Hammerstein System Identification with the Nearest Neighbor Algorithm

Włodzimierz Greblicki and Mirosław Pawlak

***Abstract*—The nonlinear characteristic in a Hammerstein system, i.e., a system in which a nonlinear memoryless subsystem and a linear dynamic are connected in a cascade, is recovered with the nonparametric nearest neighbor regression estimate. The *apriori* information is nonparametric, both the nonlinear characteristic and the impulse response are completely unknown and can be of any form. Local and global properties of the estimate are examined. Whatever the probability density of the input signal, the estimate converges at every continuity point of the characteristic as well as in the global sense. We derive the asymptotic bias and variance of the proposed estimate. As a result the optimal rate of convergence is established that additionally is independent of the shape of the input density. Results of numerical simulations are also presented.**

***Index Terms*—System identification, Hammerstein system, nearest neighbor, nonparametric regression, dependent data, rate of convergence.**

I. Introduction

In this paper we examine the commonly used in numerous applications nonlinear Hammerstein system. The problem of identification of this system has been extensively studied in the signal processing literature, see [18] and the references cited therein. The classic parametric theory of identification of the Hammerstein system is reviewed in [18].

The nonparametric regression approach for recovering a nonlinear characteristic of the system is summarized in [24]. Two types of nonparametric regression function estimates have been applied to recover the nonlinear characteristic in the Hammerstein system, i.e., the kernel methods [21], [22], [34], and orthogonal series estimates [19], [30], [33], [44]. A comprehensive review of the nonparametric estimation methods and algorithms in the context of nonlinear system identification can be found in [24]. The statistical theory of nonparametric curve estimation is examined in [15], [27], [37]. In this paper, we propose to employ the nearest neighbor algorithm being a useful and flexible tool in the theory of nonparametric estimation of a regression function [27], machine learning [14] and nonparametric econometrics [37]. We also refer to [6] for recent systematic studies of the nearest neighbor estimation method including the problem of regression estimation.

From the statistical viewpoint, we also recover a regression function. The novelty is that, because of dynamics, observations are dependent which is a source of specific problems. In fact, there have been a few studies on the nearest neighbor estimate in the context of dependent data. We mention [56], [7] where the asymptotic properties of the estimate, under the strong mixing dependence assumption, are examined. In [35] the nearest neighbor regression estimate is examined for a static memoryless system with the dependent input signal.

The Hammerstein system consists of a linear dynamic subsystem and a memoryless nonlinear element. Functional forms of the estimated nonlinear characteristic, input probability density and an impulse response function of the linear subsystem are completely unknown. Imposing no assumptions on them, we prove that the algorithm recovers the nonlinear characteristic. We examine pointwise and global convergence rates and establish their optimality property. The global properties are assessed by the L_∞, L_1 and L_2 risks. The evaluated rates are independent of the shape of the input signal density. This is an important advantage over the mentioned kernel and orthogonal series algorithms where the rate of convergence depends critically on the roughness of the input density.

It is also worth mentioning that the data generated from the Hammerstein system do not necessarily meet the strong mixing condition. The latter being a common assumption in the nonparametric inference for time series models [15] ands also used in the previous studies on the nearest neighbor method [7], [56].

The rest of the paper is organized as follows. Section II formulates the nonparametric estimation problem, whereas Section III defines the nonparametric nearest neighbor estimate of the nonlinearity of the Hammerstein system. Moreover, in Section III a brief discussion of the linear subsystem identification is also given. In Section IV we establish the conditions for the pointwise and global convergence of the estimate including the uniform convergence. The corresponding convergence rates of the estimate are evaluated in Section V. Some numerical examples that support the theory are given in Section VI. The appendix contains the basic mathematical facts pertinent to the findings of this paper.

II. Nonparametric Hammerstein System Identification Problem

The Hammerstein system shown in Fig. 1 consists of a memoryless subsystem with a nonlinear characteristic $m(\cdot)$ and a linear dynamic subsystem with the impulse response $\{\lambda_i; i = 0, 1, \ldots\}$. The input signal $\{\ldots, U_{-1}, U_0, U_1, \ldots\}$ is a sequence of independent identically distributed random variables with the unknown input density $f(\cdot)$. Independent of the input signal, random disturbance $\{\ldots, \varepsilon_{-1}, \varepsilon_0, \varepsilon_1, \ldots\}$ is stationary white noise with zero mean and unknown variance σ_ε^2.

W. Greblicki is with Wrocław School of Information Technology Horizon, Wejherowska 28, 54-239 Wrocław, Poland, email: wlodzimierz.greblicki@gmail.com

M. Pawlak is with the Department of Electrical and Computer Engineering, University of Manitoba, Canada, email: Miroslaw.Pawlak@umanitoba.ca

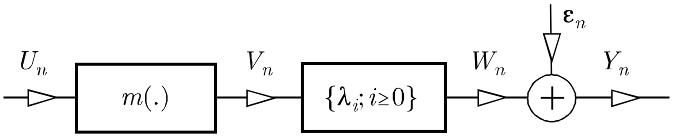

Fig. 1. The Hammerstein system.

The characteristic $m(\cdot)$ is just a Borel measurable function. Thus, $V_n = m(U_n)$ is a random variable. We assume, moreover, that

$$E\{V_n^2\} < \infty. \tag{1}$$

Concerning the linear subsystem

$$W_n = \sum_{i=0}^{\infty} \lambda_i V_{n-i} \tag{2}$$

we assume that it satisfies

$$\sum_{i=0}^{\infty} |\lambda_i| < \infty \tag{3}$$

which is the case, e.g., for every stable system described by a state equation. Owing to that, we find W_n to be a random variable and $\{\ldots, W_{-1}, W_0, W_1, \ldots\}$ is a stationary process. Consequently, $\{\ldots, Y_{-1}, Y_0, Y_1, \ldots\}$ is also a random stationary signal. Finally, signals U_n, V_n, W_n, Y_n and ε_n shown in Fig. 1 are all random variables while their sequences are stationary random processes. It means that the problem of recovering $m(\cdot)$ is well posed, i.e., can be described in statistical mathematics terms. In addition, with no loss of generality, to make our considerations easier, we assume that $E\{m(U_n)\} = 0$ and $\lambda_0 = 1$. All the above assumptions hold in the remaining of the paper and will not be repeated.

Our goal is to recover $m(\cdot)$ from a training sequence

$$\{(U_1, Y_1), (U_2, Y_2), \ldots, (U_n, Y_n)\} \tag{4}$$

of input-output observations. The intermediate signal V_n is not observed.

We want to stress that functional forms of both $f(\cdot)$ and $m(\cdot)$ are completely unknown, in particular they can be not continuous. We do not impose any restriction on $m(\cdot)$ nor $f(\cdot)$ which means that the classes of admissible characteristics and densities are as wide as possible and by no means can be parameterized. In the light of this, the *a priori* information about the system is nonparametric.

The tool we propose to use is the nearest neighbor regression estimate. This estimate forms the most intuitively appealing and adaptive of all nonparametric regression estimates. We show its pointwise and global consistency and establish the convergence rate that is argued to be optimal. Contrary to kernel and orthogonal series techniques, the algorithm can converge at points at which the input probability density $f(\cdot)$ is not continuous and, moreover, its convergence rate is independent of the shape of $f(\cdot)$ which is an obvious advantage. We also briefly discuss the problem of identifying the linear subsystem. This problem has been extensively examined in the system identification literature subject the constrain that the nonlinear part is specified parametrically [18]. The proposed algorithms are computationally dependent, i.e., the parameters in one part are held constant while the parameters in the other are determined. On the other hand, the correlation and regression function strategies allow us to decouple the linear and nonlinear system identification problems.

There have been a large number of studies on the nonparametric regression estimation for dependent data, see [15], [37] and the references cited therein. These contributions, however, have been mostly focused on various kernel methods and it has been commonly assumed that the dependence structure of the observed data meets the strong mixing condition. The dependence structure for our identification problem is determined by the linear dynamic system in (2). The conditions for the strong mixing dependence of linear processes were established in [53]. A quick inspection of the result in [53], see also [16], shows that the linear process $\{W_n\}$ in (2) with the restriction (1) is strong mixing if $|\lambda_n| = O(n^{-\gamma})$ for $\gamma > 5/2$. Note, however, the condition in (3) is satisfied under the following weaker requirement $|\lambda_n| = O(n^{-\gamma})$ for $\gamma > 1$.

There is yet another peculiar property of the data generated by the Hammerstein system. This results from the fact that the process $\{V_n\}$ is obtained as a nonlinear mapping of the input signal U_n. Consequently, this can produce the output process $\{Y_n\}$ that is neither strong mixing nor has an absolutely continuous distribution. In fact, let us assume that U_n is Gaussian with zero mean and let $V_n = \text{sgn}(U_n)$, i.e., V_n is equal to 1, -1 with probability 1/2. Furthermore, suppose that the linear system is represented by the autoregressive process $W_n = aW_{n-1} + V_n$, where $|a| < 1$. Then, the result of [1] reveals that the process $\{W_n\}$ is not strong mixing for any $a \in (0, 1/2]$. Moreover, when $a = 1/2$ then the distribution of W_n is absolutely continuous with the uniform density on $[-2, 2]$. For $0 < a < 1/2$ the distribution of W_n is singular. The case $1/2 < a < 1$ is much more complicated as we can have either absolutely continuous or purely singular distribution of W_n, see [47] and the references cited therein on the above described so-called Bernoulli convolution problem.

