*Bernoulli* **12**(2), 2006, 333–350

# An exponential inequality under weak dependence

RAOUL S. KALLABIS and MICHAEL H. NEUMANN

*Technical University of Braunschweig, Institute of Mathematical Stochastics, Pockelsstrasse 14, D-38106 Braunschweig, Germany.*

Doukhan and Louhichi introduced a covariance-based concept of weak dependence which is more general than classical mixing concepts. We prove a Bernstein-type inequality under this condition which is similar to the well-known inequality in the independent case. We apply this tool to derive asymptotic properties of penalized least-squares estimators in Barron's classes.

*Keywords:* Barron's classes; Bernstein-type inequality; cumulants; neural networks; nonparametric autoregression; penalized least-squares; weak dependence

## Introduction

Mixing conditions are the classical concept for imposing a restriction on the dependence between time series data. However, many classes of processes which are of interest in statistics do not satisfy any such conditions. Examples include processes driven by discrete innovations as they appear with model-based time series bootstrap methods. Doukhan and Louhichi (1999) proposed a new condition of weak dependence which uses covariances rather than the total variation norm as in the case of mixing. It has been shown that this concept is more general than mixing and includes, under natural conditions on the parameters, essentially all classes of processes of interest in statistics. Ango Nze *et al*. (2002) provide many examples, including Markov processes and so-called Bernoulli shifts. The case of ARCH($p$) processes is discussed in Neumann and Paparoditis (2004).

It became readily apparent that the concept of weak dependence allows in many instances the same tools to be used as in the independent or mixing case. For example, versions of a central limit theorem are derived in Doukhan and Louhichi (1999) for sequences of random variables, in Coulon-Prieur and Doukhan (2000) for situations as they appear with nonparametric curve estimators, and in Neumann and Paparoditis (2004) for triangular schemes. A first exponential inequality was obtained in Doukhan and Louhichi (1999) and a Bennett inequality was proved in Dedecker and Prieur (2004).

The major contribution of this paper is a new Bernstein-type inequality for weakly dependent random variables. Our conditions are slightly stronger than those in Dedecker and Prieur (2004); in particular, we assume an exponential decay of the coefficients of weak dependence. For typical time series models, one often has such an exponential decay under conditions on the parameters which are almost necessary for ergodicity. On the other hand, in contrast to a Bernstein-type inequality which follows from the Bennett inequality of

1350–7265 

Dedecker and Prieur (2004), the variance of the sum appears at the same place in the exponent as in the classical Bernstein inequality for independent random variables. The second (asymptotically often negligible) term differs from that in the independent case, which is an effect of the dependence. It is remarkable that the proofs of the exponential inequalities are completely different. Dedecker and Prieur (2004) use the recently developed tool of replacing weakly dependent blocks of random variables by independent ones and then refer to an exponential inequality for independent random variables. Here we use cumulant techniques mainly developed by the Lithuanian school; see, for example, Saulis and Statulevicius (1991) for an overview.

We present the Bernstein-type inequality in the next section and its proof in Section 3. As a statistical application, we derive in Section 4 rates of convergence for penalized least-squares estimators based on a neural network with one hidden layer and sinusoids as activation functions. We suppose that the function to be estimated is a member of Barron's class. It is known from related settings that this particular restriction of the complexity allows a function to be estimated with a rate which does not depend on the nominal dimension of the space. This is in sharp contrast to more traditional smoothness classes such as Hölder, Sobolev or Besov for which minimax rates of convergence deteriorate rather fast as the dimension increases. Using the Bernstein-type inequality, we derive a penalization scheme which produces an estimator whose rate of convergence does not depend on the nominal dimension but on an intrinsic measure of complexity. The proofs of some approximation-theoretic and of our statistical results are contained in Section 5.

## 2. A Bernstein-type inequality

In this section we state the main result, a Bernstein-type inequality for weakly dependent random variables.

**Theorem 2.1.** *Suppose that $X_1, \ldots, X_n$ are real-valued random variables with $\mathrm{E}X_i = 0$ and $P(|X_i| \leqslant M) = 1$, for all $i = 1, \ldots, n$ and some $M < \infty$. Let $\sigma_n^2 = \mathrm{var}(X_1 + \ldots + X_n)$. Assume, furthermore, that there exist $K < \infty$ and $\beta > 0$ such that, for all $u$-tuples $(s_1, \ldots, s_u)$ and all $v$-tuples $(t_1, \ldots, t_v)$ with $1 \leqslant s_1 \leqslant \ldots \leqslant s_u \leqslant t_1 \leqslant \ldots \leqslant t_v \leqslant n$, the following inequality is fulfilled:*

$$|\mathrm{cov}(X_{s_1} \ldots X_{s_u}, X_{t_1} \ldots X_{t_v})| \leqslant K^2 \, M^{u+v-2} v \mathrm{e}^{-\beta(t_1 - s_u)}. \tag{2.1}$$

*Then*

$$P\left(\sum_{i=1}^{n} X_i \geqslant t\right) \leqslant \exp\left(-\frac{t^2/2}{A_n + B_n^{1/3} t^{5/3}}\right), \tag{2.2}$$

*where $A_n$ can be chosen as any number greater than or equal to $\sigma_n^2$ and*

$$B_n = \left(\frac{16nK^2}{9\,A_n(1 - \mathrm{e}^{-\beta})} \vee 1\right) \frac{2\,(K \vee M)}{1 - \mathrm{e}^{-\beta}}.$$

***Remark 1.*** (i) Inequality (2.2) resembles the classical Bernstein inequality for independent random variables. Asymptotically, $\sigma_n^2$ is usually of order $O(n)$ and $A_n$ can be chosen equal to $\sigma_n^2$ while $B_n$ is usually $O(1)$ and hence negligible. In cases where $\sigma_n^2$ is very small, it might, however, be better to choose $A_n$ considerably larger than $\sigma_n^2$ since it also appears in the denominator of the constant $B_n$. It follows from (2.1) that a rough bound for $\sigma_n^2$ is given by

$$\sigma_n^2 \leqslant \frac{2nK^2}{1-\mathrm{e}^{-\beta}}. \tag{2.3}$$

Hence, taking $A_n = 2nK^2/(1-\mathrm{e}^{-\beta})$, we obtain from (2.2) that

$$P\left(\sum_{i=1}^{n} X_i \geqslant t\right) \leqslant \exp\left(-\frac{t^2}{C_1 n + C_2 t^{5/3}}\right), \tag{2.4}$$

where $C_1 = 4K^2/(1-\mathrm{e}^{-\beta})$ and $C_2 = [2(K \vee M)/(1-\mathrm{e}^{-\beta})]^{1/3}$. Inequality (2.4) is more of Hoeffding type.

(ii) Dedecker and Prieur (2004) proved a Bennett inequality for weakly dependent random variables. This also implies a Bernstein-type inequality, however, with different constants. In particular, the leading term in the denominator of the exponent differs from $\sigma_n^2$ which is a possible choice of $A_n$ in our case. This is a consequence of their method of proof which consists of replacing weakly dependent blocks of random variables by independent ones. On the other hand, our proof makes essential use of the exponential decay of the right-hand side of (2.1), while Dedecker and Prieur's (2004) result is valid under weaker assumptions regarding the coefficients of weak dependence.

(iii) Condition (2.1) is in fact fulfilled for truncated versions of random variables from many time series models. The constant $K$ in (2.1) is included to possibly take advantage of the sparsity of data as it appears, for example, in nonparametric curve estimation. The constant $\upsilon$ in (2.1) often appears when one derives bounds for covariances for time series models; see Doukhan and Louhichi (1999) and Ango Nze *et al.* (2002) for numerous examples.

