

Nonparametric identification of nonlinearities in block-oriented systems by orthogonal wavelets with compact support

(revised version)

Zygmunt Hasiewicz, Mirosław Pawlak and Przemysław Śliwiński

Abstract

The paper addresses the problem of identification of nonlinear characteristics in a certain class of discrete-time block-oriented systems. The systems are driven by random stationary white processes (i.i.d. input sequences) and disturbed by stationary, white or coloured, random noise. The prior information about nonlinear characteristics is nonparametric. To construct identification algorithms the orthogonal wavelets of compact support are applied, and a class of wavelet-based models is introduced and examined. It is shown that under moderate assumptions the proposed models converge almost everywhere (in probability) to the identified nonlinear characteristics, irrespective of the noise model. The rule for optimum model size selection is given and the asymptotic rate of convergence of the model error is established. It is demonstrated that in some circumstances the wavelet models are, in particular, superior to classical trigonometric and Hermite orthogonal series models worked out earlier.

Index Terms

block-oriented systems, nonlinearity recovering, nonparametric approach, wavelet-based models, convergence analysis.

I. INTRODUCTION

We consider the problem of identification of nonlinear characteristics in a class of discrete-time block-oriented systems, i.e. structured objects where nonlinear static elements are separated from the rest of the system and embedded in a composite structure containing discrete-time linear dynamic blocks and

Zygmunt Hasiewicz and Przemysław Śliwiński are with the Institute of Engineering Cybernetics, Wrocław University of Technology, Wrocław, Poland. Mirosław Pawlak is with the Department of Electrical and Computer Engineering, The University of Manitoba, Winnipeg, Canada.

other 'nuisance' nonlinearities ([1]). It is assumed that a priori information about subsystems is small and in particular the nonlinear characteristics to be identified are not given in a parametric form.

The problem of recovering static characteristics in interconnected composite systems has been extensively studied in the literature. For steady state systems it has been investigated in a number of papers using parametric and nonparametric approach (e.g. [19], [14]). For block-oriented dynamical systems, the problem has been initially examined under the assumption that the nonlinear characteristics are known up to the parameters and typically are polynomials of known degree. Various parametric approaches have been developed, with particular attention to cascade and parallel systems. Thorough overview of the derived parametric identification methods can be found in [18].

Less demanding nonparametric approach, discarding the restrictive assumption of parametric prior knowledge of the characteristics, has been initiated in [16] for Hammerstein systems. The approach, originated from nonparametric estimation of a regression function, was next developed bringing a collection of papers where two types of nonparametric identification algorithms were studied: kernel algorithms (e.g. [16], [17]) and orthogonal series algorithms applying conventional orthogonal series expansions of characteristics (e.g. [13], [27]). The algorithms were elaborated for Hammerstein and Wiener systems, i.e. cascade connections of static nonlinearities and linear dynamic blocks in a suitable order.

In this paper, we propose and examine a class of identification algorithms based on wavelet approximation of functions. The algorithms use for building of the wavelet models orthogonal wavelets with compact supports. Orthogonality of wavelets enables easy and convenient improvement of the models by stepwise expansion (adding new details) and compactness of the supports ensures obtainment of simultaneously accurate and parsimonious representations of nonlinear characteristics, with good localization ability. In the models we apply both father and mother wavelets, and hence fully exploit the multiresolution idea and extend the results of [28] and [20] where to the construction of wavelet models only scaling functions (father wavelets) have been used.

We assume that identified systems operate in stochastic conditions. By assumption they are excited by stationary white random processes (i.i.d. random input sequences) for which there exist probability density functions (unknown in our approach). The external stationary random noises acting on the systems can be white or coloured, with arbitrary correlation structure. Owing to the zooming capability of wavelets, we focus on the analysis of local, pointwise, properties of the wavelet models. We show that under weak assumptions concerning the underlying nonlinearities and input probability density functions, the wavelet-based models provide consistent estimates of nonlinear characteristics which rapidly converge to the true nonlinear functions, regardless of whether the disturbing noise is white or coloured. We

establish that the rates of convergence can approach the best possible nonparametric convergence rates [29] and that the wavelet models may, for example, outperform classical trigonometric and Hermite orthogonal series models worked out earlier. Up to our best knowledge, such an analysis in the context of system identification has not been presented yet. This is perhaps because in spite of numerous successful applications of wavelets in many fields of engineering, for instance signal processing, image analysis, communication systems, as to quote a few (see e.g. [26], [2] and the references cited therein), there are still few papers concerning wavelets employment in system identification. Actually, the collection of contributions in this area does not exceed significantly [24], [30], [4], [11], [21] and the papers mentioned above.

The paper is organized as follows. Section II presents the class of systems under consideration and states the system identification task. Section III gives examples of popular block-oriented systems belonging to the studied class. The general wavelet-based identification algorithm of nonlinear characteristics is constructed in Section IV. In Section V, we consider convergence of the wavelet models and provide conditions ensuring weak pointwise consistency. In Section VI, we discuss the convergence speed. The guidelines for optimum model size selection (optimizing the convergence rate) are given and we propose simple sub-optimal rule for selection of the size of the wavelet model, helpful in the case of insufficient prior knowledge about identified nonlinearity and input density. Results of computer simulations are reported in Section VII, and conclusions in Section VIII complete the paper. For the sake of convenience, pertinent facts concerning wavelet approximation of functions by orthogonal wavelets with compact support are collected in Appendix A. Appendices B, C and D contain technical derivations and lemmas.

II. PROBLEM FORMULATION

We consider a class of discrete-time block-oriented systems which input-output behaviour can be expressed by the equation

$$y_k = R(x_k) + \xi(x_{k-1}, x_{k-2}, \dots) + z_k \quad (1)$$

where $R(x)$ is a static nonlinearity to be identified, $\{(x_k, y_k)\}$ is the (scalar input, scalar output) sequence, $\{\xi_k = \xi(x_{k-1}, x_{k-2}, \dots)\}$ is the 'system noise' induced by the system dynamics, and $\{z_k\}$ is the external disturbance (Fig. 1). We assume that the system in (1) is asymptotically stable and operates in steady state. Moreover, the following conditions are imposed.

- A1.** The nonlinearity $R(x)$ is such that $|R(x)| \leq C_{1R}|x| + C_{2R}$, some $C_{1R}, C_{2R} > 0$.
- A2.** The input process $\{x_k\}_{k \in \mathbf{Z}}$ (\mathbf{Z} - the set of integers) is a sequence of independent and identically distributed (i.i.d.) random variables with finite variance σ_x^2 , and there exists a probability density

function $f(x)$ such that $f(x) \leq M_f < \infty$, some $M_f > 0$. The density $f(x)$ and the variance σ_x^2 are unknown.

A3. The system noise $\{\xi_k\}_{k \in \mathbf{Z}}$ is of the form

$$\xi_k = \xi_{1k} + \xi_{2k} + \dots + \xi_{Jk} \quad (2)$$

where

$$\xi_{jk} = \sum_{i=1}^{\infty} \lambda_{ji} \zeta_j(x_{k-i}) \quad (3)$$

and where for $j = 1, 2, \dots, J$ it holds that $\sum_{i=1}^{\infty} |\lambda_{ji}| < \infty$ and $\zeta_j(x)$ is a nonlinear function such that $E \zeta_j(x_1) = 0$ and $E \zeta_j^2(x_1) = \sigma_{j\zeta}^2 < \infty$. The impulse responses $\{\lambda_{ji}\}$ and functions $\zeta_j(x)$ are not known.

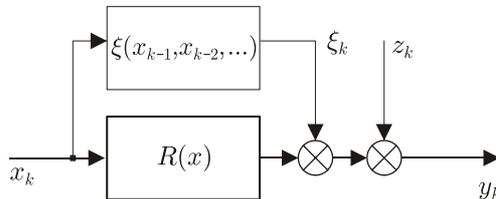


Fig. 1. General system under consideration.

A4. The external noise $\{z_k\}_{k \in \mathbf{Z}}$ is, in general, a correlated process generated from stationary white noise processes $\{\varepsilon_{lk}\}_{k \in \mathbf{Z}}$, $l = 1, 2, \dots, L$, with $E \varepsilon_{l1} = 0$ and $E \varepsilon_{l1}^2 = \sigma_{l\varepsilon}^2 < \infty$, such that

$$z_k = z_{1k} + z_{2k} + \dots + z_{Lk} \quad (4)$$

where

$$z_{lk} = \sum_{i=0}^{\infty} \omega_{li} \varepsilon_{l,k-i} \quad (5)$$

and $\sum_{i=0}^{\infty} |\omega_{li}| < \infty$ for $l = 1, 2, \dots, L$. The impulse responses $\{\omega_{li}\}$ and the noise variances $\sigma_{l\varepsilon}^2$ are unknown.

A5. Processes $\{x_k\}$ and $\{\varepsilon_{lk}\}$, $l = 1, 2, \dots, L$, are mutually statistically independent.

A6. Only the external (input, output) measurement data $\{(x_k, y_k)\}$ are available and $R(x)$ should be identified in the bounded region $[a, b]$.

A7. For $x \in [a, b]$, it holds that

$$0 < \delta \leq f(x) \quad (6)$$

some $\delta > 0$.

The aim is to recover $R(x)$ in the region $[a, b]$ using solely the input-output data $\{(x_k, y_k)\}_{k=1}^N$.

We note that Assumptions A1-A7 are rather weak and represent realistic system identification conditions. First, prior knowledge about $R(x)$ is poor. We merely presume that $|R(x)|$ grows with $|x|$ not faster than linearly (Assumption A1) which yields a broad class of admissible nonlinearities. As such nonlinearities cannot be parameterized, our identification problem is nonparametric. Second, the system input can be any i.i.d. random process possessing bounded and nonvanishing in the identification region probability density function, with finite variance (Assumptions A2 and A7). Third, the external noise can be correlated and can be any type of coloured noise (Assumption A4). Moreover, the disturbances have not to be bounded. As regards Assumption A6, it reflects standard practical situations where (i) the inner signals in the block-oriented systems are not accessible for measurements (see, e.g., [1], [18]) and (ii) the nonlinearity $R(x)$ is required to be known only in some bounded region. We observe for further use that under Assumptions A1-A4 the system output $\{y_k\}$ is a correlated second-order stationary random process of finite variance.

III. EXAMPLES OF SYSTEMS

Although class of systems in (1)-(5) looks rather specific, there is quite a lot of exemplary block-oriented structures which conform to this description and are often considered in the literature. In the following examples $\mu(x)$, $\eta(x)$ and $\nu(x)$ are static nonlinearities, $\{\lambda_i\}$, $\{\gamma_i\}$ and $\{\rho_i\}$ denote impulse responses of linear dynamical elements such that $\sum_{i=0}^{\infty} |\lambda_i| < \infty$, $\sum_{i=0}^{\infty} |\gamma_i| < \infty$, $\sum_{i=0}^{\infty} |\rho_i| < \infty$, and z_k is coloured noise produced by linear noise filter $\{\omega_i\}$ from a stationary white noise process $\{\varepsilon_k\}$ according to the equation $z_k = \sum_{i=0}^{\infty} \omega_i \varepsilon_{k-i}$ where $\sum_{i=0}^{\infty} |\omega_i| < \infty$, $E \varepsilon_1 = 0$, $E \varepsilon_1^2 = \sigma_\varepsilon^2 < \infty$ and $\{\varepsilon_k\}$ is independent of the input $\{x_k\}$. Moreover, we denote $d_{\mu\lambda} = E \mu(x_1) \sum_{i=1}^{\infty} \lambda_i$, $d_{\eta\lambda} = E \eta(x_1) \sum_{i=1}^{\infty} \lambda_i$, $d_{\eta\gamma} = E \eta(x_1) \sum_{i=1}^{\infty} \gamma_i$ and $\mu_0(x) = \mu(x) - E \mu(x_1)$, $\eta_0(x) = \eta(x) - E \eta(x_1)$.

a) *Parallel system* [3]: For parallel system as in Fig. 2a and $\lambda_0 = 0$ (one step delay), we get

$$y_k = \mu(x_k) + d + \sum_{i=1}^{\infty} \lambda_i (x_{k-i} - E x_1) + z_k$$

where $d = E x_1 \sum_{i=1}^{\infty} \lambda_i$ which maps (1)-(5) for $R(x) = \mu(x) + d$ and $J = L = 1$ with $\lambda_{1i} = \lambda_i$, $\omega_{1i} = \omega_i$, $\varepsilon_{1k} = \varepsilon_k$ and $\zeta_1(x) = x - E x_1$, provided that $E x_1^2 < \infty$ (Assumption A2).

b) *Hammerstein system* [13]: For Hammerstein system (Fig. 2b) we have

$$y_k = \lambda_0 \mu(x_k) + d_{\mu\lambda} + \sum_{i=1}^{\infty} \lambda_i \mu_0(x_{k-i}) + z_k$$

This gives (1)-(5) for $R(x) = \lambda_0 \mu(x) + d_{\mu\lambda}$ and $J = L = 1$ with the same specifications as above and $\zeta_1(x) = \mu_0(x)$, provided that $E \mu^2(x_1) < \infty$ (Assumptions A1 and A2).

