view of the Lipschitz condition **A1**, the density K is (uniformly) continuous. Thus it is bounded and satisfies $\int K^{2+\nu}(t)\,dt < \infty$. Thus, using **A1**–**A3** and Lemma 2.1, $\operatorname{Var}\hat{f}(x) \to 0$ as $nb_n^{d\left(1+\frac{\nu}{2+\nu}\right)} \to \infty$. The first part of (2.7) is shown.

Now we will show that $\hat{\varphi}(x) = \varphi(x) + o_P(1)$. The second part of (2.7) will follow, using the Slutsky lemma. With the arguments used to handle $\operatorname{Var}\hat{f}(x)$ we obtain

$$\begin{aligned} \operatorname{Var}\hat{\varphi}(x) &\leq \frac{C}{nb_n^{2d}}\left\|g(Y_t)K\left(\frac{x-X_t}{b_n}\right)\right\|_{2+\nu}^2 \\ &= \frac{C}{nb_n^{2d}}\left\{b_n^d\int |g(y)|^{2+\nu}K^{2+\nu}(t)f_Z(y,x-b_nt)dydt\right\}^{\frac{2}{2+\nu}} \\ &\leq \frac{C}{nb_n^{d(1+\frac{\nu}{2+\nu})}}\left\{\sup_{u\in\mathbb{R}^d}\int_{\mathbb{R}^{d_0}}|g(y)|^{2+\nu}f_Z(y,u)dy\int_{\mathbb{R}^d}K^{2+\nu}(t)\,dt\right\}^{\frac{2}{2+\nu}} = o(1) \end{aligned}$$

under **A1**–**A4**. Now note that

$$\begin{aligned} E\,\hat{\varphi}(x) &= b_n^{-d}\int_{\mathbb{R}^d}\left\{\int_{\mathbb{R}^{d_0}} g(y)\frac{f_{Z_t}(y,u)}{f(u)}dy\right\}K\left(\frac{x-u}{b_n}\right)f(u)du \\ &= b_n^{-d}\int_{\mathbb{R}^d} r(u)f(u)K\left(\frac{x-u}{b_n}\right)du = \int_{\mathbb{R}^d}K(t)\varphi(x-b_nt)dt \\ &= \int_{\mathbb{R}^d}K(t)\left\{\varphi(x) - b_nt'\frac{\partial\varphi}{\partial x}(x) + \frac{b_n^2}{2}t'\frac{\partial^2\varphi}{\partial x\partial x'}(x^*)t\right\}dt, \end{aligned}$$

where x^* is between x and $x - b_n t$, which shows that the bias of $\hat{\varphi}(x)$ tends to zero. The proof of (2.7) is complete.

A Taylor expansion yields

$$f(x - b_n t) = f(x) - b_n t' \frac{\partial f}{\partial x}(x) + \frac{b_n^2}{2} t' \frac{\partial^2 f}{\partial x \partial x'}(x^*) t,$$

where x^* is between x and $x - b_n t$. Using (A.1), **A1** and **A3**, we obtain

$$E\hat{f}(x) - f(x) = O(b_n^2) \quad \text{and} \quad \operatorname{Var} \hat{f}(x) = O\left(n^{-1} b_n^{-d\left(1+\frac{\nu}{2+\nu}\right)}\right). \tag{A.2}$$

Thus

$$E\left\{\hat{f}(x) - f(x)\right\}^2 = O\left(b_n^4 + n^{-1} b_n^{-d\left(1+\frac{\nu}{2+\nu}\right)}\right)$$

is asymptotically minimal for $b_n = O\left(n^{-1/\{4+d+d\nu/(2+\nu)\}}\right)$. We obtain the result for the regression estimator by the same arguments. We deduce that (2.8) holds. □

B Proof of (2.16).

Write

$$\sqrt{nb_n^d}\{\hat{r}(x) - r(x)\} = \frac{\sqrt{nb_n^d}\{\hat{\varphi}(x) - \varphi(x)\}}{\hat{f}(x)} - \varphi(x) \frac{\sqrt{nb_n^d}\{\hat{f}(x) - f(x)\}}{\hat{f}(x) f(x)}, \tag{B.1}$$

and let $H_n(x) = \left(\sqrt{nb_n^d}\{\hat{f}(x) - f(x)\}, \sqrt{nb_n^d}\{\hat{\varphi}(x) - \varphi(x)\}\right)'$. For any $c = (c_1, c_2)' \in \mathbb{R}^2$, letting $g_c(Y_t) = c_1 + c_2 g(Y_t)$, we have

$$\begin{aligned} c' H_n(x) &= \frac{1}{\sqrt{nb_n^d}} \sum_{t=1}^{n} g_c(Y_t) K\left(\frac{x - X_t}{b_n}\right) \\ &\quad - \sqrt{nb_n^d}\{c_1 f(x) + c_2 \varphi(x)\} \\ &= \frac{1}{\sqrt{n}} S_n + R_n, \end{aligned} \tag{B.2}$$

where $S_n = \sum_{t=1}^n x_{n,t}$,

$$x_{n,t} = \frac{1}{b_n^{d/2}} \left\{ g_c(Y_t) K\left(\frac{x - X_t}{b_n}\right) - E\left[g_c(Y_t) K\left(\frac{x - X_t}{b_n}\right)\right]\right\},$$

and

$$R_n = \frac{1}{\sqrt{nb_n^d}} \sum_{t=1}^{n} \left\{ E\left[g_c(Y_t) K\left(\frac{x - X_t}{b_n}\right)\right] - b_n^d \{c_1 f(x) + c_2 \varphi(x)\}\right\}.$$

To establish the asymptotic normality of $c' H_n(x)$, for $c \neq 0$, we will verify the conditions for a Central Limit Theorem for triangular arrays. Notice that the strong mixing coefficients $\alpha_n(h)$ of the process $(x_{n,t})_{t \in \mathbb{Z}}$ are such that $\alpha_n(h) \leq \alpha_Z(h)$. By stationarity, we have

$$\frac{1}{n} \operatorname{Var} S_n = \sum_{h=-n+1}^{n-1} \left(\frac{n - |h|}{n}\right) \operatorname{Cov}(x_{n,t}, x_{n,t-|h|}). \tag{B.3}$$

