Automatica 46 (2010) 1595–1603

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

# Representation and identification of non-parametric nonlinear systems of short term memory and low degree of interaction☆

Er-Wei Bai [a,*], Roberto Tempo [b]

[a] Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, United States
[b] IEIIT-CNR, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

*Article history:*
Received 19 August 2009
Received in revised form
4 February 2010
Accepted 1 June 2010
Available online 15 July 2010

*Keywords:*
System identification
Nonlinear systems
Volterra system

ABSTRACT

In this paper, we propose a new representation which is particularly useful for a class of non-parametric nonlinear systems that have short term memory and low degree of interaction. Advantages and disadvantages of this representation are discussed and compared to existing methods both theoretically and numerically. Furthermore, results regarding structural estimation based on the analysis of variance and on full scale identification are also provided.



## 1. Introduction

Unlike linear systems, representation and identification of an unknown nonlinear system is a very hard task (Ljung, 1999; Soderstrom & Stoica, 1989). It is safe to say that, to represent an unknown nonlinear system effectively, one has to exploit the structure of the system's nonlinearity. There exist several ways to represent an unknown nonlinear system with the objective to identify it. For example, basis function approach, including polynomial, Fourier series, orthogonal basis and, to a certain degree, wavelet, are popular methods (Akcay & Heuberger, 2001; Bai, 2008; Bai, Tempo, & Liu, 2007; Chen, Billings, & Luo, 1989; Harris, Hong, & Gan, 2002; Li, Peng, & Irwin, 2005; Ninness, Hjalmarsson, & Gustafsson, 1999; Van den Hof, Heuberger, & Bokor, 1995; Zhu & Billings, 1996) often used in the identification community. The performance of these methods requires a priori information of the unknown system and the choice of the basis, but finding a suitable basis function is always a serious concern (Bai, 2008). Kernel and local polynomial methods are other attractive ways for nonlinear identification with a minimal knowledge of the unknown nonlinear system (Bai et al., 2007; Fan & Gijbels, 1996; Nadaraya, 1989). Under some mild assumptions, various asymptotic convergence results can be established. However, these methods are non-parametric and do not exploit the existing structure of the system. Thus, their convergence is usually slow and they suffer from the so-called "curse of dimensionality" problem (Tempo, Calafiore, & Dabbene, 2005) when the dimension becomes large. Another approach is a semi-basis function technique, which includes neural networks (Zhang, 1997), that casts a nonlinear system representation and identification problem directly into an optimization problem, which is usually non-convex. Subsequently, numerical search is carried out to find a solution. The idea is fairly clean and simple but the search is often trapped in a local minimum. In the last few years, a new non-parametric approach, called the direct weight optimization, has been proposed (Bai & Liu, 2007; Roll, Nazin, & Ljung, 2005). The goal is to find an estimator that is optimal for a given finite number of data observations. This approach could be important because in practice, the number of data points is always finite and is not very large in many applications. Volterra series (Rugh, 1981) is another method which is quite general to represent most smooth nonlinear systems. On the other hand, because of its generality, representation and identification of an all-purpose Volterra series is almost impossible, so that in practice the Volterra series approach is often limited to second order systems.

Recently, the idea of additive nonlinear systems has been introduced in the identification community (Bai, 2005; Bai & Chan, 2008). This idea is well known in statistics and it is probably the most used nonlinear system structure in the literature. In this paper, we follow and extend the additive nonlinear system

☆ The work was supported in part by NSF ECS-0555394. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Martin Enqvist under the direction of Editor Torsten Söderström.

* Corresponding author. Tel.: +1 3193355949; fax: +1 3193356028.
*E-mail addresses:* er-wei-bai@uiowa.edu (E.-W. Bai), tempo@polito.it (R. Tempo).

0005-1098/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2010.06.031

technique to include interactive terms. The important difference between Bai (2005), Bai and Chan (2008) and the current paper is that no interaction among variables are allowed in Bai (2005) and Bai and Chan (2008) focuses on a first order system. In particular, the current paper is a continuation of Bai and Deistler (in press) and shows that such a structure is efficient for a class of nonlinear systems having short term memory and low degree of interaction between variables. Corresponding structural estimation methods are developed for such systems based either on simple (but ad hoc) visual inspection, or on sophisticated statistical tests. Furthermore, specific identification methods are also developed for such systems. Finally, the paper demonstrates advantages of such a representation compared to existing methods, both theoretically and numerically, for nonlinear systems that have short term memory and low degree of interaction between variables.

## 2. System description and structural estimation

The systems that we consider in this paper have short term memory, say Finite Impulse Response (FIR) with a time lag $n \leq 10$, and low degree of interaction, e.g. with up to two or three factor terms. In particular, a detailed derivation and analysis of the system with two factor terms is provided. Extensions to higher order terms is tedious but straightforward. Note that though this analysis is restricted to FIR nonlinear systems, many IIR systems can be approximated well by some FIR systems with a modest time lag $n \leq 10$. This is certainly the case when a Volterra series representation is used.

A FIR nonlinear non-parametric system with up to two factor terms can be written as

$$\begin{aligned} y[t] &= f(u[t-1], u[t-2], \ldots, u[t-n]) + v[t] \\ &= \bar{c} + \sum_{i=1}^{n} \bar{f}_i(u[t-i]) \\ &\quad + \sum_{1\leq i<j\leq n} \bar{f}_{ij}(u[t-i], u[t-j]) + v[t] \end{aligned} \tag{2.1}$$

where $y[t]$ and $u[t]$ are output and input sequences respectively. The noise $v[t]$ is a sequence of iid (independent identically distributed) random variables with zero mean and bounded variance $\sigma^2$. The exact time lag is unknown and only an upper bound on the time lag $n$ is available. The structure of the function $f$ as well as $\bar{f}_i$'s and $\bar{f}_{ij}$'s, referred to as 1 and 2-factor terms respectively, are unknown.

It is clear that even for short term memory, say $n \leq 10$, there are possibly quite a few terms $M = 1 + n(n+1)/2$ in the system. It is a fact that for many applications, only few terms are active and the others are negligible. Estimating the terms that are negligible amounts to the structural estimation of the system.

There are two ways to estimate the structure of the system. The first one is full scale system identification. The idea is to identify the system including each $\bar{f}_i$ and $\bar{f}_{ij}$ and then enumerate all possible models for different combinations of $\bar{f}_i$ and $\bar{f}_{ij}$ as well as $n$. Some performance measures are calculated and the model that achieves the best performance is chosen. Then, the corresponding $n$ is the estimate of time lag and the surviving terms of $\bar{f}_i$ and $\bar{f}_{ij}$ are retained in the system. All other $\bar{f}_i$'s and $\bar{f}_{ij}$'s are considered to be negligible. The method does not distinguish between model structural estimation and full scale system identification. Note that the system is non-parametric and nonlinear. Hence, identification is usually computationally expensive and the optimization algorithm could be stuck in a local minimum. It is certainly advantageous if the structure of the system can be estimated before a full scale system identification is performed. To this end, we propose two different methods.

