Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Identification of Non-parametric Nonlinear Systems with Low Degree Interactive Terms ⋆

Er-Wei Bai * **Kung-Sik Chan** ** **Colbin Erdahl** *

* *Dept. of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242 USA (e-mail: er-wei-bai, cerdahl@uiowa.edu).*
** *Dept. of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242 USA (e-mail: kung-sik-chan@uiowa.edu).*

Abstract: In this paper, an interactive term identification approach is proposed for identification of non-parametric nonlinear systems. The idea is to make a high-dimensional nonlinear identification problem into a number of low-dimensional problems and thus to effectively combat the problem of the curse of dimensionality. Convergence results are established in the paper and numerical results support the theoretical analysis and demonstrate that the proposed approach is an attractive alternative to existing nonlinear identification methods.

1. INTRODUCTION

Nonlinear system identification is usually the first step in nonlinear system analysis and design. Despite progress made in recent years in Haber et al. (1990); Juditsky et al. (1995); Ljung et al. (2005); Sjoberg et al. (1995); Soderstrom et al. (2005), development of nonlinear system identification is still in its early stage. In particular, non-parametric nonlinear system identification without a priori structural information poses a very tough problem. This is partially because the nonlinear structure is too rich and no single representation could cover all possibilities.

Consider a general non-parametric nonlinear finite impulse response (FIR) system

$$\begin{aligned} y[k] &= f(u[k-1],u[k-2],...,u[k-n])+v[k] \\ &= \bar{c}+\sum_{j=1}^{n}\bar{f}_j(u[k-j])+\sum_{1\le j_1<j_2\le n}\bar{f}_{j_1j_2}(u[k-j_1],u[k-j_2])+... \\ &\quad +\sum_{1\le j_1<j_2<...<j_{n-1}\le n}\bar{f}_{j_1\ldots j_{n-1}}(u[k-j_1],...,u[k-j_{n-1}]) \\ &\quad +\bar{f}_{12\ldots n}(u[k-1],u[k-2],...,u[k-n])+v(k), \\ & k=1,2,...,N \end{aligned} \tag{1.1}$$

where $y[k]$ and $u[k]$ are output and input measurements with the input $u[k]$ being iid random sequence in a possibly unknown interval $I \in R$ with a (unknown) probability density function $\psi(\cdot)$, and the noise $v[k]$ is a sequence of independent random variables (not necessarily identically distributed) with zero mean and uniformly bounded variance. The functions $\bar{f}_{j_1j_2\ldots j_l}$'s, referred to as l-factor terms, are unknown and describe interactions of variables $u(k-j_1), u(k-j_2), ..., u(k-j_l)$. No structural prior information on $\bar{f}_{j_1j_2\ldots j_l}$, $l=1,2,...,n$ is assumed.

A common aim of most methods in literature is to find directly the nonlinearity f representing the input-output relationship of the system. This amounts to solving a high dimensional nonlinear identification problem directly and is usually difficult if the order or dimension n is not small. One of the main challenges is the curse of dimensionality in non-parametric identification. To illustrate the situation, let $u(\cdot)$ be uniformly distributed in $I=[-0.5,0.5]$. Suppose one wants to estimate $f(x_1,x_2,...,x_n)$ at a point $(x_1,x_2,...,x_n)\in I^n$. Since any identification scheme is in some form of local smoother or weighted average based on the measurement data in the neighborhood of $(x_1,x_2,...,x_n)$, there must be enough local data in the neighborhood to average out the effects of noise and the uncertainty due to lack of structural information. For simplicity, suppose the neighborhood is a hyper-box with the side length 0.1. Then, the volume of I^n is $1^n=1$ and the volume of the neighborhood is 0.1^n. This implies the probability that a measurement data $(u[k-j_1],u[k-j_2],...,u[k-j_n])$ is in the neighborhood of $(x_1,x_2,...,x_n)$ is $0.1^n/1=0.1^n$ that goes to zero exponentially as the order or dimension n gets larger. Let N be the number of total data measurements. For a large N, it is likely there are $N\cdot 0.1^n$ measurements in the neighborhood. Unless N is huge, there is not enough data in a neighborhood for identification purpose for moderately large n.

What we are interested in this paper is not a general nonlinear system as in (1.1) but nonlinear systems with a low degree of interactions, i.e., the systems that contain no more than 3-factor interaction terms:

$$\begin{aligned} y[k] &= \bar{c}+\sum_{j=1}^{n}\bar{f}_j(u[k-j]) \\ &+\sum_{1\le j_1<j_2\le n}\bar{f}_{j_1j_2}(u[k-j_1],u[k-j_2]) \\ &+\sum_{1\le j_1<j_2<j_3\le n}\bar{f}_{j_1j_2j_3}(u[k-j_1],u[k-j_2],u[k-j_3]) \\ &+v(k). \end{aligned} \tag{1.2}$$

⋆ This work was supported in part by NSF ECS-0555394 and NIH/NIBIB EB004287.

