2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

New Approach to Variable Selection for Nonparametric Nonlinear Systems*

Xiaotao Ren[1], Wenxiao Zhao[1] and Jinwu Gao[2]

***Abstract*— Let the observation $\{u_k, y_k\}$ be generated by $y_{k+1} = f(y_k, \cdots, y_{k+1-p}, u_k, \cdots, u_{k+1-q}) + \varepsilon_{k+1}$, where (p, q) are the system orders, ε_k is the system noise, and $f(\cdot)$ is an unknown nonlinear function. A new method for variable selection of $f(\cdot)$ at any interested points is introduced. In contrast to most of the existing results, the new method is not based on optimizing a certain criterion, and estimates from the variable selection algorithm given in this paper are easy to update computationally in comparison with the criterion-optimization-based methods when new data arrive. Under reasonable conditions the estimates are proved to converge to the true contributing variables with probability one.**

I. INTRODUCTION

Due to the complexity of practical systems in nature, identification of nonlinear systems has received much attention in recent years [1][2][3][4][5][6]. A classical approach for nonlinear system identification is known as the *model on demand* approach, for which the values of the nonlinear function within the system at interested points are to be estimated[1][3][6]. Other types of nonparametric identification methods include the Gaussian random field approach [2], the reproducing kernel Hilbert spaces (RKHSs) method [7], etc. In this paper, following the idea of the *model on demand* approach, we consider the nonparametric identification of the nonlinear autoregressive system with exogenous inputs (NARX),

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

where u_k, y_k, and ε_k are the system input, output and driven noise, respectively.

Note that almost all of the *model on demand* approaches for nonparametric identification, c.f.,[1][3][6], are in some form of weighted local average. That is, weights are assigned to the observed data according to their distances to the fixed point and only those data that are near to the fixed point take effective roles. This often leads to the problem of *curse of dimensionality* if the system dimension (p, q) is high[8][9]. Thus, for nonparametric identification, it is of importance to have an estimate not only for the dimension of $[y_k, \cdots, y_{k+1-p}, u_k, \cdots, u_{k+1-q}]^T$ but also the contributing variables among $[y_k, \cdots, y_{k+1-p}, u_k, \cdots, u_{k+1-q}]^T$ at the point of interest $\varphi^* \in \mathbb{R}^{p+q}$, since this will lead to a more concise model which suffers less from the dimension problem.

In fact, the variable selection of nonparametric nonlinear systems has been investigated in systems and control literature, for example, the averaging derivative method in [10], the kernel-based Lasso-type convex optimization algorithm in [11][20], the linearisation sub-region division procedure in [12], the additive nonparametric model in [13], the inverse and contour regression approach in [14], etc. There are also concerns from areas of statistics and machine learning, see, e.g., [15][16]. To the authors' knowledge, although the algorithm design in the above literature is different from each other, a common feature lies in that, after collecting a number N of data, these algorithms find a set of possible contributing variables by optimization or computation of a certain objective function, and when data length N changes, one has to re-optimize the objective function to obtain new estimates. This is time-consuming for dynamic systems and online estimation.

The contributions of the paper are summarized as follows. First, based on the idea of local linear approximation and the kernel-based local linear estimator (LLE) [6][8], we propose a new method for variable selection of system (1) at the point of interest, which is not based on optimizing a certain criterion, and estimates from the variable selection algorithm are easy to update computationally in comparison with the criterion-optimization-based methods when new data arrive. Second, we prove that estimates generated from the proposed algorithm correctly identify the contributing variables with probability one. As a byproduct, the strong consistency of LLE as well as the convergence rate are established in the paper by using the Bernsteins inequality for φ-mixing processes [17] and the estimation theorem for double array martingales [18]. To the authors' knowledge, only convergence in probability or in mean square sense of LLE is established in existing literature, see, e.g., [6][8][19].

The rest of the paper is organized as follows. The problem formulation and the variable selection algorithm are given in Section II. The theoretical results are presented in Section III. A numerical example is given in section IV, and some conclusion remarks are provided in section V.

*The research of Xiaotao Ren and Wenxiao Zhao was supported by National Key Research and Development Program of China (2018YFA0703800), CAS Project for Young Scientists in Basic Research under Grant No. YSBR-008, and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27000000.

[1]Xiaotao Ren and Wenxiao Zhao are with the Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China renxiaotao16@mails.ucas.ac.cn, wxzhao@amss.ac.cn

[2]Jinwu Gao is with the Department of Control Science and Engineering, Jilin University, Changchun 130025, China gaojw@jlu.edu.cn

II. VARIABLE SELECTION ALGORITHM FOR NONLINEAR NONPARAMETRIC SYSTEMS

Denote $\varphi_k \triangleq [y_k, \cdots, y_{k+1-p}, u_k, \cdots, u_{k+1-q}]^T \in \mathbb{R}^{p+q}$ and the observed data set by $\{\varphi_k, y_{k+1}\}_{k=1}^N$. Denote by $\varphi^* = [y_1^*, \cdots, y_p^*, u_1^*, \cdots, u_q^*] \in \mathbb{R}^{p+q}$ the point of interest.

Definition 1: ([20][11]) Assume that $f(\cdot)$ is differentiable at the given point φ^*. If some partial differentials $\frac{\partial f}{\partial y_i^*}$ or $\frac{\partial f}{\partial u_j^*}$, $i = 1, \cdots, p, j = 1, \cdots, q$, are nonzero at φ^*, then y_{k+1-i} and u_{k+1-j} are said to be the contributing variables of system (1) at φ^*.

Algorithm 1 Nonparametric Variable Selection Algorithm

Initialization: Observations $\{\varphi_k, y_{k+1}\}_{k=1}^N$, constants $\alpha \in (0,1)$ and $\eta > 1$, a positive sequence $\{b_k\}_{k=1}^N$ tending to zero as $N \to \infty$ and a probability density function (pdf) $w(\cdot) : \mathbb{R}^{p+q} \to \mathbb{R}$.

