Dynamic System Identification with Order Statistics

Wlodzimierz Greblicki, *Member, IEEE*, and Miroslaw Pawlak, *Member, IEEE*

***Abstract*—Systems consisting of linear dynamic and memoryless nonlinear subsystems are identified. The paper deals with systems in which the nonlinear element is followed by a linear element, as well as systems in which the subsystems are connected in parallel. The goal of the identification is to recover the nonlinearity from noisy input–output observations of the whole system; signals interconnecting the elements are not measured. Observed values of the input signal are rearranged in increasing order, and coefficients for the expansion of the nonlinearity in trigonometric series are estimated from the new sequence of observations obtained in this way. Two algorithms are presented, and their mean integrated square error is examined. Conditions for pointwise convergence are also established. For the nonlinearity satisfying the Lipschitz condition, the error converges to zero. The rate of convergence derived for differentiable nonlinear characteristics is insensitive to the roughness of the probability density of the input signal. Results of numerical simulation are also presented.**

***Index Terms*—Nonlinear system identification, nonparametric regression, order statistics, spacings, orthogonal series.**

I. Introduction

Regression analysis is a standard tool used for recovering a general relationship between two random variables. Applied to nonlinear system identification, the analysis makes possible the recovery of the nonlinear characteristic of the system, provided that the nonlinearity can be represented as a regression function. Classically, the relationship between the random variables, i.e., between input and output signals of the identified system, is postulated to be of a parametric form. There is no particular reason, however, to assume that the data observed in a certain real system are actually related parametrically. One is obviously closer to real situations when one assumes that the identified characteristic, i.e., a regression function in terms of mathematical statistics, is just bounded, or continuous, or satisfies the Lipschitz condition, or is differentiable, or square integrable, and so on. Each of those assumptions is clearly very mild, and leads to nonparametric inference since the class of all possible characteristics cannot be represented in a parametric form. In this paper, our *a priori* knowledge about the unknown characteristic is nonparametric, and we apply the nonparametric regression methodology to recover it. We refer to [9], [10], [19], [22], [30], [33], [35], [37], [39] for a summary and large bibliography concerning nonparametric techniques.

In this paper, we identify nonlinear dynamic systems consisting of memoryless nonlinear elements and linear dynamic subsystems. We consider two types of such systems, i.e., a system in which the nonlinear subsystem is followed by the dynamic one, and a system in which the subsystems are connected in parallel. In order to recover the characteristic of the nonlinear subsystem in both cases, we propose two identification algorithms employing Fourier expansions. In order to estimate the coefficients of the trigonometric expansion of the unknown characteristic, i.e., the Fourier regression coefficients, we rearrange input observations in increasing order. In this way, we partition the range of the input signal into intervals whose ends are determined by consecutive ordered observations. All the intervals (called spacings in the statistical literature) have random length; see [32]. As a result, the estimates of the coefficients, as well as the identification algorithms themselves, are of the form of a certain combination of order statistics.

In this way, we obtain computationally simple estimates of the unknown nonlinear characteristic. We examine both global and pointwise properties of our estimates. Imposing some smoothness restrictions on the identified characteristic, we give the rate of convergence of the algorithms of the characteristic. Our estimates converge at a rate insensitive to the roughness of the probability density of the input signal. In fact, we only require for the density to be bounded away from zero on its support.

The algorithms are also used to identify a nonlinearity in a system in which the memoryless and the dynamic systems are connected in parallel. We also present some results of simulation experiments.

The nonparametric approach to the identification of Hammerstein cascade systems, i.e., systems in which a nonlinear memoryless element is followed by a linear dynamic subsystem, has been proposed by Greblicki and Pawlak [14]–[16]; see also Greblicki [12], Pawlak [31], and Krzyzak [25]–[26]. Wiener systems, i.e., systems in which the subsystems are connected in reverse order, also can be identified nonparametrically; see Greblicki [17]. Identi-

Manuscript received June 17, 1992; revised May 15, 1993. The work of W. Greblicki was supported by NSERC Grant A8131 and a KBN Grant. The work of M. Pawlak was supported by NSERC Grant A8131 and the Humboldt Foundation. This paper was presented in part at the IEEE Conference on Information Theory, San Antonio, TX, January 1993.

W. Greblicki is with the Institute of Engineering Cybernetics, Technical University of Wroclaw, 50-370 Wroclaw, Poland.

M. Pawlak is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Man., Canada R3T 5V6.

IEEE Log Number 9404932.

fication algorithms introduced in this paper are simpler than those applying orthogonal series examined by Greblicki and Pawlak [15]–[16], Greblicki [12], Pawlak [31], and Krzyzak [25]. We directly expand the unknown nonlinear characteristic into trigonometric series, while the authors mentioned above expanded some functions depending on the unknown characteristic and then recovered the nonlinearity from the expansions. As a consequence, algorithms proposed by them are of the form of a fraction. Moreover, the convergence rate of our estimates is insensitive to the roughness of the input probability density, a property not observed for their estimates.

From another viewpoint, our approach is called block-oriented; see, e.g., Billings [5], since we assume that the nonlinear dynamic system consists of relatively simple elements and has a known structure. The goal is to recover properties of subsystems from input–output observations of the whole system. Because signals interconnecting the subsystems are not measured, the subsystems can be identified up to some unknown constant, which is also the case with this paper. This approach is an interesting proposition in the area of identification of nonlinear systems; see Bendat [4] and Brillinger [6]. A nonparametric approach to other classes of nonlinear systems has been made by Georgiev [11] who, however, imposed on the system some restrictions rather difficult to be verified. For related problems of nonparametric inference from dependent data, we refer to [19], [22], 23], and references cited therein.

II. Preliminaries

In this section, we give some basic results concerning order statistics which will be used in the paper. Suppose that $\{U_n;\ n = \cdots, -1, 0, 1, 2, \cdots\}$ is a sequence of independent random variables, bounded to the interval $[-\pi, \pi]$, and having a probability density denoted by f. We assume, however, that there exists $\delta > 0$ such that

$$\delta \le f(u), \tag{1}$$

all $u \in (-\pi, \pi)$, i.e., it is assumed that f is bounded away from zero on $(-\pi, \pi)$. In further parts of the paper, $\{U_n\}$ will be the input signal of the identified system. We rearrange the sequence $U_1, U_2, \cdots, U_n$ into a new one:

$$U_{(1)}, U_{(2)}, \cdots U_{(n)}$$

in which $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$. Ties have zero probability since the U_n's have a density. Moreover, we define $U_{(0)} = -\pi$ and $U_{(n+1)} = \pi$. The sequence $U_{(0)}, U_{(1)}, U_{(2)}, \cdots, U_{(n)}, U_{(n+1)}$ is called the order statistics of $U_1, U_2, \cdots, U_n$, while $U_{(n+1)} - U_{(n)}, U_{(n)} - U_{(n-1)}, \cdots, U_{(1)} - U_{(0)}$ are called spacings. Spacings play important role in the paper, and we now give some useful results (see [8], [32], [40] for a review of the theory of order statistics and spacings). We shall need the following.

Lemma 1: Let f satisfy (1). Then, for any real $p > 0$ and any $n \ge 1$,

$$E(U_{(1)} - U_{(i-1)})^p \le \tau_p \delta^{-p} n^{-p}$$

any $i = 1, 2, \cdots, n+1$, some $\tau_p > 0$,

$$E\left\{\sum_{j=1}^{n+1} (U_{(j)} - U_{(j-1)})^{p+1}\right\}^2 \le 4\pi\tau_{2p+1}\delta^{-(2p+1)} n^{-2p}.$$

Moreover,

$$E(U_{(i)} - U_{(i-1)})(U_{(j)} - U_{(j-1)}) \le \delta^{-2}((n+1)(n+2))^{-1},$$

any $i, j = 1, 2, \cdots, n+1$ such that $i \ne j$.

Remark 1: Owing to (B.3) in Appendix B, we note that $\tau_p = (\sqrt{2\pi}/(\sqrt{2\pi} - 1)) p\Gamma(p) \le 1.67 p\Gamma(p)$, where $\Gamma(p)$ is the gamma function.

Proof of Lemma 1: Since, by virtue of (1),

$$U_{(i)} - U_{(i-1)} = \int_{U(i-1)}^{U_{(i)}} du \le \delta^{-1} \int_{U_{(i-1)}}^{U_{(i)}} f(u)\, du,$$

application of (B.4) and (B.5) given in Appendix B yields the first and third parts of the lemma. In order to verify the second, observe

$$\begin{aligned}\left[\sum_{j=1}^{n+1} (U_{(j)} - U_{(j-1)})^{1+p}\right]^2 &\le \sum_{j=1}^{n+1} (U_{(j)} - U_{(j-1)}) \sum_{j=1}^{n+1} (U_{(j)} - U_{(j-1)})^{2p+1} \\ &\le 2\pi \sum_{j=1}^{n+1} (U_{(j)} - U_{(j-1)})^{2p+1},\end{aligned}$$

which completes the proof. □

III. Identification Algorithms

We now identify a nonlinear dynamic system consisting of a nonlinear memoryless subsystem followed by a linear dynamic one, i.e., a Hammerstein system. We present and examine two algorithms recovering the nonlinearity from noisy input–output observations.

The identified cascade system, Fig. 1, is driven by a random process defined in Section II, i.e., by a strictly stationary white random noise $\{U_n;\ n = \cdots, -1, 0, 1, 2, \cdots\}$. Input random variables U_n's are bounded to the interval $[-\pi, \pi]$. Their probability density denoted by f is unknown. It nevertheless satisfies (1).

The system comprises two elements, the first of which is nonlinear memoryless and its characteristic is denoted by m. It means that

$$W_n = m(U_n). \tag{2}$$

We assume that m is Borel measurable and satisfies the Lipschitz condition of order α, i.e., that

$$|m(u) - m(v)| \le \gamma |u - v|^\alpha, \tag{3}$$

some $0 < \alpha \le 1$ and $\gamma > 0$, all u, v in $[-\pi, \pi]$.

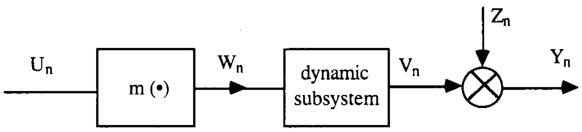

Fig. 1. The cascade system.

The dynamic subsystem is described by a state-space equation

$$\left.\begin{aligned} X_{n+1} &= AX_n + bW_n \\ V_n &= c^T X_n + dW_n \end{aligned}\right\} \tag{4}$$

where X_n is a state vector at time n and where the matrix A as well as vectors b, c, and d are all unknown. We assume nevertheless that A is an asymptotically stable matrix, i.e., that all its eigenvalues are inside the unit circle. The system is disturbed by a stationary white random noise $\{Z_n;\ n = \cdots, -1, 0, 1, 2, \cdots\}$ and

$$Y_n = V_n + Z_n.$$

The noise is independent of the input signal and has zero mean and finite variance. Thus, $\{Y_n;\ n = \cdots, -1, 0, 1, 2, \cdots\}$ is a stationary stochastic process.

Our goal is to recover m from observations (U_1, Y_1), $(U_2, Y_2), \cdots, (U_n, Y_n)$ taken at the input and output of the system.