Hammerstein system with two-segment nonlinearity [32]: For Hammerstein system with two-segment nonlinearity (Fig. 3a) we obtain

$$y_k = \lambda_0 [\mu(x_k) + \eta(x_k)] + (d_{\mu\lambda} + d_{\eta\lambda}) + \sum_{i=1}^{\infty} \lambda_i \mu_0(x_{k-i}) + \sum_{i=1}^{\infty} \lambda_i \eta_0(x_{k-i}) + z_k$$

This falls into the description (1)-(5) for $R(x) = \lambda_0 [\mu(x) + \eta(x)] + d$ where $d = d_{\mu\lambda} + d_{\eta\lambda}$, $J = 2$ with $\lambda_{1i} = \lambda_{2i} = \lambda_i$, $\zeta_1(x) = \mu_0(x)$, $\zeta_2(x) = \eta_0(x)$ and $L = 1$ with the noise filter defined above, provided that $E \mu^2(x_1) < \infty$, $E \eta^2(x_1) < \infty$ (Assumptions A1 and A2).

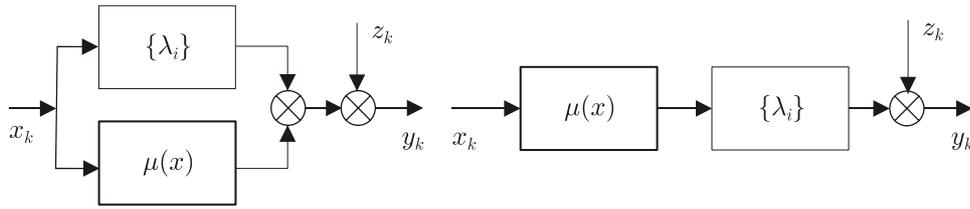


Fig. 2. **a)** Parallel system, **b)** Hammerstein system.

Hammerstein system with nuisance nonlinearity: For Hammerstein system with nuisance nonlinearity $\nu(\cdot)$ (Fig. 3b) it holds that

$$y_k = \lambda_0 \mu(x_k) + \eta(x_k) + d_{\mu\lambda} + \sum_{i=1}^{\infty} \lambda_i \mu_0(x_{k-i}) + z_k$$

where $\eta(x) = \nu(\mu(x))$. This conforms with the description (1)-(5) for $R(x) = \lambda_0 \mu(x) + \eta(x) + d_{\mu\lambda}$ and further specifications as for Hammerstein system.

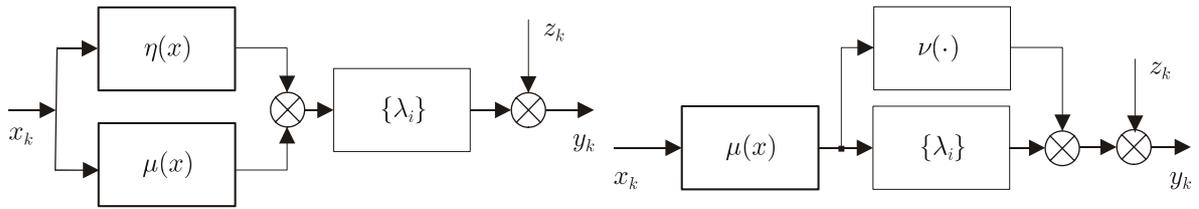


Fig. 3. **a)** Hammerstein system with two-segment nonlinearity, **b)** Hammerstein system with nuisance nonlinearity.

Hammerstein system with nuisance dynamics: For Hammerstein system with nuisance dynamics $\{\rho_i\}$ (Fig. 4a) and $\rho_0 = 0$ we have

$$y_k = \lambda_0 \mu(x_k) + d + \sum_{i=1}^{\infty} \lambda_i \mu_0(x_{k-i}) + \sum_{i=1}^{\infty} \gamma_i (x_{k-i} - E x_1) + z_k$$

where $d = d_{\mu\lambda} + \mathbb{E} x_1 \sum_{i=1}^{\infty} \gamma_i$ and $\gamma_i = \sum_{p=0}^{\infty} \lambda_p \rho_{i-p}$. This fits (1)-(5) for $R(x) = \lambda_0 \mu(x) + d$, $J = 2$ with $\lambda_{1i} = \lambda_i$, $\zeta_1(x) = \mu_0(x)$, $\lambda_{2i} = \gamma_i$, $\zeta_2(x) = x - \mathbb{E} x_1$ (provided that $\mathbb{E} \mu^2(x_1) < \infty$ and $\mathbb{E} x_1^2 < \infty$), and $L = 1$ with the previous noise model.

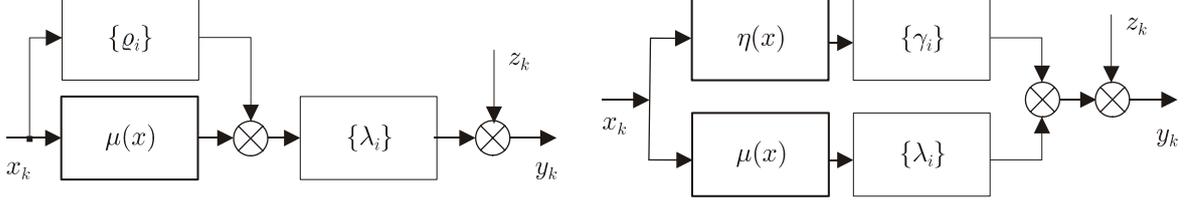


Fig. 4. **a)** Hammerstein system with nuisance dynamics, **b)** Uryson system.

Uryson system [18]: For Uryson system in Fig. 4b we obtain

$$y_k = \lambda_0 \mu(x_k) + \gamma_0 \eta(x_k) + d + \sum_{i=1}^{\infty} \lambda_i \mu_0(x_{k-i}) + \sum_{i=1}^{\infty} \gamma_i \eta_0(x_{k-i}) + z_k$$

where $d = d_{\mu\lambda} + d_{\eta\gamma}$. This fulfils (1)-(5) for $R(x) = \lambda_0 \mu(x) + \gamma_0 \eta(x) + d$, $J = 2$, $L = 1$ and obvious further specifications.

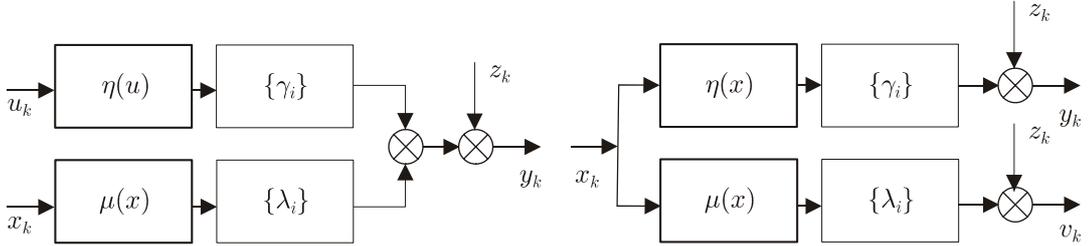


Fig. 5. **a)** Two-channel system, **b)** Bunch system.

Two-channel system [31]: For two-channel system (Fig. 5a) with stationary white inputs $\{x_k\}$ and $\{u_k\}$ (mutually independent and independent of the noise $\{\varepsilon_k\}$) we get

$$y_k = \lambda_0 \mu(x_k) + d + \sum_{i=1}^{\infty} \lambda_i \mu_0(x_{k-i}) + \sum_{i=0}^{\infty} \gamma_i (\eta(u_{k-i}) - \mathbb{E} \eta(u_1)) + z_k$$

where $d = d_{\mu\lambda} + \mathbb{E} \eta(u_1) \sum_{i=0}^{\infty} \gamma_i$. This can be put in the form of (1)-(5) for $R(x) = \lambda_0 \mu(x) + d$, $J = 1$ with $\lambda_{1i} = \lambda_i$, $\zeta_1(x) = \mu_0(x)$ and $L = 2$ with $\omega_{1i} = \gamma_i$, $\varepsilon_{1k} = \eta(u_k) - \mathbb{E} \eta(u_1)$, $\omega_{2i} = \omega_i$, $\varepsilon_{2k} = \varepsilon_k$, provided that $\mathbb{E} \mu^2(x_1) < \infty$ and $\mathbb{E} \eta^2(u_1) < \infty$.

Bunch system: For each branch of the system in Fig. 5b it holds that

$$y_k = \lambda_0 \mu(x_k) + d_{\mu\lambda} + \sum_{i=1}^{\infty} \lambda_i \mu_0(x_{k-i}) + z_k$$

$$v_k = \gamma_0 \eta(x_k) + d_{\eta\gamma} + \sum_{i=1}^{\infty} \gamma_i \eta_0(x_{k-i}) + z_k$$

This clearly maps the description (1)-(5) (see Hammerstein system). The last three examples can be easily generalized.

The above systems can be met in diverse areas as for example nonlinear control, telecommunication, acoustics, signal processing, chemical engineering, or biomedical engineering. A comprehensive bibliography concerning practical applications of such and other structures can be found in [12]. We refer also to [18, Chapter 7] and the references cited therein for some specific case studies.

The presented examples reveal that the nonlinearity $R(x)$ in equation (1) can generally differ from any true nonlinear characteristic existing in a system. However, as it follows from (1), only $R(x)$ of such type as in examples can be tried to be identified from just external input-output observations $\{(x_k, y_k)\}$ unless some additional structural conditions are fulfilled. For instance, for Hammerstein system we gain $R(x) = \mu(x)$ if $\lambda_0 = 1$ and $E\mu(x_1) = 0$ where the latter is fulfilled if, e.g., the system nonlinearity $\mu(x)$ is an odd and the input density $f(x)$ is an even function. This physical limitation, following from composite structure of block-oriented systems, is well realized in the system identification literature, cf. [1], [13], [18].

IV. IDENTIFICATION ALGORITHM

We shall derive a scheme for identification of the nonlinearity $R(x)$ in equation (1) from the measurement data $\{(x_k, y_k)\}$, using orthogonal wavelets with compact support presented in Appendix A. To this end we exploit the following.

1. For each point x such that the input probability density $f(x) > 0$, the nonlinear characteristic $R(x)$ can be expressed as the fraction

$$R(x) = g(x)/f(x) \tag{7}$$

where $g(x) = R(x) f(x)$ (compare e.g. [13]).

2. Under Assumptions A1 and A2, $g(x), f(x) \in L^2(\mathbf{R})$:

$$\int_{-\infty}^{+\infty} g^2(x) dx \leq M_f \int_{-\infty}^{+\infty} R^2(x) f(x) dx \leq 2M_f C_R^2 (\sigma_x^2 + m_x^2 + 1) < \infty$$

$$\int_{-\infty}^{+\infty} f^2(x) dx \leq M_f \int_{-\infty}^{+\infty} f(x) dx = M_f < \infty$$

where $C_R = \max\{C_{1R}, C_{2R}\}$, $m_x = E x_1$, $\sigma_x^2 = \text{var}x_1$.

3. The functions $g(x)$ and $f(x)$ can be approximated by the wavelet-based models (cf. (50) in Appendix A):

$$g(x; K) = \sum_{n=n_{\min}(\varphi, x, M)}^{n_{\max}(\varphi, x, M)} \alpha_{Mn}^g \varphi_{Mn}(x) + \sum_{m=M}^{K-1} \sum_{n=n_{\min}(\psi, x, m)}^{n_{\max}(\psi, x, m)} \beta_{mn}^g \psi_{mn}(x) \quad (8)$$

and

$$f(x; K) = \sum_{n=n_{\min}(\varphi, x, M)}^{n_{\max}(\varphi, x, M)} \alpha_{Mn}^f \varphi_{Mn}(x) + \sum_{m=M}^{K-1} \sum_{n=n_{\min}(\psi, x, m)}^{n_{\max}(\psi, x, m)} \beta_{mn}^f \psi_{mn}(x) \quad (9)$$

where the summation limits $n_{\min}(\varphi, x, M)$, $n_{\max}(\varphi, x, M)$, $n_{\min}(\psi, x, m)$, $n_{\max}(\psi, x, m)$ are as in (52) and the wavelet coefficients α_{Mn}^g , β_{mn}^g , α_{Mn}^f , β_{mn}^f are calculated according to (51) in Appendix A.