978-1-1234-7890-2/08/$20.00 © 2008 IFAC **10.3182/20080706-5-KR-1001.1192**

The 3-factor terms, $\bar{f}_{j_1j_2j_3}$'s, are zero if the system is known to contain at most 2-factor terms. For the system (1.2), we propose a radically different framework for non-parametric nonlinear system identification by fully utilizing the fact that the interaction among the variables $u[k-1], u[k-2], ..., u[k-n]$ is of low degree. Our aim is not to estimate the high dimensional f directly but to estimate the unknown interactive terms $\bar{f}_j$, $\bar{f}_{j_1j_2}$ and $\bar{f}_{j_1j_2j_3}$ as well as the unknown constant, $\bar{c}$ based on the input and output measurements. Moreover, identification of each interactive term must be decoupled with each other in some sense. This is very beneficial. For instance, suppose the system is known to contain at most 2-factor terms, for example, bilinear systems. Continue the example discussed above with $u(\cdot)$ uniformly in $I = [-0.5, 0.5]$ and $n = 5$. Then, the problem becomes identification of five 1-dimensional 1-factor terms $\bar{f}_j(u[k-j])$, $j = 1, 2..., 5$, and ten 2-dimensional 2-factor terms $\bar{f}_{j_1j_2}(u[k-j_1], u[k-j_2])$, $1 \le j_1 < j_2 \le 5$. Though the number of identifications is increased, the complexity of identification is reduced drastically. In addition to decoupling the identification of those fifteen 1-factor or 2-factor terms, identification of each interactive term is much simpler. Because of decoupling, the probability of an $u[k-j]$ in the neighborhood of x_j for one-dimensional identification is $0.1/1 = 0.1$ and the probability of $(u[k-j_1], u[k-j_2])$ in the neighborhood of (x_{j_1}, x_{j_2}) is $0.1^2/1 = 0.1^2$ for two-dimensional identification. Suppose the total number of data points is $N = 10^4$. This implies that it is likely there are 10^3 or 10^2 measurements in the neighborhood for identification of 1-factor or 2-factor terms, respectively. Recall that if the 5-dimensional $f(x_1, x_2, x_3, x_4, x_5)$ is identified directly, the probability that a data vector is in the neighborhood of $(x_1, x_2, x_3, x_4, x_5)$ is 0.1^5. With $N = 10^4$, the probability that there is one measurement in a neighborhood is 0.1. That makes that identification is nearly impossible in the presence of noise, or the identification error will be large. Clearly, the performance of identification of the 1-factor or 2-factor term can be substantially improved for the same N, compared to the identification of a five-dimensional problem f. This effectively combats the curse of dimensionality. In a sense, the approach proposed here is to replace a difficult high dimensional problem by a number of less-difficult and manageable low dimensional problems.

The contribution of this paper is four-fold:

- A model is proposed for a general FIR nonlinear system with a low degree of interaction term that emphasize the interactions between variables.
- A normalization procedure is established that makes identification of each interactive term separable in some sense.
- An identification algorithm is proposed which is convergent and is effective in combating the curse of dimensionality for nonlinear systems with low degree of interaction terms.
- A relative contribution method for order determination and regressor selection is proposed and tested.

1.1 System and identification

The purpose of identification is to estimate $\bar{c}$, $\bar{f}_j$, $\bar{f}_{j_1j_2}$ and $\bar{f}_{j_1j_2j_3}$ based on the input and output measurements. Immediately, we notice that the representation of (1.2) is actually not unique and ill-defined for identification purposes. For instance, $\bar{f}_1(\cdot)+c$ and $\bar{f}_2(\cdot)-c$, for any constant c, would produce identical input-output measurements. Hence, the system has to be normalized for identification purposes. To this end, we propose an normalization process which guarantees identifiability and moreover makes separation of each term possible. The idea can be illustrated on a system with only 1-factor terms

$$\begin{aligned} y[k] &= f(u[k-1], ..., u[k-n]) + v[k] \\ &= \bar{c} + \sum_{j=1}^{n} \bar{f}_j(u[k-j]) + v[k]. \end{aligned}$$

Let $\mathbf{E}$ denote the expectation operator

$$\mathbf{E}f(u[k-1], ..., u[k-n]) = \int_I ... \int_I f(x_1, ..., x_n)\psi(x_1)...\psi(x_n)dx_1...dx_n$$

where $\psi(\cdot)$ is the unknown probability density function of $u(\cdot)$ and I is the interval in which the input lies, and $\mathbf{E}_j$ the expectation operator with respect to the variable $u[k-j]$ that averages out $u[k-j]$ from the argument list,

$$\mathbf{E}_j f(u[k-1], ..., u[k-n]) = \int_I f(u[k-1], ..., u[k-j+1], x, u[k-j-1], ..., u[k-n])\psi(x)dx.$$

Now, apply the identity operator to the system

$$\begin{aligned} y[k] &= \prod_{\gamma=1}^{n}(I_d - \mathbf{E}_\gamma + \mathbf{E}_\gamma)(\bar{c} + \sum_{j=1}^{n} \bar{f}_j(u[k-j])) + v[k] \\ &= \bar{c} + \sum_{j=1}^{n} \mathbf{E}_j \bar{f}_j(u[k-j]) \\ &+ \sum_{j=1}^{n}\{\bar{f}_j(u[k-j]) - \mathbf{E}_j \bar{f}_j(u[k-j])\} + v[k] \\ &= \underbrace{\bar{c} + \sum_{j=1}^{n} \mathbf{E}\bar{f}_j(u[k-j])}_{c} \\ &+ \sum_{j=1}^{n} \underbrace{\{\bar{f}_j(u[k-j]) - \mathbf{E}\bar{f}_j(u[k-j])\}}_{f_j(u[k-j])} + v[k] \end{aligned}$$

where I_d is the identity operator. It is trivially verified that the f_j's are orthogonal

$$\mathbf{E}f_j(u[k-j]) = 0 \ \text{ and }$$
$$\mathbf{E}\{f_{j_1}(u[k-j_1])f_{j_2}(u[k-j_2])\} = 0, \ j_1 \neq j_2.$$

