Systems & Control Letters 111 (2018) 1–8

Contents lists available at ScienceDirect

Systems & Control Letters

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

Local variable selection of nonlinear nonparametric systems by first order expansion☆

Wenxiao Zhao [a,b,*], Han-Fu Chen [a], Er-Wei Bai [c], Kang Li [d]

[a] Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[c] Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA
[d] School of Electronics, Electrical Engineering and Computer Science, Queen's University, Belfast, UK

ARTICLE INFO

Article history:
Received 25 March 2017
Received in revised form 25 August 2017
Accepted 8 October 2017

Keywords:
Nonlinear ARX system
Variable selection
Local linear estimator
Strong consistency

ABSTRACT

Local variable selection by first order expansion for nonlinear nonparametric systems is investigated in the paper. By substantially modifying the algorithms developed in our earlier work (Bai et al., 2014), the previous results have been considerably strengthened under much less restrictive conditions. Firstly, the estimates generated by the modified algorithms are shown to have both the set and parameter convergence with probability one, rather than only the set convergence in probability given in our earlier work. Secondly, several technical assumptions, e.g., the lower and upper bounds on the growth of some random sequences, which practically are uncheckable, have been removed. Thirdly, not only the consistency but also the convergence rate of estimates have been established. Besides, a generalization of the proposed algorithms is also introduced.

1. Introduction

Consider the following nonlinear autoregressive system with exogenous inputs (NARX):

$$y_{k+1} = f(y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-q}) + \varepsilon_{k+1}, \tag{1}$$

where u_k and y_k are the system input and output, respectively, ε_k is the system noise, p and q are the upper bounds of system orders, and $f(\cdot)$ is an unknown nonparametric nonlinear function. By nonparametrization it means that no *a prior* information is assumed on the model structure such as $f(\cdot) = f(\cdot, \theta)$ with θ being the unknown parameters. In this case, the value of $f(\cdot)$ is estimated point by point. This is therefore often referred to as *Model on Demand* [1–3]. In this paper, we consider variable selection by the first order expansion of the NARX system (1) under nonparametric setting.

Variable selection plays an important role in many areas including systems and control [4–8], signal processing [9–11], statistics [12,13], and machine learning [14–16], etc., since it provides a way to discover relevant predictive variables and to ensure more reliable predictions. A problem closely related to variable selection is the order estimation, which has been extensively studied for linear systems, for example, the well-known Akaike's Information Criterion (AIC) and its variants [17,18]. Compared with order estimation, variable selection in fact goes further: with the orders p and q being determined, one seeks the contributing variables among $y_k, \ldots, y_{k+1-p}$ and $u_k, \ldots, u_{k+1-q}$. Variable selection has been investigated in the literature, for example, MDS [19], LASSO and its variants [12,13], and more recently the compressive sensing techniques [9]. Other methods include the correlation coefficient method [16], mutual information method [20], Bayesian method [6], and kernel-based method [21] in which the systems are supposed to be linear and often *a priori* probability distribution of the collected data is required. In [5,7,8], variable selection of nonlinear systems is investigated, where the nonlinearity in the system is represented as a linear-in-parameter form, i.e. $y_{k+1} = F(\varphi_k)^T\theta + \varepsilon_{k+1}$, with θ being the unknown parameter, φ_k the regression vector and $F(\cdot)$ the known basis functions. The problem is by no means trivial, but with such a system formulation, ideas from variable selection of linear systems can be adopted. Variable selection problems have also been studied in the machine learning area, c.f., [14,15]. In [14] the collected data are assumed to be iid and some *a prior* knowledge on the sample probability distribution

☆ The research of Wenxiao Zhao and Han-Fu Chen was supported by National Key Research and Development Program of China (2016YFB0901902, 2016YFB0901904), the 973 program of China under Grant No. 2014CB845301 and the NSF of China under Grant Nos. 61573345 and 61227902. The research of Kang Li was supported by the NSF of China under Grant No. 61673256.