Step 1. Local linear estimator of $f(\cdot)$ at φ^*

Define the kernel function $w_{k,N}(\varphi^*)$

$$w_{k,N}(\varphi^*) = \frac{1}{b_N^{p+q}} w(\frac{\varphi_k - \varphi^*}{b_N}),\ k = 1, \cdots, N. \quad (2)$$

Compute the local linear estimates of $f(\cdot)$ at φ^*

$$\theta_{N+1} = [\theta_{0,N+1}, \theta_{1,N+1}^T]^T \triangleq \underset{\theta_0 \in \mathbb{R}, \theta_1 \in \mathbb{R}^{p+q}}{\operatorname{argmin}} \sum_{k=1}^N w_{k,N}(\varphi^*)(y_{k+1} - \theta_0 - \theta_1^T(\varphi_k - \varphi^*))^2. \quad (3)$$

Step 2. Variable selection of $f(\cdot)$ at φ^*

Denote $\theta_{1,N+1} \triangleq [\theta_{1,N+1}(1), \cdots, \theta_{1,N+1}(p+q)]^T$, define the decision numbers

$$Q_{j,N+1} \triangleq \frac{|\theta_{1,N+1}(j)| + b_N^\alpha}{b_N^\alpha},\quad j = 1, \cdots, p+q, \quad (4)$$

and compute the estimates for variable selection

$$\tilde{\theta}_{N+1}(j) \triangleq \begin{cases} \theta_{1,N+1}(j), & \text{if } Q_{j,N+1} \geq \eta, \\ 0, & \text{if } Q_{j,N+1} < \eta. \end{cases} \quad (5)$$

Remark 1: Denote the gradient of $f(\cdot)$ at φ^* by $\nabla f(\varphi^*)$, if it exists. The estimates $\theta_{0,N+1} \in \mathbb{R}$ and $\theta_{1,N+1} \in \mathbb{R}^{p+q}$ generated from (3) serve as the estimate of $f(\varphi^*)$ and $\nabla f(\varphi^*)$, respectively. In the following we will prove that $\|\theta_{1,N+1} - \nabla f(\varphi^*)\| = O(b_N)$ almost surely and hence $Q_{j,N+1}$ will diverge to ∞ if the j-th entry of $\nabla f(\varphi^*)$ is nonzero and will converge 1 otherwise, by noting that $b_N \to 0$ as $N \to \infty$ and $\alpha \in (0,1)$. Then $\tilde{\theta}_{1,N+1}(j), j = 1, \cdots, p+q$ by (5) generate consistent estimates for variable selection of $f(\cdot)$ at φ^* as well as values of the nonzero entries in $\nabla f(\varphi^*)$. In contrast to the existing nonparametric variable selection algorithm, see, e.g., the penalized convex optimization algorithm in [20] and [11], algorithm (4)–(5) is not based on optimizing a certain criterion, and estimates from (4)–(5) are easy to update computationally in comparison with the criterion-optimization-based methods when new data arrive.

Remark 2: The estimate θ_{N+1} by (3) can be formulated by

$$\theta_{N+1} = \left(\sum_{k=1}^N w_{k,N}(\varphi^*)\varphi_k\varphi_k^T\right)^{-1}\left(\sum_{k=1}^N w_{k,N}(\varphi^*)\varphi_k y_{k+1}\right) \quad (6)$$

provided that the matrix $\sum_{k=1}^N w_{k,N}(\varphi^*)\varphi_k\varphi_k^T$ is invertible. The local linear estimator (LLE) (2)–(3) is a classical nonparametric identification algorithm widely studied in statistics as well as systems and control. To the authors' knowledge, for LLE only convergence in probability or in mean square sense is reported in literature, see, e.g., [6][8] and references therein. However, such kinds of results are insufficient for almost sure convergence of algorithm (4)–(5) and the strong consistency of LLE will be investigated in this paper.

III. THEORETICAL PROPERTIES OF ALGORITHM

Set $\Theta^* \triangleq \nabla f(\varphi^*) = [\Theta^*(1), \cdots, \Theta^*(p+q)]^T$. Without losing generality, we assume that there are $d(\leq p+q)$ contributing variables of $f(\cdot)$ at φ^* and $\Theta^*(i) \neq 0, i = 1, \cdots, d, \Theta^*(j) = 0, j = d+1, \cdots, p+q$.

We first introduce assumptions to be used in the paper.

A1) The pdf $w(\cdot)$ in (2) satisfies $w(x) = O(\rho^{\|x\|})$ for some $0 < \rho < 1$ as $\|x\| \to \infty$ and the integral $\int_{\mathbb{R}^{p+q}} w(x)xx^T dx > 0$. The bandwidth b_N satisfies $b_N \to 0$ and $Nb_N^{4(p+q+2)} \geq c_1 N^\sigma \to \infty$ for some $c_1 > 0$ and $0 < \sigma < 1$.

A2) The noise sequence $\{\varepsilon_k\}_{k\geq 0}$ is iid with $E\varepsilon_k = 0, E|\varepsilon_k|^2 < \infty$ and with a pdf, denoted by $f_\varepsilon(\cdot)$, which is positive and uniformly continuous on $\mathbb{R}$. ε_{k+1} is independent of φ_k for each $k \geq 0$.

A3) The observation sequence $\{\varphi_k\}_{k\geq 0}$ is ϕ-mixing and stationary with mixing coefficients $\{\phi_k\}_{k\geq 0}$ satisfying $\phi_k \leq c_2\rho^k, k \geq 0$ for some $c_2 > 0$ and $0 < \rho < 1$. Further, $\{\varphi_k\}_{k\geq 0}$ is with a pdf $p(\cdot)$ which is bounded on $\mathbb{R}^{p+q}$ and continuous and positive at φ^*.

A4) The function $f(\cdot)$ in (1) is measurable, continuous at φ^* and $|f(s)| \leq c_3(\|s\|^m + 1), \forall s \in \mathbb{R}^{p+q}$ for some constant $c_3 > 0$ and $m > 0$.

A5) $f(\cdot)$ and $p(\cdot)$ have second order derivatives which are continuous at φ^*.

Denote the Hessian matrices of $f(\cdot)$ and $p(\cdot)$ at φ^* by $\frac{\partial^2 f}{\partial \varphi^{*2}}$ and $\frac{\partial^2 p}{\partial \varphi^{*2}}$, respectively.