The idea can be easily extended to a general system with arbitrary higher factor terms by repeatedly applying the identify operator $\prod_{\gamma=1}^{n}(I_d - \mathbf{E}_\gamma + \mathbf{E}_\gamma)$ and grouping proper terms together. Before presenting the main results of the section, some notation has to be defined. Let the conditional expectations or marginal integrations be represented by

$$\mathbf{E}\{f_j(u[k-j]) \mid u[k-i] = x_i\} = \begin{cases} f_i(x_i) & i = j \\ \mathbf{E}f_j(u[k-j]) & i \neq j \end{cases}$$

Other conditional expectations are similarly defined.

Theorem 1.1. Consider the system (1.2) with up to 3-factor interaction terms. Then we have:

(1) The system (1.2) can be represented by

$$\begin{aligned} y[k] &= c + \sum_{j=1}^{n} f_j(u[k-j]) \\ &+ \sum_{1\le j_1<j_2\le n} f_{j_1j_2}(u[k-j_1],u[k-j_2]) \\ &+ \sum_{1\le j_1<j_2<j_3\le n} f_{j_1j_2j_3}(u[k-j_1],u[k-j_2],u[k-j_3]) \\ &+v[k] \end{aligned} \qquad (1.3)$$

for some constant c and some functions f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$, where the expectation and the conditional expectations satisfy

$$\begin{aligned} &\mathbf{E}y[k] = c, \\ &\mathbf{E}f_j(u[k-j]) = 0, \\ &\qquad 1 \le j \le n \\ &\mathbf{E}\{f_{j_1j_2}(u[k-j_1],u[k-j_2]) \mid u[k-j]\} = 0, \\ &\qquad 1 \le j_1 < j_2 \le n,\ 1 \le j \le n \\ &\mathbf{E}\left\{ \begin{array}{c} f_{j_1j_2j_3}(u[k-j_1],u[k-j_2],u[k-j_3]) \\ \mid u[k-j] \end{array} \right\} = 0, \\ &\qquad 1 \le j_1 < j_2 < j_3 \le n,\ 1 \le j \le n \\ &\mathbf{E}\left\{ \begin{array}{c} f_{j_1j_2j_3}(u[k-j_1],u[k-j_2],u[k-j_3]) \\ \mid u[k-i_1],u[k-i_2] \end{array} \right\} = 0, \\ &\qquad 1 \le j_1 < j_2 < j_3 \le n,\ 1 \le i_1 < i_2 \le n \end{aligned} \qquad (1.4)$$

(2) The 1, 2 and 3-factor terms, f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$, are orthogonal. i.e.,

$$\begin{aligned} &\mathbf{E}f_j(u[k-j]) = \mathbf{E}f_{j_1j_2}(u[k-j_1],u[k-j_2]) \\ &\quad = \mathbf{E}f_{j_1j_2j_3}(u[k-j_1],u[k-j_2],u[k-j_3]) = 0, \end{aligned}$$

$$\begin{aligned} &\mathbf{E}\{f_{j_1}(u[k-j_1])f_{j_2}(u[k-j_2])\} = \\ &\qquad \mathbf{E}\left\{ \begin{array}{l} f_{i_1i_2}(u[k-i_1],u[k-i_2]) \\ \cdot f_{j_1j_2}(u[k-j_1],u[k-j_2]) \end{array} \right\} \\ &\quad = \mathbf{E}\left\{ \begin{array}{l} f_{i_1i_2i_3}(u[k-i_1],u[k-i_2],u[k-i_3]) \\ \cdot f_{j_1j_2j_3}(u[k-j_1],u[k-j_2],u[k-j_3]) \end{array} \right\} = 0 \end{aligned}$$

for $j_1 \neq j_2$, $(i_1,i_2) \neq (j_1,j_2)$ and $(i_1,i_2,i_3) \neq (j_1,j_2,j_3)$, and

$$\begin{aligned} &\mathbf{E}\{f_j(u[k-j])f_{j_1j_2}(u[k-j_1],u[k-j_2])\} = \\ &\quad \mathbf{E}\{f_j(u[k-j])f_{j_1j_2j_3}(u[k-j_1],u[k-j_2],u[k-j_3])\} \\ &\quad = \mathbf{E}\left\{ \begin{array}{l} f_{i_1i_2}(u[k-i_1],u[k-i_2]) \\ \cdot f_{j_1j_2j_3}(u[k-j_1],u[k-j_2],u[k-j_3]) \end{array} \right\} = 0. \end{aligned}$$

for any $j, j_1, j_2, j_3, i_1, i_2$.

