



UNIVERSITY OF LEEDS

This is a repository copy of *Kernel-based local order estimation of nonlinear nonparametric systems*.

White Rose Research Online URL for this paper:  
<http://eprints.whiterose.ac.uk/143809/>

Version: Accepted Version

---

**Article:**

Zhao, W, Chen, H-F, Bai, E-W et al. (1 more author) (2015) Kernel-based local order estimation of nonlinear nonparametric systems. *Automatica*, 51. pp. 243-254. ISSN 0005-1098

<https://doi.org/10.1016/j.automatica.2014.10.069>

---

© 2014 Elsevier Ltd. This manuscript version is made available under the Creative Commons CC-BY-NC ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

# Kernel-Based Local Order Estimation of Nonlinear Nonparametric Systems <sup>★</sup>

Wenxiao Zhao <sup>a</sup>, Han-Fu Chen <sup>a</sup>, Er-wei Bai <sup>b</sup>, Kang Li <sup>c</sup>,

<sup>a</sup>*Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.*

<sup>b</sup>*Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA.*

<sup>c</sup>*School of Electronics, Electrical Engineering and Computer Science, Queen's University, Belfast, UK.*

---

## Abstract

We consider the local order estimation of nonlinear autoregressive systems with exogenous inputs (NARX), which may have different local dimensions at different points. By minimizing the kernel-based local information criterion introduced in the paper, the strongly consistent estimates for the local orders of the NARX system at points of interest are obtained. The modification of the criterion and a simple procedure of searching the minimum of the criterion, are also discussed. The theoretical results derived here are tested by simulation examples.

*Key words:* Nonlinear ARX system, recursive local linear estimator, order estimation, strong consistency.

---

## 1 Introduction

Consider a single-input single-output (SISO) nonlinear autoregressive system with exogenous input (NARX),

$$y_{k+1} = f(y_k, \dots, y_{k+1-M}, u_k, \dots, u_{k+1-M}) + \varepsilon_{k+1}, \quad (1)$$

where  $u_k$  and  $y_k$  are the system input and output, respectively,  $\varepsilon_k$  is the driven noise,  $M$  is the known upper bound of the true system order and  $f(\cdot)$  is the unknown function representing the system dynamics.

In recent years identification of system (1) has been an active research topic, estimating not only the nonlinear

function  $f(\cdot)$  itself ([3], [5], [16], [21], [23], [26], [27]) but also the system orders ([2], [4], [12], [13], [19], [20]). As far as the estimation of the nonlinear function  $f(\cdot)$  is of concern, the approaches can roughly be divided into two categories, the parametric approach ([16], [23], [24], [25], [28], [29]) and the nonparametric approach ([3], [5], [21], [26], [27]), according to the description of  $f(\cdot)$ . In the former, it is usually assumed that  $f(\cdot) = f(\cdot, \theta)$  with a known structure of  $f(\cdot)$  and an unknown parameter  $\theta$ , and consequently identification of  $f(\cdot)$  is transformed into a parametric optimization problem for  $\theta$ . While in the latter approach, it is often to estimate the values of  $f(\cdot)$  at the points of interest referred to as *Model on Demand* in the literature (see, e.g., [3][5][10][16][21][26]). The direct weight optimization [21], the local linear estimator [3] and its recursive version [27], the stochastic approximation algorithm [26] all belong to this class. Notice that most of the nonparametric identification algorithms are the weighted local average algorithms in a certain sense, and in order to derive the reliable estimates it requires to obtain the adequate measurements around the given points. In some applications, a global description of an unknown nonlinear system is too complicated both in structure and in dimension. This makes identification unreliable and the obtained model practically useless. Typical examples can be easily found in the fields of biology, atmospheric, geophysics, economy, engineering, communication, etc. An efficient and practical way is to

---

<sup>★</sup> The research of Wenxiao Zhao was supported by the National Key Basic Research Program of China (973 program) under grant No. 2014CB845301 and the National Natural Science Foundation (NSF) of China under Grants No. 61104052, 61273193, 61227902, and 61134013. The research of Han-Fu Chen was supported by the 973 program of China under grant No. 2014CB845301 and the NSF of China under Grants No. 61273193, 61120106011, and 61134013. The research of Er-Wei Bai was supported in part by NSF, DoE, IEC and NIH grants.

*Email addresses:* wxzhao@amss.ac.cn (Wenxiao Zhao), hfchen@iss.ac.cn (Han-Fu Chen), er-wei-bai@uiowa.edu (Er-wei Bai), k.li@qub.ac.uk (Kang Li).

split the task into a number of manageable pieces either in structure/dimension or in both. This is the idea of local modeling including local polynomial modeling, a hot topic in statistics. This paper studies the problem of the order of the local modeling.

Over the last a few decades considerable progress has been made on the order estimation as well as variable selection of linear stochastic systems. For example, the Akaike's information criterion (AIC) [1], Bayesian information criterion (BIC) and their generalizations [8], the recursive algorithms [9], the so-called LASSO [30], are a few among many others. But these approaches are not applicable to system (1) due to its nonparametric and nonlinear description. The order estimation for nonlinear systems has also been studied in recent years, e.g., [2], [6], [12], [14], [15], [17], [19], [20]. In [2] an approach to estimating the orders of the linearized nonlinear system is introduced. The so-called Lipschitz number approach and false nearest neighbors approach are proposed in [12] and [14], respectively, and successive research appeared in [6], [18], [19], etc. These two approaches do not identify the nonlinearity  $f(\cdot)$  itself, while estimating the orders. The methods in [2], [12], and [14] are however sensitive to the system noises, and, to the authors' knowledge, their convergence and consistency are unclear. The stepwise approach and the analysis of variance (ANOVA) approach are suggested in [17] and [20] based on hypothesis tests for the parameterized nonlinear systems. For these approaches a review is given in [13]. Note that the order estimation in the above papers is in a global sense, i.e., the true order is unique over the whole function domain. In contrast to this, sometimes the true orders of a nonlinear system are not unique and may vary from point to point. To this end, let us consider examples given below.

*Example (i):* A piecewise linear system is defined by

$$y_{k+1} = f_1(y_k, \dots, y_{k+1-M}, u_k, \dots, u_{k+1-M}) + \varepsilon_{k+1}, \quad (2)$$

with

$$f_1(y_k, \dots, y_{k+1-M}, u_k, \dots, u_{k+1-M}) = \begin{cases} a_1^{(1)} y_k + \dots + a_{p_1}^{(1)} y_{k+1-p_1} + b_1^{(1)} u_k + \dots + b_{q_1}^{(1)} u_{k+1-q_1}, & \text{if } [y_k, \dots, y_{k+1-M}, u_k, \dots, u_{k+1-M}]^T \in \mathcal{X}_1, \\ \vdots \\ a_1^{(s)} y_k + \dots + a_{p_s}^{(s)} y_{k+1-p_s} + b_1^{(s)} u_k + \dots + b_{q_s}^{(s)} u_{k+1-q_s}, & \text{if } [y_k, \dots, y_{k+1-M}, u_k, \dots, u_{k+1-M}]^T \in \mathcal{X}_s, \end{cases}$$

where  $\mathcal{X}_i$ ,  $i = 1, \dots, s$  is a partition of  $\mathbb{R}^{2M}$ .

*Example (ii):* The finite impulse response system is given by

$$y_{k+1} = f_2(u_k, u_{k-1}, u_{k-2}) + \varepsilon_{k+1}, \quad (3)$$

where  $f_2(u_k, u_{k-1}, u_{k-2}) = u_k u_{k-1} u_{k-2}$ , if  $u_k > 1$ ;  $= u_k u_{k-1}$ , if  $-1 \leq u_k \leq 1$ ; and  $= u_k$ , if  $u_k < -1$ .

These two examples demonstrate a need for the local order estimation at points of interest. To the authors' knowledge, there has not much been done on this topic, though in [4] a forward/backward approach was proposed. The numerical simulations seem to suggest that the forward/backward approach works well in terms of variable selection, but determination of the system order and its theoretical study remain open.

The contribution of the paper is as follows. First, a kernel-based local information criterion, for simplicity of reference, named as the local information criterion (LIC), is proposed for the local order estimation of system (1). Under moderate conditions, the estimates generated from LIC converge almost surely to the true local orders of system (1) at the points of interest. Second, a modification of LIC and a simple procedure of searching the minimum of LIC are suggested, and the strong consistency of the estimates is established as well.

The rest of the paper is arranged as follows. The LIC and the strong consistency of the estimates are given in Section II. A modification of LIC is discussed in Section III. Two simulation examples are given in Section IV and some concluding remarks are addressed in Section V. Some technical proofs are placed in Appendix.

*Notations.* Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be the basic probability space. Let  $\mathcal{B}^m$  denote the Borel  $\sigma$ -algebra on  $\mathbb{R}^m$ . For a vector  $\mathbf{x}(m) = [x_1 \dots x_m]^T \in \mathbb{R}^m$ , denote its Euclidean norm by  $\|\mathbf{x}(m)\|$  and its sub-vector by  $\mathbf{x}(i:j) \triangleq [x_i \dots x_j]^T \in \mathbb{R}^{j-i+1}$ . Denote by  $\|\nu(\cdot)\|_{\text{var}}$  the total variation norm of a signed measure  $\nu(\cdot)$ . For two positive sequences  $\{a_N\}_{N \geq 1}$  and  $\{b_N\}_{N \geq 1}$ , by  $a_N \sim b_N$  it means that  $c_1 b_N \leq a_N \leq c_2 b_N$ ,  $\forall N \geq 1$ , for some positive constants  $c_1$  and  $c_2$ . Denote by  $\nabla f(\cdot)$  the gradient of the function  $f(\cdot)$  if it exists.