Davydov's inequality (1968) and a direct extension of Lemma 2.1 entail

$$\begin{aligned} \left|\mathrm{Cov}(x_{n,t}, x_{n,t-|h|})\right| &\leq \frac{C}{b_n^d} \left\| g_c(Y_t) K\left(\frac{x - X_t}{b_n}\right)\right\|_{2+\nu}^2 \{\alpha_Z(|h|)\}^{\frac{\nu}{2+\nu}} \\ &\leq C b_n^{\frac{-d\nu}{2+\nu}} \{\alpha_Z(|h|)\}^{\frac{\nu}{2+\nu}} . \qquad \text{(B.4)} \end{aligned}$$

Now we will show that

$$\left|\mathrm{Cov}(x_{n,t}, x_{n,t-|h|})\right| \leq \begin{cases} C & \text{if } h = 0 \\ C b_n^d & \text{if } |h| > 0. \end{cases} \qquad \text{(B.5)}$$

First considering the case $h = 0$, we have

$$\begin{aligned} \mathrm{Var}(x_{n,t}) &\leq \frac{1}{b_n^d} \int g_c^2(y) K^2\left(\frac{x-u}{b_n}\right) f_Z(y,u) dy du \\ &= \int g_c^2(y) K^2(v) f_Z(y, x - v b_n) dy dv \\ &\leq C \int g_c^2(y) K^2(v) f_Y(y) dy dv \quad < \quad +\infty, \end{aligned}$$

where the second inequality follows from **A3'** and the last one from **A1** and **A3**. This establishes (B.5) in the case $h = 0$. Now we consider the case $h > 0$ which will be sufficient to conclude. Note that

$$\begin{aligned} E g_c(Y_t) K\left(\frac{x - X_t}{b_n}\right) &= c_1 b_n^d \int K(v) f(x - v b_n) dv \\ &\quad + c_2 b_n^d \int g(y) K(v) f_Z(y, x - v b_n) dy dv \\ &= c_1 b_n^d f(x) + c_2 b_n^d \varphi(x) + o(b_n^d). \qquad \text{(B.6)} \end{aligned}$$

It follows that

$$\begin{aligned} &|\mathrm{Cov}(x_{n,t}, x_{n,t-h})| \\ \leq\ & \frac{1}{b_n^d} \int |g_c(y_1)| K\left(\frac{x-u_1}{b_n}\right) |g_c(y_2)| K\left(\frac{x-u_2}{b_n}\right) f_{Z_h,Z_0}(y_1,u_1,y_2,u_2) dy_1 du_1 dy_2 du_2 + C b_n^d \\ \leq\ & C b_n^d \left\{ \int |g_c(y_1)| K(v_1) |g_c(y_2)| K(v_2) f_{Z_h,Z_0}(y_1, x - v_1 b_n, y_2, x - v_2 b_n) dy_1 dv_1 dy_2 dv_2 + 1 \right\} \\ \leq\ & C b_n^d \left\{ \int |g_c(y_1)| K(v_1) |g_c(y_2)| K(v_2) f_{Y_h,Y_0}(y_1, y_2) dy_1 dv_1 dy_2 dv_2 + 1 \right\} \leq C b_n^d, \end{aligned}$$

using **A3'** and the Schwarz inequality. Now we consider a truncation of the right-hand side of (B.3). Let $\varsigma = d/(4+d)$. We have, by (B.5) and **A4'**

$$\sum_{|h|=1}^{[n^\varsigma]} \left(\frac{n-|h|}{n}\right) \mathrm{Cov}(x_{n,t}, x_{n,t-|h|}) \leq C b_n^d n^\varsigma = o(1).$$

Moreover, by (B.4) and **A4'**

$$\begin{aligned} \sum_{|h|=[n^\varsigma]+1}^{n} \left(\frac{n-|h|}{n}\right) \mathrm{Cov}(x_{n,t}, x_{n,t-|h|}) &\leq C b_n^{\frac{-d\nu}{2+\nu}} \sum_{|h|>[n^\varsigma]} \{\alpha_Z(|h|)\}^{\frac{\nu}{2+\nu}} \\ &\leq C b_n^{\frac{-d\nu}{2+\nu}} n^{-\varrho\varsigma} = o(1). \end{aligned}$$

The last inequality follows from $\{\alpha_Z(|h|)\}^{\frac{\nu}{2+\nu}} = O(h^{-(\varrho+1)})$, which is a consequence of **A2'**, and from a standard comparison with an integral. It follows that

$$\lim_{n\to\infty} \frac{1}{n} \operatorname{Var} S_n = \lim_{n\to\infty} \operatorname{Var} x_{n,t} = E(g_c^2(Y_t) \mid X_t = x) f(x) \int_{\mathbb{R}^d} K^2(u) du,$$

where the second equality is a consequence of (B.6). Thus, applying a CLT for triangular sequences of mixing sequences (see e.g. the book by Davidson (1994) and the references therein),

$$n^{-1/2} S_n \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, \{c_1^2 f(x) + 2c_1 c_2 \varphi_1(x) + c_2^2 \varphi_2(x)\} \int_{\mathbb{R}^d} K^2(u) du\right).$$

Now we have, in view of (A.2) and a similar expression for the difference $E\hat{\varphi}(x) - \varphi(x)$, by **A4'**,

$$R_n = c_1 \sqrt{nb_n^d}\{E\hat{f}(x) - f(x)\} + c_2 \sqrt{nb_n^d}\{E\hat{\varphi}(x) - \varphi(x)\} = o(1).$$

Thus, (B.2) entails that $c'H_n(x)$ has the same asymptotic distribution as $n^{-1/2}S_n$. Finally, in view of (B.1) and (2.7) $\sqrt{nb_n^d}\,\{\hat{r}(x) - r(x)\}$ has the same asymptotic distribution as $c'H_n(x)$ with $c = (-r(x)/f(x), 1/f(x))'$, which completes the proof of (2.16). □

C Simulations of the model of Section 4.1.

Figure 3 plots simulations of the noise (ϵ_t) and of the weak MA(1) process (W_t). The empirical autocorrelation functions are in accordance with the theoretical second-order structure of the two simulated processes. In particular, on the basis of the correlogramm, a practitioner would certainly select the MA(1) as a plausible model. One can observe that the distribution of the noise is symmetric, whereas that of the MA(1) is clearly asymmetric. Such an asymmetry is not possible for a MA(1) process with an iid symmetric noise. The simulated trajectories displayed in Figure 3 correspond to a noise with a deterministic proportion $\tau = 92.6\%$. The asymmetry is of course less marked when the deterministic proportion τ is smaller or when the MA(1) parameter θ is close to 0.

