[3] F. M. Callier and C. A. Desoer, *Linear System Theory*. New York: Springer-Verlag, 1991.
[4] C. A. R. Crusius and A. Trofino, "Sufficient LMI conditions for output feedback control problems," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 1053–1057, June 1999.
[5] C. Edwards and S. K. Spurgeon, "Linear matrix inequality methods for designing sliding mode output feedback controllers," *Proc. Inst. Elect. Eng. Control Theory Applications*, vol. 150, pp. 539–545, 2003.
[6] D. W. Gu and F. W. Poon, "A robust state observer scheme," *IEEE Trans. Automat. Contr.*, vol. 46, pp. 1958–1963, Nov. 2001.
[7] C. H. Kuo and C. H. Fang, "Stabilization of uncertain linear systems via static output feedback control," in *Proc. Automatic Control Conf.*, vol. 1, Taiwan, 2003, pp. 1607–1611.
[8] C. H. Lien and J. D. Chen, "Discrete-delay-independent and discrete-delay-dependent criteria for a class of neutral systems," *ASME J. Dyna. Syst. Measure. Control*, vol. 125, pp. 33–41, 2003.
[9] P. Shi, El. K. Boukas, Y. Shi, and R. K. Agarwal, "Optimal guaranteed cost control of uncertain discrete time-delay systems," *J. Comput. Appl. Math.*, vol. 157, pp. 435–451, 2003.
[10] C. P. Tan and C. Edwards, "An LMI approach for designing sliding mode observers," *Int. J. Control*, vol. 74, pp. 1559–1568, 1998.
[11] P. Zitek, "Anisochronic state observers for hereditary systems," *Int. J. Control*, vol. 71, pp. 581–599, 1998.

Combined Parametric–Nonparametric Identification of Hammerstein Systems

Zygmunt Hasiewicz and Grzegorz Mzyk

***Abstract*—A novel, parametric–nonparametric, methodology for Hammerstein system identification is proposed. Assuming random input and correlated output noise, the parameters of a nonlinear static characteristic and finite impulse-response system dynamics are estimated separately, each in two stages. First, the inner signal is recovered by a nonparametric regression function estimation method (Stage 1) and then system parameters are solved independently by the least squares (Stage 2). Convergence properties of the scheme are established and rates of convergence are given.**

***Index Terms*—Convergence analysis, least squares, nonparametric regression, parameter estimation.**

I. Introduction

The methods elaborated for the identification of the Hammerstein system (Fig. 1) can be roughly divided into two categories: parametric and nonparametric ones. The former assume prior knowledge of the system components up to a finite number of parameters (e.g., a polynomial form of the nonlinear static characteristic $\mu(u)$;; see, for example, [1] and the references therein) while the latter do not impose any specific structure on the system description (e.g., [2]–[4]). In this note, a new parametric–nonparametric technique is proposed to solve the Hammerstein system identification task. Like in a parametric setting, we assume that the nonlinear static characteristic $\mu(u)$ is known with accuracy to the parameters and, further, that the linear dynamic block is a finite impulse-response (FIR) filter of the known order. In the proposed approach the identification is performed in two stages.

Manuscript received June 26, 2003; revised April 27, 2004. Recommended by Associate Editor E. Bai.

The authors are with the Institute of Engineering Cybernetics, Wroclaw University of Technology, 50-372 Wroclaw, Poland (e-mail: zhas@ict.pwr.wroc.pl; grmz@ict.pwr.wroc.pl).

Digital Object Identifier 10.1109/TAC.2004.832662

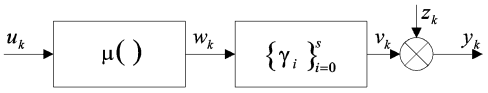

Fig. 1. Hammerstein system.

First, exploiting a nonparametric regression function estimation technique, the unmeasurable inner signal $\{w_k\}$ is estimated from the measurement data (u_k, y_k). Then, the least squares method is used to the independent estimation of the two subsystems parameters using, respectively, the pairs $(u_k, \hat{w}_k)$ and $(\hat{w}_k, y_k)$ where $\{\hat{w}_k\}$ is the estimate of the interaction sequence obtained by a nonparametric method. As compared to the parametric identification techniques developed to date, the potential advantages of the approach are that: 1) we get simple estimates of both subsystems, given by the explicit formulas; 2) the routine is not using any type of alternate updating; 3) the method works with systems having nonpolynomial static characteristics; 4) the algorithm operates efficiently for both white and colored noise, without the need of recovering the noise model; 5) each part of the system is identified separately making the estimates robust against lack or falsity of a priori information about the other part; and 6) convergence properties are established and rates of convergence are given. We refer to [5]–[9] and the references cited therein for representative examples of the parametric identification methods worked out for the Hammerstein system and to [1] for a comprehensive overview of the subject.

II. Statement of the Problem

In the Hammerstein system in Fig. 1, u_k, y_k, and z_k are respectively the input, output, and noise processes at instant k, and w_k is the internal signal (interaction input) not accessible for measurement (see [1] for a discussion). We assume the following.

Assumption 1: The nonlinear static characteristic $\mu(u)$ is known up to a finite number of parameters $c_1, \ldots, c_m$ and is a generalized polynomial [10]

$$\mu(u) = \sum_{l=1}^{m} c_l f_l(u) \tag{1}$$

where $f_1(u), \ldots, f_m(u)$ is a set of known linearly independent basis functions, such that

$$|f_l(u)| \le p_{\max}, \qquad l = 1, 2, \ldots, m \tag{2}$$

some $p_{\max} > 0$ for u in the operation region $|u| \le u_{\max}$ (see Assumption 3).

Assumption 2: Linear dynamics is an FIR filter

$$v_k = \sum_{i=0}^{s} \gamma_i w_{k-i} \tag{3}$$

some finite known order s, with the unknown impulse response $\{\gamma_i\}_{i=0}^{s}$.

Assumption 3: The input signal $\{u_k\}$ is a bounded i.i.d. random process, $|u_k| \le u_{\max}$, some $u_{\max} > 0$.

Assumption 4: The output noise $\{z_k\}$ is a correlated process generated from a bounded zero-mean white noise $\{\varepsilon_k\}$ ($E\varepsilon_k = 0, |\varepsilon_k| \le \varepsilon_{\max}$, some $\varepsilon_{\max} > 0$) by an asymptotically stable linear filter with unknown impulse response $\{\omega_i\}_{i=0}^{\infty} (\sum_{i=0}^{\infty} |\omega_i| < \infty)$

$$z_k = \sum_{i=0}^{\infty} \omega_i \varepsilon_{k-i} \tag{4}$$

independent of the input signal $\{u_k\}$. Consequently, $\{z_k\}$ is a zero-mean and bounded process; $Ez_k = 0$, $|z_k| \leq z_{\max}$ where $z_{\max} = \varepsilon_{\max} \sum_{i=0}^{\infty} |\omega_i| < \infty$.