Remark 3: For $w(\cdot)$ applied in the kernel function, A1) includes the Gaussian pdf, the uniformly distributed pdf as special cases. In A3), the mixing property of $\{\varphi_k\}_{k\geq 0}$ indicates that φ_k and φ_{k+h} are asymptotically independent as the time interval h increases. If $\{\varphi_k\}_{k\geq 0}$ is an iid sequence, then the mixing property of $\{\varphi_k\}_{k\geq 0}$ follows directly.

There have been many studies on the asymptotical properties of the kernel function $w_{k,N}(\varphi^*)$, most of which are given with convergence in probability or in mean square sense, see [6][8] and references therein. For the almost sure convergence of the kernel, we have the following lemmas and theorem.

Lemma 1: Assume that A1)–A5) hold. Then

$$Ew_{k,N}(\varphi^*) = p(\varphi^*) + O(b_N^2), \tag{7}$$

$$Ew_{k,N}(\varphi^*)(\varphi_k - \varphi^*) = b_N^2 \int_{\mathbb{R}^{p+q}} w(s) s s^T \mathrm{d}s \cdot \nabla p(\varphi^*) + o(b_N^2), \tag{8}$$

$$Ew_{k,N}(\varphi^*)(\varphi_k - \varphi^*)(\varphi_k - \varphi^*)^T = b_N^2 p(\varphi^*) \int_{\mathbb{R}^{p+q}} w(s) s s^T \mathrm{d}s + o(b_N^2), \tag{9}$$

$$Ew_{k,N}(\varphi^*)[f(\varphi_k) - f(\varphi^*) - \nabla f(\varphi^*)^T(\varphi_k - \varphi^*)] = \frac{1}{2} b_N^2 p(\varphi^*) \int_{\mathbb{R}^{p+q}} w(s) s^T \frac{\partial^2 f}{\partial \varphi^{*2}} s \mathrm{d}s + o(b_N^2), \tag{10}$$

$$Ew_{k,N}(\varphi^*)(\varphi_k - \varphi^*)[f(\varphi_k) - f(\varphi^*) - \nabla f(\varphi^*)^T(\varphi_k - \varphi^*)] = \frac{1}{2} b_N^3 p(\varphi^*) \int_{\mathbb{R}^{p+q}} w(s) s s^T \frac{\partial^2 f}{\partial \varphi^{*2}} s \mathrm{d}s + o(b_N^3). \tag{11}$$

and

$$\frac{1}{N}\sum_{k=1}^{N} w_{k,N}(\varphi^*) = p(\varphi^*) + o\left(\frac{1}{N^\kappa b_N^{\kappa(p+q)}}\right) + O(b_N^2) \quad \text{a.s.}, \tag{12}$$

$$\frac{1}{N}\sum_{k=1}^{N} w_{k,N}(\varphi^*)(\varphi_k - \varphi^*) = b_N^2 \int_{\mathbb{R}^{p+q}} w(s) s s^T \mathrm{d}s \cdot \nabla p(\varphi^*) + o\left(\frac{1}{N^\kappa b_N^{\kappa(p+q-1)}}\right) + o(b_N^2) \quad \text{a.s.}, \tag{13}$$

$$\frac{1}{N}\sum_{k=1}^{N} w_{k,N}(\varphi^*)(\varphi_k - \varphi^*)(\varphi_k - \varphi^*)^T = b_N^2 p(\varphi^*) \int_{\mathbb{R}^{p+q}} w(s) s s^T \mathrm{d}s + o\left(\frac{1}{N^\kappa b_N^{\kappa(p+q-2)}}\right) + o(b_N^2) \quad \text{a.s.}, \tag{14}$$

$$\frac{1}{N}\sum_{k=1}^{N} w_{k,N}(\varphi^*)[f(\varphi_k) - f(\varphi^*) - \nabla f^T(\varphi^*)(\varphi_k - \varphi^*)] = \frac{1}{2} b_N^2 p(\varphi^*) \int_{\mathbb{R}^{p+q}} w(s) s^T \frac{\partial^2 f}{\partial \varphi^{*2}} s \mathrm{d}s + o\left(\frac{1}{N^\kappa b_N^{\kappa(p+q)}}\right) + o(b_N^2) \quad \text{a.s.}, \tag{15}$$

$$\frac{1}{N}\sum_{k=1}^{N} w_{k,N}(\varphi^*)(\varphi_k - \varphi^*)[f(\varphi_k) - f(\varphi^*) - \nabla f^T(\varphi^*)(\varphi_k - \varphi^*)] = \frac{1}{2} b_N^3 p(\varphi^*) \int_{\mathbb{R}^{p+q}} w(s) s s^T \frac{\partial^2 f}{\partial \varphi^{*2}} s \mathrm{d}s + o\left(\frac{1}{N^\kappa b_N^{\kappa(p+q-1)}}\right) + o(b_N^3) \quad \text{a.s.} \tag{16}$$

for any $\kappa \in \left(\frac{1}{p+q+3}, \frac{1}{2}\right)$.

Proof: Due to space limitation, here we only prove (15) and the others can be treated similarly. The proofs are based on the Bernstein's inequality for ϕ-mixing processes [17, Lemma 1] and the Borel-Cantelli Lemma. The details are given in Appendix. ∎

Lemma 2: Assume that A1)–A5) hold. Then

$$\frac{1}{N}\sum_{k=1}^{N} w_{k,N}(\varphi^*)\varepsilon_{k+1} = o\left(\frac{1}{N^{\frac{1}{4}-\varepsilon} b_N^{p+q}}\right) \quad \text{a.s.}, \tag{17}$$

$$\frac{1}{N}\sum_{k=1}^{N} (\varphi_k - \varphi^*) w_{k,N}(\varphi^*)\varepsilon_{k+1} = o\left(\frac{1}{N^{\frac{1}{4}-\varepsilon} b_N^{p+q-1}}\right) \quad \text{a.s.} \tag{18}$$

for any $\varepsilon \in (0, \frac{1}{4})$.

Proof: By applying the estimation theorem for double array martingales [18, Theorem 2.9], we can obtain the results. The detailed proofs are given in Appendix. ∎

The strong consistency of LLE as well as the convergence rate can be derived immediately by the following theorem.