4. As $f(x)$ is a probability density function, the α_{Mn}^g 's, β_{mn}^g 's, α_{Mn}^f 's and β_{mn}^f 's in (8) and (9) are the following expectations (see (45) and (51) in Appendix A):

$$\alpha_{Mn}^g = \int_{-\infty}^{+\infty} R(x) \varphi_{Mn}(x) f(x) dx = E [R(x_1) \varphi_{Mn}(x_1)] \quad (10)$$

$$\beta_{mn}^g = \int_{-\infty}^{+\infty} R(x) \psi_{mn}(x) f(x) dx = E [R(x_1) \psi_{mn}(x_1)] \quad (11)$$

$$\alpha_{Mn}^f = \int_{-\infty}^{+\infty} \varphi_{Mn}(x) f(x) dx = E [\varphi_{Mn}(x_1)] \quad (12)$$

$$\beta_{mn}^f = \int_{-\infty}^{+\infty} \psi_{mn}(x) f(x) dx = E [\psi_{mn}(x_1)] \quad (13)$$

Owing to (1) and Assumptions A2-A5 it moreover holds that

$$E [R(x_1) \varphi_{Mn}(x_1)] = E [y_1 \varphi_{Mn}(x_1)] \quad (14)$$

$$E [R(x_1) \psi_{mn}(x_1)] = E [y_1 \psi_{mn}(x_1)] \quad (15)$$

5. The approximators $g(x; K)$ and $f(x; K)$ when substituted for the exact $g(x)$ and $f(x)$ in the ratio (7) yield a wavelet-based model of $R(x)$:

$$R(x; K) = g(x; K) / f(x; K) \quad (16)$$

for each point x where $f(x; K) \neq 0$.

Based on the above observations, we arrive at the following three-step algorithm to identify the nonlinearity $R(x)$ from empirical (input,output) data $\{(x_k, y_k)\}_{k=1}^N$.

Step 1: Compute the sample means (see (10)-(15))

$$\hat{\alpha}_{Mn}^g = N^{-1} \sum_{k=1}^N y_k \varphi_{Mn}(x_k) ; \quad \hat{\beta}_{mn}^g = N^{-1} \sum_{k=1}^N y_k \psi_{mn}(x_k) \quad (17)$$

and

$$\hat{\alpha}_{Mn}^f = N^{-1} \sum_{k=1}^N \varphi_{Mn}(x_k) ; \quad \hat{\beta}_{mn}^f = N^{-1} \sum_{k=1}^N \psi_{mn}(x_k) \quad (18)$$

Step 2: Plug in $\hat{\alpha}_{Mn}^g$, $\hat{\beta}_{mn}^g$, $\hat{\alpha}_{Mn}^f$, $\hat{\beta}_{mn}^f$ into the approximators in (8) and (9) obtaining the empirical wavelet-based models of $g(x)$ and $f(x)$:

$$\hat{g}(x; K) = \sum_{n=n_{\min}(\varphi, x, M)}^{n_{\max}(\varphi, x, M)} \hat{\alpha}_{Mn}^g \varphi_{Mn}(x) + \sum_{m=M}^{K-1} \sum_{n=n_{\min}(\psi, x, m)}^{n_{\max}(\psi, x, m)} \hat{\beta}_{mn}^g \psi_{mn}(x) \quad (19)$$

and

$$\hat{f}(x; K) = \sum_{n=n_{\min}(\varphi, x, M)}^{n_{\max}(\varphi, x, M)} \hat{\alpha}_{Mn}^f \varphi_{Mn}(x) + \sum_{m=M}^{K-1} \sum_{n=n_{\min}(\psi, x, m)}^{n_{\max}(\psi, x, m)} \hat{\beta}_{mn}^f \psi_{mn}(x) \quad (20)$$

Step 3: Put the models $\hat{g}(x; K)$ and $\hat{f}(x; K)$ in (16) getting the empirical wavelet-based model of $R(x)$:

$$\hat{R}(x; K) = \hat{g}(x; K) / \hat{f}(x; K) \quad (21)$$

for each point x where $\hat{f}(x; K) \neq 0$.

Remark 1: Because of the fractional form of the model $\hat{R}(x; K)$ and compactness of the support of wavelet functions, the coefficients $\hat{\alpha}_{Mn}^g$, $\hat{\beta}_{mn}^g$, $\hat{\alpha}_{Mn}^f$, $\hat{\beta}_{mn}^f$ can be calculated using simplified rules (cf. (46) and (48)):

$$\hat{\alpha}_{Mn}^g = 2^{M/2} \sum_{\{k: u_{Mn,k} \in [s_1, s_2]\}} y_k \varphi(u_{Mn,k}) \quad \text{and} \quad \hat{\beta}_{mn}^g = 2^{m/2} \sum_{\{k: u_{mn,k} \in [t_1, t_2]\}} y_k \psi(u_{mn,k}) \quad (22)$$

and

$$\hat{\alpha}_{Mn}^f = 2^{M/2} \sum_{\{k: u_{Mn,k} \in [s_1, s_2]\}} \varphi(u_{Mn,k}) \quad \text{and} \quad \hat{\beta}_{mn}^f = 2^{m/2} \sum_{\{k: u_{mn,k} \in [t_1, t_2]\}} \psi(u_{mn,k}) \quad (23)$$

and freely scaled, if needed, where $u_{mn,k} = 2^m x_k - n$ and $[s_1, s_2]$ and $[t_1, t_2]$ are, respectively, supports of the wavelet functions $\varphi(x)$ and $\psi(x)$.

Remark 2: For $R(x)$ to be estimated over the interval $[a, b]$ the total number of needed wavelet coefficients $\hat{\alpha}_{Mn}^g$, $\hat{\beta}_{mn}^g$, $\hat{\alpha}_{Mn}^f$, $\hat{\beta}_{mn}^f$ does not pass (see (19), (20) and (52) in Appendix A)

$$\bar{L}(b-a, M, K) = 2 [S(K-M+1) + 2^K [b-a]] \quad (24)$$

where $S = \lfloor s \rfloor + 1$ and s is the support size of $\varphi(x)$ and $\psi(x)$ (see (53) in Appendix A).

We remark that computations in the algorithm are the same for white and coloured noise $\{z_k\}$ (see Assumption A4 in Section II). This stands in contrast to parametric methods where correlation of disturbances acting on the system usually results in substantial modification of identification routines yielding far more demanding computation procedures than for the white noise case; see, e.g., [25], [18] for appropriate examples concerning parametric methods.

V. CONVERGENCE ANALYSIS

We shall show that the models $\hat{g}(x; K)$, $\hat{f}(x; K)$ and $\hat{R}(x; K)$ converge pointwise, in a probabilistic sense, to the true functions $g(x)$, $f(x)$ and $R(x)$ as the number N of data points (x_k, y_k) grows large and the scale factor K is appropriately fitted to the number of data. To prove consistency we examine in turn 1) convergence of the empirical wavelet coefficients $\hat{\alpha}_{Mn}^g, \hat{\beta}_{mn}^g, \hat{\alpha}_{Mn}^f, \hat{\beta}_{mn}^f$ to the true coefficients $\alpha_{Mn}^g, \beta_{mn}^g, \alpha_{Mn}^f, \beta_{mn}^f$, and 2) conditions on the scale factor K guaranteeing consistency of the models $\hat{g}(x; K)$, $\hat{f}(x; K)$ and $\hat{R}(x; K)$.

A. Convergence of the empirical coefficients

Let $(\hat{c}, c) \in \left\{ (\hat{\alpha}_{Mn}^g, \alpha_{Mn}^g), (\hat{\beta}_{mn}^g, \beta_{mn}^g), (\hat{\alpha}_{Mn}^f, \alpha_{Mn}^f), (\hat{\beta}_{mn}^f, \beta_{mn}^f) \right\}$. Owing to (10)-(15) and (17), (18) along with stationarity of the processes $\{x_k\}$ and $\{y_k\}$, we get

$$E\hat{c} = c \quad (25)$$

and hence $E(\hat{c} - c)^2 = \text{var}(\hat{c})$. As for $\hat{c} \in \left\{ \hat{\alpha}_{Mn}^g, \hat{\beta}_{mn}^g, \hat{\alpha}_{Mn}^f, \hat{\beta}_{mn}^f \right\}$ it holds that $\text{var}(\hat{c}) \leq C_c N^{-1}$ some $C_c > 0$ independent of N (see Appendix B), thus $\hat{c} \rightarrow c$ in mean square as $N \rightarrow \infty$ for each pair $(\hat{c}, c) \in \left\{ (\hat{\alpha}_{Mn}^g, \alpha_{Mn}^g), (\hat{\beta}_{mn}^g, \beta_{mn}^g), (\hat{\alpha}_{Mn}^f, \alpha_{Mn}^f), (\hat{\beta}_{mn}^f, \beta_{mn}^f) \right\}$ of wavelet coefficients. Because in the sample means in (17) the y_k 's as outputs of a dynamical system, moreover corrupted by coloured noise, are dependent quantities the proof in Appendix B concerning $\hat{\alpha}_{Mn}^g$ and $\hat{\beta}_{mn}^g$ is rather involved.

B. Convergence of the empirical wavelet models

Let $(\hat{F}(x; K), F(x; K), F(x))$ belong to $\left\{ (\hat{g}(x; K), g(x; K), g(x)), (\hat{f}(x; K), f(x; K), f(x)) \right\}$. In view of (25) we have that (cf. (8), (9) and (19), (20))

$$E\hat{F}(x; K) = F(x; K) \quad (26)$$

Since moreover for each $x \in [a, b]$ and each fixed scale factor K (see (71) and (72) in Appendix C),

$$\text{var} \left\{ \hat{F}(x; K) \right\} \leq (C_F 2^K) N^{-1} \quad (27)$$

some $C_F > 0$ independent of K and N , therefore because of (26) and the above bound, we obtain

$$E \left[\hat{F}(x; K) - F(x) \right]^2 \leq AE^2(F; x; K) + (C_F 2^K) N^{-1} \quad (28)$$

where $AE(F; x; K)$ is the approximation error as in (55) in Appendix A. This bound and (56) in Appendix A imply directly the following.

Theorem 1: If the scale factor K depends on the number N of measurement data, $K = K(N)$, and

$$K(N) \rightarrow \infty, \quad 2^{K(N)}/N \rightarrow 0 \quad \text{as } N \rightarrow \infty \quad (29)$$

then $\hat{g}(x; K(N)) \rightarrow g(x)$ and $\hat{f}(x; K(N)) \rightarrow f(x)$ in mean square, and consequently $\hat{R}(x; K(N)) \rightarrow R(x)$ in probability as $N \rightarrow \infty$ for almost all $x \in [a, b]$.

Proof: Due to foregoing arguments and the fractional form of the model $\hat{R}(x; K)$ in (21), along with $g(x) = R(x)f(x)$, the convergence is obvious. ■

Thus in order to assure consistency of the model $\hat{R}(x; K)$ as the number N of data points increases it simply suffices that $K = K(N)$ and the wavelet models $\hat{g}(x; K(N))$ and $\hat{f}(x; K(N))$ in the numerator and denominator duly expand with the number of data, according to (29). Such a condition is fulfilled for $K(N) = \lfloor c \log_2 N \rfloor$ with $0 < c < 1$.

Remark 3: Under weak assumptions as in Section II, from the mean square consistency of the models $\hat{g}(x; K(N))$ and $\hat{f}(x; K(N))$ we can only infer weak consistency of the ratio model $\hat{R}(x; K(N))$, i.e. that $P \left\{ \left| \hat{R}(x; K(N)) - R(x) \right| > \varepsilon \right\} \rightarrow 0$ as $N \rightarrow \infty$ for each $\varepsilon > 0$. However, as it follows from the Markov inequality and the Lebesgue dominated convergence theorem, convergence in probability is equivalent to more intuitive convergence of $\hat{R}(x; K(N))$ in mean, i.e. $E \left| \hat{R}(x; K(N)) - R(x) \right| \rightarrow 0$ as $N \rightarrow \infty$, provided that for the model $\hat{R}(x; K(N))$ it holds that $E \left\{ \sup_N \left| \hat{R}(x; K(N)) \right| \right\} \leq C_R < \infty$ almost everywhere on $[a, b]$, some $C_R > 0$. This is achieved if $E \left\{ \sup_N |\hat{g}(x; K(N))| \right\} \leq C_g < \infty$ and $\left| \hat{f}(x; K(N)) \right| \geq \delta_f$ (compare (6) in Assumption A7) almost everywhere over $[a, b]$, some $C_g, \delta_f > 0$ each N . The former is obtained under more restrictive demand that the overall noise corrupting the system (1) is bounded (see (1), Assumption A1, (17), (19) and (49), (53) in Appendix A), and the latter is ensured after slight modification of the model $\hat{f}(x; K(N))$, namely by considering $\hat{f}_{\text{mod}}(x; K(N)) = \hat{f}(x; K(N)) I_{\{|f(x; K(N))| \geq \delta_f\}}(x) + \delta_f I_{\{|f(x; K(N))| < \delta_f\}}(x)$ where $0 < \delta_f < \delta$ (see (6)) instead of $\hat{f}(x; K(N))$. The examination of the appropriate modified model $\hat{R}_{\text{mod}}(x; K(N)) = \hat{g}(x; K(N)) / \hat{f}_{\text{mod}}(x; K(N))$ is beyond the scope of this paper. See [33] for some hints concerning analysis of the modified density model $\hat{f}_{\text{mod}}(x; K(N))$.

