Identification of IIR Nonlinear Systems Without Prior Structural Information

Er-Wei Bai, *Fellow, IEEE*, Roberto Tempo, *Fellow, IEEE*, and Yun Liu

***Abstract*—In this paper, we propose a kernel method for identification of infinite impulse response (IIR) nonlinear systems without prior structural information. The main result of the paper is to establish the conditions for asymptotic convergence in terms of the input to output exponential stability. The performance of the method is tested on real world applications and computer simulations. The results obtained show the efficacy of the method.**

***Index Terms*—Kernel method, nonlinear system identification, parameter estimation, system identification.**

I. Introduction

SYSTEM identification is a critical part of system analysis and control [28]. The theory of linear system identification is relatively mature and a number of well developed techniques [11], [22] exist and are applicable to various problems. On the other hand, despite of a long history, there exist few results for identification of general nonlinear systems. Several classical survey papers [8], [10], [20] provide valuable insight into this problem, while more recent developments are presented in [7]. We also recall that two special issues on system identification [23] and [24] have appeared recently.

Nonlinear system identification can be roughly divided into two categories, known structure and unknown structure. If the system structure is available a priori, the nonlinear system identification becomes a nonlinear parameter estimation problem. There exist a large number of papers which deal with this topic. The most noticeable research along this direction is the so-called block oriented nonlinear system identification [1], [26]. The well-known Hammerstein and Wiener models belong to this class. Identification of nonlinear systems without prior structural information is a much harder problem and remains to be mostly intractable. Traditional methods are the Volterra and Wiener series representations [18] which require correlation-based higher order statistics. These methods are theoretically elegant, but applications are limited to nonlinear systems that can be approximated well by very low order kernels. Another approach to nonlinear system identification without prior structural information is to assume that the unknown system belongs to a certain class parameterized by a fixed basis functions, e.g., polynomials, Fourier series and orthonormal functions [1], [2], [19]. Then, again the problem is reduced to parameter estimation. A difficulty with this approach is that it is sensitive to the assumptions on the class that the actual but unknown nonlinear system belongs and to the basis functions chosen. Without enough prior structural information of the unknown system, a fixed basis function approach often needs a very large number of terms to reasonably approximate the unknown system. To overcome this difficulty, some basis functions dependent on tunable parameters, such as wavelets and neural networks, have been proposed [29]. These basis functions are much more powerful than the fixed ones. However, they still require adequate prior structural information on the unknown system so that the structure of the wavelets or the neural network is "rich" enough to include the class of nonlinear systems which contains the actual but unknown system. For instance, in neural network approximations, it is impossible to determine how many hidden nodes are needed if no prior information on the unknown nonlinear system is available. There is an additional difficulty with the tunable basis function approach, i.e., the minimization of the adjustable weights is usually non-convex and thus, the performance deteriorates due to the fact that the minimization algorithm often traps in a local minimum. There also exist a few other methods for nonlinear system identification without prior structural information, e.g., approximation by piecewise linear functions or splines [26], [27]. These methods can be effective for very low dimensional nonlinear systems. Finally, the membership set identification method has been recently extended to nonlinear systems without prior structural information [12].

In this paper, we propose a kernel method for nonlinear system identification without prior structural information, which is especially useful in this case because of its nonparametric nature. The idea of the method is not new and can be traced back to [15] for its use in probability density estimation. Since then, it has been studied and extended to nonlinear system identification without structural information along two main directions. The first one is application-oriented [3] and no attempt is made to derive analytically convergence conditions. The other direction is to find conditions so that the convergence of the kernel estimator can be achieved. The early work along this direction is reported in [13] where convergence results have been obtained for identification of a static nonlinear system. Convergence was obtained under certain mixing conditions on the input and output sequences [5]. The discussions were also extended to nonlinear systems with a state–space representation $X_{k+1} = f(X_k, U_k) + V_k$ in [6], [9], [16]. In particular in [6], it was assumed that the input vector U_k and the noise

Manuscript received October 26, 2004; revised September 15, 2005 and February 4, 2006. Recommended by Associate Editor W. X. Zheng. This work was supported in part by NSF ECS-0098181, NSF ECS-0555394 and NIH/NIBIB EB004287.

E.-W. Bai and Y. Liu are with the Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242 USA (e-mail: er-wei-bai@uiowa.edu).

R. Tempo is with the IEIIT-CNR, Politecnico di Torino, 10129 Torino, Italy (e-mail: tempo@polito.it).

Y. Liu is with Servo Tech, Inc., Chicago, IL 60608 USA (e-mail: yun_liu@servotechinc.com).

Digital Object Identifier 10.1109/TAC.2007.892385

vector V_k are both i.i.d. sequences. It is further assumed that the combined random sequence $X(k), U(k)$ is asymptotically stationary. Then, under the so-called G_2 condition, a variant of the mixing conditions, convergence results can be achieved. G_2 is a condition on the random process $X(k), U(k)$. For a given nonlinear system, the conditions that $X(k), U(k)$ is asymptotically stationary and G_2 holds were not discussed and answered in [6] or in related works. In other words, the conditions on a nonlinear system so the kernel method can be used were open. This paper takes a different approach by establishing convergence results in terms of the system properties, especially the system stability conditions. We show in this paper that traditionally defined stability is not sufficient for convergence and the input to output exponential stability is required.

The outline of the paper is as follows. Section II provides preliminaries and heuristic interpretations for the kernel estimation method. Assumptions and supporting lemmas are also given in this section. The kernel method is proposed in Section III along with its convergence proofs. The method is tested on a real world application in Section IV showing its practical effectiveness. Finally, several issues for future consideration are discussed in Section V.

II. Preliminaries

Consider a discrete time, scalar and time-invariant nonlinear infinite impulse response (IIR) system

$$y(k) = f(y(k-1), \ldots, y(k-n), u(k-1), \ldots, u(k-n)) + v(k) \quad (1)$$

$k = 1, 2, \ldots$, where the initial conditions $\{y(0), \ldots, y(-n+1)\}$ are unknown but are assumed to be bounded, i.e.,

$$\max\{|y(0)|, \ldots, |y(-n+1)|\} < M < \infty$$

for some constant M. The input $u(\cdot)$ is an iid random sequence over the interval $[\underline{u}, \bar{u}]$ with an unknown probability distribution, $v(\cdot)$ is a bounded iid random noise of zero mean and unknown variance σ_v^2 and $y(\cdot)$ is the output which is obviously a sequence of random variables. The input $u(\cdot)$ and the noise $v(\cdot)$ are statistically independent. The order n of the system is assumed to be known, but no prior structural information on f is known. The goal is to construct an estimate $\hat{f}$ based on the input-output data $\{y(k), u(k)\}_{k=1}^{N}$ so that $\hat{f}$ converges to f in some probability sense.

A function $K(\cdot)$ is said to be a kernel function if it satisfies

$$K(z) = \begin{cases} > 0, & z \in (\underline{u}, \bar{u}) \\ 0, & z \notin [\underline{u}, \bar{u}] \end{cases} \quad \text{and}$$
$$\int_{\underline{u}}^{\bar{u}} K(z)\, dz = 1 \quad (2)$$

and is continuous, bounded and symmetric with respect to the center of the interval $[\underline{u}, \bar{u}]$. There exist many such types of kernel functions, e.g., let $\underline{u} = -\bar{u}$, the second order and the bi-weight kernels are well-known

$$K(z) = \begin{cases} \frac{3}{4\bar{u}}\left[1 - \left(\frac{z}{\bar{u}}\right)^2\right], & z \in [\underline{u}, \bar{u}] \\ 0, & z \notin [\underline{u}, \bar{u}] \end{cases}$$
$$K(z) = \begin{cases} \frac{15}{16\bar{u}}\left[1 - \left(\frac{z}{\bar{u}}\right)^2\right]^2 & z \in [\underline{u}, \bar{u}] \\ 0, & z \notin [\underline{u}, \bar{u}]. \end{cases}$$

The kernel estimator is a smooth version of a conditional mean. Consider a static nonlinear system

$$y(k) = f(u(k)).$$

We estimate $f(x)$ by the empirical mean of $y(k)$'s associated with $u(k)$'s in a neighborhood of x and the size of the neighborhood decreases as the number of observations increases. It is well known that the kernel approach is convergent [4], [16], [17] for a static nonlinear system because

- the probability density functions of $y(k)$ and $y(i)$, for all k and i, are identical. In fact, the probability distribution of $y(\cdot)$ is stationary and is independent of time k;
- $y(k)$ and $y(i), i \neq k$, are independent. Simply put, the system $y(k) = f(u(k))$ has no memory.

These two key properties allow us to evaluate the statistical mean by the sample mean leading to the convergence of the kernel method. A general IIR nonlinear system of the form (1) does not have these properties even with the traditionally defined stable systems. For example

$$y(k) = 0.28y(k-1) + (y(k-1)^2 + 1)u(k-1)$$

is exponentially stable when $u(k) \equiv 0$. With $u(k) = (1/(\sqrt{2k+1}))$ that is bounded and goes to zero as $k \to \infty$, however, the solution $y(k) \to \infty$ if $y(0) = 1$, and $y(k) \to 0$ if $y(0) = 0.99$. With a small perturbation in the initial state, the solutions "remember" this difference forever and eventually depart exponentially. Clearly, to extend the results from $y(k) = f(u(k))$ to a general IIR nonlinear system (1), some conditions have to be imposed so that stationary and forgetting properties are retained.