### 2.1. Visual inspection method

Recall that in structural estimation we are interested not in full scale system identification, but rather in finding a simple and reliable way to estimate the structure, in particular to determine the terms $\bar{f}_i$ and $\bar{f}_{ij}$ which contribute significantly. In this section, we assume that the input is at our disposal (which admittedly may be restrictive in some applications). Under such an assumption, the first problem is to find an input sequence that is simple and has the ability to separate the contributions of $\bar{f}_i$ and $\bar{f}_{ij}$,

$$U_{2^3} = \begin{pmatrix} u(1) & u(0) & u(-1) \\ u(2) & u(1) & u(0) \\ u(3) & u(2) & u(1) \\ u(4) & u(3) & u(2) \\ u(5) & u(4) & u(3) \\ u(6) & u(5) & u(4) \\ u(7) & u(6) & u(5) \\ u(8) & u(7) & u(6) \\ u(9) & u(8) & u(7) \end{pmatrix} = \begin{pmatrix} a_1 & a_1 & a_1 \\ a_2 & a_1 & a_1 \\ a_2 & a_2 & a_1 \\ a_2 & a_2 & a_2 \\ a_1 & a_2 & a_2 \\ a_2 & a_1 & a_2 \\ a_1 & a_2 & a_1 \\ a_1 & a_1 & a_2 \\ a_1 & a_1 & a_1 \end{pmatrix}. \tag{2.2}$$

To this end, let $l$ be a prime number that indicates the number of levels of input, i.e., $u[t] = \{a_1, a_2, \ldots, a_l\}$, usually $|a_i| \neq |a_j|$ to avoid ambiguity for quadratic nonlinearities. To excite the system to the maximum extent, the input sequence should contain all possible combinations of $n$-tuple $(a_{i_1}, a_{i_2}, \ldots, a_{i_n})$, $a_{i_j} = a_1, \ldots, a_l$. The minimum length of such a generating sequence is $n + l^n - 1$. The Galois sequence is such a sequence which has been investigated in Godfrey (1993); Makila (1991) for worst-case identification. Galois sequence has many desirable properties. It is a periodic pseudo-random sequence with period $l^n$ (Makila, 1991) and can be easily generated (Godfrey, 1993). More importantly, within one period, it produces each $n$-tuple combination exactly once (Makila, 1991). Note that the Galois sequence defined here is slightly different from the traditional one (Godfrey, 1993) as we need all the $n$-tuples to be included. This small difference can be easily taken care of and in fact this definition is exactly the same as in Makila (1991). An example of $G(l^n)$ for $n = 3$ and $l = 2$ is given in (2.2). To average out the effect of noise, the input sequence is repeated $L$ times, i.e.,

$$U_{L^n} = \left.\begin{pmatrix} U_{l^n} \\ U_{l^n} \\ \vdots \\ U_{l^n} \end{pmatrix}\right\} L \text{ times}. \tag{2.3}$$

Before performing structural estimation, it is interesting to observe that the representation (2.1) of the system is actually not unique. For instance, $\bar{f}_1 \to \bar{f}_1 + c$ and $\bar{f}_2 \to \bar{f}_2 - c$ for any constant $c$ would not change the input–output relationship which implies that the structure of the system, as represented in (2.1), is not identifiable. To overcome this problem, we normalize the system to make the averages of $\bar{f}_i$ and $\bar{f}_{ij}$ with respect to the input equal to zero. Let

$$g_{j,ij}(u[t-j]) = \frac{1}{l}\sum_{m=1}^{l} \bar{f}_{ij}(a_m, u[t-j]),$$

$$g_{i,ij}(u[t-i]) = \frac{1}{l}\sum_{m=1}^{l} \bar{f}_{ij}(u[t-i], a_m)$$

be the partial average of $\bar{f}_{ij}$ with respect to the 1st and 2nd variables respectively and

$$\check{c}_{ij} = \frac{1}{l^2}\sum_{m_1=1}^{l}\sum_{m_2=1}^{l} \bar{f}_{ij}(a_{m_1}, a_{m_2})$$

be the total average. Define,

$$\begin{aligned} \check{f}_{ij}(u[t-i], u[t-j]) &= \bar{f}_{ij}(u[t-i], u[t-j]) - g_{j,ij}(u[t-j]) \\ &\quad - g_{i,ij}(u[t-i]) + \check{c}_{ij}. \end{aligned} \tag{2.4}$$

Obviously, the average of this new function is zero,

$$\sum_{m=1}^{l} \check{f}_{ij}(a_m, u[t-j]) = \sum_{m=1}^{l} \check{f}_{ij}(u[t-i], a_m) = 0. \tag{2.5}$$

To make the average of $\bar{f}_i$ equal to zero, let, for each $1 \le i \le n$,

$$\begin{aligned}
\check{f}_1(u[t-1]) &= \bar{f}_1(u[t-1]) + \sum_{i=2}^{n} g_{1,1i}(u[t-1]) \\
&\quad - \underbrace{\frac{1}{l}\sum_{m=1}^{l}\left[\bar{f}_1(a_m) + \sum_{i=2}^{n} g_{1,1i}(a_m)\right]}_{\check{c}_1}, \\
\check{f}_{n-1}(u[t-n+1]) &= \bar{f}_{n-1}(u[t-n+1]) \\
&\quad + \sum_{i=1}^{n-2} g_{(n-1),i(n-1)}(u[t-n+1]) + g_{(n-1),(n-1)n}(u[t-n+1]) \\
&\quad - \underbrace{\frac{1}{l}\sum_{m=1}^{l}\left[\bar{f}_{n-1}(a_m) + \sum_{i=1}^{n-2} g_{(n-1),i(n-1)}(a_m) + g_{(n-1),(n-1)n}(a_m)\right]}_{\check{c}_{n-1}}, \\
\check{f}_n(u[t-n]) &= \bar{f}_n(u[t-n]) + \sum_{i=1}^{n-1} g_{n,in}(u[t-n]) \\
&\quad - \underbrace{\frac{1}{l}\sum_{m=1}^{l}\left[\bar{f}_n(a_m) + \sum_{i=1}^{n-1} g_{n,in}(a_m)\right]}_{\check{c}_n}.
\end{aligned} \tag{2.6}$$

Since,

$$\sum_{m=1}^{l} \check{f}_i(a_m) = 0, \quad \forall i \tag{2.7}$$

by taking $\check{c} = \bar{c} - \sum_{1\le i<j\le n} \check{c}_{ij} + \sum_{i=1}^{n} \check{c}_i$, it follows that the system (2.1) can be rewritten as

$$\begin{aligned}
y[t] &= \check{c} + \sum_{i=1}^{n} \check{f}_i(u[t-i]) + \sum_{1\le i<j\le n} \check{f}_{ij}(u[t-i], u[t-j]) + v[t], \\
&\quad t = 1,\ 2, \ldots, Ll^n.
\end{aligned} \tag{2.8}$$

This makes the representation unique. For each $1 \le i < j \le n$, $m_i, m_j = 1, \ldots, l$ and $k = 1, 2, \ldots, L$, define the partial averages of the output,

$$\begin{aligned}
Z^{ij}_{m_i m_j k} &= \frac{1}{l^{n-2}} \sum_{\substack{t=1 \\ u[t-i]=a_{m_i},\ u[t-j]=a_{m_j}}}^{l^n} y[(k-1)l^n + t] \\
Z^{ij}_{m_i m_j \cdot} &= \frac{1}{L}\sum_{k=1}^{L} Z^{ij}_{m_i m_j k} \\
Z^{ij}_{m_i \cdot \cdot} &= \frac{1}{l}\sum_{m_j=1}^{l} Z^{ij}_{m_i m_j \cdot} \\
Z^{ij}_{\cdot m_j \cdot} &= \frac{1}{l}\sum_{m_i=1}^{l} Z^{ij}_{m_i m_j \cdot} \\
Z^{ij}_{\cdots} &= \frac{1}{l}\sum_{m_i=1}^{l} Z^{ij}_{m_i \cdot \cdot} = \frac{1}{l}\sum_{m_j=1}^{l} Z^{ij}_{\cdot m_j \cdot}.
\end{aligned} \tag{2.9}$$

The subscript "dot" indicates that average has been taken with respect to this variable, e.g., $Z^{ij}_{m_i m_j \cdot}$ is the average of $Z^{ij}_{m_i m_j k}$ with respect to the last variable $k$.