VI. RATE OF CONVERGENCE

The rate of convergence in Theorem 1 strongly depends on the particular choice of a constant c in the function $K(N) = \lfloor c \log_2 N \rfloor$ and the behaviour of the approximation error $AE(g; x; K)$ and $AE(f; x; K)$. For comparison purposes with conventional orthogonal series models, we discuss convergence rate of the model $\hat{R}(x; K(N))$ for local smoothness classes of $R(x)$ and $f(x)$ typically considered in the literature. In the following, for a sequence of random variables $\{\vartheta_N\}$ and positive number sequence $\{b_N\}$ convergent to zero, by $\vartheta_N = O(b_N)$ in probability we mean that $r_N(\vartheta_N/b_N)$ tends to zero in probability as $N \rightarrow \infty$, i.e.

$$P \{ |r_N| (|\vartheta_N|/b_N) > \varepsilon \} \rightarrow 0 \quad \text{as } N \rightarrow \infty \quad (30)$$

each $\varepsilon > 0$, where $\{r_N\}$ is a number sequence arbitrarily slowly tending to zero.

A. Approximation error

Let $F(x) \in \{R(x), f(x)\}$. Assume that $F(x) \in C^{\nu_F}(x_0)$ where $\nu_F \in (0, 1)$ or $\nu_F \in \mathbf{N}$ - the set of natural numbers. For fractional $\nu_F \in (0, 1)$ this means that $F(x)$ is Lipschitz continuous function around x_0 with the exponent ν_F , $F(x) \in Lip(x_0, \nu_F)$, i.e.

$$|F(x) - F(x_0)| \leq L_F |x - x_0|^{\nu_F} \quad (31)$$

some $L_F > 0$. For $\nu_F \in \mathbf{N}$ the nonlinearity $F(x)$ is by definition ν_F times continuously differentiable around x_0 , and the following bound holds in the neighborhood of x_0 (by the Taylor series expansion formula)

$$|F(x) - F(x_0)| \leq L_{F1} |x - x_0| + L_{F2} |x - x_0|^2 + \dots + L_{F\nu_F} |x - x_0|^{\nu_F} \quad (32)$$

some $L_{Fd} > 0$, $d = 1, 2, \dots, \nu_F$. We note that $F(x) \in \{R(x), f(x)\}$ does not need to be smooth, and even continuous, over whole identification region $[a, b]$.

Remark 4: According to (31) and (32), the standard Lipschitz functions $Lip(x_0, 1)$ are included into the class $C^1(x_0)$, together with once continuously differentiable functions. Therefore, to avoid ambiguity, the particular smoothness of $F(x) \in C^1(x_0)$ will be further individually specified, if necessary.

Remark 5: If $R(x) \in C^{\nu_R}(x_0)$ and $f(x) \in C^{\nu_f}(x_0)$ then the product function $g(x) = R(x)f(x) \in C^{\nu_g}(x_0)$ where $\nu_g = \min\{\nu_R, \nu_f\}$, which is straightforward to see.

Consider the approximation error as in (55) in Appendix A for $F(x) \in C^{\nu_F}(x_0)$ and wavelet functions $\psi(x)$ possessing r_ψ vanishing moments (cf. (57) in Appendix A). Since for large values of m for the

wavelet coefficients β_{mn} ((51) in Appendix A) it holds that (cf. e.g. [26])

$$|\beta_{mn}| \leq C_\beta^F 2^{-(\lambda_{F\psi} + \frac{1}{2})m} \quad \text{all } n \quad (33)$$

some $C_\beta^F > 0$ independent of m where $\lambda_{F\psi} = \min\{\nu_F, r_\psi + 1\}$, the approximation error $AE(F; x; K)$ at the point x_0 for large K is bounded as follows (see (33) and (55) in Appendix A)

$$|AE(F; x_0; K)| \leq S M_\psi C_\beta^F \sum_{m=K}^{\infty} 2^{m/2} 2^{-(\lambda_{F\psi} + \frac{1}{2})m}$$

yielding

$$|AE(F; x_0; K)| \leq C_{AE, F\psi} 2^{-\lambda_{F\psi} K} \quad (34)$$

where $C_{AE, F\psi} = S M_\psi C_\beta^F / (1 - 2^{-\lambda_{F\psi}})$. Hence and Remark 6, for the approximation error $AE(g; x_0; K)$ and $AE(f; x_0; K)$ we obtain in particular the following

$$|AE(g; x_0; K)| \leq C_{AE, g\psi} 2^{-\lambda_{g\psi} K} \quad (35)$$

and

$$|AE(f; x_0; K)| \leq C_{AE, f\psi} 2^{-\lambda_{f\psi} K} \quad (36)$$

where $C_{AE, g\psi} = S M_\psi C_\beta^g / (1 - 2^{-\lambda_{g\psi}})$, $C_{AE, f\psi} = S M_\psi C_\beta^f / (1 - 2^{-\lambda_{f\psi}})$ are appropriate constants and $\lambda_{g\psi} = \min\{\nu_g, r_\psi + 1\} = \min\{\nu_R, \nu_f, r_\psi + 1\}$ whereas $\lambda_{f\psi} = \min\{\nu_f, r_\psi + 1\}$. Since $\lambda_{g\psi} \leq \lambda_{f\psi}$, the guaranteed approximation rate of $g(x)$ is not faster than the approximation rate of $f(x)$. If $\nu_R < \nu_f$ (i.e. input density $f(x)$ is a smoother function than the identified nonlinearity $R(x)$ which is rather a typical situation in practice), we get $\lambda_{g\psi} < \lambda_{f\psi}$ which prejudices slower approximation rate of $g(x)$ provided that $r_\psi + 1 > \nu_R$, i.e. applied wavelet functions are regular enough.

B. Convergence rate

Suppose that $R(x) \in C^{\nu_R}(x_0)$ and $f(x) \in C^{\nu_f}(x_0)$. Taking into account (28), (35), (36) and Lemma 3 in Appendix D, we ascertain that asymptotically (for large values of N and hence large $K(N)$; cf. (29)) it holds

$$\hat{g}(x_0; K(N)) = g(x_0) + O\left(\left(2^{-2\lambda_{g\psi}K(N)} + 2^{K(N)}/N\right)^{1/2}\right)$$

and

$$\hat{f}(x_0; K(N)) = f(x_0) + O\left(\left(2^{-2\lambda_{f\psi}K(N)} + 2^{K(N)}/N\right)^{1/2}\right)$$

in probability, and further (by Lemma 4)

$$\hat{R}(x_0; K(N)) = R(x_0) + O\left(\left(2^{-2\lambda_{g\psi}K(N)} + 2^{K(N)}/N\right)^{1/2}\right) \quad (37)$$

in probability including that $\lambda_{g\psi} \leq \lambda_{f\psi}$. Clearly, the best asymptotic convergence rate in (37) is achieved when the two (antagonistic) components of the error are in balance. This leads to the following.

Theorem 2: Let $R(x) \in C^{\nu_R}(x_0)$, $f(x) \in C^{\nu_f}(x_0)$ and let $\psi(x)$ be a wavelet function possessing r_ψ vanishing moments. If the scale factor $K(N)$ is selected according to the rule

$$K_{opt}(N) = \lfloor (1/(2\lambda_{g\psi} + 1)) \log_2 N \rfloor \quad (38)$$

then

$$\hat{R}(x_0; K_{opt}(N)) = R(x_0) + O\left(N^{-\lambda_{g\psi}/(2\lambda_{g\psi}+1)}\right) \quad (39)$$

in probability where $\lambda_{g\psi} = \min\{\nu_R, \nu_f, r_\psi + 1\}$, and this is the best guaranteed asymptotic rate of convergence of the wavelet model $\hat{R}(x; K(N))$ for the triple $(R(x), f(x), \psi(x))$.

Proof: The conclusion follows immediately from (37). ■

In view of (39), an optimal matching of the wavelets in the model $\hat{R}(x; K_{opt}(N))$ to the smoothness of $R(x)$ and $f(x)$ is obtained when $r_\psi + 1 = \lceil \min\{\nu_R, \nu_f\} \rceil$. Then we obtain the fastest rate $O\left(N^{-\min\{\nu_R, \nu_f\}/(2\min\{\nu_R, \nu_f\}+1)}\right)$ for given ν_R and ν_f . Higher regularity of wavelets (larger number of vanishing moments) does not improve convergence and lower regularity decreases the rate in (39).

Because of the effect of the approximation error $AE(g; x; K)$ ((35)) and $AE(f; x; K)$ ((36)) apparent in (37), the rate in (39) is slower than the rate $O(N^{-1/2})$ in probability achieved in the absence of such an error (compare (28)). This rate, $O\left(N^{-1/(2+1/\lambda_{g\psi})}\right)$, for large values of $\lambda_{g\psi}$ may be close to $O(N^{-1/2})$ but cannot exceed the latter (such a general phenomenon is well known in the nonparametric inference [29]). Nevertheless, the following beneficial properties of the wavelet model $\hat{R}(x; K_{opt}(N))$ can be concluded from (39).

1. If $R(x)$ and $f(x)$ are Lipschitz functions on $[a, b]$ ($\nu_R = \nu_f = 1$) then the rate in (39) is $O(N^{-1/3})$ in probability for each $x \in [a, b]$ and each wavelet family. As established in [29], this is the best possible nonparametric rate of convergence for Lipschitz nonlinearities.
2. If $R(x) \in C^{\nu_R}(x_0)$, $f(x) \in C^{\nu_f}(x_0)$, some $\nu_R, \nu_f \in \mathbf{N}$, and moreover $\nu_f \geq \nu_R$ and $r_\psi + 1 \geq \nu_R$ then the rate in (39) is $O(N^{-\nu_R/(2\nu_R+1)})$ in probability. This is the best possible nonparametric rate of convergence for differentiable nonlinearities $R(x)$ [29].

Because of these optimality properties and freedom in the choice of wavelet functions, the wavelet-based model $\hat{R}(x; K_{opt}(N))$ can outperform conventional orthogonal series models elaborated earlier in the literature. For example, for Lipschitz $R(x)$ and $f(x)$ around x_0 the attainable rate $O(N^{-1/3})$ in probability is better than $O(N^{-1/4})$ in probability guaranteed for more smooth differentiable $R(x), f(x) \in C^1(x_0)$ by the trigonometric and Hermite series models (see [15], [13]). The rate $O(N^{-1/4})$ can, in turn, be achieved

by the wavelet model $\hat{R}(x; K_{opt}(N))$ for considerably less smooth $R(x), f(x) \in C^{1/2}(x_0)$. For more smooth differentiable $R(x), f(x) \in C^d(x_0)$, $d \in \mathbb{N}$, and wavelet-based models with $r_\psi + 1 \geq d$ we can achieve the convergence rate $O(N^{-d/(2d+1)})$ which again is faster than $O(N^{-(2d-1)/4d})$ in probability achieved by the models using trigonometric or Hermite series approximations (see the references cited above). Specifically, for locally constant functions $R(x)$ and $f(x)$ around x_0 , i.e. $R(x), f(x) \in C^\infty(x_0)$ in our denotation, the model $\hat{R}(x; K_{opt}(N))$ can potentially achieve the parametric rate of convergence $O(N^{-1/2})$ in probability.

It should be mentioned that the presented asymptotic rates of convergence refer to characteristics embedded in a dynamical system which can be corrupted by arbitrarily correlated noise (Assumption A4 in Section II). The particular system and noise dynamics as well as level (variance) of the noise do not alter the order of the guaranteed asymptotic rate of convergence which constantly remains the same as for a static system contaminated by white noise. This is directly seen from (37)-(39) and derivations in Appendix B and C.