Defining $\xi_n = c^T(X_n - EX_n)$, we get

$$Y_n = \mu(U_n) + \xi_n + Z_n \tag{5}$$

where $\mu(u) = dm(u) + \beta$, $\beta = c^T EX_n$, which means that μ is observed in the presence of noise $\xi_n + Z_n$, of which the first component is correlated, while the second is white. Observing $E\{Y_n | U_n = u\} = \mu(u)$, we shall now present two algorithms for estimating the regression of Y_n on U_n, i.e., algorithms which recover μ. That we are able to identify m up to some unknown constants d and β is a simple consequence of the cascade structure of the system and the fact that the signal W_n interconnecting the subsystems is not accessible. Nevertheless, if $f(u)$ is symmetric and $m(u)$ is odd, then $Em(U_n) = 0$, and if, additionally, $d = 1$, then $EX_n = 0$, and consequently, $\mu(u) = m(u)$.

It is worth noting the (5) can be written in the input–output form

$$Y_n = \sum_{i=0}^{\infty} g_i m(U_{n-1}) + Z_n$$

where $g_0 = d$, $g_i = c^T A^{i-1} b$, $i = 1, 2, \cdots$ is the impulse response of the linear subsystem. The class of stochastic processes $\{Y_n\}$ so generated is very broad. It covers m-dependent sequences, all autoregressive-moving average processes and certain sequences which are neither ϕ mixing nor strong mixing. Examples of sequences which are not strong mixing are easily obtained. In fact, the first-order autoregressive process $Y_n = \frac{1}{2}Y_{n-1} + m(U_n)$, i.e., $Y_n = \sum_{j=0}^{\infty} 2^{-j} m(U_{n-j})$, is strongly mixing if $m(U_n)$ has a distribution with absolutely integrable characteristic function [1], [2]. Hence, if $\{m(U_n)\}$ is a sequence of i.i.d. symmetric Bernoulli random variables, then $\{Y_n\}$ is not strongly mixing [1], [2]. The latter requirement can be easily satisfied by taking $m(u) = 1$ for $u \geq 0$ and $m(u) = 0$ otherwise, with $f(u)$ being a symmetric density on $[-\pi, \pi]$. This observation should be contrasted with the most literature on the nonparametric estimation for dependent data, where some sort of mixing conditions are assumed; see, e.g., [19], [22], [23].

The identification algorithms proposed in this paper make use of the trigonometric series $\{e^{iku};\ k = 0, \pm 1, \pm 2, \cdots\}$, orthogonal on the interval $[-\pi, \pi]$. μ is obviously integrable and

$$\mu(u) \sim \sum_{k=-\infty}^{\infty} c_k e^{iku} \tag{6}$$

where

$$c_k = (1/2\pi) \int_{-\pi}^{\pi} \mu(u) e^{-iku}\, du \tag{7}$$

is the kth Fourier coefficient of the expansion of μ. It is worth noting that in [24] the problem of recovering the regression model

$$Y_n = \sum_{|k| \leq q} c_k e^{ikU_n} + Z_n$$

has been studied (clearly, e^{iku} can be replaced by other orthogonal functions). Here, one would like to estimate $\{c_k\}$ and q (model order). It has been assumed, however, that q is finite and smaller than some known number. Our case is fully nonparametric since $m(u)$ is represented by $\sum_{k=-\infty}^{\infty} c_k e^{iku}$.

We estimate the Fourier coefficients in (7) from the order statistics $U_{(0)}, U_{(1)}, U_{(2)}, \cdots, U_{(n)}, U_{(n+1)}$ rather than from the input sequence itself. In order to estimate the coefficients, we rearrange the sequence of input–output observations $(U_1, Y_1), (U_2, Y_2), \cdots, (U_n, Y_n)$ into the following one:

$$(U_{(1)}, Y_{[1]}), (U_{(2)}, Y_{[2]}), \cdots, (U_{(n)}, Y_{[n]}).$$

Note that $Y_{[i]}$'s are not ordered; they are just paired with $U_{(i)}$'s.

We propose two estimators of c_k, namely,

$$\hat{c}_k = (1/2\pi) \sum_{j=1}^{n} Y_{[j]} \int_{U_{(j-1)}}^{U_{(j)}} e^{-iku}\, du$$

and

$$\tilde{c}_k = (1/2\pi) \sum_{j=1}^{n} Y_{[j]} \int_{U_{(j-1)}}^{U_{(j)}} e^{-ikU_{(j)}}\, du$$
$$= (1/2\pi) \sum_{j=1}^{n} Y_{[j]} (U_{(j)} - U_{(j-1)}) e^{-ikU_{(j)}}.$$

As a consequence, our identification algorithms are the following:

$$\hat{\mu}(u) = \sum_{|k| \le q(n)} \hat{c}_k e^{iku} \tag{8}$$

and

$$\tilde{\mu}(u) = \sum_{|k| \le q(n)} \tilde{c}_k e^{iku} \tag{9}$$

where $\{q(n)\}$ is an integer sequence. Obviously, the quantities in (8) and (9) are estimates of $S_{q(n)}(u) = \sum_{|k| \le q(n)} c_k e^{iku}$, i.e., of the $q(n)$th partial sum of the expansion of μ in the trigonometric series. It is also worth noting that the estimates in (8) and (9) can be represented in a convolution form, i.e.,

$$\hat{\mu}(u) = \sum_{j=1}^{n} Y_{[j]} \int_{U_{(j-1)}}^{U_{(j)}} D_{q(n)}(u - v)\, dv \tag{8$'$}$$

$$\tilde{\mu}(u) = \sum_{j=1}^{n} Y_{[j]}(U_{(j)} - U_{(j-1)}) D_{q(n)}(u - U_{(j)}) \tag{9$'$}$$

where $2\pi D_q(u) = \sum_{|k| \le q} e^{-iku} = (\sin(q + \frac{1}{2})u / \sin(u/2)$ and $\pi D_q(u)$ is the Dirichlet kernel of the qth order [41]. The representations in (8$'$) and (9$'$) are useful for studying the pointwise properties of our estimates; see Section V for a short discussion of this issue.

We shall need the following crucial lemma.

Lemma 2: Let f be any density on $[-\pi, \pi]$. Let m be any bounded Borel measurable function on $[-\pi, \pi]$. For the cascade system whose linear part is asymptotically stable,

$$E\left\{\xi_{[i]}(U_{(i)} - U_{(i-1)})\right\}^2 \le \rho_1 E(U_{(i)} - U_{(i-1)})^2,$$

any i and any $n \ge 1$.

Furthermore,

$$E\left\{\xi_{[i]}\xi_{[j]}(U_{(i)} - U_{(i-1)})(U_{(j)} - U_{(j-1)})\right\} \le \rho_2 n^{-1} E\left\{(U_{(i)} - U_{(i-1)})(U_{(j)} - U_{(j-1)})\right\},$$

any i and j such that $i \ne j$ and any $n \ge 1$, where ρ_1 and ρ_2 are independent of i, j, n as well as f.

The proof of the lemma is in Appendix A. From Lemmas 1 and 2, we get the following.

Lemma 3: Let f satisfy (1). Let m be any bounded Borel measurable function on $[-\pi, \pi]$. For the cascade system whose linear part is asymptotically stable,

$$E\left\{\sum_{i=1}^{n} \xi_{[i]}(U_{(i)} - U_{(i-1)})\right\}^2 \le \rho n^{-1},$$

any $n \ge 1$ and some $\rho > 0$ independent of n.

IV. Mean Integrated Square Error

In this section, we show that the mean integrated square error (MISE) converges to zero as the number of observations tends to infinity. We also give the rate of the convergence. For algorithm (8), the MISE is defined as

$$\text{MISE}(\hat{\mu}) = \frac{1}{2\pi} E \int_{-\pi}^{\pi} |\hat{\mu}(u) - \mu(u)|^2\, du,$$

and similarly for (9). We assume that the integer sequence satisfies the restriction

$$q(n) \to \infty \qquad \text{as } n \to \infty \tag{10}$$

and some of the following:

$$q(n)/n \to 0 \qquad \text{as } n \to \infty \tag{11}$$

$$q(n)/n^{2\alpha} \to 0 \qquad \text{as } n \to \infty \tag{12}$$

$$q^3(n)/n^2 \to 0 \qquad \text{as } n \to \infty \tag{13}$$

where α is defined in (3).

We shall now prove the following.

Theorem 1: Let f satisfy (1), and let m satisfy (3) with $0 < \alpha \le 1$. Let A be an asymptotically stable matrix. If (10)–(12) hold, then

$$\text{MISE}(\hat{\mu}) \to 0 \qquad \text{as } n \to \infty.$$

Remark 2: For $1/2 \le \alpha$, i.e., in particular for $\alpha = 1$, (11) implies (12), and as a consequence, (10) and (11) are sufficient for the estimate to be consistent. For $0 < \alpha < 1/2$, (12) entails (11), and owing to that, (10) and (12) constitute a sufficient consistency condition.

Remark 3: Selecting $q(n) \approx n^\epsilon$, we find that the estimate is consistent for $0 < \epsilon < \min(1, 2\alpha)$. Here and throughout the paper $a_n \approx b_n$ denotes that $a_n/b_n \to c$ as $n \to \infty$, where c is a nonzero constant.

Proof of Theorem 1: Since

$$2\pi c_k = \sum_{j=1}^{n} \int_{U_{(j-1)}}^{U_{(j)}} \mu(u) e^{-iku}\, du + \int_{U_{(n)}}^{U_{(n+1)}} \mu(u) e^{-iku}\, du$$

we have

$$\begin{aligned} 2\pi(\hat{c}_k - c_k) = &\sum_{j=1}^{n} (\xi_{[j]} + Z_{[j]}) \int_{U_{(j-1)}}^{U_{(j)}} e^{-iku}\, du \\ &+ \sum_{j=1}^{n} \int_{U_{(j-1)}}^{U_{(j)}} \left[\mu(U_{(j)}) - \mu(u)\right] e^{-iku}\, du \\ &- \int_{U_{(n)}}^{U_{(n+1)}} \mu(u) e^{-iku}\, du \end{aligned}$$

where $\xi_{[i]}$ and $Z_{[i]}$ are paired with $U_{(i)}$.