It is also worth mentioning that there are various generalizations of the concept of time series dependence. In particular, the dependence measure based on the physical data-generating mechanism was proposed in [55]. In [32] this type of predictive dependence was employed to establish the asymptotic properties of kernel regression estimates. In the context of the linear system in (2), this dependence measure is well-defined if (3) holds with the additional requirement that V_n has a differentiable density. Since $V_n = m(U_n)$ then it is clear that there is a large class of system nonlinearities such that this restriction cannot be satisfied.

The Hammerstein model can also be employed as the parsimonious approximation of more general dynamical nonlinear stationary systems. For example, let us assume the following $p-$th order nonlinear moving average system (NMA(p))

$$Y_n = g(U_n, U_{n-1}, \ldots, U_{n-p}) + \varepsilon_n, \tag{5}$$

where it is assumed that $EY_n^2 < \infty$ and $\{U_n\}$ is the iid input process. Next, let us consider a class of Hammerstein models of order p

$$\mathcal{H}(\lambda, m(\cdot)) = \left\{ Y_n^H = \sum_{j=0}^{p} \lambda_j m(U_{n-j}) : E\{m^2(U_n)\} < \infty \right\}.$$

Then it can be shown that the L_2 error $E\{(Y_n - Y_n^H)^2\}$ is minimized by a function $m(\cdot)$ and the impulse response $\{\lambda_j\}$ defined as

$$m(u) = \sum_{k=0}^{p} \eta_k g_k(u) + \alpha, \tag{6}$$

where $g_k(u) = E\{Y_n | U_{n-k} = u\}$, $\eta_k = \lambda_k / \sum_{j=0}^{p} \lambda_j^2$ for some constant α. The impulse response $\{\lambda_k\}$ is specified by $\lambda_k = \beta \mathrm{cov}(Y_n, U_{n-k})$, some constant β. The proof of this fact can be derived from the Hajek projection lemma that gives the best L_2 approximation of a function $g(X_1, \ldots, X_d)$ of independent random variables $X_1, \ldots, X_d$ by linear combinations of the form $\sum_{i=1}^{d} g_i(X_i)$ for some univariate functions $g_i(\cdot)$ with $E\{g_i^2(X_i)\} < \infty$, $i = 1, \ldots, d$, see [51] and the references cited therein. The fact that the aforementioned orthogonal projection of the NMA(p) system onto the class $\mathcal{H}(\lambda, m(\cdot))$ is specified up to some unknown constants is due to the cascade structure of the model. If, however, we assume that $E\{m(U_n)\} = 0$ and $\lambda_0 = 1$ then $\alpha = 0$ and $\beta = 1/\mathrm{cov}(Y_n, U_n)$. The estimation methodology developed in this paper can be extended to the problem of recovering the optimal nonlinearity in (6), where one should estimate the sequence of regression functions $g_k(u) = E\{Y_n | U_{n-k} = u\}$ and correlations $\mathrm{cov}(Y_n, U_{n-k})$ for $k = 0, \ldots, p$. The aforementioned projection is an important property since one can avoid the curse of dimensionality inherent with models such as NMA(p). In fact, instead of estimating the $p+1$-dimensional function $g(\cdot)$ characterizing the NMA(p) system in (5) we can merely estimate $p+1$ one-dimensional nonlinearities $g_k(u)$, $k = 0, \ldots, p$ defined in (6). It is also worth mentioning that there exists the theory of linear approximations of nonlinear systems, see [41] and the references cited therein.

Throughout the paper we need some notation and basic definitions. Hence, let us recall that, denoted by $\mathcal{S}$, the support of $f(\cdot)$ is the set of all points $u \in (-\infty, \infty)$ such that $\int_{u-\varepsilon}^{u+\varepsilon} f(v)dv > 0$ for every $\varepsilon > 0$. The characteristic $m(\cdot)$ can be bounded or unbounded in the support. Whenever is, we denote $M = \sup_{u \in S} |m(u)|$.

For the sake of simplicity, we sometimes write U for U_n. Moreover, for a sequence of random variables X_n writing $X_n = O_P(a_n)$ for a sequence of numbers a_n, we mean that X_n / a_n is bounded in the probability sense. Also by $\mathbb{1}(A)$ we denote the indicator function of the set A.

III. Identification Algorithms

A. Nonlinear System Identification

Let us recall that

$$Y_n = \sum_{i=0}^{\infty} \lambda_i m(U_{n-i}) + \varepsilon_n. \tag{7}$$

This readily yields

$$E\left\{Y_n | U_n = u\right\} = m(u). \tag{8}$$

Thus, to recover $m(u)$, a nonparametric regression estimate can be applied, e.g., kernel or orthogonal series estimates, see, e.g., [15], [24], [27], [37]. In this paper, however, we employ the k-nearest neighbor regression estimate, see [6]. It is now obvious that the assumptions $E\{m(U)\} = 0$ and $\lambda_0 = 1$ we have made for the sake of simplicity only. If these restrictions are not satisfied, the estimate recovers $\lambda_0 m(u) + \lambda$, where $\lambda = E\{m(U)\} \sum_{j=1}^{\infty} \lambda_j$. The fact that we can recover $m(\cdot)$ only up to some unknown constants is caused by the cascade structure of the system and, under the assumed lack of the *a priori* information, cannot be overcome.

For a fixed $u \in \mathcal{S}$, let

$$\left\{(U_{(1)}, Y_{[1]}), (U_{(2)}, Y_{[2]}), \ldots, (U_{(n)}, Y_{[n]})\right\}$$

be a sequence obtained from (4) by arranging all pairs of the input signal in a way such that

$$\left|U_{(1)} - u\right| < \left|U_{(2)} - u\right| < \ldots < \left|U_{(n)} - u\right|. \tag{9}$$

Ties, i.e., events that $|U_i - u| = |U_j - u|$ for $i \neq j$ have zero probability and can be neglected. Hence, the pair $(U_{(i)}, Y_{[i]})$ is such that $U_{(i)}$ depends on u and it is the i-th nearest neighbor of u. The corresponding output signal $Y_{[i]}$ is just paired with $U_{(i)}$ and is often called the concomitant of order statistics [9]. Hence, $Y_{[i]} = Y_i$ if $|U_i - u|$ is the i'th largest among $\{|U_j - u|, j = 1, \ldots, n\}$.

Our estimate of $m(u)$ is of the following form

$$\widehat{m}(u) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{[i]}. \tag{10}$$

The motivation of the estimate is obvious, since $\widehat{m}(u)$ is just the mean value of $Y_{[1]}, \ldots, Y_{[k_n]}$, i.e., of those Y_i's in (4) whose U_i's are among the k_n closest to u, where k_n is an integer such that $1 \leq k_n \leq n$.

The k-NN estimate is the simplest and most appealing of all nonparametric methods. It has a few attractive properties such as the ability to adapt to the amount and sparsity of local information that is available. This property is not shared by the fixed bandwidth estimators like kernel and series regression estimation methods. Besides, the k-NN regression estimate is not the ratio of two random variables, as this is the case with the popular nonparametric kernel method that is defined as

$$\widetilde{m}(u) = \frac{\displaystyle\sum_{i=1}^{n} Y_i K\left(\frac{u - U_i}{h}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{u - U_i}{h}\right)}, \tag{11}$$

where $K(u)$ is the kernel function and h is the bandwidth. Among many other things, this implies that asymptotical properties (like convergence) of the k-NN estimate holds under weaker conditions than those needed for the kernel estimate. In this paper we show that the convergence and the corresponding rate are entirely independent of the input density.

The idea of the nearest neighbor method can be traced back as far as [8], [17], [26], see also [14] where it was devoted to the problem of nonparametric classification. It was next applied to estimate a probability density function [39], and then adapted to a regression and studied by a number of authors, see, e.g., [10], [13], [49], [40] or more recently in [27], [37], [3], [4], [6]. A comprehensive examination of the nearest neighbor method is given in [6].

We apply the algorithm to recover the nonlinearity in the Hammerstein system. The presence of a dynamic subsystem makes, however, our problem tougher, since Y_i and Y_j, $i \neq j$, depend on each other whereas in [6] they are independent. Nevertheless, we should mention that in [56], [7] the nearest neighbor regression estimate was examined under the strong mixing dependence assumption on the observed data. As we have already discussed in Section II this assumption is not met in our case. In [35] the asymptotic behavior of (10) for dependent input signal U_n is studied. It has been assumed, however, that dependence between the output signal Y_n and the inputs $\{U_i, i \leq n\}$ is memoryless. Hence, Y_n depends only on U_n but not on the past input signals $\{U_i, i < n\}$. Clearly, this restriction is not satisfied in our case.

A related approach for the Hammerstein system identification based on spacings obtained also from order statistics can be found in [20], [23] and [24]. It is also worth mentioning that the nearest neighbor method forms the foundation for a large class of modern local machine learning algorithms, see [4], [28], [38] and the references cited therein.

B. Linear System Identification

Although, the main goal of this paper is the problem of recovering the system nonlinearity, let us shortly discuss the issue of the linear subsystem identification. Owing to (7) and the fact that the input signal is iid we can obtain

$$\mathrm{cov}(Y_{n+k}, U_n) = \gamma\lambda_k, \tag{12}$$

where $\gamma = \mathrm{cov}(m(U), U)$.