## 2 Local Order Estimation

### 2.1 Local Information Criterion for Order Estimation

We further introduce the following notations. Notice that the nonlinear function  $f(\cdot)$  in (1) is defined on  $\mathbb{R}^{2M}$ . The regressor and the point of interest in  $\mathbb{R}^{2M}$  are denoted by  $\varphi_k(M, M)$  and  $\mathbf{x}^*(2M)$ , respectively,

$$\varphi_k(M, M) = [y_k \dots y_{k+1-M} \ u_k \dots u_{k+1-M}]^T, \quad (4)$$

$$\mathbf{x}^*(2M) = [x_1^*, \dots, x_{2M}^*]^T. \quad (5)$$

Similar to (4), for any fixed  $1 \leq p \leq M$  and  $1 \leq q \leq M$

let us define

$$\varphi_k(p, q) = [y_k \cdots y_{k+1-p} u_k \cdots u_{k+1-q}]^T, \quad (6)$$

$$\mathbf{x}^*(p, q) = [x_1^*, \dots, x_p^*, x_{M+1}^*, \dots, x_{M+q}^*]^T. \quad (7)$$

From the examples given in Introduction, it is seen that the orders of nonlinear systems may be varying from point to point. This is a different picture from linear systems. The question is how to define and estimate the local order of  $f(\cdot)$  at the given  $\mathbf{x}^*(2M)$  based on the observations  $\{y_k, u_k\}_{k \geq 1}$ . A direct approach is to define the local order of  $f(\cdot)$  at  $\mathbf{x}^*(2M)$  as the number of variables that contribute to the function value  $f(\mathbf{x}^*(2M))$ . However, if the system order is defined in such a manner, it is difficult to choose the quantitative information based on which the algorithms estimating the local order can be designed, since  $f(\cdot)$  is nonlinear and nonparametric. On the other hand, it is clear that the function  $f(x)$  can be well approximated by a local linear model if  $x$  is close to  $\mathbf{x}^*(2M)$ , i.e.,

$$f(\mathbf{x}(2M)) = f(\mathbf{x}^*(2M)) + \nabla f(\mathbf{x}^*(2M))^T \cdot (\mathbf{x}(2M) - \mathbf{x}^*(2M)) + O\left(\|\mathbf{x}(2M) - \mathbf{x}^*(2M)\|^2\right) \quad (8)$$

$\forall \|\mathbf{x}(2M) - \mathbf{x}^*(2M)\| \leq \varepsilon$  for small enough  $\varepsilon > 0$ . Denote the gradient of  $f(\cdot)$  at  $\mathbf{x}^*(2M)$  by  $\nabla f(\mathbf{x}^*(2M)) \triangleq \left[ \frac{\partial f}{\partial x_1^*} \cdots \frac{\partial f}{\partial x_M^*} \frac{\partial f}{\partial x_{M+1}^*} \cdots \frac{\partial f}{\partial x_{2M}^*} \right]^T \in \mathbb{R}^{2M}$  if it exists. It is clear that if  $f(\mathbf{x}^*(2M))$  depends only on  $(p_0 + q_0)$  variables, i.e.,

$$\begin{aligned} f(\mathbf{x}^*(2M)) &= f(x_1^*, \dots, x_{2M}^*) \\ &= f(x_1^*, \dots, x_{p_0}^*, \mathbf{x}^T(M - p_0), \\ &\quad x_{M+1}^*, \dots, x_{M+q_0}^*, \mathbf{x}^T(M - q_0)) \end{aligned} \quad (9)$$

$\forall \mathbf{x}(M - p_0) \in \mathbb{R}^{M-p_0}$  and  $\forall \mathbf{x}(M - q_0) \in \mathbb{R}^{M-q_0}$ , then  $\frac{\partial f}{\partial x_i^*} = 0$  for  $i = p_0 + 1, \dots, M$  and  $M + q_0 + 1, \dots, 2M$ , i.e.,

$$\nabla f(\mathbf{x}^*(2M)) = \begin{bmatrix} \frac{\partial f}{\partial x_1^*} \cdots \frac{\partial f}{\partial x_{p_0}^*} \underbrace{0 \cdots 0}_{M-p_0} \\ \frac{\partial f}{\partial x_{M+1}^*} \cdots \frac{\partial f}{\partial x_{M+q_0}^*} \underbrace{0 \cdots 0}_{M-q_0} \end{bmatrix}^T. \quad (10)$$

From (8) and (10) it is seen that if we can find a local linear model of  $f(\cdot)$  at  $\mathbf{x}^*(2M)$ , then we can estimate the local order by determining the biggest  $p$  and  $q$  such that  $\partial f / \partial x_p^* \neq 0$ ,  $1 \leq p \leq M$  and  $\partial f / \partial x_{M+q}^* \neq 0$ ,  $1 \leq q \leq M$ .

To this end, we further impose the following assumptions.

- A1) The finite upper bound  $M$  for orders  $(p, q)$  is known;
- A2)  $|f(x)| \leq c_1 \|x\|^r + c_2$ ,  $x \in \mathbb{R}^{2M}$  for some positive constants  $c_1$ ,  $c_2$ , and  $r$  and  $f(\cdot)$  is twice differentiable at  $\mathbf{x}^*(2M)$ . Further,  $\partial f / \partial x_p^* \neq 0$  and  $\partial f / \partial x_{M+q}^* \neq 0$  for some  $p = 1, \dots, M$  and  $q = 1, \dots, M$ .

**Definition 1** The local order of  $f(\cdot)$  at  $\mathbf{x}^*(2M)$  is defined as  $(s_0, t_0)$ , where

$$\begin{aligned} s_0 &\triangleq \max \left\{ p = 1, \dots, M \mid \frac{\partial f}{\partial x_p^*} \neq 0 \right\} \\ t_0 &\triangleq \max \left\{ q = 1, \dots, M \mid \frac{\partial f}{\partial x_{M+q}^*} \neq 0 \right\}. \end{aligned}$$

It is natural to ask why  $(s_0, t_0)$  rather than  $(p_0, q_0)$  given in (9) is defined as the local order of  $f(\cdot)$  at  $\mathbf{x}^*(2M)$ ? Do we need to take the second order derivatives into consideration? By the Taylor expansion, we know that a local linear estimator approximates  $f(\cdot)$  at  $\mathbf{x}^*(2M)$  well if  $\mathbf{x}(2M) \in \mathbb{R}^{2M}$  is close to  $\mathbf{x}^*(2M)$  and the second order terms can be neglected. In this regard, it is reasonable to find the local order of  $f(\cdot)$  at  $\mathbf{x}^*(2M)$  from its local linear approximates. On the other hand, it is clear that if  $\partial f / \partial x_{p_0}^* \neq 0$  and  $\partial f / \partial x_{M+q_0}^* \neq 0$ , then  $(s_0, t_0) = (p_0, q_0)$ . But sometimes, the local order given by Definition 1 is smaller than  $(p_0, q_0)$ . Next we provide two examples to illustrate Definition 1.

*Example (iii):* For the linear system  $y_{k+1} = a_1 y_k + \cdots + a_{p_0} y_{k+1-p_0} + b_1 u_k + \cdots + b_{q_0} u_{k+1-q_0} + \varepsilon_{k+1}$  with  $a_{p_0} \neq 0$ ,  $b_{q_0} \neq 0$  we have  $f(\mathbf{x}^*(2M)) = a_1 x_1^* + \cdots + a_{p_0} x_{p_0}^* + b_1 x_{M+1}^* + \cdots + b_{q_0} x_{M+q_0}^*$ . It is clear that  $\partial f / \partial x_{p_0}^* \neq 0$ ,  $\partial f / \partial x_{M+q_0}^* \neq 0$ , and  $\partial f / \partial x_i^* = 0$ ,  $i = p_0 + 1, \dots, M, M + q_0 + 1, \dots, 2M$ . Thus for this example the system order  $(s_0, t_0)$  derived by Definition 1 equals  $(p_0, q_0)$ , which is consistent with the linear system theory.

*Example (iv):* For the nonlinear system  $y_{k+1} = a y_k y_{k-1} + b u_k u_{k-1} + \varepsilon_{k+1}$  with  $a \neq 0$ ,  $b \neq 0$ , we have  $f(\mathbf{x}(4)) = f(x_1, x_2, x_3, x_4) = a x_1 x_2 + b x_3 x_4$ . At the fixed point  $\mathbf{x}^*(4) = [0 \ 1 \ 0 \ 1]^T \in \mathbb{R}^4$ , it is clear that  $\nabla f(\mathbf{x}^*(4)) = [a \ 0 \ b \ 0]^T$ , and by the Taylor expansion  $f(\mathbf{x}(4)) = f(\mathbf{x}^*(4)) + a \frac{\partial f}{\partial x_1^*} (x_1 - x_1^*) + b \frac{\partial f}{\partial x_3^*} (x_3 - x_3^*)$  for all  $\mathbf{x}(4)$  close to  $\mathbf{x}^*(4)$ . This implies that the local order at the given point should be  $(s_0, t_0) = (1, 1)$ .

Based on the above discussion the key step of our approach to estimating the local order is to find the local linear model of  $f(\cdot)$  at  $\mathbf{x}^*(2M)$ . In [3] and [27], the kernel function-based local linear estimator (LLE) and

its recursive version (RLLE) are considered, which estimate the values of the nonlinear function at fixed points together with their gradients. Let us first reformulate the RLLE introduced in [27], which the order estimation algorithm is essentially based on. Notice that the RLLE in [27] is with known system orders, but here the orders  $(p, q)$  in the algorithm may vary in the set  $\{(p, q) : 1 \leq p \leq M, 1 \leq q \leq M\}$ .

With the given order  $(p, q)$  and measurements  $\{u_k, y_{k+1}\}_{k=1}^N$  the RLLE estimate of  $f(\cdot)$  at time  $N + 1$  is given by

$$\begin{aligned} \theta_{N+1}(p, q) &= \left[ \theta_{0, N+1}(p, q) \ \theta_{1, N+1}^T(p, q) \right]^T \\ &\triangleq \underset{\substack{\theta_0(p, q) \in \mathbb{R} \\ \theta_1(p, q) \in \mathbb{R}^{p+q}}}{\operatorname{argmin}} \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) \left( y_{k+1} - \theta_0(p, q) \right. \\ &\quad \left. - \theta_1(p, q)^T (\varphi_k(p, q) - \mathbf{x}^*(p, q)) \right)^2, \end{aligned} \quad (11)$$

where the kernel function  $w_k(\mathbf{x}^*(2M))$  is given by

$$w_k(\mathbf{x}^*(2M)) = \frac{1}{b_k^{2M}} w \left( \frac{1}{b_k} (\varphi_k(M, M) - \mathbf{x}^*(2M)) \right). \quad (12)$$

Notice that  $\theta_{N+1}(p, q) = \left[ \theta_{0, N+1}(p, q) \ \theta_{1, N+1}^T(p, q) \right]^T$ . With the given order  $(p, q)$ ,  $\theta_{0, N+1}(p, q)$  serves as the estimate for  $f(\mathbf{x}^*(M, M))$  while  $\theta_{1, N+1}(p, q)$  for  $\nabla f(\mathbf{x}^*(M, M))$ .

Set

$$X_k(p, q) \triangleq \begin{bmatrix} 1 \\ \varphi_k(p, q) - \mathbf{x}^*(p, q) \end{bmatrix}. \quad (13)$$

By some simple manipulations, RLLE in (11) can be expressed by

$$\begin{aligned} \theta_{N+1}(p, q) &= \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) X_k(p, q)^T \right)^{-1} \\ &\quad \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) y_{k+1} \right), \end{aligned} \quad (14)$$

if the matrices  $\sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) X_k(p, q)^T$ ,  $N \geq 1$  are nonsingular. Notice that by the matrix inverse lemma,  $\theta_{N+1}(p, q)$  given by (14) can be computed in a recursive way.

**Remark 1** A widely used kernel is the Gaussian pdf,

and in this case we have

$$w_k(\mathbf{x}^*(2M)) = \frac{1}{(2\pi)^M} \frac{1}{b_k^{2M}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^M \left( \frac{y_{k+1-i} - x_i^*}{b_k} \right)^2 - \frac{1}{2} \sum_{j=1}^M \left( \frac{u_{k+1-j} - x_{M+j}^*}{b_k} \right)^2 \right\}.$$

Other important kernels include the rectangle kernel, triangle kernel, Epanechnikov kernel, etc.

