Automatica 50 (2014) 100–113

Contents lists available at ScienceDirect

Automatica

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

Kernel based approaches to local nonlinear non-parametric variable selection☆

Er-wei Bai [a,b,1], Kang Li [b], Wenxiao Zhao [c], Weiyu Xu [a]

[a] Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, United States

[b] School of Electronics, Electrical Engineering and Computer Science, Queen's University, Belfast, UK

[c] Academy of Mathematics and Systems Science, China

ARTICLE INFO

Article history:
Received 3 September 2012
Received in revised form
3 August 2013
Accepted 4 October 2013
Available online 31 October 2013

Keywords:
Variable selection
Nonlinear identification
Kernel
Adaptive LASSO
Forward and backward stepwise selections

ABSTRACT

In this paper, we consider the variable selection problem for a nonlinear non-parametric system. Two approaches are proposed, one top-down approach and one bottom-up approach. The top-down algorithm selects a variable by detecting if the corresponding partial derivative is zero or not at the point of interest. The algorithm is shown to have not only the parameter but also the set convergence. This is critical because the variable selection problem is binary, a variable is either selected or not selected. The bottom-up approach is based on the forward/backward stepwise selection which is designed to work if the data length is limited. Both approaches determine the most important variables locally and allow the unknown non-parametric nonlinear system to have different local dimensions at different points of interest. Further, two potential applications along with numerical simulations are provided to illustrate the usefulness of the proposed algorithms.

1. Introduction

This paper concerns identification of a scalar discrete nonlinear non-parametric system

$$y(k) = f(x(k)) + v(k) = f(x_1(k), x_2(k), \ldots, x_p(k)) + v(k), \quad k = 1, 2, \ldots, N \tag{1.1}$$

where $y(\cdot)$ is the system output and $v(\cdot)$ is an iid noise sequence of zero mean and finite variance. The regressor $x(k) = (x_1(k), \ldots, x_p(k))$ consists of possible contributing variables. The structure of the nonlinear function f is unknown. The system (1.1) represents a large class of nonlinear systems including the finite impulse response nonlinear system by letting $x(k) = (u(k-1), \ldots, u(k-p))$. The well known nonlinear auto-regressive moving average system with exogenous inputs (NARX) (Mao & Billings, 2006; Sjoberg et al., 1995) is also a special case of (1.1)

$$y(k) = f(y(k-1), \ldots, y(k-m), u(k-1), \ldots, u(k-m)) + v(k) \tag{1.2}$$

for $x(k) = (y(k-1), \ldots, y(k-m), u(k-1), \ldots, u(k-m))$ and $p = 2m$, where $u(k)$'s are the inputs to the system.

The purpose of nonlinear non-parametric system identification is to estimate the unknown function $f(\cdot)$ based on the available data $\{y(k), x(k)\}_{k=1}^N$. Obviously, one of the difficulties in non-parametric system identification is that the structure of f is unknown. One approach towards this problem is to approximate the unknown system by a possibly nonlinear basis function $\phi_i(x)$'s but linear in the unknown parameters $f(x) = \sum \alpha_i \phi_i(x)$. This approach includes polynomial representation as in Volterra series, splines approximation, linearization of f and others. An advantage of this approach is that it converts approximately a non-parametric problem into a linear in parameters problem. A disadvantage is that much information of the unknown f must be available *a priori* or the number of terms in the representation to reasonably approximate the unknown f could be extremely high which makes identification very sensitive to noise and uncertainty in f.

The other approach is to estimate the value of $f(\cdot)$ point by point, say $f(x^0)$ is of interest, the value of $f(x^0)$ is estimated based on the available data, often referred to as Model on Demand. Almost all the methods in this class are in some form of weighted local averages. The celebrated kernel and local polynomial estimators (Bai, 2010; Fan & Gijbels, 1996) as well as the direct weight optimization (Bai & Liu, 2007; Roll, Nazin, & Ljung, 2005) and the stochastic approximation all belong to this class. A problem with any local average approach with a high dimensional regressor $x(k) = (x_1(k), \ldots, x_p(k)) \in R^p$ is the curse of dimensionality. To illustrate, consider a simple example of an FIR system with

☆ This work was supported in part by grants NSF CNS-1329657, DoE DE-FG52-09NA29364 and NNSF of China, 61104052, 61273193, and 61134013. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Antonio Vicino under the direction of Editor Torsten Söderström.