D Complements on the stock index data of Section 4.2.

Figure 6 displays the autocorrelation functions of the returns. In this figure, the dotted lines $\pm 1.96/\sqrt{n}$ define the standard significance band in which the autocorrelations of an iid noise should stay with asymptotic probability 95%. These significance bands, obtained from an application of the well-known standard Bartlett's formula, are not valid when the observations are uncorrelated but not independent, as it is the case for GARCH processes, and more generally for weak white noises. Significant bands obtained from the generalized Bartlett's formula recently proposed by Francq and Zakoïan (2009), are given in full lines. Given that most of the autocorrelations fall into the generalized Bartlett's bands, it is reasonable to consider the returns as weak white noises. This is in agreement with the standard economic theory which asserts that such stock returns should be martingale differences. In view of Figure 7, displaying the autocorrelation functions of the squares of the returns, it is however clear that the squares of the returns are correlated.

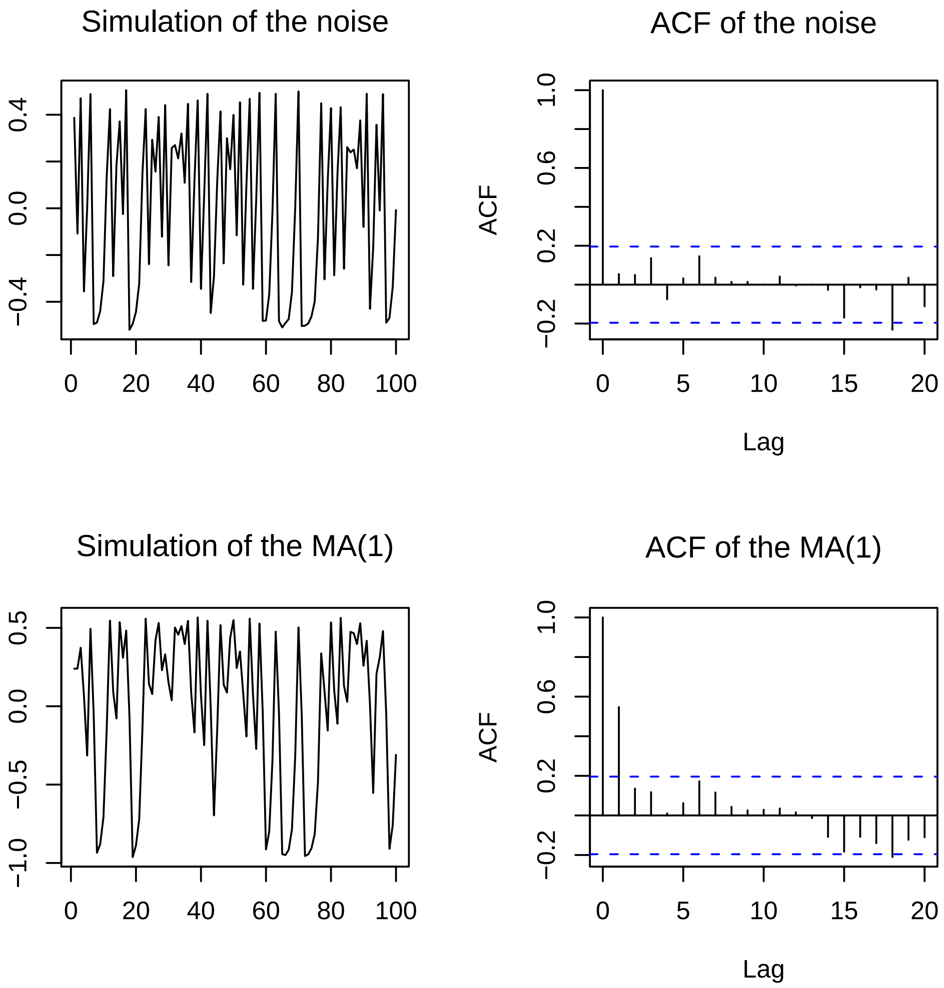

Figure 3: A simulation of the weak white noise ϵ_t defined by (4.1) with $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$ and $\sigma_\eta = 0.1$, and a simulation of the MA(1) process $W_t = \epsilon_{t-1} + 0.9\epsilon_{t-1}$. The right panels display the autocorrelation functions of the two simulations.