(iv) As kindly pointed out by Professor McCullagh, condition (2.1) could also be replaced by (3.10) below, which is a condition on the joint cumulants of the process. Such a condition is sometimes easier to verify in applications, in particular in cases where the index set is not strictly ordered.

## 3. Proof of Theorem 2.1

First, note that it is not possible to adapt the classical method of proof since it makes heavy use of independence; see Bennett (1962). One possible approach to proving an exponential inequality involves replacing blocks of weakly dependent random variables by independent ones and then applying an available inequality from the independent case. This was recently done by Dedecker and Prieur (2004), however, for reasons explained in Remark 1(ii) we do not follow this route.

Our proof of the theorem is based on a result of Bentkus and Rudzkis (1980) which we quote here for reader's convenience.

Let $\xi$ be an arbitrary real-valued random variable with $\mathrm{E}\xi = 0$ and finite moments of all orders. The $k$th cumulant of $\xi$ is defined as

$$\Gamma_k(\xi) = \frac{1}{\mathrm{i}^k} \frac{\mathrm{d}^k}{\mathrm{d}t^k} \ln \mathrm{E}\mathrm{e}^{\mathrm{i}t\xi}\bigg|_{t=0}.$$

If there exist $\gamma \geqslant 0$, $\sigma^2 > 0$ and $B \geqslant 0$ such that

$$|\Gamma_k(\xi)| \leqslant \left(\frac{k!}{2}\right)^{1+\gamma} \sigma^2 B^{k-2}, \qquad \text{for all } k = 2, 3, \ldots,$$

then, for all $t \geqslant 0$,

$$P(\xi \geqslant t) \leqslant \exp\left(-\frac{t^2/2}{\sigma^2 + B^{1/(1+\gamma)} t^{(1+2\gamma)/(1+\gamma)}}\right). \tag{3.1}$$

Note that the quotation of this result in Lemma 2.4 in Saulis and Statulevicius (1991: 19) contains a typo; it is correctly stated and proved in Bentkus and Rudzkis (1980, Lemma 2.1).

Before we proceed with the calculations, we recall some notions needed in the course of the proof. It follows from the definition of the cumulants that

$$\Gamma_k(X_1 + \ldots + X_n) = \sum_{1 \leqslant t_1, \ldots, t_k \leqslant n} \Gamma(X_{t_1}, \ldots, X_{t_k}), \tag{3.2}$$

where

$$\Gamma(X_{t_1}, \ldots, X_{t_k}) = \frac{1}{\mathrm{i}^k} \frac{\partial^k}{\partial u_{t_1} \ldots \partial u_{t_k}} \ln \mathrm{E}\mathrm{e}^{\mathrm{i}(u_1 X_1 + \ldots + u_n X_n)}\bigg|_{u_1 = \ldots = u_n = 0}$$

are mixed cumulants. For any random variable $Y$ with finite expectation, we define $\overline{Y} = Y - \mathrm{E}Y$. For $1 \leqslant t_1 \leqslant \ldots \leqslant t_k \leqslant n$, define so-called centred moments as $\overline{\mathrm{E}}(X_{t_1}, \ldots, X_{t_k}) = \mathrm{E}[X_{t_1} X_{t_2} \ldots \overline{X_{t_{k-1}} X_{t_k}}]$ ($\overline{\mathrm{E}}(X_{t_1}) = \mathrm{E}X_{t_1}$). Statulevicius (1970, Lemma 3) has shown that, for $1 \leqslant t_1 \leqslant \ldots \leqslant t_k \leqslant n$, the mixed cumulants can be expressed in terms of centred moments as

$$\Gamma(X_{t_1}, \ldots, X_{t_k}) = \sum_{\nu=1}^{k} (-1)^{\nu-1} \sum_{\bigcup_{p=1}^{\nu} I_p = I} N_\nu(I_1, \ldots, I_\nu) \prod_{p=1}^{\nu} \overline{\mathrm{E}} X_{I_p}, \tag{3.3}$$

where $\sum_{\cup_{p=1}^{\nu} I_p = I}$ denotes the summation over all unordered partitions in disjoint subsets $I_1, \ldots, I_\nu$ of the set $I = \{1, \ldots, k\}$; see also equation (1.63) in Saulis and Statulevicius (1991), as a more easily available reference. Given such a partition, $\overline{\mathrm{E}} X_{I_p}$ stands for $\overline{\mathrm{E}}(X_{t_{i_1^{(p)}}}, \ldots, X_{t_{i_{k_p}^{(p)}}})$ if $I_p = \{i_1^{(p)}, \ldots, i_{k_p}^{(p)}\}$ with $i_1^{(p)} < \ldots < i_{k_p}^{(p)}$. We arrange the subsets in the partitions such that $i_1^{(1)} < \ldots < i_1^{(\nu)}$. $N_\nu(I_1, \ldots, I_\nu)$ are certain non-negative integers defined as follows. For $i \in I$, let $n_i(I_1, \ldots, I_\nu) = \#\{p : i_1^{(p)} < i < i_{k_p}^{(p)}\}$. Then

$$N_1(I) = 1$$

and, for $\nu \geqslant 2$,

$$N_\nu(I_1, \ldots, I_\nu) = \prod_{p=2}^{\nu} n_{i_1^{(p)}}(I_1, \ldots, I_\nu);$$

see equations (4.36) and (4.37) in Saulis and Statulevicius (1991: 80). According to this, it follows that $N_\nu(I_1, \ldots, I_\nu) \neq 0$ if and only if $\{I_1, \ldots, I_\nu\}$ is connected, that is, $n_{i_1^{(p)}}(I_1, \ldots, I_\nu) > 0$ for all $p = 2, \ldots, \nu$. Furthermore, we have that

$$\sum_{\nu=1}^{k} \sum_{\bigcup_{p=1}^{\nu} I_p = I} N_\nu(I_1, \ldots, I_\nu) = (k-1)!; \tag{3.4}$$

see Saulis and Statulevicius (1991, equation (4.43)).

As a first step to deriving estimates for the cumulants of $X_1 + \ldots + X_n$, we derive estimates for the centred moments.

**Lemma 3.1.** *Suppose that the assumptions of Theorem 2.1 are fulfilled. Then, for* $1 \leqslant t_1 \leqslant \ldots \leqslant t_k \leqslant n$, $k \geqslant 2$ *and* $i \in \{1, \ldots, k-1\}$,

$$|\overline{\mathrm{E}}(X_{t_1}, \ldots, X_{t_k})| \leqslant 2^{k-1} K^2 M^{k-2} \, \mathrm{e}^{-\beta(t_{i+1} - t_i)}.$$

***Proof.*** For $t_1 \leqslant \ldots \leqslant t_k$, $k \in N$, we define the shorthand notation $Y_k = X_{t_k}$ and, for $1 \leqslant j < k$, $Y_j = X_{t_j} X_{t_{j+1}} \ldots \overline{X_{t_{k-1}} \overline{X_{t_k}}}$.