Hence,

$$(2\pi)^2 E|\hat{c}_k - c_k|^2 \le 4(V_1 + V_2 + V_3 + V_4) \tag{14}$$

where

$$V_1 = \sigma_z^2 \sum_{j=1}^{n} E(U_{(j)} - U_{(j-1)})^2 \tag{15}$$

$$V_2 = E\left\{\sum_{j=1}^{n} |\xi_{[j]}|(U_{(j)} - U_{(j-1)})\right\}^2 \tag{16}$$

$$V_3 = E\left\{\sum_{j=1}^{n} \int_{U_{(j-1)}}^{U_{(j)}} |\mu(U_{(j)}) - \mu(u)|\, du\right\}^2 \tag{17}$$

and

$$V_4 = E\left\{\int_{U_{(n)}}^{U_{(n+1)}} |\mu(u)|\, du\right\}^2. \tag{18}$$

Applying Lemma 1, we get $V_1 \leq 2\sigma_Z^2\delta^{-2}n^{-1}$. Lemma 2 leads to $V_2 \leq \delta_1 n^{-1}$, some $\delta_1 > 0$. Owing to (3), the integral in the expression defining V_3 is not greater than

$$d\gamma \int_{U_{(j-1)}}^{U_{(j)}} |u - U_{(j)}|^{\alpha}\, du \leq d\gamma(1+\alpha)^{-1}(U_{(j)} - U_{(j-1)})^{1+\alpha}. \tag{19}$$

Thus, due to Lemma 1, we get $V_3 \leq 4\pi(d\gamma)^2(\alpha+1)^{-2}\tau_{2\alpha+1}\delta^{-(2\alpha+1)}n^{-2\alpha}$. Finally, since μ is bounded, $V_4 \leq 2\bar{\mu}^2\delta^{-2}n^{-2}$, where $\bar{\mu} = \sup_u |\mu(u)|$. Note that $\bar{\mu} \leq dM + \beta$, where $M = \sup_u |m(u)|$. Therefore, the quantity in (14) is bounded by $\delta_2 n^{-1} + \delta_3 n^{-2\alpha}$, some δ_2 and δ_3 independent of n.

Since

$$\mathrm{MISE}(\hat{\mu}) = \sum_{|k|\leq q(n)} E|\hat{c}_k - c_k|^2 + \sum_{|k|>q(n)} |c_k|^2$$

we get finally

$$\mathrm{MISE}(\hat{\mu}) \leq \delta_4 q(n)n^{-1} + \delta_5 q(n)n^{-2\alpha} + \sum_{|k|>q(n)} |c_k|^2,$$

some positive δ_4 and δ_5, which completes the proof. □

Remark 4: The result of Theorem 1 can be extended to the case when the noise process Z_n in (5) is not homogeneous, i.e., the input–output relationship is given by

$$Y_n = \mu(U_n) + \xi_n + \sigma(U_n)Z_n$$

where $\sigma(u)$ is a measurable bounded function on $(-\pi, \pi)$ and $\{Z_n\}$ is independent on $\{U_n\}$. It is obvious that all terms in (14), except V_1, remain unaltered. The term V_1 is now of the form

$$V_1 = EZ_1^2 \sum_{j=1}^{n} E\left(\sigma^2(U_{(j)})(U_{(j)} - U_{(j-1)})^2\right).$$

This does not exceed

$$EZ_1^2\,(\sup_u |\sigma(U)|)^2 \sum_{j=1}^{n} E(U_{(j)} - U_{(j-1)})^2 = O(n^{-1}).$$

The next theorem concerns algorithm (9).

Theorem 2: Let f satisfy (1), and let m satisfy (3) with $0 < \alpha \leq 1$. Let A be an asymptotically stable matrix. If (10), (12), and (13) hold, then

$$\mathrm{MISE}(\tilde{\mu}) \to 0 \qquad \text{as } n \to \infty.$$

Remark 5: Selecting $q(n) \approx n^{\epsilon}$, we find the estimate consistent for $0 < \epsilon \leq \min(2/3, 2\alpha)$.

Proof of Theorem 2: We have

$$\begin{aligned}2\pi(\tilde{c}_k - c_k) &= \sum_{j=1}^{n} (\xi_{[j]} + Z_{[j]}) \int_{U_{(j-1)}}^{U_{(j)}} e^{-ikU_{(j)}}\, du \\ &\quad + \sum_{j=1}^{n} \int_{U_{(j-1)}}^{U_{(j)}} \left[\mu(U_{(j)})e^{-ikU_{(j)}} - \mu(u)e^{-iku}\right] du \\ &\quad - \int_{U_{(n)}}^{U_{(n+1)}} \mu(u)e^{-iku}\, du.\end{aligned}$$

Therefore,

$$(2\pi)^2 E|\tilde{c}_k - c_k|^2 \leq 4(V_1 + V_2 + V_3' + V_4)$$

where V_1, V_2, and V_4 are defined by (15), (16), and (18), respectively, and where

$$V_3' = E\left\{\sum_{j=1}^{n} \int_{U_{(j-1)}}^{U_{(j)}} |\mu(U_{(j)})e^{-ikU_{(j)}} - \mu(u)e^{-iku}|\, du\right\}^2.$$

Since the quantity under the sign of integration is not greater than

$$d\gamma|u - U_{(j)}|^{\alpha} + \bar{\mu}k|u - U_{(j)}|,$$

where $\bar{\mu} = \sup_u |\mu(u)|$, V_3' is bounded by

$$2(d\gamma)^2(1+\alpha)^{-2} E\left\{\sum_{j=1}^{n} (U_{(j)} - U_{(j-1)})^{1+\alpha}\right\}^2 + (1/2)\bar{\mu}^2k^2 E\left\{\sum_{j=1}^{n} (U_{(j)} - U_{(j-1)})^2\right\}^2.$$

Applying Lemma 1, we find $V_3' \leq \delta_3' n^{-2\alpha} + \delta_3'' k^2 n^{-2}$, some δ_3', δ_3'', and finally,

$$\mathrm{MISE}(\tilde{\mu}) \leq \delta_9 q(n)n^{-1} + \delta_{10} q(n)n^{-2\alpha} + \delta_{11} q^3(n)n^{-2} + \sum_{|k|>q(n)} |c_k|^2, \tag{20}$$

Some δ_9, δ_{10}, δ_{11}. Observing that (13) means that $[q(n)/n^{2/3}]^3 \to \infty$ as $n \to \infty$, i.e., that (13) implies (11), we can complete the proof. □

Hence, if $\alpha \geq 1/2$, then $q(n)/n \to 0$ and $q^3(n)/n^2 \to 0$ are sufficient for convergence of $\mathrm{MISE}(\hat{\mu})$ and $\mathrm{MISE}(\tilde{\mu})$ to zero, respectively. Clearly, $q^3(n)/n^2 \to 0$ is more restrictive than $q(n)/n \to 0$. This is due to the fact that $\tilde{c}_k$ defined in (5) is a less accurate approximation of c_k than $\hat{c}_k$ in (4). However, this scheme fails if $m(u)$ is so rough that $\alpha < 1/2$. Then $q(n)/n^{2\alpha} \to 0$ is required for $\hat{\mu}(u)$ and $\max(q^3(n)/n^2, q(n)/n^{2\alpha}) \to 0$ for $\tilde{\mu}(u)$. Hence, for $\alpha < 1/3$, the condition $q(n)/n^{2\alpha} \to 0$ has to be assumed for both estimates. If $1/3 < \alpha < 1/2$, the condition $q^3(n)/n^2 \to 0$ is less restrictive than $q(n)/n^{2\alpha} \to 0$.

V. Pointwise Convergence

In this section, we study the pointwise properties of our estimates. These are summarized in the following theorem.

Theorem 3: Let f satisfy (1), and let m satisfy (3) with $0 < \alpha \le 1$. Let A be an asymptotically stable matrix. If (10)–(12) hold, then

$$E(\hat{\mu}(u) - \mu(u))^2 \to 0 \qquad \text{as } n \to \infty.$$

If, in turn, (10), (12), and (13) hold, then

$$E(\tilde{\mu}(u) - \mu(u))^2 \to 0 \qquad \text{as } n \to \infty.$$

Proof: Without loss of generality, let us consider $\hat{\mu}(u)$. Owing to (8′), one can decompose the estimate as follows:

$$\begin{aligned}\hat{\mu}(u) = &\sum_{j=1}^{n} \int_{U_{(j-1)}}^{U_{(j)}} \left[\mu(U_{(j)}) - \mu(v)\right] D_{q(n)}(u - v)\, dv\\ &+ \sum_{j=1}^{n} \xi_{[j]} \int_{U_{(j-1)}}^{U_{(j)}} D_{q(n)}(u - v)\, dv\\ &+ \sum_{j=1}^{n} Z_{[j]} \int_{U_{(j-1)}}^{U_{(j)}} D_{q(n)}(u - v)\, dv\\ &- \int_{U_{(n)}}^{U_{(n+1)}} D_{q(n)}(u - v)\mu(v)\, dv + S_{q(n)}(u)\end{aligned}$$

where $S_{q(n)}(u) = \int_{-\pi}^{\pi} D_{q(n)}(u - v)\mu(v)\, dv$ is the $q(n)$th partial sum of the expansion of $\mu(u)$ in the trigonometric series. This yields the following bound:

$$E(\hat{\mu}(u) - \mu(u))^2 \le 8(V_1 + V_2 + V_3 + V_4) + 2\left(S_{q(n)}(u) - \mu(u)\right)^2$$

where

$$V_1 = \operatorname{var}(Z_1) \sum_{j=1}^{n} E\left(\int_{U_{(j-1)}}^{U_{(j)}} D_{q(n)}(u - v)\, dv\right)^2,$$

$$V_2 = E\left\{\sum_{j=1}^{n} \xi_{[j]} \int_{U_{(j-1)}}^{U_{(j)}} D_{q(n)}(u - v)\, dv\right\}^2,$$

$$V_3 = E\left\{\sum_{j=1}^{n} \int_{U_{(j-1)}}^{U_{(j)}} \left[\mu(U_{(j)}) - \mu(v)\right] D_{q(n)}(u - v)\, dv\right\}^2,$$

$$V_4 = E\left\{\int_{U_{(n)}}^{U_{(n+1)}} \mu(v) D_{q(n)}(u - v)\, dv\right\}^2.$$

Observe that due to Dini's convergence criterion, $S_q(u) \to \mu(u)$ as $q \to \infty$ at every point $u \in (-\pi, \pi)$ where the condition in (3) is satisfied [3]. For later purposes, we note the following easily verified facts: $\int_{-\pi}^{\pi} D_q^2(v)\, dv = \pi^{-1}(q + 1/2)$ and $|D_q(u)| \le \pi^{-1}(q + 1/2)$. By this, Lemma 1, and by virtue of the mean value theorem of integration, we get

$$V_4 = E\left\{D_{q(n)}^2(u - \theta)\mu^2(\theta)(U_{(n+1)} - U_{(n)})^2\right\} = O\left(\left(\frac{q(n)}{n}\right)^2\right)$$

where $\theta \in (U_{(n)}, U_{(n+1)})$.

Concerning the term V_3, note that by the Cauchy–Schwarz inequality, we obtain

$$V_3 \le \sum_{j=1}^{n} \int_{U_{(j-1)}}^{U_{(j)}} \left[\mu(U_{(j)}) - \mu(v)\right]^2 dv \cdot \sum_{j=1}^{n} \int_{U_{(j-1)}}^{U_{(j)}} D_{q(n)}^2(u - v)\, dv.$$

Owing to (3) and Lemma 1, this is further bounded by

$$\frac{3\gamma^2}{2\pi(2\alpha + 1)} q(n) \sum_{j=1}^{n} E(U_{(j)} - U_{(j-1)})^{2\alpha+1} = O\left(\frac{q(n)}{n^{2\alpha}}\right).$$

To evaluate V_1, let us note that due to (1) and the Cauchy–Schwarz inequality, we get

$$V_1 \le \operatorname{var}(Z_1)\delta^{-1} \sum_{j=1}^{n} E\left\{d_j \int_{U_{(j-1)}}^{U_{(j)}} D_{q(n)}^2(u - v)\, dv\right\}$$

where d_j is the uniform spacing.