Assumption 5: $\mu(u_0)$ is known at some point u_0 and $\gamma_0 = 1$.

The last requirement is strictly connected with the method presented in the note and will be explained in Section III.

Owing to (1), (3), and the relationship $w_k = \mu(u_k)$, we get

$$w_k = \phi^T(u_k)c \quad y_k = \vartheta_k^T \gamma + z_k \tag{5}$$

where $c = (c_1, c_2, \ldots, c_m)^T, \gamma = (\gamma_0, \gamma_1, \ldots, \gamma_s)^T, \phi(u_k) = (f_1(u_k), f_2(u_k), \ldots, f_m(u_k))^T$ and $\vartheta_k = (w_k, w_{k-1}, \ldots, w_{k-s})^T$. Our objective is to recover unknown c and γ using input–output measurements $\{(u_k, y_k)\}_{k=1}^M$ of the *whole* system.

III. Background of the Approach

Denote $\Phi_{N_0} = (\phi(u_1), \phi(u_2), \ldots, \phi(u_{N_0}))^T$, $W_{N_0} = (w_1, w_2, \ldots, w_{N_0})^T$ and $\Theta_N = (\vartheta_{1+s}, \vartheta_{2+s}, \ldots, \vartheta_{N+s})^T$, $Y_N = (y_{1+s}, y_{2+s}, \ldots, y_{N+s})^T$, and $Z_N = (z_{1+s}, z_{2+s}, \ldots, z_{N+s})^T$. From (5), we have

$$W_{N_0} = \Phi_{N_0} c \quad Y_N = \Theta_N \gamma + Z_N \tag{6}$$

Hence, we can get the following least squares solution for the parameter vector c:

$$c = \left(\Phi_{N_0}^T \Phi_{N_0}\right)^{-1} \Phi_{N_0}^T W_{N_0} \tag{7}$$

and the least squares estimate of the vector γ

$$\hat{\gamma}_N = \left(\Theta_N^T \Theta_N\right)^{-1} \Theta_N^T Y_N \tag{8}$$

(weakly consistent in the problem in question [11]), provided that Φ_{N_0} and Θ_N are of full-column rank. To this end let us assume that $N_0 \geq m = \dim c$ and $N \geq s + 1 = \dim \gamma$. The obvious difficulty in carrying out this simple approach is that the w_k's entering W_{N_0} and Θ_N cannot be measured. A way to overcome this drawback and utilize the solutions (7) and (8) is to recover the unknown w_k's from the measurement data $\{(u_k, y_k)\}_{k=1}^M$ and to use the obtained estimates $\hat{w}_{k,M}$ instead of w_k. Observing, however, that [see (3) and (4), along with the relations $w_k = \mu(u_k)$ and $y_k = v_k + z_k$]

$$y_k = \gamma_0 \mu(u_k) + \sum_{i=1}^{s} \gamma_i \mu(u_{k-i}) + \sum_{i=0}^{\infty} \omega_i \varepsilon_{k-i}$$

or, equivalently

$$y_k = (\gamma_0 \mu(u_k) + d) + \sum_{i=1}^{s} \gamma_i [\mu(u_{k-i}) - E\mu(u_1)] + \sum_{i=0}^{\infty} \omega_i \varepsilon_{k-i} \tag{9}$$

where $d = E\mu(u_1) \sum_{i=1}^{s} \gamma_i$ and viewing the tail in (9) as the aggregate zero-mean output noise. We realize that only scaled and shifted version $\mu_b(u) = \gamma_0 \mu(u) + d$ of the nonlinear characteristic $\mu(u)$ is at most accessible from the input–output data $\{(u_k, y_k)\}$. Hence, only scaled and shifted values $\gamma_0 w_k + d$ of the interactions $w_k = \mu(u_k)$ can be recovered in a general case. To get consistent estimates $\hat{w}_{k,M}$ of w_k, we require the technical Assumption 5. Then, for $\gamma_0 = 1$, the bias d can be clearly eliminated by taking, for a given u_k, the estimate $\hat{w}_{k,M} = \hat{\mu}_{b,M}(u_k) - \hat{\mu}_{b,M}(u_0) + \mu_0$ where $\hat{\mu}_{b,M}(u)$ is an estimate of the characteristic $\mu_b(u)$. While the knowledge of $\mu(u)$ in at least one point u_0, $\mu(u_0) = \mu_0$, is necessary for removing the bias (unless $E\mu(u_1) = 0$ or $\sum_{i=1}^{s} \gamma_i = 0$) and this requirement cannot be dropped or relaxed, the requirement that $\gamma_0 = 1$ can be weakened to the demand that $\gamma_0 \neq 0$ as the parameter sets $(\gamma_0 c, (1/\gamma_0)\gamma)$ and (c, γ) are equivalent (indistinguishable) from the input–output data point of view for each $\gamma_0 \neq 0$ [see (9) and [12] and [13] for the discussion of this issue.] Introducing, in particular, the regression function

$$R(u) = E\{y_k \mid u_k = u\} \tag{10}$$

we get for $\gamma_0 = 1$ that $R(u) = \mu(u) + d$ [cf. (9) and Assumptions 3 and 4] and further $R(u) - R(u_0) = \mu(u) - \mu(u_0)$ or $\mu(u) = R(u) - R(u_0) + \mu(u_0)$. For technical reasons, without any loss of generality, we assume, henceforth, $u_0 = 0$ and $\mu(0) = 0$ getting

$$\mu(u) = R(u) - R(0). \tag{11}$$

The fundamental formula (11) relating the nonlinearity $\mu(u)$ with the regression function $R(u)$ was given in [3]. We note that $R(u) = \mu_b(u)$.

Equations (10) and (11) along with $w_k = \mu(u_k)$ suggest that estimation of interactions w_k from the data $\{(u_k, y_k)\}_{k=1}^M$ can in particular be performed by a nonparametric regression function estimation method, without the use of parametric prior knowledge of the system components. The class of nonparametric regression function estimates elaborated for the Hammerstein system to date comprises kernel estimates [3], [14], orthogonal series estimates [2], and wavelet estimates [15], [4]. For a general treatment of nonparametric regression function estimation methods, see, for example, [16]–[19].