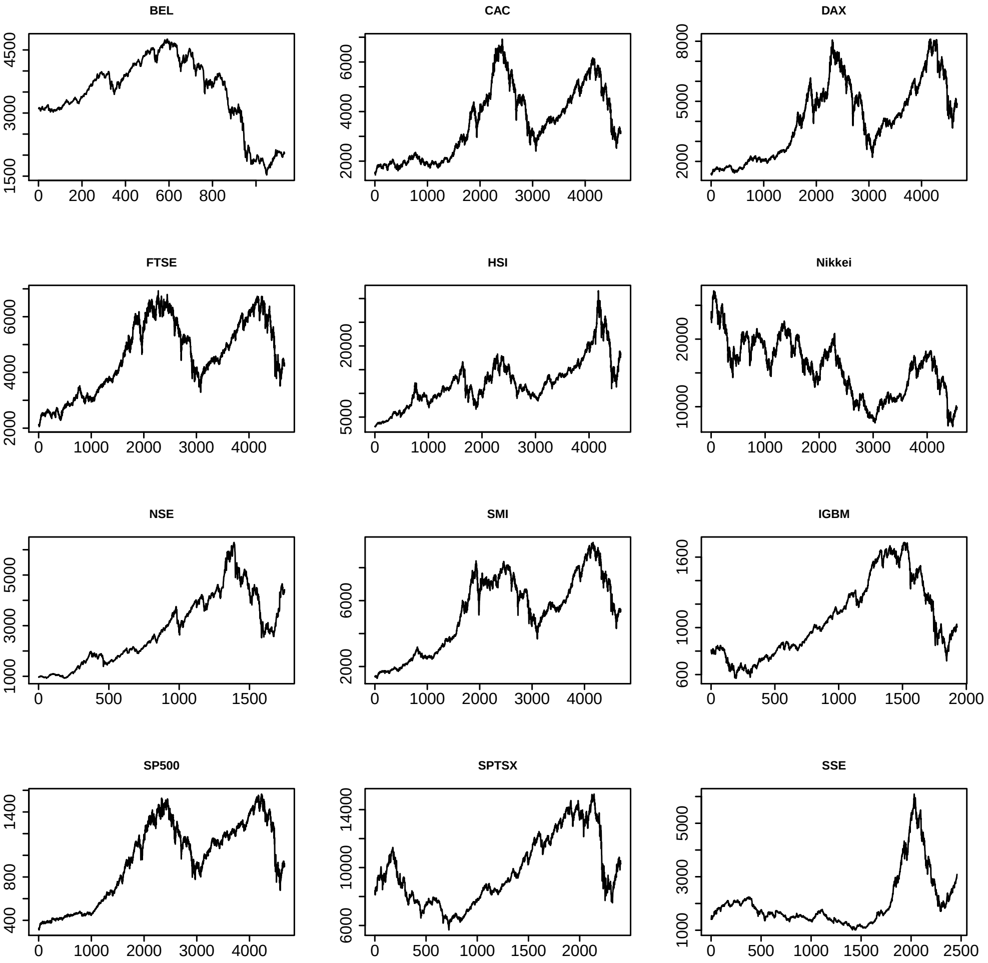

Figure 4: World financial indices between January 2, 1991 and July 3, 2009.

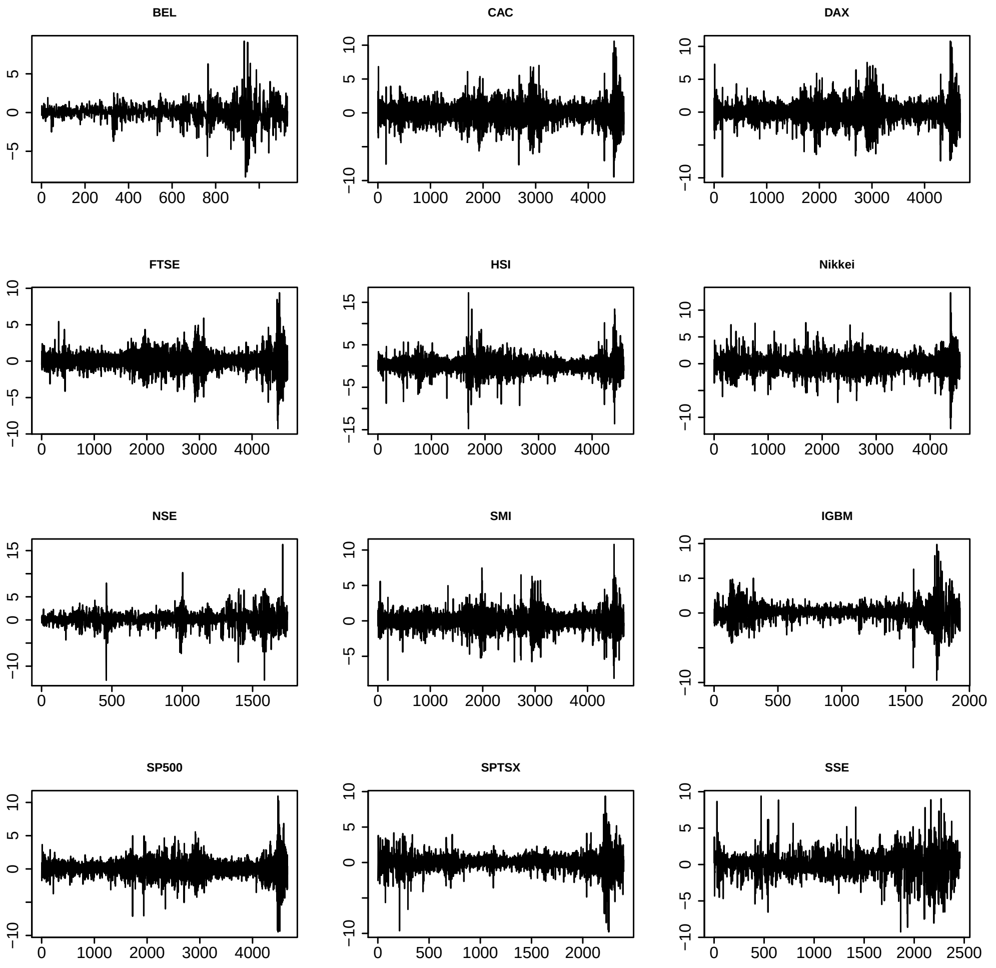

Figure 5: Returns of the indices of Figure 4.