Theorem 1: Assume that A1)–A5) hold. Then for LLE, θ_{N+1} converges almost surely to the true value $[f(\varphi^*), \nabla f^T(\varphi^*)]^T$ with the convergence rate

$$\begin{bmatrix} \theta_{0,N+1} \\ \theta_{1,N+1} \end{bmatrix} - \begin{bmatrix} f(\varphi^*) \\ \nabla f(\varphi^*) \end{bmatrix} = \begin{bmatrix} O(b_N^2) \\ O(b_N) \end{bmatrix} \quad \text{a.s.} \tag{19}$$

with b_N specified in A1).

Proof: Set $\theta^0 = [f(\varphi^*), \nabla f^T(\varphi^*)]^T$. Then from (6) we have

$$\theta_{N+1} - \theta^0 = \begin{bmatrix} b_N & 0 \\ 0 & I \end{bmatrix} A_N^{-1} B_N, \tag{20}$$

where

$$A_N = \begin{bmatrix} A_N(1,1) & A_N(1,2) \\ A_N(2,1) & A_N(2,2) \end{bmatrix}, \quad B_N = \begin{bmatrix} B_N(1) \\ B_N(2) \end{bmatrix},$$

$$A_N(1,1) = \frac{1}{N}\sum_{k=1}^{N} w_{k,N}(\varphi^*),$$

$$A_N(1,2) = A_N(2,1)^T = \frac{1}{N b_N}\sum_{k=1}^{N} w_{k,N}(\varphi^*)(\varphi_k - \varphi^*)^T,$$

$$A_N(2,2) = \frac{1}{N b_N^2}\sum_{k=1}^{N} w_{k,N}(\varphi^*)(\varphi_k - \varphi^*)(\varphi_k - \varphi^*)^T,$$

and

$$B_N(1) = \frac{1}{N b_N}\sum_{k=1}^{N} w_{k,N}(\varphi^*)[f(\varphi_k) - f(\varphi^*) - \nabla f^T(\varphi^*)(\varphi_k - \varphi^*) + \varepsilon_{k+1}],$$

$$B_N(2) = \frac{1}{N b_N^2}\sum_{k=1}^{N} w_{k,N}(\varphi^*)(\varphi_k - \varphi^*)[f(\varphi_k) - f(\varphi^*) - \nabla f^T(\varphi^*)(\varphi_k - \varphi^*) + \varepsilon_{k+1}].$$

Noting A1) that $N b_N^{4(p+q+2)} \geq c_1 N^\sigma$, for $0 < \varepsilon < \frac{\sigma}{4}$ we have

$$\frac{1}{N b_N}\sum_{k=1}^{N} w_{k,N}(\varphi^*)\varepsilon_{k+1} = o\left(\frac{1}{N^{\frac{1}{4}-\varepsilon} b_N^{p+q+1}}\right) = o(b_N) \quad \text{a.s.} \tag{21}$$

and

$$\frac{1}{N b_N^2}\sum_{k=1}^{N} w_{k,N}(\varphi^*)(\varphi_k - \varphi^*)\varepsilon_{k+1}$$

$$= o\left(\frac{1}{N^{\frac{1}{4}-\varepsilon} b_N^{p+q+1}}\right) = o(b_N) \quad \text{a.s.} \tag{22}$$

Combining (21)–(22) with Lemma 1, we obtain that

$$A_N \xrightarrow[N\to\infty]{} p(\varphi^*) \begin{bmatrix} 1 & 0 \\ 0 & \int_{\mathbb{R}^{p+q}} w(s) s s^T \mathrm{d}s \end{bmatrix} > 0 \quad \text{a.s.} \tag{23}$$

$$B_N(1) = \frac{1}{2} b_N p(\varphi^*) \int_{\mathbb{R}^{p+q}} w(s) s^T \frac{\partial^2 f}{\partial \varphi^{*2}} s \mathrm{d}s + o(b_N) \quad \text{a.s.}, \tag{24}$$

and

$$B_N(2) = \frac{1}{2} b_N p(\varphi^*) \int_{\mathbb{R}^{p+q}} w(s) s s^T \frac{\partial^2 f}{\partial \varphi^{*2}} s \mathrm{d}s + o(b_N) \quad \text{a.s.} \tag{25}$$

Based on (20)–(25), we have

$$\|\theta_{N+1} - \theta^0\| = \begin{bmatrix} b_N & 0 \\ 0 & I \end{bmatrix} \cdot O(1) \cdot \begin{bmatrix} O(b_N) \\ O(b_N) \end{bmatrix} = \begin{bmatrix} O(b_N^2) \\ O(b_N) \end{bmatrix}. \tag{26}$$

■

For consistency of the variable selection algorithm, we have the following result.

Theorem 2: Assume that A1)–A5) hold. Then for $\{\tilde{\theta}_{N+1}(j), j = 1, \cdots, p+q\}$ generated by (5), there exists an ω-set Ω_0 with $\Pr(\Omega_0) = 1$ such that for any $\omega \in \Omega_0$, there exists an integer $N_0(\omega)$ such that

$$\tilde{\theta}_{N+1}(d+1) = \cdots = \tilde{\theta}_{N+1}(p+q) = 0 \quad N \geq N_0(\omega), \tag{27}$$

and

$$[\tilde{\theta}_{N+1}(1), \cdots, \tilde{\theta}_{N+1}(d)] \xrightarrow[N\to\infty]{} [\theta_1(1), \cdots, \theta_1(d)]. \tag{28}$$

Proof: Since the results in Lemmas 1, 2 and Theorem 1 hold almost surely, there exists an ω-set Ω_0 with $\Pr(\Omega_0) = 1$ such that Lemmas 1, 2 and Theorem 1 hold for any $\omega \in \Omega_0$. In the following, we consider a fixed sample path $\omega \in \Omega_0$.