(3) The unknown c, f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$ are the expectation and conditional expectations of the output,

$$\begin{aligned} &c = \mathbf{E}y[k], \\ &f_j(x_j) = \mathbf{E}\{y(k) \mid u[k-j] = x_j\} - c, \\ &\qquad 1 \le j \le n \\ &f_{j_1j_2}(x_{j_1},x_{j_2}) = \\ &\quad \mathbf{E}\{y[k] \mid \ u[k-j_1] = x_{j_1}, u[k-j_2] = x_{j_2}\} \\ &\quad - f_{j_1}(x_{j_1}) - f_{j_2}(x_{j_2}) - c, \\ &\qquad 1 \le j_1 < j_2 \le n \\ &f_{j_1j_2j_3}(x_{j_1},x_{j_2},x_{j_3}) = \\ &\quad \mathbf{E}\left\{ \begin{array}{l} y[k] \mid u[k-j_1] = x_{j_1}, \\ \qquad u[k-j_2] = x_{j_2}, u[k-j_3] = x_{j_3} \end{array} \right\} \\ &\quad - f_{j_1j_2}(x_{j_1},x_{j_2}) - f_{j_1j_3}(x_{j_1},x_{j_3}) - f_{j_2j_3}(x_{j_2},x_{j_3}) \\ &\quad - f_{j_1}(x_{j_1}) - f_{j_2}(x_{j_2}) - f_{j_3}(x_{j_3}) - c, \\ &\qquad 1 \le j_1 < j_2 < j_3 \le n. \end{aligned} \qquad (1.5)$$

From the above theorem, the unknown c, f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$ can be calculated from expectation and conditional expectation values or marginal integrations. Now the question is how to calculate these expectation values by empirical averages based on the available input-output measurement data set $\{y[k], u[k-1], u[k-2], ..., u[k-n]\}_1^N$. In this paper, we adopt a fairly simple yet efficient kernel approach. To this end, let $(x_1, x_2, .., x_n) \in R^n$ and each $x_j \in I$ that is the interval in which the input $u(\cdot)$ lies. Because $u[k]$'s are iid and the law of large number applies which implies

$$\frac{1}{N}\sum_{k=1}^{N} f_j^2(u[k-j]) \to \mathbf{E}f_j^2(u[k-j])$$

$$\frac{1}{N}\sum_{k=1}^{N} f_{j_1j_2}^2(u[k-j_1],u[k-j_2]) \to \mathbf{E}f_{j_1j_2}^2(u[k-j_1],u[k-j_2])$$

$$\frac{1}{N}\sum_{k=1}^{N} f_{j_1j_2j_3}^2(u[k-j_1],u[k-j_2],u[k-j_3]) \to \mathbf{E}f_{j_1j_2j_3}^2(u[k-j_1],u[k-j_2],u[k-j_3]).$$

Now, for any given $x_j \in I$, define

$$\phi_j(k) = \|u[k-j] - x_j\|_2.$$

Let $\delta > \min \phi_j(k)$ be any positive constant and define

$$M_j = \{m_j(1), m_j(2), ..., m_j(l_j)\}$$

be a set such that $k \in M_j \Leftrightarrow \delta > \phi_j(k)$. Now, let

$$w_j(k) = \begin{cases} \dfrac{\delta - \phi_j(k)}{l_j\delta - \sum_{i=1}^{l_j}\phi_j(m_j(i))} & k \in M_j \\ 0 & k \notin M_j \end{cases}.$$

Obviously, $w_j(k) \ge 0$ for all k and $\sum_{k=1}^{l_j} w_j(k) = 1$. Similarly, for a given pair $0 \le j_1 < j_2 \le n$ and $(x_{j_1}, x_{j_2}) \in I^2$, define

$$\phi_{j_1j_2}(k) = \|(u[k-j_1],u[k-j_2]) - (x_{j_1},x_{j_2})\|_2.$$

If $\delta > \min \phi_{j_1j_2}(k)$, let $M_{j_1j_2} = \{m_{j_1j_2}(1), m_{j_1j_2}(2), ..., m_{j_1j_2}(l_{j_1j_2})\}$ be a set such that $k \in M_{j_1j_2} \Leftrightarrow \delta > \phi_{j_1j_2}(k)$. Define

$$w_{j_1j_2}(k) = \begin{cases} \dfrac{\delta - \phi_{j_1j_2}(k)}{l_{j_1j_2}\delta - \sum_{i=1}^{l_{j_1j_2}}\phi_j(m_{j_1j_2}(i))} & k \in M_{j_1j_2} \\ 0 & k \notin M_{j_1j_2} \end{cases}.$$

Notice that the same properties hold

$$w_{j_1j_2}(k) \ge 0, \ \sum_{k=1}^{l_{j_1j_2}} w_{j_1j_2}(k) = 1.$$

Again, for $1 \le j_1 < j_2 < j_3 \le n$ and $(x_{j_1}, x_{j_2}, x_{j_3}) \in I^3$, define

$$\phi_{j_1j_2j_3}(k) = \|(u[k-j_1],u[k-j_2],u[k-j_3]) - (x_{j_1},x_{j_2},x_{j_3})\|_2.$$

If $\delta > \min \phi_{j_1j_2j_3}(k)$, let $M_{j_1j_2j_3} = \{m_{j_1j_2j_3}(1), m_{j_1j_2j_3}(2), ..., m_{j_1j_2j_3}(l_{j_1j_2j_3})\}$ be a set such that $k \in M_{j_1j_2j_3} \Leftrightarrow \delta > \phi_{j_1j_2j_3}(k)$. Now, define

$$w_{j_1j_2j_3}(k) = \begin{cases} \dfrac{\delta - \phi_{j_1j_2j_3}(k)}{l_{j_1j_2j_3}\delta - \sum_{i=1}^{l_{j_1j_2j_3}}\phi_j(m_{j_1j_2j_3}(i))} & k \in M_{j_1j_2j_3} \\ 0 & k \notin M_{j_1j_2j_3} \end{cases}.$$