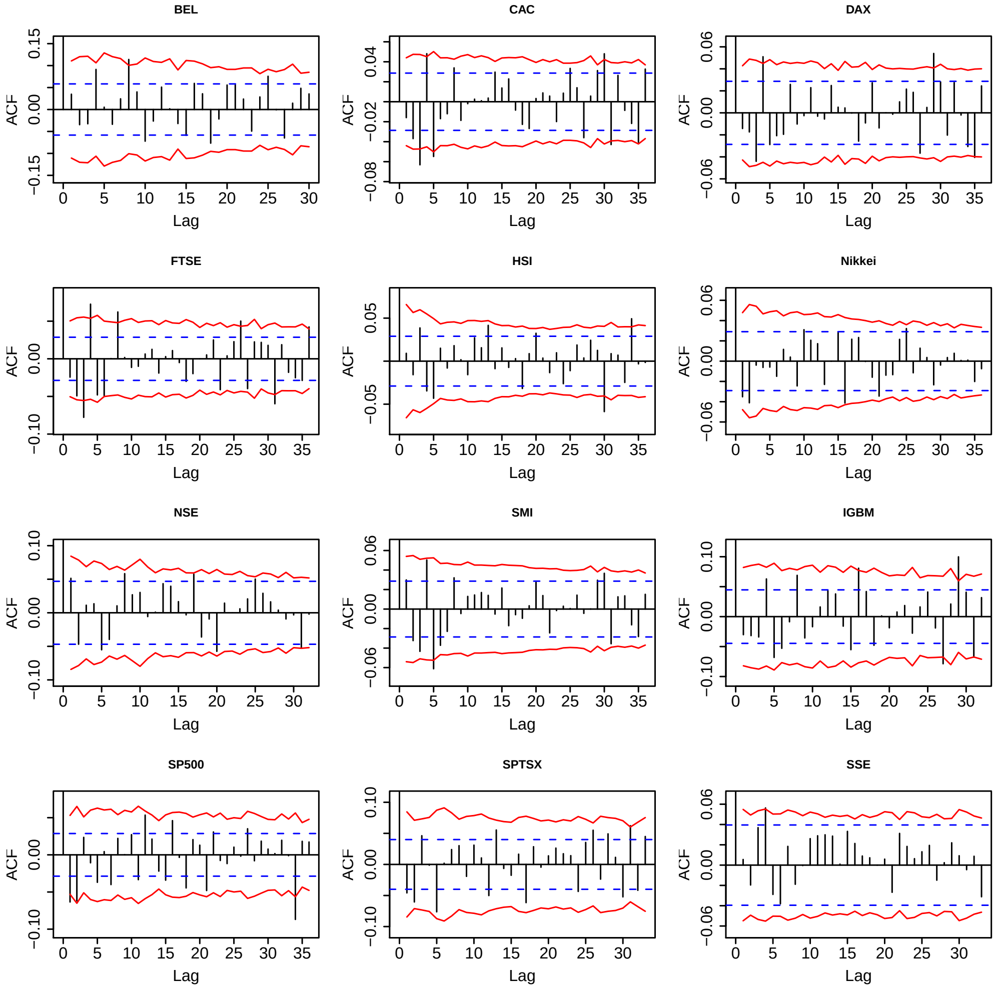

Figure 6: Autocorrelation functions of the series of the returns.

Figures 8-10 display, for each index, the Nadaraya-Watson estimator $\hat{r}(x)$ of the regression $r(x) = E(W_t \mid W_{t-k} = x)$ for the lag $k = 1$ and $k = 7$. The dotted lines are 95% confidence bands for the regression, deduced from (2.16). It is interesting to note that the general allure of all these regressions is that of increasing functions. This is in accordance with the usual finding of strong positive correlations for the squared returns.

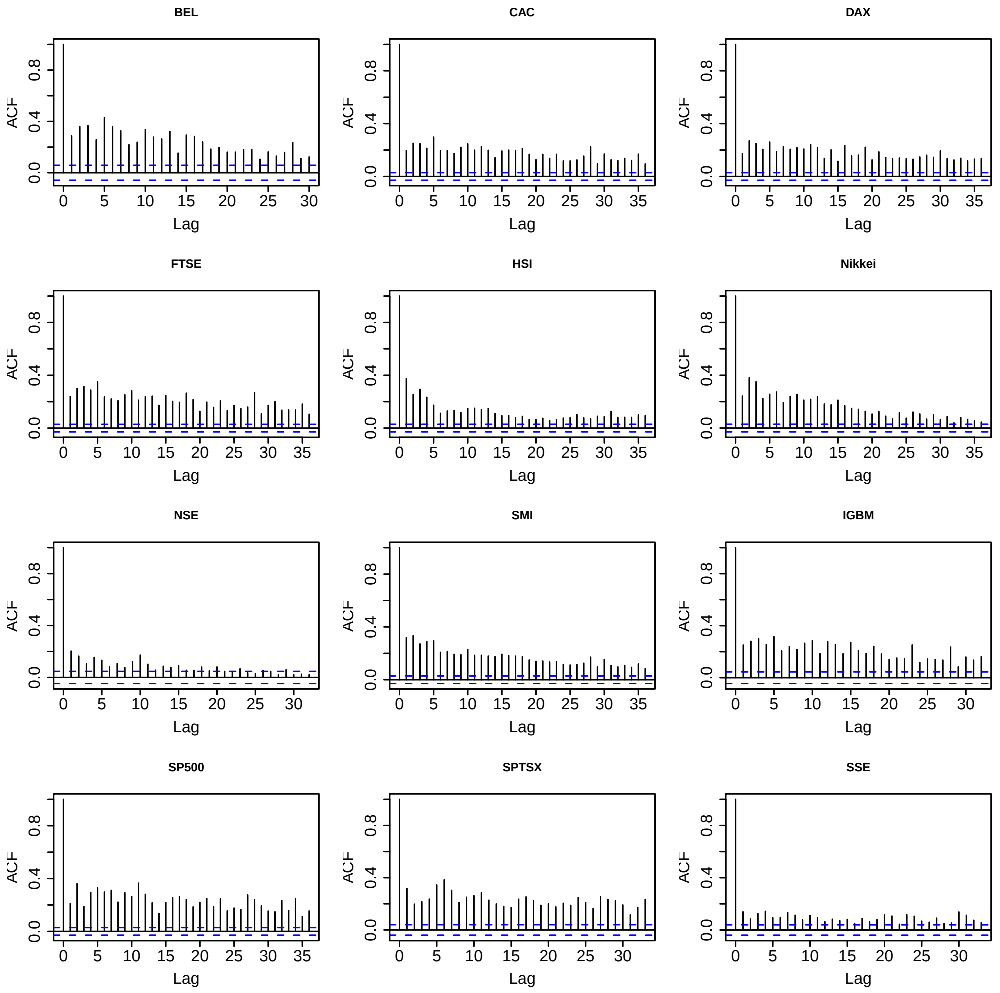

Figure 7: Autocorrelation functions of the squares of the returns.