Elementary calculations show, for $1 \leqslant j < i < k$, that

$$\begin{aligned} Y_j &= X_{t_j} Y_{j+1} - X_{t_j} \mathrm{E}[Y_{j+1}] \\ &= \ldots = X_{t_j} \ldots X_{t_i} Y_{i+1} - \sum_{l=j}^{i} X_{t_j} \ldots X_{t_l} \mathrm{E}[Y_{l+1}] \\ &= X_{t_j} \ldots X_{t_i} \overline{Y_{i+1}} - \sum_{l=j}^{i-1} X_{t_j} \ldots X_{t_l} \mathrm{E}[Y_{l+1}]. \end{aligned} \tag{3.5}$$

Since $\mathrm{E}X_{t_k} = 0$, in the special case of $i = k-1$ this becomes

$$Y_j = X_{t_j} \ldots X_{t_k} - \sum_{l=j}^{k-1} X_{t_j} \ldots X_{t_l} \mathrm{E}[Y_{l+1}]. \tag{3.6}$$

Without making use of the weak dependence assumption, we obtain, for $3 \leqslant j < k$, that

$$\begin{aligned} \mathrm{E}|Y_j| &= \mathrm{E}|X_{t_j} \overline{Y_{j+1}}| \leqslant M \mathrm{E}|\overline{Y_{j+1}}| \leqslant 2M \mathrm{E}|Y_{j+1}| \\ &\leqslant \ldots \leqslant (2M)^{k-j} \mathrm{E}|X_{t_k}| \leqslant 2^{k-j} M^{k-j+1}. \end{aligned} \tag{3.7}$$

Hence, we obtain in conjunction with (3.6) that

$$
\begin{aligned}
C_{j,i} &:= |\mathrm{cov}\big(X_{t_j} \dots X_{t_i}, Y_{i+1}\big)| \\
&\leqslant |\mathrm{cov}\big(X_{t_j} \dots X_{t_i}, X_{t_{i+1}} \dots X_{t_k}\big)| + \sum_{l=i+1}^{k-1} |\mathrm{cov}\big(X_{t_j} \dots X_{t_i}, X_{t_{i+1}} \dots X_{t_l} \mathrm{E}[Y_{l+1}]\big)| \\
&\leqslant K^2 M^{k-j-1}(k-i)\mathrm{e}^{-\beta(t_{i+1}-t_i)} + \sum_{l=i+1}^{k-1} K^2 M^{l-j-1}(l-i)\mathrm{e}^{-\beta(t_{i+1}-t_i)} 2^{k-l-1} M^{k-l} \\
&= K^2 M^{k-j-1}\left((k-i) + \sum_{l=i+1}^{k-1} (l-i)2^{k-l-1}\right) \mathrm{e}^{-\beta(t_{i+1}-t_i)} \\
&\leqslant K^2 M^{k-j-1} 2^{k-i} \mathrm{e}^{-\beta(t_{i+1}-t_i)}. \qquad (3.8)
\end{aligned}
$$

(The last inequality follows from $(k-i) + \sum_{l=i+1}^{k-1} (l-i)2^{k-l-1} = \sum_{j=1}^{k-i-1} \times j2^{k-i-1-j} + (k-i)\sum_{j=1}^{\infty} 2^{-j} \leqslant 2^{k-i-2}\left(\sum_{j=1}^{\infty} j2^{1-j}\right) = 2^{k-i}$.)

On the other hand, we obtain from (3.5) that

$$
|\mathrm{E}[Y_j]| \leqslant C_{j,i} + \sum_{l=j}^{i-1} |\mathrm{E}[X_{t_j} \dots X_{t_l}]| \cdot |\mathrm{E}[Y_{l+1}]|.
$$

Therefore, we obtain recursively that

$$
\begin{aligned}
|\overline{\mathrm{E}}(X_{t_1}, \dots, X_{t_k})| &= |\mathrm{E}[Y_1]| \\
&\leqslant C_{1,i} + \sum_{l=1}^{i-1} M^l \, |\mathrm{E}[Y_{l+1}]| \\
&\leqslant \dots \leqslant C_{1,i} + \sum_{1 \leqslant l_1 \leqslant i-1} M^{l_1} C_{l_1+1,i} + \sum_{1 \leqslant l_1 < l_2 \leqslant i-1} M^{l_2} C_{l_2+1,i} \\
&\quad + \dots + \sum_{1 \leqslant l_1 < \dots < l_{i-1} \leqslant i-1} M^{i-1} C_{i,i}.
\end{aligned}
$$

From (3.8) we now obtain that

$$
\begin{aligned}
|\overline{\mathrm{E}}(X_{t_1}, \dots, X_{t_k})| &\leqslant K^2 M^{k-2} 2^{k-i} \mathrm{e}^{-\beta(t_{i+1}-t_i)} \sum_{l=0}^{i-1} \binom{i-1}{l} \\
&= K^2 M^{k-2} 2^{k-1} \mathrm{e}^{-\beta(t_{i+1}-t_i)}.
\end{aligned}
$$

□

Equations (3.2), (3.3) and the result of Lemma 3.1 can now be used to derive estimates for the cumulants of $X_1 + \dots + X_n$.

**Lemma 3.2.** *Suppose that the assertions of Theorem 2.1 are fulfilled. Then, for* $k \geqslant 3$,

$$|\Gamma_k(X_1 + \ldots + X_n)| \leqslant nk!((k-1)!)^2 K^2 (K \vee M)^{k-2} \left(\frac{2}{1-\mathrm{e}^{-\beta}}\right)^{k-1}.$$

***Proof.*** Our proof deviates from the proof of similar results in Saulis and Statulevicius (1991) since we were not able to follow all of their arguments. In particular, we could not verify their equation (4.55) on p. 94 which was crucial to their approach.

From (3.2) we obtain that

$$|\Gamma_k(X_1 + \ldots + X_n)|\mathrm{H} \leqslant k! \sum_{1 \leqslant t_1 \leqslant \ldots \leqslant t_k \leqslant n} |\Gamma(X_{t_1}, \ldots, X_{t_k})|. \tag{3.9}$$

According to (3.3) and Lemma 3.1, we have, for $1 \leqslant t_1 \leqslant \ldots \leqslant t_k \leqslant n$, that

$$|\Gamma(X_{t_1}, \ldots, X_{t_k})| \leqslant \sum_{\nu=1}^{k} \sum_{\bigcup_{p=1}^{\nu} I_p = I} N_\nu(I_1, \ldots, I_\nu) \prod_{p=1}^{\nu} |\overline{\mathrm{E}}(\mathrm{X}_{I_p})|$$

$$\leqslant \sum_{\nu=1}^{k} \sum_{\bigcup_{p=1}^{\nu} I_p = I} N_\nu(I_1, \ldots, I_\nu) \prod_{p=1}^{\nu} 2^{k_p - 1} \mathrm{K}^2 M^{k_p - 2} \min_{1 < j \leqslant k_p} \exp\left(-\beta(t_{i_j^{(p)}} - t_{i_{j-1}^{(p)}})\right).$$

Note that we have, for any connected partition,

$$\max_{1 \leqslant p \leqslant \nu} \max_{1 < j \leqslant k_p} \{t_{i_j^{(p)}} - t_{i_{j-1}^{(p)}}\} \geqslant \max_{1 < i \leqslant k} \{t_i - t_{i-1}\}.$$

Since $N_\nu(I_1, \ldots, I_\nu) = 0$ if $\{I_1, \ldots, I_\nu\}$ is not connected we therefore obtain, in conjunction with (3.4), that

$$|\Gamma(X_{t_1}, \ldots, X_{t_k})| \leqslant \sum_{\nu=1}^{k} \sum_{\bigcup_{p=1}^{\nu} I_p = I} N_\nu(I_1, \ldots, I_\nu)\, 2^{k-1} K^2 (K \vee M)^{k-2} \min_{1 < i \leqslant k} \mathrm{e}^{-\beta(t_i - t_{i-1})}$$

$$\leqslant (k-1)!\, 2^{k-1} K^2 (K \vee M)^{k-2} \min_{1 < i \leqslant k} \mathrm{e}^{-\beta(t_i - t_{i-1})}.$$