C. Practical scale selection strategy

The scale selection rule in (38) needs the values of local smoothness indices ν_R and ν_f and hence (i) requires advanced prior knowledge about regularity of the nonlinearity $R(x)$ and the input density $f(x)$ and (ii) can vary from point to point, along with the local smoothness of $R(x)$ and $f(x)$, yielding unstable values of $K_{opt}(N)$. Because of this disadvantage and since the indices ν_R and ν_f are usually unknown, we propose the following 'rule of thumb' for selecting the scale in the wavelet model, independent of local regularity of $R(x)$ and $f(x)$:

$$\bar{K}(N) = \lfloor (1/3) \log_2 N \rfloor \quad (40)$$

Owing to (37) and obvious inequality $\min\{\nu_R, \nu_f, r_\psi + 1\} \geq \min\{\nu_R, \nu_f, 1\}$, for this rule we obtain the asymptotic guaranteed rate

$$\hat{R}(x_0; \bar{K}(N)) = R(x_0) + O\left(N^{-\min\{\nu_R, \nu_f, 1\}/3}\right) \quad \text{in probability} \quad (41)$$

irrespective of the applied wavelet family, which ensures convergence $O(N^{-1/3})$ in probability provided that $\nu_R, \nu_f \geq 1$, i.e. $R(x)$ and $f(x)$ are at least Lipschitz functions around x_0 . For $\min\{\nu_R, \nu_f\} \geq 1$ the rule assures balance of both error components in (37), and no other rule $K(N) = \lfloor (1/n) \log_2 N \rfloor$, integer $n > 1$, yields better warranted convergence of the error for each wavelet model (each $r_\psi \geq 0$). Since Lipschitz regularity of nonlinearity $R(x)$ and input density $f(x)$ can be reliably expected in most

real-life situations and the rate $O(N^{-1/3})$ seems quite satisfactory as compared with the theoretically best possible rate of convergence $O(N^{-1/2})$ in probability (see Section VI-B), we recommend the rule (40) for practical use in the absence of prior knowledge of ν_R and ν_f . When considering rougher functions $R(x)$ or $f(x)$ around x_0 , with $\min\{\nu_R, \nu_f\} < 1$, the guaranteed rate in (41) is deteriorated to the order $O(N^{-\min\{\nu_R, \nu_f\}/3})$ in probability. For fractional $\min\{\nu_R, \nu_f\} = k/l$ with $k \ll l$ we get very slow rate $O(N^{-k/3l})$. Such situation seems however to be a very special case in practice. Practical utility of the rule (40) is verified empirically in Section VII. The results presented there confirm fair efficiency of the strategy.

Remark 6: Convergence properties of specific wavelet models can be easily concluded from (37), Theorem 2 and discussion in Section VI-C. For example, if we applied wavelets collected in Table I in Appendix A, they follow from the mentioned facts and the specifications concerning the number of vanishing moments r_ψ given in Table I. As an instance, for the models using Haar wavelets (a particular case of Daubechies wavelets for the wavelet number $p = 1$) we may conclude the following: if around x_0 , $R(x) \in C^{\nu_R}(x_0)$, $f(x) \in C^{\nu_f}(x_0)$ and the scale factor is selected according to the rule (40) then $\hat{R}_H(x_0; \bar{K}(N)) = R(x_0) + O(N^{-1/3})$ in probability for $\nu_R, \nu_f \geq 1$ and this is the best possible rate of convergence of the Haar wavelet model $\hat{R}_H(x; K(N))$. Another conclusion, concerning whole class of Daubechies wavelet-based models, is that for $R(x)$ and $f(x)$ with $\min\{\nu_R, \nu_f\} < 1$ employing Daubechies wavelets with the wavelet numbers $p > 1$ is pointless from the rate of convergence viewpoint as the rate in (39) is then not faster than for the Haar wavelets with $p = 1$.

VII. NUMERICAL EXPERIMENT

In the experiment, we examine the performance of the models $\hat{R}_D^p(x; K(N))$ obtained for the Daubechies wavelets $(\varphi_D^p(x), \psi_D^p(x))$ with $p \geq 1$ (see Table I in Appendix A). Since the Daubechies wavelets have the shortest supports among all orthogonal compactly supported wavelet functions with a given number of vanishing moments and hence require the smallest number of components in the generic wavelet models (19) and (20), the Daubechies wavelet models $\hat{R}_D^p(x; K(N))$ are most parsimonious within the class considered in the paper (see [5], [26], Table I and (53) in Appendix A). In our simulations we assume $[a, b] = [0, 1]$ (identification region) and use Hammerstein systems which represent well the systems reported in Section III, with three nonlinear characteristics $\mu(x)$ shown in $[0, 1]$ in Fig. 6:

$$\begin{aligned} \text{polynomial} & : \mu_1(x) = 10(2x^3 - 3x^2 + x) \\ \text{cube root} & : \mu_2(x) = \sqrt[3]{x - 1/2} \end{aligned} \quad (42)$$

$$\text{quantizer : } \mu_3(x) = 1/8 + \lfloor 8x - 4 \rfloor / 4$$

The polynomial nonlinearity illustrates smooth infinitely times continuously differentiable functions with finite power series representation, for which approximation error is particularly sensitive to the number of vanishing moments of the applied wavelet functions (see (51), (55) and (57) in Appendix A). The cube root nonlinearity is a smooth almost everywhere differentiable function with infinite power series representation and hence the approximation error less sensitive to the number of vanishing moments of applied wavelets. In turn, the quantizer nonlinearity (piecewise-constant) is an example of discontinuous functions with jumps which match well the dyadic grid points and thus are well-fitted to the implementation of the Haar wavelet models ($p = 1$). The linear output dynamics was selected as (i) FIR element with the impulse

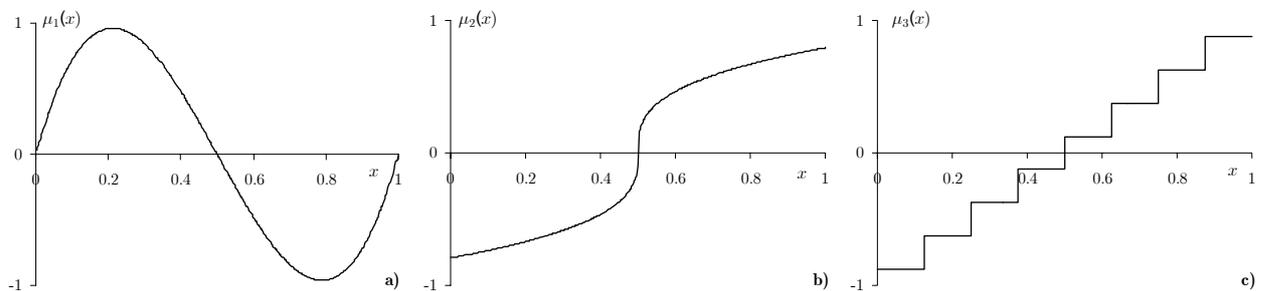


Fig. 6. Non-linear characteristics of the simulated Hammerstein systems in $[0, 1]$ (identification region): **a)** polynomial, **b)** cube root, **c)** quantizer non-linearity.

response $\{\lambda_i = 1 - i/4\}_{i=0}^5$ and (ii) IIR subsystem with the description $v_k - 0.75v_{k-1} + 0.5v_{k-2} = w_k + 0.75w_{k-1} + 0.5w_{k-2}$ (where $w_k = \mu(x_k)$ is the input and v_k is the output of the dynamics). The systems were driven by white, stationary random sequence $\{x_k\}$ of the uniform distribution $x_k \sim U[0, 1]$ (Assumption A2 in Section II), and thus the input probability density function $f(x)$ was infinitely times continuously differentiable in each internal point of the identification region. Because in our tests the nonlinearities $\mu(x)$ are anti-symmetric with respect to the center $x = 0.5$ of the region $[0, 1]$ (Fig. 6) and the input density is symmetric, we get in each case $E\mu(x_1) = 0$. Since moreover $\lambda_0 = 1$, we have $R(x) = \mu(x)$ (see Section III) i.e. in our experiment we identify true nonlinear characteristics of the Hammerstein systems. Finally, the external correlated zero mean output noise $\{z_k\}$ was generated as an output of MA(l) filter $z_k = \sum_{i=0}^l \omega_i \varepsilon_{k-i}$, driven by the uniformly distributed white noise process $\varepsilon_k \sim U[-c, c]$, independent of $\{x_k\}$ (see Assumption A4 for $L = 1$ and Assumption A5). In simulations we assumed $\{\omega_i = 1 - i/(l+1)\}_{i=0}^l$ and $l = 0, 1, \dots, 4$ obtaining $\{z_k\}$ white for $l = 0$ and correlated for $l \geq 1$, with increasing correlation for growing l . For each identified nonlinearity $R(x) = \mu(x)$ and

each value of l , the white noise parameter c was selected as to give constant the noise-to-signal ratio:

$$NSR \stackrel{\text{def}}{=} \frac{\max |z_k|}{\max_{x \in [0,1]} |R(x)|}$$

In all tests we set $NSR = 10\%$. In fact, according to (1), the noise blurring the nonlinearity $R(x)$ was enlarged by the system dynamics, and factually $y_k = R(x_k) + \xi_k + z_k$ with $\xi_k = \sum_{i=1}^{\infty} \lambda_i \mu(x_{k-i})$ (see Section III) yielding a substantial additional disturbance for large values of $\sum_{i=1}^{\infty} |\lambda_i|$. Such influence was discussed in [21].

The experiments were performed using the models $\hat{R}_D^p(x; K(N))$ with the wavelet numbers $p = 1, 2, \dots, 5$. We assumed the initial scale $M = 2$ and the scale factor K was computed according to the scale selection rule $\bar{K}(N)$ in (40). In what follows, we denote $\hat{R}_D^p(x; \bar{K}(N)) = R_N(x; p)$ for shortness. Accuracy of the models was evaluated using the empirical average pointwise identification error of the form

$$Q(N; p) = \frac{1}{T} \sum_{r=1}^T \left\{ \frac{1}{1000} \sum_{q=1}^{1000} [R(x_q) - R_N^r(x_q; p)]^2 \right\} \quad (43)$$

where T is the number of independent trials for each sample size N , $R_N^r(x; p)$ is the empirical model computed in the r -th run for N data points, and $x_q = q/1000$, $q = 1, 2, \dots, 1000$, are the equidistant estimation points from the interval $[0, 1]$. In our experiment we assumed $T = 100$ and $N = 50, 100, \dots, 500$.

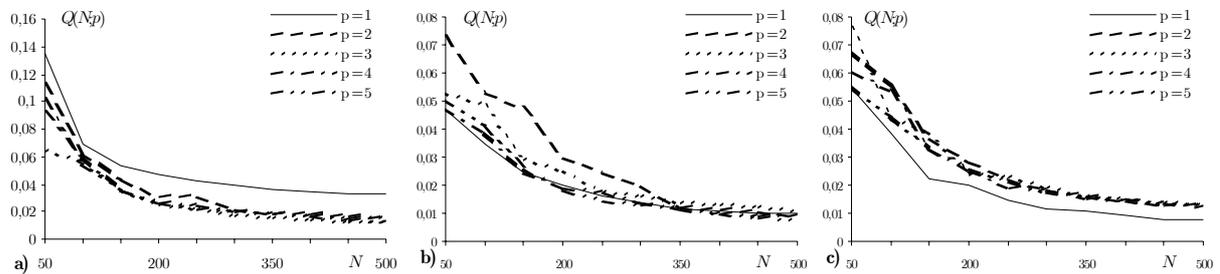


Fig. 7. Accuracy of the models $R_N(x; p)$ against the sample size N and the wavelet number p for **a)** polynomial, **b)** cube root, **c)** quantizer non-linearity; FIR dynamics and MA(2) output noise.

Precision of the models is illustrated in Figs. 7 and 8 where the $Q(N; p)$ error is shown jointly for $p = 1, 2, \dots, 5$ and growing number N of data for each test nonlinearity and, respectively, FIR and IIR dynamics. We see that the error decreases rapidly with increasing N and good fit between the nonlinear characteristics and the wavelet models is guaranteed for quite moderate N . For $N \geq 350$ the error is almost the same for each wavelet number $p \geq 2$ and each sample nonlinear function. For $p = 1$ (Haar wavelets) we obtain larger (for smooth polynomial nonlinearity) and smaller (for quantizer nonlinearity)

identification error, however still the same regarding the order of the error rate of convergence (shape of the plots). This is in a good agreement with the asymptotic convergence rate formula in (41). Nevertheless, in each case one can point out $p = p^*$ minimizing the empirical error $Q(N; p)$. In our experiment, $p^* = 3$ for polynomial nonlinearity, $p^* = 5$ for cube root nonlinearity, and $p^* = 1$ for quantizer nonlinearity, in accordance with expectations (see the discussion concerning test nonlinearities). Similarity of the shape of plots in Figs. 7 and 8 confirms robustness of the order of the identification error against system dynamics, established in Section VI.