To provide a physical interpretation of the above variables, let us focus on the system (2.8) with $n = 3, l = 2$ and the Galois sequence $GF(2^3)$ as in (2.2) and (2.3). Within one period, it is clear that for any fixed column of $U_{2^3}$, half of the entries have values at $a_1$ and the other half are at $a_2$. Further, it is straightforward using (2.5) and (2.7) to show that for $i = 1$ and $j = 2$,

$$\begin{aligned}
Z^{12}_{11k} &= \check{c} + \check{f}_1(a_1) + \check{f}_2(a_1) + \check{f}_{12}(a_1, a_1) \\
&\quad + (v[(k-1)2^3 + 1] + v[(k-1)2^3 + 8])/2, \\
Z^{12}_{12k} &= \check{c} + \check{f}_1(a_1) + \check{f}_2(a_2) + \check{f}_{12}(a_1, a_2) \\
&\quad + (v[(k-1)2^3 + 5] + v[(k-1)2^3 + 7])/2, \\
Z^{12}_{21k} &= \check{c} + \check{f}_1(a_2) + \check{f}_2(a_1) + \check{f}_{12}(a_2, a_1) \\
&\quad + (v[(k-1)2^3 + 2] + v[(k-1)2^3 + 6])/2, \\
Z^{12}_{22k} &= \check{c} + \check{f}_1(a_2) + \check{f}_2(a_2) + \check{f}_{12}(a_2, a_2) \\
&\quad + (v[(k-1)2^3 + 3] + v[(k-1)2^3 + 4])/2.
\end{aligned}$$

Moreover,

$$\begin{aligned}
Z^{12}_{11\cdot} &= \check{c} + \check{f}_1(a_1) + \check{f}_2(a_1) + \check{f}_{12}(a_1, a_1) \\
&\quad + \frac{1}{L}\sum_{k=1}^{L}(v[(k-1)2^3 + 1] + v[(k-1)2^3 + 8])/2, \\
Z^{12}_{12\cdot} &= \check{c} + \check{f}_1(a_1) + \check{f}_2(a_2) + \check{f}_{12}(a_1, a_2) \\
&\quad + \frac{1}{L}\sum_{k=1}^{L}(v[(k-1)2^3 + 2] + v[(k-1)2^3 + 7])/2, \\
Z^{12}_{1\cdot\cdot} &= \check{c} + \check{f}_1(a_1) + \frac{1}{2L}\sum_{k=1}^{L}\{(v[(k-1)2^3 + 2] \\
&\quad + v[(k-1)2^3 + 7])/2 + (v[(k-1)2^3 + 1] \\
&\quad + v[(k-1)2^3 + 8])/2\}, \\
Z^{12}_{\cdots} &= \check{c} + \frac{1}{4L}\sum_{t=1}^{L2^3} v[t].
\end{aligned}$$

Clearly, an estimate $\check{c}$ is obtained by $Z^{12}_{\cdots}$ and an estimate $\check{f}_1(a_1)$ is obtained by $Z^{12}_{1\cdot\cdot} - Z^{12}_{\cdots}$. The results can be trivially but cumbersomely extended to the system (2.8) with any $n \ge 2, l \ge 2$ and $i, j$ as summarized in the following theorem.

**Theorem 2.1.** *Consider the system (2.8) for any $n \ge 2, l \ge 2$ with the Galois input as in (2.3) and the variables defined in (2.9). Then, for any $1 \le i < j \le n$ and $m_i, m_j = 1, \ldots, l$, we have*

$$Z^{ij}_{m_i m_j k} = \check{c} + \check{f}_i(a_{m_i}) + \check{f}_j(a_{m_j}) + \check{f}_{ij}(a_{m_i}, a_{m_j}) + \epsilon^{ij}_{m_i m_j k}$$

*where $\epsilon^{ij}_{m_i m_j k}$'s are iid with zero mean and variance $\sigma^2/l^{n-2}$ and*

$$\begin{aligned}
Z^{ij}_{m_i m_j \cdot} &= \check{c} + \check{f}_i(a_{m_i}) + \check{f}_j(a_{m_j}) + \check{f}_{ij}(a_{m_i}, a_{m_j}) + \frac{1}{L}\sum_{k=1}^{L}\epsilon^{ij}_{m_i m_j k}, \\
Z^{ij}_{m_i \cdot \cdot} &= \check{c} + \check{f}_i(a_{m_i}) + \frac{1}{lL}\sum_{m_j=1}^{l}\sum_{k=1}^{L}\epsilon^{ij}_{m_i m_j k}, \\
Z^{ij}_{\cdot m_j \cdot} &= \check{c} + \check{f}_j(a_{m_j}) + \frac{1}{lL}\sum_{m_i=1}^{l}\sum_{k=1}^{L}\epsilon^{ij}_{m_i m_j k}, \\
Z^{ij}_{\cdots} &= \check{c} + \frac{1}{ll^{n-2}L}\sum_{t=1}^{Ll^n} v[t].
\end{aligned}$$

Therefore, for a large $L$, very good estimates of $\check{c}$, $\check{f}_i$, and $\check{f}_{ij}$ are available from $Z^{ij}_{m_i m_j \cdot}$, $Z^{ij}_{m_i \cdot \cdot}$, $Z^{ij}_{\cdot m_j \cdot}$ and $Z^{ij}_{\cdots}$ that are computable from the input–output measurements. The implication of the above

result is that the graph of $\check{f}_i(a_{m_i})$ ($\check{f}_j(a_{m_j})$) versus $a_{m_i}$ ($a_{m_j}$) is obtained by the graph of its estimate

$$\tilde{f}_i(a_{m_i}) = Z^{ij}_{m_i\cdot\cdot} - Z^{ij}_{\cdots} \text{ vs } a_{m_i} \quad \text{or}$$
$$\tilde{f}_j(a_{m_j}) = Z^{ij}_{\cdot m_j\cdot} - Z^{ij}_{\cdots} \text{ vs } a_{m_j}$$

and the graph of $\check{f}_{ij}(a_{m_i}, a_{m_j})$ versus $(a_{m_i}, a_{m_j})$ is obtained by $\tilde{f}_{ij}(a_{m_i}, a_{m_j}) = (Z^{ij}_{m_im_j\cdot} - Z^{ij}_{m_i\cdot\cdot} - Z^{ij}_{\cdot m_j\cdot} + Z^{ij}_{\cdots})$ and

$\tilde{f}_{ij}(a_{m_i}, a_{m_j})$ vs $(a_{m_i}, a_{m_j})$.

Accordingly, the contribution of $\check{f}_i(a_{m_i})$ and $\check{f}_{ij}(a_{m_i}, a_{m_j})$ can be visually inspected by the graphs of $\tilde{f}_i(a_{m_i})$ and $\tilde{f}_{ij}(a_{m_i}, a_{m_j})$. We make two comments here.

- Structural estimation is similar to model validation in identification. One can never validate a model unless all possible inputs have been applied. This is clearly impossible in practice. In structural estimation, one can only say that the contribution of $\check{f}_i(a_{m_i})$ or $\check{f}_{ij}(a_{m_i}, a_{m_j})$ is negligible with respect to the applied input. Therefore, the values $a_1, \ldots, a_l$ are important and have to be chosen judiciously.
- In general, increasing the level $l$ excites the system at more points and this is quite useful for nonlinear system identification. However, there is a balance between the number of levels $l$ and the complexity of the implementation. For $l = 2$ or any binary input, the minimum length of the sequence to cover all possible $n$-tuple combinations is $2^n$ and for an $l$ level input, the minimum length becomes $l^n$. Thus, the complexity increases quickly as $l$ gets larger.
- In general, a visual inspection works only for 2-factor terms.

### 2.2. Analysis of variance (ANOVA)

The visual inspection approach discussed above is intuitive, efficient but ad hoc. If an estimate $\tilde{f}_i$ is non-zero but small, it is hard to determine if the term should be retained or discarded because of noise. To make the idea mathematically rigorous, in this section, we develop a statistical hypothesis test based on the well-known analysis of variance (ANOVA) and $F$ distribution tests. To this end we make an assumption.

**Assumption 2.1.** The noise $v[\cdot]$ is iid Gaussian with zero mean and variance $\sigma^2$.

The Gaussian assumption is needed for the mathematical derivation. However, it has been well documented in the literature Lind and Ljung (2005) that ANOVA is quite robust against violation of the Gaussian assumption. Consider the system (2.8), the input (2.3) and the variables (2.9). Let, for each $1 \le i < j \le n$,

$$\begin{aligned}
SS^{ij}_T &= \sum_{m_i=1}^{l}\sum_{m_j=1}^{l}\sum_{k=1}^{L}(Z^{ij}_{m_im_jk} - Z^{ij}_{\cdots})^2\\
SS^{ij}_{m_i\cdot} &= \sum_{m_i=1}^{l} lL(Z^{ij}_{m_i\cdot\cdot} - Z^{ij}_{\cdots})^2\\
SS^{ij}_{\cdot m_j} &= \sum_{m_j=1}^{l} lL(Z^{ij}_{\cdot m_j\cdot} - Z^{ij}_{\cdots})^2\\
SS^{ij}_{\cdot\cdot} &= \sum_{m_i=1}^{l}\sum_{m_j=1}^{l} L(Z^{ij}_{m_im_j\cdot} - Z^{ij}_{\cdot m_j\cdot} - Z^{ij}_{m_i\cdot\cdot} + Z^{ij}_{\cdots})^2\\
SS^{ij}_E &= \sum_{m_i=1}^{l}\sum_{m_j=1}^{l}\sum_{k=1}^{L}(Z^{ij}_{m_im_jk} - Z^{ij}_{m_im_j\cdot})^2.
\end{aligned} \tag{2.10}$$

The following theorem can be shown by some algebraic manipulations and the Cochran Theorem (Papoulis & Pillai, 2002).