Next, using the assumption that $\lambda_0 = 1$ we can write

$$\lambda_k = \frac{\mathrm{cov}(Y_{n+k}, U_n)}{\mathrm{cov}(Y_n, U_n)}, \tag{13}$$

where we assume that $\gamma \neq 0$. The identity in (13) allows us to define a consistent correlation type estimate of λ_k that is entirely independent on the identification algorithm used for the nonlinear subsystem. Hence, we have

$$\widehat{\lambda}_k = \frac{\widehat{\theta}_k}{\widehat{\theta}_0}, \tag{14}$$

where $\widehat{\theta}_k = n^{-1}\sum_{j=1}^{n-k}(Y_{j+k} - \overline{Y})(U_j - \overline{U})$. Here $\overline{Y}$ and $\overline{U}$ are the average values of the output and input signals, respectively.

Since the output process $\{Y_n\}$ is stationary and ergodic we can easily argue that $\widehat{\lambda}_k \to \lambda_k$ in probability, as $n \to \infty$. Further properties of the correlation based and other identification algorithms for recovering the linear part of the Hammerstein system can be found in [24], [18].

IV. Convergence

In this section we show that the estimate $\widehat{m}(u)$ in (10) is consistent with respect to various distances. This includes the pointwise convergence as well as the consistency in the L_∞ (uniform), L_1 and L_2 norms. We begin with the pointwise consistency.

Theorem 1: If

$$k_n \to \infty \text{ as } n \to \infty, \tag{15}$$

$$\frac{k_n}{n} \to 0 \text{ as } n \to \infty, \tag{16}$$

then

$$\widehat{m}(u) \to m(u) \text{ as } n \to \infty \text{ in probability} \tag{17}$$

at every point u belonging to $\mathcal{S}$ - the support of $f(\cdot)$ at which $m(\cdot)$ is continuous.

Proof: In the system, $Y_n = m(U_n) + \psi_n + Z_n$ with $\psi_n = \sum_{i=1}^{\infty}\lambda_i m(U_{n-i})$. Hence, $Y_{[i]} = m(U_{(i)}) + \psi_{[i]} + Z_{[i]}$, where $\psi_{[i]}$ and $Z_{[i]}$ are paired with $U_{(i)}$. Therefore

$$\widehat{m}(u) - m(u) = \Phi_n(u) + \Psi_n + \Omega_n \tag{18}$$

with

$$\Phi_n(u) = \frac{1}{k_n}\sum_{i=1}^{k_n}\left(m(U_{(i)}) - m(u)\right), \tag{19}$$

$\Psi_n = (1/k_n)\sum_{i=1}^{k_n}\psi_{[i]}$, and $\Omega_n = (1/k_n)\sum_{i=1}^{k_n}\varepsilon_{[i]}$. Clearly $E\Omega_n = 0$ and $\mathrm{var}[\Omega_n] = \sigma_\varepsilon^2/k_n$. Hence $E\Omega_n^2 = \sigma_\varepsilon^2/k_n$. Passing to Ψ_n, we write $E\Psi_n^2 = \mathrm{var}[\Psi_n] = A_n + B_n$ with

$$\begin{aligned} A_n &= \frac{1}{k_n^2}\sum_{i=1}^{k_n}\mathrm{var}[\psi_{[i]}] = \frac{1}{k_n^2}\sum_{i=1}^{k_n}\mathrm{var}[\psi_i] \\ &= \frac{1}{k_n}\mathrm{var}\left[m(U)\right]\sum_{i=1}^{\infty}\lambda_i^2 \end{aligned}$$

and

$$\begin{aligned} B_n &= \frac{1}{k_n^2}\sum_{i=1}^{k_n}\sum_{j=1,j\neq i}^{k_n}\mathrm{cov}\left[\psi_{[i]},\psi_{[j]}\right] \\ &= \frac{1}{k_n^2}\sum_{i=1}^{k_n}\sum_{j=1,j\neq i}^{k_n}E\left\{\psi_{[i]}\psi_{[j]}\right\}. \end{aligned}$$

To evaluate B_n let us consider for $i \neq j$ the term $E\left\{\psi_{[i]}\psi_{[j]}\right\}$. Hence,

$$\begin{aligned} E\left\{\psi_{[i]}\psi_{[j]}\right\} = &\sum_{p=1}^{n}\sum_{q=1,q\neq p}^{n}E\{\psi_{[i]}\psi_{[j]}|U_{(i)} \text{ came } p\text{-th}, \\ &U_{(j)} \text{ came } q\text{-th}\} \\ &\times P\left\{U_{(i)} \text{ came } p\text{-th}, U_{(j)} \text{ came } q\text{-th}\right\} \\ &= \frac{1}{n(n-1)}\sum_{p=1}^{n}\sum_{q=1,q\neq p}^{n}E\left\{\psi_p\psi_q\right\}. \end{aligned}$$

The last formula results from the fact that the joint probability of the ranks of $U_{(i)}$ and $U_{(j)}$, i.e., their positions in the ordered data in (9) is equal to $1/(n(n-1))$, i.e., $P\left\{U_{(i)} \text{ came } p\text{-th}, U_{(j)} \text{ came } q\text{-th}\right\} = 1/(n(n-1))$. Next, let us note that for $q \geq p$ we have

$$E\{\psi_p\psi_q\} = \text{var}\,[m(U)] \sum_{i=1}^{\infty} \lambda_i\lambda_{i+q-p}.$$

Denoting the right-hand-side of the above formula by $\rho(q-p)$ we observe that

$$\sum_{p=1}^{n-1}\sum_{q=p+1}^{n} \rho(q-p) = n\sum_{l=1}^{n-1}\left(1-\frac{l}{n}\right)\rho(l).$$

This and the fact that

$$\sum_{p=1}^{n}\sum_{q=1,q\neq p}^{n} E\left\{\psi_p\psi_q\right\} = 2\sum_{p=1}^{n-1}\sum_{q=p+1}^{n} E\left\{\psi_p\psi_q\right\}$$

readily imply that

$$E\left\{\psi_{[i]}\psi_{[j]}\right\} = \frac{2}{n-1}\sum_{l=1}^{n-1}\left(1-\frac{l}{n}\right)\rho(l)$$

and consequently that

$$|E\left\{\psi_{[i]}\psi_{[j]}\right\}| \le \frac{2}{n-1}\sum_{l=1}^{\infty}|\rho(l)|.$$

Next let us note that

$$\sum_{l=1}^{\infty}|\rho(l)| \le \text{var}\,[m(U)]\sum_{l=1}^{\infty}\sum_{i=1}^{\infty}|\lambda_i\lambda_{i+l}| = \text{var}\,[m(U)]\left(\sum_{i=1}^{\infty}|\lambda_i|\right)^2.$$

This, due to the assumption in (3), implies that $B_n = O(1/n)$. This yields

$$E\{\Psi_n^2\} = d_0/k_n + O(1/n),$$

where $d_0 = \text{var}\,[m(U)]\sum_{i=1}^{\infty}\lambda_i^2$. Finally we obtain,

$$\text{var}[\Psi_n + \Omega_n] = d_1/k_n + O(1/n), \tag{20}$$

where $d_1 = \sigma_\varepsilon^2 + \text{var}\,[m(U)]\sum_{i=1}^{\infty}\lambda_i^2$.

Next by virtue of Lemma 1 in Appendix A and the condition in (16) we find that $\left|U_{(k_n)} - u\right| \to 0$ as $n \to \infty$ in probability at every point u belonging to the support of $f(\cdot)$. Assuming, moreover, that $m(\cdot)$ is continuous at the point u, we conclude that, for $i = 1,\ldots,k_n$, $m(U_{(i)}) \to m(u)$ as $n \to \infty$ in probability. Consequently, if (16) holds then $\Phi_n(u) \to 0$ as $n \to \infty$ in probability which completes the proof. ■

Remark 1: If (15) and (16) hold, then $\widehat{m}(u) \to 0$ as $n \to \infty$ in probability at every point u outside the support of $f(\cdot)$.

Remark 2: To satisfy (15) and (16), one can select $k_n = cn^\alpha$, $c > 0$, with $0 < \alpha < 1$.

Theorem 1 says that, whatever the density $f(\cdot)$, the algorithm converges at every continuity point of $m(\cdot)$ belonging to the support of $f(\cdot)$. To have a closer look, we assume that $m(\cdot)$ is continuous and that $f(u)$ is defined as $f(u) = (1/2)\mathbb{1}(-1 \le u \le 0) + (1-u)\mathbb{1}(0 < u \le 1)$. Since the interval $[-1,1]$ is the support of $f(\cdot)$, the algorithm converges at every point $u \in [-1,1]$, in particular at points $u = 0$ and $u = 1$, despite the fact that $f(\cdot)$ is not continuous at $u = 0$ and is equal zero at $u = 1$. Kernel and orthogonal series algorithms fail at both points because the kernel estimate requires pointwise continuity of $f(\cdot)$, while the orthogonal series method needs more than differentiability, not mentioning the fact that $f(\cdot)$ should be positive at the point of estimation of $m(\cdot)$, see [24]. This example demonstrates an important advantage of the nearest neighbor estimate.

Corollary 1: Let $m(\cdot)$ be bounded in the support of $f(\cdot)$. If (15) and (16) hold, then

$$E\left(\widehat{m}(u) - m(u)\right)^2 \to 0 \text{ as } n \to \infty$$

at every point $u \in \mathcal{S}$ - the support of $f(\cdot)$ at which $m(\cdot)$ is continuous.