**Remark 2** From the above example we see that the kernel function plays the role like a weight: The regressors  $\varphi_k(M, M)$  close to  $\mathbf{x}^*(M, M)$  are taken into considerably higher account in comparison with those far away from  $\mathbf{x}^*(M, M)$ , because the kernel  $w_k(\mathbf{x}^*(2M))$  rapidly vanishes as the regressors deviates from  $\mathbf{x}^*(M, M)$ . As for the sequence  $\{b_k\}_{k \geq 1}$ , it is usually required that  $b_N \rightarrow 0$  but  $b_N^{2M} N \rightarrow \infty$  as  $N \rightarrow \infty$ . Thus, the number of data around  $\mathbf{x}^*(2M)$  is increasing, and the estimates  $\theta_{0, N}(p, q)$  and  $\theta_{1, N}(p, q)$  generated by (11)–(12) are approaching to  $f(\mathbf{x}^*(M, M))$  and  $\nabla f(\mathbf{x}^*(M, M))$ , respectively, as  $N \rightarrow \infty$ , provided the orders  $(p, q)$  match the true system orders well.

We introduce the the following assumption which is adopted in [27] for the convergence analysis of RLLE.

- A3) Select  $b_k = 1/k^\delta$  for some  $\delta \in (0, 1/(2(2M+1)))$ ;  $w(\cdot)$  is chosen as a symmetric probability density function (pdf) with  $w(x) = O(\rho^{\|x\|})$  for some  $0 < \rho < 1$  as  $\|x\| \rightarrow \infty$ , and  $\int_{\mathbb{R}^{2M}} w(x) x x^T dx > 0$ .

For estimating the local order  $(s_0, t_0)$ , we introduce the following local information criterion (LIC)  $L_{N+1}(p, q)$ :

$$L_{N+1}(p, q) \triangleq \sigma_{N+1}(p, q) + a_N \cdot (p + q), \quad (15)$$

where

$$\begin{aligned} \sigma_{N+1}(p, q) &\triangleq \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) \left( y_{k+1} - \theta_{0, N+1}(p, q) \right. \\ &\quad \left. - \theta_{1, N+1}(p, q)^T (\varphi_k(p, q) - \mathbf{x}^*(p, q)) \right)^2, \end{aligned} \quad (16)$$

$\{a_N\}_{N \geq 1}$  is a positive sequence tending to infinity as  $N \rightarrow \infty$ , and  $\theta_{0, N+1}(p, q)$  and  $\theta_{1, N+1}(p, q)$  are RLLE generated by (11) with the given order  $(p, q)$ .

The order estimate  $(p_{N+1}, q_{N+1})$  of  $(s_0, t_0)$  is defined by minimizing  $L_{N+1}(p, q)$ :

$$(p_{N+1}, q_{N+1}) \triangleq \underset{\substack{1 \leq p \leq M \\ 1 \leq q \leq M}}{\operatorname{argmin}} L_{N+1}(p, q). \quad (17)$$

**Remark 3** Notice that  $\partial f/\partial x_i^* = 0$ ,  $i = s_0 + 1, \dots, M, M+t_0+1, \dots, 2M$ . Thus if  $RLLE \theta_N(p, q) = [\theta_{0,N}(p, q) \theta_{1,N}^T(p, q)]^T$  approximates the true value well, then the function  $\sigma_{N+1}(p, q)$  decreases as  $p$  and  $q$  increase but the performance may not change much for  $p \geq s_0$  and  $q \geq t_0$ . On the other hand,  $(p+q)$  increases as  $p$  and  $q$  increases. This indicates that (17) with appropriately chosen  $\{a_N\}_{N \geq 1}$  defines a reasonable estimate for  $(s_0, t_0)$ .

We list some further conditions used for convergence analysis of the order estimates. Note that (1) is an infinite impulse response nonlinear system, and the second order statistics may not contain adequate information for its identification. So ergodicity and mixing properties are often required, see, e.g., [10] in statistics literature.

- A4)  $\{\varepsilon_k\}_{k \geq 0}$  is a sequence of independent and identically distributed (iid) random variables with  $E\varepsilon_k = 0$ ,  $0 < E|\varepsilon_k|^{2+\eta} < \infty$  for some  $\eta \in (0, 2]$ ;  $\varphi_k(M, M)$  and  $\varepsilon_{k+1}$  are mutually independent for each  $k \geq 1$ .
- A5) The sequence  $\{\varphi_k(M, M)\}_{k \geq 1}$  is geometrically ergodic, i.e., there exists an invariant probability measure  $P_{IV}(\cdot)$  on  $(\mathbb{R}^{2M}, \mathcal{B}^{2M})$  and some constants  $c_1 > 0$  and  $0 < \rho_1 < 1$  such that  $\|P_k(\cdot) - P_{IV}(\cdot)\|_{\text{var}} \leq c_1 \rho_1^k$ , where  $P_k(\cdot)$  is the marginal distribution of  $\varphi_k(M, M)$ .  $P_{IV}(\cdot)$  is with a bounded pdf, denoted by  $f_{IV}(\cdot)$ , which is with a continuous second order derivative at  $\mathbf{x}^*(2M)$ .
- A6)  $\{\varphi_k(M, M)\}_{k \geq 1}$  is an  $\alpha$ -mixing with mixing coefficients  $\{\alpha(k)\}_{k \geq 1}$  satisfying  $\alpha(k) \leq c_2 \rho_2^k$  for some  $c_2 > 0$  and  $0 < \rho_2 < 1$  and  $E\|\varphi_k(M, M)\|^r < \infty$  for  $k \geq 1$ , where the constant  $r$  is specified in assumption A2).
- A7) The sequence  $\{a_N\}_{N \geq 1}$  satisfies

$$N^{1-4\delta}/a_N \xrightarrow{N \rightarrow \infty} 0, \quad a_N/N^{1-2\delta} \xrightarrow{N \rightarrow \infty} 0, \quad (18)$$

where  $\delta > 0$  is given in A3).

The conditions A5) and A6), in fact, are on the asymptotical independency and stationarity of the sequence  $\{\varphi_k(M, M)\}_{k \geq 1}$ , and they can be guaranteed by assuming stability of the system with input excited in a certain sense as shown in [26] and [27]. The conditions given in [26] and [27] cover a large class of systems, including the ARX system, the Hammerstein systems, and the Wiener system, etc. So for ease of presentation, in this paper we assume that  $\{\varphi_k(M, M)\}_{k \geq 1}$  is a mixing process with an asymptotically stationary distribution.

The convergence of (17) is considered in the next section.

## 2.2 Strong Consistency of Estimates

For any fixed  $1 \leq p \leq M$  and  $1 \leq q \leq M$ , define

$$\nabla f(\mathbf{x}^*(p, q)) \triangleq \left[ \frac{\partial f}{\partial x_1^*} \cdots \frac{\partial f}{\partial x_p^*} \frac{\partial f}{\partial x_{M+1}^*} \cdots \frac{\partial f}{\partial x_{M+q}^*} \right]^T, \quad (19)$$

and

$$\begin{aligned} \bar{\theta}_{1,N+1}(p, q) &\triangleq \begin{bmatrix} \theta_{1,N+1}(p, q)(1:p)^T & \underbrace{0 \cdots 0}_{M-p} \\ \theta_{1,N+1}(p, q)(p+1:p+q)^T & \underbrace{0 \cdots 0}_{M-q} \end{bmatrix}^T \in \mathbb{R}^{2M}, \quad (20) \\ \tilde{\theta}_{N+1}(p, q) &\triangleq [f(\mathbf{x}^*(2M)) - \theta_{0,N+1}(p, q) \\ \nabla f(\mathbf{x}^*(2M))^T - \bar{\theta}_{1,N+1}(p, q)^T]^T \in \mathbb{R}^{1+2M}. \quad (21) \end{aligned}$$

Denote the maximal and minimal eigenvalues of  $\sum_{i=1}^N w_i(\mathbf{x}^*(2M)) X_i(p, q) X_i(p, q)^T$  by  $\lambda_{\max}^{(p,q)}(N)$  and  $\lambda_{\min}^{(p,q)}(N)$ , respectively.

**Theorem 1** Under conditions A1)-A7), the order estimate  $(p_N, q_N)$  given by (17) is strongly consistent,

$$(p_N, q_N) \xrightarrow{N \rightarrow \infty} (s_0, t_0) \text{ a.s.} \quad (22)$$

*Proof:* See Appendix.  $\square$

## 2.3 A Simple Procedure for Searching the Minimum of LIC

To obtain estimates defined by (17) it is required to calculate  $M^2$  function values of  $L_{N+1}(p, q)$  and then to find the minimum among them. In this section we introduce a simple procedure for searching the minimum of (17) for which the computational complexity is  $O(M)$ .

Define

$$\hat{p}_{N+1} \triangleq \underset{1 \leq p \leq M}{\operatorname{argmin}} L_{N+1}(p, M), \quad (23)$$

$$\hat{q}_{N+1} \triangleq \underset{1 \leq q \leq M}{\operatorname{argmin}} L_{N+1}(\hat{p}_{N+1}, q), \quad (24)$$

where  $L_{N+1}(p, q)$  is defined by (15).

**Theorem 2** Assume A1)-A7) hold. Then

$$\hat{p}_N \xrightarrow{N \rightarrow \infty} s_0 \text{ a.s.} \quad (25)$$

$$\hat{q}_N \xrightarrow{N \rightarrow \infty} t_0 \text{ a.s.} \quad (26)$$

*Proof:* Here we just sketch the proof. The proof is divided into two steps. First, the strong consistency of  $\hat{p}_N$  is proved. This can be done by carrying out almost the same discussion as that given in Theorem 1. Second, based on that  $\hat{p}_N = s_0$  and hence  $L_{N+1}(\hat{p}_{N+1}, q) = L_{N+1}(s_0, q)$  for all  $N$  large enough, the convergence of  $\hat{q}_N$  is established via a similar derivation as that for (25).  $\square$

**Remark 4** The order estimates can also be defined by

$$\hat{q}_{N+1} \triangleq \operatorname{argmin}_{1 \leq q \leq M} L_{N+1}(M, q), \quad (27)$$

$$\hat{p}_{N+1} \triangleq \operatorname{argmin}_{1 \leq p \leq M} L_{N+1}(p, \hat{q}_{N+1}), \quad (28)$$

which are strongly consistent under A1)-A7).

### 3 Modified LIC

In the last section, based on LIC the strongly consistent estimate for the system order at a fixed point is obtained. We now introduce a modified LIC as follows:

$$\bar{L}_{N+1}(p, q) \triangleq N \log \sigma_{N+1}(p, q) + a_N \cdot (p + q), \quad (29)$$

where  $\sigma_{N+1}(p, q)$  is given by (16).

The estimate  $(\bar{p}_{N+1}, \bar{q}_{N+1})$  for  $(s_0, t_0)$  is given by minimizing  $\bar{L}_{N+1}(p, q)$ , i.e.,

$$(\bar{p}_{N+1}, \bar{q}_{N+1}) \triangleq \operatorname{argmin}_{\substack{1 \leq p \leq M \\ 1 \leq q \leq M}} \bar{L}_{N+1}(p, q). \quad (30)$$

**Theorem 3** Under conditions A1)-A7), the order estimate  $(\bar{p}_N, \bar{q}_N)$  given by (30) is strongly consistent,

$$(\bar{p}_N, \bar{q}_N) \xrightarrow{N \rightarrow \infty} (s_0, t_0) \text{ a.s.} \quad (31)$$

*Proof:* See Appendix.  $\square$

Define

$$\tilde{p}_{N+1} \triangleq \operatorname{argmin}_{1 \leq p \leq M} \bar{L}_{N+1}(p, M), \quad (32)$$

$$\tilde{q}_{N+1} \triangleq \operatorname{argmin}_{1 \leq q \leq M} \bar{L}_{N+1}(\tilde{p}_{N+1}, q), \quad (33)$$

where  $\bar{L}_{N+1}(p, q)$  is defined by (29).