Applying Abel's transform [3] we can write

$$\sum_{j=1}^{n} d_j \int_{U_{(j-1)}}^{U_{(j)}} D_{q(n)}^2(u - v)\, dv = \sum_{k=1}^{n-1} (d_k - d_{k+1}) L_k + d_n L_n$$

where $L_k = \sum_{j=1}^{k} \int_{U_{(j-1)}}^{U_{(j)}} D_{q(n)}^2(u - v)\, dv$. Noting that $L_k \le (3/2\pi)q(n)$ and applying results of Appendix B.1, we can obtain that $V_1 = O(q(n)/n)$. Proceeding in a similar way and using the result of Lemma 3, we can also conclude that $V_2 = O(q(n)/n)$. The proof of Theorem 3 is thus complete. □

VI. Convergence Rates

In order to present the convergence rate of the algorithms, we need the following lemma.

Lemma 4: (Lorentz's inequality [28]; see also Bary [3, pp. 215–217]): Let g defined on $[-\pi, \pi]$ satisfy the Lipschitz condition of order α, where $0 < \alpha \le 1$, i.e., let

$$|g(x) - g(y)| \le c|x - y|^\alpha,$$

some $c > 0$, all $x, y \in [-\pi, \pi]$. Then

$$\sum_{|k| \ge q} |a_k|^2 \le dq^{-2\alpha},$$

some $d > 0$, where, a_k's are the Fourier coefficients of g.

If, in turn, $\alpha > 1/2$, then

$$|S_q(u) - g(u)| \le bq^{-(\alpha - 1/2)},$$

some b, where $S_q(u)$ is the qth partial sum of the expansion of $g(u)$ in the Fourier series.

From Lorentz's inequality and (3), it follows that $\sum_{|k| \ge q(n)} |c_k|^2 \le \rho/q(n)^{2\alpha}$, some $\rho > 0$, provided that $0 < \alpha \le 1$. From this, (3), (19), and (20), we easily get the following.

Theorem 4: Suppose the f satisfies (1) and m satisfies (3). Let $0 < \alpha \leq 1/2$. For $q(n) \approx n^{2\alpha/(1+2\alpha)}$.

$$\mathrm{MISE}(\hat{\mu}) = O(n^{-4\alpha^2/(1+2\alpha)}) \quad \text{and}$$
$$\mathrm{MISE}(\tilde{\mu}) = O(n^{-4\alpha^2/(1+2\alpha)}).$$

Let $1/2 \leq \alpha \leq 1$. For $q(n) \approx n^{1/(1+2\alpha)}$,

$$\mathrm{MISE}(\hat{\mu}) = O(n^{-2\alpha/(1+2\alpha)}) \quad \text{and}$$
$$MISE(\tilde{\mu}) = \mathrm{O}(\mathrm{n}^{-2\alpha/(1+2\alpha)}).$$

Furthermore, Theorem 3 and Lemma 4 yield the following.

Theorem 5: Suppose that f satisfies (1) and m satisfies (3). Let $1/2 < \alpha \leq 1$. For $q(n) \approx n^{1/2\alpha}$,

$$E(\hat{\mu}(u) - \mu(u))^2 = O(n^{-(2\alpha-1)/2\alpha}) \quad \text{and}$$
$$E(\tilde{\mu}(u) - \mu(u))^2 = O(n^{-(2\alpha-1)/2\alpha}).$$

From the theorem, it follows that the algorithms behave asymptotically very alike. In particular, for $\alpha = 1$, MISE for both estimates is $O(n^{-2/3})$, provided that $q(n) \approx n^{1/3}$, whereas the mean-square error is $O(n^{-1/2})$ for $q(n) \approx n^{1/2}$.

Imposing further smoothness restrictions on m, we can improve the convergence rate. We shall, however, need the following.

Lemma 5: Let g have $s - 1$ absolutely continuous derivatives, and let $g^{(r)}(-\pi) = g^{(r)}(\pi)$ for $r = 0, 1, \cdots, s-1, s \geq 1$. Suppose that

$$|g^{(s)}(x) - g^{(s)}(y)| \leq c|x - y|^{\alpha},$$

some $c > 0$, $0 < \alpha \leq 1$, all $x, y \in [-\pi, \pi]$. Then

$$\sum_{|k|\geq q} |a_k|^2 \leq d/q^{2(s+\alpha)},$$

some $d > 0$, where a_k's are the Fourier coefficients of g.

The proof of this lemma results from [28] and integration by parts; see also [9]. There is also a pointwise counterpart of Lemma 5; see [13, Lemma 2]. It is worth noting that the periodicity assumption $g^{(r)}(-\pi) = g^{(r)}(\pi)$, $r = 0, 1, \cdots, s - 1$ must be required due to the well-known edge effect (Gibbs phenomenon) of trigonometric series; see [3]. In fact, if this condition is not assumed, then

$$\sum_{|k|>q} |a_k|^2 = O(q^{-2}),$$

regardless of how many derivatives of $g(x)$ exist. As a result, the rate of convergence cannot be faster than $O(n^{-2/3})$. This phenomenon need not occur for other orthogonal series. See also [20] for some techniques for removing edge effects in the context of kernel regression estimates.

Observe now that if $m(u)$ satisfies the conditions of Lemma 5, then also $\mu(u) = dm(u) + \beta$ does. Hence, $\sum_{|k|\geq q(n)} |c_k|^2 \leq d/q^{2p}(n)$, $p = s + \alpha$, some $d > 0$. Hence, for $q(n) \approx n^{1/(1+2p)}$,

$$\mathrm{MISE}(\hat{\mu}) = O(n^{-2p/(1+2p)}) \quad \text{and}$$
$$\mathrm{MISE}(\tilde{\mu}) = O(n^{-2p/(1+2p)}).$$

For a smooth characteristic m, i.e., for large p, the rate becomes close to n^{-1}, i.e., the rate typical for parametric inference. Moreover, the rate is independent of the roughness of the input density f since the result holds for any density satisfying (1). The density can be an arbitrarily rough function and, in particular, can have no derivative. Such a property has not been observed for other regression estimates, even when a regression is observed in the presence of only a white noise. It should be stressed here that in our identification problem, the noise disturbing the regression has two components, white and correlated.

It it also interesting to compare the rate with those achieved by other estimates. Of those, the kernel estimate is most thoroughly examined in the literature, and therefore results obtained for it are not easily equaled. It is known, see e.g., Härdle [13], that the classical kernel estimate (Nadaraya–Watson estimate)

$$\sum_{j=1}^{n} Y_j K\big((u - U_j)/h(n)\big) \Big/ \sum_{j=1}^{n} K\big((u - U_j)/h(n)\big), \quad (21)$$

where K is a kernel function and $\{h(n)\}$ a number sequence, has MISE of order $O(n^{-\beta})$, where $\beta = 2\min[p/(1 + 2p), r/(2 + 2r), s/(1 + s)]$. In the expression, p and r are the number of existing derivatives of m and f, respectively, while $s - 1$ is the number of vanishing moments of the kernel.

In [29] (see also [7]), the kernel estimate of the form

$$\sum_{j=1}^{n-1} Y_{[j]}(U_{(j+1)} - U_{(j)}) \frac{1}{h(n)} K\big((u - U_{(j)})/h(n)\big) \quad (22)$$

has been studied. This is clearly a counterpart of our estimate (9). The pointwise properties of this estimate have been studied, and it has been demonstrated that it has the mean-squared error of order $O(n^{-\gamma})$, where $\gamma = 2\min[p/(2p + 1), s/(2s + 1)]$. The density of U is assumed to be defined on a compact set, continuously differentiable and to satisfy condition (1).

These rates have been obtained for conditions more comfortable than ours, i.e., for a regression observed in the presence of white noise only, and for independent pairs (U_j, Y_j)'s. Observe that the rate equals that derived by us, provided that the input density has p derivatives [for the estimate in (21)] and that the kernel is suitably selected. Our estimate achieves that rate for any density satisfying (1), i.e., for a very wide class of densities. Therefore, if the density does not have a derivative, the rate obtained for our estimate seems to be better. For example, for m and f having three and one derivatives, respectively, and for a symmetric kernel function (a common choice in practice), the kernel estimates in (21) and (22) converge as fast as $O(n^{-2/3})$ and $O(n^{-4/5})$,respectively, while ours have the convergence rate $O(n^{-6/7})$. It should be also noted that the rate derived by us equals the best possible one obtained by Stone [38].

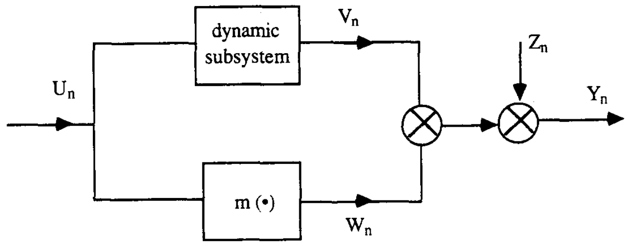

Fig. 2. The parallel system.

Finally we mention that, so far, the following orthogonal series algorithm (in random design setting) has been studied:

$$\bar{\mu}(u) = \frac{\sum_{|k| \le q(n)} \bar{a}_k e^{iku}}{\sum_{|k| \le q(n)} \bar{b}_k e^{iku}} \tag{23}$$

where

$$\bar{a} = \frac{1}{2\pi n} \sum_{j=1}^{n} Y_j e^{-ikU_j} \quad \text{and} \quad \bar{b}_k = \frac{1}{2\pi n} \sum_{j=1}^{n} e^{-ikU_j}$$

are estimates of the coefficients of expansions of $f(\cdot)\mu(\cdot)$ and $f(\cdot)$ in the trigonometric series, respectively; see [12], [13], [15], [16], [21], [25], [31]. The estimate is of a fractional form, and its pointwise convergence rate is sensitive to the roughness of the input density; see [13] and [21] for some specific convergence results concerning $\bar{\mu}(u)$ in the case of independent data.

VII. Other Systems

Some authors have examined dynamic systems described by (4) and assumed that $d = 0$. In such a case,

$$Y_n = \nu(U_{n-1}) + \zeta_n + Z_n$$

where $\nu(u) = c^T bm(u) + c^T AEX_n$ and $\zeta_n = c^T A(X_n - EX_n)$. We can say that the regression ν is observed in the presence of noise $\zeta_n + Z_n$. Therefore, one can recover ν from pairs $(U_1, Y_2), (U_2, Y_3), \cdots, (U_n, Y_{n+1})$, i.e., from order statistics

$$(U_{(1)}, Y_{[2]}), (U_{(2)}, Y_{[3]}), \cdots, (U_{(n)}, Y_{[n+1]}).$$

In the definition of our estimates, $Y_{[j]}$ should be replaced by $Y_{[j+1]}$. One can easily verify that all results in Appendix A given for ξ_n hold also for ζ_n.

Another extension would include a general description of the cascade model of the form

$$Y_n = \sum_{j=-\infty}^{n} g_{n-j} W_j + Z_n$$

where $\{g_j\}$ is the impulse response sequence satisfying $\sum_{j=0}^{\infty} g_j^2 < \infty$. Clearly, the description in (4) is of this form with $g_0 = d$, $g_j = c^T A^{j-1} b$, $j \ge 1$.