**Theorem 2.2.** *Consider the variables defined in (2.10). Then,*

- $SS^{ij}_T = SS^{ij}_{m_i\cdot} + SS^{ij}_{\cdot m_j} + SS^{ij}_{\cdot\cdot} + SS^{ij}_E$.
- $SS^{ij}_{m_i\cdot}$, $SS^{ij}_{\cdot m_j}$, $SS^{ij}_{\cdot\cdot}$ *and* $SS^{ij}_E$ *are statistically independent.*
- $\frac{l^{n-2}}{\sigma^2}SS^{ij}_E \sim \chi^2(l^2(L-1))$ *is* $\chi^2$ *distributed with* $l^2(L-1)$ *degrees of freedom.*
- *If* $\check{f}_{ij}(a_{m_i}, a_{m_j}) = 0$ *for all* $m_i, m_j = 1, \ldots, l$, *then*
$$\frac{l^{n-2}}{\sigma^2}SS^{ij}_{\cdot\cdot} \sim \chi^2((l-1)^2).$$
- *If* $\check{f}_i(a_{m_i}) = 0$ *for all* $m_i = 1, \ldots, l$, *then*
$$\frac{l^{n-2}}{\sigma^2}SS^{ij}_{m_i\cdot} \sim \chi^2(l-1).$$
- *If* $\check{f}_j(a_{m_j}) = 0$ *for all* $m_j = 1, \ldots, l$, *then*
$$\frac{l^{n-2}}{\sigma^2}SS^{ij}_{\cdot m_j} \sim \chi^2(l-1).$$

This theorem sets the foundation for the test of three null hypotheses,

$$\begin{aligned}
&H_{0ij}\colon \check{f}_{ij}(a_{m_i}, a_{m_j}) = 0, \quad \forall a_{m_i}, a_{m_j} = 1, \ldots, l,\\
&H_{0i\cdot}\colon \check{f}_i(a_{m_i}) = 0, \quad \forall a_{m_i} = 1, \ldots, l,\\
&H_{0\cdot j}\colon \check{f}_j(a_{m_j}) = 0, \quad \forall a_{m_j} = 1, \ldots, l,
\end{aligned}$$

by the $F$-test because if $H_{0ij}$ is true then

$$T^{ij} = \frac{SS^{ij}_{\cdot\cdot}/(l-1)^2}{SS^{ij}_E/(l^2(L-1))} \sim F((l-1)^2, l^2(L-1)),$$
for all $1 \le i < j \le n$,

is $F$-distributed with $(l-1)^2$ and $l^2(L-1)$ degrees of freedom. Similarly, if $H_{0i\cdot}$ is true,

$$T^1 = \frac{SS^{12}_{m_i\cdot}/(l-1)}{SS^{12}_E/(l^2(L-1))} \sim F(l-1, l^2(L-1))$$

and if $H_{0\cdot j}$ is true, $\forall j = 2, \ldots, n$,

$$T^j = \frac{SS^{1j}_{\cdot m_j}/(l-1)}{SS^{1j}_E/(l^2(L-1))} \sim F(l-1, l^2(L-1)).$$

The null hypothesis $H_{0ij}$ is rejected if $T^{ij} > F_\alpha((l-1)^2, l^2(L-1))$ where $\alpha$ denotes the level of significance, usually in the range 0.01–0.1. The tests for $H_{0i\cdot}$ and $H_{0\cdot j}$ are similar. The results from the hypothesis tests are used to determine which $f_i$ or $f_{ij}$ should be retained with a certain confidence in probability.

## 3. Full scale identification

For full scale system identification, using the Galois sequence is not appropriate because the Galois sequence only excites the system at a finite points. We assume in this section that the input $u[t]$ is an iid random sequence in a (unknown) open interval $I \in R$ with a (unknown) probability density function $\psi(\cdot)$. Then, the results of Bai (2008) can be used. Similar to the structural estimation case, the system (2.1) needs to be normalized for

identification purposes. Let $\mathbf{E}$ be the expectation operator. Define the partial averages,

$$c_{ij} = \mathbf{E}\{\bar{f}_{ij}(u[t-i], u[t-j])\},$$

$$c_1 = \mathbf{E}\left\{\bar{f}_1(u[t-1]) + \sum_{j=2}^{n} \mathbf{E}(\bar{f}_{1j}(u[t-1], u[t-j]) \mid u[t-1] = x_1)\right\},$$

$$c_i^1 = \mathbf{E}\left\{\bar{f}_i(u[t-i]) + \sum_{k=i+1}^{n} \mathbf{E}(\bar{f}_{ik}(u[t-i], u[t-k]) \mid u[t-i] = x_i)\right\},$$

$$c_i^2 = \sum_{k=1}^{i-1} \mathbf{E}(\bar{f}_{ki}(u[t-k], u[t-i]) \mid u[t-i] = x_i),$$

$$c_n = \mathbf{E}\left\{\bar{f}_n(u[k-n]) + \sum_{k=1}^{n-1} \mathbf{E}(\bar{f}_{kn}(u[t-k], u[t-n]) \mid u[t-n] = x_n)\right\}.$$

Now, for every $x_i$ and $x_j \in I$, define

$$\begin{aligned} f_{ij}(x_i, x_j) &= \bar{f}_{ij}(x_i, x_j) - \mathbf{E}(\bar{f}_{ij}(u[t-i], u[t-j]) \mid u[t-j] = x_j) \\ &\quad - \mathbf{E}(\bar{f}_{ij}(u[t-i], u[t-j]) \mid u[t-i] = x_i) \\ &\quad + c_{ij}, \quad 1 \le i < j \le n, \end{aligned}$$

$$\begin{aligned} f_1(x_1) &= \bar{f}_1(x_1) \\ &\quad + \sum_{j=2}^{n} \mathbf{E}(\bar{f}_{1j}(u[t-1], u[t-j]) \mid u[t-1] = x_1) - c_1, \end{aligned}$$

$$\begin{aligned} f_i(x_i) &= \bar{f}_i(x_i) + \sum_{k=i+1}^{n} \mathbf{E}(\bar{f}_{ik}(u[t-i], u[t-k]) \mid u[t-i] = x_i) \\ &\quad + \sum_{k=1}^{i-1} \mathbf{E}(\bar{f}_{ki}(u[t-k], u[t-i]) \mid u[t-i] = x_i) \\ &\quad - c_i^1 - c_i^2, \quad i = 2, 3, \ldots, n-1, \end{aligned}$$

$$f_n(x_n) = \bar{f}_n(x_n) + \sum_{i=1}^{n-1} \mathbf{E}(\bar{f}_{in}(u[t-i], u[t-n]) \mid u[t-n] = x_n) - c_n. \tag{3.11}$$

Next, with $c = \bar{c} - \sum_{1 \le i < j \le n} c_{ij} + \sum_{i=1}^{n} c_i$, $c_i = c_i^1 + c_i^2$ the system (2.1) can be written as

$$y[t] = c + \sum_{i=1}^{n} f_i(u[t-i]) + \sum_{1 \le i < j \le n} f_{ij}(u[t-i], u[t-j]) + v[t], \quad t = 1, 2, \ldots, N \tag{3.12}$$

with

$$\begin{aligned} \mathbf{E} f_i(u[t-i]) &= \mathbf{E}(f_{ij}(u[t-i], u[t-j]) \mid u[t-i] = x_i) \\ &= \mathbf{E}(f_{ij}(u[t-i], u[t-j]) \mid u[t-j] = x_j) = 0. \end{aligned}$$