Similarly,

$$w_{j_1j_2j_3}(k) \geq 0, \quad \sum_{k=1}^{l_{j_1j_2j_3}} w_{j_1j_2j_3}(k) = 1.$$

Now, we are in a position to define the estimates $\hat{c}$, $\hat{f}_j$, $\hat{f}_{j_1j_2}$ and $\hat{f}_{j_1j_2j_3}$ of c, f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$ respectively.

$$\hat{c} = \frac{1}{N}\sum_{k=1}^{N} y[k],$$

$$\hat{f}_j(x_j) = \sum_{k=1}^{l_j} w_j(k)y[k] - \hat{c},$$

$$\hat{f}_{j_1j_2}(x_{j_1}, x_{j_2}) = \sum_{k=1}^{l_{j_1j_2}} w_{j_1j_2}(k)y[k] - \hat{f}_{j_1}(x_{j_1}) - \hat{f}_{j_2}(x_{j_2}) - \hat{c}$$

$$\hat{f}_{j_1j_2j_3}(x_{j_1}, x_{j_2}, x_{j_3}) = \sum_{k=1}^{l_{j_1j_2j_3}} w_{j_1j_2j_3}(k)y[k]$$
$$-\hat{f}_{j_1}(x_{j_1}) - \hat{f}_{j_2}(x_{j_2}) - \hat{f}_{j_3}(x_{j_3})$$
$$-\hat{f}_{j_1j_2}(x_{j_1}, x_{j_2}) - \hat{f}_{j_1j_3}(x_{j_1}, x_{j_3}) - \hat{f}_{j_2j_3}(x_{j_2}, x_{j_3}) - \hat{c}. \quad (1.6)$$

Theorem 1.2. Consider the system (1.3) and the estimates above. For given $x_{j_1}, x_{j_2}, x_{j_3} \in I$, assume

- The unknown functions f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$ are differentiable with the Lipschitz constant L for $x_{j_1}, x_{j_2}, x_{j_3} \in I$.
- Let $\psi(\cdot)$ be the (unknown) probability density function of the input $u(\cdot)$. Then, the density function is positive at $x_{j_1}, x_{j_2}, x_{j_3}$, i.e.,
$$\psi(x_{j_1}) \neq 0, \ \psi(x_{j_2}) \neq 0, \ \psi(x_{j_3}) \neq 0.$$
- $\delta \to 0$ and $\delta^3 N \to \infty$ as $N \to \infty$.

Then, as $N \to \infty$, we have in probability

$$\hat{c} \to c$$
$$\hat{f}_j(x_j) \to f_j(x_j)$$
$$\hat{f}_{j_1j_2}(x_{j_1}, x_{j_2}) \to f_{j_1j_2}(x_{j_1}, x_{j_2})$$
$$\hat{f}_{j_1j_2j_3}(x_{j_1}, x_{j_2}, x_{j_3}) \to f_{j_1j_2j_3}(x_{j_1}, x_{j_2}, x_{j_3}).$$

The choice of the bandwidth δ in the estimates (1.6) and generally in kernel identification is important. The idea of the kernel method is to represent the unknown nonlinearities locally. In fact, all measurements so that $\phi[k] > \delta$, are not used to construct the estimates. A small δ does not necessarily imply that the achieved estimation error is small. The choice of δ balances the trade off between the bias and the variance. A large δ implies a large bandwidth interval and accordingly more data is used that results in a small variance. On the other hand, because more data points area used even those not in a close vicinity, the approximation error gets large, which gives rise to a large bias term. A small δ produces just the opposite, a large variance and a small bias. Hence, increasing δ tends to reduce the variance but at the same time increases the bias. The best choice is to balance the bias and the variance. Some guidelines are provided in Nadaraya (1989) for the choice of the bandwidth δ.

2. ORDER AND REGRESSOR SELECTION

In this paper, only the upper bound n on the order of the system is assumed. It is natural in identification to ask how to determine the actual order. A closely related issue is the regressor selection. Once the order n is determined and f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$ are estimated, the question is which f_j, $f_{j_1j_2}$ or $f_{j_1j_2j_3}$ should be included in the model and which ones should not. An easy way is to visually inspect each f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$. A more reliable way is to carry out a statistical hypothesis test to check if the interested term is zero or not. Notice that in identification, what we are interested in is not if a particular term f_j, $f_{j_1j_2}$ or $f_{j_1j_2j_3}$ contributes or not, but whether the contribution is significant or not. Identification or modelling is always a balance between model accuracy and model parsimony. In other words, a relative contribution is more important for the order and regressor selection. Also, notice that the output contains contributions from noises and the constant term c. To truly determine the relative contribution, the noise and the constant term effects should be removed in the analysis. To this end, we propose a relative contribution approach. Consider the system (1.3). Again, it is easily verified from Theorem (1.1) that in the absence of the noise, we have

$$\eta = \mathbf{E}(y[k] - c)^2 = \sum_{j=1}^{n} \mathbf{E}f_j^2(u[k-j]) +$$
$$\sum_{1 \leq j_1 < j_2 \leq n} \mathbf{E}f_{j_1j_2}^2(u[k-j_1], u[k-j_2])$$
$$+ \sum_{1 \leq j_1 < j_2 < j_3 \leq n} \mathbf{E}f_{j_1j_2j_3}^2(u[k-j_1], u[k-j_2], u[k-j_3]).$$