To this end, for given initial time k_0, initial conditions $\{y(k_0), \ldots, y(k_0 - n + 1)\}$, input and noise sequences $\{u(\cdot)\}_{k_0-n+1}^{k-1}$ and $\{v(\cdot)\}_{k_0+1}^{k}$, let the solution of (1) at time k be denoted by

$$y(k) = \xi\left(k, \{y(k_0), \ldots, y(k_0 - n + 1)\}, \{u(i)\}_{k_0-n+1}^{k-1}, \{v(i)\}_{k_0+1}^{k}\right).$$

Assumption II.1:

1) The nonlinear system (1) is assumed to be exponentially input-to-output stable [21], i.e.,
   - for any $k > k_0$ and any initial conditions $\{y(k_0), \ldots, y(k_0 - n + 1)\}$

$$\left|\xi\left(k, \{y(k_0), \ldots, y(k_0 - n + 1)\}, \{u(i)\}_{k_0-n+1}^{k-1}, \{v(i)\}_{k_0+1}^{k}\right)\right| \le M_1(y(k_0), \ldots, y(k_0 - n + 1))\lambda^{k-k_0} + \gamma(\max_{k_0 \le i \le k}\{|u(i)|, |v(i)|\})$$

for some $0 \leq \lambda < 1$ and some bounded positive functions $M_1 > 0$ and γ.

- Consider two solutions at the initial time k_0 with different initial conditions but the same input and noise sequences

$$\xi\left(k, \{y(k_0), \ldots, y(k_0 - n + 1)\}, \{u(i)\}_{k_0-n+1}^{k-1}, \{v(i)\}_{k_0+1}^{k}\right)$$

and

$$\xi\left(k, \{\hat{y}(k_0), \ldots, \hat{y}(k_0 - n + 1)\}, \{u(i)\}_{k_0-n+1}^{k-1}, \{v(i)\}_{k_0+1}^{k}\right).$$

Then

$$\begin{aligned}&\Big|\xi\left(k, \{y(k_0), \ldots, y(k_0 - n + 1)\}, \{u(i)\}_{k_0-n+1}^{k-1}, \{v(i)\}_{k_0+1}^{k}\right)\\ &\quad - \xi\left(k, \{\hat{y}(k_0), \ldots, \hat{y}(k_0 - n + 1)\}, \{u(i)\}_{k_0-n+1}^{k-1}, \{v(i)\}_{k_0+1}^{k}\right)\Big|\\ &\quad \leq M_2(y(k_0), \ldots, y(k_0 - n + 1), \hat{y}(k_0), \ldots \hat{y}(k_0 - n + 1))\lambda^{k-k_0}\end{aligned}$$

for some bounded positive function $M_2 > 0$ and $0 \leq \lambda < 1$. In other words, the contribution of the initial condition is forgotten exponentially if the input and the noise for the two solutions are the same after the initial time k_0.

2) The function $f(y_1, \ldots, y_n, u_1, \ldots, u_n)$ is unknown but locally Lipschitz.
3) For all k, the probability density function of the random output $y(k)$ and the joint probability density function of $y(k-1), \ldots, y(k-n)$ exist and are continuous. Moreover, the joint probability density function of the random variables $y(k-1), \ldots, y(k-n), u(k-1), \ldots, u(k-n)$, denoted by $q_k(y_1, \ldots, y_n, u_1, \ldots, u_n)$ is locally Lipschitz in $y_1, \ldots, y_n$, i.e., for sufficiently small $\Delta y_1, \ldots, \Delta y_n$, we have

$$\begin{aligned}&|q_k(y_1 + \Delta y_1, \ldots, y_n + \Delta y_n, u_1, \ldots, u_n)\\ &\quad - q_k(y_1, \ldots, y_n, u_1, \ldots, u_n)| \leq M_3 \max_{1 \leq i \leq n} |\Delta y_i|\end{aligned}$$

for some bounded positive function $M_3 > 0$.

We remark that M_1, M_2 and $M_3 > 0$ in the assumption are not used in the algorithm and only enter in the convergence proof. Therefore, the conditions above are existence conditions, see additional comments in the discussion section.

To show a key lemma of this section, let $y(k)$ and $y(l)$ be the solutions of (1) with the initial time $k_0 = 0$ and initial conditions

$$\{y(0), \ldots, y(-n+1)\}, \{\hat{y}(0), \ldots, \hat{y}(-n+1)\}$$

respectively, i.e.,

$$\begin{aligned}y(k) &= \xi\left(k, \{y(0), \ldots, y(-n+1)\}, \{u(i)\}_{-n+1}^{k-1}, \{v(i)\}_1^k\right)\\ y(l) &= \xi\left(l, \{\hat{y}(0), \ldots, \hat{y}(-n+1)\}, \{u(i)\}_{-n+1}^{l-1}, \{v(i)\}_1^l\right).\end{aligned}$$

Clearly, the random variables $y(k)$ and $y(l)$ depend on the initial conditions $\{y(0), \ldots, y(-n+1)\}, \{\hat{y}(0), \ldots, \hat{y}(-n+1)\}$ and the input and noise sequences. Denote the probability density functions of $y(k)$ and $y(l)$ by

$$p_{k,\{y(0),\ldots,y(-n+1)\}}(\cdot) \quad \text{and} \quad p_{l,\{\hat{y}(0),\ldots,\hat{y}(-n+1)\}}(\cdot)$$

respectively. Now, we are in a position to show a key lemma.

Lemma II.1: Consider the system (1) with initial time $k_0 = 0$ under Assumption 2.1. Then, for any x, the probability density functions $p_{k,\{y(-1),\ldots,y(-n)\}}(\cdot)$ of $y(k)$ and $p_{l,\{\hat{y}(0),\ldots,\hat{y}(-n+1)\}}(\cdot)$ of $y(l)$ satisfy

$$|p_{k,\{y(0),\ldots,y(-n+1)\}}(x) - p_{l,\{\hat{y}(0),\ldots,\hat{y}(-n+1)\}}(x)| \to 0$$

as $\min\{k, l\} \to \infty$, for any $y(0), \ldots, y(-n+1), \hat{y}(0), \ldots, \hat{y}(-n+1)$ provided that

$$\max\{|y(0)|, \ldots, |y(-n+1)|, |\hat{y}(0)|, \ldots, |\hat{y}(-n+1)|\} < M < \infty.$$

Proof: Without loss of generality, we may assume that $l - k = j$ for some $j \geq 0$. We now make three observations.

1) Consider (1). For $j - m \geq 1$, let $\hat{y}(j-m)$ be the solution with the initial conditions $\{\hat{y}(0), \ldots, \hat{y}(-n+1)\}$

$$\hat{y}(j-m) = \xi\left(j - m, \{\hat{y}(0), \ldots, \hat{y}(-n+1)\}, \{u(i)\}_{-n+1}^{j-m-1}, \{v(i)\}_1^{j-m}\right).$$

We may now write $y(l)$ in terms of the initial time j and initial conditions $\{\hat{y}(j), \ldots, \hat{y}(j-n+1)\}$

$$\begin{aligned}y(l)&\xi\left(l, \{\hat{y}(0), \ldots, \hat{y}(-n+1)\}, \{u(i)\}_{-n+1}^{l-1}, \{v(i)\}_1^l\right)\\ &= \xi\left(k + j, \{\hat{y}(j), \ldots, \hat{y}(j-n+1)\}, \{u(i)\}_{j-n+1}^{k+j-1}, \{v(i)\}_{j+1}^{k+j}\right).\end{aligned}$$

The system (1) is time invariant. By shifting the time by j units, it follows that

$$y(l) = \xi\left(k, \{\bar{y}(0), \ldots, \bar{y}(-n+1)\}, \{\bar{u}(i)\}_{-n+1}^{k-1}, \{\bar{v}(i)\}_1^k\right)$$

where $\bar{y}(0) = \hat{y}(j), \ldots, \bar{y}(-n+1) = \hat{y}(j-n+1), \bar{u}_i = u_{i+j}, \bar{v}_i = v_{i+j}$.

2) From the input to output exponential stability and the previous observation, we have

$$\begin{aligned}&|y(l) - \xi\left(k, \{0, \ldots, 0\}, \{\bar{u}(i)\}_{-n+1}^{k-1}, \{\bar{v}(i)\}_1^k\right)|\\ &\quad = |\xi\left(k, \{\bar{y}(0), \ldots, \bar{y}(-n+1)\}, \{\bar{u}(i)\}_{-n+1}^{k-1}, \{\bar{v}(i)\}_1^k\right)\\ &\qquad - \xi\left(k, \{0, \ldots, 0\}, \{\bar{u}(i)\}_{-n+1}^{k-1}, \{\bar{v}(i)\}_1^k\right)|\\ &\quad = |\delta_1(k, \bar{y}(0), \ldots, \bar{y}(-n+1))|\\ &\quad \leq M_1\lambda^k \to 0, \text{ as } k \to \infty\end{aligned}$$

and as $k \to \infty$

$$\begin{aligned}&\left|y(k) - \xi\left(k, \{0, \ldots, 0\}, \{u(i)\}_{-n+1}^{k-1}, \{v(i)\}_1^k\right)\right|\\ &\quad = \left|\xi\left(k, \{y(0), \ldots, y(-n+1)\}, \{u(i)\}_{-n+1}^{k-1}, \{v(i)\}_1^k\right)\right.\\ &\qquad \left.- \xi\left(k, \{0, \ldots, 0\}, \{u(i)\}_{-n+1}^{k-1}, \{v(i)\}_1^k\right)\right|\\ &\quad = |\delta_2(k, y(0), \ldots, y(-n+1))| \leq M_1\lambda^k \to 0.\end{aligned}$$

3) The sequences $\{u_i\}$ and $\{\bar{u}_i\} = \{u_{i+j}\}$ are iid with identical distributions. Further, $\{v_i\}$ and $\{\bar{v}_i\} = \{v_{i+j}\}$ are two iid noise sequences with identical distributions. Thus, the solutions

$$\xi\left(k, \{0, \ldots, 0\}, \{u(i)\}_{-n+1}^{k-1}, \{v(i)\}_1^k\right)$$

and

$$\xi\left(k, \{0, \ldots, 0\}, \{\bar{u}(i)\}_{-n+1}^{k-1}, \{\bar{v}(i)\}_1^k\right)$$

must have the identical probability density function, say $p_{k,\{0,\ldots,0\}}(\cdot)$.