Remark 1: The idea of recovering unmeasurable inner signal in the Hammerstein and Wiener system has been recently realized in [12], [13] by using frequency domain method and a nonparametric method to support the recovery from independent data of the linear regression parameters was used in [20].

IV. Two-Stage Estimation of Nonlinearity Parameters

Due to the aforementioned information, we propose the following two-stage estimation procedure of vector c.

Stage 1 (Nonparametric): Using input-output data $\{(u_k, y_k)\}_{k=1}^M$, for a set of input points $\{u_n; n = 1, 2, \ldots, N_0\}$ such that $M > N_0 \geq m = \dim c$ and $\Phi_{N_0} = (\phi(u_1), \phi(u_2), \ldots, \phi(u_{N_0}))^T$ is of full-column rank, estimate the corresponding interactions $\{w_n = \mu(u_n); n = 1, 2, \ldots, N_0\}$ as

$$\hat{w}_{n,M} = \hat{R}_M(u_n) - \hat{R}_M(0) \tag{12}$$

where $\hat{R}_M(u)$ is a nonparametric estimate of the regression function $R(u)$, computed for $u \in \{0, u_n; n = 1, 2, \ldots, N_0\}$. Here, the estimation points $\{u_n\}_{n=1}^{N_0}$ can be the measured input data or the points freely selected by the experimenter; we do not distinguish these two situations because of no formal importance of the difference.

Stage 2 (Parametric): Compute the estimate of parameter vector c as [cf. (7)]

$$\hat{c}_{N_0,M} = \left(\Phi_{N_0}^T \Phi_{N_0}\right)^{-1} \Phi_{N_0}^T \hat{W}_{N_0,M} \tag{13}$$

where Φ_{N_0} and $\hat{W}_{N_0,M} = (\hat{w}_{1,M}, \hat{w}_{2,M}, \ldots, \hat{w}_{N_0,M})^T$ are established in Stage 1.

Remark 2: The requirement that $\operatorname{rank}\Phi_{N_0} = m = \dim c$ can be fulfilled because of linear independence of $f_1(u), \ldots, f_m(u)$

and the fact that the estimation points u_n in Stage 1 may in particular be selected in an arbitrary manner. Such a condition is for instance automatically satisfied if $N_0 \geq m$, $u_1, u_2, \ldots, u_{N_0}$ are distinct points, and $f_1(u), \ldots, f_m(u)$ in (1) is a Tchebycheff system (i.e., satisfies the Haar condition). Appropriate examples are $\{1, u, u^2, \ldots, u^{m-1}\}$ [which yields standard polynomial characteristic in (1)], $\{1, \sin u, \sin 2u, \ldots, \sin(m-1)u\}$ (on $[0, 2\pi]$) or $\{e^{\lambda_1 u}, e^{\lambda_2 u}, \ldots, e^{\lambda_m u}\}$ and nondegenerate linear combinations of these functions [10].

Defining the estimation error in Stage 1 as $\varsigma_{n,M} = \hat{w}_{n,M} - w_n$ and including (5), we see that Stage 2 may be considered as the identification of a static element $\hat{w}_{n,M} = \phi^T(u_n)c + \varsigma_{n,M}$ from the "data" $\{(u_n, \hat{w}_{n,M})\}_{n=1}^{N_0}$ by means of least squares. We note that estimation of vector c in (13) is performed without the use of *a priori* knowledge of the system dynamics, in contrast to the alternate updating methods [5], [6], [9].

The following theorem holds.

Theorem 1: If for $u \in \{0, u_n; n = 1, 2, \ldots, N_0\}$ it holds that $\hat{R}_M(u) \to R(u)$ in probability as $M \to \infty$ (Stage 1) then $\hat{c}_{N_0,M} \to c$ in probability as $M \to \infty$ (Stage 2).

Proof: See Appendix A.

Remark 3: Any kind of probabilistic convergence can be considered in Theorem 1. We examine convergence in probability as this particular type of convergence is usually studied in the convergence analysis of nonparametric regression function estimates for Hammerstein systems (see [2], [3], [14], and [4]).

In the next theorem establishing asymptotic rate of convergence of the estimate $\hat{c}_{N_0,M}$, for a sequence of random variables $\{\varsigma_M\}$ and positive number sequence $\{a_M\}$ convergent to zero, $\varsigma_M = O(a_M)$ in probability means that $d_M \varsigma_M / a_M \to 0$ in probability as $M \to \infty$ for any number sequence $\{d_M\}$ tending to zero [14, p. 140].

Theorem 2: If in Stage 1 $|\hat{R}_M(u) - R(u)| = O(M^{-\tau})$ in probability as $M \to \infty$ for each $u \in \{0, u_n; n = 1, 2, \ldots, N_0\}$ then also $\|\hat{c}_{N_0,M} - c\| = O(M^{-\tau})$ in probability as $M \to \infty$.

Proof: See Appendix B.

As we see, in the method the rate of convergence for estimating vector c is determined by the nonparametric rates for the regression function. Since, as is well known, for each nonparametric method $0 < \tau < 1/2$ (e.g., [17]), the rate $O(M^{-\tau})$ is in general of slower order than the best possible parametric rate of convergence $O(M^{-1/2})$ in probability. However, for polynomial and other smooth nonlinearities the convergence rate can be made arbitrarily close to $O(M^{-1/2})$ by applying nonparametric estimates being able to adapt to smooth functions, e.g., wavelet estimates [15], [4].

V. Two-Stage Identification of Linear Dynamics

Considering (8), we propose similar two-stage scheme to estimate γ.

Stage 1 (Nonparametric): Using input–output measurement data $\{(u_k, y_k)\}_{k=1}^{M}$, estimate the entries $\{w_{t-r}; t = n + s; n = 1, 2, \ldots, N; r = 0, 1, \ldots, s\}$ of Θ_N by a nonparametric method, i.e., as $\hat{w}_{t-r,M} = \hat{R}_M(u_{t-r}) - \hat{R}_M(0)$ where $\{u_{t-r}\}$ are the input data points corresponding to the output measurements $\{y_t\}_{t=1+s}^{N+s}$ collected in vector Y_N [see (6)] and $\hat{R}_M(u)$ is a nonparametric estimate of the regression function $R(u)$ computed for $u \in \{0, u_{t-r}; t = n + s; n = 1, 2, \ldots, N; r = 0, 1, \ldots, s\}$ [cf. (10)–(12)].