E-mail addresses: er-wei-bai@uiowa.edu (E.-w. Bai), k.li@qub.ac.uk (K. Li), wxzhao@amss.ac.cn (W. Zhao), weiyu-xu@uiowa.edu (W. Xu).

[1] Tel.: +1 3193355949; fax: +1 3193356028.

0005-1098/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.automatica.2013.10.010

$x(k) = (u(k-1), \ldots, u(k-p))$. Let the input $u(\cdot)$ be iid and uniform in $[-1, 1]$. Suppose the value $f(x^0)$ with $x^0 = (0, 0, \ldots, 0)$ is of interest. Then, there must be adequate measurements $x(k)$'s in the neighborhood of $x^0 = (0, 0, \ldots, 0)$ to reliably estimate $f(x^0)$ due to noise and uncertainty. For simplicity, suppose the neighborhood of $x^0 = (0, 0, \ldots, 0)$ is a ball of radius 0.1 centered at the origin. Thus the probability that $x(k)$ for each k is inside the neighborhood of $x^0 = (0, 0, \ldots, 0)$ is $\frac{\pi^{p/2}0.1^p}{\Gamma(p/2+1)}\frac{1}{2^p}$ where Γ is the Gamma function. In particular, $\Gamma(p/2+1) = p/2!$ for an even integer p. Suppose 10 points in the neighborhood are adequate. Then on average for a large N to have 10 or more points in the neighborhood, the total data length N has to satisfy $N\frac{\pi^{p/2}0.1^p}{\Gamma(p/2+1)}\frac{1}{2^p} \geq 10$ or

$$N \geq 10 \cdot (20)^p \cdot (p/2)!/(\pi^{p/2}) = \begin{cases} 1.24 \cdot 10^8, & p = 6 \\ 4.02 \cdot 10^{13}, & p = 10. \end{cases}$$

This implies that in a practical situation, the required data length N is gigantic even for a modest p. The curse of dimensionality is a fundamental problem for all local average approaches in many fields, not limited to system identification.

Fortunately, for a number of practical applications, systems are sparse in the sense that not all $x_i(k)$'s, $i = 1, 2, \ldots, p$ contribute to the output $y(k)$ or contribute little. If these variables $x_i(k)$'s that do not contribute can be identified and removed, the dimension could be smaller. This is referred to as the variable selection problem in the literature. The variable selection problem has been extensively investigated in the literature in a linear setting including MDS (Cox & Cox, 2000), LASSO, LARS and their variants (Zou, 2006). More recently, compressive sensing techniques are also developed for this purpose. In addition, there exists a large literature in machine learning to deal with the dimension reduction problem, e.g., PCA (Jolliffe, 2002), LLE (Roweis & Saul, 2000), tree structured algorithms and others. Most works in the machine learning literature however project data to a lower dimensional space based on some features ignoring the output variable y, e.g., by nonlinear PCA and LLE. Thus these methods are not ideal for system identification of the system (1.1) where the output error is one of the major concerns. Even algorithms that take output error into consideration, e.g., the partial least squares (Rosipal & Kramer, 2006) developed in a linear setting do not seem to work in a nonlinear non-parametric setting. It should be emphasized that besides obvious nonlinearity, there are some fundamental differences in variable selection between a linear setting and a nonlinear setting. In a linear setting, a variable contributes or not is a global concept while in a nonlinear setting, it can be a local concept. For instance, consider a nonlinear system,

$$y(k) = f(u(k-1), u(k-2), u(k-3), u(k-4)) = \begin{cases} u(k-4) & u(k-1) \geq 0 \\ u(k-4)u(k-2) & u(k-4) < 0, u(k-2) \geq 0 \\ u(k-4)u(k-3) & u(k-4) < 0, u(k-2) < 0 \\ u(k-1) & \text{otherwise.} \end{cases} \quad (1.3)$$

All 4 variables $u(k-1), u(k-2), u(k-3), u(k-4)$ contribute. In other words, $f(\cdot)$ is not sparse globally but sparse locally and the sparsity varies from one location to another. Clearly, in general, algorithms developed for variable selection in a linear setting do not directly apply to a nonlinear setting. In fact, variable selection in a nonlinear non-parametric setting in a system identification content has received only scattered attention in the identification literature. Sjoberg et al. (1995) provides an excellent survey. Many variable selection methods proposed represent the nonlinear system in a linear-in-parameters setting so that the methods developed for linear systems can be readily applied but approximately.