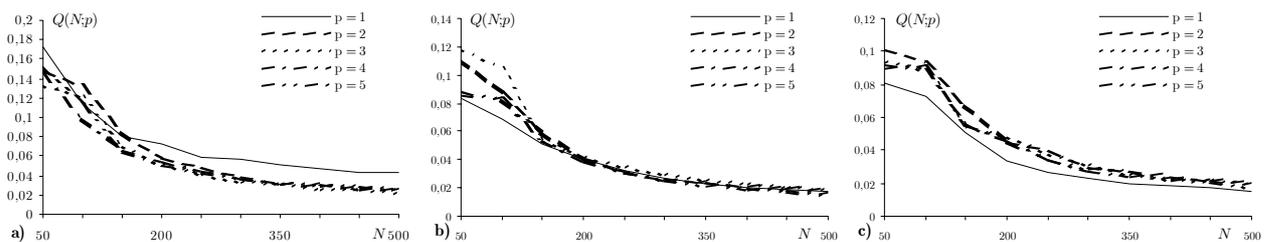


Fig. 8. Accuracy of the models $R_N(x; p)$ against the sample size N and the wavelet number p for **a)** polynomial, **b)** cube root, **c)** quantizer non-linearity; IIR dynamics and MA(2) output noise.

To show the performance of the models for various correlation patterns of the noise, Fig. 9 presents the plots of the $Q(N; p^*)$ error against the order l of the external noise MA(l) for FIR dynamics and each test nonlinearity. It is seen that for $N \geq 100$ and l varying over the range $0 \leq l \leq 4$ the error is nearly the same, i.e. the computed models are insensitive to correlation of the output noise $\{z_k\}$ yielding, in particular, similar identification error for white noise ($l = 0$) and the noise MA(4).

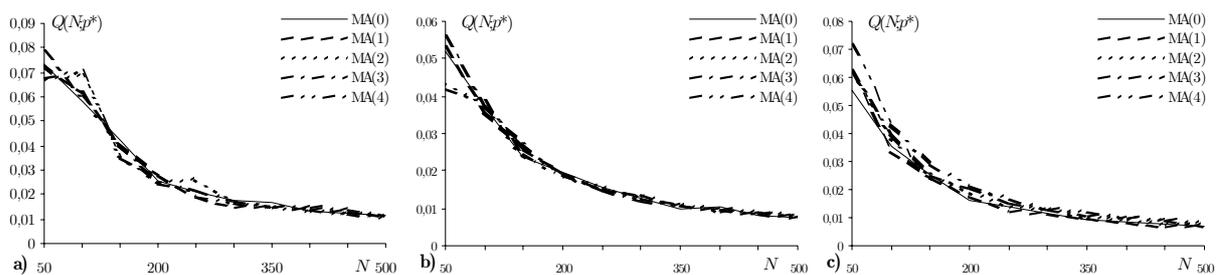


Fig. 9. Effect of the correlation of the output noise on the accuracy of the models $R_N(x; p^*)$ for **a)** polynomial, **b)** cube root, **c)** quantizer non-linearity; MA(l) output noise model, FIR dynamics.

VIII. CONCLUSIONS

We have focused on the analysis of local, pointwise, behaviour of the wavelet models of nonlinear characteristics of a class of block-oriented dynamical systems operating under random excitations and random disturbances. It was demonstrated that for appropriate choice of the scale factor K in relation to the number of data N , the models converge in probability almost everywhere to the true nonlinear characteristics and the convergence holds irrespective of system and noise dynamics. This means in particular that the proposed models are consistent, regardless of whether the noise contribution on the sampled data can be modeled exactly. Under suitable smoothness conditions concerning the nonlinear characteristic and the input probability density function we have determined the rule for optimum size selection of the wavelet model, minimizing the asymptotic identification error, and established asymptotic rate of convergence. It was shown that the models can attain the best possible nonparametric rate of convergence and may perform better than trigonometric and Hermite series models worked out earlier in the literature. It is noteworthy that for the use of the models only a relatively small number of coefficients $\hat{\alpha}_{Mn}^g, \hat{\beta}_{mn}^g, \hat{\alpha}_{Mn}^f, \hat{\beta}_{mn}^f$ must be calculated from experimental data. For the nonlinearity $R(x)$ to be identified in the bounded region $[a, b]$, the needed number of coefficients is of order $O(2^K)$ (equation (24)). For the scale factor K chosen as in Theorem 1 this number is much smaller than the number N of measured raw data and the 'data compression' rapidly increases with growing N as $2^K/N \rightarrow 0$ for $N \rightarrow \infty$.

In this paper we have not considered nonlinear in-the-parameters wavelet models which might be obtained by applying thresholding to wavelet model coefficients (see [26] or [22] for a review of standard thresholding techniques). The reason is that proper evaluation of such techniques in system identification framework requires preliminary entire analysis of the behaviour in this context of linear in-the-parameters wavelet models, which has not been done yet and hence was the aim of the present paper. Various thresholding methods are at present widely and successfully used in statistics in nonparametric regression [8] and density estimation [9], and in engineering in, for instance, computer graphics [10] and signal and image processing for denoising and compression; see, e.g., [7], [2], [6], [22] and the references cited therein. In the aforementioned problems no dynamics is however present, and in particular all measured data are statistically independent, provided that blurring noise is white. This is not the case in system identification tasks where system dynamics is an inherent feature of the problem. The dynamics next conveyed to the measured data makes in particular the output measurements $\{y_k\}$ statistically dependent (correlated) quantities (see equation (1) and Assumptions A3 and A4). This converts in turn into much

more complicated correlation patterns of wavelet coefficients. Since the situation is quite different, it is not clear if shrinkage strategy, very successful for 'static' problems, can be simply conveyed to systems with dynamics producing dependent data, and whether wavelet thresholding may play in system identification any significant role, particularly in non-Gaussian settings. The experiments with Hammerstein system in the settings as in Section VII show for example that conventional soft thresholding technique ([7]) yields questionable results. Fig. 10 presents jointly appropriate $Q(N; p^*)$ error (43) for the thresholded wavelet model $R_{N,thr}(x; p^*)$ along with the error for the linear in-the-parameters model $R_N(x; p^*)$ for visual evaluation of the models. Because of clear deterioration of the quality of the thresholded model,

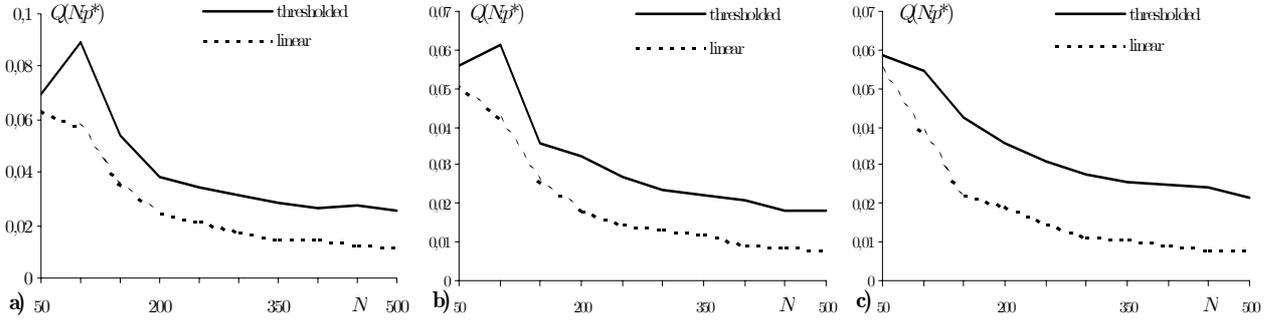


Fig. 10. Accuracy of the linear $R_N(x; p^*)$ and soft thresholded $R_{N,thr}(x; p^*)$ models against the sample size N for **a)** polynomial, **b)** cube root, **c)** quantizer non-linearity; FIR dynamics and MA(2) output noise.

the problem appears essential and worth a separate investigation. It is left for future work.

APPENDIX A

We report below basic facts concerning wavelet approximation of functions, used in the paper.

1. Any square integrable function $F(x) \in L^2(\mathbf{R})$ can be approximated with the help of orthogonal wavelets $\{\varphi_{Mn}(x)\}_{n \in \mathbf{Z}} \cup \{\psi_{mn}(x)\}_{M \leq m \leq K-1, n \in \mathbf{Z}}$, \mathbf{Z} —the set of integers, and the wavelet approximator (wavelet model) in the adequate approximation (model) space $V_K = V_M \oplus W_M \oplus W_{M+1} \oplus \dots \oplus W_{K-1} \subset L^2(\mathbf{R})$ has the form

$$F(x; K) = \sum_{n=-\infty}^{\infty} \alpha_{Mn} \varphi_{Mn}(x) + \sum_{m=M}^{K-1} \sum_{n=-\infty}^{\infty} \beta_{mn} \psi_{mn}(x) \quad (44)$$

where

$$\alpha_{Mn} = \int_{-\infty}^{+\infty} F(x) \varphi_{Mn}(x) dx; \quad \beta_{mn} = \int_{-\infty}^{+\infty} F(x) \psi_{mn}(x) dx \quad (45)$$

and

$$\varphi_{Mn}(x) = 2^{M/2} \varphi(2^M x - n) ; \quad \psi_{mn}(x) = 2^{m/2} \psi(2^m x - n) \quad (46)$$

are scaled (factor m) and translated (factor n) versions of father and mother wavelet $\varphi(x)$ and $\psi(x)$; they span the initial approximation space V_M and the orthogonal detail (wavelet) spaces W_m :

$$V_M = \text{span} \{ \varphi_{Mn}(x), n \in \mathbf{Z} \} ; \quad W_m = \text{span} \{ \psi_{mn}(x), n \in \mathbf{Z} \} .$$

2. The approximator (44) can be step by step refined by adding further details

$$F(x; K+1) = F(x; K) + \sum_{n=-\infty}^{\infty} \beta_{Kn} \psi_{Kn}(x)$$

3. For bounded wavelets with compact support

$$|\varphi(x)| \leq M_\varphi I_{[s_1, s_2]}(x) ; \quad |\psi(x)| \leq M_\psi I_{[t_1, t_2]}(x) \quad (48)$$

some $M_\varphi, M_\psi > 0$, and hence (by (46))

$$|\varphi_{Mn}(x)| \leq 2^{M/2} M_\varphi I_{[\frac{s_1+n}{2^M}, \frac{s_2+n}{2^M}]}(x) ; \quad |\psi_{mn}(x)| \leq 2^{m/2} M_\psi I_{[\frac{t_1+n}{2^m}, \frac{t_2+n}{2^m}]}(x) \quad (49)$$

the approximator (44) takes the form

$$F(x; K) = \sum_{n=n_{\min}(\varphi, x, M)}^{n_{\max}(\varphi, x, M)} \alpha_{Mn} \varphi_{Mn}(x) + \sum_{m=M}^{K-1} \sum_{n=n_{\min}(\psi, x, m)}^{n_{\max}(\psi, x, m)} \beta_{mn} \psi_{mn}(x) \quad (50)$$

where

$$\alpha_{Mn} = \int_{(s_1+n)/2^M}^{(s_2+n)/2^M} F(x) \varphi_{Mn}(x) dx ; \quad \beta_{mn} = \int_{(t_1+n)/2^m}^{(t_2+n)/2^m} F(x) \psi_{mn}(x) dx \quad (51)$$

and

$$\begin{aligned} n_{\min}(\varphi, x, M) &= \lfloor 2^M x - s_2 \rfloor ; & n_{\max}(\varphi, x, M) &= \lceil 2^M x - s_1 \rceil \\ n_{\min}(\psi, x, m) &= \lfloor 2^m x - t_2 \rfloor ; & n_{\max}(\psi, x, m) &= \lceil 2^m x - t_1 \rceil \end{aligned} \quad (52)$$

indicate active (nonvanishing) wavelets at point x ($\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are respectively the 'floor' and 'ceiling' function).

4. The number of components (nonzero wavelet coefficients) in (50) depends on the support size $s = s_2 - s_1 = t_2 - t_1$ of $\varphi(x)$ and $\psi(x)$ (the same for both wavelet functions, where moreover $t_1 = -(s_2 - s_1 - 1)/2$ and $t_2 = (s_2 - s_1 + 1)/2$ [26]) and for each M, m and x it holds that

$$\begin{aligned} n_{\max}(\varphi, x, M) - n_{\min}(\varphi, x, M) + 1 &\leq S \\ n_{\max}(\psi, x, m) - n_{\min}(\psi, x, m) + 1 &\leq S \end{aligned} \quad (53)$$

where $S = \lfloor s \rfloor + 1$.