By Theorem 1 and noting that the parameter α in Algorithm 1 is chosen as $\alpha \in (0, 1)$, we have $\frac{\|\theta_{1,N+1}(j) - \theta_1^*(j)\|}{b_N^\alpha} = o(1)$, $j = 1, \cdots, p+q$. This combining with the definition of $Q_{j,N+1}$ in (4) leads to

$$\lim_{N\to\infty} Q_{j,N+1} = \begin{cases} \infty, & \text{if } \theta_1(j) \neq 0, \\ 1, & \text{if } \theta_1(j) = 0, \end{cases} \tag{29}$$

for $j = 1, \cdots, p+q$.

By noting that $\Theta^* = \nabla f(\varphi^*)$, $\Theta^*(i) \neq 0, i = 1, \cdots, d$, $\Theta^*(j) = 0, j = d+1, \cdots, p+q$ and the value of η being bigger than 1, from (29) we know that there exists a positive integer N_0 such that for all $N > N_0$, $Q_{j,N+1} \geq \eta$, $j = 1, \cdots, d$ and $Q_{j,N+1} < \eta$, $j = d+1, \cdots, p+q$. By the definition of $\tilde{\theta}_{N+1}(j)$ in (5), we have $\tilde{\theta}_{N+1}(j) = 0$, $j = d+1, \cdots, p+q$ for all $N \geq N_0$ and $\tilde{\theta}_{N+1}(j) \xrightarrow[N\to\infty]{} \theta_1(j)$, $j = 1, \cdots, d$. This finishes the proof. ■

IV. SIMULATION

Consider the following nonlinear system ([11]):

$$\begin{aligned} y_{k+1} = {} & \alpha_1 \sin(u_{1,k} u_{2,k}) + \alpha_2 (u_{3,k} - 0.5)^2 \\ & + \alpha_3 u_{4,k} + \alpha_4 u_{5,k} + \alpha_5 u_{6,k} u_{7,k} + \alpha_6 u_{7,k}^2 \\ & + \alpha_7 \cos(u_{6,k} u_{8,k}) + \alpha_8 \exp\{-|u_{8,k}|\} + \varepsilon_{k+1}, \end{aligned} \tag{30}$$

with $\alpha_1 = \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 = 0$, $a_2 = 2$, $a_3 = 1$ and $\alpha_4 = 0$ if $u_{5,k} \leq 0$, while $\alpha_4 = 1$ if $u_{5,k} > 0$.

The test point φ^* is chosen as $\varphi^* = [0\ 0\ 0\ 0\ 0.5\ 0\ 0\ 0]^T$. It can be directly verified that $u_{3,k}$, $u_{4,k}$ and $u_{5,k}$ are contributing variables at φ^* since $\frac{\partial f}{\partial u_{3,k}}\big|_{u_{3,k}=0} = -2$, $\frac{\partial f}{\partial u_{4,k}}\big|_{u_{4,k}=0} = 1$, and $\frac{\partial f}{\partial u_{5,k}}\big|_{u_{5,k}=0} = 1$. In simulation, the input $\{u_k\}_{k\geq 1}$ is chosen as iid variables uniformly distributed over[-1,1], and $\{\varepsilon_k\}_{k\geq 1}$ is a sequence of iid Gaussian random variables with distribution $\mathcal{N}(0, 0.1^2)$. The simulation is performed on a Lenovo desktop with an Intel 1.30GHz i7-CPU.

For Algorithm 1, we choose $b_N = N^{-0.024}$, $\alpha = \frac{1}{2}$, and $\eta = 2$ and perform 100 simulations. Table 1 shows the estimates for values of the contributing variables from one of the simulations. It can be found that as the number of data N increases, the estimates converge to the true values. To further testify the performance of the algorithm, we compare the performance of Algorithm 1 with the Lasso-type optimization-based variable selection algorithm in [20] through their correct rates and computation time. To be specific, Fig.1 shows the rates of the 100 simulations that the contributing variables are correctly identified as a function of the data length N and Table 2 lists the computation time of 100 simulations as data length N increases. It can be found that as N increases, both the correct rates of Algorithm 1 and the algorithm in [20] converge to 100%. On the other hand, as data length N increases, algorithm in [20] requires much more computation time to obtain the estimates.

TABLE I

ESTIMATES FOR VALUES OF CONTRIBUTING VARIABLES

True values	-2.0000	1.0000	1.0000
N=200	-1.8381	0.8485	0.8839
N=1000	-1.9781	0.9755	1.0312
N=2000	-1.9824	1.0301	1.0297
N=3000	-2.0110	0.9937	1.0226

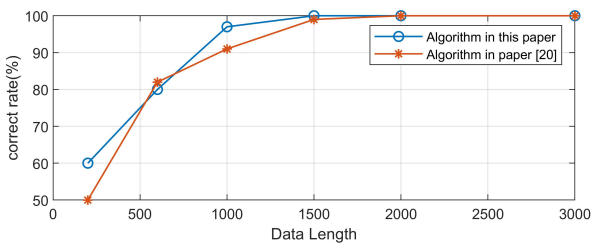

Fig. 1. Rates that Contributing Variables Being Correctly Identified

TABLE II
COMPUTATION TIME OF ALGORITHM 1 AND ALGORITHM IN [20] (UNIT: SECOND)

	Algorithm 1	algorithm in [20]
N=200	0.4178	3.2171
N=1000	1.4382	20.8815
N=2000	3.2841	69.1936
N=3000	5.5861	187.7395

V. CONCLUSION

In this paper, variable selection of the nonparametric nonlinear systems is considered and the strongly consistent estimates for variable selection as well as values of contributing variables are established. Compared with the criterion-optimization-based algorithms, the algorithms proposed in this paper is easy to update when new observed data are available. As a byproduct, the almost sure convergence and the convergence rate of the local linear estimator are established, which, to the authors' knowledge, have been not reported in literature.

For future research, it is of interest to consider the global variable selection for nonlinear systems. It is also of interest to combine the compressive sensing technology with the variable selection of dynamic systems.

APPENDIX

We first introduce the result on the estimation of double array martingales.