Similar to Theorem 2, the following result holds.

**Theorem 4** Assume A1)-A7) hold. Then

$$\tilde{p}_N \xrightarrow{N \rightarrow \infty} s_0 \text{ a.s.} \quad (34)$$

$$\tilde{q}_N \xrightarrow{N \rightarrow \infty} t_0 \text{ a.s.} \quad (35)$$

**Remark 5** The order estimates can also be defined by

$$\bar{\bar{q}}_{N+1} \triangleq \operatorname{argmin}_{1 \leq q \leq M} \bar{L}_{N+1}(M, q), \quad (36)$$

$$\bar{\bar{p}}_{N+1} \triangleq \operatorname{argmin}_{1 \leq p \leq M} \bar{L}_{N+1}(p, \bar{\bar{q}}_{N+1}), \quad (37)$$

which are strongly consistent under A1)-A7).

**Remark 6** LIC and its modification considered in the paper look similar to the well known AIC, BIC, and their generalizations. However, AIC, BIC, and others are in a global sense and thus they are inapplicable to the local order estimation. While for LIC the kernel function  $w_k(\mathbf{x}^*(2M))$  plays a bandwidth like role to stress those measurements which are close to the given point and to take their average. The sequence  $\{a_N\}$  in AIC, BIC, and their generalizations can be chosen as  $N^\alpha$  for any  $0 < \alpha < 1$ , or  $\log^{1+\beta} N$  for some  $\beta \geq 0$ , or even a constant ([7], [8]), but here in LIC the choice of  $\{a_N\}$  is more delicate.

### 4 Discussions and Simulations

In the above sections, we have introduced two criteria, i.e.,  $L_N(p, q)$  defined by (15) and  $\bar{L}_N(p, q)$  defined by (29), respectively. Theoretically, any  $a_N$  that meets the requirement in assumption A7), for example,  $a_N = cN^{1-3\delta}$  for any constant  $c > 0$ , guarantees the a.s. convergence of the estimates generated by (15) and (29). However, from the numerical calculation point of view, there exists some difference between  $L_N(p, q)$  and  $\bar{L}_N(p, q)$ .

- (i) Let us take  $a_N = cN^{1-3\delta}$  for some constant  $c > 0$  as an example. As required in assumption A3), the parameter  $\delta$  usually is small and thus even for the integer  $N > 0$  large enough it still holds that  $N^{1-3\delta} \approx N$ . On the other hand, since the kernel function  $w_k(\mathbf{x}^*(2M))$  is involved in the residual term, i.e.,

$$\sigma_{N+1}(p, q) \triangleq \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) \left( y_{k+1} - \theta_{0,N+1}(p, q) - \theta_{1,N+1}(p, q)^T (\varphi_k(p, q) - \mathbf{x}^*(p, q)) \right)^2,$$

it often holds that  $\sigma_{N+1}(p, q) = o(N)$  and thus  $a_N(p+q)$  is the dominated term in  $L_{N+1}(p, q)$ , i.e.,

$$L_{N+1}(p, q) = \sigma_{N+1}(p, q) + a_N \cdot (p+q) \approx O(N \cdot (p+q)).$$

This indicates that for convergence of the estimates generated from the criterion  $L_N(p, q)$ , in order to balance the penalty term  $a_N(p+q)$  it usually requires the number of data be large enough, and thus the convergence rate is slow. To speed up the convergence rate, one may choose, for example,  $a_N = cN^{1-3\delta}$  for some  $c > 0$  small enough to reduce the effect of the penalty term  $a_N \cdot (p+q)$  in  $L_N(p, q)$ .

- (ii) By noticing the first term in  $\bar{L}_N(p, q)$  defined by (29), it can be found that  $N = o(N \log \sigma_{N+1}(p, q))$  and thus  $a_N \cdot (p+q)$  with  $a_N$  satisfying A7) is a moderate penalty term in  $\bar{L}_N(p, q)$ . So the convergence rate of estimates generated from  $\bar{L}_N(p, q)$  should be faster than that of (17).

In the following we present the numerical simulations to verify the theoretical analysis.

*Example 1.* Consider an FIR system

$$y_{k+1} = f(u_k, u_{k-1}, u_{k-2}) + \varepsilon_{k+1}, \quad (38)$$

$$f(u_k, u_{k-1}, u_{k-2}) = \begin{cases} u_k + u_{k-1} + u_{k-2}, & \text{if } u_k > 1, \\ u_k + u_{k-1}, & \text{if } -1 \leq u_k \leq 1, \\ u_k, & \text{if } u_k < -1, \end{cases}$$

where the inputs  $\{u_k\}_{k \geq 1}$  and the noises  $\{\varepsilon_k\}_{k \geq 1}$  are mutually independent iid Gaussian random variables with distributions  $\mathcal{N}(0, 2^2)$  and  $\mathcal{N}(0, 0.1^2)$ , respectively. It is noticed that the right-hand side of (38) is free of the system output, so  $\mathbf{x}^*(2M)$  defined by (5) changes to  $\mathbf{x}^*(M) \triangleq [u_1^*, \dots, u_M^*]^T$ , and  $L_{N+1}(p, q)$  and  $\sigma_{N+1}(p, q)$  defined by (15) and (16) correspondingly change to functions  $L_{N+1}(q)$  and  $\sigma_{N+1}(q)$ , respectively. Assume the upper bound  $M$  of system orders is 4. Thus,  $\mathbf{x}^*(M) = [u_1^*, \dots, u_4^*]^T$ . We choose two points for test,  $\mathbf{x}_1^*(M) = [2 \ 1 \ 1 \ 0]^T$  and  $\mathbf{x}_2^*(M) = [0 \ 0 \ 0 \ 0]^T$ . Note that the true local orders are different at the two points.

More than 30 simulations have been performed. Here we only present one of them since the performance of others is almost the same. Tables 1 and 2 show the performance of the proposed estimator with the data set  $\{u_k, y_{k+1}\}_{k=1}^N$  for  $N = 1000, 2000, 3000, 4000$ , and  $5000$ , respectively, where

$$L_{N+1}(q) = \sigma_{N+1}(q) + 0.005N^{1-3\delta} \cdot q, \quad (39)$$

$$\bar{L}_{N+1}(q) = N \log \sigma_{N+1}(q) + 0.5N^{1-3\delta} \cdot q \quad (40)$$

with  $\delta = 0.05$ .

It can be found that the criteria  $L_{N+1}(q)$  and  $\bar{L}_{N+1}(q)$  always give the correct order estimates 3 and 2 for the local orders of the system at  $\mathbf{x}_1^*(M)$  and  $\mathbf{x}_2^*(M)$ , respectively. It can also be found that the convergence rate of estimates generated from  $\bar{L}_{N+1}(q)$  is faster than that generated from  $L_{N+1}(q)$ .

*Example 2.* Consider a benchmark problem for nonlinear system identification ([5], [27]):

$$x_1(k+1) = \left( \frac{x_1(k)}{1+x_1^2(k)} + 1 \right) \sin x_2(k),$$

$$x_2(k+1) = x_2(k) \cos x_2(k) + x_1(k) \exp \left( -\frac{x_1^2(k)+x_2^2(k)}{8} \right) + \frac{u_k^3}{1+u_k^2+0.5 \cos(x_1(k)+x_2(k))},$$

$$y_k = \frac{x_1(k)}{1+0.5 \sin x_2(k)} + \frac{x_2(k)}{1+0.5 \sin x_1(k)} + \varepsilon_k,$$

where  $u_k$  and  $y_k$  are the system input and output, respectively,  $\varepsilon_k$  is the system noise with Gaussian distribution  $\varepsilon_k \in \mathcal{N}(0, \sigma^2)$ ,  $\sigma = 0.1$ , and  $x_1(k)$  and  $x_2(k)$  are the unmeasured system states.

The NARX system

$$y_{k+1} = f(y_k, \dots, y_{k-M}, u_k, \dots, u_{k-M}) + \varepsilon_{k+1}$$

is used to approximate the unknown system. Notice that in existing literature [5][27], a common choice for the order  $M$  is  $M = 3$ . Here we adopt  $M = 3$  as the upper bound for the system order.

First,  $N(= 1000)$  samples  $\{u_k, y_k\}_{k=1}^{1000}$  are generated by iid  $u_k$  with Gaussian distribution  $u_k \in \mathcal{N}(0, 1)$ . The local orders as well as the values of the function  $f(\cdot)$  and its gradients  $\nabla f(\cdot)$  are estimated based on  $\{u_k, y_k\}_{k=1}^{1000}$ . Then the input signals  $u_k = \sin \frac{\pi k}{5} + \sin \frac{2\pi k}{25}$ ,  $k = N+1, \dots, N+100$  are fed into the estimated model to calculate the one-step predicted output. Specifically, the intervals  $[-3, 3]$  and  $[-2, 2]$  are equally divided into 5 and 4 subintervals, respectively and the domain of interest  $S = \{(y_3, y_2, y_1, u_3, u_2, u_1) \in \mathbb{R}^6 \mid y_3 \in [-3, 3], y_2 \in [-3, 3], y_1 \in [-3, 3], u_3 \in [-2, 2], u_2 \in [-2, 2], u_1 \in [-2, 2]\}$  is uniformly divided into 8000 disjoint small cubics  $S = \bigcup_{i=1}^{8000} S_i$  and from each  $S_i$  a point  $\varphi_i^*$  is randomly chosen,  $i = 1, \dots, 8000$ . Then with  $\delta = 0.04$  and  $a_N = N^{1-3\delta}$ , the algorithms (14) and (29) are applied to estimate the local orders denoted by  $(p_{N,i}, q_{N,i})$ , and parameters denoted by  $f_N(\varphi_i^*(p_{N,i}, q_{N,i}))$  and  $\nabla f_N(\varphi_i^*(p_{N,i}, q_{N,i}))$  at each  $\varphi_i^*$ ,  $i = 1, \dots, 8000$ , where  $\varphi_i^*(p_{N,i}, q_{N,i})$  is a  $(p_{N,i} + q_{N,i})$ -vector defined by (7). Then the one-step predictions are given as follows,

$$\hat{y}_{k+1} = \hat{f}_N(\varphi_i^*(p_{N,i}, q_{N,i})) + \nabla \hat{f}_N(\varphi_i^*(p_{N,i}, q_{N,i}))^T (\hat{\varphi}_k(p_{N,i}, q_{N,i}) - \varphi_i^*(p_{N,i}, q_{N,i})), \quad (41)$$

with regressor

$$\hat{\varphi}_k(p_{N,i}, q_{N,i}) = [\hat{y}_k, \dots, \hat{y}_{k-p_{N,i}}, u_k, \dots, u_{k-q_{N,i}}]^T,$$

if  $\hat{\varphi}_k(3, 3) \in S_i$  for some  $i = 1, \dots, 8000$  where  $k = N+1, \dots, N+100$ .