This implies that

$$\sum_{1 \leqslant t_1 \leqslant \ldots \leqslant t_k \leqslant n} |\Gamma(X_{t_1}, \ldots, X_{t_k})|$$

$$\leqslant n(k-1)! 2^{k-1} K^2 (K \vee M)^{k-2} \sum_{s_2, \ldots, s_k = 0}^{\infty} \min_{2 \leqslant i \leqslant k} \mathrm{e}^{-\beta s_i}. \tag{3.10}$$

Since $\#\{(s_2, \ldots, s_k) : 0 \leqslant s_i \leqslant s,\ \max\{s_2, \ldots, s_k\} = s\} \leqslant (k-1)(s+1)^{k-2}$ and

$$\sum_{s=0}^{\infty}(s+1)^{k-2}\mathrm{e}^{-\beta s} \leqslant \sum_{s=0}^{\infty}(s+1)\ldots(s+k-2)\mathrm{e}^{-\beta s}$$

$$= \frac{\mathrm{d}^{k-2}}{\mathrm{d}p^{k-2}}\left(\frac{1}{1-p}\right)\Big|_{p=\mathrm{e}^{-\beta}} = (k-2)!\ \frac{1}{(1-\mathrm{e}^{-\beta})^{k-1}},$$

we obtain that

$$\sum_{1\leqslant t_1\leqslant \ldots \leqslant t_k\leqslant n}|\Gamma(X_{t_1},\ldots,X_{t_k})| \leqslant n((k-1)!)^2 2^{k-1}K^2(K\vee M)^{k-2}\frac{1}{(1-\mathrm{e}^{-\beta})^{k-1}},$$

which, in conjunction with (3.9), proves the assertion of the lemma. □

***Proof of Theorem 2.1.*** From Lemma 3.2 we obtain, for $k \geqslant 3$, that

$$|\Gamma_k(X_1+\ \ldots\ +X_n)| \leqslant \left(\frac{k!}{2}\right)^3 \frac{16nK^2}{9(1-\mathrm{e}^{-\beta})}\left(\frac{2(K\vee M)}{1-\mathrm{e}^{-\beta}}\right)^{k-2},$$

which implies that

$$|\Gamma_k(X_1+\ \ldots\ +X_n)| \leqslant \left(\frac{k!}{2}\right)^3 A_n B_n^{k-2}$$

holds for all $k \geqslant 2$. The assertion of the theorem follows now from (3.1). □

# 4. Nonparametric autoregression

We suppose that we observe $X_1, \ldots, X_n$ from a real-valued and (strong) stationary process $(X_t)_{t\in\mathbb{Z}}$. We intend to construct an estimator $\widehat{f^{(d)}}$ of the $m$-step-ahead autoregression function based on $d$ lagged variables, $f^{(d)}(x_1,\ldots,x_d) = \mathrm{E}(X_{n+m}|X_n = x_1, \ldots, X_{n-d+1} = x_d)$. We suppose that the following condition of weak dependence is fulfilled.

(A1) There exist constants $C < \infty$ and $\beta > 0$ such that, for any $u$-tuple $(s_1,\ldots,s_u)$ and any $v$-tuple $(t_1,\ldots,t_v)$ with $s_1 \leqslant \ldots \leqslant s_u \leqslant t_1 \leqslant \ldots \leqslant t_v$ and arbitrary functions $g : \mathbb{R}^u \to \mathbb{R}$, $h : \mathbb{R}^v \to \mathbb{R}$ with $\mathrm{E}[g^2(X_{s_1},\ldots,X_{s_u})] < \infty$, $\mathrm{E}[h^2(X_{t_1},\ldots,X_{t_v})] < \infty$, the following inequalities are fulfilled:

$$|\mathrm{cov}(g(X_{s_1},\ldots,X_{s_u}),\ h(X_{t_1},\ldots,X_{t_v}))|$$

$$\leqslant Cv\sqrt{\mathrm{E}[g^2(X_{s_1},\ldots,X_{s_u})]}\ \mathrm{Lip}(h)\ \mathrm{e}^{-\beta(t_1-s_u)} \tag{4.1}$$

and, for $s < t$,

$$|\mathrm{cov}(g(X_s,\ldots,X_{s-d+1})X_{s+m},\ h(X_t,\ldots,X_{t-d+1})X_{t+m})|$$

$$\leqslant C\sqrt{\mathrm{E}[g^2(X_s,\ldots,X_{s-d+1})]}\ (\mathrm{Lip}(h)+\|h\|_\infty)\mathrm{e}^{-\beta(t-s-m)}$$

where $\mathrm{Lip}(h) = \sup_{x \neq y}\{|h(x) - h(y)|/\|x - y\|_{l_1}\}$ denotes the Lipschitz modulus of continuity of $h$.

***Remark 2.*** (i) Condition (4.1) is similar to a condition used in Doukhan and Louhichi (1999) where only $\|g\|_\infty$ rather than $\sqrt{\mathrm{E}[g^2(X_{s_1}, \ldots, X_{s_u})]}$ appears on the right-hand side. Doukhan and Louhichi (1999) and Ango Nze *et al*. (2002) discuss many examples of processes which satisfy such a condition including Markov processes and so-called Bernoulli shifts. With a slight modification of their proofs, it can be shown that our condition (A1) is also fulfilled by these processes. The case of ARCH($p$) processes is studied in Neumann and Paparoditis (2004).

(ii) Alternative conditions which are more closely related to classical mixing concepts are proposed by Dedecker and Doukhan (2003). A condition similar to theirs,

$$\mathrm{E}[\mathrm{E}(h(X_{t_1}, \ldots, X_{t_v})X_s, X_{s-1}, \ldots) - \mathrm{E}h(X_{t_1}, \ldots, X_{t_v})]^2 \leqslant Cv\,\mathrm{Lip}(h)\mathrm{e}^{-\beta(t_1 - s)},$$

for all $s < t_1 < \ldots < t_k$ and Lipschitz continuous $h$, is indeed equivalent to our condition (4.1).

As a criterion for evaluating the performance of $\widehat{f^{(d)}}$, we use the loss function

$$L(\widehat{f^{(d)}}, f^{(d)}) = \int (\widehat{f^{(d)}}(\underline{x}) - f^{(d)}(\underline{x}))^2\, \mu^{(d)}(d\underline{x}), \tag{4.2}$$

where $\underline{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ and $\mu^{(d)}$ denotes the $d$-dimensional stationary distribution of the process $(X_t)_{t \in \mathbb{Z}}$.

It is well known that achievable rates of convergence in traditional smoothness classes (Hölder, Sobolev, Besov) deteriorate rather fast as the dimension $d$ increases. For example, if some $d$-dimensional curve possesses $s$ bounded derivatives, then the corresponding minimax rate of convergence for the mean squared error is typically $n^{-2s/(2s+d)}$. This phenomenon is connected with the notion of the 'curse of dimensionality'. On the other hand, this viewpoint often seems too pessimistic. The difficulty of estimation in the above smoothness classes comes from sequences of functions that contain an increasing (with $n$) number of 'features' on increasingly fine scales. In contrast, in practical examples, multivariate functions are often of simpler structure. Scott (1992, Chapter 2) claims: 'Multivariate data in $\mathbb{R}^d$ are almost never $d$-dimensional. That is, the *underlying structure* of data in $\mathbb{R}^d$ is almost always of dimension lower than $d$.'

While many popular estimators such as standard kernel estimators suffer from the curse of dimensionality, there is some hope for a reasonably good asymptotic behaviour if the true function does indeed have some 'simple structure' and if the method of estimation is able to exploit this simplicity. To get good rates of convergence even in high dimensions, one sometimes imposes *structural* assumptions such as an additive structure on the function to be estimated. Corresponding estimators attain favourable rates of convergence if the true function obeys the assumed structure; however, they fail completely if these structural assumptions are not fulfilled. In contrast, we choose a truly nonparametric approach which originates from the seminal paper by Barron (1993).