From these three observations, we have

$$\begin{aligned} &p_{l,\{\hat{y}(0),\ldots,\hat{y}(-n+1)\}}(x) \\ &\quad= \frac{d}{dx}\text{Prob}\{y(l) \le x\} \\ &\quad= \frac{d}{dx}\text{Prob}\left\{\xi\left(k, \{0, \ldots, 0\}, \{\bar{u}(i)\}_{-n+1}^{k-1}, \{\bar{v}(i)\}_1^k\right) \right. \\ &\quad\quad \left. \le x + \delta_1\right\} \\ &\quad= \frac{d}{dx}\int_{-\infty}^{x+\delta_1} p_{k,\{0,\ldots,0\}}(s)ds \\ &\quad= p_{k,\{0,\ldots,0\}}(x + \delta_1). \end{aligned}$$

Similarly

$$\begin{aligned} &p_{k,\{y(0),\ldots,y(-n+1)\}}(x) \\ &\quad= \frac{d}{dx}\text{Prob}\{y(k) \le x\} \\ &\quad= \frac{d}{dx}\text{Prob}\left\{\xi\left(k, \{0, \ldots, 0\}, \{u(i)\}_{-n+1}^{k-1}, \{v(i)\}_1^k\right) \right. \\ &\quad\quad \left. \le x + \delta_2\right\} \\ &\quad= \frac{d}{dx}\int_{-\infty}^{x+\delta_2} p_{k,\{0,\ldots,0\}}(s)ds \\ &\quad= p_{k,\{0,\ldots,0\}}(x + \delta_2). \end{aligned}$$

Since the density function $p_{k,\{0,\ldots,0\}}(\cdot)$ is assumed to be continuous and δ_1 and $\delta_2 \to 0$ as $k \to \infty$, the conclusion follows.

We make a few remarks here.

1) The above result by no means implies that $y(k)$ and $y(l)$ are independent. In fact, for all k and $l, y(k)$ and $y(l)$ are always dependent.
2) The result implies that the probability density function of $y(\cdot)$ does not depend on time k if k is large enough. In other words, in the steady state $k \to \infty$ or the initial time $k_0 = -\infty$, the contribution due to the initial condition vanishes and the probability density functions of $y(k)$ and $y(l)$ satisfy
$$q_{k,\{y(-1),\ldots,y(-n)\}}(\cdot) = q_{l,\{\hat{y}(-1),\ldots,\hat{y}(-n)\}}(\cdot) = q(\cdot)$$
independent of k, l and initial conditions. By the same token, the joint probability density functions of the random variables $y(k-1), \ldots, y(k-n), u(k-1), \ldots, u(k-n)$, denoted by $q_k(y_1, \ldots, y_n, u_1, \ldots, u_n)$ and the random variables $y(l-1), \ldots, y(l-n), u(l-1), \ldots, u(l-n)$, denoted by $q_l(y_1, \ldots, y_n, u_1, \ldots, u_n)$ satisfy
$$\begin{aligned} q_k(y_1, \ldots, y_n, u_1, \ldots, u_n) &= q_l(y_1, \ldots, y_n, u_1, \ldots, u_n) \\ &= q(y_1, \ldots, y_n, u_1, \ldots, u_n) \end{aligned}$$
independent of k and l.
3) The probability distribution of $y(\cdot)$ is unknown. In the steady state, however, the probability distributions of $y(k)$ and $y(l)$ are identical. This implies that the range $[\underline{y}, \bar{y}]$ in which $y(\cdot)$ lies can be easily estimated from the observations of $y(k)$'s.

To avoid unnecessary complications, in the rest of the paper, we assume that the steady state has been reached or the initial time is at $-\infty$ so that the probability density functions do not depend on k. The practical implication is that to carry out identification using the method proposed in this paper, one has to wait until the contribution of the initial condition becomes insignificant. This is the case even for linear system identification.

Lemma II.2: Let $q(y_1, \ldots, y_n, u_1, \ldots, u_n)$ be the joint probability density function of $y(k-1), \ldots, y(k-n), u(k-1), \ldots, u(k-n)$ and

$$q(y_1, \ldots, y_n, u_1, \ldots, u_n \,|\, y(i-1), \ldots, y(i-n), u(i-1), \ldots, u(i-n))$$

be the conditional probability density function of $y(k-1), \ldots, y(k-n), u(k-1), \ldots, u(k-n)$ given $y(i-1), \ldots, y(i-n), u(i-1), \ldots, u(i-n)$. Then, under Assumption 2.1, there exists $M_4, M_5 < \infty$ and $0 \le \lambda < 1$ so that for $k - i \ge M_5$

$$\begin{aligned} &|q(y_1, \ldots, y_n, u_1, \ldots, u_n | y(i-1), \ldots, y(i-n) \\ &u(i-1), \ldots, u(i-n)) \\ &\quad - q(y_1, \ldots, y_n, u_1, \ldots, u_n)| \le M_4\lambda^{k-i}. \end{aligned}$$

Proof: Let $y(k)$ be the solution of (1) at time k for given $y(i-1), \ldots, y(i-n), u(i-1), \ldots, u(i-n)$. When $k - i > n$, $y(k)$ may be written as

$$\begin{aligned} y(k) &= \xi\left(k, \{y(i-1), \ldots, y(i-n)\}, \{u(l)\}_{i-n}^{k-1}, \{v(l)\}_i^k\right) \\ &= \xi\left(k, \{y(i+n-1), \ldots, y(i)\}, \{u(l)\}_i^{k-1}, \{v(l)\}_{i+n}^k\right) \end{aligned}$$

where

$$\begin{aligned} &y(i) \\ &\quad= f(y(i-1), \ldots, y(i-n), u(i-1), \ldots \\ &\quad\quad u(i-n)) + v(i) \\ &y(i+1) \\ &\quad= f(y(i), \ldots, y(i-n+1), u(i), \ldots \\ &\quad\quad u(i-n+1)) + v(i+1) \\ &\quad\quad \cdots \\ &y(i+n-1) \\ &\quad= f(y(i+n-2), \ldots \\ &\quad\quad y(i-1), u(i+n-2), \ldots \\ &\quad\quad u(i-1)) + v(i+n-1). \end{aligned}$$

Let $\bar{y}(k)$ be a solution for arbitrary initial conditions $\{\bar{y}(i-1), \ldots, \bar{y}(i-n)\}$

$$\begin{aligned} \bar{y}(k) &= \xi\left(k, \{\bar{y}(i-1), \ldots, \bar{y}(i-n)\}, \{u(l)\}_{i-n}^{k-1}, \{v(l)\}_i^k\right) \\ &= \xi\left(k, \{\bar{y}(i+n-1), \ldots, \bar{y}(i)\}, \{u(l)\}_i^{k-1}, \{v(l)\}_{i+n}^k\right) \end{aligned}$$

for some $\bar{y}(i+n-1), \ldots, \bar{y}(i)$. By the exponential input-to-output stability

$$|y(k) - \bar{y}(k)| \le M_2\lambda^{k-(i+n)} = \frac{M_2}{\lambda^n}\lambda^{k-i} = M_6\lambda^{k-i}.$$

In other words

$$\bar{y}(k) = y(k) + \Delta y(k) \quad |\Delta y(k)| \le M_6\lambda^{k-i}.$$

Hence, these exists $M_7 > 0$ and for $k-i \geq M_7$, we have

$$\begin{aligned}
&q(y_1,\ldots,y_n,u_1,\ldots,u_n \,|\, y(i-1),\ldots,y(i-n) \\
&\quad u(i-1),\ldots,u(i-n)) \\
&\quad = \frac{d^{2n}}{dy_1\ldots dy_n du_1 \ldots du_n} \\
&\qquad \cdot \int_{\underline{y}}^{y_1} \cdots \int_{\underline{y}}^{y_n} \int_{\underline{u}}^{u_1} \cdots \int_{\underline{u}}^{u_n} q(s_1,\ldots,s_n,w_1,\ldots \\
&\qquad w_n \,|\, y(i-1),\ldots \\
&\qquad y(i-n),u(i-1),\ldots \\
&\qquad u(i-n))ds_1\ldots ds_n dw_1 \ldots dw_n \\
&= \frac{d^{2n}}{dy_1\ldots dy_n du_1 \ldots du_n} \\
&\qquad \cdot \mathrm{Prob}\{y(k-1) \leq y_1,\ldots \\
&\qquad y(k-n) \leq y_n, u(k-1) \leq u_1,\ldots \\
&\qquad u(k-n) \leq u_n \,|\, y(i-1),\ldots \\
&\qquad y(i-n),u(i-1),\ldots,u(i-n)\} \\
&\quad = \frac{d^{2n}}{dy_1\ldots dy_n du_1 \ldots du_n} \\
&\quad \cdot \mathrm{Prob}\{\bar{y}(k-1) \leq y_1 + \Delta y(k-1),\ldots \\
&\qquad \bar{y}(k-n) \leq y_n + \Delta y(k-n), u(k-1) \leq u_1,\ldots \\
&\qquad u(k-n) \leq u_n\} \\
&\quad = \frac{d^{2n}}{dy_1\ldots dy_n du_1 \ldots du_n} \\
&\qquad \cdot \int_{\underline{y}}^{y_1+\Delta y(k-1)} \cdots \int_{\underline{y}}^{y_n+\Delta y(k-n)} \int_{\underline{u}}^{u_1} \cdots \\
&\qquad \int_{\underline{u}}^{u_n} q(s_1,\ldots,s_n \\
&\qquad w_1,\ldots,w_n)ds_1\ldots ds_n dw_1 \ldots dw_n \\
&\quad = q(y_1+\Delta y(k-1),\ldots,y_n+\Delta y(k-n),u_1,\ldots,u_n).
\end{aligned}$$