The variable selection problem actually consists of two parts, determining the number of variables that contribute to the output $y(k)$ and once the number is determined, finding the contributing variables among $x_1, \ldots, x_p$. The variable selection problem is closely related to the problem of the order determination. For instance, in a setting of the NARX system (1.2), it is to find the minimum m in (1.2) so that for all $i > m$, $u(k-i)$ and $y(k-i)$ do not contribute. Obviously, in this setting, once m is determined, the contributing variables are $y(k-1), \ldots, y(k-m), u(k-1), \ldots, u(k-m)$. To determine the order m, several ways have been reported. One is based on hypothesis tests (Hong et al., 2008; Peduzzi, 1980). A null hypothesis is specified by assuming the minimum m. Then, the hypothesis is tested. A difficulty of this approach is that the statistics depend on the unknown f. Since f is unknown, some assumptions have to be made. Another form of hypothesis tests are ANOVA which seem to work well under the Gaussian assumption (Bai & Chan, 2008; Lind & Ljung, 2008). The order determination for a nonlinear non-parametric system is investigated in Pillonetto, Quang, and Chiuso (2011) and Su and Yang (2002) by semi-parametric approaches. In Su and Yang (2002), the unknown nonlinear system is modeled by a neural fuzzy network. If the prescribed neural fuzzy network is rich enough and contains the true but unknown nonlinear system, the approach is expected to work if the data length N is large. How to assign a neural fuzzy network without knowing the system is not a trivial question. In Pillonetto et al. (2011), the unknown system is cleverly modeled as a Gaussian random process and thus the unknown quantities are the covariance and the hyperparameters that specify the covariance. A key part is the choice of the covariance that is usually based on prior information of the unknown system. A right or a poor choice of the covariance critically affects the performance. The proposed approaches in this paper try to minimize the use of a prior information on the unknown system and in fact only smoothness is assumed.

The other interesting methods in determining the order m include the Lipschitz numbers (He & Asada, 2003), the false nearest neighbor approach (Bomerger & Seborg, 1998; He & Asada, 2003) and their variants (Cao, 1997; Kennel, Brown, & Abarbanel, 1992). The idea is simple and elegant. For a given $x(k) = (y(k-1), \ldots, y(k-m), u(k-1), \ldots, u(k-m))$, the nearest neighbor $x(j) = (y(j-1), \ldots, y(j-m), u(j-1), \ldots, u(j-m))$ satisfies

$$\|x(k) - x(j)\| \leq \|x(k) - x(i)\|, \quad \forall i \neq k.$$

Then determine if

$$\frac{|y(k) - y(j)|}{\|x(k) - x(j)\|} \leq R \quad (1.4)$$

for some threshold R. If the above inequality holds then the neighbor $x(j)$ is a true neighbor of $x(k)$. Otherwise, the neighbor is a false one. Continue the process for all $k = 1, 2, \ldots, N$ and calculate the percentage of false nearest neighbors. The minimum m that has zero or a small percentage of the false nearest neighbors is considered to be the right dimension n of the system. Note that the idea is global since the dimension m is tested globally.

The variable selection problem is different from the order determination and actually goes further. Even in the setting of the NARX system (1.2) with a known dimension m, the variable selection determines if $y(k-i)$ and $u(k-j)$ $i, j \leq m$ contribute to $y(k)$ or not. If not, these variables could be removed from $f(\cdot)$. Recall in the order determination, once m is determined, all variables $y(k-i)$, $u(k-i)$'s, $i \leq m$, are considered to be contributing variables.

The current paper discusses the variable selection problem. Given a point of interest, the goal is to determine the number n of variables that contribute locally and once n is determined, then to find those n variables. In this paper, we study the variable selection problem in two directions. First, given a point of interest $x^0 = (x_1^0, x_2^0, \ldots, x_p^0)$, the importance of $x_i(k)$ in the neighborhood