Algorithms proposed in the paper also can be used to identify systems other than cascade systems. A model of some physical interest is pictured in Fig. 2, where the nonlinear memoryless element (2) and the linear dynamic system (4) are connected in parallel; see [4]. That is,

$$\left.\begin{aligned} X_{n+1} &= AX_n + bU_n \\ V_n &= c^T X_n + dU_n \end{aligned}\right\}$$

and

$$Y_n = V_n + W_n + Z_n \tag{24}$$

where $W_n = m(U_n)$.

Clearly, this model can also be expressed in the signal-plus-noise form as in (5), i.e.,

$$Y_n = \eta(U_n) + \xi_n + Z_n$$

where $\eta(n) = m(u) + du + \beta$, $\beta = c^T EX_n$ and $\xi_n = c^T(X_n - EX_n)$. Thus, $E\{Y_n | U_n = u\} = \eta(u)$, and one can recover $m(u)$ up to a linear function. Hence, let $\hat{\eta}(u)$ and $\tilde{\eta}(u)$ be estimates of $\eta(u)$ defined as in (8) and (9), respectively, where now Y_j is generated from the model (24). In order to show that Theorems 1 and 2 hold for $\hat{\eta}(u)$ and $\tilde{\eta}(u)$, let us first observe that Lemmas 2 and 3 are in force. In fact, since now $\xi_n = \sum_{j=-\infty}^{n-1} c^T A^{n-j-1} b(U_j - EU_j)$, therefore the proof of Lemma 2 is valid here. To evaluate other terms in the proofs of Theorems 1 and 2, it remains to observe that if $m(u)$ satisfies (3), then

$$|\eta(u) - \eta(v)| \le \left(\gamma + d(2\pi)^{1-\alpha}\right)|u - v|^{\alpha}, \tag{25}$$

i.e., $\eta(u)$ is also Lipschitz with the same order as $m(u)$. All of these considerations yield the following consistency result concerning the identification of the parallel model.

Theorem 6: Let f satisfy (1), and let m satisfy (3) with $0 < \alpha \le 1$. Let the linear dynamic subsystem of the parallel model be an asymptotic stable. If (10)–(12) hold, then

$$\text{MISE}(\hat{\eta}) \to 0 \qquad \text{as } n \to \infty.$$

If, on the other hand, (10), (12), and (13) are satisfied, then

$$\text{MISE}(\tilde{\eta}) \to 0 \qquad \text{as } n \to \infty.$$

If one assumes that the density of U_n is symmetric and $m(u)$ is odd, then $EX_n = 0$, and if, additionally, $d = 1$, then we can recover $m(u) + u$, i.e., $\hat{\eta}(u) - u$, $\tilde{\eta}(u) - u$ are consistent estimates of $m(u)$.

Concerning the rate of convergence, we can note that due to (25), the results of Theorems 4 and 5 remain valid.

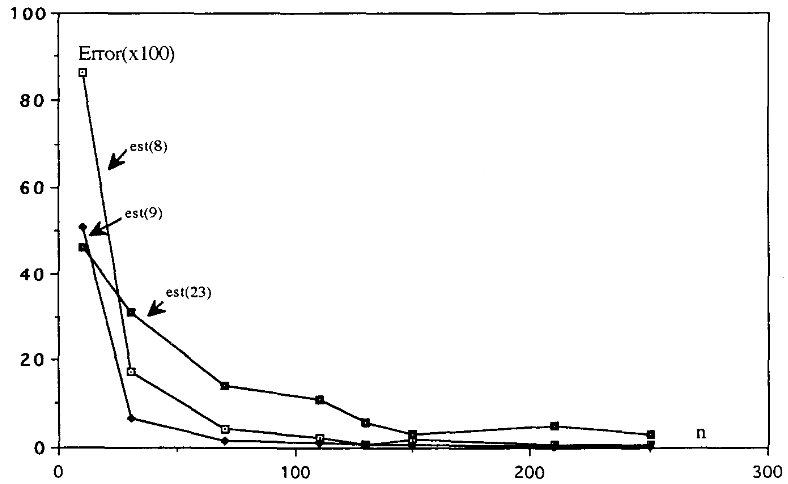

Fig. 3. Error versus n for estimates (8), (9), (23); $m(u) = m_1(u)$.

This is not the case, however, under the assumptions of Lemma 5 since $\eta(\pi) - \eta(-\pi) = 2d\pi$ and the global rate does not exceed $O(n^{-2/3})$. Nevertheless, if d is known, then $\eta(u) - du$ satisfies the conditions of Lemma 5 and $\hat{\eta}(u) - du, \bar{\eta}(u) - du$ are consistent estimates of $m(u) + \beta$, and they can reach the global rate of order $O(n^{-2p/(1+2p)})$. In particular, this is the case if $d = 0$.

It is necessary to note here that the case where A, b, d equal zero in (24), a situation in which the linear subsystem has no dynamics, $(E\{Y_n|U_n = u\} = m(u))$, has been extensively examined in the statistical literature; see, e.g., Eubank [10], Härdle [22], Muller [30], Rao [35], and Wahba [39]. Such a situation is simpler than that in this paper since the nonlinearity is then recovered from independent pairs (U_n, Y_n)'s. Orthogonal series regression estimators have been studied by Greblicki and Pawlak [13]; see also [21], but their estimators are, however, different from ours since their form is fractional. Pointwise convergence rates derived by them are sensitive to the roughness of f. We also mention that the case of deterministic regressors has been treated with the help of orthogonal series by Rafajlowicz [34] and Rutkowski [36] (see also [10, sect. 3]).

VIII. Simulation Study

To illustrate the small sample properties of the identification algorithms presented in the paper, let us consider some numerical examples. Throughout the simulation study, the following error has been empirically evaluated:

$$\text{error} = \frac{1}{n}\sum_{j=1}^{n} E|\hat{\mu}(U_j) - m(U_j)|^2$$

where $\hat{\mu}(u)$ is a certain estimate of $m(u)$. In all our experiments, the error has been determined from 20 different sets of the input–output data of size n.

The following particular descriptions for systems in (4) and (24) have been used:

$$\begin{cases} X_n = aX_{n-1} + m(U_n) \\ Y_n = X_n + Z_n \end{cases} \tag{26}$$

$$\begin{cases} X_n = aX_{n-1} + U_n \\ Y_n = X_n + m(U_n) + Z_n \end{cases} \tag{27}$$

where $|a| < 1$ and $m(u)$ is the characteristic is to be identified. To get some insight into small sample properties of our estimation techniques, let us first consider the memoryless system. Hence, the situation when $a = 0$ in (26), i.e., $Y_n = m(U_n) + Z_n$. As a nonlinear characteristic, $m_1(u) = 0.5 + 0.5\cos(u) + 2\sin(u) + 0.5\cos(2u) + \sin(2u)$ has been used; compare with [23]. Note that $m_1(u)$ is a periodic function, and therefore it seems to be ideally suited to our estimates. Fig. 3 shows error as a function of n for estimates (8), (9), and (23). It was assumed that Z_n is Gaussian $(0, 0.1)$ and U_n is uniformly distributed on $[-\pi, \pi]$. The truncation parameter $q(n)$ was selected as the minimizer of the error for $n = 250$. This value is equal to 2 for algorithms (8) and (9), while it equals 4 for the estimate in (23). Surprisingly, the estimate in (9) exhibits the best rate of convergence for small values of n. For large n, all techniques have virtually the same performance and the error becomes negligibly small for $n \geq 200$.

In the next experiment, the sensitivity of the estimates to the selection of $q(n)$ as well as various input distributions is presented. It is assumed that $n = 128$. Fig. 4 depicts the error of the estimate $\hat{\mu}(u)$ in (8) versus q for three different densities of U_n. Hence, $f1(u)$ is a uniform density of $[-\pi, \pi]$, $f2(u)$ is a piecewise constant, and is equal to $3/4\pi$ and $1/4\pi$ for $|u| \leq \pi/2$ and $\pi/2 \leq |u| \leq \pi$, respectively, and $f3(u)$ is a symmetric triangular density on $[-\pi, \pi]$. Note that $f2(u)$ is a discontinuous function, while $f3(u)$, $u \in [-\pi, \pi]$ does not satisfy the assumption in (1). Remarkably, the behavior of $\hat{\mu}(u)$ is almost insensitive to various distributions of U_n. The error

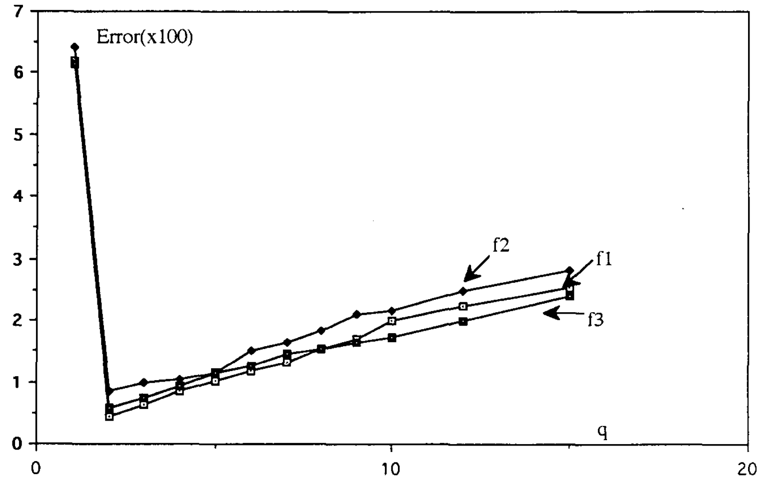

Fig. 4. Error versus q for the estimate in (8), for different distributions of U_n; $m(u) = m_1(u)$, $n = 128$.

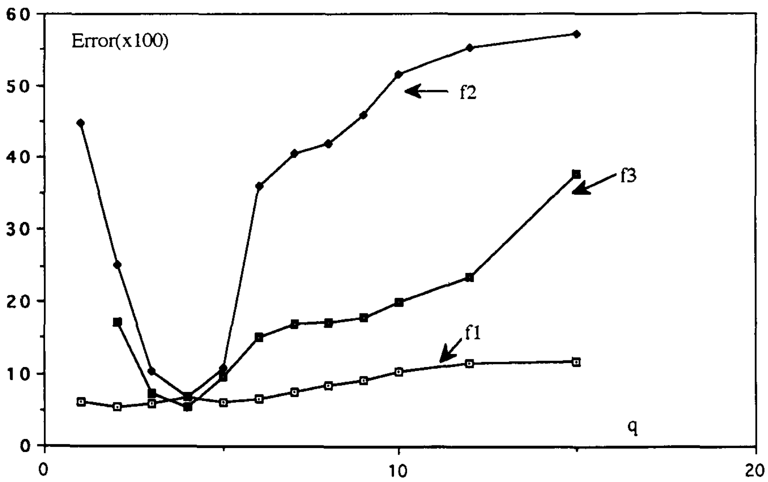

Fig. 5. Error versus q for the estimate in (23), for different distributions of U_n; $m(u) = m_1(u)$, $n = 128$.

is minimized at $q = 2$ in all cases. Fig. 5 shows an analogous result for the estimate in (23). A strong sensitivity to different input distributions can be observed. In particular, the estimate exhibits a poor performance for the discontinuous density $f2(u)$. The error is minimized at $q = 4$ in all cases.