From the assumption that the joint density function is locally Lipschitz, it follows that there exist $M_4, M_5 < \infty$ and $0 \leq \lambda < 1$ such that for $k-i \geq M_5$

$$\begin{aligned}
&|q(y(k-1),\ldots,y(k-n),u(k-1),\ldots \\
&\quad u(k-n)|y(i-1),\ldots,y(i-n),u(i-1),\ldots,u(i-n)) \\
&\qquad - q(y(k-1),\ldots,y(k-n),u(k-1),\ldots,u(k-n))| \\
&\qquad\quad = |q(y(k-1)+\Delta y(k-1),\ldots,y(k-n) \\
&\qquad\qquad + \Delta y(k-n),u(k-1),\ldots,u(k-n)) \\
&\qquad\qquad - q(y(k-1),\ldots,y(k-n),u(k-1),\ldots \\
&\qquad\qquad\quad u(k-n))| \leq M_4\lambda^{k-i}.
\end{aligned}$$

This completes the proof. □

A similar proof gives the following corollary.

Corollary II.1: Let $q(y_1,\ldots,y_n,u_1,\ldots,u_n)$ be the joint probability density function of $y(k-1),\ldots,y(k-n),u(k-1),\ldots,u(k-n)$ and

$$\begin{aligned}
&q(y_1,\ldots,y_n,u_1,\ldots,u_n \,|\, y(i-1),\ldots, \\
&\quad y(i-n),u(i-1),\ldots,u(i-n),v(i))
\end{aligned}$$

be the conditional probability density function of $y(k-1),\ldots,y(k-n),u(k-1),\ldots,u(k-n)$ given $y(i-1),\ldots,y(i-n),u(i-1),\ldots,u(i-n)$ and $v(i)$. Then, under Assumption 2.1, there exists $M_8, M_9 < \infty$ and $0 \leq \lambda < 1$ so that for $k-i \geq M_9$

$$\begin{aligned}
&|q(y_1,\ldots,y_n,u_1,\ldots,u_n \,|\, y(i-1),\ldots \\
&\quad y(i-n),u(i-1),\ldots,u(i-n),v(i)) \\
&\qquad - q(y_1,\ldots,y_n,u_1,\ldots,u_n)| \leq M_8\lambda^{k-i}.
\end{aligned}$$

III. The Kernel Method and Its Convergence Analysis

We recall again that the main objective of this paper is to estimate the nonlinear function $f(y_1,\ldots,y_n,u_1,\ldots,u_n)$ based on the input-output measurements $y(k),u(k) \in [\underline{y},\bar{y}] \times [\underline{u},\bar{u}]$. We are now in a position to define the kernel estimate and to prove its convergence. Let the multi-dimensional kernel function $K(y_1,\ldots,y_n,u_1,\ldots,u_n)$ be defined as in (2), i.e., $K(\cdot,\ldots,\cdot)$ is continuous, bounded and symmetric with each variable satisfying

$$K(y_1,\ldots,y_n,u_1,\ldots,u_n) = \begin{cases} >0, & y_i \in (\underline{y},\bar{y}) \text{ and } u_i \in (\underline{u},\bar{u}) \quad i=1,\ldots,n \\ 0, & y_i \notin [\underline{y},\bar{y}] \text{ or } u_i \notin [\underline{u},\bar{u}] \text{ for some } i \end{cases}$$

and

$$\int_{\underline{y}}^{\bar{y}} \cdots \int_{\underline{y}}^{\bar{y}} \int_{\underline{u}}^{\bar{u}} \cdots \int_{\underline{u}}^{\bar{u}} K(y_1,\ldots,y_n,u_1,\ldots,u_n) \cdot dy_1 \cdots dy_n du_1 \cdots du_n = 1.$$

Next, given the measurements $\{y(k),u(k)\}_1^N$, define the estimate $\hat{f}_N(y_1,\ldots,y_n,u_1,\ldots,u_n)$ of $f(y_1,\ldots,y_n,u_1,\ldots,u_n)$ as shown at the bottom of the page for some sufficiently small $r > 0$.

We now show the convergence of the estimate (1).

Theorem III.1: Consider the system (1) and the kernel estimate at the bottom of the page under Assumption 2.1. Then, for every $(y_1,\ldots,y_n,u_1,\ldots,u_n) \in (\underline{y},\bar{y})^n \times (\underline{u},\bar{u})^n$ so that the probability density function

$$q(y_1,\ldots,y_n,u_1,\ldots,u_n) \neq 0$$

we have

$$\hat{f}_N(y_1,\ldots,y_n,u_1,\ldots,u_n) \to f(y_1,\ldots,y_n,u_1,\ldots,u_n)$$

in probability as $N \to \infty$, provided that $r \to 0, r^{2n}N \to \infty$ as $N \to \infty$.

$$\hat{f}_N(y_1,\ldots,y_n,u_1,\ldots,u_n) = \frac{\sum_{j=1}^N K\left(\frac{y_1-y(j-1)}{r},\ldots,\frac{y_n-y(j-n)}{r},\frac{u_1-u(j-1)}{r},\frac{u_n-u(j-n)}{r}\right)y(j)}{\sum_{j=1}^N K\left(\frac{y_1-y(j-1)}{r},\ldots,\frac{y_n-y(j-n)}{r},\frac{u_1-u(j-1)}{r},\ldots,\frac{u_n-u(j-n)}{r}\right)}$$

Proof: Since $q(y_1,\ldots,y_n,u_1,\ldots,u_n)\neq 0$, it suffices to show that in probability

$$nu_N = \frac{1}{r^{2n}N}\sum_{j=1}^{N}K\left(\frac{y_1-y(j-1)}{r},\ldots,\frac{u_n-u(j-n)}{r}\right)y(j) \to q(y_1,\ldots,y_n,u_1,\ldots,u_n)f(y_1,\ldots,y_n,u_1,\ldots,u_n) \quad (2)$$

and

$$de_N = \frac{1}{r^{2n}N}\sum_{j=1}^{N}K\left(\frac{y_1-y(j-1)}{r},\ldots,\frac{u_n-u(j-n)}{r}\right) \to q(y_1,\ldots,y_n,u_1,\ldots,u_n). \quad (3)$$

To simplify notation, denote the unknown density function of $v(i)$ by $q_v(i)$ and define

$$\begin{aligned}
K(l) &= K\left(\frac{y_1-y(l-1)}{r},\ldots,\frac{u_n-u(l-n)}{r}\right)\\
f(l) &= f(y(l-1),\ldots,y(l-n)\\
&\quad u(l-1),\ldots,u(l-n))\\
q(l) &= q(y(l-1),\ldots,y(l-n)\\
&\quad u(l-1),\ldots,u(l-n))\\
q(i,j) &= q(y(i-1),\ldots,y(i-n)\\
&\quad u(i-1),\ldots,u(i-n),y(j-1),\ldots,y(j-n)\\
&\quad u(j-1),\ldots,u(j-n))\\
q(i\,|\,j) &= q(y(i-1),\ldots,y(i-n)\\
&\quad u(i-1),\ldots,u(i-n)|y(j-1),\ldots,y(j-n)\\
&\quad u(j-1),\ldots,u(j-n))\\
q(i\,|\,j,v(j)) &= q(y(i-1),\ldots,y(i-n)\\
&\quad u(i-1),\ldots,u(i-n)\,|\,y(j-1),\ldots,y(j-n)\\
&\quad u(j-1),\ldots,u(j-n),v(j))\\
d\vec{y}_l &= dy(l-1)dy(l-2)\ldots dy(l-n)\\
d\vec{u}_l &= du(l-1)du(l-2)\ldots du(l-n).
\end{aligned}$$

Now, to show (2), we write

$$\begin{aligned}
&\mathbf{E}|nu_N - f(y_1,\ldots,y_n,u_1,\ldots,u_n)\\
&\quad q(y_1,\ldots,y_n,u_1,\ldots,u_n)|^2\\
&\qquad = [\mathbf{E}(nu_N)^2-(\mathbf{E}nu_N)^2]\\
&\qquad\quad + [\mathbf{E}nu_N - f(y_1,\ldots,y_n,u_1,\ldots,u_n)\\
&\qquad\quad \cdot q(y_1,\ldots,y_n,u_1,\ldots,u_n)]^2
\end{aligned} \quad (4)$$

where $\mathbf{E}$ stands for the expectation operator. First, we have

$$\begin{aligned}
\mathbf{E}(nu_N)^2 &= \mathbf{E}\left[\frac{1}{r^{2n}N}\sum_{i=1}^{N}K(i)(f(i)+v(i))\right]^2\\
&= \mathbf{E}\frac{1}{r^{2n}N}\sum_{i=1}^{N}K(i)f(i)\cdot\frac{1}{r^{2n}N}\sum_{j=1}^{N}K(j)f(j)\\
&\quad + \mathbf{E}\frac{1}{r^{2n}N}\sum_{i=1}^{N}K(i)v(i)\\
&\quad \cdot\frac{1}{r^{2n}N}\sum_{j=1}^{N}K(j)v(j) + 2\mathbf{E}\frac{1}{r^{2n}N}\sum_{i=1}^{N}K(i)f(i)\\
&\quad \cdot\frac{1}{r^{2n}N}\sum_{j=1}^{N}K(j)v(j).
\end{aligned} \quad (5)$$