Proof: Recalling Lebesgue's dominated convergence theorem it suffices to notice that $|\Phi_n(u)| \le 2M$. ■

In applications of the Hammerstein system to control engineering [18] one needs to evaluate accuracy of the nonlinearity estimation in terms of the sup-norm. The following theorem establishes such consistency under the global Lipschitz condition and slightly stronger requirements for the sequence k_n than those in (15) and (16).

Theorem 2: Let

$$\delta \le f(u),$$

some $\delta > 0$, for all $u \in \mathcal{S}$. Let, moreover,

$$|m(u) - m(v)| \le \alpha|u - v|, \tag{21}$$

some $\alpha > 0$, for all $u, v \in \mathcal{S}$. If (15) holds and

$$\frac{k_n \log n}{n} \to 0 \text{ as } n \to \infty, \tag{22}$$

then

$$\sup_{u\in\mathcal{S}}|\widehat{m}(u) - m(u)| \to 0 \text{ as } n \to \infty \text{ in probability.}$$

Proof: Inspecting the proof of Theorem 1, we conclude that it suffices to show that

$$\sup_{u\in\mathcal{S}}|\Phi_n(u)| \to 0 \text{ as } n \to \infty \text{ in probability,}$$

with $\Phi_n(u)$ as in (19). Obviously due to (21)

$$|\Phi_n(u)| \le \frac{\alpha}{k_n}\sum_{i=1}^{k_n}\left|U_{(i)} - u\right| \le \frac{\alpha}{k_n\delta}\sum_{i=1}^{k_n}\left|\int_u^{U_{(i)}} f(v)dv\right|.$$

To find an appropriate upper bound for the integral in the above expression, we denote $\xi_i = U_i$, $i = 1,\ldots,n$, and arrange ξ_i's in the increasing order. In this way, we obtain a sequence $\xi_{(1)}, \xi_{(2)}, \ldots, \xi_{(n)}$ such that $\xi_{(1)} < \xi_{(2)} < \cdots < \xi_{(n)}$. Hence, the sequence $\xi_{(1)}, \xi_{(2)}, \ldots, \xi_{(n)}$ represents the order statistics of the original input data $U_1,\ldots,U_n$. Let, moreover, $\xi_{(0)} = \inf_{u\in S} u$ and $\xi_{(n+1)} = \sup_{u\in S} u$. Let us denote

$$D_i = \int_{\xi_{(i)}}^{\xi_{(i+1)}} f(v)dv.$$

It should be noted that the sequence $\{D_i\}$ represents the uniform spacings, i.e., the gaps between the ordered uniform random variables defined on $[0,1]$. Let also $\mathbb{D}_n = \max_{0\le i\le n} D_i$ be the maximal gap or the largest uniform spacing, see [54],

[9], [46] for the basic theory of spacings. In this way, we obtain the desired bound

$$\left| \int_u^{U_{(i)}} f(v)dv \right| \le k_n \mathbb{D}_n$$

owing to which we can write

$$\sup_{u\in\mathcal{S}} |\Phi_n(u)| \le \frac{\alpha}{\delta} k_n \mathbb{D}_n. \tag{23}$$

In [48], see also [12], we find

$$n\mathbb{D}_n - \log n = O(\log\log n) \text{ as } n \to \infty \text{ almost surely.}$$

This readily implies that

$$\frac{n}{\log n}\mathbb{D}_n \to 1 \text{ as } n \to \infty \text{ almost surely}$$

and consequently

$$k_n \mathbb{D}_n \frac{n}{k_n \log n} \to 1 \text{ as } n \to \infty \text{ almost surely.}$$

This fact along with the bound in (23) and the condition in (22) allow us to complete the proof. ■

We shall now examine further global properties of our algorithm measured in terms of the L_1 and L_2 norms.

Theorem 3: Let $m(\cdot)$ be bounded in the support of $f(\cdot)$. If (15) and (16) hold, then

$$\int_{-\infty}^{\infty} |\widehat{m}(u) - m(u)| f(u)du \to 0 \text{ as } n \to \infty$$

in probability.

Proof: Going through the proof of Theorem 1, we conclude that it suffices to verify that $\int_{-\infty}^{\infty} |\Phi_n(u)| f(u)du \to 0$ as $n \to \infty$ in probability, i.e., that

$$\frac{1}{k_n}\sum_{i=1}^{k_n} \int_{-\infty}^{\infty} \left| m(U_{(i)}) - m(u) \right| f(u)du \to 0 \text{ as } n \to \infty$$

in probability. It is clear that, for $i = 1, 2, \ldots, k_n$,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \left| m(U_{(i)}) - m(u) \right| f(u)du \\
&\le \sup_{|v|\le r(i)} \int_{-\infty}^{\infty} |m(u-v) - m(u)| f(u)du \\
&\le \sup_{|v|\le r(k_n)} \int_{-\infty}^{\infty} |m(u-v) - m(u)| f(u)du
\end{aligned}$$

which, by virtue of Lemma 4 in Appendix B converges to zero as $n \to \infty$ in probability, where $r(k) = |U_{(k)} - u|$. In fact, due to Lemma 1 in Appendix A, $r(k_n) \to 0$ as $n \to \infty$ in probability. ■

Invoking Lebesgue's dominated convergence theorem, we get

Corollary 2: Let $m(\cdot)$ be bounded in the support of $f(\cdot)$. If (15) and (16) hold, then

$$E\int_{-\infty}^{\infty} (\widehat{m}(u) - m(u))^2 f(u)du \to 0 \text{ as } n \to \infty.$$

Notice that Theorems 1, 2 and 3 as well as Corollaries 1 and 2 hold for any probability density $f(\cdot)$. In turn, $m(\cdot)$ is also of any form in Theorem 1 while bounded in Theorem 3 and Corollaries 1 and 2. On the other hand, Theorem 2 imposes the Lipschitz condition on $m(\cdot)$ over the whole support S of the input density $f(\cdot)$. It is possible to extend this result to the case when the Lipschitz condition holds only over a compact subset of S. This would give the localized version of the result in Theorem 2.

V. Convergence rate

In order to evaluate the rate of convergence of our estimate $\widehat{m}(\cdot)$ we need some mild smoothness condition on the system nonlinearity $m(\cdot)$. Hence, let us fix $u \in \mathcal{S}$ and assume that $f(\cdot)$ is bounded from zero in the neighborhood of u, i.e., that $f(v) > \delta$, some $\delta > 0$, for all $v \in A$, where $A = [u-a, u+a]$ with some $a > 0$. Observe that u can be a discontinuity point of $f(\cdot)$.

Moreover, let us assume that $m(\cdot)$ satisfies the local Lipschitz condition at the point u, i.e.,

$$|m(v) - m(u)| \le \alpha |v - u| \tag{24}$$

for some α, for all v in some neighborhood of u. With no loss of generality, let A be the neighborhood. In addition, we assume that $m(\cdot)$ is bounded in the support of $f(\cdot)$. Owing to the decomposition in (18) and the asymptotic behavior of the estimate variance established in (20) it suffices to evaluate the term $\Phi_n(u)$ in (18) that controls the estimate bias.

Hence, we have

$$\begin{aligned}
E\{\Phi_n^2(u)\} &= E\left\{\Phi_n^2(u)|U_{(k_n)} \in A\right\} P\left\{U_{(k_n)} \in A\right\} \\
&\quad + E\left\{\Phi_n^2(u)|U_{(k_n)} \notin A\right\} P\left\{U_{(k_n)} \notin A\right\} \\
&\le E\left\{\Phi_n^2(u)|U_{(k_n)} \in A\right\} \\
&\quad + E\left\{\Phi_n^2(u)|U_{(k_n)} \notin A\right\} P\left\{U_{(k_n)} \notin A\right\}.
\end{aligned}$$

Suppose that $U_{(k_n)} \in A$. In such a case

$$\begin{aligned}
|\Phi_n(u)| &\le \frac{\alpha}{k_n}\sum_{i=1}^{k_n} \left| U_{(i)} - u \right| = \frac{\alpha}{k_n}\sum_{i=1}^{k_n} r(i) \\
&\le \alpha r(k_n) \le \frac{\alpha}{2\delta} B(k_n).
\end{aligned}$$

This is due to the fact that $r(k_n) = (1/2)\int_{u-r(k_n)}^{u+r(k_n)} dv \le (1/2\delta)B(k_n)$, where $r(k_n)$ and $B(k_n)$ are defined in (40) and (39) in Appendix A, respectively. Applying (41) from Appendix A, we find

$$E\left\{\Phi_n^2(u)|U_{(k_n)} \in A\right\} = \frac{\alpha^2}{4\delta^2}\left(\frac{k_n}{n}\right)^2 + O(n^{-1}). \tag{25}$$

In turn, employing Lemma 2 from Appendix A, we get $P\left\{U_{(k_n)} \notin A\right\} \le 2\exp(-\beta n)$ with some $\beta > 0$, provided that n is large enough. Observing that $E\left\{\Phi_n^2(u)|U_{(k_n)} \notin A\right\} \le 4M^2$ and using (25) we obtain

$$E\{\Phi_n^2(u)\} = \frac{\alpha^2}{4\delta^2}\left(\frac{k_n}{n}\right)^2 + O(n^{-1}). \tag{26}$$

Next, inspecting the proof of Theorem 1, see (20), we can write

$$E(\widehat{m}(u) - m(u))^2 = \frac{d_1}{k_n} + d_2\left(\frac{k_n}{n}\right)^2 + O(n^{-1}), \tag{27}$$

where

$$d_1 = \sigma_\varepsilon^2 + \mathrm{var}\,[m(U)] \sum_{i=1}^{\infty} \lambda_i^2 \text{ and } d_2 = \frac{\alpha^2}{4\delta^2}. \tag{28}$$

All these considerations yield the following result on the pointwise convergence rate of the k_n-nearest neighbor estimate for identification of the nonlinearity in the Hammerstein system.