So far, we have employed the characteristic $m_1(u)$ which is well behaved at the boundary points $u = \pm\pi$, i.e., $m_1(-\pi) = m_1(\pi)$ and $m_1^{(r)}(-\pi) = m_1^{(r)}(\pi)$, $r = 1, 2, \cdots$.

In order to see a possible deterioration in the estimate's performance for nonperiodic characteristics, let us consider the following function:

$$m_2(u) = \left(0.6 - \frac{u}{2\pi}\right)^{1/4}\left(0.6 + \frac{u}{2\pi}\right)^{1/2}.$$

Assuming that $n = 128$ and U_n is uniformly distributed, Fig. 6 shows $m_2(u)$ along with estimates (8), (9), and (23). Optimal values of $q(n)$ were used, and they are equal to 3, 2, and 5, respectively. All estimates exhibit poor behavior at the end points; see also Section V for a theoretical discussion concerning the boundary problem.

Now let $a = -0.2$ in (26) and (27) and $n = 128$. The nonlinear characteristic is an odd function of the form $m_3(u) = 0.5u + 0.5u\cos(u) + 2\sin(u) + 0.5u\cos(2u) + \sin(2u)$ and Z_n is Gaussian (0; 0.1). The error versus q for the cascade and parallel systems is depicted in Figs. 7 and 8, respectively. Note, moreover, that the error for the parallel model is calculated as $(1/n)\sum_{j=1}^{n} E|\hat{\mu}(U_j) - U_j - m(U_j)|^2$. This is due to the fact that $\hat{\mu}(u) - u$ is a consistent estimate of $m(u)$:

Clearly, the estimate (8) exhibits the best performance with, however, a much larger value of q minimizing the error (it is equal to 16 and 25 for the cascade and parallel models, respectively). Furthermore, the error for the

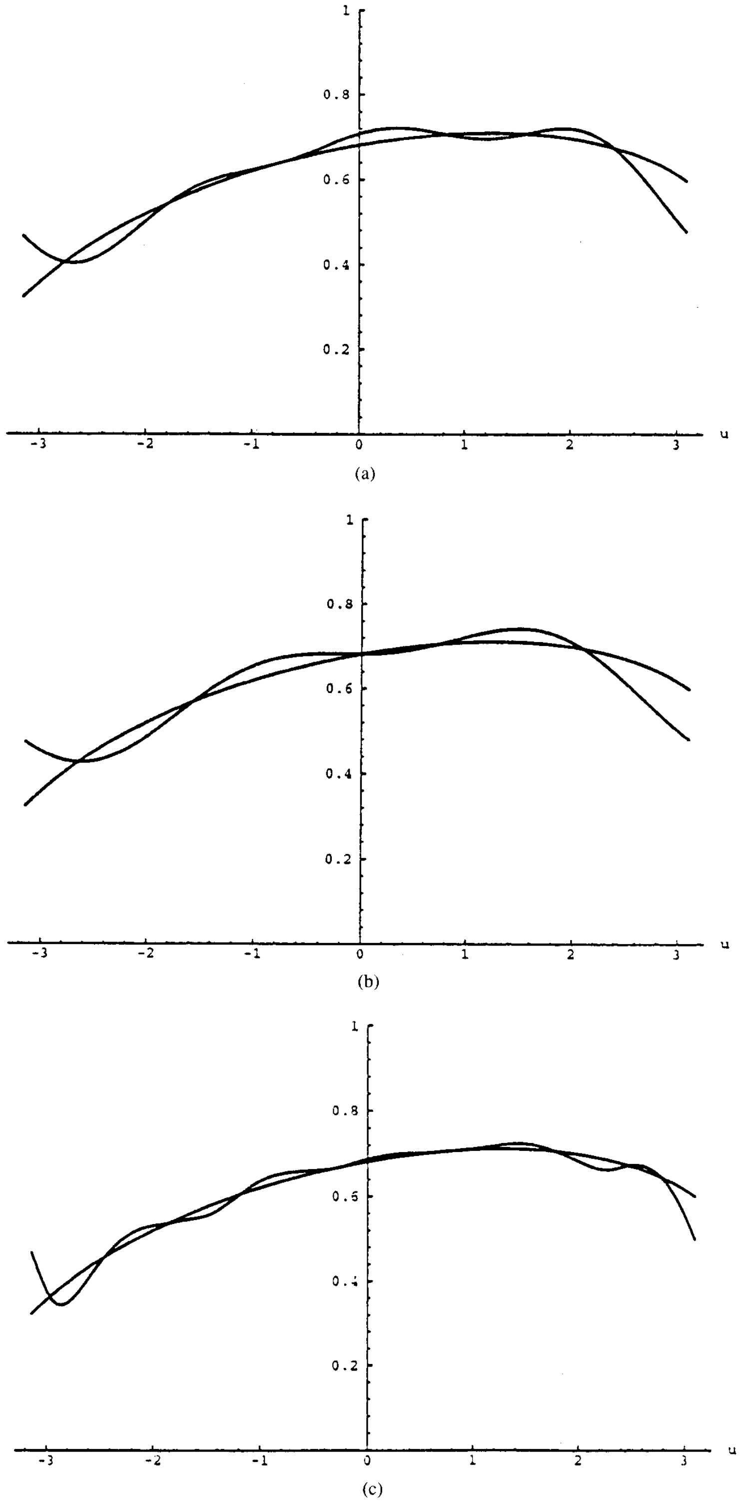

Fig. 6. (a) $m_2(u)$ and $\hat{\mu}(u)$, (b) $m_2(u)$ and $\tilde{\mu}(u)$, (c) $m_2(u)$ and $\overline{\mu}(u)$; $n = 128$.

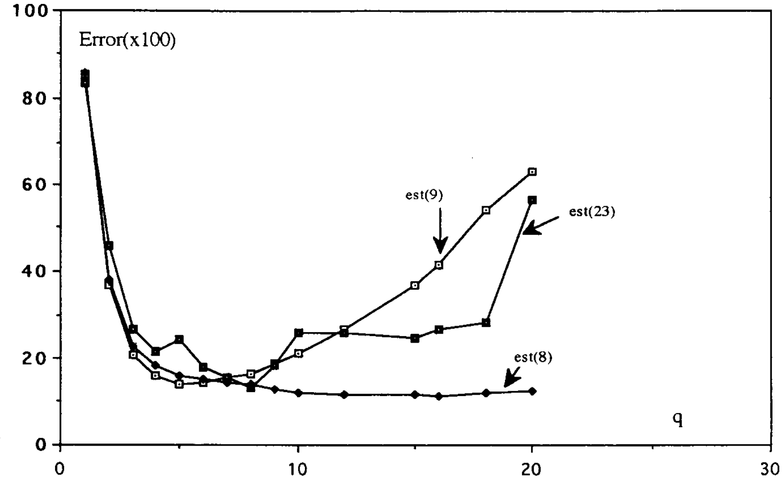

Fig. 7. Cascade dynamical system. Error versus q for estimates (8), (9), (23); $m(u) = m_3(u)$, $n = 128$.

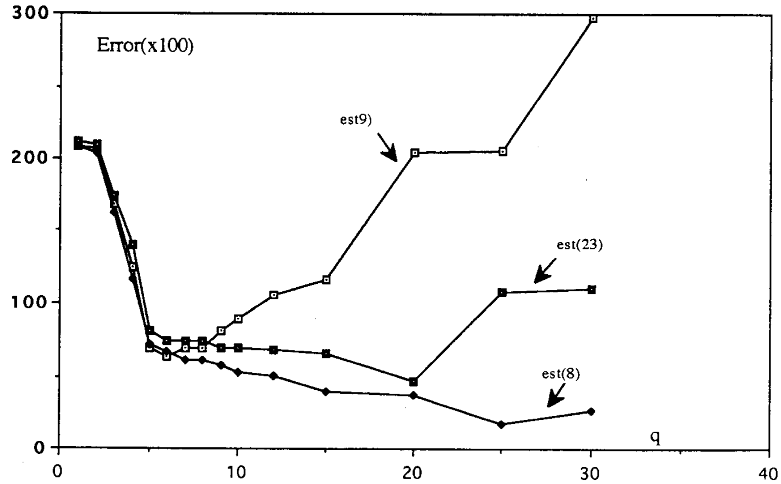

Fig. 8. Parallel dynamical system. Error versus q for estimates (8), (9), (23); $m(u) = m_3(u)$, $n = 128$.

parallel model is about twice as large as the one for the cascade model. To explain this phenomenon, let us write the state equations for both systems in the signal-plus-noise form, i.e., $Y_n = m(U_n) + \zeta_n$, where $\zeta_n = \sum_{j=-\infty}^{n-1} a^{n-j} m(U_j) + Z_n$ for the cascade model and $\zeta_n = \sum_{j=-\infty}^{n-1} a^{n-j} U_j + Z_n$ for the parallel one. From this we can easily calculate the variance of ζ_n for both systems. This is equal to $\sigma^2 + \operatorname{var}(m(U))(1 - a^2)^{-1}$ and $\sigma^2 + \operatorname{var}(U)(1 - a^2)^{-1}$, respectively, where $\sigma^2 = \operatorname{var}(Z)$. In our particular case, i.e., $m(u) = m_3(u)$, a simple algebra yields $\operatorname{var}(U) > \operatorname{var}(m_3(U))$. Hence, the parallel model has a smaller signal-to-noise ratio than the cascade one. Clearly, the inverse property can occur as well. In fact, if $m(u)$ is odd and satisfies $c_1 u \le m(u) \le c_2 u$ for some positive $c_1, c_2, c_2 \ge c_1$, then $\operatorname{var}(U) < \operatorname{var}(m(U))$ if only $c_1 > 1$. All the aforementioned considerations reveal the importance of the selection of the sequence $q(n)$. One has to specify the $q(n)$ before our estimators can be applied. The prescription for $q(n)$ given in Theorem 3 is optimal, but only in the asymptotic sense, and it depends on some unknown characteristics of the system. Thus, the natural question arises how to select an optimal (or nearly optimal) $q(n)$ directly from the data. This problem can be carried out by some form of cross-validation techniques. For example, the so-called future prediction error (see [10], [19], [21], [22], [26]) $CV(q) = (1/n)\sum_{j=1}^{n} |Y_j - \hat{\mu}_{-j}(U_j)|^2$ can be a good candidate for the choice of $q(n)$. Here, $\hat{\mu}_{-j}(u)$ is the version of $\hat{\mu}(u)$ calculated from all of the data points except the jth. This technique requires a considerable amount of computing as the estimate has to be formed n times in the computation of $CV(q)$. A simpler and equally efficient technique aims at minimizing the penalized version of the residual error (this is often called the generalized cross-validation technique):

$$G(q) = \frac{1}{n} \sum_{j=1}^{n} |Y_j - \hat{\mu}(U_j)|^2 \Omega(q)$$

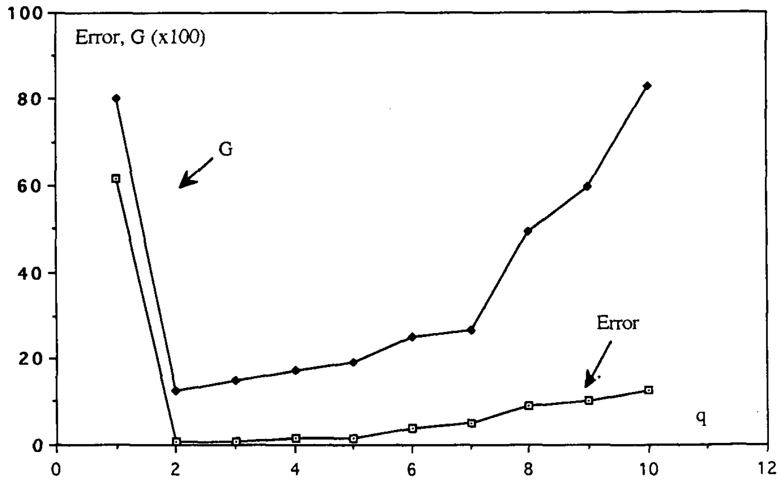

Fig. 9. Error and G versus q for $\tilde{\mu}(u)$; $m(u) = m_1(u)$, $n = 128$.