Note that if $i\neq j$, say $i>j$, then $v(i)$ is independent of $K(i)$, $K(j)$ and $v(j)$. This implies $\mathbf{E}K(i)K(j)v(i)v(j)=0$ if $i\neq j$. Thus, the middle term of (5) is equal to $(\sigma_v^2)/(r^{2n}N^2)\sum_{j=1}^{N}\mathbf{E}(1/(r^{2n}))K^2(j)$. Now

$$\begin{aligned}
\mathbf{E}\frac{1}{r^{2n}}K^2(i) &= \frac{1}{r^{2n}}\int_{\underline{y}}^{\bar{y}}\cdot\int_{\underline{y}}^{\bar{y}}\int_{\underline{u}}^{\bar{u}}\cdot\int_{\underline{u}}^{\bar{u}}K^2\\
&\quad\cdot\left(\frac{y_1-y(i-1)}{r},\ldots,\frac{y_n-y(i-n)}{r},\right.\\
&\quad\left.\frac{u_1-u(i-1)}{r},\ldots,\frac{u_n-u(i-n)}{r}\right)q(i)\,d\vec{y}_i\,d\vec{u}_i.
\end{aligned}$$

Observe for $r>0$

$$\underline{y}\le\frac{y_j-y(i-j)}{r}\le\bar{y}\Longleftrightarrow y_j-r\bar{y}\le y(i-j)\le y_j-r\underline{y}$$

$$\underline{u}\le\frac{u_j-u(i-j)}{r}\le\bar{u}\Longleftrightarrow u_j-r\bar{u}\le u(i-j)\le u_j-r\underline{u}$$

where $j=1,\ldots,n$. By defining two new variables s_j and w_j and making changes of variables

$$y_j-y(i-j)=rs_j \quad u_j-u(i-j)=rw_j, \qquad j=1,2,\ldots,n$$

it is easily verified that $\mathbf{E}(1/(r^{2n}))K^2(i)$ is bounded. This implies that the middle term of (5) is bounded by

$$\left|\frac{\sigma_v^2}{r^{2n}N^2}\sum_{j=1}^{N}\mathbf{E}\frac{1}{r^{2n}}K^2(j)\right| = O\left(\frac{1}{r^{2n}N}\right)$$

and

$$\begin{aligned}
&\mathbf{E}(nu_N)^2-(\mathbf{E}nu_N)^2\\
&\quad = O\left(\frac{1}{r^{2n}N}\right)+\frac{1}{r^{4n}N^2}\\
&\qquad\cdot\sum_{i=1}^{N}\sum_{j=1}^{N}\{\mathbf{E}K(i)K(j)f(i)f(j)\\
&\qquad - \mathbf{E}K(i)f(i)\mathbf{E}K(j)f(j)\}\\
&\qquad + \frac{2}{r^{4n}N^2}\mathbf{E}\sum_{i=1}^{N}K(i)f(i)\cdot\sum_{j=1}^{N}K(j)v(j).
\end{aligned} \quad (6)$$

The middle term of (6) can be written as

$$\begin{aligned}
&\frac{1}{r^{4n}N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int_{\underline{y}}^{\bar{y}}\cdots\int_{\underline{y}}^{\bar{y}}\int_{\underline{u}}^{\bar{u}}\cdots\int_{\underline{u}}^{\bar{u}}K(i)f(i)K(j)f(j)\\
&\quad\cdot[q(i,j)-q(i)q(j)]d\vec{y}_i d\vec{u}_i d\vec{y}_j d\vec{u}_j\\
&\qquad = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\{\int_{\underline{y}}^{\bar{y}}\cdot\int_{\underline{y}}^{\bar{y}}\int_{\underline{u}}^{\bar{u}}\cdot\int_{\underline{u}}^{\bar{u}}\frac{1}{r^{2n}}\right.\\
&\qquad\quad\cdot K(i)f(i)\frac{1}{r^{2n}}K(j)\cdot f(j)[q(i\,|\,j)\\
&\qquad\quad\left. -\,q(i)]q(j)d\vec{y}_i d\vec{u}_i d\vec{y}_j d\vec{u}_j\right\}
\end{aligned} \quad (7)$$

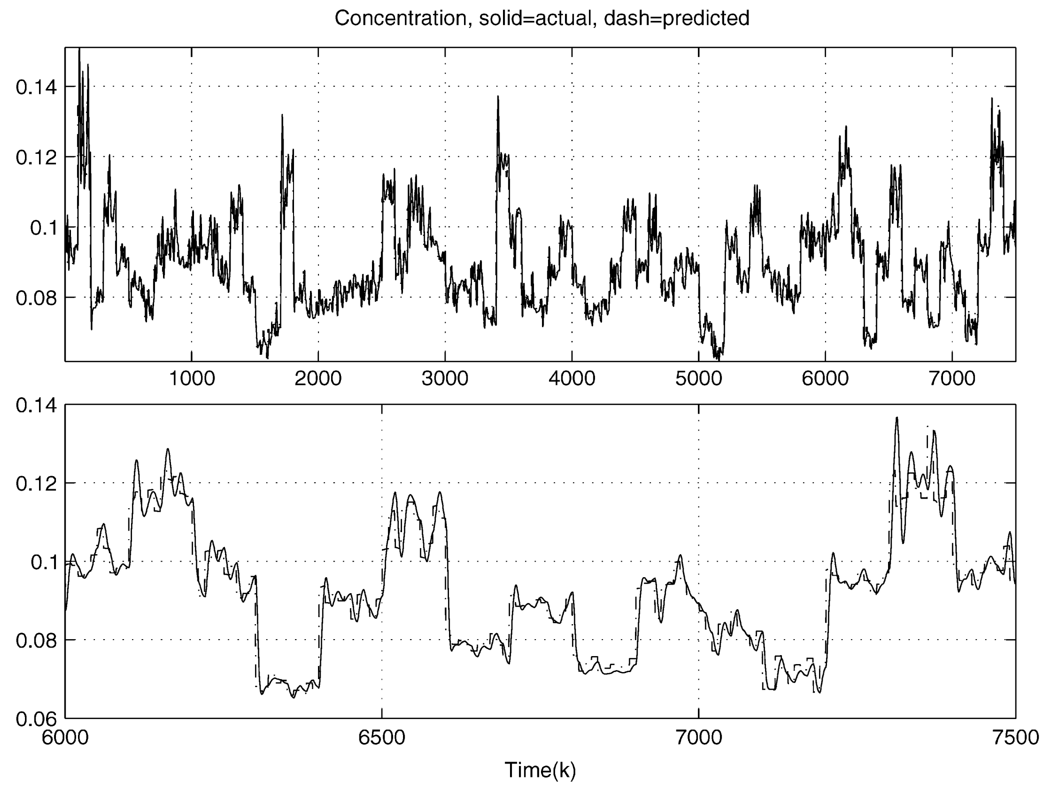


Fig. 1. Actual concentration (solid) and the predicted concentration (dash-dotted).

and the last term of (6) becomes, because $v(j)$ is independent of $K(i)$ and $K(j)$ for $j \geq i$

$$\begin{aligned}
&\frac{2}{r^{4n}N^2}\mathbf{E}\sum_{i=1}^{N}K(i)f(i)\cdot\sum_{j=1}^{i-1}K(j)v(j)\\
&=\frac{2}{r^{4n}N^2}\mathbf{E}\sum_{i=1}^{N}K(i)f(i)\\
&\quad\cdot\sum_{j=1}^{i-1}K(j)v(j)-\frac{2}{r^{4n}N^2}\\
&\quad\cdot\mathbf{E}\sum_{i=1}^{N}K(i)f(i)\mathbf{E}\sum_{j=1}^{i-1}K(j)v(j)\\
&=\frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\int_{\underline{y}}^{\bar{y}}\cdot\int_{\underline{y}}^{\bar{y}}\int_{\underline{u}}^{\bar{u}}\cdot\int_{\underline{u}}^{\bar{u}}\int_{\underline{v}}^{\bar{v}}\frac{1}{r^{4n}}\\
&\quad\cdot K(i)f(i)K(j)v(j)\\
&\quad\cdot[q(i\,|\,j,v(j))-q(i)]\\
&\quad\cdot q_v(j)q(j)d\vec{y}_i d\vec{u}_i d\vec{y}_j d\vec{u}_j dv(j). \qquad (8)
\end{aligned}$$

We make two observations.