The problem is how to identify $f_i$ and $f_{ij}$. Observe that these variables are conditional expectations and thus can be calculated by empirical data easily, for instance using the kernel estimation method (Bai, 2008). To this end, we define the kernel functions. A continuous, bounded and radially symmetric function $K(\cdot)$ is said to be a kernel function if

$$K(z) = \begin{cases} > 0, & z \in [-1, 1] \\ 0, & z \notin [-1, 1] \end{cases} \quad \text{and} \quad \int_{-1}^{1} K(z)\mathrm{d}z = 1. \tag{3.13}$$

Now, the estimates of $c$, $f_i$ and $f_{ij}$ can be defined for each $x_i, x_j \in I$ in which the input $u[\cdot]$ lies,

$$\hat{c} = \frac{1}{N} \sum_{t=1}^{N} y[t] \tag{3.14}$$

$$\hat{f}_i(x_i) = \frac{\sum_{t=1}^{N} K\left(\frac{x_i - u[t-i]}{\delta}\right) y[t]}{\sum_{t=1}^{N} K\left(\frac{x_i - u[t-i]}{\delta}\right)} - \hat{c}, \quad i = 1, \ldots, n$$

$$\begin{aligned} \hat{f}_{ij}(x_i, x_j) &= \frac{\sum_{t=1}^{N} K\left(\frac{\|(x_i, x_j) - (u[t-i], u[t-j])\|}{\delta}\right) y[t]}{\sum_{t=1}^{N} K\left(\frac{\|(x_i, x_j) - (u[t-i], u[t-j])\|}{\delta}\right)} \\ &\quad - \hat{f}_i(x_i) - \hat{f}_j(x_j) - \hat{c}, \quad 1 \le i < j \le n \end{aligned}$$

where $\delta > 0$ is the bandwidth. The following result, which is a standard exercise, follows from Bai (2008).

**Theorem 3.1.** *Consider the system (3.12) with differentiable $f_i$ and $f_{ij}$, and any kernel function defined above. Then, for any $x_i, x_j \in I$, provided that the input density function is positive at $x_i, x_j$, i.e., $\psi(x_i), \psi(x_j) > 0$ and $\delta \to 0$, $\delta^2 N \to \infty$ as $N \to \infty$, we have*

$$\hat{c} \to c$$
$$\hat{f}_i(x_i) \to f_i(x_i)$$
$$\hat{f}_{ij}(x_i, x_j) \to f_{ij}(x_i, x_j)$$

*in probability as $N \to \infty$.*

## 4. Comparisons with existing methods

A new representation for a class of nonlinear non-parametric system with short term memory and low degree of interaction has been proposed in (2.8). Further, structural estimation and full scale identification have been discussed in the previous section. Naturally, two questions arise. The first one is what are the advantages of the representation (2.8) as compared to some existing methods, in particular the fixed basis approach and the Volterra series? Secondly, even if one accepts the representation (2.8), why use the structural estimation and system identification techniques discussed in the previous section as compared to the traditional approach of identifying $f(u[t-1], \ldots, u[t-n])$ directly? We address these two issues in this section.

### 4.1. Relation with the Volterra series

If the system (2.8) is smooth with an upper bound $n$ on the time lag, its Volterra series is given by

$$\begin{aligned} y[t] &= h_0 + \sum_{l=1}^{n} \sum_{i_1=0}^{\infty} \sum_{i_2=i_1}^{\infty} \cdots \sum_{i_l=i_{l-1}}^{\infty} h_l(i_1, \ldots, i_l) \\ &\quad \times u[t-i_1]u[t-i_2]\ldots u[t-i_l] + v[t]. \end{aligned}$$

Two of the major advantages of the Volterra series are (1) it is in a closed form and (2) it is parametric. In other words, any smooth nonlinear non-parametric system can always be written in the above form. Further, identification becomes a linear estimation of the coefficients $h_l$'s. However, the Volterra series also has some disadvantages. In this paper, we are mainly interested in verifying if the Volterra series is a good candidate for the system of short term memory and low degree of interaction as in (2.1) or (3.12). To this end, we need to understand the differences between a

system of low degree of interaction and a system of low order in the classical sense. Traditionally, a system is said to be of low order if it can be written as or at least can be well approximated by a low order multi-dimensional polynomial. For instance, a system is said to be first order if it is linear

$$y[k] = f(u[k-1], \ldots, u[k-n]) = c + \sum_{i=1}^{n} \alpha_i u[k-i]$$

or to be of second order if

$$y[k] = c + \sum_{i=1}^{n} \alpha_i u[k-i] + \sum_{1 \le j_1 \le j_2 \le n} \gamma_{j_1 j_2} u[k-j_1]u[k-j_2].$$

Clearly, in both cases, the system is of 1-factor or 2-factor terms. In general, a system of low order in the traditional sense implies low degree of interaction. The other way around is however incorrect. For example, $e^{u[k-1]}$ is an 1-factor term that is not necessarily of low order depending on the input magnitude. Also, $(u[k-1]u[k-2])^{10}$ is a 2-factor term which may not be approximated well by a second order polynomial. Therefore, nonlinear systems of low order in the traditional sense are low degree interaction systems but the reverse implication is not necessarily true. Now, we consider a Volterra series approach. A second order Volterra series is a model that contains all the first and second order kernels $u[k-i]$'s and $u[k-j_1]u[k-j_2]$'s. This model is a 2-factor interaction system. However, a 2-factor system $y[k] = e^{u[k-1]} + (u[k-1]u[k-2])^{10}$ is definitely not represented well by a low order Volterra series.

In summary, if a nonlinear system of short term memory and low degree of interaction resembles the structure of a low order multi-dimensional polynomial, the Volterra series is a good candidate. If the system is far away from a polynomial or the order of the polynomial is high, the Volterra series is not a good candidate simply because too many terms are needed to approximate the given system. In such a case, i.e., the unknown system is of low degree of interaction but not necessarily a low order polynomial, the proposed representation is a vital choice. This observation is not surprising because the Volterra series is an extension of Taylor polynomial expansion of an analytic function. The advantages of the proposed representation for systems of short memory and low degree of interaction will be further illustrated in the simulation section.

### 4.2. Basis function approach

Without structural information, a fixed basis function approach is often used in nonlinear system identification. Typical basis functions are Fourier series, polynomials and some orthogonal versions. Obviously, the success of a basis function approach relies on how much a priori information is available on the unknown structure. If the chosen basis functions resembles the structure of the unknown nonlinear system, only a few terms are needed to represent the unknown system. In this case, identification is likely to be successful. Otherwise, a fixed basis function approach requires a large number of terms which has a considerable negative effect on the identification step. The advantage of the proposed representation is that, if a nonlinear system has short term memory and low degree of interaction which fits (3.12), then no additional structural information is required. In other words, there is no need to choose any basis functions and whether a chosen basis function resembles the unknown structure is no longer an issue.

### 4.3. Traditional one shoot kernel approach

Once the representation of (2.1) or (3.12) is accepted, the second question is why to use the identification method proposed in the previous section and why not to identify the nonlinear function $f(u[k-1], \ldots, u[k-n])$ directly, which is a traditional approach. The difference is that the identification method proposed in this paper decomposes a potentially high dimensional nonlinear identification problem into a number of one or two dimensional problems. Since the method proposed in the paper is kernel based, we compare it with the one shoot kernel based identification method.

First, for the one shoot kernel estimation of $f(u[k-1], \ldots, u[k-n])$ under iid inputs, the asymptotic convergence rate (Fan & Gijbels, 1996) is $O(N^{-\frac{\alpha}{2\alpha+n}})$, where $N$ is the total number of data points and $\alpha$ depends on the choices of the kernel functions and the bandwidth. For the method proposed in the paper, because identification is one or two dimensional, the asymptotic convergence rate is $O(N^{-\frac{\alpha}{2\alpha+n}}|_{n=2}) = O(N^{-\frac{\alpha}{2\alpha+2}})$ (Fan & Gijbels, 1996). Thus, asymptotically, there is an advantage to use the proposed method.