Apparently, an appropriate measure of the relative contribution can be defined as

$$R(f_j) = \frac{\mathbf{E}f_j^2(u[k-j])}{\eta}$$

for f_j

$$R(f_{j_1j_2}) = \frac{\mathbf{E}f_{j_1j_2}^2(u[k-j_1], u[k-j_2])}{\eta}$$

for $f_{j_1j_2}$ and

$$R(f_{j_1j_2j_3}) = \frac{\mathbf{E}f_{j_1j_2j_3}^2(u[k-j_1], u[k-j_2], u[k-j_3])}{\eta}$$

for $f_{j_1j_2j_3}$. Since the square term is proportional to energy, the meaning of the regressor contribution is the relative contribution of a particular term to the overall output in terms of energy.

Of course, in reality, f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$ are unavailable. However, their estimates $\hat{f}_j$, $\hat{f}_{j_1j_2}$ and $\hat{f}_{j_1j_2j_3}$ are available and converge to f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$ respectively. Also we have that, as $N \to \infty$,

$$\frac{1}{N}\sum f_j^2(u[k-j]) \to \mathbf{E}f_j^2(u[k-j])$$
$$\frac{1}{N}\sum f_{j_1j_2}^2(u[k-j_1], u[k-j_2])$$
$$\to \mathbf{E}f_{j_1j_2}^2(u[k-j_1], u[k-j_2])$$
$$\frac{1}{N}\sum f_{j_1j_2j_3}^2(u[k-j_1], u[k-j_2], u[k-j_3])$$
$$\to \mathbf{E}f_{j_1j_2j_3}^2(u[k-j_1], u[k-j_2], u[k-j_3])$$

Therefore, we define the estimates of η, $R(f_j)$, $R(f_{j_1j_2})$ and $R(f_{j_1j_2j_3})$ as, respectively,

$$\hat{\eta} = \frac{1}{N}\sum_{k=1}^{n}(y[k]-\hat{c})^2 = \hat{\eta} = \sum_{j=1}^{n}\frac{1}{N}\sum \hat{f}_j^2(u[k-j])$$

$$+ \sum_{1\le j_1<j_2\le n}\frac{1}{N}\sum \hat{f}_{j_1j_2}^2(u[k-j_1],u[k-j_2])$$

$$+ \sum_{1\le j_1<j_2<j_3\le n}\frac{1}{N}\sum \hat{f}_{j_1j_2j_3}^2(u[k-j_1],u[k-j_2],u[k-j_3]),$$

$$\hat{R}(f_j) = \frac{\frac{1}{N}\sum \hat{f}_j^2(u[k-j])}{\hat{\eta}}$$

$$\hat{R}(f_{j_1j_2}) = \frac{\frac{1}{N}\sum \hat{f}_{j_1j_2}^2(u[k-j_1],u[k-j_2])}{\hat{\eta}},$$

and

$$\hat{R}(f_{j_1j_2j_3}) = \frac{\frac{1}{N}\sum \hat{f}_{j_1j_2j_3}^2(u[k-j_1],u[k-j_2],u[k-j_3])}{\hat{\eta}}.$$

To determine if $\hat{f}_j$, $\hat{f}_{j_1j_2}$ or $\hat{f}_{j_1j_2j_3}$ should be included in the model, we compute $\hat{R}(f_j)$, $\hat{R}(f_{j_1j_2})$ and $\hat{R}(f_{j_1j_2j_3})$. Let the threshold d, for example d=0.05 or 5% be chosen. If $\hat{R}(f_j)$, $\hat{R}(f_{j_1j_2})$ or $\hat{R}(f_{j_1j_2j_3}) \ge d$, $\hat{f}_j$, $\hat{f}_{j_1j_2}$ or $\hat{f}_{j_1j_2j_3}$ is included. Otherwise the term is discarded. Because of the convergence, this test is very reliable for large N.

3. NUMERICAL SIMULATION

Consider a nonlinear system

$$y[k] = f\begin{pmatrix} u[k-1],u[k-2],u[k-3] \\ ,u[k-4],u[k-5] \end{pmatrix} + v(k) \qquad (3.7)$$

$$= \underbrace{1.25/3}_{c} + \underbrace{u[k-1]}_{f_1} + \underbrace{10\cdot u[k-2]^3}_{f_2} + \underbrace{5\cdot u[k-3]^2 - 1.25/3}_{f_3}$$

$$+\underbrace{0}_{f_4}+\underbrace{0}_{f_5}++\underbrace{5\cdot u[k-1]*u[k-2]}_{f_{12}}$$

$$+\underbrace{0}_{f_{13}}+\underbrace{0}_{f_{14}}+\underbrace{0}_{f_{15}}-\underbrace{0.5\cdot cos(2\pi u[k-2]+u[k-3])}_{f_{23}}$$

$$+\underbrace{0}_{f_{24}}+\underbrace{0}_{f_{25}}+\underbrace{0}_{f_{34}}+\underbrace{0}_{f_{35}}+\underbrace{0}_{f_{45}}+v[k], \quad k=1,2,...,N.$$

The prior information on the system is that it is a nonlinear system with up to 3-factor terms. No prior structural information on f, f_j,$f_{j_1j_2}$ and $f_{j_1j_2j_3}$ are available. The order of the system is also unknown and only an upper bound of $n = 5$ is assumed. For identification, one can either identify the unknown 5-dimensional system (3.7) directly, or use the interactive term method (1.6) proposed in the paper. For simulation, $N = 20,000$ and $\delta = 0.1$. The input $u[\cdot]$ is independent and uniformly distributed in $[-0.5, 0.5]$, and the noise $v[\cdot]$ is iid Gaussian with $SNR = 20dB$.