- Note that $K(i) \neq 0$ only if $\underline{y} \leq ((y_l - y(i-l))/r) \leq \bar{y}$ and $((\underline{u} \leq u_l - u(i-l))/r) \leq \bar{u}$, $l = 1, \ldots, n$. By defining s_{il}, s_{jl}, w_{il}, and w_{jl}
$$y_l - y(i-l) = rs_{il} \quad u_l - u(i-l) = rw_{il}$$
$$y_l - y(j-l) = rs_{jl} \quad u_l - u(j-l) = rw_{jl}$$
and substituting new variables in the integrals, it can be easily verified that both integrals (7) and (8) are bounded, say bounded by a constant C.
- Because of exponential input-to-output stability as shown in Lemma II.2 and Corollary II.1, there exist some M_4, M_5 and $0 \leq \lambda < 1$, and for $|i-j| \geq M_5$
$$|q(i|j) - q(i)| \leq M_4\lambda^{|i-j|}, |q(i|j, v(j)) - q(i)| \leq M_4\lambda^{|i-j|}.$$

It follows that

$$\begin{aligned}
&|\mathbf{E}(nu_N)^2 - (\mathbf{E}nu_N)^2|\\
&\leq O\left(\frac{1}{r^{2n}N}\right)+\frac{1}{N^2}\sum\sum_{i,j=1,\ldots,N,|i-j|<M_5}C\\
&\quad+\frac{1}{N^2}\sum\sum_{i,j=1,\ldots,N,|i-j|\geq M_5}M_4\lambda^{|i-j|}\\
&\quad+\frac{2}{N^2}\sum_{i=1}^{N}\sum_{j>i-M_5}C+\frac{2}{N^2}\sum_{i=1}^{N}\sum_{j\leq i-M_5}M_4\lambda^{|i-j|}\\
&\leq O\left(\frac{1}{r^{2n}N}\right)+O\left(\frac{1}{N}\right)+\frac{M_4}{N}\frac{2\lambda^{M_5}}{1-\lambda}+O\left(\frac{1}{N}\right)\\
&\quad+\frac{2M_4}{N}\frac{2\lambda^{M_5}}{1-\lambda}=O\left(\frac{1}{r^{2n}N}\right)+O\left(\frac{1}{N}\right).
\end{aligned}$$

This bounds the first term in (4). We now consider the second term in (4)

$$\begin{aligned}
&\mathbf{E}nu_N\\
&=\mathbf{E}\frac{1}{r^{2n}N}\sum_{i=1}^{N}K\left(\frac{y_1-y(i-1)}{r},\ldots,\frac{y_n-y(i-n)}{r},\right.\\
&\quad\left.\frac{u_1-u(i-1)}{r},\ldots,\frac{u_n-u(i-n)}{r}\right)\\
&\quad\cdot(f(y(i-1),\ldots,y(i-n)\\
&\quad\cdot u(i-1),\ldots,u(i-n))+v(i))
\end{aligned}$$

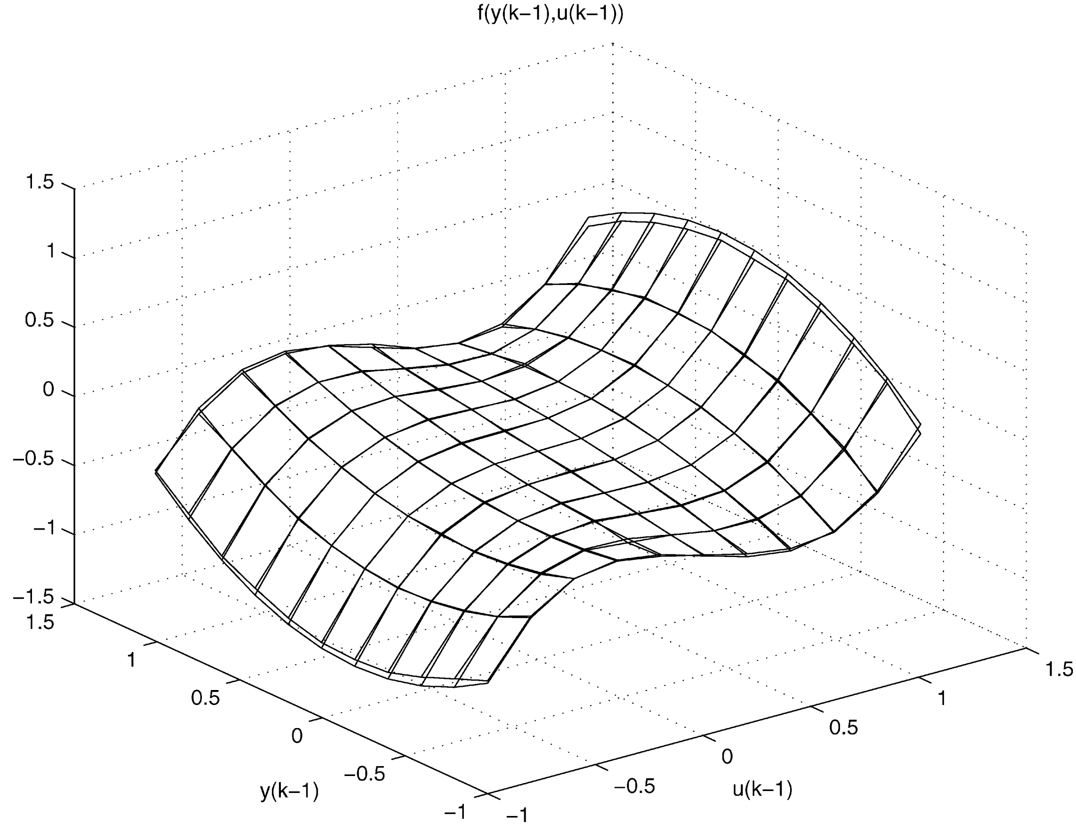


Fig. 2. $f(y,u)$ superimposed with its estimate $\hat{f}(y,u)$.

$$= \frac{1}{r^{2n}N}\sum_{i=1}^{N}\int_{\underline{y}}^{\bar{y}}\cdots\int_{\underline{y}}^{\bar{y}}\int_{\underline{u}}^{\bar{u}}\cdots\int_{\underline{u}}^{\bar{u}} K(i)f(i)q(i)d\vec{y}_i d\vec{u}_i$$
$$= \frac{1}{r^{2n}}\int_{y_1-r\bar{y}}^{y_1-r\underline{y}}\cdots\int_{y_n-r\bar{y}}^{y_n-r\underline{y}}\int_{u_1-r\bar{u}}^{u_1-r\underline{u}}\cdots\int_{u_n-r\bar{u}}^{u_n-r\underline{u}} K(i) \cdot f(i)q(i)d\vec{y}_i d\vec{u}_i$$
$$= \int_{\underline{y}}^{\bar{y}}\cdots\int_{\underline{y}}^{\bar{y}}\int_{\underline{u}}^{\bar{u}}\cdots\int_{\underline{u}}^{\bar{u}} K(s_1,\ldots,s_n,w_1,\ldots,w_n) \cdot f(y_1-rs_1,\ldots,y_n-rs_n,u_1-rw_1,\ldots,u_n-rw_n) \cdot q(y_1-rs_1,\ldots,y_n-rs_n,u_1-rw_1,\ldots,u_n-rw_n) \cdot ds_1\cdots ds_n dw_1\cdots dw_n$$
$$= f(y_1,\ldots,y_n,u_1,\ldots,u_n)q(y_1,\ldots,y_n,u_1,\ldots,u_n) \cdot \int_{\underline{y}}^{\bar{y}}\cdots\int_{\underline{y}}^{\bar{y}}\int_{\underline{u}}^{\bar{u}}\cdots\int_{\underline{u}}^{\bar{u}} K(s_1,\ldots,s_n,w_1,\ldots,w_n) \cdot ds_1\cdots ds_n dw_1\cdots dw_n + \int_{\underline{y}}^{\bar{y}}\cdots\int_{\underline{y}}^{\bar{y}}\int_{\underline{u}}^{\bar{u}}\cdots\int_{\underline{u}}^{\bar{u}} K(s_1,\ldots,s_n,w_1,\ldots,w_n) \cdot O(r)ds_1\cdots ds_n dw_1\cdots dw_n$$
$$= f(y_1,\ldots,y_n,u_1,\ldots,u_n) \cdot q(y_1,\ldots,y_n,u_1,\ldots,u_n) + O(r)$$

for sufficiently small r. This implies that the second term in (4) is given by

$$[\mathbf{E}nu_N - f(y_1,\ldots,y_n,u_1,\ldots,u_n) \cdot q(y_1,\ldots,y_n,u_1,\ldots,u_n)]^2 = O(r).$$

Combing the first and the second term in (4), we have

$$\mathbf{E}|(nu_N) - f(y_1,\ldots,y_n,u_1,\ldots,u_n) \cdot q(y_1,\ldots,y_n,u_1,\ldots,u_n)|^2 \le O\left(\frac{1}{r^{2n}N}\right) + O\left(\frac{1}{N}\right) + O(r)$$

and this implies

$$nu_N \to q(y_1,\ldots,y_n,u_1,\ldots,u_n)f(y_1,\ldots,y_n,u_1,\ldots,u_n).$$

Convergence of (3) can be similarly established. This completes the proof. □

IV. Examples

To show the efficacy of the proposed algorithm without prior structural information, we tested the proposed method on three examples. The first one is a real world problem, a liquid-saturated steam heat exchanger, where the input is the liquid flow rate and the output is the outlet liquid temperature. The data set is obtained from the classical Identification Database DaISy (www.esat.kuleuven.ac.be/sista/daisy/). Advantages using real data are obvious. A disadvantage is that no information on the actual nonlinear function f is available and thus it is not easy to compare the actual f with the estimated $\hat{f}$. An indirect way is to compare the predicted output $\hat{y}(k)$ based on the estimated nonlinear function $\hat{f}$ with the actual output $y(k)$. To be able to compare f and $\hat{f}$ directly, we also include a computer simulation example, where f is known exactly, though no information about it was used in simulation. Finally, a three dimensional nonlinear benchmark system [14] is used.