He imposed smoothness constraints of the form

$$\int_{\mathbb{R}^d} \|\omega\| |\widetilde{f}(\omega)| \, \mathrm{d}\omega \leqslant L, \tag{4.3}$$

where $\widetilde{f}$ is the Fourier transform of $f$, and studied approximation-theoretic properties of such function classes in certain overcomplete bases built from sigmoidal functions. It has been shown in several statistical papers – see, for example, Barron (1994) and Juditsky and Nemirovski (2000) for nonparametric regression, Modha and Masry (1996) for nonparametric regression with dependent data, Barron *et al*. (1999) for nonparametric regression and density estimation, Delyon and Juditsky (2000) for nonparametric autoregression – that favourable approximation-theoretic properties of certain bases (sigmoids, sinusoids, etc.) transfer to statistical rates of convergence that do not depend on the nominal dimension $d$. Despite many similarities, details of the corresponding methods are different. While Delyon and Juditsky (2000) and Juditsky and Nemirovski (2000) supposed that the upper bound in an inequality similar to (4.3) is known in advance, Barron (1994), Modha and Masry (1996), and Barron *et al*. (1999) developed penalized minimum distance methods which do not require such prior information.

In the present paper, we devise an estimator of the $m$-step-ahead autoregression function in the context of general stationary processes satisfying a condition of weak dependence. This generalizes previous work of Delyon and Juditsky (2000) who assumed that data are generated by an autoregressive process of finite order, and also of Modha and Masry (1996) who used a stronger restriction on the dependence. Furthermore, in contrast to Delyon and Juditsky (2000) and Juditsky and Nemirovski (2000), we wish to avoid the assumption that an upper bound as in (4.3) is known in advance. To this end, we propose a penalized least-squares method which does not use knowledge of any constant that is usually not available in practice. The proposed estimator can be considered as an artificial neural network estimator with one hidden layer and sinusoids as activation functions.

Similarly to condition (4.3), we assume that

(A2) $f^{(d)}$ can be represented as

$$f^{(d)}(\underline{x}) = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\underline{x}^{\mathrm{T}}} \widetilde{F^{(d)}}(\mathrm{d}\omega), \tag{4.4}$$

where the Fourier distribution $\widetilde{F^{(d)}}$ of $f^{(d)}$ satisfies

$$\int_{\mathbb{R}^d} \left| \widetilde{F^{(d)}}(\mathrm{d}\omega) \right| \leqslant L_0 \tag{4.5}$$

and

$$\int_{\mathbb{R}^d} \|\omega\|_{l_\infty} \left| \widetilde{F^{(d)}}(\mathrm{d}\omega) \right| \leqslant L_1, \tag{4.6}$$

for some $L_0, L_1 < \infty$.

We decided to choose this particular restriction of the complexity of $f^{(d)}$ for several reasons. On the one hand, it is strong enough to allow rates of convergence that do not

suffer from the curse of dimensionality. On the other hand, (A2) is more general than the assumption of an additive structure and includes many cases that may not be unrealistic; see Barron (1993) for examples. Finally, it motivates a method of estimation that is quite flexible in taking advantage of different types of simple structures of $f^{(d)}$.

The complexity bounds (4.5) and (4.6) imply that $f^{(d)}$ can be approximated by a moderate number of basis functions. Using basis functions with *variable frequencies* $\omega$, one can find an approximation with $N$ terms, $\widetilde{f^{(d)}}(\underline{x}) = \sum_{k=1}^{N} \theta_k e^{i\omega_k^{\mathrm{T}}\underline{x}}$, such that

$$\int |\widetilde{f^{(d)}}(\underline{x}) - f^{(d)}(\underline{x})|^2 \, \mu^{(d)}(\mathrm{d}\underline{x}) = O(N^{-1});$$

see Barron (1993) for similar results. We will use an alternative approach based on fixed basis functions where the possible frequencies are taken from a sufficiently fine grid. Since

$$\left| f^{(d)}(\underline{x}) - \int_{\{\omega: \|\omega\|_{l_\infty} \leqslant n^{1/4}\}} e^{i\underline{x}^{\mathrm{T}}\omega} \, \widetilde{F^{(d)}}(d\omega) \right| \leqslant \int_{\{\omega: \, \|\omega\|_{l_\infty} > n^{1/4}\}} \left| \widetilde{F^{(d)}}(\mathrm{d}\omega) \right|$$

$$\leqslant n^{-1/4} L_1 \tag{4.7}$$

and the desired rate of convergence for the loss is $(\ln(n)/n)^{1/2}$, it follows that it suffices to include only parameters $\omega$ with $\|\omega\|_{l_\infty} \leqslant n^{1/4}$. Let $\Omega^{(n)} = \{\omega_0^{(n)}, \ldots, \omega_{N_n}^{(n)}\}$ be a sequence of $n^{-1/4}$ nets of cardinality $N_n + 1 = O(n^{d/2})$ for $\{\omega \in \mathbb{R}^d : \|\omega\|_{l_\infty} \leqslant n^{1/4}\}$. In particular, we choose $\omega_0^{(n)} = (0, \ldots, 0)^{\mathrm{T}}$. Since the target function $f^{(d)}$ is real-valued, it is natural to use sine and cosine functions for the approximation. We set $\phi_0^{(n)} \equiv 1$ and, for $1 \leqslant k \leqslant N_n$, $\phi_k^{(n)}(\underline{x}) = \cos(\underline{x}^{\mathrm{T}}\omega_k^{(n)})$, $\phi_{-k}^{(n)}(\underline{x}) = \sin(\underline{x}^{\mathrm{T}}\omega_k^{(n)})$. To simplify notation, we set $\omega_{-k}^{(n)} = \omega_k^{(n)}$, for $k = 1, \ldots, N_n$. The following result characterizes the ability of the chosen basis to approximate $f^{(d)}$ and serves as the basis for the derivation of statistical properties of our method of estimation.

**Lemma 4.1.** *Suppose that* (A2) *is fulfilled. Then there exist real parameters* $\tilde{\theta}_{-N_n}^{(n)}, \ldots, \tilde{\theta}_{N_n}^{(n)}$ *such that*

$$\int_{\mathbb{R}^d} \left| \sum_{k=-N_n}^{N_n} \tilde{\theta}_k^{(n)} \phi_k^{(n)}(\underline{x}) - f^{(d)}(\underline{x}) \right|^2 \mu^{(d)}(\mathrm{d}\underline{x}) \leqslant n^{-1/2} \left(2L_1^2 + 8L_0^2 \, d^2 \, \mathrm{E}X_t^2\right)$$

*and*

$$\sum_{k=-N_n}^{N_n} |\tilde{\theta}_k^{(n)}| \left( \|\omega_k^{(n)}\|_{l_\infty} + 1 \right) \leqslant 2\left(n^{-1/4} + 1\right) L_0 + 2L_1.$$