We use the interactive term method to identify each f_j, $f_{j_1j_2}$ and $f_{j_1j_2j_3}$ and calculate their relative contributions as shown in the third column of Table 1 ($N = 20,000$).

To determine the order of the system as well as which term should be included in the model, let the threshold d=5%. If $\hat{R}_j$, $\hat{R}_{j_1j_2}$, $\hat{R}_{j_1j_2j_3} \ge d$, we include the corresponding term in the model. Otherwise the contribution of the corresponding term is deemed to be insignificant and omitted in the model. Clearly, from the third column ($N = 20,000$), only the terms f_1, f_2, f_3, f_{12} and f_{23} contribute significant and should be included in the model. Simply put, the system order is determined to be $n = 3$, though the upper bound is assumed to be 5. Further, it is determined that the system contains only 5 terms, f_1, f_2,

N	30,000	20,000	15,000	10,000	5,000	$\ge d$?
$\hat{R}_1$	0.1130	0.1117	0.1040	0.1088	0.0827	$\surd$
$\hat{R}_2$	0.2770	0.2743	0.2641	0.2472	0.2100	$\surd$
$\hat{R}_3$	0.1676	0.1631	0.1539	0.1471	0.1259	$\surd$
$\hat{R}_4$	0.0001	0.0003	0.0005	0.0004	0.0012	
$\hat{R}_5$	0.0002	0.0003	0.0004	0.0006	0.0010	
$\hat{R}_{12}$	0.2295	0.2228	0.2179	0.2055	0.1724	$\surd$
$\hat{R}_{13}$	0.0010	0.0015	0.0018	0.0026	0.0037	
$\hat{R}_{14}$	0.0010	0.0016	0.0020	0.0027	0.0053	
$\hat{R}_{15}$	0.0010	0.0014	0.0021	0.0024	0.0047	
$\hat{R}_{23}$	0.1525	0.1433	0.1452	0.1322	0.1131	$\surd$
$\hat{R}_{24}$	0.0009	0.0011	0.0017	0.0022	0.0040	
$\hat{R}_{25}$	0.0009	0.0012	0.0015	0.0020	0.0046	
$\hat{R}_{34}$	0.0010	0.0013	0.0017	0.0023	0.0056	
$\hat{R}_{35}$	0.0010	0.0016	0.0023	0.0031	0.0053	
$\hat{R}_{45}$	0.0011	0.0016	0.0023	0.0031	0.0054	
$\hat{R}_{123}$	0.0019	0.0029	0.0038	0.0053	0.0084	
$\hat{R}_{124}$	0.0038	0.0056	0.0074	0.0101	0.0207	
$\hat{R}_{125}$	0.0038	0.0056	0.0074	0.0104	0.0208	
$\hat{R}_{134}$	0.0064	0.0093	0.0125	0.0178	0.0302	
$\hat{R}_{135}$	0.0065	0.0087	0.0129	0.0174	0.0303	
$\hat{R}_{145}$	0.0081	0.0103	0.0141	0.0197	0.0386	
$\hat{R}_{234}$	0.0042	0.0059	0.0078	0.0115	0.0208	
$\hat{R}_{235}$	0.0041	0.0062	0.0079	0.0114	0.0207	
$\hat{R}_{245}$	0.0064	0.0085	0.0112	0.0158	0.0316	
$\hat{R}_{345}$	0.0068	0.0100	0.0138	0.0191	0.0336	

Table 1. Relative contribution of each term for different N.

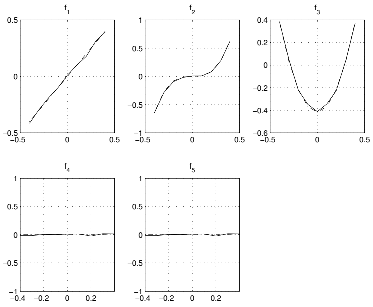

Fig. 1. $f_j(u[k-j])$'s (solid) and their estimates $\widehat{f}_j(u[k-j])$ (dashdot), $j = 1, 2, 3, 4, 5$.

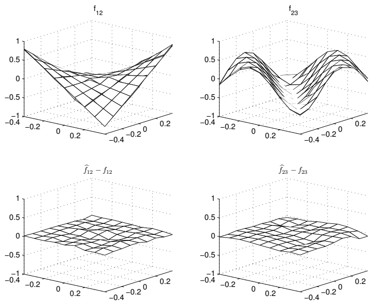

Fig. 2. f_{12}, f_{23} and $\widehat{f}_{12}$, $\widehat{f}_{23}$.

$f = 1.9564$	direct 5-dim	interactive term	improvement ratio
average estimation error	$\lvert\widehat{f} - f\rvert = 0.7623$	$\lvert\widehat{f} - f\rvert = 0.0263$	29
variance	0.9121	0.0016	573

Table 2. Average estimation error and variance of two identification methods.

f_3, f_{12} and f_{23} and all other terms including all 3-factor terms are zero. The conclusion is consistent with the true but unknown system.