* Corresponding author at: Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

E-mail addresses: wxzhao@amss.ac.cn (W. Zhao), hfchen@iss.ac.cn (H.-F. Chen), er-wei-bai@uiowa.edu (E.-W. Bai), k.li@qub.ac.uk (K. Li).

https://doi.org/10.1016/j.sysconle.2017.10.001
0167-6911/

is required. It is clear that for the NARX system (1), the input–output data are not iid and the nonlinear function $f(\cdot)$ cannot always be formulated in a linear-in-parameter form. Besides, in the above papers, the system contributing variables are in a global sense, that is, they are effective over the whole operating domain. This is true for linear systems or nonlinear systems formulated in linear-in-parameter forms. But for nonlinear systems, we need to investigate the problem from a different angle. Consider the following example [4]:

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

$$f(u_k,\cdots,u_{k-3}) = \begin{cases} u_{k-3}, & \text{if } u_k \geq 0 \\ u_{k-3}u_{k-1}, & \text{if } u_k < 0,\ u_{k-1} > 1 \\ u_{k-3}u_{k-2}, & \text{if } u_k < 0,\ u_{k-1} < -1 \\ u_k, & \text{otherwise.} \end{cases} \tag{3}$$

From (3) it is clear that there is no domain where all variables collectively contribute to generating a value of $f(\cdot)$, e.g., variable u_{k-3} is the only contributing variable in domain $u_k \geq 0$. This is quite different from linear systems and for nonlinear systems it is therefore meaningful to select locally the contributing variables.

The first problem is how to define which variables are contributing for the system (1). An intuitive but effective approach is to take the *first order expansion*. Let $\varphi^* \triangleq [y(1), \ldots, y(p), u(1), \ldots, u(q)]^T$ be the given point. Assume $f(\cdot)$ is differentiable at φ^* and set $\nabla f(\varphi^*) \triangleq \left[\frac{\partial f}{\partial y(1)} \cdots \frac{\partial f}{\partial y(p)} \frac{\partial f}{\partial u(1)} \cdots \frac{\partial f}{\partial u(q)}\right]^T$. It is clear that $f(\varphi) \approx f(\varphi^*) + \nabla f(\varphi^*)^T(\varphi - \varphi^*)$ for all φ close to φ^*. Thus, the importance of $y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-q}$ in the neighbourhood of φ^* can be captured by the magnitudes of $|\partial f/\partial y(i)|$, $i = 1, \ldots, p$ and $|\partial f/\partial u(j)|$, $j = 1, \ldots, q$. If y_{k-i} or u_{k-j} does not contribute locally, then $\partial f/\partial y(i) = 0$ or $\partial f/\partial u(j) = 0$. In fact, the idea of the first order expansion is not new and has been used in order detection/variable selection/local embedding of nonlinear systems. See, e.g. [4,10,11], etc. In [10], the idea of first order expansion is introduced, which leads to successive studies on the modelling and identification of nonlinear systems [16,20]. However, in [10,16,20], only algorithms are given, while the theoretical properties are not addressed. In [11], the system under consideration is static. How to extend the idea to dynamic systems remains open. In [4], with the similar idea as in [10] for the variable selection of nonlinear systems, a kernel-based Lasso-type penalized convex optimization algorithm is proposed to locally estimate the contributing variables at fixed points. It is shown that this kind of algorithms has the set convergence in probability, i.e., the Lasso-type algorithms correctly identifying which variables contribute locally and which do not. In comparison to our previous work, the main contributions of this paper are summarized as follows. Firstly, we show that with probability one the estimates generated by a modified version of the algorithms in [4] not only have the set convergence but also the parameter convergence. Secondly, several restrictive assumptions imposed in [4], for example, the boundedness assumption on the conditional number of the data matrix, have been removed. Thirdly, in addition to convergence, the convergence rate of estimates is also given in the paper.

The rest of the paper is organized as follows. The Lasso-type penalized convex optimization algorithm for variable selection of the NARX system (1) is formulated in Section 2, and the main results of the paper are presented in Section 3. A generalization of the algorithm is introduced in Section 4. A simulation study is given in Section 5. Finally, some concluding remarks are addressed in Section 6.

Notations. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be the basic probability space and ω be an element in Ω. Denote the 2-norm of a matrix M by $\|M\|$, and its (i,j)-element by $M(i,j)$. The ith element of a vector m is denoted by $m(i)$. Denote by $\|\nu(\cdot)\|_{\text{var}}$ the total variation norm of a signed measure $\nu(\cdot)$. The invariant probability measure and density of a Markov chain are denoted by $P_{\text{IV}}(\cdot)$ and $f_{\text{IV}}(\cdot)$, respectively, if they exist. Denote by $\nabla f(\cdot)$ the gradient of a function $f(\cdot)$. By $\text{sgn}(x)$ we denote the sign function, i.e., $\text{sgn}(x) = 1$ if $x \geq 0$ and $\text{sgn}(x) = -1$ if $x < 0$.