Theorem 4: Let $m(\cdot)$ satisfy the local Lipschitz condition in (24). Then, by selecting

$$k_n^* = cn^{2/3}, \tag{29}$$

for some $c > 0$, we obtain

$$E(\widehat{m}(u) - m(u))^2 = O(n^{-2/3}). \tag{30}$$

It is worth noting that due to (27) the constant c in (29) can be specified as

$$c = \left(\frac{d_1}{2d_2}\right)^{1/3} \tag{31}$$

with the corresponding asymptotic mean squared error of order $3(d_1^2 d_2/4)^{1/3} n^{-2/3}$, where d_1, d_2 are defined in (28). The form of the constant c in (31) reveals that the asymptotical optimal k_n^* increases with the noise variance and the memory length of the linear system represented by the sum $\sum_{i=1}^{\infty} \lambda_i^2$.

By virtue of Lebesgue's dominated convergence theorem, for $f(\cdot)$ bounded from zero and $m(\cdot)$ bounded and Lipschitz on the whole support of $f(\cdot)$, and the result in (30) we obtain the rate for the mean integrated squared error

$$E\int_{-\infty}^{\infty} (\widehat{m}(u) - m(u))^2 f(u) du = O(n^{-2/3}).$$

Let us now turn into the uniform rate of convergence. From the proof of Theorem 2, see (23), we have that

$$\sup_{u \in \mathcal{S}} |\Phi_n(u)| = O_P\left(\frac{k_n}{n} \log n\right).$$

This combined with (20) yields

$$\sup_{u \in \mathcal{S}} |\widehat{m}(u) - m(u)| = O_P\left(\frac{1}{\sqrt{k_n}}\right) + O_P\left(\frac{k_n}{n} \log n\right).$$

The resulting uniform rate of convergence is given in the following theorem.

Theorem 5: Let $m(\cdot)$ satisfy the global Lipschitz condition in (21). Then, by selecting

$$k_n^* = c\frac{n^{2/3}}{\log^{2/3} n}, \tag{32}$$

for some $c > 0$, we obtain

$$\sup_{u \in \mathcal{S}} |\widehat{m}(u) - m(u)| = O_P\left(\frac{\log^{1/3} n}{n^{1/3}}\right). \tag{33}$$

Thus, the uniform rate in (33) is slower than the pointwise and L_2 rates but merely by the logarithmic factor $\log^{2/3} n$. On the other hand, the asymptotically optimal k_n in (32) is slightly smaller than the one in (29).

The result in (30) gives the the upper bound on the mean squared error at the point u where $m(u)$ meets the local Lipschitz condition in (24). Nevertheless, the rate in (30) agrees with the optimal rate established in [50]. We shall now show that the rate in (30) is in fact optimal, i.e., it cannot be improved for all nonlinearities satisfying (24). In order to do so we choose a specific form of the Hammerstein system and prove that the rate $O(n^{-2/3})$ must hold.

Hence, suppose that $\{\lambda_n\} = \{1, 0, 0, \ldots\}$. In addition, suppose that $\{U_n\}$ is distributed uniformly over the interval $[-d, d]$, i.e., that $f(u) = (2d)^{-1}\mathbb{1}(|u| \leq d)$. Let, moreover, $m(u) = \alpha|u|$, $\alpha > 0$. Now, $E\{Y_n|U_n = u\} = m(u)$ and $\widehat{m}(u) - m(u) = \Phi_n(u) + \Omega_n$, see the proof of Theorem 1. This yields $E(\widehat{m}(u) - m(u))^2 = E\{\Phi_n^2(u)\} + \sigma_\varepsilon^2/k_n$. Let $u^* = 0$. Clearly,

$$\begin{aligned}\Phi_n(u^*) &= \frac{\alpha}{k_n}\sum_{i=1}^{k_n} \left|U_{(i)} - u^*\right| = \frac{\alpha}{k_n}\sum_{i=1}^{k_n} r(i) \\ &= \frac{\alpha d}{k_n}\sum_{i=1}^{k_n} \int_{u^*-r(i)}^{u^*+r(i)} f(v)dv = \frac{\alpha d}{k_n}\sum_{i=1}^{k_n} B(i),\end{aligned}$$

where $B(i)$ and $r(i)$ are as in (39) and (40) in Appendix A with u replaced by u^*. Hence, in order to evaluate $E(\widehat{m}(u^*) - m(u^*))^2$ we need to consider $E\{\sum_{i=1}^{k_n} B(i)\}^2$. First let us note that

$$\begin{aligned}E\left\{\sum_{i=1}^{k_n} B(i)\right\}^2 &= \sum_{i=1}^{k_n} E\{B^2(i)\} \\ &\quad + 2\sum_{i=1}^{k_n-1}\sum_{j=i+1}^{k_n} E\{B(i)B(j)\}.\end{aligned}$$

By this, the identities (41) and (42) in Appendix A and the following elementary facts

$$\sum_{i=1}^{k_n} i(i+1) = \frac{1}{3}k_n(k_n+1)(k_n+2),$$

$$\sum_{i=1}^{k_n-1}\sum_{j=i+1}^{k_n} i(j+1) = \frac{1}{8}(k_n-1)k_n(k_n+1)(k_n+2),$$

we can evaluate $E\{\Phi_n^2(u^*)\}$. Hence, we obtain

$$\begin{aligned}&E(\widehat{m}(u^*) - m(u^*))^2 \\ &\quad = \alpha^2 d^2 \frac{(k_n+1)(k_n+2)\,(3k_n+1)}{12k_n(n+1)(n+2)} + \sigma_\varepsilon^2 \frac{1}{k_n}.\end{aligned} \tag{34}$$

Thus, in order to find k_n minimizing the error, we have to know α, d as well as σ_ε^2 which are all assumed unknown. Nevertheless, selecting k_n as in (29), i.e., $k_n = cn^{2/3}$ we get

$$n^{2/3}E(\widehat{m}(u^*) - m(u^*))^2 \to \gamma \text{ as } n \to \infty,$$

where $\gamma = (\alpha dc)^2/4 + \sigma_\varepsilon^2/c$ is a strictly positive constant.

Combining this necessary result with the sufficient one established in Theorem 4 we have the following result for the optimality of the obtained rate of convergence.

Theorem 6: Let $m(\cdot)$ satisfy the condition in (24). If $k_n = cn^{2/3}$, then $O(n^{-2/3})$ is the optimal pointwise rate convergence in the mean squared error sense.

It is worth noting that the optimal rate is independent of the shape of $f(\cdot)$. It is obvious that the choice in (29) is optimal in the asymptotic sense. All that is an important

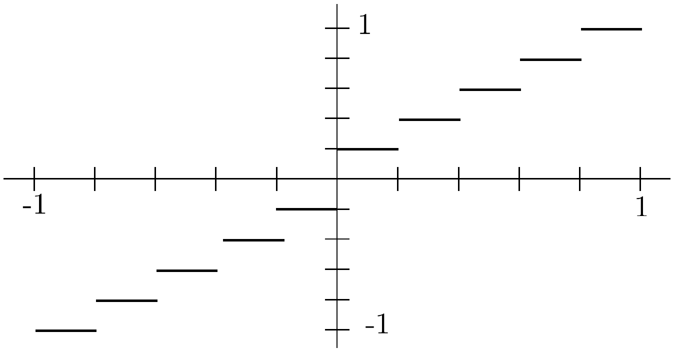

Fig. 2. The nonlinear characteristic $m(\cdot)$ with multiple jump points.

advantage over kernel and orthogonal series algorithms, since results concerning their convergence rate demand some kind of smoothness of both $f(\cdot)$ and $m(\cdot)$, e.g., differentiability, see [24]. Besides, from (34), it follows that the assumptions in (15) and (16) are not only sufficient, but also necessary for the mean squared convergence. This complements the characterization of (15) and (16) given in [13] in the case of iid data, see also [6].

Thus far, we have examined the convergence rate of $\widehat{m}(u)$ under the local Lipschitz condition implying the continuity of $m(u)$ at u. It is an interesting option to consider the convergence rate under the bounded variation assumption that allows discontinuities in $m(\cdot)$. We conjecture that the rate $O(n^{-2/3})$ still holds as it is known that the mean squared error of estimating the size of discontinuity in nonparametric regression is of the order $O(n^{-2/3})$, see [25] for such results in the case of iid data. The another extension of the practical importance would attempt to establish the limit law for $\widehat{m}(u) - E\widehat{m}(u)$ at $u \in \mathcal{S}$. Such a result would be useful for employing the estimate $\widehat{m}(u)$ for nonparametric testing a hypothesis on the restricted shape of the Hammerstein system nonlinearity such as monotonicity or its parametric form. We refer to [45] for some preliminary results into this direction, where a nonparametric test for correct parametric specification of the nonlinearity was proposed. The limit law for the nearest neighbor regression estimate in the noiseless iid case has been established in [6].