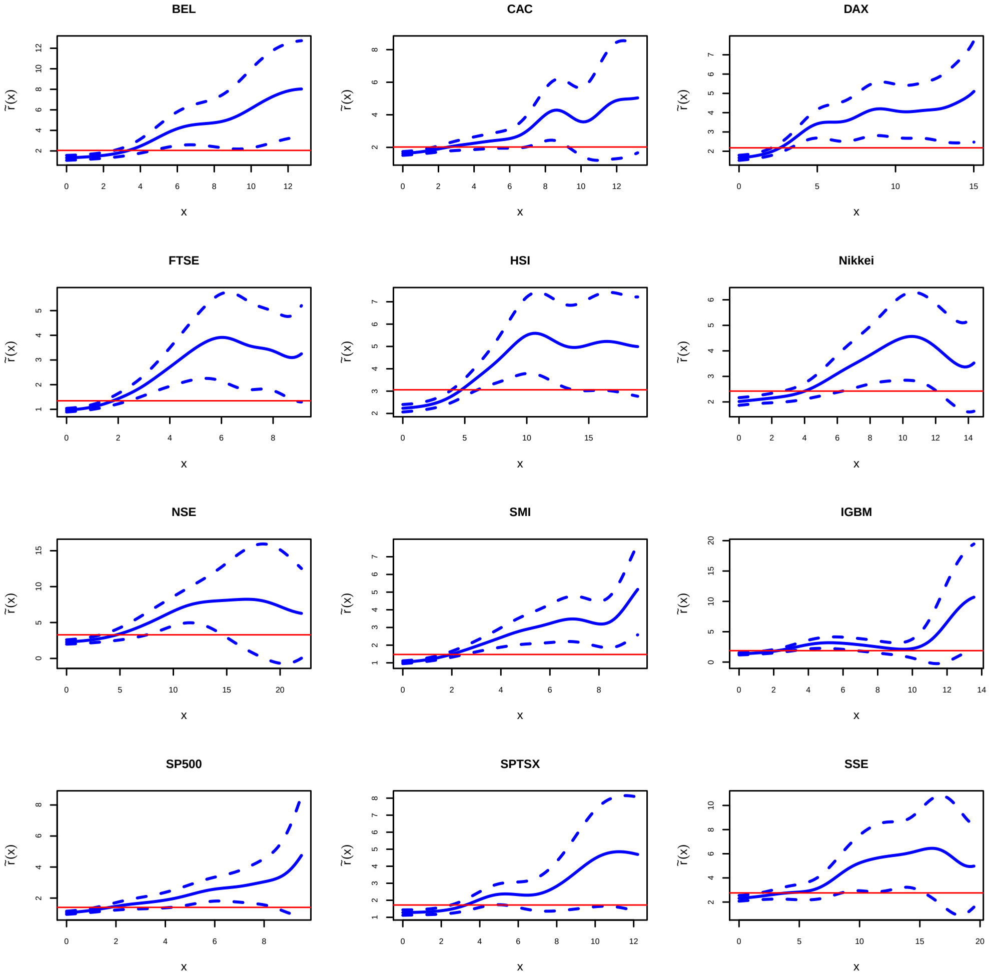

Figure 8: Nadaraya-Watson estimator $\hat{r}(x)$ of the regression $r(x) = E(W_t \mid W_{t-1} = x)$, where W_t denotes the squares of the returns.

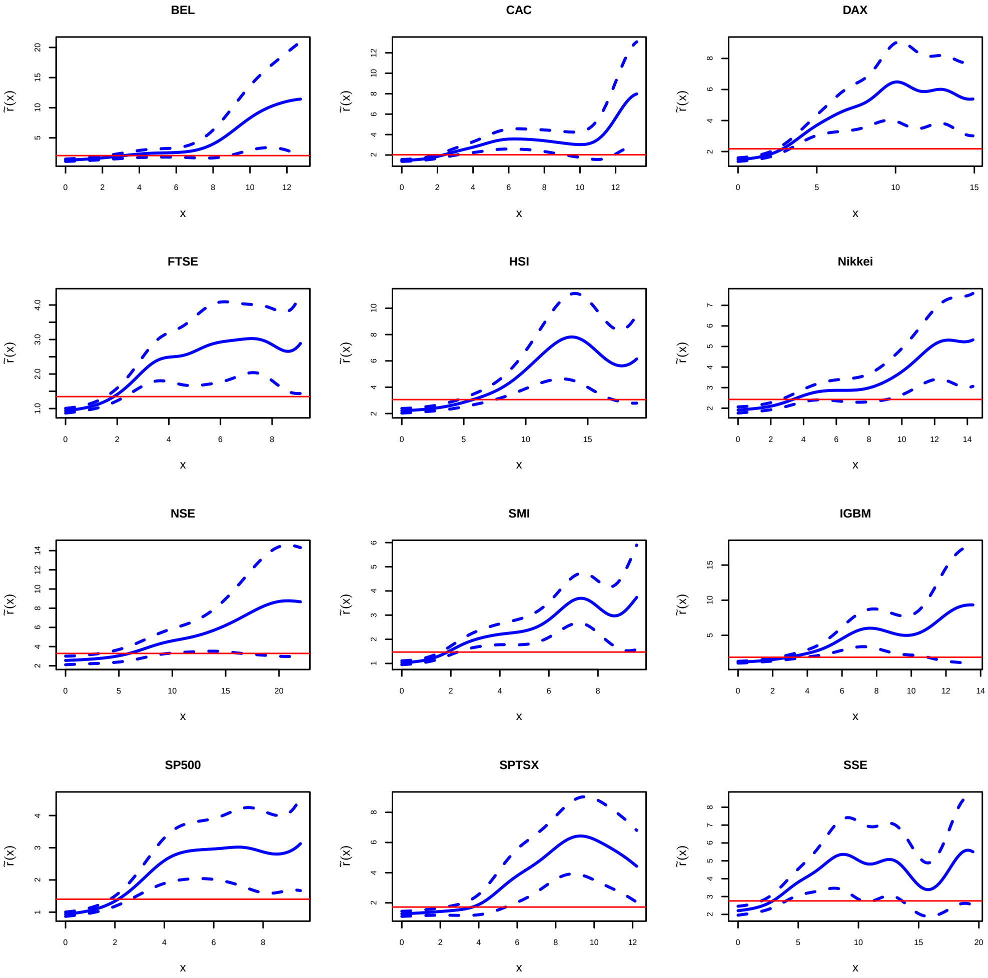

Figure 9: As Figure 8, but for the lag 2 regression $r(x) = E(W_t \mid W_{t-2} = x)$.