Ten simulations are performed. Figures 1 and 2 shows one of the simulations. In Figure 1 the solid lines are the actual output  $y_k$ ,  $k = N+1, \dots, N+100$ , the dotted line the predicted output generated by (41) and the dashed line the predicted output generated by (42) without order estimation, i.e.,

$$\hat{y}_{k+1} = \hat{f}_N(\varphi_i^*(3, 3)) + \nabla \hat{f}_N(\varphi_i^*(3, 3))^T (\hat{\varphi}_k(3, 3) - \varphi_i^*(3, 3)). \quad (42)$$

Figure 2 shows the estimated orders at the 8000 given points. Notice that the blocks at the bottom of the figure represent the 1st to the 100th points while those at the top of the figure represent the 7901st to the 8000th points. The estimated orders are indicated with different depths of color.

To test the performance of algorithm, the following quality of fit (QOF) is calculated

$$\left(1 - \frac{\sum_{k=N+1}^{N+100} (y_k - \hat{y}_k)^2}{\sum_{k=N+1}^{N+100} \left(y_k - \frac{1}{N} \sum_{t=N+1}^{N+100} y_t\right)^2}\right) \times 100\%, \quad (43)$$

where  $\hat{y}_k$  is the predicted output.

Table 3 shows the average of QOF of the ten simulations and the standard deviation. From Figures 1, 2 and Table 3 we find that the performance of algorithm (41) is similar to that of algorithm (42). However, from Figure 2 we find that the estimated local orders are reduced at many of the 8000 given points. Thus the benefit of algorithm (41) is that to apply the order estimation technique the complexity of the identified model for the benchmark problem is reduced and therefore a more precise system model is obtained.

## 5 Concluding Remarks

In the paper LIC is suggested for the local order estimation of NARX systems and the consistency of the estimates is established. Some important issues connected with LIC are summarized as follows.

1. Theoretically, the order estimation algorithm requires to compute the local order at each point of interest. For some special systems, for example, the piecewise-defined systems, the number of data points needed can be significantly reduced. In this case, by implementing the proposed algorithms, fewer local orders have to be estimated and better models of the system can be obtained.
2. LIC is based on the recursive locally linear estimator introduced in [27]. We can also use its nonrecursive version investigated in [3] to construct LIC and to carry out corresponding convergence analysis.
3. The results in the paper can easily be extended to the case  $1 \leq s_0 \leq M_1$  and  $1 \leq t_0 \leq M_2$  for some known but different  $M_1$  and  $M_2$ . For future research, it is of interest to remove the upper bound assumption on the true system orders.
4. The order estimation algorithms in the paper are non-recursive, i.e., for each  $N \geq 1$  we need to calculate the function  $L_{N+1}(p, q)$ ,  $1 \leq p \leq M$ ,  $1 \leq q \leq M$  and then to find the minimum to serve as the estimate. It is interesting to consider the recursive way to obtain the order estimates.
5. The closed-loop order estimation of NARX systems also deserves further research.

## Appendix

**Lemma 1** Assume that A1)-A7) hold. Then the following estimates take place:

$$\begin{aligned} & \frac{1}{N} \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) \left( f(\varphi_k(M, M)) - f(\mathbf{x}^*(2M)) \right. \\ & \quad \left. - \nabla f(\mathbf{x}^*(2M))^T (\varphi_k(M, M) - \mathbf{x}^*(2M)) \right) \\ &= \frac{1}{2(1-2\delta)} b_N^2 \int_{\mathbb{R}^{2M}} w(x) x^T \frac{\partial^2 f}{\partial \mathbf{x}^*(2M)^2} x dx \cdot f_{\text{IV}}(\mathbf{x}^*(2M)) \\ & \quad + o(b_N^2) + o\left(1/(N^{\frac{1}{2}-\epsilon} b_N^M)\right) \quad \text{a.s.} \end{aligned} \quad (44)$$

$$\begin{aligned} & \frac{1}{N} \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) (\varphi_k(M, M) - \mathbf{x}^*(2M)) \\ & \quad \cdot \left( f(\varphi_k(M, M)) - f(\mathbf{x}^*(2M)) \right. \\ & \quad \left. - \nabla f(\mathbf{x}^*(2M))^T (\varphi_k(M, M) - \mathbf{x}^*(2M)) \right) \\ &= \frac{1}{2(1-3\delta)} b_N^3 \int_{\mathbb{R}^{2M}} w(x) x x^T \frac{\partial^2 f}{\partial \mathbf{x}^*(2M)^2} x dx \cdot f_{\text{IV}}(\mathbf{x}^*(2M)) \\ & \quad + o(b_N^3) + o\left(1/(N^{\frac{1}{2}-\epsilon} b_N^{M-1})\right) \quad \text{a.s.} \end{aligned} \quad (45)$$

$$\sum_{k=1}^{N-1} w_k(\mathbf{x}^*(2M)) \varepsilon_{k+1} = O\left(N^{\frac{1}{2}+M\delta+\epsilon}\right), \quad \text{a.s.} \quad (46)$$

$$\begin{aligned} & \sum_{k=1}^{N-1} w_k(\mathbf{x}^*(2M)) (\varphi_k(M, M) - \mathbf{x}^*(2M)) \varepsilon_{k+1} \\ &= O\left(N^{\frac{1}{2}+(M-1)\delta+\epsilon}\right), \quad \text{a.s.} \end{aligned} \quad (47)$$

for any  $\epsilon > 0$ ,

$$\begin{aligned} & \frac{1}{N} \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) g(\varphi_k(M, M)) \\ & \xrightarrow{N \rightarrow \infty} g(\mathbf{x}^*(2M)) f_{\text{IV}}(\mathbf{x}^*(2M)), \quad \text{a.s.} \end{aligned} \quad (48)$$

for any measurable function  $g(x)$  being continuous at  $\mathbf{x}^*(2M)$  and  $|g(x)| \leq c_1 \|x\|^t + c_2$ ,  $x \in \mathbb{R}^{2M}$  for some positive  $c_1$ ,  $c_2$  and  $t$ , and

$$E w_k^\alpha(\mathbf{x}^*(2M)) = O\left(\frac{1}{b_k^{2M(\alpha-1)}}\right), \quad (49)$$

for any fixed  $\alpha > 0$ .

*Proof:* The results similar to (44)–(48) are in [27] where the exact orders of the NARX system are not required and only their upper bounds are assumed to be available. Here we consider (49). By the definition of  $w_k(\mathbf{x}^*(2M))$ ,

we have

$$\begin{aligned} & Ew_k^\alpha(\mathbf{x}^*(2M)) \\ &= \int_{\mathbb{R}^{2M}} \frac{1}{b_k^{2M\alpha}} w^\alpha \left( \frac{1}{b_k} (x - \mathbf{x}^*(2M)) \right) P_k(dx) \\ &= I_{1,k} + I_{2,k}, \end{aligned} \quad (50)$$

where

$$I_{1,k} = \int_{\mathbb{R}^{2M}} \frac{1}{b_k^{2M\alpha}} w^\alpha \left( \frac{1}{b_k} (x - \mathbf{x}^*(2M)) \right) f_{\text{IV}}(x) dx, \quad (51)$$

$$I_{2,k} = \int_{\mathbb{R}^{2M}} \frac{1}{b_k^{2M\alpha}} w^\alpha \left( \frac{1}{b_k} (x - \mathbf{x}^*(2M)) \right) \cdot (P_k(dx) - P_{\text{IV}}(dx)). \quad (52)$$

By denoting  $s = (x - \mathbf{x}^*(2M))/b_k$  and then changing coordinates in (51), it follows that  $I_{1,k} = O(1/b_k^{2M(\alpha-1)})$ . By the geometrical ergodicity of  $\varphi_k(M, M)$ , it follows that  $I_{2,k} = O(\rho^k/b_k^{2M\alpha})$  for some  $0 < \rho < 1$ . Combining (51) and (52) leads to (49).  $\square$

**Lemma 2** [27] *Assume that A1)-A7) hold. Then*

$$\begin{aligned} & \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(M, M) X_k(M, M)^T \\ &= N \begin{bmatrix} 1 & 0 \\ 0 & N^{-\delta} I \end{bmatrix} A_N \begin{bmatrix} 1 & 0 \\ 0 & N^{-\delta} I \end{bmatrix}, \end{aligned} \quad (53)$$

and

$$A_N \xrightarrow{N \rightarrow \infty} f_{\text{IV}}(\mathbf{x}^*(2M)) \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{1-2\delta} \int_{\mathbb{R}^{2M}} w(x) x x^T dx \end{bmatrix} > 0 \text{ a.s.} \quad (54)$$

where  $I \in \mathbb{R}^{2M \times 2M}$  and  $A_N = \begin{bmatrix} A_N(1,1) & A_N(1,2) \\ A_N(2,1) & A_N(2,2) \end{bmatrix}$  with elements  $A_N(1,1)$ ,  $A_N(1,2)$ ,  $A_N(2,1)$ , and  $A_N(2,2)$  defined as follows:

$$\begin{aligned} A_N(1,1) &= \frac{1}{N} \sum_{k=1}^N w_k(\mathbf{x}^*(2M)), \quad A_N(2,1) = A_N(1,2)^T \\ A_N(1,2) &= \frac{1}{N^{1-\delta}} \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) (\varphi_k(M, M) - \mathbf{x}^*(2M))^T, \\ A_N(2,2) &= \frac{1}{N^{1-2\delta}} \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) (\varphi_k(M, M) - \mathbf{x}^*(2M)) \\ &\quad \cdot (\varphi_k(M, M) - \mathbf{x}^*(2M))^T. \end{aligned}$$

At given  $\mathbf{x}^*(2M)$ , define

$$\begin{aligned} \xi_{k+1} &\triangleq f(\varphi_k(M, M)) - f(\mathbf{x}^*(2M)) \\ &\quad - \nabla f(\mathbf{x}^*(2M))^T (\varphi_k(M, M) - \mathbf{x}^*(2M)) + \varepsilon_{k+1}. \end{aligned} \quad (55)$$

**Lemma 3** *Assume A1) holds. Then the function  $\sigma_{N+1}(p, q)$  defined by (16) with any  $1 \leq p \leq M$  and  $1 \leq q \leq M$  takes the following expression:*

$$\begin{aligned} \sigma_{N+1}(p, q) &= \tilde{\theta}_{N+1}(p, q)^T \\ &\quad \cdot \sum_{i=1}^N w_i(\mathbf{x}^*(2M)) X_i(M, M) X_i(M, M)^T \tilde{\theta}_{N+1}(p, q) \\ &\quad + 2\tilde{\theta}_{N+1}(p, q)^T \sum_{i=1}^N w_i(\mathbf{x}^*(2M)) X_i(M, M) \xi_{i+1} \\ &\quad + \sum_{i=1}^N w_i(\mathbf{x}^*(2M)) \xi_{i+1}^2, \end{aligned} \quad (56)$$

where  $\xi_{k+1}$  is defined in (55). Further, if A1)-A7) hold and if  $p \geq s_0$  and  $q \geq t_0$ , then

$$\begin{aligned} & \sigma_{N+1}(p, q) \\ &= - \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) \xi_{k+1} \right)^T \\ &\quad \cdot \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) X_k(p, q)^T \right)^{-1} \\ &\quad \cdot \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) \xi_{k+1} \right) \\ &\quad + \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) \xi_{k+1}^2. \end{aligned} \quad (57)$$