VI. Simulation Examples

In the first numerical example, the nonlinear characteristic $m(\cdot)$ shown in Fig. 2 has a number of discontinuity points which makes the examined algorithm in (10) converge slowly. The dynamic subsystem is described by the following equation: $W_n = aW_{n-1} + V_n$, with $a = 0.5$, where $V_n = m(U_n)$. Since $\lambda_n = a^n$, $\lambda_0 = 1$, therefore the algorithm recovers $m(u)$. The input signal is distributed uniformly on the interval $[-1, 1]$, disturbance has a normal distribution with variance 0.2.

The global error defined as MISE $= E\int_{-1}^{1}(\widehat{m}(u) - m(u))^2 f(u)du$ has been calculated empirically. The error versus k, where k plays the role of k_n, is shown in Fig. 3. Notice that the larger n, the less sensitivity of the error to not optimal choice of k. Hence, the optimal choice of k_n is easier for larger values of n. In turn, the minimal (with respect to k) MISE versus n is presented in Fig. 4. The error gets small rather fast. In fact, the error for $n = 1280$ is more than 8 times smaller than the error for $n = 40$. The optimal k, i.e., the optimal number of nearest neighbors, varies from 3 to 20 for n between 40 and 1280, respectively.

In practice, one would like to specify k directly from the observed data set in (4). To do so, we can apply the classical cross-validation strategy that selects k as the minimizer of the following criterion

$$CV(k) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{m}_{-i}(U_i)\right)^2,$$

where $\widehat{m}_{-i}(u)$ is the leave-one-out version of $\widehat{m}(u)$, i.e., $\widehat{m}_{-i}(u)$ is the version of $\widehat{m}(u)$ where the observation (U_i, Y_i) is left out in constructing $\widehat{m}(u)$. This is a reasonable method with some optimality conditions as it was proved in [36]. This is true as long as we have the iid data set which is not the case in our set-up. For dependent data the cross-validation needs to be modified and some kind of leave-$(2s+1)$-out version should be used. The choice of s depends on the memory of the dynamic subsystem. The value $s = 0$ corresponds to the iid case. This strategy was examined so far only in the classical density estimation situation using the kernel estimate [29]. The issue of the data-driven choice of k in the context of nonlinear dynamic system identification is postponed to future research.

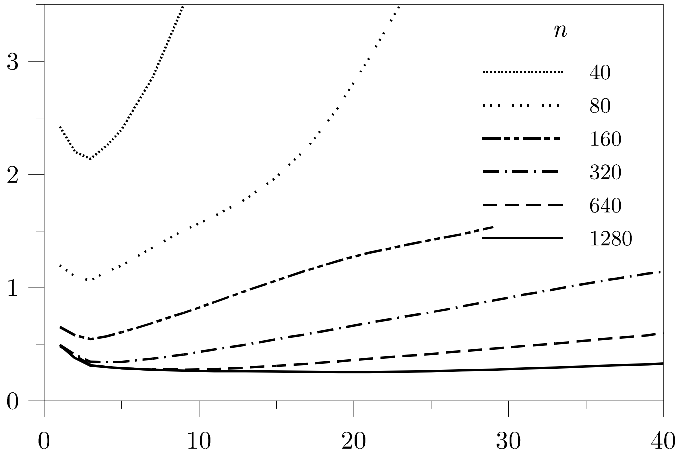

Fig. 3. 100×MISE versus k.

In the second experiment we wish to verify our asymptotic results on the invariance of the nearest neighbor estimate on the shape of the input density. In particular, the issue of the estimate adaptivity to the sparsity of data is to be exhibited. For the comparison studies we consider the classical kernel estimate defined in (11). The kernel estimate is a smooth estimate and to have a fair comparison we modify our nearest neighbor estimate to the following distance weighted form

$$\widehat{m}(u) = \frac{\displaystyle\sum_{i=1}^{k_n} Y_{[i]} K\left(\frac{u - U_{(i)}}{r(k_n)}\right)}{\displaystyle\sum_{i=1}^{k_n} K\left(\frac{u - U_{(i)}}{r(k_n)}\right)},$$

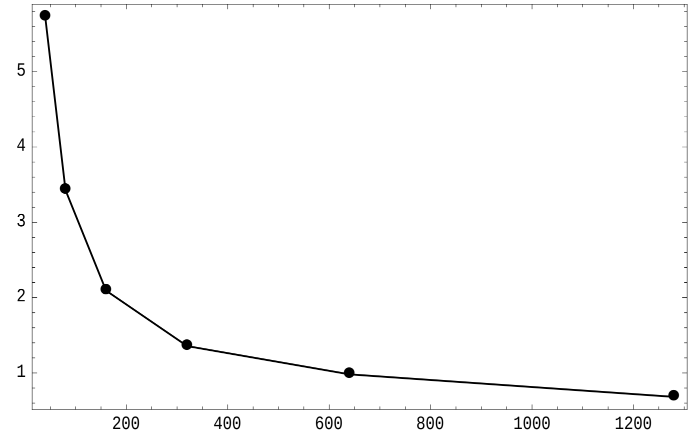

Fig. 4. 100×MISE versus n.

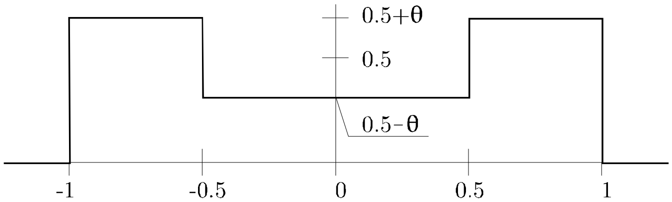

Fig. 5. Input density $f(u;\theta)$.

where $r(k_n)$ is the distance from u to its k_nth nearest neighbor. We choose the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)\mathbb{1}(|u| \leq 1)$ and the linear system identical as in the first experiment with $a = 0.5$.

The input density is of the form

$$f(u;\theta) = \left(\frac{1}{2} - \theta\right) \mathbb{1}\left(|u| \leq \frac{1}{2}\right) + \left(\frac{1}{2} + \theta\right) \mathbb{1}\left(\frac{1}{2} < |u| \leq 1\right),$$

where $0 \leq \theta \leq 1/2$, see Fig. 5. This class of densities can model a number of the sparsity situations regarding the distribution of input data points over the interval $[-1,1]$. In fact, the case $\theta = 0$ corresponds to the uniform density on $[-1,1]$. The extreme case $\theta = 1/2$ gives the input density that has no probability mass within the interval $[-1/2, 1/2]$. Hence, the value of θ close to 1/2 reflects the case of a very sparse input data set over the compact subset of the support of the input density. Assuming the system nonlinearity $m(u) = 3\sin(\pi u)$ we plot the mean squared pointwise error (MSE) versus θ ranging from $\theta = 0$ to $\theta = 0.45$. The error is evaluated at the 100 equally spaced points over the interval $[-1,1]$. Fig. 6 plots the error for both the kernel and nearest neighbor estimates. The global values of smoothing parameters h and k were selected as the minimizers of the MISE for the sample size $n = 200$. The advantage of the nearest neighbor estimate over the kernel method for larger values of θ is apparent. Fig. 7 gives the plots of the both estimates corresponding to the highly sparse input data with $\theta = 0.375$.

VII. Final Remarks

In this paper we have shown that the nearest neighbor estimate can be successfully applied to recover the nonlinearity in Hammerstein systems. It converges in local as well as global sense. The latter is measured in terms of the L_∞, L_1 and L_2 risks. Its advantage is that, whatever $f(\cdot)$, it converges at every continuity point of $m(\cdot)$ belonging to the support of $f(\cdot)$, not only at points at which both $f(\cdot)$ and $m(\cdot)$ are continuous or differentiable which is the case with other nonparametric algorithms. Moreover, the pointwise and global convergence rates are independent of the shape of $f(\cdot)$. The nearest neighbor algorithm has yet another useful property, i.e., it is invariant for data scaling. In fact, let T_n denote the training sequence in (4) and let $T_n^\alpha = \{(\alpha U_1, Y_1), (\alpha U_2, Y_2), \ldots, (\alpha U_n, Y_n)\}$, for $\alpha \neq 0$, be the transformed data set. Let also $\widehat{m}(u; T_n)$ and $\widehat{m}(u; T_n^\alpha)$ denote the nearest neighbor estimates based on T_n and T_n^α, respectively. Then, it is clear that we have $\widehat{m}(\alpha u; T_n^\alpha) = \widehat{m}(u; T_n)$. This invariance property is not shared by both kernel and orthogonal series regression estimates. The invariant properties of the $k-$NN regression estimate has been thoroughly examined recently in [5].

The input signal in this paper was assumed to be iid with an arbitrary (unknown) density function. If the input $\{U_n\}$ is a dependent stationary process, then we can argue that the fundamental relationship between the regression function $E\{Y_n|U_n = u\}$ and the system nonlinearity $m(u)$ established in (8) takes now the following integral equation form

$$E\{Y_n|U_n = u\} = m(u) + \int_{-\infty}^{\infty} L(v,u)m(v)dv, \qquad (35)$$

where $L(v,u) = \sum_{i=1}^{\infty} \lambda_i f_i(v|u)$ with $f_i(v|u)$ being the conditional density of U_i on U_0. Hence, the system nonlinearity is related to the regression function through a linear Fredholm integral equation of the second kind, see [42] for some recent studies of solutions of such equations for noisy data.