Stage 2 (Parametric): Compute [see (8)]

$$\hat{\gamma}_{N,M} = \left(\hat{\Theta}_{N,M}^T \hat{\Theta}_{N,M}\right)^{-1} \hat{\Theta}_{N,M}^T Y_N \tag{14}$$

where $\hat{\Theta}_{N,M} = (\hat{\vartheta}_{1+s,M}, \hat{\vartheta}_{2+s,M}, \ldots, \hat{\vartheta}_{N+s,M})^T$, $\hat{\vartheta}_{t,M} = (\hat{w}_{t,M}, \hat{w}_{t-1,M}, \ldots, \hat{w}_{t-s,M})^T$ and Y_N is a noisy output vector.

In Stage 1, we assume that $M - s > N \geq s + 1 = \dim \gamma$. We accentuate that estimation of γ in (14) is independent of estimating parameter vector c of the static subsystem.

Remark 4: Recalling that the "input" data are contaminated by the additive (estimation) errors, $\hat{w}_{t-r,M} = w_{t-r} + \varsigma_{t-r,M}$ (cf. Section IV), we note that recovering of γ from the data $(\hat{\Theta}_{N,M}, Y_N)$ is in fact the errors-in-variables estimation problem. This is unlike the aforementioned [20], where input data are accurate and only output measurements are corrupted by the noise (white).

The following theorem holds.

Theorem 3: If for $u \in \{0, u_{t-r}; t = n + s; n = 1, 2, \ldots, N; r = 0, 1, \ldots, s\}$ the estimate $\hat{R}_M(u)$ is bounded and the asymptotic nonparametric estimation error in Stage 1 behaves like

$$|\hat{R}_M(u) - R(u)| = O(M^{-\tau}) \text{ in probability} \tag{15}$$

then $\hat{\gamma}_{N,M} \to \gamma$ in probability in Stage 2 provided that $N, M \to \infty$ and $NM^{-\tau} \to 0$.

Proof: See Appendix C.

The condition $NM^{-\tau} \to 0$ is fulfilled if $M = \text{const} \cdot N^{(1+\alpha)/\tau}$, or equivalently $N = \text{const} \cdot M^{\tau/(1+\alpha)}$, any $\alpha > 0$. It is noteworthy that since in general $0 < \tau < 1/2$ (cf. Section IV), to get consistency of the estimate $\hat{\gamma}_{N,M}$ far more data points $\{(u_k, y_k)\}_{k=1}^{M}$ must be used in Stage 1 for nonparametric estimation of interactions $\{w_{t-r}\}$ than the "observations" $\{(\hat{\vartheta}_{t,M}, y_t)\}_{t=1+s}^{N+s}$ in Stage 2 for computing $\hat{\gamma}_{N,M}$. This can be explained by the necessity of effective reduction of the 'input' errors in $\hat{\vartheta}_{t,M}$'s and slower convergence of nonparametric methods. From the data length M viewpoint, small α are clearly preferred. In contrast, $\alpha \geq 1/2$ are desirable from the $\hat{\gamma}_{N,M}$ convergence rate point of view.

Theorem 4: For $M \sim N^{(1+\alpha)/\tau}$, equivalently $N \sim M^{\tau/(1+\alpha)}$, $\alpha > 0$, the asymptotic convergence rate in Stage 2 is $\|\hat{\gamma}_{N,M} - \gamma\| = O(N^{-\min(1/2,\alpha)})$ in probability.

Proof: See Appendix D.

If $\alpha \geq 1/2$ we attain for the estimate $\hat{\gamma}_{N,M}$ the best possible parametric rate of convergence $O(N^{-1/2})$ in probability (w.r.t. N). This means that for $\alpha \geq 1/2$ the influence of the input (estimation) errors in $\hat{\vartheta}_{t,M}$'s on the accuracy of the estimate $\hat{\gamma}_{N,M}$ is dominated by the standard effect of the output measurement noise. In the case of $\alpha < 1/2$, we get slower guaranteed convergence rate of order $O(N^{-\alpha})$.

VI. Example

Let us use, for example, in Stage 1 of the schemes the kernel regression estimate studied in [3] and [14], i.e.,

$$\hat{R}_M(u) = \frac{\sum_{k=1}^{M} y_k K\left(\frac{u-u_k}{h(M)}\right)}{\sum_{k=1}^{M} K\left(\frac{u-u_k}{h(M)}\right)} \tag{16}$$

where $K(u)$ is a kernel function and $h(M)$ is a bandwidth parameter. Standard examples are $K(u) = I_{[-0.5,0.5]}(u)$, $(1 - |u|)I_{[-1,1]}(u)$ or $(1/\sqrt{2\pi})e^{-u^2/2}$ and $h(M) = \text{const} \cdot M^{-\alpha}$ with a positive constant and $0 < \alpha < 1$; see [19]. Owing to the convergence results presented there, we find out that for each of the aforementioned $K(u)$ it holds that $\hat{R}_M(u) \to R(u)$ in probability as $M \to \infty$, and that the convergence takes place at every $u \in \text{Cont}(\mu, \nu)$, the set of continuity points of $\mu(u)$ and $\nu(u)$, at which $\nu(u) > 0$ where $\nu(u)$ is a probability density function of the system input (assumed to exist). Taking in particular the Gaussian kernel $K(u) = (1/\sqrt{2\pi})e^{-u^2/2}$ and, according to the recommendation in [14], $h(M) \sim M^{-1/5}$ we get the convergence rate $|\hat{R}_M(u) - R(u)| = O(M^{-2/5})$ in probability and hence $\|\hat{c}_{N_0,M} - c\| = O(M^{-2/5})$ in probability ($\tau = 2/5$) in Stage 2 for a static element, provided that $\mu(u)$ and $\nu(u)$ are at