2. Kernel-based nonparametric variable selection algorithm

Define $\varphi_k \triangleq [y_k, \ldots, y_{k+1-p}, u_k, \ldots, u_{k+1-q}]^T$.

Definition 1 ([4,10,22]). Assume $f(\cdot)$ is differentiable at a given $\varphi^* = [y(1), \ldots, y(p), u(1), \ldots, u(q)]^T$. If some of partial differentials $\partial f/\partial y(i)$, $i = 1, \ldots, p$ or $\partial f/\partial u(j)$, $j = 1, \ldots, q$ are nonzero at φ^*, then y_{k-i} or u_{k-j} in the neighbourhood of φ^* are defined as the contributing variables.

Based on the above definition, the key steps towards variable selection for the NARX system (1) are to find a local linear model of $f(\cdot)$ at φ^* and then to determine which coefficients in the local linear model are zero. We now introduce the algorithms as follows. Based on the measurements $\{u_k, y_{k+1}\}_{k=1}^N$, the local linear estimator (LLE) θ_{N+1} is given by

$$J_{1,N+1}(\theta) \triangleq \sum_{k=1}^{N} w_k(\varphi^*)\Big(y_{k+1} - \theta_0 - \theta_1^T(\varphi_k - \varphi^*)\Big)^2 \tag{4}$$

$$\theta_{N+1} = \begin{bmatrix}\theta_{0,N+1} & \theta_{1,N+1}^T\end{bmatrix}^T \triangleq \underset{\theta_0 \in \mathbb{R}, \theta_1 \in \mathbb{R}^{p+q}}{\text{argmin}} J_{1,N+1}(\theta), \tag{5}$$

with the kernel function

$$w_k(\varphi^*) = \frac{1}{b_k^{p+q}} w\left(\frac{\varphi_k - \varphi^*}{b_k}\right), \tag{6}$$

where $b_k = \frac{1}{k^\delta}$ for some $\delta \in (0, 1)$ and $w(\cdot)$ is a probability density function (pdf).

Remark 1. A widely used $w(\cdot)$ in the kernel is the Gaussian pdf. The estimates derived from (5)–(6) correspond to a local linear model for $f(\cdot)$ at φ^*: $\theta_{0,N+1}$ and $\theta_{1,N+1}$ given by LLE serve as the estimates for $f(\varphi^*)$ and $\nabla f(\varphi^*)$, respectively.

Define

$$X_k \triangleq \begin{bmatrix}1 & (\varphi_k - \varphi^*)^T\end{bmatrix}^T. \tag{7}$$

The estimates based on (4)–(6) can be expressed as

$$\theta_{N+1} = \left(\sum_{k=1}^{N} w_k(\varphi^*) X_k X_k^T\right)^{-1} \left(\sum_{k=1}^{N} w_k(\varphi^*) X_k y_{k+1}\right) \tag{8}$$

given that $\sum_{k=1}^N w_k(\varphi^*) X_k X_k^T$ is nonsingular. Further, define $\theta_{1,N+1} \triangleq [\theta_{1,N+1}(1), \ldots, \theta_{1,N+1}(p+q)]^T$. Then, the penalized convex optimization algorithm for variable selection is given as

$$J_{2,N+1}(\beta) \triangleq \sum_{k=1}^{N} w_k(\varphi^*)(y_{k+1} - \theta_{0,N+1} - \beta^T(\varphi_k - \varphi^*))^2 + \gamma_N \sum_{j=1}^{p+q} \frac{1}{|\bar{\theta}_{1,N+1}(j)|}|\beta_j|, \tag{9}$$

$$\beta_{N+1} = [\beta_{N+1}(1), \cdots, \beta_{N+1}(p+q)]^T \triangleq \underset{\beta \in \mathbb{R}^{p+q}}{\text{argmin}} J_{2,N+1}(\beta), \tag{10}$$

where the kernel $w_k(\varphi^*)$ is introduced in (6), $\{\gamma_N\}_{N \geq 1}$ is a positive sequence tending to infinity, and

$$\bar{\theta}_{1,N+1}(j) \triangleq \theta_{1,N+1}(j) + \frac{1}{N^\varepsilon}\text{sgn}(\theta_{1,N+1}(j)) \tag{11}$$

for $j = 1, \ldots, p+q$ and $\varepsilon \in (0, \delta)$.