Concerning the linear subsystem identification let us note that the counterpart of the identity in (12) reads as

$$\mathrm{cov}[Y_{n+k}, U_n] = \sum_{j=0}^{\infty} \lambda_j \mathrm{cov}[m(U_{k-j}), U_0]. \qquad (36)$$

Assuming additionally that the input is a stationary Gaussian process we can evaluate both the conditional density $f_i(v|u)$ as well as $\mathrm{cov}[m(U_{k-j}), U_0]$. The latter, due to Bussgang identity, see [43] and also [24], can be rewritten as

$$\mathrm{cov}[m(U_{k-j}), U_0] = E\{m^{(1)}(U_0)\}\mathrm{cov}[U_{k-j}, U_0], \qquad (37)$$

where $m^{(1)}(u)$ is the derivative of $m(u)$ assumed to exist. The identities in (36) and (37) allow us to write the following convolution equation for the impulse response function $\{\lambda_j\}$ in terms of the covariance functions $R_{YU}(k) = \mathrm{cov}[Y_{n+k}, U_n]$ and $R_{UU}(k) = \mathrm{cov}[U_k, U_0]$.

$$R_{YU}(k) = \kappa \sum_{j=0}^{\infty} \lambda_j R_{UU}(k-j), \qquad (38)$$

where $\kappa = E\{m^{(1)}(U_0)\}$. The formulas established in (35) and (38) would provide a starting point for investigating the problem of identification of the Hammerstein system under

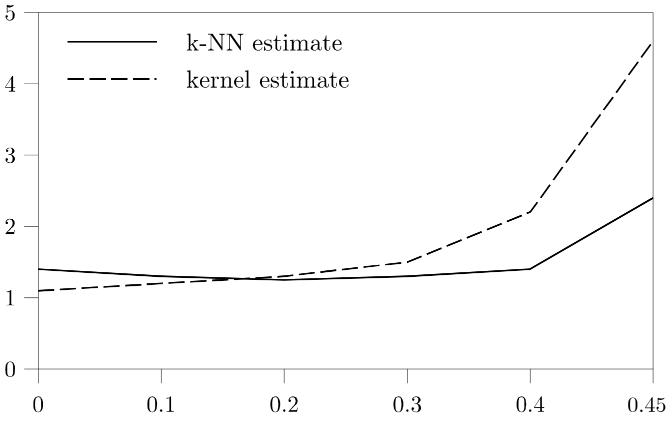

Fig. 6. 100×MSE versus θ, $n = 200$.

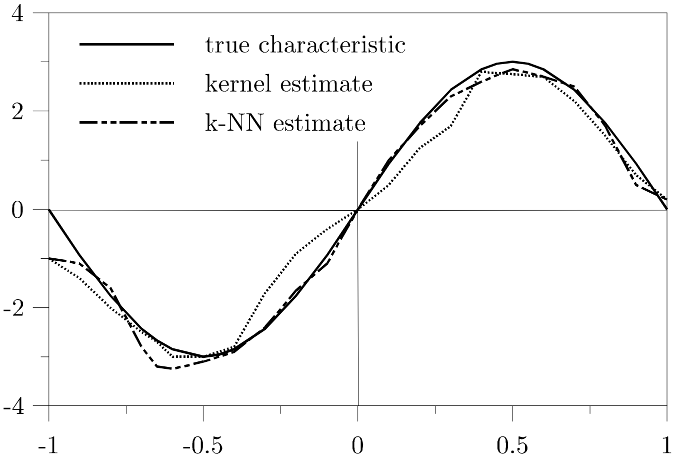

Fig. 7. The kernel and k-NN estimates of the nonlinearity $m(u) = 3\sin(\pi u)$ for $\theta = 0.375$, $n = 200$.

the dependent input. We leave this issue, however, for future research.

Let us finally comment on the computational complexity of the nearest neighbor estimates as it requires operation of ordering. It has to be performed for each point u separately. Notice, however, that, for close points, say u_1 and u_2, the distances $|U_{(i)} - u_1|$ and $|U_{(i)} - u_2|$ are also close and this fact can be taken into account in the process of ordering. To save computer time, the ordered sequence obtained for u_1 can be used as a starting point in a recursive process of ordering with respect to u_2. Another idea is to use a series ordered in a natural way, i.e., from the smallest to the greatest input observation, as the starting point. A computationally efficient recursive nearest neighbor regression estimate was studied in [11]. Only the case of iid data was taken into account. The all aforementioned issues reveal that the class of $k-$NN algorithms is worth further studies.

Appendix A

Next results and lemmas are related to order statistics. It is known that, whatever the density $f(\cdot)$ of the independent random sequence $U_1, U_2, \ldots, U_n$, we have that the random variable

$$B(k) = \int_{u-r(k)}^{u+r(k)} f(v)dv, \qquad (39)$$

where

$$r(k) = \left|U_{(k)} - u\right|, \qquad (40)$$

with $U_{(k)}$ as defined in (9), has a beta distribution with parameters k and $n-k+1$, see [39] or [54, Chapter 8]. Hence

$$EB^2(k) = \frac{k(k+1)}{(n+1)(n+2)}. \qquad (41)$$

Moreover,

$$E\left\{B(i)B(j)\right\} = \frac{i(j+1)}{(n+1)(n+2)}, \qquad (42)$$

for $1 \le i < j \le n$, see, e.g., [9] or [2].

Lemma 1: If (16) holds, then $r(k_n) \to 0$ as $n \to \infty$ in probability at every point u belonging to the support of $f(\cdot)$.

Proof: For every u in the support of $f(\cdot)$, $\int_{u-\varepsilon}^{u+\varepsilon} f(v)dv \to 0$ implies $\varepsilon \to 0$. From (16) and (41), it follows that $B(k_n) \to 0$ in probability as $n \to \infty$ which completes the proof. ■

Lemma 2: Let $A = [u-a, u+a]$, some $a > 0$, and let $P\{U \in A\} > 0$. If (16) holds, then $P\left\{U_{(k_n)} \notin A\right\} \le 2e^{-\beta n}$ with some $\beta > 0$, for n large enough.

Proof: Let $\eta_i = \mathbb{1}(U_i \in A)$ and denote $p = P\{U_i \in A\} = E\eta_i$. Notice that, in view of (16), $k_n/n < p/2$ for n large enough. Thus,

$$\begin{aligned} P\{U_{(k_n)} \notin A\} &= P\left\{\sum_{i=1}^{n} \eta_i < k_n\right\} \\ &\le P\left\{\frac{1}{n}\sum_{i=1}^{n} \eta_i < \frac{p}{2}\right\} \end{aligned}$$

which equals

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} \eta_i - p < -\frac{p}{2}\right\} \le P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} \eta_i - p\right| > \frac{p}{2}\right\}$$

which, by virtue of Hoeffding's inequality, see Lemma 3, is bounded by $2e^{-\beta n}$, $\beta = p^2/2$. ■

The next lemma is due to Hoeffding, see [31].

Lemma 3: Let $X_1, \ldots, X_n$ be independent identically distributed random variables such that $0 \le X_i \le 1$ and let $p = EX_i$. Then, for any positive t,

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| > t\right\} \le 2\exp\left(-2nt^2\right).$$

Appendix B

The following lemma is of general interest.

Lemma 4: If $m(\cdot)$ is a bounded function in the support of the probability density function $f(\cdot)$, then

$$\int_{-\infty}^{\infty} |m(u+h) - m(u)|\, f(u)du \to 0 \text{ as } |h| \to 0.$$

Proof: Clearly,

$$\int_{-\infty}^{\infty} |m(u+h) - m(u)|\, f(u) du$$
$$\leq \int_{-\infty}^{\infty} |m(u+h)|\, |f(u+h) - f(u)|\, du$$
$$+ \int_{-\infty}^{\infty} |m(u+h)f(u+h) - m(u)f(u)|\, du$$

which is bounded by

$$M \int_{-\infty}^{\infty} |f(u+h) - f(u)|\, du$$
$$+ \int_{-\infty}^{\infty} |m(u+h)f(u+h) - m(u)f(u)|\, du.$$

Since both $f(\cdot)$ and $m(\cdot)f(\cdot)$ are absolutely integrable functions then by virtue of Lemma 5 we can conclude the proof of Lemma 4. ■

In [52, Theorem 8.19] we find

Lemma 5: If $\int_{-\infty}^{\infty} |\varphi(u)| du < \infty$, then

$$\int_{-\infty}^{\infty} |\varphi(u+h) - \varphi(u)|\, du \to 0 \text{ as } |h| \to 0.$$

ACKNOWLEDGEMENT

The authors thank to the reviewers for their comments.

REFERENCES

[1] D.W.K. Andrews, "Non-strong mixing autoregressive processes," *Journal of Applied Probability*, vol. 21, pp. 930–934, 1984.

[2] N. Balakrishnan and A.C. Cohen, *Order Statistics and Inference*, Boston: Adademic Press, 1990.

[3] G. Biau, F. Chazal, D.Cohen-Steiner, L. Devroye and C. Rodrigues, "A weighted k-nearest neighbor density estimate for geometric inference," *Electronic Journal of Statistics*, vol. 5, pp. 204–237, 2011.

[4] G. Biau, F. Cérou and A. Guyader, "On the rate of convergence of the bagged nearest neighbor estimate," *Journal of Machine Learning Research*, vol. 11, pp. 687–712, 2010.