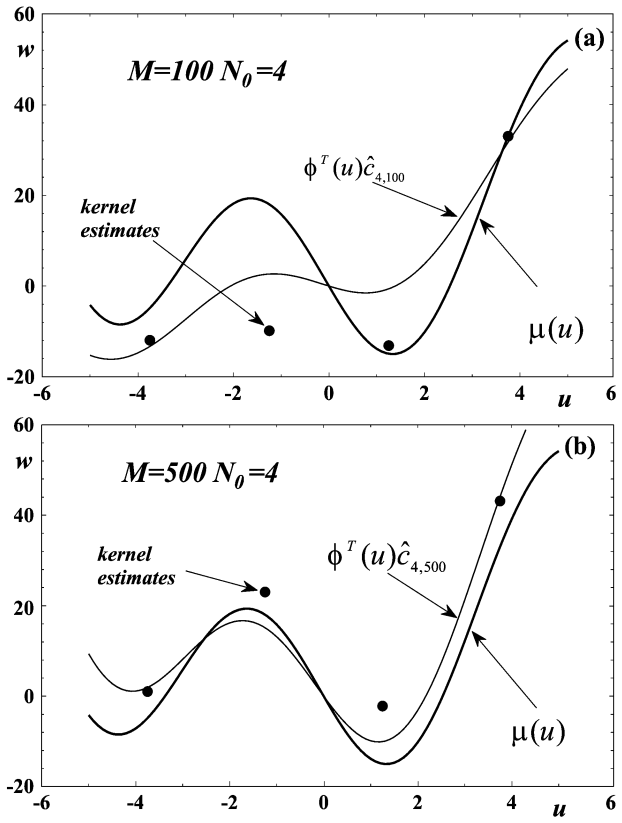

Fig. 2. True versus estimated static characteristic and the "data" points $(u_n, \hat{w}_{n,M})$ (bold-faced) computed in Stage 1 by the kernel nonparametric method for (a) $M = 100$ and (b) $M = 500$ measurement data.

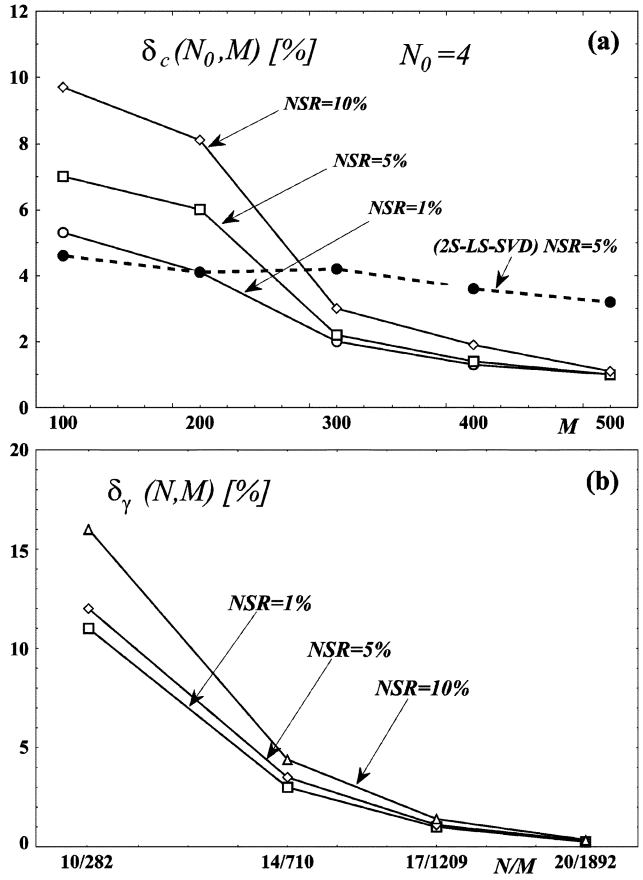

Fig. 3. Average relative estimation error versus number of data for (a) vector c (dashed line: 2S-LS-SVD algorithm) and (b) vector γ.

$u \in \{0, u_n; n = 1, 2, \ldots, N_0\}$ at least two times continuously differentiable functions and $\nu(u) > 0$ there (cf. [3], [14], and Theorem 2). As regards a dynamical part, since in the problem in question the estimate (16) with the Gaussian kernel is bounded (see (5) and Assumptions 1–4), using this estimate we get in Stage 2 the convergence $\hat{\gamma}_{N,M} \to \gamma$ in probability as $N, M \to \infty$ provided that $M \sim N^{5(1+\alpha)/2}$, or equivalently $N \sim M^{2/5(1+\alpha)}$, $\alpha > 0$ (cf. Theorem 3). For $\alpha = 1/2$ the asymptotic convergence rate is $\|\hat{\gamma}_{N,M} - \gamma\| = O(N^{-1/2})$ in probability (Theorem 4). The behavior of the estimates $\hat{c}_{N_0,M}$ and $\hat{\gamma}_{N,M}$ for small and moderate number of data is demonstrated in the simulation example.

VII. Simulation Study

The Hammerstein system from (5), with nonpolynomial static characteristic, was simulated for $m = 3$ and $s = 2$ taking $\phi(u_k) = (u_k, u_k^2, \sin u_k)^T$, $c = (2, 1, -20)^T$, $\gamma = (1, -1, 1)^T$, and $z_k = 2^{-1} z_{k-1} + \varepsilon_k$, i.e., $\{\omega_i = 2^{-i}\}_{i=0}^{\infty}$ [cf. (4)]. Random input and white noise processes $\{u_k\}$ and $\{\varepsilon_k\}$ were generated according to the uniform distributions $u_k \sim U[-5; 5]$ and $\varepsilon_k \sim U[-\varepsilon_{\max}; \varepsilon_{\max}]$ (cf. Assumptions 3 and 4) where $\varepsilon_{\max}$ was changed as to give the noise-to-signal ratio $\mathrm{NSR} = (z_{\max}/v_{\max}) \cdot 100\%$ equal to 1%, 5%, and 10%, where $z_{\max} = \varepsilon_{\max} \sum_{i=0}^{\infty} 2^{-i} = 2\varepsilon_{\max}$ is the noise magnitude (Assumption 4) and $v_{\max} = w_{\max} \sum_{i=0}^{2} |\gamma_i| = 3w_{\max}$ with $w_{\max} = \max_{u_k \in [-5;5]} |\phi^T(u_k)c|$ [cf. (3) and (5)] is the magnitude of the noiseless output signal; in our experiment $v_{\max} = 165$. To estimate interactions in Stage 1, the nonparametric kernel estimate outlined in Section VI was applied, with the Gaussian kernel $K(u) = (1/\sqrt{2\pi})e^{-u^2/2}$ and the globally optimal bandwidth $h(M) = e(\varepsilon_{\max})M^{-1/5}$, $e(\varepsilon_{\max}) = (8.873 + 0.006\varepsilon_{\max}^2)^{1/5}$, computed according to the rule recommended in [14, Sec. 8, p. 145]. For the identification of the static element we assumed $N_0 = 4$ and the estimation points u_n in Stage 1 were chosen arbitrarily as $u_1 = -3.75$, $u_2 = -1.25$, $u_3 = 1.25$, and $u_4 = 3.75$. For the identification of the dynamic part, we put $M = \lceil 0.5 \cdot N^{5(1+0.1)/2} \rceil = \lceil 0.5 \cdot N^{2.75} \rceil$, i.e., $N = \lceil (2M)^{1/2.75} \rceil = \lceil (2M)^{0.36} \rceil$ (cf. Section VI). For each number M of data the experiment was repeated $P = 10$ times, and accuracy of the estimates $\hat{c}_{N_0,M}$ and $\hat{\gamma}_{N,M}$ was evaluated using the average relative estimation error $\delta_\theta(N, M) = [(1/P)\sum_{p=1}^{P} \|\hat{\theta}_{N,M}^{(p)} - \theta\|_2^2 / \|\theta\|_2^2] \cdot 100\%$ where $\hat{\theta}_{N,M}^{(p)}$ is the estimate of $\theta \in \{c, \gamma\}$ obtained in the pth run, and $\| \cdot \|_2$ is the Euclidean vector norm. Exemplary results of two-stage identification of the nonlinear static characteristic for $M = 100$ and $M = 500$ measurement data and $\mathrm{NSR} = 5\%$, along with the true characteristic and the "data" points $\{(u_n, \hat{w}_{n,M})\}_{n=1}^{N_0=4}$ (bold-faced) computed in Stage 1 by the kernel method, are visualized for a single trial in Fig. 2. The bold-faced points are the only outcome of the nonparametric estimation in the scheme. Thanks to the parametric prior knowledge of the characteristic we are able to derive from this small set of data the models $\hat{\mu}(u) = \phi^T(u)\hat{c}_{N_0,M}$ $(N_0 = 4)$ of good visual quality in the whole range of inputs (particularly for $M = 500$), which is beyond the reach of any nonparametric method.