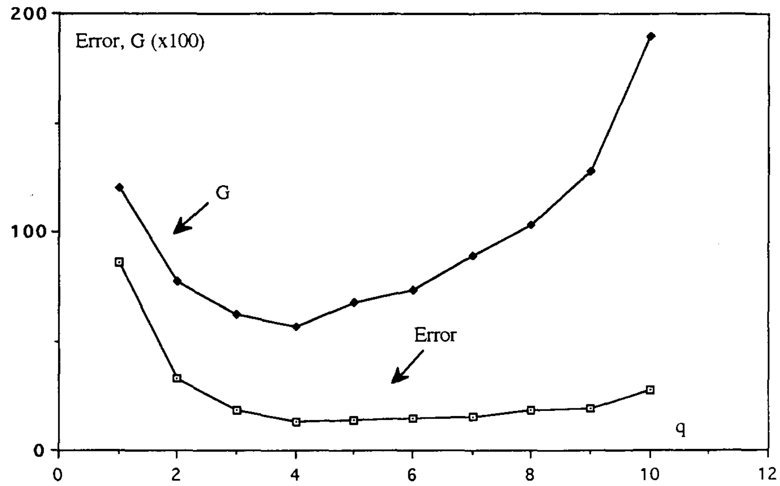

Fig. 10. Error and G versus q for $\tilde{\mu}(u)$; $m(u) = m_3(u)$, $a = -0.2$, $n = 128$.

where $\Omega(q)$ is a suitable increasing function of q. One possible prescription for $\Omega(q)$ is $(1 - 2\pi(q + 1/2)/n)^{-2}$, see [10], [19], [21], [22], [26] for some other choices.

Fig. 9 illustrates the above approach for the memoryless system using the estimate in (9). The error (q) and $G(q)$ measures are plotted with $m(u) = m_1(u)$ and $n = 128$. It is seen the the minimum of $G(q)$ agrees with the minimum of error (q). Note, however, that $G(q)$ is much larger than error (q).

Fig. 10 shows a similar result in the case of the cascade dynamical system. The estimate (9) was used with $m(u) = m_3(u)$, $a = -0.2$, $n = 128$.

The problem of optimality of q selected with respect to $CV(q)$ or $G(q)$ is left for further studies; see [22] for some related results in the context of kernel estimates.

IX. Final Remarks

In this paper, we have proposed two nonparametric algorithms to recover a nonlinear characteristic of nonlinear dynamic systems. Since the characteristic has been expressed as a regression function, the algorithms are, in fact, nonparametric estimates of a regression function observed in the presence of additive disturbance. The disturbance is a sum of white and correlated noise; the source of the latter is the dynamic subsystem. The algorithms differ from each other by the way the Fourier coefficients are estimated. Clearly, $\tilde{c}_k$ is easier to calculate than $\hat{c}_k$. On the other hand, the latter is a more accurate approximation of c_k. Moreover, conditions imposed on $\{q(n)\}$ are more restrictive for the estimate using $\tilde{c}_k$. Nevertheless, both algorithms reach the same convergence rate.

We emphasize that the algorithms presented in the paper cannot be directly extended to multiinput systems since the notion of spacings would have to be generalized to vector spaces. It is, however, known that even a multidimensional space of observations can be partitioned into subsets whose properties are the same as those of uniform spacings presented in Appendix B. This property can be used to redefine our algorithms.

Appendix A

Proof of Lemma 2: In the proof, for a matrix $P = [p_{ij}]$, P^+ denotes the matrix $|p_{ij}|$. From the asymptotic stability of A, it follows that there exist matrices P and Q such that

$$\sup_{n\in\{0,1,\cdots\}} (A^n)^+ = P \quad \text{and} \quad \sup_{n\in\{0,1,\cdots\}} \sum_{k=0}^{n} (A^k)^+ = Q$$

Taking into account that

$$\xi_q = \sum_{j=-\infty}^{q-1} c^T A^{q-j-1} b\lambda_j \tag{A.1}$$

where $\lambda_j = m(U_j) - Em(U_j)$, and that $|\lambda_j| \le 2M$, where $M = \sup|m(u)|$, we have

$$\sup_j |\xi_{[j]}| = \sup_j |\xi_j| \le 2M(c^T)^+ Qb^+$$

and the first part of the lemma follows.

Verification of the second part is much more difficult. Let us first observe that

$$\xi_q = \sum_{j=-\infty}^{0} c^T A^{q-j-1} b\lambda_j + \sum_{j=1}^{q-1} c^T A^{q-j-1} b\lambda_j$$

and

$$\xi_p = \sum_{j=-\infty}^{0} c^T A^{p-j-1} b\lambda_j + \sum_{j=1}^{p-1} c^T A^{p-j-1} b\lambda_j.$$

This yields

$$\begin{aligned} E\{\xi_p\xi_q|U_p,U_q,U_r,U_s\} &= \operatorname{var}(m(U_1)) \sum_{i=0}^{\infty} c^T A^{p+1-1} bc^T A^{q+i-1} b \\ &\quad + E\left\{ \sum_{i=1}^{p-1} c^T A^{p-i-1} b\lambda_i \right. \\ &\quad \left. \cdot \sum_{j=1}^{q-1} c^T A^{q-j-1} b\lambda_j | U_p,U_q,U_r,U_s \right\} \end{aligned} \tag{A.2}$$

for any $p,q,r,s \ge 1$. It is evident that the absolute value of the first term on the right-hand side of (A.2) does not exceed

$$\operatorname{var}(m(U_1))(c^T)^+ (A^{p-1})^+ Q(b)^+ (c^T)^+ Pb. \tag{A.3}$$

Concerning the second term on the right-hand side of (A.2), let us assume, without loss of generality, that $p > q+1$. Then, this term can be decomposed as follows:

$$E\{V_1|U_p,U_q,U_r,U_s\} + E\{V_2|U_p,U_q,U_r,U_s\}$$

where

$$V_1 = c^T A^{p-q-1} \sum_{i=1}^{q} A^{q-i} b\lambda_i \sum_{j=1}^{q-1} c^T A^{q-j-1} b\lambda_j$$

and

$$V_2 = \sum_{i=q+1}^{p-1} c^T A^{p-i-1} b\lambda_i \sum_{j=1}^{q-1} c^T A^{q-j-1} b\lambda_j.$$

Obviously,

$$\begin{aligned} &|E\{V_1|U_p,U_q,U_r,U_s\}| \\ &\quad \le (2M)^2 (c^T)^+ (A^{|p-q|-1})^+ Qb^+ (c^T)^+ Qb^+ \\ &\quad = c_1 (c^T)^+ (A^{|p-q|-1})^+ Qb^+, \end{aligned} \tag{A.4}$$

some c_1 dependent only on A, b, and c, any p, any q. Since $E\lambda_i = 0$, for p,q,r,s all different, we find

$$\begin{aligned} &E\{V_2|U_p,U_q,U_r,U_s\} \\ &= \begin{cases} c^T A^{p-r-1} bc^T A^{q-s-1} b\lambda_r\lambda_s, & \text{if } s < q < r < p \\ c^T A^{p-s-1} bc^T A^{q-r-1} b\lambda_r\lambda_s, & \text{if } r < q < s < p \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Since the cases when $p = q+1$ and $q > p+1$ can be treated in the same way, therefore, by (A.3), (A.4), for any different $p,q,r,s > 1$, we have

$$|E\{\xi_p\xi_q|U_p,U_q,U_r,U_s\}| \le \gamma(p,q,r,s), \tag{A.5}$$

some $\gamma(p,q,r,s)$ where

$$\sum_{p,q,r,s=1}^{n} \gamma(p,q,r,s) \le c_2 n^3, \tag{A.6}$$

p,q,r,s all different, some c_2.

Let us now fix i and j, and for p,q,r,s all different, define the following event:

$$\begin{aligned} A_{pqrs} = \{&U_r \text{ is the } (i-1) \text{ smallest, } U_q \text{ is the } i\text{th smallest,} \\ &U_s \text{ is the } (j-1)\text{th smallest, } U_p \text{ is the } j\text{th smallest}\}, \end{aligned}$$

i.e., U_r, U_q, U_s, U_p have ranks $(i-1), i, (j-1), j$, respectively. Owing to (A.5), we have

$$\begin{aligned} &|E\{\xi_{[i]}\xi_{[j]}(U_{(i)} - U_{(i-1)}) \\ &\quad \cdot (U_{(j)} - U_{(j-1)})|A_{pqrs}, U_{(i-1)}, U_{(i)}, U_{(j-1)}, U_{(j)}\}| \\ &\quad \le \gamma(p,q,r,s)(U_{(i)} - U_{(i-1)})(U_{(j)} - U_{(j-1)}). \end{aligned}$$

It is obvious that

$$\begin{aligned}&E\{\xi_{[i]}\xi_{[j]}(U_{(i)} - U_{(i-1)})\\ &\quad\cdot(U_{(j)} - U_{(j-1)})|U_{(i-1)}, U_{(i)}, U_{(j-1)}, U_{(j)}\}\\ &= \sum_{p,q,r,s=1}^{n} E\{\xi_{[i]}\xi_{[j]}(U_{(i)} - U_{(i-1)})\\ &\quad\cdot(U_{(j)} - U_{(j-1)})|A_{pqrs}, U_{(i-1)}, U_{(i)}, U_{(j-1)}, U_{(j)}\}\\ &\quad\cdot P(A_{pqrs}).\end{aligned}$$

By this, the fact that $P(A_{pqrs}) = (n(n-1)(n-2)(n-3))^{-1}$, and (A.6), we get

$$\begin{aligned}&|E\{\xi_{[i]}\xi_{[j]}(U_{(i)} - U_{(i-1)})\\ &\quad\cdot(U_{(j)} - U_{(j-1)})|U_{(i-1)}, U_{(i)}, U_{(j-1)}, U_{(j)}\}|\\ &\quad\le c_3 n^{-1}(U_{(i)} - U_{(i-1)})(U_{(j)} - U_{(j-1)})\end{aligned}$$

for some c_3 independent on n, i, j.