[5] G. Biau, L. Devroye, V. Dujmović and A. Krzyżak, "An affine invariant k-nearest neighbor regression estimate," *Journal of Multivariate Analysis*, vol. 112, pp. 24–34, 2012.

[6] G. Biau and L. Devroye, *Lectures on the Nearest Neighbour Method*, New York: Springer, 2015.

[7] G. Boente and R. Fraiman, "Robust nonparametric regression estimation for dependent observations," *The Annals of Statistics*, vol. 17, pp. 1242-1256, 1989.

[8] T.M. Cover, "Estimation by the nearest neighbor rule," *IEEE Transactions on Information Theory*, vol. 14, pp. 50–55, 1968.

[9] H.A. David and H.N. Nagaraja, *Order Statistics*, New York: Wiley, 2003, Third Edition.

[10] L. Devroye, "The uniform convergence of the nearest neighbor regression function estimators and their application in optimization," *IEEE Transactions on Information Theory*, vol. 24, pp. 142–151, 1978.

[11] L. Devroye and G.L. Wise, "Consistency of a recursive nearest neighbor regression estimate," *Journal of Multivariate Analysis*, vol. 10, pp. 539–550, 1980.

[12] L. Devroye, "Laws of the iterated logarithm for order statistics and uniform spacings," *The Annals of Probability*, vol. 9, pp. 860–867, 1981.

[13] L. Devroye, "Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 61, pp. 467–481, 1982.

[14] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, New York: Springer, 1996.

[15] J. Fan and Q. Yao, *A Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer, 2003.

[16] Y.A. Doukham, *Mixing: Properties and Examples*, New York: Springer-Verlag, 1994.

[17] E. Fix and J.L. Hodges, "Discriminatory analysis. Nonparametric discrimination: Consistency properties," *Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine*, Randolph Field, Texas, 1951.

[18] F. Giri and E.W. Bai (Eds.), *Block-Oriented Nonlinear System Identification*, New York: Springer-Verlag, 2010.

[19] W. Greblicki, "Non-parametric orthogonal series identification of Hammerstein systems," *International Journal of Systems Science*, vol. 20, pp. 2355–2367, 1989.

[20] W. Greblicki, "Nonlinearity estimation in Hammerstein systems based on ordered observations,"*IEEE Transactions on Signal Processing*, vol. 44, pp. 1224–1233, 1996.

[21] W. Greblicki and M. Pawlak, "Identification of discrete Hammerstein systems using kernel regression estimates," *IEEE Transactions on Automatic Control*, vol. 31, pp. 74–77, 1986.

[22] W. Greblicki, M. Pawlak, "Nonparametric identification of Hammerstein systems," *IEEE Transactions on Information Theory*, vol. 35, pp. 409-418, 1989.

[23] W. Greblicki and M. Pawlak, "Dynamic system identification with order statistics," *IEEE Transactions on Information Theory*, vol. 40, pp. 1474-1489, 1994.

[24] W. Greblicki and M. Pawlak, *Nonparametric System Identification*. Cambridge: Cambridge University Press, 2008.

[25] G. Gregoire and Z. Hamrouni, "Change point estimation by local linear smoothing," *Journal of Multivariate Analysis*, vol. 83, pp. 56-83, 2002.

[26] L. Györfi, "On the rate of convergence of nearest neighbor rule," *IEEE Transactions on Information Theory*, vol. 24, pp. 509-512, 1978.

[27] L. Györfi, M. Kohler, A. Krzyżak and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer, 2002.

[28] R. Hable, "Universal consistency of localized versions of regularized kernel methods," *Journal of Machine Learning Research*, vol. 14, pp. 153-186, 2013.

[29] J.D. Hart and P. Vieu, "Data-driven bandwidth choice for density estimation based on dependent data," *The Annals of Statistics*, vol. 18, pp. 873-890, 1990.

[30] Z. Hasiewicz,"Hammerstein system identification by the Haar multiresolution approximation," *International Journal of Adaptive Control and Signal Processing*, vol.13, pp. 691–717, 1999.

[31] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

[32] Y. Huang, X. Chen and W.B. Wu, "Recursive nonparametric estimation for time series," *IEEE Transactions on Information Theory*, vol. 60, pp. 1301-1312, 2014.

[33] A. Krzyżak, "Identification of discrete Hammerstein systems by the Fourier series regression estimate," *International Journal of Systems Science*, vol. 20, pp. 1729-1744, 1989.

[34] A. Krzyżak, "On estimation of a class of nonlinear systems by the kernel regression estimate," *IEEE Transactions on Information Theory*, vol. 36, pp. 141-152, 1990.

[35] S.R. Kulkarni and S.E. Posner, "Rates of convergence of nearest neighbor estimation under arbitrary sampling," *IEEE Transactions on Information Theory*, vol. 41, pp. 1028-1039, 1995.

[36] K.C. Li, "Consistency for cross-validated nearest neighbor estimates in nonparametric regression," *The Annals of Statistics*, vol. 12, pp. 230-240, 1984.

[37] Q. Li and J.S. Racine, *Nonaprametric Econometrics*, Princeton: Princeton University Proess, 2007.

[38] Y. Lin and Y. Jeon, "Random forests and adaptive nearest neighbors," *Journal of the American Statistical Association*, vol. 101, pp. 578–590, 2006.

[39] D.O. Loftsgaarden and C.P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Annals of Mathematical Statistics*, vol. 36, pp. 1049-1051, 1965.

[40] Y.P. Mack, "Local properties of k-NN regression estimates," *SIAM Journal on Algebraic Discrete Methods*, vol. 2, pp. 311-323, 1981.

[41] P.M. Mäkilä, "On optimal LTI approximation of nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 49, pp. 1178–1182, 2004.

[42] E. Mammen and K. Yu, "Nonparametric estimation of noisy integral equations of the second kind," *Journal of the Korean Statistical Society*, vol. 38, pp. 99-110, 2009.

[43] A. Papoulis and S.U. Pillai, *Probability, Random Variables and Stochastic Processes*, Boston: McGraw-Hill, 2002.

[44] M. Pawlak, "On the series expansion approach to the identification of Hammerstein systems," *IEEE Transactions on Automatic Control*, vol. 36, pp. 763–767, 1991.

[45] M. Pawlak and J. Lv, ”Nonparametric specification testing for Hammerstein systems,” *17th IFAC Symposium on System Identification SYSID 2015*, vol. 48, pp. 392–397, 2015.

[46] R.Pyke, ”Spacings,” *Journal of the Royal Statistical Society*, vol. 27, pp. 395-436, 1965.

[47] P. Shmerkin, ”On the exceptional set for absolutely continuity of Bernoulli convolutions,” *Geometric and Functional Analysis*, vol. 24, pp. 946-958, 2014.

[48] E. Slud, ”Entropy and maximal spacings for random partitions,” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 41, pp. 341–352, 1978.

[49] C.J. Stone, ”Consistent nonparametric regression,” *Annals of Statistics*, vol. 5, pp. 595-645, 1977.

[50] C.J. Stone, ”Optimal rates of convergence for nonparametric estimators,” *Annals of Statistics*, vol. 8, pp. 1348-1360, 1980.

[51] A.W. van der Vaart, *Asymptotic Statistics*, Cambridge: Cambridge University Press, 1998.

[52] R.L. Wheeden and A. Zygmund, *Measure and Integral*, New York: Dekker, 1977.

[53] C.S. Withers, ”Conditions for linear processes to be strong-mixing,” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 57, pp. 477–480, 1981.

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

[55] W.B. Wu, ”Nonlinear system theory: Another look at dependence,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 14150-14154, 2005.

[56] S. Yakowitz, ”Nearest-neighbour methods for time series,” *Journal of Time Series*, vol. 8, pp. 235–247, 1987.

Włodzimierz Greblicki was born in 1943 in Poland. He received the M.Eng., Ph.D., and D.Sc. degrees in automatic control from Wrocław University of Technology, Wrocław, Poland, in 1966, 1971, and 1975, respectively.

From 1966 to 2013 he held academic positions with the University. He is currently Full Professor at Wrocław School of Information Technology “Horyzont”, Wrocław, Poland. His research interests include mathematical statistics, pattern recognition and system identification.

Prof. Greblicki is a coauthor (with Prof. M. Pawlak) of the book *Nonparametric System Identification*, Cambridge University Press, 2008.

Mirosław Pawlak received the Ph.D. and D.Sc. degrees in computer engineering from Wrocław University of Technology, Wrocław, Poland.

He held research and teaching positions at Wrocław University of Technology and Concordia University, Montreal. He is currently a Professor at the Department of Electrical and Computer Engineering, University of Manitoba. He has held a number of visiting positions in North American, Australian, and European Universities. He was at the University of Ulm and Georg-August University in Göttingen, Germany, as the Alexander von Humboldt Foundation Fellow. His research interests include statistical aspects of signal/image processing, machine learning, and nonparametric modelling. Among his publications in these areas are the books Image Analysis by Moments: Reconstruction and Computational Aspects (Wrocław University of Technology Press, 2006), and Nonparametric System Identification (Cambridge University Press, 2008), coauthored with Włodzimierz Greblicki. Dr. Pawlak has been an Associate Editor of the Journal of Pattern Recognition and Applications, Pattern Recognition, International Journal on Sampling Theory in Signal and Image Processing, Opuscula Mathematica, and Statistics in Transition-New Series.