Figure 1 shows the actual but unknown $f_j(u[k-j])$(solid), $j = 1, ..., 5$ and their estimates. $\widehat{f}_j(u[k-j])$ (dashdot), $j = 1, ..., 5$, respectively. The top diagrams in Figure 2 show $f_{12}(u[k-1], u[k-2])$ and $f_{23}(u[k-2], u[k-3])$ superimposed on their estimates $\widehat{f}_{12}(u[k-1], u[k-2])$ and $\widehat{f}_{23}(u[k-2], u[k-3])$ respectively. The estimation errors $\widehat{f}_{12}(u[k-1], u[k-2]) - f_{12}(u[k-1], u[k-2])$ and $\widehat{f}_{23}(u[k-2], u[k-3]) - f_{12}(u[k-2], u[k-3])$ are in the bottom diagrams. All other terms $f_{j_1 j_2}$'s are zero. It can be seen that the estimates fit the actual functions well.

Notice that in theory the estimates of the relative contributions $\hat{R}_j$, $\hat{R}_{j_1 j_2}$ and $\hat{R}_{j_1 j_2 j_3}$ converge to the actual relative contributions R_j, $R_{j_1 j_2}$ and $R_{j_1 j_2 j_3}$ as $N \to \infty$. This implies that the estimates are reliable if N is large. In practice, the question is always how large is large enough or how large does N need to be before these estimates become reliable. To this end, Table 1 shows the estimates of the relative contribution for different N. From Table 1, it is seen that as long as N is large, e.g., $N \geq 10,000$, the results are fairly robust.

To compare the results with the method that directly identifies the 5-dimensional nonlinear function (3.7), we consider identification of the nonlinearity at an arbitrary point in $[-0.5, 0.5]^5$, say at

$$(u[k-1], u[k-2], u[k-3], u[k-4], u[k-5]) = (0.3801, 0.2940, -0.4541, -0.1866, -0.3090)$$

based on the available measurement data set $\{y(k), u(k-1), ..., u(k-n)\}_1^N$. The true but unknown

$$f(0.3801, 0.2940, -0.4541, -0.1866, -0.3090) = 1.9564.$$

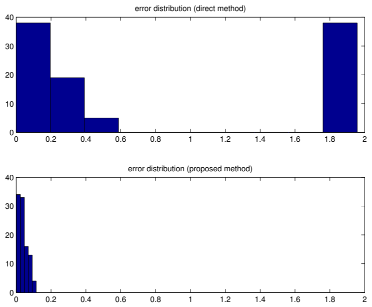

Fig. 3. Error histogram of 100 Monte Carlo runs. The horizontal axis is the identification error $|\hat{f} - f|$ and the vertical axis is the number of occurrence.

The direct method is an kernel method as used in (1.6) but with a 5-dimensional kernel. The average estimation error and variance of 100 Monte Carlo runs for both the direct method and the interactive method are listed in Table 2, having been produced under the exact simulation conditions as described above. Clearly, the interactive term method outperforms the direct identification method in terms of both bias and variance drastically. In fact, the variance is improved by a factor of 0.9121/0.0016 = 573. The reason for this improvement is that for the direct 5-dimensional identification, only a very small number of measurements are in the neighborhood and that makes identification unreliable. In fact, in almost half of the Monte Carlo runs, the estimated value of $\hat{f}$ for the direct method is zero which implies that no measurement is in the neighborhood, see Figure 3 for error histogram of the 100 Monte Carlo runs. This curse of dimensionality is unavoidable for a dimension that is not small. In the proposed interactive method, identification is projected into lower dimensional f_j's and $f_{j_1 j_2}$'s, that are much less problematic. Obviously one expects even higher improvement ratios between two methods if the dimension n gets larger.

REFERENCES

K. Godfrey. *Perturbation Signals for Identification.* Prentice-Hall, Hemel Hempstead, Hertfordshire, UK, 1993.

R. Haber, and H. Unbehauen. Structure identification of nonlinear dynamic systems-A survey on input/output approaches. *Automatica*, volume 26, pages 651-677, 1990.

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*, volume 31, pages 1725-1750, 1995.

I. Lind, and L. Ljung. Regressor selection with the ANOVA. *Automatica*, volume 41, pages 51-56, 2005.

L. Ljung, and A. Vicino. Special issue on identification. *IEEE Trans. on Auto. Contr.*, volume 50, number 10, pages 1477-1634, 2005.

E. Nadaraya. *Nonparametric Estimation of Probability Densities and Regression Curves.* Kluwer Academic Pub., Dordrecht, The Netherlands, 1989.

R. Pintelon, and J. Schoukens. *System Identification: A Frequency Domain Approach.* IEEE Press, New Jersey, 2001.

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*, volume 31, pages 1691-1724, 1995.

T. Soderstrom, P. Van den Hof, B. Wahlberg, and S. Weiland, editors. Special issue on data-based modeling and system identification. *Automatica*, volume 41, number 3, pages 357-562, 2005.

S. Sperlich, D. Tjostheim and L. Yang. Non-parametric estimation and testing of interaction in additive models. *Econometric Theory*, volume 18, pages 197-251, 2002.