*Proof:* By the definition of  $\sigma_{N+1}(p, q)$  we have that

$$\begin{aligned} & \sigma_{N+1}(p, q) \\ &= \sum_{i=1}^N w_i(\mathbf{x}^*(2M)) \left( y_{i+1} - \theta_{0, N+1}(p, q) \right. \\ &\quad \left. - \theta_{1, N+1}(p, q)^T (\varphi_i(p, q) - \mathbf{x}^*(p, q)) \right)^2 \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^N w_i(\mathbf{x}^*(2M)) \left( f(\varphi_i(M, M)) + \varepsilon_{i+1} - \theta_{0, N+1}(p, q) \right. \\
&\quad \left. - \theta_{1, N+1}(p, q)^T (\varphi_i(p, q) - \mathbf{x}^*(p, q)) \right)^2 \\
&= \sum_{i=1}^N w_i(\mathbf{x}^*(2M)) \left( f(\mathbf{x}^*(2M)) \right. \\
&\quad + \nabla f(\mathbf{x}^*(2M))^T (\varphi_i(M, M) - \mathbf{x}^*(2M)) \\
&\quad - \theta_{0, N+1}(p, q) - \bar{\theta}_{1, N+1}(p, q)^T (\varphi_i(M, M) - \mathbf{x}^*(2M)) \\
&\quad + f(\varphi_i(M, M)) - f(\mathbf{x}^*(2M)) \\
&\quad \left. - \nabla f(\mathbf{x}^*(2M))^T (\varphi_i(M, M) - \mathbf{x}^*(2M)) + \varepsilon_{i+1} \right)^2 \\
&= \sum_{i=1}^N w_i(\mathbf{x}^*(2M)) \left( \tilde{\theta}_{N+1}(p, q)^T X_i(M, M) + \xi_{i+1} \right)^2, \tag{58}
\end{aligned}$$

where  $\xi_{i+1}$  is defined by (55) and  $\tilde{\theta}_{N+1}(p, q)$  is given by (21). From (58) we then obtain (56).

We now consider the case  $p \geq s_0$  and  $q \geq t_0$ . By Lemma 2, the matrices  $\sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(M, M) X_k(M, M)^T$  are nonsingular for all  $N$  large enough. This ensures  $\sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) X_k(p, q)^T > 0$ , for  $1 \leq p \leq M$ ,  $1 \leq q \leq M$  and all  $N$  large enough by noticing the definition of  $X_k(p, q)$ . Without losing generality, we assume that these matrices are positive definite for  $N \geq 1$ . Then the formula (14) takes place. From (14) and by noticing  $p \geq s_0$  and  $q \geq t_0$  we have

$$\begin{aligned}
&\theta_{N+1}(p, q) \\
&= \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) X_k(p, q)^T \right)^{-1} \\
&\quad \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) y_{k+1} \right) \\
&= \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) X_k(p, q)^T \right)^{-1} \\
&\quad \cdot \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) \left( f(\mathbf{x}^*(2M)) \right. \right. \\
&\quad \left. \left. + \nabla f(\mathbf{x}^*(p, q))^T (\varphi_k(p, q) - \mathbf{x}^*(p, q)) \right. \right. \\
&\quad \left. \left. + f(\varphi_k(M, M)) - f(\mathbf{x}^*(2M)) \right. \right. \\
&\quad \left. \left. - \nabla f(\mathbf{x}^*(p, q))^T (\varphi_k(p, q) - \mathbf{x}^*(p, q)) + \varepsilon_{k+1} \right) \right), \tag{59}
\end{aligned}$$

which, by noticing the definition of  $\xi_{k+1}$  given by (55), implies

$$[f(\mathbf{x}^*(2M)) \quad \nabla f(\mathbf{x}^*(p, q))^T]^T - \theta_{N+1}(p, q)$$

$$\begin{aligned}
&= - \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) X_k(p, q)^T \right)^{-1} \\
&\quad \cdot \left( \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) \xi_{k+1} \right). \tag{60}
\end{aligned}$$

Noticing (21) we find that

$$\begin{aligned}
&\left( [f(\mathbf{x}^*(2M)) \quad \nabla f(\mathbf{x}^*(p, q))^T]^T - \theta_{N+1}(p, q) \right)^T X_k(p, q) \\
&= \tilde{\theta}_{N+1}(p, q)^T X_k(M, M), \tag{61}
\end{aligned}$$

which combining with (56) and (60) yields (57).  $\square$

*Proof of Theorem 1:*

The proof is motivated by [7] for the order estimation of linear systems. Here we present it in detail in a non-linear and nonparametric description. By Lemma 2, we may assume  $\sum_{k=1}^N w_k(\mathbf{x}^*(2M)) X_k(p, q) X_k(p, q)^T > 0$ , for  $1 \leq p \leq M$ ,  $1 \leq q \leq M$  and all  $N \geq 1$ .

Because all  $p$ ,  $q$ ,  $s_0$ , and  $t_0$  are positive integers between 1 and  $M$ , for (22) it suffices to show that any limit point of  $\{(p_N, q_N)\}_{N \geq 1}$  coincides with  $(s_0, t_0)$ . Assume that  $(p', q')$  is a limit point of  $\{(p_N, q_N)\}_{N \geq 1}$ , i.e., there exists a subsequence of  $\{(p_N, q_N)\}_{N \geq 1}$  denoted by  $\{(p_{N_k}, q_{N_k})\}_{k \geq 1}$ , such that  $(p_{N_k}, q_{N_k}) \rightarrow (p', q')$  as  $k \rightarrow \infty$ . Since  $\{(p_N, q_N)\}_{N \geq 1}$  and  $(p', q')$  are nonnegative integers, there exists  $K > 0$  such that

$$(p_{N_k}, q_{N_k}) = (p', q'), \quad \forall k \geq K. \tag{62}$$

For (22) we need to prove the impossibility of the following cases: (i)  $p' < s_0$ ; (ii)  $q' < t_0$ ; (iii)  $p' + q' > s_0 + t_0$ .

We first consider case (i). By Lemmas 1 and 2, it follows that

$$\lambda_{\max}^{(M, M)}(N) \sim N, \quad \lambda_{\min}^{(M, M)}(N) \sim N^{1-2\delta}. \tag{63}$$

Define

$$\begin{aligned}
M_{N_k+1} &\triangleq \tilde{\theta}_{N_k+1}(p', q')^T \\
&\cdot \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) X_i(M, M)^T \tilde{\theta}_{N_k+1}(p', q') \\
&+ 2\tilde{\theta}_{N_k+1}(p', q')^T \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) \xi_{i+1}, \tag{64}
\end{aligned}$$

and

$$\alpha_{N_k+1} \triangleq \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) X_i(M, M)^T \right) \cdot \tilde{\theta}_{N_k+1}(p', q'). \quad (65)$$

From Lemma 3, for all  $k \geq K$  it follows that

$$\sigma_{N_k+1}(p', q') = M_{N_k+1} + \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) \xi_{i+1}^2. \quad (66)$$

By the definition of  $\tilde{\theta}_{N_k+1}(p', q')$  and noticing  $p' < s_0$ , we know that

$$\|\tilde{\theta}_{N_k+1}(p', q')\|^2 \geq \left( \frac{\partial f}{\partial x_{s_0}^*} \right)^2 \quad (67)$$

and the following equality takes place:

$$\begin{aligned} & M_{N_k+1} \\ &= \alpha_{N_k+1}^T \left( \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) X_i(M, M)^T \right)^{-1} \right. \\ & \quad + 2 \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) X_i(M, M)^T \right)^{-1} \\ & \quad \cdot \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) \xi_{i+1} \right) \\ & \quad \cdot \|\tilde{\theta}_{N_k+1}(p', q')\|^{-2} \cdot \tilde{\theta}_{N_k+1}(p', q')^T \\ & \quad \left. \cdot \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) X_i(M, M)^T \right)^{-1} \right) \alpha_{N_k+1}. \end{aligned} \quad (68)$$

For RLLE, we have

$$\begin{aligned} & \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) X_i(M, M)^T \right)^{-1} \\ & \cdot \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) \xi_{i+1} \right) \\ &= \begin{bmatrix} N^{-\delta} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_N(1, 1) & A_N(1, 2) \\ A_N(2, 1) & A_N(2, 2) \end{bmatrix}^{-1} \begin{bmatrix} B_N(1) \\ B_N(2) \end{bmatrix}, \end{aligned} \quad (69)$$

where  $A_N(i, j)$ ,  $i, j = 1, 2$  are defined in Lemma 2 and

$$\begin{aligned} B_N(1) &= \frac{1}{N^{1-\delta}} \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) \\ & \quad \cdot \left( f(\varphi_k(M, M)) - f(\mathbf{x}^*(2M)) \right. \\ & \quad \left. - \nabla f(\mathbf{x}^*(2M))^T (\varphi_k(M, M) - \mathbf{x}^*(2M)) + \varepsilon_{k+1} \right) \\ B_N(2) &= \frac{1}{N^{1-2\delta}} \sum_{k=1}^N w_k(\mathbf{x}^*(2M)) \\ & \quad \cdot (\varphi_k(M, M) - \mathbf{x}^*(2M)) \\ & \quad \cdot \left( f(\varphi_k(M, M)) - f(\mathbf{x}^*(2M)) \right. \\ & \quad \left. - \nabla f(\mathbf{x}^*(2M))^T (\varphi_k(M, M) - \mathbf{x}^*(2M)) + \varepsilon_{k+1} \right). \end{aligned}$$

By Lemmas 1 and 2, it follows that

$$\begin{aligned} & \begin{bmatrix} A_N(1, 1) & A_N(1, 2) \\ A_N(2, 1) & A_N(2, 2) \end{bmatrix} \xrightarrow{N \rightarrow \infty} f_{IV}(\mathbf{x}^*(2M)) \\ & \cdot \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{1-2\delta} \int_{\mathbb{R}^{2M}} w(x) x x^T dx \end{bmatrix} > 0 \text{ a.s.} \end{aligned} \quad (70)$$

and

$$\begin{bmatrix} B_N(1) \\ B_N(2) \end{bmatrix} \xrightarrow{N \rightarrow \infty} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ a.s.} \quad (71)$$