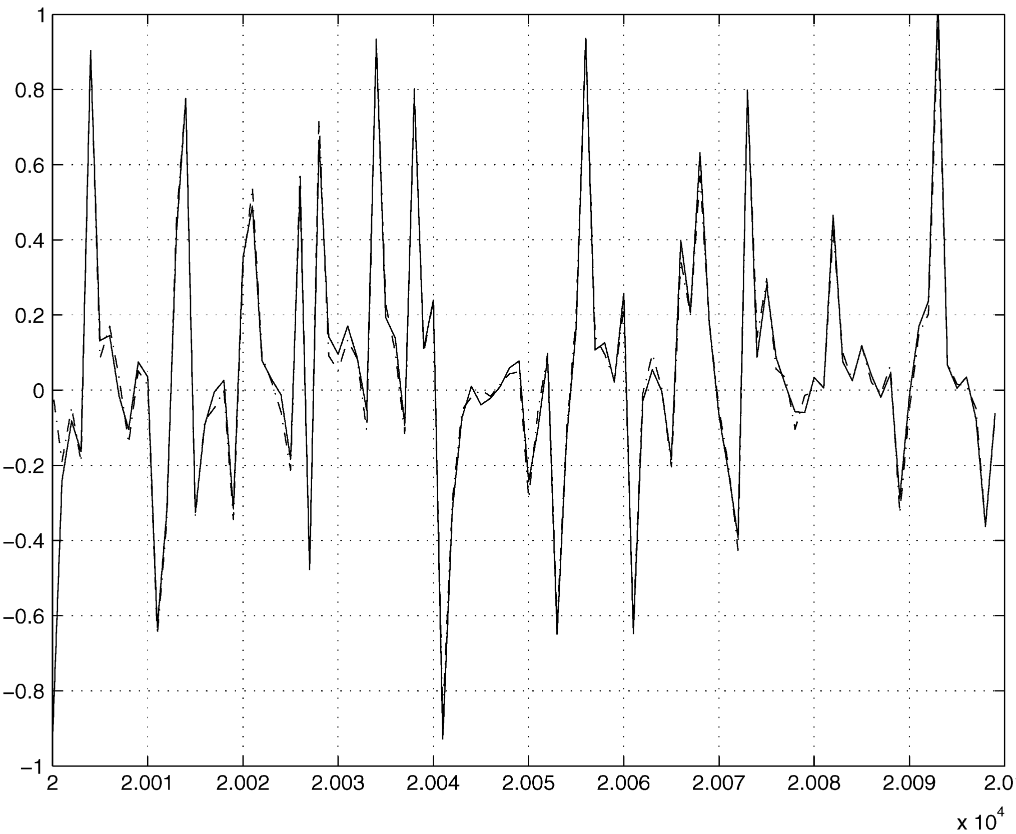


Fig. 3. Actual output $y(k)$ (solid) and the predicted output $\hat{y}(k)$ (dashed–dotted) based on the estimate $\hat{f}(y, u)$.

Example 1: This is an input-output record of a continuous stirring tank reactor, where the input is the coolant flow (l/min) and the output is the concentration (mol/l). The data set consists of 7500 samples. No *a priori* knowledge on the actual model, including the structure and the order, is available. We model this tank reactor by a first order IIR nonlinear system

$$y(k) = f(y(k-1), u(k-1)).$$

The first 6000 data points were used to obtain the estimate $\hat{f}$ with $r = 0.1$. Then, the output estimates, $k = 1, \ldots, 7500$

$$\hat{y}(k) = \hat{f}(y(k-1), u(k-1))$$

were calculated and compared to the actual output $y(k)$. The results are given in Fig. 1. The top figure shows the whole range and the bottom one focuses on $y(k)$ (solid) and $\hat{y}(k)$ (dash-dotted) for $k = 6001, \ldots, 7500$. Note that $y(k)$'s, $k = 6001, \ldots, 7500$ were not used in identification and their estimates $\hat{y}(k)$'s do predict $y(k)$'s very well. This validates the identification method proposed in this paper.

Example 2: Let the unknown nonlinear system be

$$\begin{aligned} y(k) &= f(y(k-1), u(k-1)) + v(k) \\ &= 0.2y(k-1) - 0.5y(k-1)^2 u(k-1) \\ &\quad + u(k-1)^3 + v(k) \end{aligned}$$

where the inputs $u(k)$'s are iid uniformly in $[-1, 1]$ and the noise is iid uniformly in $[-0.05, 0.05]$. For simulation purposes, we take $N = 20000$ and $r = 0.05$. Fig. 2 shows $f(y, u)$ and its estimate $\hat{f}(y, u)$ which is very close to the true but unknown f, as expected. To further test the obtained estimate $\hat{f}$, we generate a new data set $k = 20001, \ldots, 20100$ and define the predicted output as

$$\hat{y}(k) = \hat{f}(y(k-1), u(k-1))$$

where $\hat{f}$ was estimated from the previous data set $k = 1, \ldots, 20000$. Fig. 3 shows the actual output $y(k)$ (solid) and its estimate (dash-dotted) $\hat{y}(k)$. The actual output and its estimate almost coincide.

Example 3: The system is defined in state–space equation form [14] and is a benchmark problem for identification

$$\begin{aligned} x_1(k+1) &= \left(\frac{x_1(k)}{1 + x_1^2(k)} + 1\right) \sin x_2(k) \\ x_2(k+1) &= x_2(k) \cos x_2(k) + x_1(k) e^{-\frac{x_1^2(k) + x_2^2(k)}{8}} \\ &\quad + \frac{u^3(k)}{1 + u^2(k) + 0.5 \cos(x_1(k) + x_2(k))} \\ y(k) &= \frac{x_1(k)}{1 + 0.5 \sin x_2(k)} + \frac{x_2(k)}{1 + 0.5 \sin x_1(k)} + v(k). \end{aligned}$$

The Gaussian noise $v(k)$ with $\sigma = 0.1$ is added as in [14]. $x_1(k)$ and $x_2(k)$ are not measurable and only $u(k)$ and $y(k)$ are available. A three-dimensional nonlinear system

$$\begin{aligned} y(k) = f(&y(k-1), y(k-2), y(k-3), u(k-1), \\ &u(k-2), u(k-3)) + v(k) \end{aligned}$$

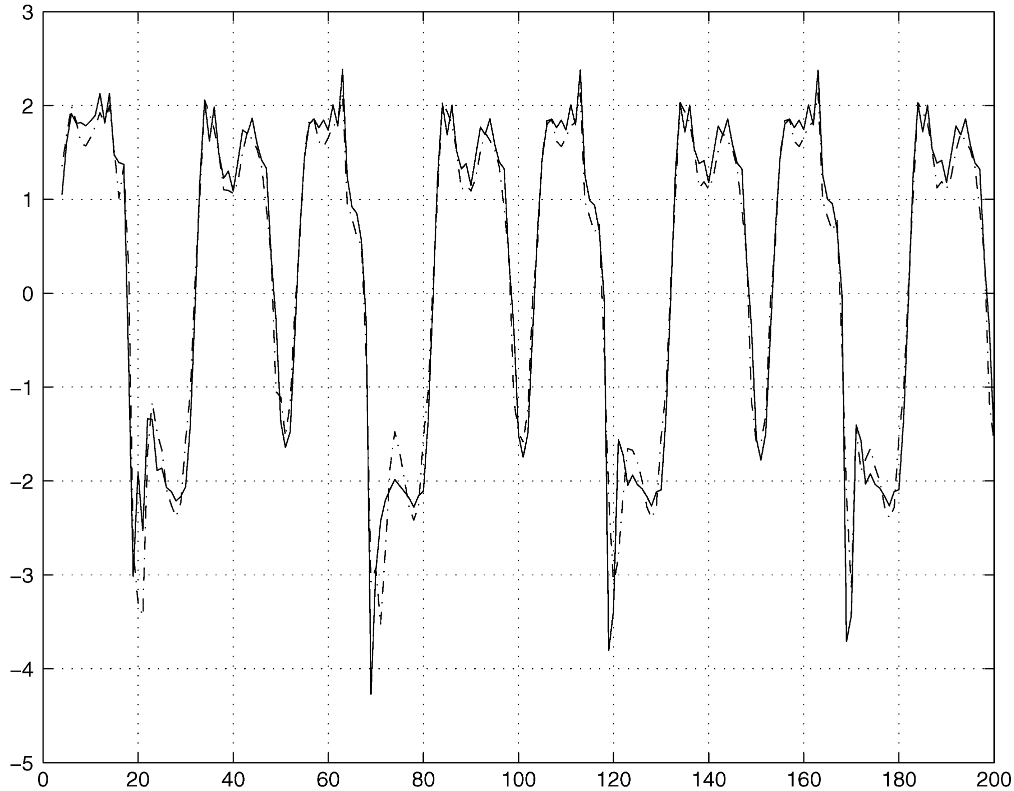


Fig. 4. Predicted (dashed–dotted) and actual outputs.

is used to model the system. As in [14], $N = 20000$ samples were generated using a uniformly distributed random input $u(k) \in [-2.5, 2.5]$. For validation, the input signal

$$u(k) = \sin\frac{2\pi k}{10} + \sin\frac{2\pi k}{25}, \qquad k = 1, \ldots, 200$$

was used. Fig. 4 shows the actual output (solid) and the predicted output (dashed–dotted) by the kernel method with $r = 0.2$. A reasonable fit is obtained.

V. Discussion

In this section, we summarize several important issues which are open and worth studying in the future.