Next, we consider the case that $N$ is large but fixed. For nonlinear system identification, the curse of dimensionality is always a concern even for a modest $n$. We use similar arguments and examples as in Bai (2008) to illustrate the situation. Let $u[\cdot]$ be uniformly distributed in $I = [-1, 1]$. Suppose one wants to estimate $f(x_1, x_2, \ldots, x_n)$ at a point $(x_1, x_2, \ldots, x_n) \in I^n$. Since any non-parametric identification scheme, including the kernel approach, is in some form of local smoother or weighted average based on the measurement data in the neighborhood of $(x_1, x_2, \ldots, x_n)$, there must be enough 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 $2^n$ and the volume of the neighborhood is $0.1^n$. This implies that the probability that a measurement data $(u[k-1], u[k-2], \ldots, u[k-n])$ is in the neighborhood of $(x_1, x_2, \ldots, x_n)$ is $(1/20)^n$ that goes to zero exponentially as $n$ gets large. For a large $N$, there are likely $N \cdot (1/20)^n$ measurements in the neighborhood. Unless $N$ is huge, there is not enough data in a neighborhood for identification purpose even for a modest $n$. For the proposed method, however, the maximum dimension is two. The curse of dimensionality is not a problem. For instance, let $n = 8$. Then, the problem becomes identification of 8 1-factor terms $f_j(u[k-j]), j = 1, 2, \ldots, 8$, and 28 2-factor terms $f_{j_1 j_2}(u[k-j_1], u[k-j_2])$. Though the number of identification steps increases, the complexity of identification is reduced drastically. Because of decoupling, the probability of an $u[k-j]$ in the neighborhood of $x_j$ for one-dimensional identification is 0.05 and the probability of $(u[k-j_1], u[k-j_2])$ in the neighborhood of $(x_{j_1}, x_{j_2})$ is 0.0025. Suppose that the total number of data points is $N = 10^5$. This implies that likely there are 5000 or 250 measurements in the neighborhood for identification of 1-factor or 2-factor terms, respectively. Recall that if the 8-dimensional $f(x_1, \ldots, x_8)$ is identified directly, the probability that a data vector is in the neighborhood of $(x_1, \ldots, x_8)$ is $(1/20)^8$. With $N = 10^5$, the probability that there is one measurement in a neighborhood is $(1/2)^8 \cdot 10^{-3} = \frac{1}{2^8 10^3}$ that makes identification nearly impossible. 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 8-dimensional problem $f$. This effectively combats the curse of dimensionality.

## 5. Numerical comparisons

In this section, we discuss two numerical examples that shed lights on the efficiency of the proposed representation and identification method in the context of existing methods.

**Table 1**
Calculated $T^i$ and $T^{ij}$ for polynomial input nonlinearity.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $T^i$ | 3986 | 4617 | 371.5 | 23.3 | 2.5 | 1 | 1.2 | 1.1 |
| $T^{ij}$ | | $j$ | | | | | | |
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $i$ | 1.0 | 56 | 49 | 1.2 | 1.1 | 0.9 | 0.6 | 1.1 |
| | 2 | | 5.8 | 1.5 | 1.0 | 0.9 | 0.7 | 1.0 |
| | 3 | | | 1.2 | 1.3 | 0.7 | 1.2 | 0.9 |
| | 4 | | | | 1.3 | 1.0 | 1.0 | 0.9 |
| | 5 | | | | | 1.0 | 0.9 | 0.8 |
| | 6 | | | | | | 0.9 | 0.7 |
| | 7 | | | | | | | 0.8 |

**Table 2**
Goodness-of-fits for the polynomial input nonlinearity.

| | Proposed method | 4th order Volterra | 2nd order fixed basis | Traditional one shoot |
|---|---|---|---|---|
| Gof | 0.9470 | 0.9563 | 0.8121 | 0.6762 |

## Example 1.

$$w[t] = u[t] - 0.3u[t]^3$$
$$x[t] = 0.3x[t-1] - 0.02x[t-2] + 0.5w[t-1] + 0.4w[t-2]$$
$$y[t] = x[t] + 0.4x[t]^2 + v[t].$$

The noise $v[t]$ is an iid zero mean and unit variance Gaussian random variable multiplied by 0.2. The actual nonlinear system is IIR and therefore there are no exact $f_i$ and $f_{ij}$. We represent the system by (3.12) assuming that the maximum time lag $n \leq 8$. Note determination of the order of an unknown nonlinear system is an interesting and open problem which is out of scope of the paper. Here we just assume that the upper bound $n = 8$ is available (admittedly it could be restrictive in some applications).

First, structural estimation is carried out by using a binary Galois sequence $GF(2^8)$ with $n = 8$, $l = 2$ and $L = 11$ and $a_1 = 1, a_2 = 0$. ANOVA was used to calculate $T^{ij}$ and $T^i$ as shown in Table 1 that are the averages of 50 Monte Carlo simulations.

For the hypothesis tests, we choose $\alpha = 0.1$. From the $F$ distribution, we have $F_{0.1}(1, 40) = 2.84$. By the $F$-tests, we have $T^1, T^2, T^3, T^4, T^{12}, T^{13}, T^{23} > 2.84$ and all other $T^i, T^{ij} < 2.84$ as can be seen in Table 1. Thus, we reject the hypotheses that $f_1, f_2, f_3, f_4, f_{12}, f_{13}$ and $f_{23}$ are negligible and assume that all other terms are zero. Secondly, these non-negligible terms are identified with iid input uniformly in $[-1.5, 1.5]$, a triangle kernel (Bai, 2008) with $\delta = 0.4$ and the total number of data points $N = 5000$. Further, their estimates are used to construct the model

$$\begin{aligned}\hat{y}[t] = \hat{c} &+ \hat{f}_1(u[t-1]) + \hat{f}_2(u[t-2]) + \hat{f}_3(u[t-3]) \\ &+ \hat{f}_4(u[t-4]) + \hat{f}_{12}(u[t-1], u[t-2]) \\ &+ \hat{f}_{13}(u[t-1], u[t-3]) + \hat{f}_{23}(u[t-2], u[t-3]).\end{aligned}$$

To validate the model, the input is generated

$$u[t] = 1.5\sin(t/10)\cos(t/20), \quad t = 1, \ldots, 160$$

as well as the corresponding actual outputs $y[t]$ and predicted outputs $\hat{y}[t]$'s. To quantify the error between $\hat{y}[t]$ and $y[t]$, a standard goodness-of-fit (gof) is adopted

$$\left(1 - \sqrt{\frac{\sum_t (y[t] - \hat{y}[t])^2}{\sum_t \left(y[t] - 1N\sum_t y[t]\right)^2}}\right) \times 100\%.$$

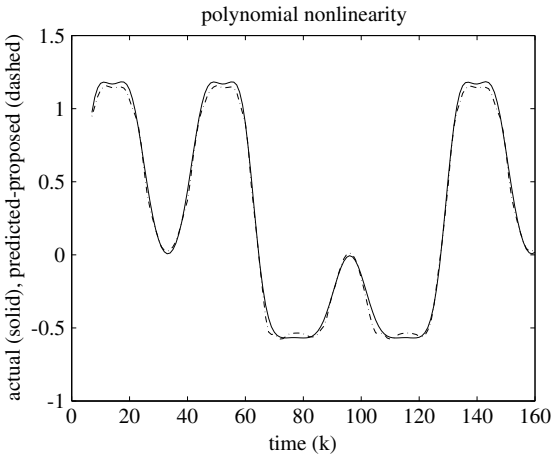


**Fig. 1.** Actual $y[t]$ and predicted $\hat{y}[t]$ by the proposed method with gof = 0.9470 (polynomial nonlinearity).

Figs. 1–4 show $y[t]$, $\hat{y}[t]$'s predicted by the proposed method, the Volterra series of 4th order, a fixed basis of polynomial up to the second order $y(t)y(k), u(t)u(k), y(t)u(k)$ and the one shoot method respectively as well as their gof's. Since the actual nonlinearity is a polynomial, the proposed method, the Volterra series and the fixed basis of polynomial all perform satisfactory, significantly better than the one shoot method as expected. An overview of the performances is given in Table 2.