This completes the proof of Lemma 2. □

Appendix B

B.1 Uniform Spacings

Suppose that $X_1, X_2, \cdots, X_n$ are independent random variables distributed uniformly in [0, 1]. Let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ be a sequence obtained by arranging X_i's in increasing order. It means that $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. Ties, i.e., events that $X_{(i)} = X_{(j)}$ for $i \neq j$, have zero probability. Define, moreover, $X_{(0)} = 0$ and $X_{(n+1)} = 1$, and denote $d_i = X_{(i)} - X_{(i-1)}$, $i = 1, 2, \cdots, n$. Obviously, $d_1 + d_2 + \cdots + d_{n+1} = 1$. In this way, interval [0, 1] has been split randomly into $n+1$ subintervals called spacings.

Clearly (see, e.g., David [8], Pyke [32], or Wilks [40]),

$$f_i(x) = \begin{cases} n(1-x)^{n-1}, & \text{for } 0 \le x \le 1\\ 0, & \text{otherwise}\end{cases}$$

where f_i is the density of d_i. Thus, d_i has a beta distribution with parameters 1 and n.

From this, one easily obtains

$$E d_i^p = p\Gamma(p)\frac{n!}{\Gamma(n+p+1)}, \tag{B.1}$$

any i, any real $p > 0$, and in particular, $E d_i = 1/(n+1)$, $E d_i^2 = 2/(n+1)(n+2)$, and so on. Here, $\Gamma(x)$ is the gamma function. In the proof of Theorems 1 and 2, we need a bound for $E d_i^p$, any real $p > 0$ and any $n \ge 1$. Using (B.1), it is clear that if p is an integer, then

$$E d_i^p \le p\Gamma(p) n^{-p}$$

for any i and any $n \ge 1$.

Now let p be a real number. Then p can be represented as $p = s + \gamma$, where s is an integer and $0 < \gamma < 1$. By invoking (B.1), we can get

$$\begin{aligned}E d_i^p &= p\Gamma(p)\frac{n}{(n+s+\gamma)\cdots(n+\gamma)}\frac{\Gamma(n)}{\Gamma(n+\gamma)}\\ &\le p\Gamma(p)n^{-s}\frac{\Gamma(n)}{\Gamma(n+\gamma)}.\end{aligned} \tag{B.2}$$

It is known that for $0 < \gamma < 1$, $\Gamma(n+\gamma) \ge n^{\gamma}\Gamma(n) - e^{-n}n^{\gamma+n-1}$, and as a result, we have

$$\frac{\Gamma(n)}{\Gamma(n+\gamma)} \le n^{-\gamma}(1 - e^{-n}n^{n-1}/\Gamma(n))^{-1}.$$

Stirling's formula [18, p. 47] yields

$$e^{-n}n^{n-1}/\Gamma(n) \le (2\pi)^{-1/2}n^{-1/2}.$$

This gives

$$\frac{\Gamma(n)}{\Gamma(n+\gamma)} \le \frac{\sqrt{2\pi}}{\sqrt{2\pi}-1}n^{-\gamma},$$

and due to (B.2), we finally obtain

$$E d_i^p \le p\Gamma(p)\frac{\sqrt{2\pi}}{\sqrt{2\pi}-1}n^{-p} \le 1.67p\Gamma(p)n^{-p} \tag{B.3}$$

for any i, any $p > 0$, and any $n \ge 1$. It is worth noting that by the Euler–Gauss Formula

$$\Gamma(p) = \lim_{n\to\infty} n^p \frac{n!}{p(p+1)\cdots(p+n)}$$

and (B.1), we can get the following asymptotic expression $E d_i^p \sim p\Gamma(p)n^{-p}$ as $n \to \infty$, i.e., $n^p E d_i^p \to p\Gamma(p)$ as $n \to \infty$. Recalling that the joint density of d_i and d_j, see [32, p. 398], is given by

$$f_{i,j}(x,y) = \begin{cases} n(n-1)(1-x-y)^{n-2} & \text{for } x > 0,\ y > 0,\ x + y \le 1\\ 0, & \text{otherwise}\end{cases}$$

we can get $E d_i d_j = 1/(n+1)(n+2)$, any i any j such that $i \neq j$. Furthermore, $E\{d_i^2 d_j^2\} = 4/(n+1)(n+2)(n+3)(n+4)$ for $i \neq j$.

B.2 General Spacings

In this section, $U_1, U_2, \cdots, U_n$ are input random variables, i.e., independent random variables defined on $[-\pi, \pi]$ and having density f. Denote

$$b_i = \int_{U_{(i-1)}}^{U_{(i)}} f(x)\,dx,$$

$i = 1, 2, \cdots, n+1$. Clearly, $b_1, b_2, \cdots, b_{n+1}$ are random and add up to 1. The joint distribution of $\{b_1, b_2, \cdots, b_{n+1}\}$ is the same, see [40], as that of $\{d_1, d_2, \cdots, d_{n+1}\}$ discussed in the previous section for uniform spacings, and therefore

$$E b_i^p \le \tau_p n^{-p} \tag{B.4}$$

where $\tau_p = 1.67p\Gamma(p)$ for any i any $p > 0$, and any $n \ge 1$.

Furthermore,

$$E b_i b_j = 1/(n+1)(n+2), \tag{B.5}$$

any i, j, such that $i \neq j$.

Acknowledgment

The authors thank the reviewers for their useful comments.

References

[1] D. W. K. Andrews, "Non-strong mixing autoregressive processes," *J. Appl. Prob.*, vol. 21, pp. 930–934, 1984.
[2] ——, "A nearly independent, but non-strong mixing, triangular array," *J. Appl. Prob.*, vol. 22, pp. 729–731, 1985.
[3] N.K. Barry, *A Treatise on Trigonometric Series*. Oxford: Pergamon, 1964.

[4] J. S. Bendat, *Nonlinear System Analysis and Identification.* New York: Wiley, 1990.

[5] S. A. Billings, "Identification of nonlinear systems—A survey," *Proc. IEE*, vol. 127, pp. 272–285, 1980.

[6] D. R. Brillinger, "The identification of a particular nonlinear time series system," *Biometrika*, vol. 64, pp. 509–515, 1977.

[7] C. K. Chu and J. S. Marron, "Choosing a kernel regression estimator," *Statist. Sci.*, vol. 6, pp. 404–436, 1991.

[8] H. A. David, *Order Statistics*, 2nd ed. New York: Wiley, 1981.

[9] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1 View.* New York: Wiley, 1985.

[10] R. L. Eubank, *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker, 1988.

[11] A. Georgiev,"Nonparametric kernel algorithm for recovering of functions from noisy measurement with applications," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 782–784, 1985.

[12] W. Greblicki, "Non-parametric orthogonal series identification of Hammerstein systems," *Int. J. Syst. Sci.*, vol. 20, pp. 2355–2367, 1989.

[13] W. Greblicki and M. Pawlak, "Fourier and Hermite series estimates of regression functions," *Ann. Inst. Statist. Math.*, vol. 37, pp. 443–459, 1985.

[14] ——, "Nonparametric identification of Hammerstein systems," *IEEE Trans. Inform. Theory*, vol. 35, pp. 409–418, 1989.

[15] ——, "Nonparametric identification of a cascade nonlinear time series system," *Signal Processing*, vol. 22, pp. 61–75, 1991.

[16] ——, "Nonparametric identification of a particular nonlinear time series system," *IEEE Trans. Signal Processing*, vol. 40, pp. 985–989, 1992.

[17] W. Greblicki, "Nonparametric identification of Wiener systems," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1487–1493, 1992.

[18] D. H. Greene and D. E. Knuth, *Mathematics for the Analysis of Algorithms*, 3rd ed. Boston: Birkhäuser, 1990.

[19] L. Györfi, W. Härdle, P. Sarda, and P. Vieu, *Nonparametric Curve Estimation from Time Series.* Berlin: Springer-Verlag, 1989.

[20] P. Hall and T. E. Wehrly, "A geometric method for removing edge effects from kernel-type nonparametric regression estimators," *J. Amer. Statist. Ass.*, vol. 86, pp. 665–672, Sept. 1991.

[21] W. Härdle, "A law of the iterated logarithm for nonparametric regression function estimators," *Ann. Statist.*, vol. 12, pp. 624–635, 1984.

[22] ——, *Applied Nonparametric Regression.* Cambridge: Cambridge Univ. Press, 1990.

[23] W. Härdle and P. Vieu, "Kernel regression smoothing of time series," *J. Time Series*, vol. 13, pp. 209–232, 1992.

[24] R. Kannurpatti and G. W. Hart, "Memoryless nonlinear system identification with unknown model order," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1440–1450, 1991.

[25] A. Krzyzak, "Identification of discrete Hammerstein systems by the Fourier series regression estimate," *Int. J. Syst. Sci.*, vol. 20, pp. 1729–1744, 1989.

[26] ——, "On estimation of a class of nonlinear systems by the kernel regression estimate," *IEEE Trans. Inform. Theory*, vol. 36, pp. 141–152, 1990.

[27] K. C. Li "Asymptotic optimality of C_p, C_L, cross-validation and generalized cross-validation: Discrete index set," *Ann. Statis.*, vol. 15, pp. 958–975, 1987.

[28] G. G. Lorentz, "Fourier-Koeffizienten und Funktionenklassen," *Math. Z.*, vol. 51, pp. 135–149, 1948.

[29] Y. P. Mack and H. G. Müller, "Convolution type estimators for nonparametric regression," *Statis. Prob. Lett.*, vol. 7, pp. 229–239, 1989.

[30] H. G. Müller, *Nonparametric Analysis of Longitudinal Data* (*Lecture Notes in Statistics*, Vol. 46). New York: Springer, 1988.

[31] M. Pawlak, "On the series expansion approach to the identification of Hammerstein system," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 763–767, 1991.

[32] R. Pyke, "Spacings," *J. Roy. Statist. Soc. B*, vol. 27, pp. 395–436, 1965.

[33] G. Roussas, Ed., *Nonparametric Functional Estimation and Related Topics* (NATO Advanced Study Institute). Dordrecht: Kluwer Academic, 1991.

[34] E. Rafajlowicz, "Nonparametric orthogonal series estimators of regression: A class attaining the optimal convergence rate in L_2," *Statist. Prob. Lett.*, vol. 5, pp. 219–224, 1987.

[35] B. L. S. Rao, *Nonparametric Functional Estimation.* Orlando: Academic, 1983.

[36] L. Rutkowski, "Identification of MISO nonlinear regression in the presence of a wide class of disturbances," *IEEE Trans. Inform. Theory*, vol. 37, pp. 214–216, 1991.

[37] C. J. Stone, "Consistent nonparametric regression," *Ann. Statist.*, vol. 5, pp. 595–620, 1977.

[38] ——, "Optimal global rates of convergence for nonparametric regression," *Ann. Statist.*, vol. 10, pp. 1040–1053, 1982.

[39] G. Wahba, *Spline Models for Observational Data.* Philadelphia: SIAM, 1990.

[40] S. S. Wilks, *Mathematical Statistics.* New York: Wiley, 1962.

[41] A. Zygmund, *Trigonometric Series.* Cambridge: Cambridge Univ., 1959.