Then by (69), (70), and (71), we have

$$\begin{aligned} & \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) X_i(M, M)^T \right)^{-1} \\ & \cdot \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) \xi_{i+1} \right) = o(1), \end{aligned} \quad (72)$$

and by noticing (67),

$$\begin{aligned} & M_{N_k+1} \\ &= \alpha_{N_k+1}^T \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(M, M) X_i(M, M)^T \right)^{-1} \\ & \quad \cdot \alpha_{N_k+1} \cdot (1 + o(1)) \\ & \geq \frac{1}{2} \lambda_{\min}^{(M, M)}(N_k) \|\tilde{\theta}_{N_k+1}(p', q')\|^2 \\ & \geq \frac{1}{2} \lambda_{\min}^{(M, M)}(N_k) \left( \frac{\partial f}{\partial x_{s_0}^*} \right)^2, \end{aligned} \quad (73)$$

from which and (66) we have

$$\begin{aligned} \sigma_{N_k+1}(p', q') &\geq \frac{1}{2} \lambda_{\min}^{(M,M)}(N_k) \left( \frac{\partial f}{\partial x_{s_0}^*} \right)^2 \\ &\quad + \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) \xi_{i+1}^2. \end{aligned} \quad (74)$$

Now we consider  $\sigma_{N_k+1}(s_0, t_0)$ . By Lemma 3, it holds that

$$\begin{aligned} &\sigma_{N_k+1}(s_0, t_0) \\ &= - \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(s_0, t_0) \xi_{i+1} \right)^T \\ &\quad \cdot \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(s_0, t_0) X_i(s_0, t_0)^T \right)^{-1} \\ &\quad \cdot \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(s_0, t_0) \xi_{i+1} \right) \\ &\quad + \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) \xi_{i+1}^2 \\ &\leq \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) \xi_{i+1}^2. \end{aligned} \quad (75)$$

By the definition of  $\xi_k$  given by (55), we have

$$\begin{aligned} &\sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) \xi_{i+1}^2 \\ &\leq 2 \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) \left( f(\varphi_i(M, M)) - f(\mathbf{x}^*(2M)) \right. \\ &\quad \left. - \nabla f(\mathbf{x}^*(2M))^T (\varphi_i(M, M) - \mathbf{x}^*(2M)) \right)^2 \\ &\quad + 2 \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) \varepsilon_{i+1}^2. \end{aligned} \quad (76)$$

By Lemma 1, it follows that

$$\begin{aligned} &\sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) \left( f(\varphi_i(M, M)) - f(\mathbf{x}^*(2M)) \right. \\ &\quad \left. - \nabla f(\mathbf{x}^*(2M))^T (\varphi_i(M, M) - \mathbf{x}^*(2M)) \right)^2 = O(N_k), \end{aligned} \quad (77)$$

and

$$\sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) E \varepsilon_{i+1}^2 = O(N_k). \quad (78)$$

We now prove that

$$\sum_{k=1}^{\infty} \frac{1}{k} w_k(\mathbf{x}^*(2M)) (\varepsilon_{k+1}^2 - E \varepsilon_{k+1}^2) < \infty, \quad \text{a.s.} \quad (79)$$

For this by noticing A4) it suffices to show that

$$\sum_{k=1}^{\infty} \frac{1}{k^{\frac{2+\eta}{2}}} E w_k^{\frac{2+\eta}{2}}(\mathbf{x}^*(2M)) < \infty. \quad (80)$$

By Lemma 1, we have

$$\begin{aligned} &\sum_{k=1}^{\infty} \frac{1}{k^{\frac{2+\eta}{2}}} E w_k^{\frac{2+\eta}{2}}(\mathbf{x}^*(2M)) \\ &= O \left( \sum_{k=1}^{\infty} \frac{1}{k^{\frac{2+\eta}{2}}} \cdot \frac{1}{b_k^{2M(\frac{2+\eta}{2}-1)}} \right) \\ &= O \left( \sum_{k=1}^{\infty} \frac{1}{k^{\frac{2+\eta}{2}}} \cdot \frac{1}{b_k^{M\eta}} \right) = O \left( \sum_{k=1}^{\infty} \frac{1}{k^{\frac{2+\eta}{2}}} \cdot k^{M\eta\delta} \right) \\ &= O(1), \end{aligned} \quad (81)$$

since  $\delta \in \left(0, \frac{1}{2(2M+1)}\right]$  and hence  $0 < M\delta < 1/2$ . The estimate (81) implies (80) and hence (79).

Combining (75), (77), (78), and (79), we have

$$\sigma_{N_k+1}(s_0, t_0) \leq \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) \xi_{i+1}^2 = O(N_k). \quad (82)$$

By (63), (74), and (75) and paying attention to A7) we have the following

$$\begin{aligned} 0 &\geq L_{N_k+1}(p', q') - L_{N_k+1}(s_0, t_0) \\ &= \sigma_{N_k+1}(p', q') - \sigma_{N_k+1}(s_0, t_0) \\ &\quad + a_{N_k}(p' + q' - s_0 - t_0) \\ &\geq c \lambda_{\min}^{(M,M)}(N_k) + a_{N_k}(p' + q' - s_0 - t_0) \\ &= \lambda_{\min}^{(M,M)}(N_k) \left( c + \frac{a_{N_k}}{\lambda_{\min}^{(M,M)}(N_k)} (p' + q' - s_0 - t_0) \right) \xrightarrow[k \rightarrow \infty]{} \infty, \end{aligned} \quad (83)$$

where  $c > 0$  may depend on sample paths. The contradiction ensures that  $p' \geq s_0$ . Similarly, we can prove that  $q' \geq t_0$ .

Finally, we consider the case (iii):  $p' + q' > s_0 + t_0$ . Since we have established  $p' \geq s_0$  and  $q' \geq t_0$ , by Lemma 3, it

follows that

$$\begin{aligned}
& \sigma_{N_k+1}(p', q') \\
&= - \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(p', q') \xi_{i+1} \right)^T \\
&\quad \cdot \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(p', q') X_i(p', q')^T \right)^{-1} \\
&\quad \cdot \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(p', q') \xi_{i+1} \right) \\
&\quad + \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) \xi_{i+1}^2. \tag{84}
\end{aligned}$$

By Lemmas 1 and 2, we have

$$\begin{aligned}
& \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(p', q') \xi_{i+1} \right)^T \\
&\quad \cdot \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(p', q') X_i(p', q')^T \right)^{-1} \\
&\quad \cdot \left( \sum_{i=1}^{N_k} w_i(\mathbf{x}^*(2M)) X_i(p', q') \xi_{i+1} \right) \\
&= O \left( \begin{bmatrix} N_k^{1-2\delta} & N_k^{1-3\delta} \mathbf{1}^T \\ 0 & N_k^{-\delta} I \end{bmatrix} \begin{bmatrix} N_k & 0 \\ 0 & N_k^{-\delta} I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & N_k^{-\delta} I \end{bmatrix} \right)^{-1} \\
&\quad \cdot \begin{bmatrix} N_k^{1-2\delta} \\ N_k^{1-3\delta} \mathbf{1} \end{bmatrix} \Big) = O(N_k^{1-4\delta}), \tag{85}
\end{aligned}$$

where  $\mathbf{1} \in \mathbb{R}^{p'+q'}$  and  $I \in \mathbb{R}^{(p'+q') \times (p'+q')}$ .

From (75), (84), and (85) we have

$$\begin{aligned}
0 &\geq L_{N_k+1}(p', q') - L_{N_k+1}(s_0, t_0) \\
&= \sigma_{N_k+1}(p', q') - \sigma_{N_k+1}(s_0, t_0) \\
&\quad + a_{N_k}(p' + q' - s_0 - t_0) \\
&\geq -cN_k^{1-4\delta} + a_{N_k}(p' + q' - s_0 - t_0) \\
&= a_{N_k} \left( p' + q' - s_0 - t_0 - c \frac{N_k^{1-4\delta}}{a_{N_k}} \right) \xrightarrow{k \rightarrow \infty} \infty, \tag{86}
\end{aligned}$$

where the limit takes place by noticing A7) and  $p' + q' > s_0 + t_0$ . The obtained contradiction indicates that  $p' = s_0$  and  $q' = t_0$ . This completes the proof.  $\square$

*Proof of Theorem 3:*

We only sketch the proof.

It suffices to show that any limit point of  $\{(\bar{p}_N, \bar{q}_N)\}_{N \geq 1}$  coincides with  $(s_0, t_0)$ . Assume that  $(p', q')$  is a limit point of  $\{(\bar{p}_N, \bar{q}_N)\}_{N \geq 1}$ , i.e., there exists a subsequence of  $\{(\bar{p}_N, \bar{q}_N)\}_{N \geq 1}$  denoted by  $\{(\bar{p}_{N_k}, \bar{q}_{N_k})\}_{k \geq 1}$ , and  $K > 0$  such that

$$(\bar{p}_{N_k}, \bar{q}_{N_k}) = (p', q'), \quad \forall k \geq K. \tag{87}$$

For (31) we need to prove the impossibility of the following cases: (i)  $p' < s_0$ ; (ii)  $q' < t_0$ ; (iii)  $p' + q' > s_0 + t_0$ .

We first consider the case (i). By (63), (74), and (82) we have

$$\begin{aligned}
0 &\geq \bar{L}_{N_k+1}(p', q') - \bar{L}_{N_k+1}(s_0, t_0) \\
&= N_k \log \left( 1 + \frac{\sigma_{N_k+1}(p', q') - \sigma_{N_k+1}(s_0, t_0)}{\sigma_{N_k+1}(s_0, t_0)} \right) \\
&\quad + a_{N_k}(p' + q' - s_0 - t_0) \\
&\geq N_k \log \left( 1 + \frac{c\lambda_{\min}^{(M,M)}(N_k)}{N_k} \right) \\
&\quad + a_{N_k}(p' + q' - s_0 - t_0) \\
&= N_k \cdot \frac{c\lambda_{\min}^{(M,M)}(N_k)}{N_k} + N_k \cdot o \left( \frac{c\lambda_{\min}^{(M,M)}(N_k)}{N_k} \right) \\
&\quad + a_{N_k}(p' + q' - s_0 - t_0) \\
&= \lambda_{\min}^{(M,M)}(N_k) \left( c + o(1) + \frac{a_{N_k}}{\lambda_{\min}^{(M,M)}(N_k)} (p' + q' - s_0 - t_0) \right) \xrightarrow{k \rightarrow \infty} \infty, \tag{88}
\end{aligned}$$

where  $c > 0$  may depend on sample paths. The obtained contradiction ensures that  $p' \geq s_0$ . Similarly, we can prove that  $q' \geq t_0$ .