Lemma 3: [18, Theorem 2.9] Let $\{\omega_t, \mathscr{F}_t\}_{t\geq 0}$ be an m-dimensional martingale difference sequence satisfying $\|\omega_t\| = o(\varphi(t))$, where $\varphi(x)$ is a positive, deterministic and nondecreasing function that satisfies$\sup_k \varphi(e^{k+1})/\varphi(e^k) < \infty$. Consider the $p \times m$-dimensional double array random matrix sequence $\{f_t(k),\ k = 1, 2, \cdots\},\ t \geq 1$ and suppose that $\{f_t(k),\ k = 1, 2, \cdots,\}$ is $\mathscr{F}_t$-measurable and for some $A > 0, \|f_t(k)\| \leq A < \infty$ a.s. for all $t,\ k$. Then for $h_n = O([\log n]^\alpha)$ with $\alpha > 0$, as $n \to \infty$ it holds

$$\max_{1\leq k\leq h_n} \max_{1\leq i\leq n} \left\| \sum_{j=1}^{i} f_j(k)\omega_{j+1} \right\| = O\left(\max_{1\leq k\leq h_n} \sum_{j=1}^{n} \|f_j(k)\|^2 \right) + o(\varphi(n)\log\log n) \quad \text{a.s.} \tag{31}$$

provided that$\sup_j E(\|\omega_{j+1}\|^2 |\mathscr{F}_j) < \infty$ a.s. Besides, if h_n is only assumed to satisfy $h_n = O(n^\alpha)$ with $\alpha > 0$, then (31) with "$\log\log n$" replaced by "$\log n$" still holds.

A. Proof of Lemma 1

Denote

$$\begin{aligned}&\Delta_{k,N}(\varphi^*)\\ &\triangleq \frac{1}{Nr_N}\Big(w_{k,N}(\varphi^*)[f(\varphi_k) - f(\varphi^*) - \nabla f^T(\varphi^*)(\varphi_k - \varphi^*)]\\ &\quad - E\Big[w_{k,N}(\varphi^*)[f(\varphi_k) - f(\varphi^*) - \nabla f^T(\varphi^*)(\varphi_k - \varphi^*)]\Big]\Big)\end{aligned} \tag{32}$$

where $r_N = \frac{1}{N^\kappa b_N^{\kappa(p+q)}}$ with any fixed $\kappa \in (\frac{1}{p+q+3}, \frac{1}{2})$.

We first show that

$$\sum_{k=1}^{N} \Delta_{k,N}(\varphi^*) \xrightarrow[N\to\infty]{\text{a.s.}} 0. \tag{33}$$

It is direct to verify that

$$E\Delta_{k,N}(\varphi^*) = 0 \tag{34}$$

and by assumption A4)

$$\begin{aligned}&\Big\| w_{k,N}(\varphi^*)\Big[f(\varphi_k) - f(\varphi^*) - \nabla f^T(\varphi^*)(\varphi_k - \varphi^*)\Big]\Big\|\\ &\leq \frac{1}{b_N^{p+q}} w\left(\frac{\varphi_k - \varphi^*}{b_N}\right)\\ &\quad \times \Big[\|f(\varphi_k)\| + \|f(\varphi^*)\| + \|\nabla f(\varphi^*)\| \cdot \|\varphi_k - \varphi^*\|\Big]\\ &\leq \frac{c}{b_N^{p+q}} w\left(\frac{\varphi_k - \varphi^*}{b_N}\right)\\ &\quad \times \left[\|\varphi_k\|^m + \|\varphi^*\|^m + \left\|\frac{\varphi_k - \varphi^*}{b_N}\right\| b_N + 1\right]\\ &\leq \frac{c}{b_N^{p+q}} w\left(\frac{\varphi_k - \varphi^*}{b_N}\right)\\ &\quad \times \left[\left\|\frac{\varphi_k - \varphi^*}{b_N}\right\|^m b_N^m + \|\varphi^*\|^m + \left\|\frac{\varphi_k - \varphi^*}{b_N}\right\| b_N + 1\right]\\ &\leq \frac{c}{b_N^{p+q}}\end{aligned} \tag{35}$$

for some $c > 0$ which may change among different inequalities, where for the last inequality the assumption $w(s) = \rho^{\|s\|}$ as $\|s\| \to \infty$ is applied.

From (35) and (10), we obtain that for some $c > 0$

$$\|\Delta_{k,N}(\varphi^*)\| \leq \frac{c}{Nr_N b_N^{p+q}} + \frac{c}{Nr_N} \leq \frac{c}{Nr_N b_N^{p+q}} \triangleq d(N) \tag{36}$$

and

$$E\|\Delta_{k,N}(\varphi^*)\| \leq \frac{cb_N^2}{Nr_N} \triangleq \delta(N). \tag{37}$$

We now consider $E\|\Delta_{k,N}(\varphi^*)\|^2$. It follows that

$$\begin{aligned}&Ew_{k,N}^2(\varphi^*)[f(\varphi_k) - f(\varphi^*) - \nabla f^T(\varphi^*)(\varphi_k - \varphi^*)]^2\\ &= \int_{\mathbb{R}^{p+q}} \frac{1}{b_N^{2(p+q)}} w^2\left(\frac{x - \varphi^*}{b_N}\right)[f(x) - f(\varphi^*)\\ &\quad - \nabla f^T(\varphi^*)(x - \varphi^*)]^2 p(x)\mathrm{d}x\\ &= \int_{\mathbb{R}^{p+q}} \frac{1}{b_N^{p+q}} w^2(s)[f(b_N s + \varphi^*) - f(\varphi^*) - \nabla f^T(\varphi^*) b_N s]^2\end{aligned}$$

$$\times p(b_N s+\varphi^*)\mathrm{d}x$$
$$\le \frac{c}{b_N^{p+q-2}} \tag{38}$$

where for the last inequality assumption A1) is applied and hence

$$E\left\|\Delta_{k,N}(\varphi^*)\right\|^2 \le \frac{c}{N^2 r_N^2 b_N^{p+q-2}} \triangleq D(N). \tag{39}$$