5. The supports of wavelets present in (50) (active at point x) are included in the interval

$$[x_{\min}(x, M), x_{\max}(x, M)] = [x - s/2^M, x + s/2^M] \quad (54)$$

6. The pointwise approximation error of the approximator (50) is (cf. (49) and (53))

$$\begin{aligned} |AE(F; x; K)| &= |F(x) - F(x; K)| = \left| \sum_{m=K}^{\infty} \sum_{n=n_{\min}(\psi, x, m)}^{n_{\max}(\psi, x, m)} \beta_{mn} \psi_{mn}(x) \right| \leq \\ &\leq S M_{\psi} \sum_{m=K}^{\infty} 2^{m/2} \max\{|\beta_{mn}| : n_{\min}(\psi, x, m) \leq n \leq n_{\max}(\psi, x, m)\} \end{aligned} \quad (55)$$

7. For wavelet functions as in (48) it holds that

$$|AE(F; x; K)| \rightarrow 0 \quad \text{as} \quad K \rightarrow \infty \quad (56)$$

for almost all x (in the sense of Lebesgue measure), and in particular at the continuity points of $F(x)$ [23, Theorem 2.1(ii)].

8. By definition, the mother wavelet $\psi(x)$ has r_{ψ} vanishing moments if

$$\int_{t_1}^{t_2} x^k \psi(x) dx = 0 \quad \text{for} \quad k = 0, 1, \dots, r_{\psi}; \quad (57)$$

then also [26]

$$\int_{(t_1+n)/2^m}^{(t_2+n)/2^m} x^k \psi_{mn}(x) dx = 0 \quad \text{for} \quad k = 0, 1, \dots, r_{\psi}$$

for each $m, n \in \mathbf{Z}$ and the wavelet support size is $s \geq 2r_{\psi} + 1$ [26, Proposition 7.4].

	Daubechies/symmlet	coiflet
Support of $\varphi(x)$	$[0, 2p - 1]$	$[-2p, 4p - 1]$
Support of $\psi(x)$	$[1 - p, p]$	$[1 - 3p, 3p]$
$n_{\min}(\varphi, x, m)$	$\lfloor 2^m x \rfloor - 2p + 2$	$\lfloor 2^m x \rfloor - 4p + 2$
$n_{\max}(\varphi, x, m)$	$\lceil 2^m x \rceil - 1$	$\lceil 2^m x \rceil + 2p - 1$
$n_{\min}(\psi, x, m)$	$\lfloor 2^m x \rfloor - p + 1$	$\lfloor 2^m x \rfloor - 3p + 1$
$n_{\max}(\psi, x, m)$	$\lceil 2^m x \rceil + p - 2$	$\lceil 2^m x \rceil + 3p - 2$
r_{ψ}	$p - 1$	$2p - 1$

TABLE I

BASIC PROPERTIES OF TYPICAL ORTHOGONAL WAVELET FUNCTIONS WITH COMPACT SUPPORT (p – WAVELET NUMBER)

Further details can be found in the rich wavelet literature (e.g. [5], [26]). Examples of orthogonal compactly supported wavelets are Daubechies wavelets, symmlets and coiflets characterized in Table I.

APPENDIX B

For shortness, the derivations are given for $J = L = 1$ and consequently we omit in the denotations the superfluous subscript '1'. Extension to $J, L > 1$ is straightforward. Due to Assumption A6, x varies over the interval $[a, b]$.

From stationarity of the processes $\{x_k\}$ and $\{y_k\}$ (see equation (1) and Assumptions A2-A4 in Section II) it follows that (cf. (17))

$$\begin{aligned} \text{var}(\hat{\alpha}_{Mn}^g) &= \text{var} \left[N^{-1} \sum_{k=1}^N y_k \varphi_{Mn}(x_k) \right] = N^{-1} \text{var} [y_1 \varphi_{Mn}(x_1)] \\ &\quad + 2N^{-1} \sum_{k=1}^{N-1} (1 - k/N) \text{cov} [y_{k+1} \varphi_{Mn}(x_{k+1}), y_1 \varphi_{Mn}(x_1)] \\ &= N^{-1} (A + B), \quad \text{say} \end{aligned} \quad (59)$$

Since $\text{var} [y_1 \varphi_{Mn}(x_1)] \leq E [y_1^2 \varphi_{Mn}^2(x_1)]$ and $y_1^2 \leq 3(R^2(x_1) + \xi_1^2 + z_1^2)$ (by (1) and Cauchy inequality), we get for the first component

$$A \leq 3 [E [R^2(x_1) \varphi_{Mn}^2(x_1)] + (E\xi_1^2 + Ez_1^2) E [\varphi_{Mn}^2(x_1)]]$$

owing to independence of x_1 and ξ_1, z_1 (Assumptions A2, A3 and A5). Further, because of boundedness of the density function $f(x)$ (Assumption A2), orthonormality of wavelets $\{\varphi_{Mn}(x)\}$ and the fact that for x ranging over the region $[a, b]$ the supports of all active wavelets $\varphi_{Mn}(x)$ in the model (19) are contained in the interval $[x_{\min}(a, M), x_{\max}(b, M)] = [a - s/2^M, b + s/2^M]$ (see (54) in Appendix A), we recognize that for each $n_{\min}(\varphi, x, M) \leq n \leq n_{\max}(\varphi, x, M)$ where $x \in [a, b]$ (see (19)) it holds that

$$E [\varphi_{Mn}^2(x_1)] = \int_{x_{\min}(a, M)}^{x_{\max}(b, M)} \varphi_{Mn}^2(x) f(x) dx \leq M_f \quad (60)$$

$$E [R^2(x_1) \varphi_{Mn}^2(x_1)] = \int_{x_{\min}(a, M)}^{x_{\max}(b, M)} \varphi_{Mn}^2(x) R^2(x) f(x) dx \leq M_R^2 M_f \quad (61)$$

In the above expressions we have exploited that for active wavelets $\int_{x_{\min}(a, M)}^{x_{\max}(b, M)} \varphi_{Mn}^2(x) dx = 1$ and that for $x \in [x_{\min}(a, M), x_{\max}(b, M)]$ we have $|R(x)| \leq M_R$, where

$$M_R = C_{1R} \max\{|x_{\min}(a, M)|, |x_{\max}(b, M)|\} + C_{2R}$$

(see Assumption A1 in Section II). Using (60) and (61) we obtain eventually

$$A = \text{var} [y_1 \varphi_{Mn}(x_1)] \leq 3M_f [M_R^2 + \sigma_\xi^2 + \sigma_z^2] < \infty \quad (62)$$

where $\sigma_\xi^2 = E\xi_1^2 = \sigma_\zeta^2 \sum_{i=1}^{\infty} \lambda_i^2 < \infty$ and $\sigma_z^2 = Ez_1^2 = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} \omega_i^2 < \infty$ (see Assumptions A2-A4).

Concerning the second term in (59), denote $r_{Mn,\varphi}^g(k) = \text{cov}[y_{k+1}\varphi_{Mn}(x_{k+1}), y_1\varphi_{Mn}(x_1)]$. Using equation (1) and Assumptions A2-A5 (in particular, stationarity and independence of appropriate quantities along with the fact that $E\xi_1 = Ez_1 = 0$), we ascertain after standard calculation that

$$\begin{aligned} r_{Mn,\varphi}^g(k) &= \lambda_k E[\zeta(x_1)R(x_1)\varphi_{Mn}(x_1)] E[\varphi_{Mn}(x_1)] \\ &\quad + (E[\xi_{k+1}\xi_1] + E[z_{k+1}z_1]) E^2[\varphi_{Mn}(x_1)] \end{aligned}$$

Since (see the definitions of $\{\xi_k\}$ and $\{z_k\}$ in Assumptions A3 and A4 for $J = L = 1$) $\rho_\xi(k) = E[\xi_{k+1}\xi_1] = \sigma_\zeta^2 \sum_{i=1}^{\infty} \lambda_i \lambda_{i+k}$, $\rho_z(k) = E[z_{k+1}z_1] = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} \omega_i \omega_{i+k}$ and (by Schwarz inequality)

$$\begin{aligned} |E[\zeta(x_1)R(x_1)\varphi_{Mn}(x_1)]| &\leq \sigma_\zeta (E[R^2(x_1)\varphi_{Mn}^2(x_1)])^{1/2} \\ |E[\varphi_{Mn}(x_1)]| &\leq (E[\varphi_{Mn}^2(x_1)])^{1/2} \end{aligned}$$

thus, owing to (60) and (61), we get therefrom that

$$\begin{aligned} |r_{Mn,\varphi}^g(k)| &\leq M_f (M_R \sigma_\zeta |\lambda_k| + |\rho_\xi(k)| + |\rho_z(k)|) \\ &\leq M_f \left(M_R \sigma_\zeta |\lambda_k| + \sigma_\zeta^2 \sum_{i=1}^{\infty} |\lambda_i \lambda_{i+k}| + \sigma_\varepsilon^2 \sum_{i=0}^{\infty} |\omega_i \omega_{i+k}| \right) \end{aligned} \quad (63)$$

Hence (cf. (59))

$$\begin{aligned} |B| &\leq 2 \sum_{k=1}^{N-1} (1 - k/N) |r_{Mn,\varphi}^g(k)| \leq 2 \sum_{k=1}^{\infty} |r_{Mn,\varphi}^g(k)| \\ &\leq 2M_f (c_1 M_R \sigma_\zeta + c_2 \sigma_\zeta^2 + c_3 \sigma_\varepsilon^2) < \infty \end{aligned} \quad (64)$$

as under Assumptions A3 and A4 it holds that $\sum_{k=1}^{\infty} |\lambda_k| = c_1$, $\sum_{k=1}^{\infty} \sum_{i=1}^{\infty} |\lambda_i \lambda_{i+k}| = c_2$ and $\sum_{k=1}^{\infty} \sum_{i=0}^{\infty} |\omega_i \omega_{i+k}| = c_3$ some $0 < c_1, c_2, c_3 < \infty$. Putting together (59), (62) and (64) yields

$$\text{var}(\hat{\alpha}_{Mn}^g) \leq C_\alpha^g N^{-1} \quad (65)$$

where $C_\alpha^g = M_f (\bar{c}_1 M_R \max\{M_R, \sigma_\zeta\} + \bar{c}_2 \sigma_\zeta^2 + \bar{c}_3 \sigma_\varepsilon^2)$ and $\bar{c}_1 = 2c_1 + 3$, $\bar{c}_2 = 2c_2 + 3 \sum_{i=1}^{\infty} \lambda_i^2$, $\bar{c}_3 = 2c_3 + 3 \sum_{i=0}^{\infty} \omega_i^2$. After similar steps, we get (cf. (17))

$$\text{var}(\hat{\beta}_{mn}^g) \leq C_\beta^g N^{-1} \quad (66)$$

where $C_\beta^g = C_\alpha^g$. As regards the coefficients $\hat{\alpha}_{Mn}^f$ and $\hat{\beta}_{mn}^f$ (cf. (18)), from Assumption A2 we obtain immediately

$$\text{var}(\hat{\alpha}_{Mn}^f) = \text{var} \left[N^{-1} \sum_{k=1}^N \varphi_{Mn}(x_k) \right] = N^{-1} \text{var}[\varphi_{Mn}(x_1)]$$

and then using $\text{var}[\varphi_{Mn}(x_1)] \leq E[\varphi_{Mn}^2(x_1)]$ and the bound in (60), we conclude that $\text{var}(\hat{\alpha}_{Mn}^f) \leq M_f N^{-1}$. Analogously, $\text{var}(\hat{\beta}_{mn}^f) \leq M_f N^{-1}$.