Finally, we consider the case (iii). From (82), (84), and (85) we have

$$\begin{aligned}
0 &\geq \bar{L}_{N_k+1}(p', q') - \bar{L}_{N_k+1}(s_0, t_0) \\
&= N_k \log \left( 1 + \frac{\sigma_{N_k+1}(p', q') - \sigma_{N_k+1}(s_0, t_0)}{\sigma_{N_k+1}(s_0, t_0)} \right) \\
&\quad + a_{N_k}(p' + q' - s_0 - t_0) \\
&\geq N_k \log \left( 1 - c \frac{N_k^{1-4\delta}}{N_k} \right) + a_{N_k}(p' + q' - s_0 - t_0) \\
&= a_{N_k} \left( p' + q' - s_0 - t_0 - c \frac{N_k^{1-4\delta}}{a_{N_k}} + o \left( \frac{N_k^{1-4\delta}}{a_{N_k}} \right) \right) \\
&\quad \xrightarrow{k \rightarrow \infty} \infty, \tag{89}
\end{aligned}$$

where  $c > 0$  may depend on sample paths. Thus the case  $p' + q' > s_0 + t_0$  is impossible, which in turn guarantees that (31) takes place.  $\square$

Table 1  
Values of  $L_N(q)$  and  $\bar{L}_N(q)$  for  $x_1^*(M)$

$q$	1	2	3	4
$L_N(q), N = 1000$	36.9894	29.9966	<b>26.5952</b>	28.3266
$L_N(q), N = 2000$	86.5703	69.4839	<b>60.8297</b>	64.0119
$L_N(q), N = 3000$	165.2003	132.5021	<b>112.9479</b>	117.2555
$L_N(q), N = 4000$	230.3250	183.6043	<b>155.0399</b>	160.8978
$L_N(q), N = 5000$	282.8153	220.9308	<b>185.3929</b>	192.5084
$\bar{L}_N(q), N = 1000$	$3.5187 \times 10^3$	$3.4255 \times 10^3$	<b><math>3.3769 \times 10^3</math></b>	$3.5512 \times 10^3$
$\bar{L}_N(q), N = 2000$	$9.1149 \times 10^3$	$8.8993 \times 10^3$	<b><math>8.8046 \times 10^3</math></b>	$9.1306 \times 10^3$
$\bar{L}_N(q), N = 3000$	$1.5305 \times 10^4$	$1.4963 \times 10^4$	<b><math>1.4753 \times 10^4</math></b>	$1.5204 \times 10^4$
$\bar{L}_N(q), N = 4000$	$2.2067 \times 10^4$	$2.1511 \times 10^4$	<b><math>2.1141 \times 10^4</math></b>	$2.1713 \times 10^4$
$\bar{L}_N(q), N = 5000$	$2.9322 \times 10^4$	$2.8682 \times 10^4$	<b><math>2.8349 \times 10^4</math></b>	$2.9036 \times 10^4$

Table 2  
Values of  $L_N(q)$  and  $\bar{L}_N(q)$  for  $x_2^*(M)$

$q$	1	2	3	4
$L_N(q)$ at $x_2^*(M), N = 1000$	34.7826	26.0630	<b>25.1868</b>	27.0610
$L_N(q)$ at $x_2^*(M), N = 2000$	87.3153	63.7282	<b>58.7612</b>	61.6941
$L_N(q)$ at $x_2^*(M), N = 3000$	159.3853	<b>105.1250</b>	105.6590	110.3763
$L_N(q)$ at $x_2^*(M), N = 4000$	243.2833	<b>156.3480</b>	156.6344	162.2496
$L_N(q)$ at $x_2^*(M), N = 5000$	303.3111	<b>193.8610</b>	194.4921	201.5572
$\bar{L}_N(q)$ at $x_2^*(M), N = 1000$	$3.6742 \times 10^3$	<b><math>3.4690 \times 10^3</math></b>	$3.5212 \times 10^3$	$3.7036 \times 10^3$
$\bar{L}_N(q)$ at $x_2^*(M), N = 2000$	$9.1842 \times 10^3$	<b><math>8.7373 \times 10^3</math></b>	$8.7498 \times 10^3$	$9.0588 \times 10^3$
$\bar{L}_N(q)$ at $x_2^*(M), N = 3000$	$1.5579 \times 10^4$	<b><math>1.4896 \times 10^4</math></b>	$1.4923 \times 10^4$	$1.5381 \times 10^4$
$\bar{L}_N(q)$ at $x_2^*(M), N = 4000$	$2.2457 \times 10^4$	<b><math>2.1474 \times 10^4</math></b>	$2.1477 \times 10^4$	$2.2049 \times 10^4$
$\bar{L}_N(q)$ at $x_2^*(M), N = 5000$	$2.9154 \times 10^4$	<b><math>2.7808 \times 10^4</math></b>	$2.7874 \times 10^4$	$2.8573 \times 10^4$

Table 3  
Average of QOF in 10 Simulations

	QOF with order estimation	QOF without order estimation
average	70.36%	71.55%
standard deviation	0.0516	0.0423

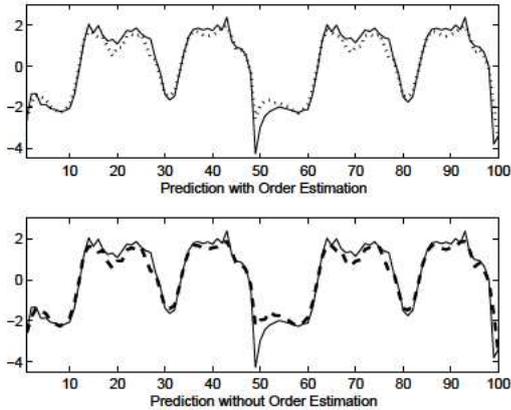


Fig. 1. Predicted and actual outputs.

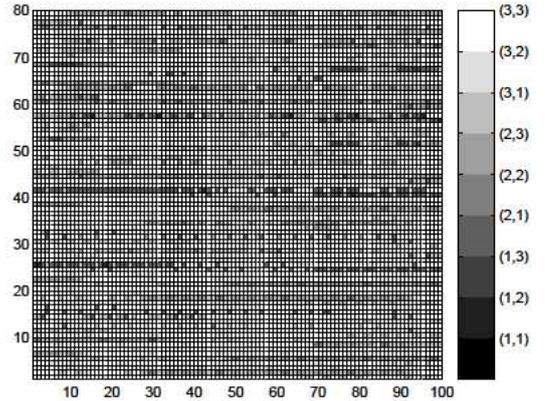


Fig. 2. Heat map of estimated system orders.

## References

- [1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, Vol. 19, pp. 716–723, 1974.
- [2] M. Autinl, M. Biey, and M. Haslerl, "Order of discrete time nonlinear systems determined from input-output signals", *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 296–299, 1992.
- [3] E. W. Bai, "Non-parametric nonlinear system identification: An asymptotical minimum mean squared error estimator," *IEEE Trans. Automatic Control*, Vol. 55, No. 7, pp. 1615–1626, 2010.
- [4] E. W. Bai, K. Li, W. X. Zhao, and W. Y. Xu, "A kernel based forward/backward stepwise approach to local nonlinear non-parametric variable selection", *Automatica*, accepted for publication.
- [5] E. W. Bai, R. Tempo and Y. Liu, "Identification of IIR nonlinear systems without prior structural information," *IEEE Trans. Automatic Control*, Vol. 52, No. 3, pp. 442–453, 2007.
- [6] J. D. Bomberger and D. Seborg, "Determination of model order for NARX models directly from input-output data", *J. Process Control*, Vol. 8, pp. 459–468, 1998.
- [7] H. F. Chen and L. Guo, "Consistent estimation of the order of stochastic control systems", *IEEE Trans. Automatic Control*, Vol. 32, pp. 531–535, 1987.
- [8] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Boston, MA: Birkhäuser, 1991.
- [9] H. F. Chen and W. X. Zhao, "New method of order estimation for ARMA/ARMAX processes", *SIAM J. Control and Optimization*, Vol. 48, pp. 4157–4176, 2010.
- [10] J. Fan and I. Gijbels, *Local Polynomial Modeling and Its Applications*. London, U.K.: Chapman & Hall/CRC, 1996.
- [11] L. S. Huang and P. W. Davidson, "Analysis of variance and F-tests for partial linear models with applications to enviromental health data", *Journal of American Statistical Association*, Vol. 105, No. 491, pp. 991–1004, 2010.
- [12] X. He and H. Asada, "A new method for identifying orders of input-output models for nonlinear dynamic systems", *Proceedings of the American Control Conference*, San Francisco, USA, pp. 2520–2523, 1993.
- [13] X. Hong, R. J. Mitchell, S. Chen, et al, "Model selection approaches for nonlinear system identification:a review", *Int. J. of System Science*, Vol. 39, pp.925–949, 2008.
- [14] M. B. Kennel, R. Brown, and H. Abarbanel, "Determining embedding dimension for phase-space reconstruction using geometrical construction", *Physical Review*, Vol. 45, pp. 3403–3411, 1992.
- [15] K. Mao and S. A. Billings, "Variable selection in nonlinear system modeling", *Mechanical Systems and Signal processing*, Vol. 13, pp. 351–366, 2006.
- [16] B. Ninness and S. J. Henriksen, "Bayesian system identification via Markov chain Monte Carlo techniques," *Automatica*, Vol. 46, No. 1, pp. 40–51, 2010.
- [17] P. Peduzzi, "A stepwise variable selection procedure for nonlinear regression methods", *Biometrics*, Vol. 36, pp. 510–516, 1980.
- [18] S. Ramdania, J. F. Castiesa, F. Boucharab, et al, "Influence of noise on the averaged false neighbors method for analyzing time series", *Physica D*, Vol. 223, pp. 229–241, 2006.
- [19] C. Rhodes and M. Morari, "Determining the model input/output order of nonlinear systems", *AIChE Journal*, Vol. 44, No. 1, pp. 151–163, 1998.
- [20] J. Roll, I. Lind, and L. Ljung, "Connections between optimisation-based regressor selection and analysis of variance", *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 4907–4914, San Diego, USA, 2006.
- [21] J. Roll, A. Nazin and L. Ljung, "Nonlinear system identification via direct weight optimization," *Automatica*, Vol. 41, No. 3, pp. 475–490, 2005.
- [22] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", *Science*, Vol. 290, pp. 2323–2326, 2000.
- [23] T. Schon, A. Wills and B. Ninness, "System identification of nonlinear state-space models," *Automatica*, Vol. 47, No. 1, pp. 39–49, 2011.
- [24] K. Y. You, L. H. Xie, and S. J. Song, "Asymptotically optimal parameter estimation with scheduled measurements", *IEEE Trans. Signal Processing*, Vol. 61, No. 14, pp. 3521–3531, 2013.
- [25] K. Y. You, "Recursive algorithms for parameter estimation with adaptive quantizer", *Automatica*, submitted.
- [26] W. X. Zhao, H. F. Chen, and W. X. Zheng, "Recursive identification for nonlinear ARX systems based on stochastic approximation," *IEEE Trans. Automatic Control*, Vol. 55, No. 6, pp.1287–1299, 2010.
- [27] W. X. Zhao, W. X. Zheng, and E. W. Bai, "A recursive local linear estimator for identification of nonlinear ARX systems: asymptotical convergence and applications," *IEEE Trans. Automatic Control*, Vol. 58, No. 12, pp. 3054–3069, 2013.
- [28] B. Zhou, W. X. Zheng, and G.-R. Duan, "An improved treatment of saturation nonlinearity with its application to control of systems subject to nested saturation", *Automatica*, Vol. 47, No. 2, pp. 306–315, 2011.
- [29] B. Zhou, G.-R. Duan, and Z. L. Lin, "A parametric periodic Lyapunov equation with application in semi-global stabilization of discrete-time periodic systems subject to actuator saturation", *Automatica*, Vol. 47, No. 2, pp. 316–325, 2011.
- [30] H. Zou, "The adaptive Lasso and its oracle properties", *Journal of the American Statistical Association*, Vol. 101, pp. 1418–1429, 2006.