By the Bernstein's inequality for ϕ-mixing processes [17, Lemma 1], for any fixed $\varepsilon>0$ it holds that

$$\Pr\left[\left\|\sum_{k=1}^{N}\Delta_{k,N}(\varphi^*)\right\|>\varepsilon\right] \le C_1 e^{-\alpha\varepsilon+\alpha^2 N C_2} \tag{40}$$

where C_1, C_2, and α are positive constants depending on N such that

$$C_1 = 2e^{3\sqrt{e}N\phi_{m_N}/m_N}, \tag{41}$$
$$C_2 = 6[D(N)+4\delta(N)d(N)\tilde{\phi}_{m_N}], \tag{42}$$

m_N is any number in $\{1,\cdots,N\}$, $\{\phi_k\}_{k\ge 0}$ is the sequence of mixing coefficients of $\{\varphi_k\}_{k\ge 0}$, $\tilde{\phi}_{m_N}=\sum_{k=1}^{m_N}\phi_k$ and α can be any positive number such that

$$\alpha\cdot m_N\cdot d(N)\le\frac{1}{4}. \tag{43}$$

Set $m_N=[N^{\frac{1}{2}}b_N^{\frac{p+q}{2}}]$. By the definition of r_N (see the definition below (32)), it can be directly verified that $\frac{b_N^2}{m_N r_N}\to 0$, $\frac{Nr_N b_N^{p+q}}{m_N\log N}\to\infty$ and $\frac{r_N}{b_N^2}\to 0$ as $N\to\infty$. By noticing (36)–(39) and assumption A3) that the mixing coefficients $\{\phi_k\}_{k\ge 0}$ satisfy $\phi_k\le c\rho^k, 0<\rho<1$, we have $C_1=O(1)$ and for some constant $c>0$

$$C_2\le c\left[\frac{1}{N^2 r_N^2 b_N^{p+q-2}}+\frac{b_N^2}{Nr_N}\cdot\frac{1}{Nr_N b_N^{p+q}}\right]\le\frac{c}{N^2 r_N^2 b_N^{p+q-2}}. \tag{44}$$

Let $\varepsilon>0$ be such that $\varepsilon c<1/4$ where c is the constant in (36) and set $\alpha=\varepsilon\frac{Nr_N b_N^{p+q}}{m_N}$. It follows that

$$\alpha\cdot m_N\cdot d(N)\le\varepsilon\frac{Nr_N b_N^{p+q}}{m_N}\cdot m_N\cdot\frac{c}{Nr_N b_N^{p+q}}=\varepsilon c\le\frac{1}{4}, \tag{45}$$

from which and by (44),

$$\begin{aligned}
&-\alpha\varepsilon+\alpha^2 N C_2\\
&\le -\varepsilon^2\frac{Nr_N b_N^{p+q}}{m_N}+\varepsilon^2\frac{(Nr_N b_N^{p+q})^2}{m_N^2}\cdot N\cdot\frac{c}{N^2 r_N^2 b_N^{p+q-2}}\\
&\le -\varepsilon^2\frac{Nr_N b_N^{p+q}}{m_N}+\varepsilon^2\frac{Nb_N^{p+q}}{m_N}\cdot\frac{cb_N^2}{m_N}\\
&= -\varepsilon^2\frac{Nr_N b_N^{p+q}}{m_N}(1-\frac{cb_N^2}{m_N r_N})\\
&\le -\frac{1}{2}\varepsilon^2\frac{Nr_N b_N^{p+q}}{m_N},
\end{aligned} \tag{46}$$

where for the last inequality $\frac{b_N^2}{m_N r_N}\to 0$ is applied.

Hence by setting $\beta_N\triangleq\frac{Nr_N b_N^{p+q}}{m_N\log N}$ which goes to $+\infty$ as $N\to\infty$, we have

$$\Pr\left[\left\|\sum_{k=1}^{N}\Delta_{k,N}(\varphi^*)\right\|>\varepsilon\right]\le ce^{-\frac{1}{2}\varepsilon^2\frac{Nr_N b_N^{p+q}}{m_N}}=cN^{-\frac{1}{2}\varepsilon^2\beta_N} \tag{47}$$

and

$$\sum_{N=1}^{\infty}\Pr\left[\left\|\sum_{k=1}^{N}\Delta_{k,N}(\varphi^*)\right\|>\varepsilon\right]\le c\sum_{N=1}^{\infty}N^{-\frac{1}{2}\varepsilon^2\beta_N}<\infty. \tag{48}$$

By using the Borel-Cantelli Lemma, we know that (33) takes place and

$$\begin{aligned}
&\frac{1}{N}\Bigg[\sum_{k=1}^{N}w_{k,N}(\varphi^*)[f(\varphi_k)-f(\varphi^*)-\nabla f^T(\varphi^*)(\varphi_k-\varphi^*)]\\
&\quad-E\sum_{k=1}^{N}w_{k,N}(\varphi^*)[f(\varphi_k)-f(\varphi^*)-\nabla f^T(\varphi^*)(\varphi_k-\varphi^*)]\Bigg]\\
&=o\left(\frac{1}{N^{\kappa}b_N^{\kappa(p+q)}}\right)\quad\text{a.s.}
\end{aligned} \tag{49}$$

Combining (49) and (10), we know that (15) holds. ■

B. Proof of Lemma 2

Define $\varphi(k)\triangleq k^{\frac{1}{2}+\delta'}, k\ge 1$ with fixed $\delta'\in(0,\frac{1}{4})$ and $f_k(N)\triangleq\frac{\varphi_k-\varphi^*}{N^{\frac{1}{4}}b_N}w(\frac{\varphi_k-\varphi^*}{b_N}), k=1,\cdots,N$. It is direct to check that $\{\varphi(k)\}_{k\ge 1}$ is a positive and nondecreasing sequence satisfying $\sup_k\varphi(e^{k+1})/\varphi(e^k)=e^{\frac{1}{2}+\delta'}<\infty$.