## Example 2.

$$w[t] = u[t] - 0.3u[t]^3 e^{1.4u[t]}$$
$$x[t] = 0.3x[t-1] - 0.02x[t-2] + 0.5w[t-1] + 0.4w[t-2]$$
$$y[t] = x[t] + 0.4x[t]^2 + v[t].$$

The only difference between Examples 1 and 2 is that the input nonlinearity now contains an exponential term. All other simulation conditions remain the same. $T^i$ and $T^{ij}$ for Example 2 are given in Table 3 for a binary test input $GF(l^n)$ with $n = 8$, $l = 2$ and $L = 11$.

With $\alpha = 0.1$ and by the $F$-test as shown in Table 3, only the terms $f_1, f_2, f_3, f_4, f_5, f_{12}, f_{13}, f_{14}, f_{23}$ and $f_{24}$ are not negligible and thus the model is given by

$$\begin{aligned}\hat{y}[t] = \hat{c} &+ \hat{f}_1(u[t-1])\hat{f}_2(u[t-2]) + \hat{f}_3(u[t-3]) + \hat{f}_4(u[t-4]) \\ &+ \hat{f}_5(u[t-5]) + \hat{f}_{12}(u[t-1], u[t-2]) \\ &+ \hat{f}_{13}(u[t-1], u[t-3]) + \hat{f}_{14}(u[t-1], u[t-4]) \\ &+ \hat{f}_{23}(u[t-2], u[t-3]) + \hat{f}_{24}(u[t-2], u[t-4]).\end{aligned}$$

Under the same validation input, the corresponding $y[t]$ and predicted $\hat{y}[t]$ by various methods are shown in Figs. 5–8. The corresponding gof's are given in Table 4.

The results of the 2nd, 3rd, 4th, 5th and 6th order Volterra series are also shown in Table 4 and Fig. 6, exhibiting a considerable performance deterioration. This is because a low order polynomial approximation in $u[\cdot]$ like the Volterra series is inefficient to model an exponential function. This demonstrates the advantage of the proposed representation along with structural estimation and system identification for nonlinear non-parametric system of short term memory and low degree of interaction. It is interesting to note that a higher order Volterra does not necessarily imply a better identification result because variance error also increases as the order gets high. The gofs of the fixed basis function for the second and third order polynomials are 0.2299 and 0.1659 respectively. Fig. 7 demonstrates the corresponding $y[t]$ and $\hat{y}[t]$ for the fixed

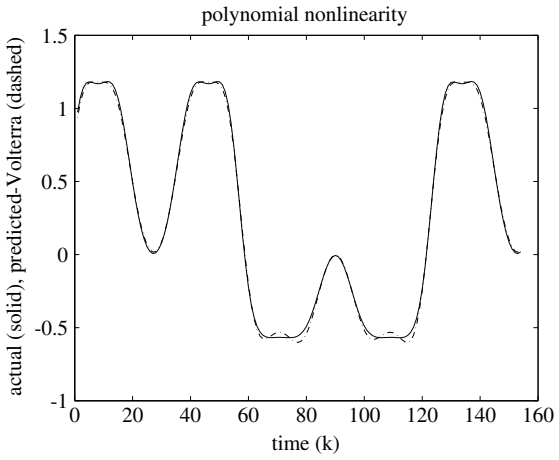


**Fig. 2.** Actual $y[t]$ and predicted $\hat{y}[t]$ by a 4th order Volterra with gof = 0.9563 (polynomial nonlinearity).

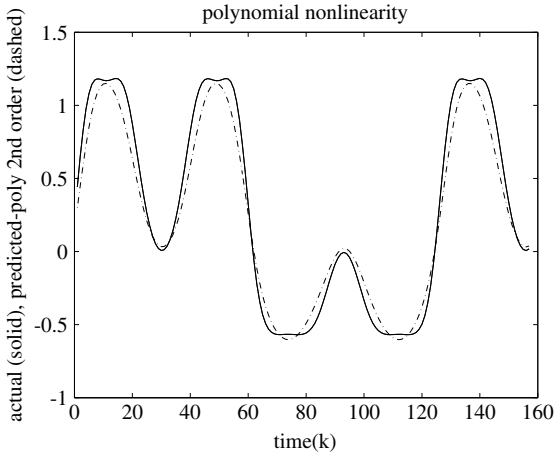


**Fig. 3.** Actual $y[t]$ and predicted $\hat{y}[t]$ by a 2nd polynomial with gof = 0.8121 (polynomial nonlinearity).

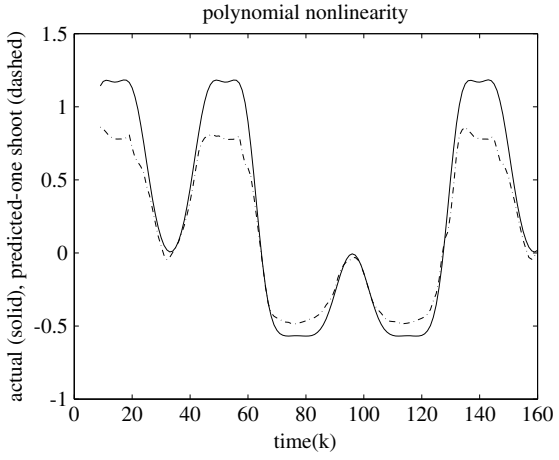


**Fig. 4.** Actual $y[t]$ and predicted $\hat{y}[t]$ by one shoot method with gof = 0.6762 (polynomial nonlinearity).

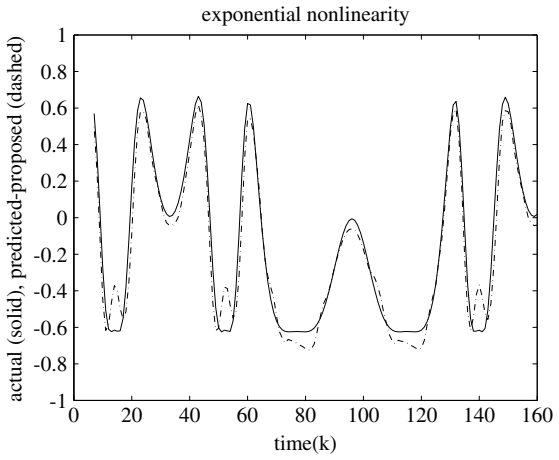


**Fig. 5.** Actual $y[t]$ and predicted $\hat{y}[t]$ by the proposed method (exponential nonlinearity).

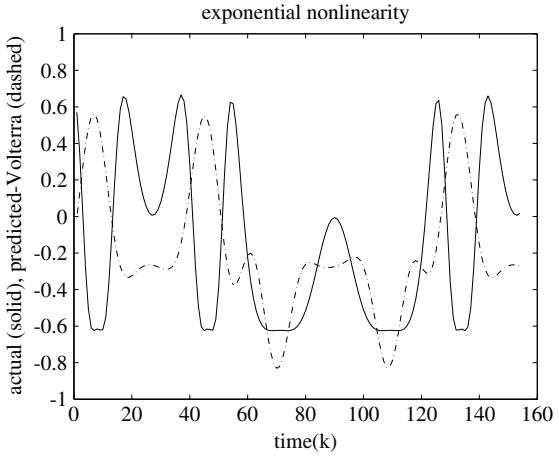


**Fig. 6.** Actual $y[t]$ and predicted $\hat{y}[t]$ by a 3rd order Volterra (exponential nonlinearity).

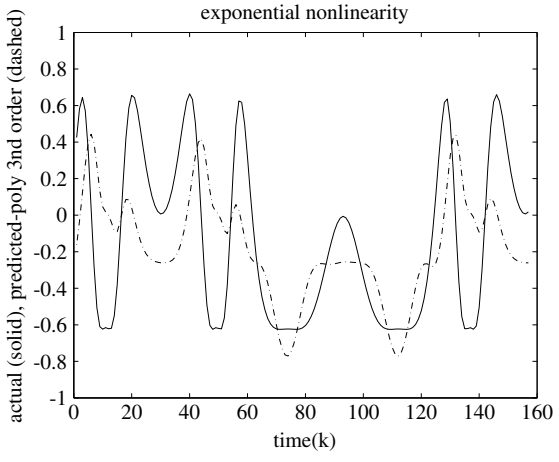


**Fig. 7.** Actual $y[t]$ and predicted $\hat{y}[t]$ by a 3rd polynomial (exponential nonlinearity).

basis function approach of 3rd order. Again, performance of a fixed basis function approach depends on if the chosen functions resemble the unknown structure or not. The result of the one shoot kernel is shown in Fig. 8 with gof = 0.1679, a poor performance. The reason is that for a higher dimension $n = 8$, the bandwidth $\delta$ has to be large or there is no data in the neighborhood that consequently increases the bias. In the simulation, bandwidth was carefully adjusted to find the best gof which is reported here. It is clear, for Example 2 which is of short term memory and low order interaction, the proposed method outperforms any other method.