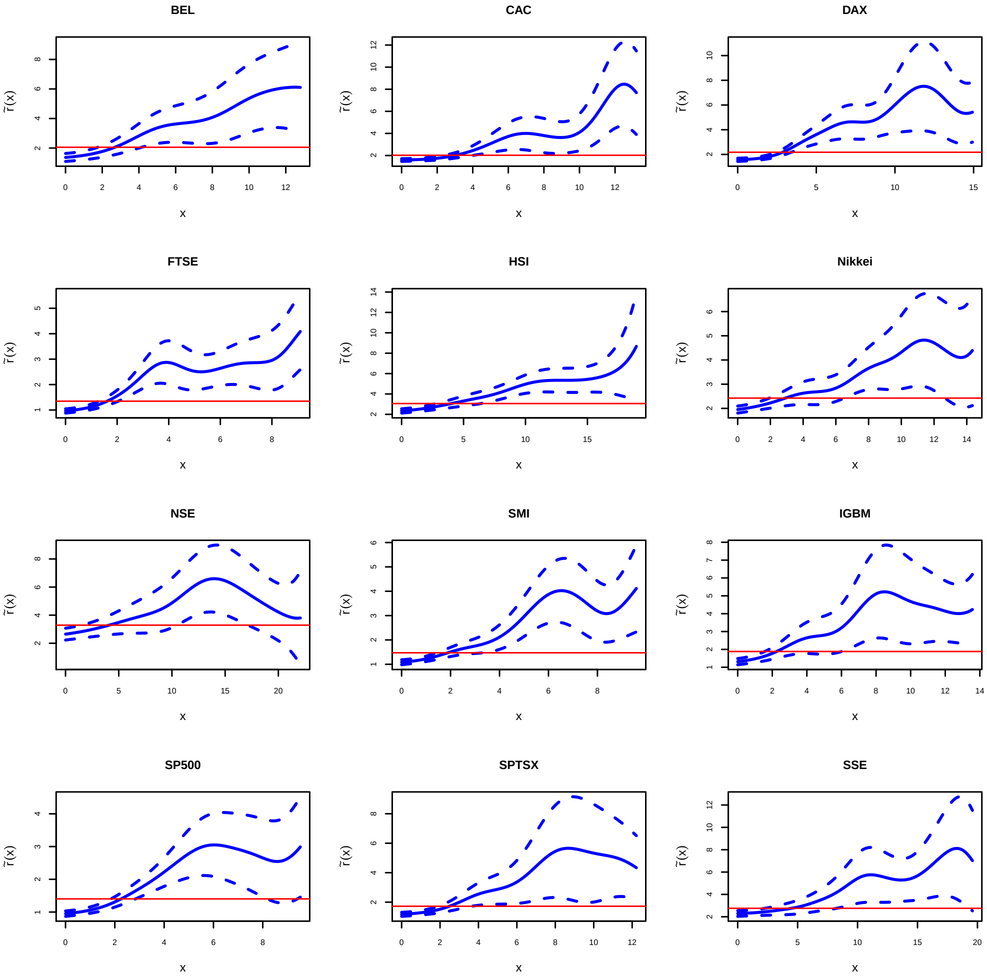

Figure 10: As Figure 8, but for the higher lag regression $r(x) = E(W_t \mid W_{t-7} = x)$.

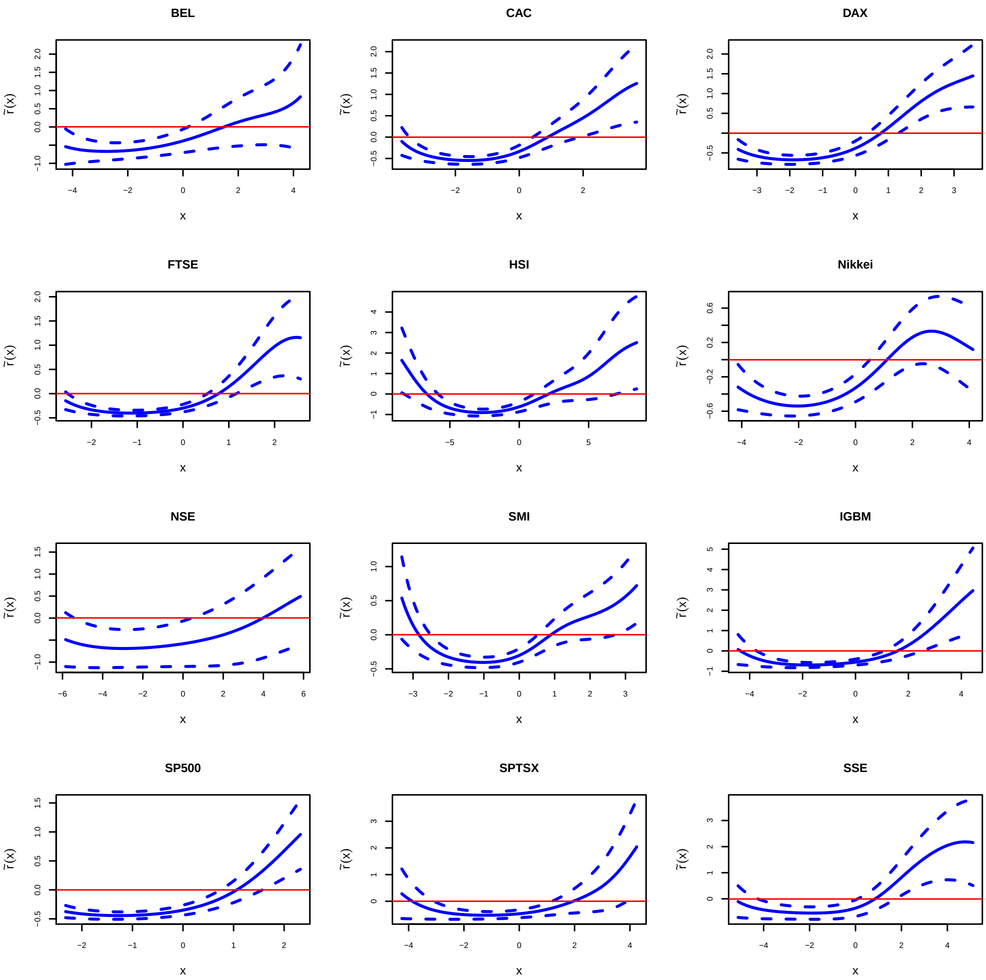

Figure 11: Non parametric estimator of $r(x) = E(\epsilon_t|\epsilon_{t-1} = x)$ where ϵ_t is the MA(1) error term.