From assumption A1), we know that $\|s\|w(s)=O(1)$ and hence

$$\|f_k(N)\|=O\left(\frac{1}{N^{\frac{1}{4}}}\right). \tag{50}$$

By assumption A2) and Chebyshev inequality, we have

$$\Pr\left\{|\varepsilon_k|\ge\varphi(k)\right\}\le\frac{E\varepsilon_k^2}{\varphi(k)^2}\le\frac{\sup_i E\varepsilon_i^2}{k^{1+2\delta'}}$$

and hence

$$\sum_{k=1}^{\infty}\Pr\left\{\|\varepsilon_k\|\ge\varphi(k)\right\}<\infty. \tag{51}$$

By using the Borel-Cantelli Lemma, we have

$$\|\varepsilon_k\|=o(\varphi(k))\quad\text{a.s.} \tag{52}$$

Define the sequence of σ-algebras $\mathscr{F}_k\triangleq\sigma\{u_1,\varepsilon_1,\cdots,u_k,\varepsilon_k\}, k\ge 1$. We know that $\{\varepsilon_k,\mathscr{F}_k\}_{k\ge 1}$ a martingale difference sequence and for any fixed $N\ge 1$, $f_k(N), k=1,\cdots,N$ are $\mathscr{F}_k$-measurable. Then by Lemma 3, we have for any $\varepsilon'>0$,

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{k=1}^{N}(\varphi_k-\varphi^*)w_{k,N}(\varphi^*)\varepsilon_{k+1}\right\|\\
&=\frac{1}{N^{\frac{3}{4}}b_N^{p+q-1}}\left\|\sum_{k=1}^{N}\frac{1}{N^{\frac{1}{4}}}\frac{\varphi_k-\varphi^*}{b_N}w(\frac{\varphi_k-\varphi^*}{b_N})\varepsilon_{k+1}\right\|\\
&=\frac{1}{N^{\frac{3}{4}}b_N^{p+q-1}}\left\|\sum_{k=1}^{N}f_k(N)\varepsilon_{k+1}\right\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{N^{\frac{3}{4}} b_N^{p+q-1}} \max_{1\leq i\leq N} \left\| \sum_{k=1}^{i} f_k(N)\varepsilon_{k+1} \right\| \\
&= \frac{1}{N^{\frac{3}{4}} b_N^{p+q-1}} \left[O\left(\sum_{k=1}^{N} f_k^2(N) \right) + o(N^{\frac{1}{2}+\delta'+\varepsilon'}) \right] \\
&= \frac{1}{N^{\frac{3}{4}} b_N^{p+q-1}} \left[O\left(N^{\frac{1}{2}} \right) + o(N^{\frac{1}{2}+\delta'+\varepsilon'}) \right] \\
&= o\left(\frac{1}{N^{\frac{1}{4}-\varepsilon} b_N^{p+q-1}} \right) \qquad (53)
\end{aligned}
$$

where $0 < \varepsilon \triangleq \delta' + \varepsilon' < \frac{1}{4}$ can be arbitrarily small. This finishes the proof of (18) and (17) can be proved similarly. ■

REFERENCES

[1] J. Roll, A. Nazin, and L. Ljung, "Nonlinear system identification via direct weight optimization," *Automatica*, vol. 41, no. 3, pp. 475–490, 2005.

[2] G. Pillonetto, M. H. Quang, and A. Chiuso, "A new kernel-based approach for nonlinearsystem identification," *IEEE Transactions on Automatic Control*, vol. 56, no. 12, pp. 2825–2840, 2011.

[3] W. Zhao, W. X. Zheng, and E. Bai, "A recursive local linear estimator for identification of nonlinear arx systems: Asymptotical convergence and applications," *IEEE Transactions on Automatic Control*, vol. 58, no. 12, pp. 3054–3069, 2013.

[4] J. Sjöberg, Q. Zhang, L. Ljung, and et al., "Nonlinear black-box modeling in system identification: a unified overview," *Automatica*, vol. 31, no. 12, pp. 1691–1724, 1995.

[5] P. Mattsson, D. Zachariah, and P. Stoica, "Recursive nonlinear-system identification using latent variables," *Automatica*, vol. 93, pp. 343–351, 2018.

[6] J. Fan and I. Gijbels, *Local Polynomial Modeling and Its Applications*. London, U.K. Chapman Hall/CRC, 1996.

[7] G. Pillonetto, "System identification using kernel-based regularization: New insights on stability and consistency issues," *Automatica*, vol. 93, pp. 321–332, 2018.

[8] J. Fan and Q. Yao, *Nonlinear Time Series: Nonparametric and Parametric Methods.* New York: Springer-Verlag, 2005.

[9] E. W. Bai, L. Kang, W. X. Zhao, B. Q. Mu, and W. X. Zheng, "Variable selection in nonlinear non-parametric system identification," *Scientia Sinica: Mathematica*, Vol. 46, No. 10, pp. 1383–1400, 2016.

[10] E. Bai, C. M. Cheng, and W. Zhao, "Variable selection of high-dimensional non-parametric nonlinear systems by derivative averaging to avoid the curse of dimensionality," *Automatica*, vol. 101, pp. 138 – 149, 2019.

[11] E. Bai, K. Li, W. Zhao, and W. Xu, "Kernel based approaches to local nonlinear non-parametric variable selection," *Automatica*, vol. 50, no. 1, pp. 100 – 113, 2014.

[12] K. Z. Mao and S. A. Billings, "Variable selection in non-linear systems modelling," *Mechanical Systems and Signal Processing*, vol. 13, no. 2, pp. 351 – 366, 1999.

[13] B. Mu, W. X. Zheng, and E.-W. Bai, "Variable selection and identification of high-dimensional nonparametric additive nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 5, pp. 2254–2269, 2017.

[14] C. Cheng, E. W. Bai, and Z. Peng, "Consistent variable selection for a nonparametric nonlinear system by inverse and contour regressions," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, pp. 2653–2664, 2019.

[15] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, p. 1157-1182, 2003.

[16] K. Bertin and G. Lecué, "Selection of variables and dimension reduction in high-dimensional non-parametric regression," *Electronic Journal of Statistics*, vol. 2, pp. 1224 – 1241, 2008.

[17] G. Collomb, "Propriétés de convergence presque complète du prédicteur à noyau," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 66, no. 3, pp. 441–460, 1984.

[18] H.-F. Chen and L. Guo, *Identfication and Stochastic Adaptive Control*. Birkhauser, Boston, MA, USA, 1991.

[19] R. David ahd M. P. Wand, "Multivariate locally weighted least squares regression," *The Annals of Statistics*, pp. 1346–1370, 1994.

[20] W. Zhao, H. Chen, E. Bai, and K. Li, "Local variable selection of nonlinear nonparametric systems by first order expansion," *Systems and Control Letters*, vol. 111, pp. 1 – 8, 2018.