The estimation error of vector c and γ is plotted in Fig. 3. The simulation results show convergence of the estimates $\hat{c}_{N_0,M}$ and $\hat{\gamma}_{N,M}$ with growing number M of measurement data. For comparison, the estimate of vector c was also computed by using the optimal two-stage least-squares SVD (2S-LS-SVD) algorithm [8], after obvious adaptation to Hammerstein systems. We assumed however incorrect prior knowledge of the system dynamics taking erroneously in the computations the dynamics order $s = 1$ instead of the true $s = 2$. Such a

disagreement has of course no influence on the estimate $\hat{c}_{N_0,M}$ in our method. The average relative estimation error for the 2S-LS-SVD algorithm and $\mathrm{NSR} = 5\%$ is depicted in Fig. 3(a) with dashed line. As we see, the 2S-LS-SVD method gives in such circumstances biased estimate of c, with about 3% systematic error.

VIII. Conclusion

Two stage parametric–nonparametric schemes for separate identification of parameters of a nonlinear static and linear FIR dynamic part of Hammerstein system have been proposed and examined. In each of the schemes, only standard nonparametric regression and least squares computations are needed. Because of parametric prior knowledge the estimated models are provided in the explicit form. For a static part, the method works with nonpolynomial characteristics, thus extending the standard range of parametric methodology. The strength of the approach is complete decoupling of identification of the subsystems, computational simplicity, and robustness against lack or incorrectness of *a priori* knowledge of the noise model. Moreover, for each part of the system, the method is immune to possible errors in prior knowledge of companion subsystem. These advantages are achieved at the expense of a bit slower than the best possible convergence rate of the obtained parameter estimates, which is a consequence of the leading role of nonparametric regression estimation in the scheme and slower convergence of nonparametric methods.

Appendix

A. Proof of Theorem 1

Denote $A_{N_0} = (\Phi_{N_0}^T \Phi_{N_0})^{-1} \Phi_{N_0}^T$. Since by (2) in Assumption 1, for each vector norm $\|\cdot\|$ and the induced matrix norm it holds that $\|A_{N_0}\| \le C$, some $C > 0$, thus [see (7) and (13)] $\|\hat{c}_{N_0,M} - c\| \le C\|\hat{W}_{N_0,M} - W_{N_0}\|$. By equivalence of vector norms, we get further $\|\hat{W}_{N_0,M} - W_{N_0}\| \le \alpha \|\hat{W}_{N_0,M} - W_{N_0}\|_1$, some $\alpha > 0$, where $\|x\|_1 \triangleq \sum_{i=1}^{\dim x} |x[i]|$.

Consequently [cf. (11) and (12)]

$$\|\hat{c}_{N_0,M} - c\| \le \alpha C \sum_{n=1}^{N_0} (|\hat{R}_M(u_n) - R(u_n)| + |\hat{R}_M(0) - R(0)|).$$

Hence

$$\begin{aligned} &P\{\|\hat{c}_{N_0,M} - c\| > \varepsilon\} \\ &\quad \le \sum_{n=1}^{N_0} P\{|\hat{R}_M(u_n) - R(u_n)| > \varepsilon / 2N_0 \alpha C\} \\ &\qquad + N_0 P\{|\hat{R}_M(0) - R(0)| > \varepsilon / 2N_0 \alpha C\} \end{aligned} \tag{17}$$

for each $\varepsilon > 0$. Since N_0 is finite and by assumption $|\hat{R}_M(u) - R(u)| \to 0$ in probability as $M \to \infty$ for each $u \in \{0, u_n; n = 1, 2, \ldots, N_0\}$, we get the conclusion.

B. Proof of Theorem 2

Denote $\varepsilon = \delta M^{-\tau} / |d_M|$, any $\delta > 0$. From (17) in Appendix A, we obtain

$$\begin{aligned} &P\left\{|d_M| \, \|\hat{c}_{N_0,M} - c\| / M^{-\tau} > \delta\right\} \\ &\quad \le \sum_{n=1}^{N_0} P\{|d_M| |\hat{R}_M(u_n) - R(u_n)| / M^{-\tau} > \delta / 2N_0 \alpha C\} \\ &\qquad + N_0 P\{|d_M| |\hat{R}_M(0) - R(0)| / M^{-\tau} > \delta / 2N_0 \alpha C\} \end{aligned}$$

which gives the conclusion.