- The first one is the choice of the bandwidth parameter r. The idea of the kernel method is to represent the unknown f locally. In fact, all measurements $y(k)$'s and $u(k)$'s so that $((y - y(k))/r) \notin [\underline{y}, \bar{y}]$ or $((u - u(k))/r) \notin [\underline{u}, \bar{u}]$ are not used to construct $\hat{f}$ in the neighborhood of (y, u). Thus, increasing r tends to reduce the variance but at the same time to increase the bias. The best choice is to balance between the bias and the variance. Some guidelines are provided in [13] for identification of a static function. It is very beneficial to investigate if these guidelines are also applicable to IIR nonlinear system identification. Some random sampling techniques may be also useful [25].
- Another topic relates to the expression of the kernel estimate. Note that the kernel estimate may be written as

$$\hat{f}_N(y_1, \ldots, y_n, u_1, \ldots, u_n) = \sum_{i=1}^{N} \alpha_i g_i$$

with

$$\alpha_i = y(i)$$

$$g_i = \frac{K\left(\frac{y_1 - y(i-1)}{r}, \ldots, \frac{u_n - u(i-n)}{r}\right)}{\sum_{j=1}^{N} K\left(\frac{y_1 - y(j-1)}{r}, \ldots, \ldots, \frac{u_n - u(j-n)}{r}\right)}.$$

This expression is reminiscent of the series representation of the unknown f by basis functions g_i's with two important distinctions. One is that the number of terms N is exactly the same as the number of the data points. Clearly, there is no reason why two numbers should coincide. If we set $\hat{f}_N = \sum_{i=1}^{m(N)} \alpha_i g_i$ where the order $m(N)$ is a function of N, is there an optimal order $m(N)$? The second distinction is that the basis functions g_i's depend on the measurements but all the basis functions g_i's have a common denominator. In other words, g_i's are globally tuned. Obviously, tuning each g_i locally would give rise to a better performance.
- The other topic is the order estimation. In this paper, the order n is assumed to be known. How to estimate n in nonlinear system identification without prior structural information is an interesting question.
- In this paper, the input range $[\underline{u}, \bar{u}]$ is assumed to be available. In some applications, the input range itself could be unknown and needs to be estimated. This would be another interesting problem.
- To achieve convergence, exponential input-to-output stability is assumed. This is a sufficient condition, but it is not a necessary condition. A large number of numerical simulations clearly shows that even for some systems that are

not exponentially input-to-output stable, the kernel method proposed in the paper still converges. Now, the question is what constitutes a necessary condition for the kernel method to be convergent. This is a tough but a very relevant question. Intuition is that the summation of the cross term contributions must grow slower than $O(N^2)$ so that the average effect converges to zero.

VI. Concluding Remarks

In this paper, identification of IIR nonlinear systems without prior structural information is studied, and asymptotic convergence properties of the kernel method are shown. To the best of our knowledge, this is the first convergent result of the kernel method based on a stability condition of a unknown nonlinear system. The numerical test results are very promising. We hope that the work reported in this paper will lead to continue research on nonlinear system identification without prior structural information.

References

[1] E. W. Bai, "Frequency domain identification of Hammerstein models," *IEEE Trans. Autom. Control*, vol. 48, no. 4, pp. 530–542, Apr. 2003.

[2] S. A. Billings, S. Chen, S. , and M. J. Kronenberg, "Identification of MIMO nonlinear systems using a forward-regression orthogonal estimator," *Int. J. Control*, vol. 49, pp. 2157–2189, 1988.

[3] S. Chen, X. Hong, C. Harris, and X. Wang, "Identification of nonlinear systems using generalized kernel models," *IEEE Trans. Control Syst. Technol.*, vol. 13, no. 3, pp. 401–411, May 2005.

[4] M. Duflo, *Random Iterative Models*. Berlin, Germany: Springer-Verlag, 1997.

[5] J. Fan and I. Gijbels, *Local Polynomial Modeling and Its Applications*. New York: Chapman and Hall, 1996.

[6] A. Georgiev, "Nonparametric system identification by kernel methods," *IEEE Trans. Autom. Control*, vol. AC-29, no. 3, pp. 356–358, Apr. 1984.

[7] R. Haber and L. Keviczky, *Nonlinear System Identification: Input-Output Modeling Approach*. Norwell, MA: Kluwer, 2000.

[8] R. Haber and H. Unbehauen, "Structure identification of nonlinear dynamic systems—A survey on input/output approaches," *Automatica*, vol. 26, pp. 651–677, 1990.

[9] N. Hilgert, R. Senoussi, and J. Vila, "Nonparametric identification of controlled nonlinear time varying processes," *SIAM J. Control Optim.*, vol. 39, pp. 950–960, 2000.

[10] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjoberg, and Q. Zhang, "Nonlinear block-box models in system identification: Mathematical foundations," *Automatica*, vol. 31, pp. 1725–1750, 1995.

[11] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[12] M. Milanese and C. Novara, "Set membership identification of nonlinear systems," *Automatica*, vol. 40, pp. 957–975, 2004.

[13] E. A. Nadaraya, *Nonparametric Estimation of Probability Densities and Regression Curves*. Dordrecht, The Netherlands: Kluwer, 1989.

[14] K. Narendra and S. Li, "Neural networks in control systems," in *Mathematical Perspectives on Neural Networks*, P. Smolensky, M. Mozer, and D. Rumelhart, Eds. Mahwah, NJ: Lawrence Erlbaum, 1996, ch. 11.

[15] E. Parzen, "An estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.

[16] B. Portier and A. Oulidi, "Nonparametric estimation and adaptive control of functional autoregressive models," *SIAM J. Control Optim.*, vol. 39, pp. 411–432, 2000.

[17] J. Poggi and B. Portier, "Nonlinear adaptive tracking using kernel estimators: Estimation and test for linearity," *SIAM J. Control Optim.*, vol. 39, pp. 707–727, 2000.

[18] W. Rugh, *Nonlinear System Theory*. Baltimore, MD: Johns Hopkins Univ. Press, 1981.

[19] I. Scott and B. Mulgrew, "Nonlinear system identification and prediction using orthonormal functions," *IEEE Trans. Signal Process.*, vol. 45, no. 7, pp. 1842–1853, Jul. 1997.

[20] J. Sjoberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarsson, and A. Juditsky, "Nonlinear black-box modeling in system identification: A unified overview," *Automatica*, vol. 31, pp. 1691–1724, 1995.

[21] E. Sontag, "The ISS philosophy as a unifying framework for stability-like behavior," in *Nonlinear Control in the Year 2000*, A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, Eds. Berlin, Germany: Springer-Verlag, 2000, vol. 2, pp. 443–468.

[22] T. Soderstrom and P. Stoica, *System Identification*. New York: Prentice-Hall, 1989.

[23] T. Soderstrom, P. Van den Hof, B. Wahlberg, and S. Weiland, Eds., "Special issue on data-based modeling and system identification," *Automatica*, vol. 41, no. 3, pp. 357–562, 2005.

[24] L. Ljung and A. Vicino, "Special issue on identification," *IEEE Trans. Autom. Control*, vol. 50, no. 10, pp. 1477–1634, Oct. 2005.

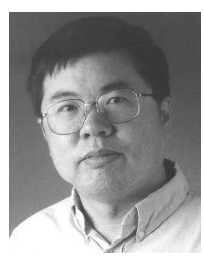
[25] R. Tempo, G. Calafiore, and F. Dabbene, *Randomized Algorithms for Analysis and Control of Uncertain System*. London, U.K.: Spring-Verlag, 2005.

[26] J. Voros, "Parameter identification of discontinuous Hammerstein systems," *Automatica*, vol. 33, no. 6, pp. 1141–1146, 1997.

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

[28] L. L. Xie and L. Guo, "How much uncertainty can be dealt with by feedback," *IEEE Trans. Autom. Control*, vol. 45, no. 12, pp. 2203–2217, Dec. 2000.

[29] Q. Zhang, "Using wavelet network in nonparametric estimation," *IEEE Trans. Neural Netw.*, vol. 8, no. 2, pp. 227–236, Mar. 1997.



Er-Wei Bai (M'90–SM'02–F'03) was educated at Fudan University and Shanghai Jiaotong University, both in Shanghai, China, and the University of California, Berkeley.

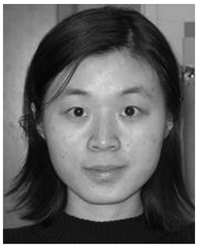
He is a Professor of Electrical and Computer Engineering at the University of Iowa, Iowa City, where he teaches and conducts research in identification, control, and their applications in medicine and communication.

Dr. Bai was a recipient of the President's Award for Teaching Excellence.

Roberto Tempo (M'90–SM'98–F'00) was born in Cuorgne, Italy, in 1956. He graduated in electrical engineering from the Politecnico di Torino, Torino, Italy, in 1980.

After a period spent at the Dipartimento di Automatica e Informatica, Politecnico di Torino, he joined the National Research Council of Italy (CNR) at the research institute IEIIT, Torino, where he has been a Director of Research of Systems and Computer Engineering since 1991. He has held visiting and research positions at Kyoto University, Kyoto, Japan, the University of Illinois at Urbana-Champaign, German Aerospace Research Organization, Oberpfaffenhofen, Germany, and Columbia University, New York. His research activities are mainly focused on complex systems with uncertainty, and related applications. He is the author or coauthor of more than 130 research papers published in international journals, books, and conferences. He is also a coauthor of the book *Randomized Algorithms for Analysis and Control of Uncertain Systems* (London, U.K.: Springer-Verlag, 2005).

Dr. Tempo is a recipient of the "Outstanding Paper Prize Award" from the International Federation of Automatic Control (IFAC) for a paper published in *Automatica*, and of the "Distinguished Member Award" from the IEEE Control Systems Society. He is currently the Editor for Technical Notes and Correspondence of the IEEE Transactions on Automatic Control. He is also an Editor and Deputy Editor-in-Chief of *Automatica*. He was Vice-President for Conference Activities of the IEEE Control Systems Society from 2002 to 2003 and a member of the EUCA Council from 1998 to 2003.



Yun Liu received the B.E. and M.E. degrees from Shanghai Jiao Tong University, Shangai, China, in 1998 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Iowa, Iowa City, in 2006.

Her research interests are dynamical system identification, signal processing, and control methods for vehicle applications. She is now a control engineer with Servo Tech, Inc., Chicago, IL.