This lemma states that there exist parameter values which provide a rate of approximation of order $n^{-1/2}$. We intend to choose such parameters by some penalized least-squares method. As a starting point, we take the residual sum of squares,

$$\mathrm{RSS}(\theta) = \frac{1}{n'} \sum_{t=d}^{n-m} \left( X_{t+m} - \sum_{k=-N_n}^{N_n} \theta_k \phi_k^{(n)}(X_t, \ldots, X_{t-d+1}) \right)^2, \tag{4.8}$$

where $n' = n - m - d + 1$ is the number of summands in (4.8). Note that, up to a summand not dependent on $\theta$, $\mathrm{RSS}(\theta)$ is equivalent to

$$\begin{aligned}
&\widetilde{\mathrm{RSS}}(\theta) \\
&:= \mathrm{RSS}(\theta) - \mathrm{E}\big(X_{t+m} - f^{(d)}(X_t, \ldots, X_{t-d+1})\big)^2 - \frac{1}{n'}\sum_{t=d}^{n-m}(X_{t+m}^2 - \mathrm{E}[X_{t+m}^2]) \\
&= \int_{\mathbb{R}^d}\left(\sum_{k=-N_n}^{N_n}\theta_k\phi_k^{(n)}(\underline{x}) - f^{(d)}(\underline{x})\right)^2 \mu^{(d)}(\mathrm{d}\underline{x}) \\
&\quad - 2\sum_{k=-N_n}^{N_n}\theta_k\frac{1}{n'}\sum_{t=d}^{n-m}\Big(X_{t+m}\phi_k^{(n)}(X_t, \ldots, X_{t-d+1}) - \mathrm{E}[X_{t+m}\phi_k^{(n)}(X_t, \ldots, X_{t-d+1})]\Big). \\
&\quad + \sum_{k,l=-N_n}^{N_n}\theta_k\theta_l\frac{1}{n'}\sum_{t=d}^{n-m}\Big(\phi_k^{(n)}(X_t, \ldots, X_{t-d+1})\phi_l^{(n)}(X_t, \ldots, X_{t-d+1}) \\
&\quad - \mathrm{E}[\phi_k^{(n)}(X_t, \ldots, X_{t-d+1})\phi_l^{(n)}(X_t, \ldots, X_{t-d+1})]\Big). \qquad (4.9)
\end{aligned}$$

Hence, it is sufficient to choose the penalty term in such a way that it compensates for the last two terms in (4.9). Since the $X_t$ are not necessarily bounded, we also impose the following condition which, in particular, allows us to apply the Bernstein inequality given in Theorem 2.1 in the course of the proof of Lemma 4.2 below.

(A3) For all $M < \infty$, $\mathrm{E}|X_t|^M$ is finite.

A hint for an appropriate choice of the penalty function can be derived from the next lemma.

**Lemma 4.2.** *Suppose that* (A1) *and* (A3) *are fulfilled. For arbitrary $\lambda > 0$, there exist finite constants $K_1$, $K_2$ such that*

$$\begin{aligned}
\text{(i)}\ P\Bigg(&\left|\frac{1}{n'}\sum_{t=d}^{n-m}\Big(X_{t+m}\phi_k^{(n)}(X_t, \ldots, X_{t-d+1}) - \mathrm{E}[X_{t+m}\phi_k^{(n)}(X_t, \ldots, X_{t-d+1})]\Big)\right| \\
&> K_1\sqrt{\frac{\ln(n)}{n}}\ (\|\omega_k^{(n)}\|_{l_\infty} \vee 1) \text{ for any } k \in \{-N_n, \ldots, N_n\}\Bigg) \\
&= O(N_n\ n^{-\lambda}),
\end{aligned}$$

$$\text{(ii)}\ P\Bigg(\Bigg|\frac{1}{n'}\sum_{t=d}^{n-m}\Big(\phi_k^{(n)}(X_t,\ldots,X_{t-d+1})\phi_l^{(n)}(X_t,\ldots,X_{t-d+1}) - \mathrm{E}[\phi_k^{(n)}(X_t,\ldots,X_{t-d+1})\phi_l^{(n)}(X_t,\ldots,X_{t-d+1})]\Big)\Bigg| > K_2\sqrt{\frac{\ln(n)}{n}}\,(\|\omega_k^{(n)}\|_{l_\infty}+\|\omega_l^{(n)}\|_{l_\infty}) \text{ for any } k,\, l\in\{-N_n,\ldots,N_n\}\Bigg) = O(N_n^2\, n^{-\lambda}).$$

According to this lemma, it follows from (4.9) that

$$\int_{\mathbb{R}^d}\left(\sum_{k=-N_n}^{N_n}\theta_k\phi_k^{(n)}(\underline{x})-f^{(d)}(\underline{x})\right)^2\mu^{(d)}(\mathrm{d}\underline{x}) \leqslant \widetilde{\mathrm{RSS}}(\theta)+Z_n\sqrt{\frac{\ln(n)}{n}}\left[\sum_k|\theta_k|\sqrt{\|\omega_k^{(n)}\|_{l_\infty}\vee 1}+\sum_{k,l}|\theta_k|\,|\theta_l|\sqrt{\|\omega_k^{(n)}\|_{l_\infty}}\right] \tag{4.10}$$

holds for all $\theta\in\mathbb{R}^{2N_n+1}$, where $Z_n$ is some random variable *not dependent on* $\theta$ with $Z_n=O_P(1)$. Since, up to a summand not depending on $\theta$, $\widetilde{\mathrm{RSS}}(\theta)$ is equivalent to $\mathrm{RSS}(\theta)$, we can use inequality (4.10) to find an appropriate penalty term. Since it seems to be very cumbersome to find appropriate explicit expressions approximating the constants $K_1$ and $K_2$ from Lemma 4.2 without prior knowledge of the dependence structure, we choose an arbitrary continuous function $\rho:[0,\infty)\to[0,\infty)$ with $\rho(x)/x\to_{n\to\infty}\infty$, and define the penalty function as

$$\mathrm{Pen}_n(\theta)=\sqrt{\frac{\ln(n)}{n}}\,\rho\left(\left[\sum_k|\theta_k|\sqrt{\|\omega_k^{(n)}\|_{l_\infty}\vee 1}+\sum_{k,l}|\theta_k|\,|\theta_l|\sqrt{\|\omega_k^{(n)}\|_{l_\infty}}\right]\right). \tag{4.11}$$

Since, for arbitrary $c<\infty$, $\sup_{x\leqslant c}\sup_{y\geqslant 0}\{x\cdot y-\rho(y)\}<\infty$, it follows from $Z_n=O_P(1)$ that

$$\sup_{\theta\in\mathbb{R}^{2N_n+1}}\left\{Z_n\left[\sum_k|\theta_k|\sqrt{\|\omega_k^{(n)}\|_{l_\infty}\vee 1}+\sum_{k,l}|\theta_k|\,|\theta_l|\sqrt{\|\omega_k^{(n)}\|_{l_\infty}}\right] -\rho\left(\left[\sum_k|\theta_k|\sqrt{\|\omega_k^{(n)}\|_{l_\infty}\vee 1}+\sum_{k,l}|\theta_k|\,|\theta_l|\sqrt{\|\omega_k^{(n)}\|_{l_\infty}}\right]\right)\right\}=O_P(1). \tag{4.12}$$

This implies, in conjunction with (4.10), that

$$\int_{\mathbb{R}^d}\left(\sum_{k=-N_n}^{N_n}\theta_k\phi_k^{(n)}(\underline{x})-f^{(d)}(\underline{x})\right)^2\mu^{(d)}(\mathrm{d}\underline{x})=\widetilde{\mathrm{RSS}}(\theta)+\mathrm{Pen}_n(\theta)+R_n(\theta), \tag{4.13}$$

where $\sup_{\theta\in\mathbb{R}^{2N_n+1}} R_n(\theta) = O_P(\sqrt{\ln(n)/n})$.

$\hat{\theta}$ is now chosen by penalized least-squares, that is, it is a measurable function with

$$\hat{\theta} \in \arg\min_{\theta\in\mathbb{R}^{2N_{n+1}}} \{\mathrm{RSS}(\theta) + \mathrm{Pen}_n(\theta)\}. \tag{4.14}$$

(The existence of such a measurable version is guaranteed by Lemma 2 in Jennrich 1969.) This leads to the estimator

$$f_{\hat{\theta}}(\underline{x}) = \sum_{k=-N_n}^{N_n} \hat{\theta}_k \phi_k^{(n)}(\underline{x}).$$

A similar penalized minimum distance method has been developed, for example, in Barron *et al*. (1999). The following theorem is the main result in this section.