C. Proof of Theorem 3

For each vector norm $\|\cdot\|$, we have

$$\|\hat{\gamma}_{N,M} - \gamma\| \le \|\hat{\gamma}_{N,M} - \hat{\gamma}_N\| + \|\hat{\gamma}_N - \gamma\|. \tag{18}$$

Since [11] $\|\hat{\gamma}_N - \gamma\| \to 0$ in probability as $N \to \infty$, it remains to show that $\|\hat{\gamma}_{N,M} - \hat{\gamma}_N\| \to 0$ in probability as $N, M \to \infty$, provided that $NM^{-\tau} \to 0$. Owing to equivalence of norms, we further use the 1-vector norm $\|\cdot\|_1$ as in Appendix A. Let

$$\begin{aligned} \xi_{N,M} &\triangleq \left\| \frac{1}{N} \hat{\Theta}_{N,M}^T \hat{\Theta}_{N,M} - \frac{1}{N} \Theta_N^T \Theta_N \right\|_{1,1} \\ \varkappa_{N,M} &\triangleq \left\| \left(\frac{1}{N} \hat{\Theta}_{N,M}^T \hat{\Theta}_{N,M} \right)^{-1} - \left(\frac{1}{N} \Theta_N^T \Theta_N \right)^{-1} \right\|_{1,1} \\ \chi_{N,M} &\triangleq \left\| \frac{1}{N} \hat{\Theta}_{N,M}^T Y_N - \frac{1}{N} \Theta_N^T Y_N \right\|_1 \end{aligned}$$

and, moreover, $A_N = ((1/N)\Theta_N^T \Theta_N)^{-1}$, $A = (E\vartheta_1 \vartheta_1^T)^{-1}$, $\varepsilon_N = 1/\|A_N\|_{1,1}$, $r_{N,M} = \xi_{N,M}/(\varepsilon_N(\varepsilon_N - \xi_{N,M}))$ where $\|\cdot\|_{1,1}$ is the matrix norm induced by $\|\cdot\|_1$, and let ϑ_{nl} and $\hat{\vartheta}_{nl,M} (n = 1, 2, \ldots, N; l = 1, 2, \ldots, s+1)$ be the elements of Θ_N, $\hat{\Theta}_{N,M} \in R^{N \times (s+1)}$. Since $\vartheta_{nl} = w_{(n+s)-(l-1)}$, $\hat{\vartheta}_{nl,M} = \hat{w}_{(n+s)-(l-1),M}$ and

$$\begin{aligned} |\hat{\vartheta}_{nl,M} - \vartheta_{nl}| &\le |\hat{R}_M(u_{(n+s)-(l-1)}) - R(u_{(n+s)-(l-1)})| \\ &\quad + |\hat{R}_M(0) - R(0)| \end{aligned} \tag{19}$$

thus owing to (15) and using the Banach inverse map theorem, we get for M large

$$\varkappa_{N,M} \le r_{N,M}. \tag{20}$$

In turn, because the w_n's are bounded, $|w_n| \le a$, some $a > 0$, by the ergodic law of large numbers (e.g., [11, Lemma B.1]) $\varepsilon_N \to \bar{a}$ in probability as $N \to \infty$, where $\bar{a} = 1/\|A\|_{1,1}$ [see (5) and Assumptions 1 and 3]. Further, after standard calculations we obtain

$$\begin{aligned} \xi_{N,M} &\le (s+1) \max_{n,l} |\hat{\vartheta}_{nl,M} - \vartheta_{nl}|^2 \\ &\quad + 2a(s+1) \max_{n,l} |\hat{\vartheta}_{nl,M} - \vartheta_{nl}| \\ \chi_{N,M} &\le b(s+1) \max_{n,l} |\hat{\vartheta}_{nl,M} - \vartheta_{nl}| \end{aligned}$$

as $|y_{n+s}| \le b$, some $b > 0$ [see (5) and Assumptions 1–4]. This yields, respectively

$$\begin{aligned} P\{\xi_{N,M} > \varepsilon\} &\le \sum_{n,l} P\left\{ |\hat{\vartheta}_{nl,M} - \vartheta_{nl}| > \sqrt{\frac{\varepsilon}{2(s+1)}} \right\} \\ &\quad + \sum_{n,l} P\left\{ |\hat{\vartheta}_{nl,M} - \vartheta_{nl}| > \frac{\varepsilon}{4a(s+1)} \right\} \end{aligned} \tag{21}$$

$$P(\chi_{N,M} > \varepsilon) \le \sum_{n,l} P\{|\hat{\vartheta}_{nl,M} - \vartheta_{nl}| > \varepsilon / (b(s+1))\}. \tag{22}$$

Using Markov's inequality, including (19) and the fact that (by boundedness of $\hat{R}_M$) for $u \in \{0, u_{(n+s)-(l-1)}; n =$

$1,2,\ldots,N; l = 1,2,\ldots,s+1\}$ the rate (15) implies asymptotically $E|\hat{R}_M(u) - R(u)| = O(M^{-\tau})$, we observe that the right-hand side of (21) and (22) is bounded by (23) and (24)

$$C(s+1)^2(\sqrt{2/(\varepsilon(s+1))} + 4a/\varepsilon)NM^{-\tau} \quad (23)$$

$$C(s+1)^2(b/\varepsilon)NM^{-\tau} \quad (24)$$

some $0 < C < \infty$. These two expressions tend to zero as $N, M \to \infty$ for each $\varepsilon > 0$, provided that $NM^{-\tau} \to 0$. The latter along with (20) concludes the proof [see (8) and (14)].

D. Proof of Theorem 4

It holds that (here $\|\cdot\|$ means the 1-vector norm $\|\cdot\|_1$ or, respectively, the induced matrix norm $\|\cdot\|_{1,1}$ as in Appendix C)

$$\|\hat{\gamma}_{N,M} - \hat{\gamma}_N\| \le \varkappa_{N,M}\chi_{N,M} + \varkappa_{N,M}\|b_N - b\| + \chi_{N,M}\|A_N - A\| + d_1\varkappa_{N,M} + d_2\chi_{N,M} \quad (25)$$

where $\varkappa_{N,M}$, $\chi_{N,M}$, A_N, A are as in Appendix C and $b_N = \frac{1}{N}\Theta_N^T Y_N$, $b = E\vartheta_1 y_1$, $d_1 = \|b\|$, $d_2 = \|A\|$. Owing to (21)–(24), for $M \sim N^{(1+\alpha)/\tau}$, $\alpha > 0$, we get

$$\xi_{N,M} = O(N^{-\alpha}) \text{ in probability} \quad (26)$$

$$\chi_{N,M} = O(N^{-\alpha}) \text{ in probability.} \quad (27)$$

Including that $r_{N,M} = \xi_{N,M}/(\varepsilon_N(\varepsilon_N - \xi_{N,M}))$ (see Appendix C) and using (26), the rate $\varepsilon_N = \bar{a} + O(N^{-1/2})$ in probability and Lemma 5 and 6 in Appendix E, we obtain $r_{N,M} = O(N^{-\min(1/2,\alpha)})$ in probability. In view of (20), this yields

$$\varkappa_{N,M} = O\left(N^{-\min(1/2,\alpha)}\right) \text{ in probability.} \quad (28)$$