The first columns of Table 4 give the p-values of the KNY test that the regression of W_t on $W_{t-1}, \dots, W_{t-d}$ is constant. This assumption is clearly rejected, confirming the visual aspect of the regressions displayed in Figures 8–10, and also the strong autocorrelations of (W_t) displayed in Figure 7. The last columns of Table 4 concern the test described in Section 3.5 that the best predictor is the MA(1) model. Since the assumption is clearly rejected, the MA(1) predictor is likely to be beaten by a purely non parametric predictor or by a mixed-predictor. Figure 11 confirms the output of the tests, since numerous regressions have a "smile" form which leads to predict a positive ϵ_t (and thus a larger volatility) when the innovation ϵ_{t-1} is far from zero. The asymmetry of the smile indicates that the volatility increases more when W_{t-1} is higher than when it is lower than expected.

Table 4: Test that the optimal predictor is a constant and test that the optimal predictor is MA(1): p-value of the tests of null hypotheses $H_0^W(d) : E(W_t \mid W_{t-1},\ldots,W_{t-d}) \equiv EW_t$ and $H_0^\epsilon(d) : E(\epsilon_t \mid \epsilon_{t-1}\ldots,\epsilon_{t-d}) \equiv 0$, where the ϵ_t's are approximated by MA(1) residuals.

	$H_0^W(d)$				$H_0^\epsilon(d)$			
	$d=1$	$d=2$	$d=3$	$d=4$	$d=1$	$d=2$	$d=3$	$d=4$
BEL	0	0	0	0	0.010	0	0	0
CAC	0.006	0	0	0	0.084	0	0	0
DAX	0.026	0	0	0	0.370	0	0	0
FTSE	0	0	0	0	0.070	0	0	0
HSI	0	0	0	0	0	0	0	0
Nikkei	0.004	0	0	0	0.002	0	0	0
NSE	0.016	0.002	0	0	0.118	0.014	0.002	0
SMI	0	0	0	0	0.044	0	0	0
IGBM	0.004	0	0	0	0.088	0	0	0
SP500	0.016	0	0	0	0.008	0	0	0
SPTSX	0	0	0	0	0.006	0	0	0
SSE	0.178	0.006	0	0	0.998	0.128	0	0

E R code

This section contains programs written in the R language (see http://cran.r-project.org/). We begin with three auxiliary routines: the function `K()` for the multivariate gaussian kernel, the function `rn()` for the Nadaraya-Watson regressor of `Y` on the columns of `X` at the point `x` with the bandwidth `h`, and the function `bandwidth.x()` for an approximation of the local optimal bandwidth at the point `x` (as described in Bosq, 1996, Chapter 2).

```
# Gaussian Kernel
K<- function (x) prod(dnorm(x))
# Nadaraya-Watson estimator
rn<- function (x,h,X,Y) {
n<-length(Y)
dum<-sapply(1:n,function (i) K((x-X[i,])/h))
sum(Y*dum)/sum(dum) }
# approximation of the local optimal bandwidth at the point x
bandwidth.x<- function (x,X) {
n<-length(X[,1]); d<-length(X[1,]); sdv<-sd(X[,1])
f<-prod(dnorm(x,sd=sdv))
dum<-(sum(x^2)/sdv^4 -d/sdv^2)^2/d
c0<-(dum*f*(2*sqrt(pi))^2)^(-1/(d+4))
c0*n^(-1/(d+4)) }
```

Given a time series `W[1:n]`, the function `prevNP()` uses the function `rn()` to compute the non parametric Nadaraya-Watson predictor of `W[n+1]` as function of d past values.

```
# nonparametric prediction of W[n+1] as a function of x=(W[n], ..., W[n-d+1])
prevNP<- function (W,d) {
n<-length(W); x<-W[n:(n-d+1)]
X<-matrix(nrow=(n-d),ncol=d)
for (j in(1:d))X[,j]<-W[(d+1-j):(n-j-delay)]
h<- bandwidth.x(x,X)
prev<-rn(x,h,X,W[(d+1+delay):n]) }
```

The function `prevPara.arma11()` uses the function `arma()` of the library `tseries` to compute the ARMA(1,1) prediction of `W[n+1]`.

```
## ARMA(1,1) prediction
library(tseries)
prevPara.arma11<- function (W){
n<-length(W) # (y_t-c)-a*(y_{t-1}-c)=e_t+b*e_{t-1}
arma11<-arma(W,coef=c(0.9,-0.85,mean(W)*0.1))
ahat<-as.numeric(arma11$coef[1])
bhat<-as.numeric(arma11$coef[2])
chat<-as.numeric(arma11$coef[3])/(1-ahat)
prev<-chat+bhat*arma11$residuals[n]+ahat*(W[n]-chat) }
```

The function `prevMixte.arma11()` provides a mixed prediction, sum of an ARMA(1,1) prediction of `W[n+1]` and of a non parametric prediction of `res[n+1]`.

```
## mixed ARMA(1,1) + nonparametric prediction
prevMixte.arma11<- function (W,d,fact=2.5)   {
n<-length(W)
arma11<-arma(W,coef=c(0.9,-0.85,mean(W)*0.1))
ahat<-as.numeric(arma11$coef[1])
bhat<-as.numeric(arma11$coef[2])
chat<-as.numeric(arma11$coef[3])/(1-ahat)
prev<-chat+bhat*arma11$residuals[n]+ahat*(W[n]-chat)

res<-arma11$residuals[2:n]
res.tronc<-pmax(pmin(res,fact*sd(res)),-fact*sd(res))# to deal with outliers
res<-res.tronc

n<-length(res); x<-res[n:(n-d+1)]
X<-matrix(nrow=(n-d),ncol=d)
for (j in(1:d))X[,j]<-res[(d+1-j):(n-j)]
h<- bandwidth.x(x,X)
prevres<-rn(x,h,X,res[(d+1):n])

prevmixtes<-max(prevres+prev,0); prevmixtes }
```

References for the supplementary part

Davidson, J. (1994) *Stochastic Limit Theory.* Oxford University Press, New-York.

Davydov, Y.A. (1968) Convergence of Distributions Generated by Stationary Stochastic Processes. *Theory of Probability and its Applications* **13**, 691–696.

Francq, C. and J-M. Zakoïan (2009) Bartlett's formula for a general class of non linear processes. *Journal of Time Series Analysis*, 30, 449–465.