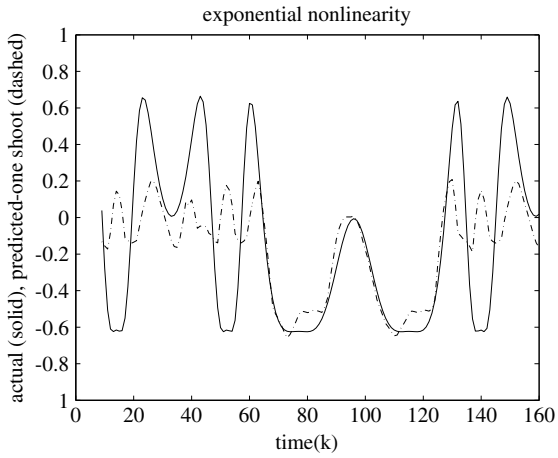


**Fig. 8.** Actual $y[t]$ and predicted $\hat{y}[t]$ by one shoot method (exponential nonlinearity).

**Table 3**
$T^i$ and $T^{ij}$ for exponential nonlinearity.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $T^i$ | 25784 | 30338 | 2336 | 123 | 8 | 1 | 1 | 1 |
| $T^{ij}$ | | $j$ | | | | | | |
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $i$ | 1 | 846 | 65 | 4 | 1 | 1 | 1 | 1 |
| | 2 | | 78 | 5 | 1 | 1 | 1 | 1 |
| | 3 | | | 1 | 1 | 1 | 1 | 1 |
| | 4 | | | | 1 | 1 | 1 | 1 |
| | 5 | | | | | 1 | 1 | 1 |
| | 6 | | | | | | 1 | 1 |
| | 7 | | | | | | | 1 |

**Table 4**
Goodness-of-fits for the exponential input nonlinearity.

| | Proposed method | 2nd order fixed basis | 3rd order fixed basis | Traditional one shoot |
|---|---|---|---|---|
| gof | 0.6855 | 0.2299 | 0.1679 | 0.2722 |

| Volterra (order) | 2nd | 3rd | 4th | 5th | 6th |
|---|---|---|---|---|---|
| gof | −0.3437 | −0.7652 | −0.6194 | −8.6657 | −7.5490 |

## 6. Concluding remarks

This paper presents a framework for a class of nonlinear systems and shows that the representation together with its structural estimation and identification are efficient for systems that have short term memory and low degree of interaction between variables. In particular, when the nonlinearity is non-polynomial or it is a high order polynomial where the Volterra series usually fails, the proposed representation is very useful. The representation can be considered as an extension of the additive system approach (Bai, 2005; Bai & Chan, 2008; Bai & Deistler, in press; Sperlich, Tjostheim, & Yang, 2002) which is itself an extension of the well-known auto-regressive moving average linear systems.

## References


Akcay, H., & Heuberger, P. (2001). A frequency domain iterative identification algorithm using general orthonormal basis functions. *Automatica*, *37*, 663–674.

Bai, E. W. (2005). Identification of additive nonlinear systems. *Automatica*, *41*, 1247–1253.

Bai, E. W. (2008). Non-Parametric nonlinear system identification: a data-driven orthogonal basis function approach. *IEEE Transactions on Automatic Control*, *53*, 2615–2626.

Bai, E. W., & Chan, K. S. (2008). Identification of additive nonlinear systems and its application in generalized Hammerstein models. *Automatica*, *44*, 430–436.

Bai, E. W., & Liu, Y. (2007). Recursive direct weight optimization in nonlinear system identification: a minimal probability approach. *IEEE Transactions on Automatic Control*, *52*, 1218–1231.

Bai, E.W., & Deistler, M. (2010). An interactive term approach to non-parametric FIR nonlinear system identification IEEE Transactions on Automatic Control (in press).

Bai, E. W., Tempo, R., & Liu, Y. (2007). Identification of nonlinear systems without prior structural information. *IEEE Transactions on Automatic Control*, *52*, 442–453.

Chen, S., Billings, S. A., & Luo, W. (1989). Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, *50*, 1873–1896.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. New York: Chapman & Hall.

Godfrey, G. (1993). *Perturbation signal for system identification*. New York: Prentice-Hall.

Harris, C. J., Hong, X., & Gan, Q. (2002). *Adaptive modeling, estimation and fusion from data: a neurofuzzy approach*. London: Springer-Verlag.

Lind, I., & Ljung, L. (2005). Regression selection with the analysis of variance method. *Automatica*, *41*, 693–700.

Li, K., Peng, J., & Irwin, G. (2005). A fast nonlinear model identification method. *IEEE Transactions on Automatic Control*, *50*, 1211–1216.

Ljung, L. (1999). *System identification: theory for the user* (2nd ed.). Upper Saddle River: Prentice-Hall.

Makila, P. M. (1991). Robust identification and Galois sequence. *International Journal of Control*, *54*, 1189–1200.

Nadaraya, E. (1989). *Nonparametric estimation of probability densities and regression curves*. Dordrecht, The Netherlands: Kluwer Academic Pub.

Ninness, B., Hjalmarsson, H., & Gustafsson, F. (1999). On the fundamental role of orthonormal bases in system identification. *IEEE Transactions on Automatic Control*, *44*, 1384–1407.

Papoulis, P., & Pillai, A. (2002). *Probability, random variables and stochastic processes* (4th ed.). Boston: McGraw Hill.

Roll, J. A., Nazin, L., & Ljung, (2005). Nonlinear system identification via direct weight optimization. *Automatica*, *41*, 475–490.

Rugh, W. (1981). *Nonlinear system theory*. London: John Hopkins University Press.

Soderstrom, T., & Stoica, P. (1989). *System identification*. New York: Prentice-Hall.

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

Tempo, R., Calafiore, G., & Dabbene, F. (2005). *Randomimized algorithms for analysis and control of uncertain*. London: Springer-Verlag.

Van den Hof, P. M. J., Heuberger, P. S. C., & Bokor, J. (1995). System identification with generalized orthonormal basis functions. *Automatica*, *31*, 1821–1834.

Zhang, Q. (1997). Using wavelet network in non-parametric estimation. *IEEE Transactions on Neural Networks*, *8*, 227–236.

Zhu, Q. M., & Billings, S. A. (1996). Fast orthogonal identification of nonlinear stochastic models and radial basis function neural networks. *International Journal of Control*, *64*, 871–886.


**Er-Wei Bai** was educated in Fudan University, Shanghai Jiaotong University, both in Shanghai, China, and the University of California at Berkeley. Dr. Bai is Professor of Electrical and Computer Engineering and Radiology at the University of Iowa where he teaches and conducts research in identification, control, signal processing and their applications in engineering and medicine.

Dr. Bai is an IEEE Fellow and a recipient of the President's Award for Teaching Excellence and the Board of Regents Award for Faculty Excellence.

**Roberto Tempo** is currently a Director of Research of Systems and Computer Engineering at the institute IEIIT, National Research Council (CNR), Politecnico di Torino, Italy. He has held visiting positions at various institutions, including Kyoto University, University of Illinois at Urbana-Champaign, German Aerospace Research Organization in Oberpfaffenhofen and Columbia University in New York.

Dr. Tempo's research activities are mainly focused on analysis and design of complex systems with uncertainty, and related control applications. He is author or co-author of more than 150 research papers published in international journals, books and conferences. He is also a co-author of the book "Randomized Algorithms for Analysis and Control of Uncertain Systems", Springer-Verlag, 2005. He 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 a Fellow of the IEEE and a Fellow of the IFAC.