**Theorem 4.1.** *Suppose that* (A1), (A2) *and* (A3) *are fulfilled. Then*

$$\int_{\mathbb{R}^d} \left(f_{\hat{\theta}}(\underline{x}) - f^{(d)}(\underline{x})\right)^2 \mu^{(d)}(\mathrm{d}\underline{x}) = O_P\left(\sqrt{\frac{\ln(n)}{n}}\right).$$

This result basically means that the rate of convergence does not depend on the dimension $d$, but is related to the actual difficulty of the problem that derives from the complexity of $f^{(d)}$. Due to the use of the function $\rho$ we do not have to worry about the right constant in the penalty term. This means that we need not estimate variances of certain sums which could turn out to be quite involved if the dependence structure is complex. However, we agree with an Associate Editor who pointed out that the result of Theorem 4.1 is therefore purely asymptotic.

# 5. Proofs of approximation-theoretic and statistical results

***Proof of Lemma 4.1.*** We can decompose the complex-valued measure $\widetilde{F^{(d)}}$ as $\widetilde{F_{\mathrm{re}}^{(d)}} - \mathrm{i}\widetilde{F_{\mathrm{im}}^{(d)}}$, where $\widetilde{F_{\mathrm{re}}^{(d)}}$ and $\widetilde{F_{\mathrm{im}}^{(d)}}$ are both real-valued measures, and obtain

$$f^{(d)}(\underline{x}) = \int_{\mathbb{R}^d} \cos(\underline{x}^{\mathrm{T}}\omega)\, \widetilde{F_{\mathrm{re}}^{(d)}}(\mathrm{d}\omega) + \int_{\mathbb{R}^d} sin(\underline{x}^{\mathrm{T}}\omega)\, \widetilde{F_{\mathrm{im}}^{(d)}}(\mathrm{d}\omega). \tag{5.1}$$

Let $\Omega_0^{(n)}, \ldots, \Omega_{N_n}^{(n)}$ be any decomposition of $\{\omega : \|\omega\|_{l_\infty} \leqslant n^{1/4}\}$ into disjoint subsets with the property $\Omega_k^{(n)} \subseteq \{\omega : \|\omega - \omega_k^{(n)}\|_{l_\infty} \leqslant n^{-1/4}\}$. (Such a decomposition exists since $\Omega^{(n)}$ is an $n^{-1/4}$ net.) We define the approximation

$$\widetilde{f^{(d)}}(\underline{x}) = \sum_{k=-N_n}^{N_n} \tilde{\theta}_k^{(n)} \phi_k^{(n)}(\underline{x}),$$

where $\tilde{\theta}_k^{(n)} = \int_{\Omega_k^{(n)}} \widetilde{F_{\mathrm{re}}^{(d)}}(\mathrm{d}\omega)$, for $k = 0, \ldots, N_n$, and $\tilde{\theta}_k^{(n)} = \int_{\Omega_{-k}^{(n)}} \widetilde{F_{\mathrm{im}}^{(d)}}(\mathrm{d}\omega)$, for $k = -N_n, \ldots, -1$.

With $f^{(d)}_{\text{trunc}}(\underline{x}) = \int_{\{\omega:\|\omega\|_{l_\infty}\leqslant n^{1/4}\}} \mathrm{e}^{\mathrm{i}\underline{x}^{\mathrm{T}}\omega}\,\widetilde{F^{(d)}}(\mathrm{d}\omega)$, we obtain that

$$\begin{aligned}
&|f^{(d)}(\underline{x}) - \widetilde{f^{(d)}}(\underline{x})| \\
&\quad\leqslant |f^{(d)}(\underline{x}) - f^{(d)}_{\text{trunc}}(\underline{x})| + |f^{(d)}_{\text{trunc}}(\underline{x}) - \widetilde{f^{(d)}}(\underline{x})| \\
&\quad\leqslant \int_{\{\omega:\|\omega\|_{l_\infty}>n^{1/4}\}} |\widetilde{F^{(d)}}(\mathrm{d}\omega)| + \sum_{k=-N_n}^{N_n} |\tilde{\theta}^{(n)}_k| \sup_{\omega\in\Omega^{(n)}_{|k|}} |\mathrm{e}^{\mathrm{i}\underline{x}^{\mathrm{T}}\omega} - \mathrm{e}^{\mathrm{i}\underline{x}^{\mathrm{T}}\omega^{(n)}_k}| \\
&\quad\leqslant n^{-1/4}(L_1 + 2L_0\|\underline{x}\|_{l_1}).
\end{aligned}$$

This implies that

$$\int |f^{(d)}(\underline{x}) - \widetilde{f^{(d)}}(\underline{x})|^2\,\mu^{(d)}(\mathrm{d}\underline{x}) \leqslant n^{-1/2}(2L_1^2 + 8L_0^2\,d^2\,\mathrm{E}X_t^2).$$

Furthermore, we obtain that

$$\sum_{k=-N_n}^{N_n} |\tilde{\theta}^{(n)}_k|\|\omega^{(n)}_k\|_{l_\infty} \leqslant 2\sum_{k=0}^{N_n} \int_{\Omega^{(n)}_k} (\|\omega\|_{l_\infty} + n^{-1/4})|\widetilde{F^{(d)}}(\mathrm{d}\omega)| \leqslant 2(L_1 + n^{-1/4}L_0)$$

as well as

$$\sum_{k=-N_n}^{N_n} |\tilde{\theta}^{(n)}_k| \leqslant 2\sum_{k=0}^{N_n} \int_{\Omega^{(n)}_k} |\widetilde{\mathrm{F}^{(d)}}(\mathrm{d}\omega)| \leqslant 2L_0,$$

which proves the second assertion. □

***Proof of Lemma 4.2.*** We have that

$$\mathrm{Lip}(\phi^{(n)}_k\phi^{(n)}_l) \leqslant \mathrm{Lip}(\phi^{(n)}_k)\cdot\|\phi^{(n)}_l\|_{L_\infty} + \mathrm{Lip}(\phi^{(n)}_l)\cdot\|\phi^{(n)}_k\|_{L_\infty} \leqslant \|\omega^{(n)}_k\|_{l_\infty} + \|\omega^{(n)}_l\|_{l_\infty},$$

which implies that

$$\mathrm{var}\left(\sum_{t=d}^{n-m} \phi^{(n)}_k(X_t, \ldots, \mathrm{X}_{t-d+1})\phi^{(n)}_l(\mathrm{X}_t, \ldots, X_{t-d+1})\right) = O(n\,(\|\omega^{(n)}_k\|_{l_\infty} + \|\omega^{(n)}_k\|_{l_\infty})).$$

Assertion (ii) now follows directly from Theorem 2.1.