Now, considering that $\|A_N - A\| = O(N^{-1/2})$ in probability, $\|b_N - b\| = O(N^{-1/2})$ in probability along with (27) and (28), and taking into account Lemma 5 in Appendix E, we conclude [see (25)]

$$\|\hat{\gamma}_{N,M} - \hat{\gamma}_N\| = O\left(N^{-\min(1/2,\alpha)}\right) \text{ in probability.} \quad (29)$$

Since $\|\hat{\gamma}_N - \gamma\| = O(N^{-1/2})$ in probability, thus owing to (18), (29), and using again Lemma 5, we obtain eventually $\|\hat{\gamma}_{N,M} - \gamma\| = O(N^{-\min(1/2,\alpha)})$ in probability. Due to equivalence of norms, the conclusion holds for each norm $\|\cdot\|$.

E. Technical Lemmas

Lemma 5: If $\alpha_N = O(a_N)$ in probability and $\beta_N = O(b_N)$ in probability then $\alpha_N\beta_N = O(\max\{a_N^2, b_N^2\})$ in probability and $\alpha_N + \beta_N = O(\max\{a_N, b_N\})$ in probability, where $\{a_N\}$, $\{b_N\}$ are positive number sequences convergent to zero.

Proof: The inequality $|a||b| \le (1/2)(a^2 + b^2)$ implies $|\alpha_N\beta_N|/\max\{a_N^2, b_N^2\} \le (1/2)(\alpha_N^2/a_N^2 + \beta_N^2/b_N^2)$ which yields

$$P\left\{|d_N|\frac{|\alpha_N\beta_N|}{\max\{a_N^2, b_N^2\}} > \varepsilon\right\} \le P\left\{|d_N|^{\frac{1}{2}}\frac{|\alpha_N|}{a_N} > \sqrt{\varepsilon}\right\} + P\left\{|d_N|^{\frac{1}{2}}\frac{|\beta_N|}{b_N} > \sqrt{\varepsilon}\right\}$$

for each $\varepsilon > 0$. Similarly, $|\alpha_N + \beta_N|/\max\{a_N, b_N\} \le |\alpha_N|/a_N + |\beta_N|/b_N$ and, hence

$$P\left\{|d_N|\frac{|\alpha_N + \beta_N|}{\max\{a_N, b_N\}} > \varepsilon\right\} \le P\left\{|d_N|\frac{|\alpha_N|}{a_N} > \frac{\varepsilon}{2}\right\} + P\left\{|d_N|\frac{|\beta_N|}{b_N} > \frac{\varepsilon}{2}\right\}$$

each $\varepsilon > 0$, which ends the proof.

Lemma 6: [15] If $\alpha_N = \alpha + O(a_N)$ in probability, $\beta_N = \beta + O(b_N)$ in probability and $\beta \ne 0$ then $\alpha_N/\beta_N = \alpha/\beta + O(\max\{a_N, b_N\})$ in probability.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their helpful comments and suggestions.

REFERENCES

[1] R. Haber and L. Keviczky, *Nonlinear System Identification: Input-Output Modeling Approach*. Norwell, MA: Kluwer, 1999.
[2] W. Greblicki, "Nonparametric orthogonal series identification of Hammerstein systems," *Int. J. Syst. Sci.*, vol. 20, pp. 2355–2367, 1989.
[3] W. Greblicki and M. Pawlak, "Identification of discrete Hammerstein systems using kernel regression estimates," *IEEE Trans. Automat.Contr.*, vol. AC-31, pp. 74–77, Jan. 1986.
[4] Z. Hasiewicz and P. Śliwiński, "Identification of nonlinear characteristics of a class of block-oriented nonlinear systems via Daubechies wavelet-based models," *Int. J. Syst. Sci.*, vol. 33, no. 14, pp. 1121–1144, 2002.
[5] K. S. Narendra and P. G. Gallman, "An iterative method for the identification of nonlinear systems using the Hammerstein model," *IEEE Trans. Automat. Contr.*, vol. AC-11, pp. 546–550, Aug. 1966.
[6] N. D. Haist, F. H. Chang, and R. Luus, "Nonlinear identification in the presence of correlated noise using Hammerstein model," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 552–555, Aug. 1973.
[7] S. A. Billings and S. Y. Fakhouri, "Identification of systems containing linear dynamic and static nonlinear elements," *Automatica*, vol. 18, no. 1, pp. 15–26, 1982.
[8] E. W. Bai, "An optimal two-stage identification algorithm for Hammerstein–Wiener nonlinear systems," *Automatica*, vol. 34, no. 3, pp. 333–338, 1998.
[9] J. Vörös, "Iterative algorithm for identification of Hammerstein systems with two-segment nonlinearities," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 2145–2149, Dec. 1999.
[10] E. W. Cheney, *Introduction to Approximation Theory*. Providence, RI: AMS Chelsea, 1982.
[11] T. Söderström and P. Stoica, *System Identification*. Upper Saddle River, NJ: Prentice-Hall, 1989.
[12] E. W. Bai, "Frequency domain identification of Hammerstein models," *IEEE Trans. Automat. Contr.*, vol. 48, pp. 530–542, Apr. 2003.
[13] ——, "Frequency domain identification of Wiener models," *Automatica*, vol. 39, pp. 1521–1530, 2003.
[14] W. Greblicki and M. Pawlak, "Cascade nonlinear system identification by a nonparametric method," *Int. J. Syst. Sci.*, vol. 25, pp. 129–153, 1994.
[15] Z. Hasiewicz, M. Pawlak, and P. Śliwiński, "Nonparametric identification of nonlinearities in block-oriented systems by orthogonal wavelets with compact support," *IEEE Trans. Circuits Syst.*, 2004, to be published.
[16] S. Efromovich, *Nonparametric Curve Estimation*. New York: Springer-Verlag, 1999.
[17] W. Härdle, *Applied Nonparametric Regression*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[18] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov, *Wavelets, Approximation, and Statistical Applications*. New York: Springer-Verlag, 1998.
[19] M. P. Wand and H. C. Jones, *Kernel Smoothing*. London, U.K.: Chapman Hall, 1995.
[20] J. A. Cristobal, P. F. Roca, and W. G. Manteiga, "A class of linear regression parameter estimators constructed by nonparametric estimation," *Ann. Statist.*, vol. 15, no. 2, pp. 603–609, 1987.