APPENDIX C

Consider the variance of $\hat{g}(x; K)$ for $x \in [a, b]$. By virtue of (26), we have

$$\text{var} \{ \hat{g}(x; K) \} = E [\hat{g}(x; K) - g(x; K)]^2 \quad (67)$$

where (cf. (8) and (19))

$$\begin{aligned} \hat{g}(x; K) - g(x; K) &= \sum_{n=n_{\min}(\varphi, x, M)}^{n_{\max}(\varphi, x, M)} (\hat{\alpha}_{Mn}^g - \alpha_{Mn}^g) \varphi_{Mn}(x) \\ &\quad + \sum_{m=M}^{K-1} \sum_{n=n_{\min}(\psi, x, m)}^{n_{\max}(\psi, x, m)} (\hat{\beta}_{mn}^g - \beta_{mn}^g) \psi_{mn}(x) \end{aligned}$$

Owing to (49) in Appendix A, we obtain immediately

$$\begin{aligned} |\hat{g}(x; K) - g(x; K)| &\leq M_\varphi 2^{M/2} \sum_{n=n_{\min}(\varphi, x, M)}^{n_{\max}(\varphi, x, M)} |\hat{\alpha}_{Mn}^g - \alpha_{Mn}^g| \\ &\quad + M_\psi \sum_{m=M}^{K-1} 2^{m/2} \sum_{n=n_{\min}(\psi, x, m)}^{n_{\max}(\psi, x, m)} |\hat{\beta}_{mn}^g - \beta_{mn}^g| \end{aligned}$$

and then (by Cauchy inequality)

$$\begin{aligned} [\hat{g}(x; K) - g(x; K)]^2 &\leq 2M_\varphi^2 2^M \left(\sum_{n=n_{\min}(\varphi, x, M)}^{n_{\max}(\varphi, x, M)} |\hat{\alpha}_{Mn}^g - \alpha_{Mn}^g| \right)^2 \\ &\quad + 2M_\psi^2 \left(\sum_{m=M}^{K-1} 2^{m/2} \sum_{n=n_{\min}(\psi, x, m)}^{n_{\max}(\psi, x, m)} |\hat{\beta}_{mn}^g - \beta_{mn}^g| \right)^2 \\ &= 2(e_1^2 + e_2^2), \quad \text{say} \end{aligned} \quad (68)$$

For the first component, using again Cauchy inequality and including (53) in Appendix A, we get

$$e_1^2 \leq M_\varphi^2 S 2^M \sum_{n=n_{\min}(\varphi, x, M)}^{n_{\max}(\varphi, x, M)} (\hat{\alpha}_{Mn}^g - \alpha_{Mn}^g)^2 \quad (69)$$

For the second component, after appropriate rearrangements, we have

$$e_2^2 = M_\psi^2 \sum_{m=M}^{K-1} 2^{m/2} \sum_{m'=M}^{K-1} 2^{m'/2} \sum_{n=n_{\min}(\psi, x, m)}^{n_{\max}(\psi, x, m)} \sum_{n'=n_{\min}(\psi, x, m')}^{n_{\max}(\psi, x, m')} |\hat{\beta}_{mn}^g - \beta_{mn}^g| |\hat{\beta}_{m'n'}^g - \beta_{m'n'}^g| \quad (70)$$

Now, taking into account (67)-(70), the fact that (see (25) in Section V)

$$\begin{aligned} E (\hat{\alpha}_{Mn}^g - \alpha_{Mn}^g)^2 &= \text{var} (\hat{\alpha}_{Mn}^g) \\ E |\hat{\beta}_{mn}^g - \beta_{mn}^g| |\hat{\beta}_{m'n'}^g - \beta_{m'n'}^g| &\leq \max \{ \text{var} (\hat{\beta}_{mn}^g), \text{var} (\hat{\beta}_{m'n'}^g) \} \end{aligned}$$

and (53) in Appendix A along with the bounds (65) and (66) in Appendix B, we see that for $x \in [a, b]$ it holds that

$$\text{var} \{\hat{g}(x; K)\} \leq 2 S^2 \max \{M_\varphi^2, M_\psi^2\} C_\alpha^g \left[2^M + \left(\sum_{m=M}^{K-1} 2^{m/2} \right)^2 \right] N^{-1}$$

Since

$$2^M + \left(\sum_{m=M}^{K-1} 2^{m/2} \right)^2 = 2^K \left[2^{-(K-M)} + \left(1/(\sqrt{2}-1) \right)^2 \left(1 - 2^{-(K-M)/2} \right)^2 \right]$$

and in the wavelet models $K > M$, we get eventually

$$\text{var} \{\hat{g}(x; K)\} \leq (C_g 2^K) N^{-1} \quad (71)$$

where $C_g = 2 \left(1 + 1/(\sqrt{2}-1) \right)^2 S^2 \max \{M_\varphi^2, M_\psi^2\} C_\alpha^g$, for each point $x \in [a, b]$ and each scale factor K . Because of similarity of the models (9), (20) to those in (8), (19), and similarity of the variance bounds of $\hat{\alpha}_{Mn}^f$ and $\hat{\beta}_{mn}^f$ to those in (65), (66), we obtain after analogous steps that

$$\text{var} \{\hat{f}(x; K)\} \leq (C_f 2^K) N^{-1} \quad (72)$$

where $C_f = 2 \left(1 + 1/(\sqrt{2}-1) \right)^2 S^2 \max \{M_\varphi^2, M_\psi^2\} M_f$, for any $x \in [a, b]$ and any K .

APPENDIX D

Let $\{\vartheta_N\}$ and $\{\eta_N\}$ be sequences of random variables and $\{a_N\}, \{b_N\}$ be sequences of positive numbers such that $a_N, b_N \rightarrow 0$ as $N \rightarrow \infty$.

Lemma 3: If $E\vartheta_N^2 \leq C a_N$ some $C > 0$ independent of N then $\vartheta_N = O(a_N^{1/2})$ in probability.

Proof: From $E\vartheta_N^2 \leq C a_N$, by Chebychev's inequality, $P\{|r_N| (|\vartheta_N|/b_N) > \varepsilon\} \leq (C/\varepsilon^2) r_N^2 (a_N/b_N^2)$ for each $\varepsilon > 0$, i.e. (30) holds for $b_N = a_N^{1/2}$. ■

Lemma 4: If $\vartheta_N = a + O(a_N)$ in probability and $\eta_N = b + O(b_N)$ in probability where $b \neq 0$, then $\vartheta_N/\eta_N = a/b + O(\max\{a_N, b_N\})$ in probability.

Proof: It holds that (see proof of Lemma 2 in [15])

$$\left| \frac{\vartheta_N}{\eta_N} - \frac{a}{b} \right| \leq \left| \frac{\vartheta_N}{\eta_N} \right| \frac{|\eta_N - b|}{|b|} + \frac{|\vartheta_N - a|}{|b|}$$

Assume $|\vartheta_N - a| \leq |a| (\varepsilon/(2+\varepsilon)) (a_N/|r_N|)$ and $|\eta_N - b| \leq |b| (\varepsilon/(2+\varepsilon)) (b_N/|r_N|)$. Making use of the above inequality, we get that

$$\left| \frac{\vartheta_N}{\eta_N} - \frac{a}{b} \right| \left(1 - \left(\frac{\varepsilon}{2+\varepsilon} \right) \frac{b_N}{|r_N|} \right) \leq 2 \left| \frac{a}{b} \right| \left(\frac{\varepsilon}{2+\varepsilon} \right) \frac{\max\{a_N, b_N\}}{|r_N|}$$

and for r_N arbitrarily slowly tending to zero this yields for large N that

$$\left| \frac{\vartheta_N}{\eta_N} - \frac{a}{b} \right| \leq \left| \frac{a}{b} \right| \varepsilon \frac{\max\{a_N, b_N\}}{|r_N|}$$

Therefore

$$P \left\{ |r_N| \left| \frac{\vartheta_N}{\eta_N} - \frac{a}{b} \right| / \max\{a_N, b_N\} > \left| \frac{a}{b} \right| \varepsilon \right\} \leq P \left\{ |r_N| \frac{|\vartheta_N - a|}{a_N} > |a| \left(\frac{\varepsilon}{2 + \varepsilon} \right) \right\} \\ + P \left\{ |r_N| \frac{|\eta_N - b|}{b_N} > |b| \left(\frac{\varepsilon}{2 + \varepsilon} \right) \right\}$$

each $\varepsilon > 0$, which - including (30) in Section V - ends the proof. ■

Acknowledgments

The authors thank the Reviewers for their helpful comments.

REFERENCES

- [1] J. S. Bendat. *Nonlinear System Analysis and Identification*. Wiley, New York, 1990.
- [2] A. Chambolle, R. A. DeVore, N. Y. Lee, and B. J. Lucier. Nonlinear wavelet image-processing - variational-problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing*, 7:319–335, 1998.
- [3] H. W. Chen. Modelling and identification of parallel non-linear systems: Structural classification and parameter estimation methods. *Proceedings of IEEE*, 83:39–66, 1995.
- [4] D. Coca and S. A. Billings. Non-linear system identification using wavelet multiresolution models. *International Journal of Control*, 74:1718–1736, 2001.
- [5] I. Daubechies. *Ten Lectures on Wavelets*. SIAM Edition, Philadelphia, 1992.
- [6] R. A. DeVore and B. J. Lucier. Fast wavelet techniques for near-optimal image processing. In *IEEE Military Communications Conference Record MILCOM'92*, pages 1129–1135, San Diego, California, October 11-14 1992. IEEE Piscataway, NJ.
- [7] D. L. Donoho. De-noising by soft thresholding. *IEEE Transactions on Information Theory*, 41:613–627, 1995.
- [8] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelets shrinkage. *Annals of Statistics*, 26:879–921, 1998.
- [9] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Annals of Statistics*, 24:508–539, 1996.
- [10] H. Y. Gao. Wavelet shrinkage denoising using the non-negative garrote. *Journal of Computer Graphics Statistics*, 7:469–488, 1998.
- [11] R. Ghanem and F. Romeo. A wavelet-based approach for model and parameter identification of non-linear systems. *International Journal of Nonlinear Mechanics*, 36:835–859, 2001.
- [12] G. B. Giannakis and E. Serpedin. A bibliography on nonlinear system identification. *Signal Processing*, 81:533–580, 2001.
- [13] W. Greblicki. Nonparametric orthogonal series identification of Hammerstein systems. *International Journal of Systems Science*, 20:2355–2367, 1989.
- [14] W. Greblicki and A. Krzyżak. Nonparametric identification of a memoryless system with a cascade structure. *International Journal of Systems Science*, 10:1301–1310, 1979.

- [15] W. Greblicki and M. Pawlak. Fourier and Hermite series estimates of regression function. *Annals of The Institute of Statistical Mathematics*, 37:443–455, 1985.
- [16] W. Greblicki and M. Pawlak. Identification of discrete Hammerstein system using kernel regression estimates. *IEEE Transactions on Automatic Control*, 31:74–77, 1986.
- [17] W. Greblicki and M. Pawlak. Nonparametric identification of Hammerstein systems. *IEEE Transactions on Information Theory*, 35:409–418, 1989.
- [18] R. Haber and L. Keviczky. *Nonlinear System Identification: Input-Output Modeling Approach*. Kluwer Academic Publishers, Dordrecht-Boston-London, 1999.
- [19] Z. Hasiewicz. Applicability of least-squares to the parameter estimation of large-scale no-memory linear composite systems. *International Journal of Systems Science*, 20:2427–2449, 1989.
- [20] Z. Hasiewicz. Non-parametric estimation of non-linearity in a cascade time series system by multiscale approximation. *Signal Processing*, 81:791–807, 2001.
- [21] Z. Hasiewicz and P. Śliwiński. Identification of non-linear characteristics of a class of block-oriented non-linear systems via Daubechies wavelet-based models. *International Journal of Systems Science*, 14:1121–1144, 2002.
- [22] M. Jansen. *Noise Reduction by Wavelet Thresholding*. Springer, New York, 2001.
- [23] S. Kelly, M. Kon, and M. Raphael. Pointwise convergence of wavelet expansions. *Bulletin of The American Mathematical Society*, 30:87–94, 1994.
- [24] Y. Kitada. Identification of nonlinear structural dynamic systems using wavelets. *Journal of Engineering Mechanics-ASCE*, 124:1059–1066, 1998.
- [25] L. Ljung. *System Identification - Theory For the User*. Prentice-Hall, Englewood Cliffs, 1987.
- [26] S. G. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 1998.
- [27] M. Pawlak. On the series expansion approach to the identification of Hammerstein systems. *IEEE Transactions on Automatic Control*, 36:763–767, 1991.
- [28] M. Pawlak and Z. Hasiewicz. Nonlinear system identification by the Haar multiresolution analysis. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 45:945–961, 1998.
- [29] C. J. Stone. Optimal rates of convergence for nonparametric regression. *Annals of Statistics*, 8:1348–1360, 1980.
- [30] N. Sureshbabu and J. A. Farrell. Wavelet based system identification for non-linear control. *IEEE Transactions on Automatic Control*, 44:412–417, 1999.
- [31] S. Verdu. *Multiuser Detection*. Cambridge University Press, London, 1998.
- [32] J. Vörös. Iterative algorithm for parameter identification of Hammerstein systems with two-segment nonlinearities. *IEEE Transactions on Automatic Control*, 44:2145–2149, 1999.
- [33] Y. Yang. Mixing strategies for density estimation. *Annals of Statistics*, 28:75–87, 2000.