It remains to prove (i). Since the random variables $X_t$ are not necessarily bounded we will replace them by the following truncated versions. We choose any $\delta \in (0, 3/8)$ and define

$$\widetilde{X}_t = \max\{\min\{X_t, n^\delta\}, -n^\delta\}, \qquad t = 1, \ldots, n.$$

It now follows by Markov's inequality from (A3) that

$$P\left(X_t \neq \widetilde{X}_t\right) = O(n^{-\lambda})$$

and

$$\mathrm{E}|X_t - \widetilde{X}_t|^r \leqslant \sqrt{P\Big(X_t \neq \widetilde{X}_t\Big)}\sqrt{\mathrm{E}|X_t - \widetilde{X}_t|^{2r}} = O(n^{-\lambda}), \qquad r = 1, 2, \tag{5.2}$$

hold for arbitrary $\lambda < \infty$. We define random variables $Z_{t,k} = X_{t+m}\phi_k^{(n)}(X_t, \ldots, X_{t-d+1})$ and $\widetilde{Z}_{t,k} = \widetilde{X}_{t+m}\phi_k^{(n)}(X_t, \ldots, X_{t-d+1})$. Furthermore, we set $K_n^2 = \|\omega_k^{(n)}\|_{l_\infty} \vee 1$ and $M_n = n^{4\delta/3}$. From (A1) and (5.2) we obtain that

$$\mathrm{var}\left(\sum_{t=d}^{n-m} \widetilde{Z}_{t,k}\right) = \mathrm{var}\left(\sum_{t=d}^{n-m} Z_{t,k}\right) + O(n^{-\lambda}) = O\big(nK_n^2\big). \tag{5.3}$$

Furthermore, according to (A1), the random variables $\widetilde{Z}_{t,k}$ are weakly dependent and it follows in conjunction with (A3), for $s_1 \leqslant \ldots \leqslant s_u \leqslant t_1 \leqslant \ldots \leqslant t_v$, and $u + v \geqslant 3$, that

$$\begin{aligned}|\mathrm{cov}(\widetilde{Z}_{s_1,k} \ldots \widetilde{Z}_{s_u,k}, \widetilde{Z}_{t_1,k} \ldots \widetilde{Z}_{t_v,k})| &\leqslant Cv(n^\delta)^v K_n^2\, \mathrm{e}^{-\beta(t_1-d+1-s_u+m)} \\ &\leqslant \widetilde{C}K_n^2\, M_n^{u+v-2}\, v\mathrm{e}^{-\beta(t_1-s_u)}. \end{aligned} \tag{5.4}$$

With the choice of $A_n = 2n\widetilde{C}K_n^2/(1-\mathrm{e}^{-\beta})$ which is possible by (2.3), we obtain by the Bernstein-type inequality from Theorem 2.1 that

$$P\left(\left|\sum_{t=d}^{n-m} \widetilde{Z}_{t,k}\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2/2}{A_n + B_n^{1/3}t^{5/3}}\right),$$

where $B_n = O(\|\omega_k^{(n)}\|_{l_\infty} \vee n^\delta) = O(n^{1/4})$. Therefore, we obtain that

$$\begin{aligned}&P\Bigg(\left|\frac{1}{n'}\sum_{t=d}^{n-m}\Big[X_{t+m}\phi_k^{(n)}(X_t, \ldots, X_{t-d+1}) - \mathrm{E}X_{t+m}\phi_k^{(n)}(X_t, \ldots, X_{t-d+1})\Big]\right| \\ &> K_1\sqrt{\frac{\ln(n)}{n}\,(\|\omega_k^{(n)}\|_{l_\infty} \vee 1)} \text{ for any } k \in \{-N_n, \ldots, N_n\}\Bigg) \\ &\leqslant P\Bigg(\left|\frac{1}{n'}\sum_{t=d}^{n-m}\Big[\widetilde{X}_{t+m}\phi_k^{(n)}(X_t, \ldots, X_{t-d+1}) - \mathrm{E}\widetilde{X}_{t+m}\phi_k^{(n)}(X_t, \ldots, X_{t-d+1})\Big]\right| \\ &> K_1\sqrt{\frac{\ln(n)}{n}\,(\|\omega_k^{(n)}\|_{l_\infty} \vee 1)} - C_\lambda n^{-\lambda} \text{ for any } k \in \{-N_n, \ldots, N_n\}\Bigg) \\ &+ P\Big(X_{t+m} \neq \widetilde{X}_{t+m} \text{ for any } t \in \{d, \ldots, n-m\}\Big) \\ &= O(N_n\, n^{-\lambda}).\end{aligned}$$

□

***Proof of Theorem 4.1.*** Let $\tilde{\theta}^{(n)} = (\tilde{\theta}_{-N_n}^{(n)}, \ldots, \tilde{\theta}_{N_n}^{(n)})^{\mathrm{T}}$ be as in Lemma 4.1. It follows from Lemmas 4.1 and 4.2 that

$$\widetilde{\mathrm{RSS}}(\tilde{\theta}^{(n)}) + \mathrm{Pen}_n(\tilde{\theta}^{(n)}) = O_P\left(\sqrt{\frac{\ln(n)}{n}}\right).$$

Since $\widetilde{\mathrm{RSS}}(\hat{\theta}) + \mathrm{Pen}_n(\hat{\theta}) \leqslant \widetilde{\mathrm{RSS}}(\tilde{\theta}^{(n)}) + \mathrm{Pen}_n(\tilde{\theta}^{(n)})$ we obtain from (4.13) that

$$\int_{\mathbb{R}^d} \left(\sum_k \hat{\theta}_k \phi_k(\underline{x}) - f^{(d)}(\underline{x})\right)^2 \mu^{(d)}(\mathrm{d}\underline{x}) = O_P\left(\sqrt{\frac{\ln(n)}{n}}\right).$$

□

## Acknowledgement

We thank two anonymous referees and an Associate Editor for their valuable comments. In particular, they pointed out to us an error in the proof of Theorem 2.1.

## References

Ango Nze, P., Bühlmann, P. and Doukhan, P. (2002) Nonparametric regression estimation under weak dependence beyond mixing and association. *Ann. Statist.*, **30**, 397–430.

Barron, A. (1993) Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, **39**, 930–945.

Barron, A. (1994) Approximation and estimation bounds for artificial neural networks. *Machine Learning*, **14**, 115–133.

Barron, A., Birgé, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, **113**, 301–413.

Bennett, G. (1962) Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.*, **57**, 33–45.

Bentkus, R. and Rudzkis, R. (1980) On exponential estimates of the distribution of random variables, *Lithuanian Math. J.*, **20**, 15–30 (in Russian).

Coulon-Prieur, C. and Doukhan, P. (2000) A triangular CLT for weakly dependent sequences. *Statist. Probab. Lett.*, **47**, 61–68.

Dedecker, J. and Doukhan, P. (2003) A new covariance inequality and applications. *Stochastic Process. Appl.*, **106**, 63–80.

Dedecker, J. and Prieur, C. (2004) Coupling for $\tau$-dependent sequences and applications. *J. Theoret. Probab.*, **17**, 861–885.

Delyon, B. and Juditsky, A. (2000) On minimax identification of nonparametric autoregressive models. *Probab. Theory Related Fields*, **116**, 21–39.

Doukhan, P. and Louhichi, S. (1999) A new weak dependence condition and application to moment inequalities. *Stochastic Process. Appl.*, **84**, 313–342.

Jennrich, R.I. (1969) Asymptotic properties of non-linear least-squares estimators. *Ann. Math. Statist.*, **40**, 633–643.

Juditsky, A. and Nemirovski, A. (2000) Functional aggregation for nonparametric regression. *Ann. Statist.*, **28**, 681–712.

Modha, D.S. and Masry, E. (1996) Minimum complexity regression estimation with weakly dependent observations. *IEEE Trans. Inform. Theory*, **42**, 2133–2145.

Neumann, M.H. and Paparoditis, E. (2004) Goodness-of-fit tests for Markovian time series models. Manuscript.

Saulis, L. and Statulevicius, V.A. (1991) *Limit Theorems for Large Deviations*. Dordrecht: Kluwer.

Scott, D.W. (1992) *Multivariate Density Estimation. Theory, Practice and Visualisation.* New York: Wiley.

Statulevicius, V.A. (1970) Limit theorems for random functions. *Lithuanian Math. J.*, **10**, 583–592, (in Russian).

Received May 2002 and revised August 2005