Systems & Control Letters 117 (2018) 30–36

Contents lists available at ScienceDirect

Systems & Control Letters

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

Stability orthogonal regression for system identification

Xiaoquan Tang [a], Long Zhang [b,*]

[a] School of Automation, Huazhong University of Science and Technology, Wuhan, 430074, China
[b] School of Electrical and Electronic Engineering, University of Manchester, Manchester, M13 9PL, United Kingdom

ARTICLE INFO

Article history:
Received 24 August 2017
Received in revised form 26 April 2018
Accepted 1 May 2018

Keywords:
Orthogonal forward regression
Stability selection
Stability orthogonal regression
System identification

ABSTRACT

Variable selection methods have been widely used for system identification. However, there is still a major challenge in producing parsimonious models with optimal model structures as popular variable selection methods often produce suboptimal model with redundant model terms. In the paper, stability orthogonal regression (SOR) is proposed to build a more compact model with fewer or no redundant model terms. The main idea of SOR is that multiple intermediate models are produced by orthogonal forward regression (OFR) using sub-sampling technique and then the final model is a combination of these intermediate model terms but does not include infrequently selected terms. The effectiveness of the proposed methods is analyzed in theory and also demonstrated using two numerical examples in comparison with some popular algorithms.

1. Introduction

The main objective of system identification is to establish a mathematical model for a system using system input and output observations. The widely used linear models include auto-regressive with eXogenous input (ARX), auto-regressive moving average with eXogenous input (ARMAX), Box Jenkins and state space models [1]. If the performance of the linear models is not satisfied, the nonlinear ARX (NARX) is an alternative option.

The most popular structure for the NARX model is a sum of nonlinear functions whose parameters are given a priori. The nonlinear functions with pre-set parameters are also referred to as terms in some literatures [2]. However, the pre-fixed values for these nonlinear parameters are not optimal, and therefore their corresponding nonlinear functions are often redundant. The simple option is to use ordinary least square methods to estimate all the coefficients of these nonlinear functions. For these redundant functions, their correct coefficients should be zeros. However, due to the noise effect and correlations between redundant and important functions, the estimated coefficients of redundant functions are often not zeros. In other words, the redundant functions are included into the estimated models, leading to unsatisfactory model performance. Alternatively, regularized least squares algorithms, such as l_1 or l_2 regularization can be used to penalize the coefficients and therefore to produce more compact models. For regularized methods, some additional parameters need to be tuned carefully [3].

Another popular option of building a nonlinear model is to select representative nonlinear functions and then determine their coefficients. The process for selecting nonlinear functions is also referred to as subset or term selection [2]. The predetermined model set may include a huge number of terms and most of terms should not be included into the final model. Therefore, it is important to determine which terms to be included into the final model. The principle of subset selection is to build a parsimonious model with as few redundant model terms as possible [2]. The ideal case is to produce an optimal model without any redundant model term. The orthogonal forward regression (OFR) is one of the most well known subset selection methods. A good review for these existing term selection and their modifications can be found in literatures [1,3–5]. This paper focuses on the subset selection which is a hard problem in the NARX model [6].

The OFR and their modifications have been successfully used in many applications and well studied within system identification community. In most applications, they can produce a parsimonious model. However, a suboptimal solution can be obtained in some applications, in particular when the following conditions happen:

- **Insufficient input–output data and non-persistent excitation**: Most existing methods are based on least square principle and they are asymptotically optimal. The training data length is too short to incorporate all the useful information, which may lead to an inaccurate model. Non-persistent input is another proper problem relating to system input data. Non-persistent excitation can cause regression matrix being ill-conditioning, which may result in poor estimation of the parameters and also poor long term prediction [7].

* Corresponding author.
E-mail address: long.zhang@manchester.ac.uk (L. Zhang).

https://doi.org/10.1016/j.sysconle.2018.05.002
0167-6911/

- **Highly correlated terms**: The adjacent lagged system inputs or outputs could be very similar in their values and therefore their corresponding nonlinear terms are highly correlated, which causes difficulty in selecting the correct terms from the similar alternatives.
- **No optimal criteria**: Most methods have to rely on the information based criteria to determine the model structure. Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and other statistical criteria are popular options [1]. These criteria are simple to use but they may not produce optimal model sizes.
- **Mixed problems**. The above problems can be coupled, which makes it more difficult to build an accurate model, especially for nonlinear systems.

The above reasons can cause sub-optimal model structure with redundant terms. Generally, there are two types of redundant terms. The first type is that the terms are highly correctly with the useful terms and they represent the useful terms when entering the models. The second type is that the terms can generally reduce the model error but tend to approximating the noise. For sparse modeling problems where the number of the useful terms is much smaller than that of the whole candidate terms, the second type, noise terms, could be more serious than the first type in terms of their number in the final model. Since the noise in the input–output data is usually unknown and very hard to estimate with a good accuracy, it is difficult to choose a proper stopping criterion or threshold to control the number of noise terms. Further, system identification usually use random data as the system input. When repeating producing models using different input–output data, the models could be significantly different in terms of the number of redundant or noise terms even if the model stopping criteria or threshold is fixed. In other words, one main difficulty in choosing the model stopping criterion using OFR in practice is to limit the model redundant model terms. If a good criterion or threshold is chosen, the resultant model has fewer or no redundant terms. If not chosen well, the model could have a large number of redundant terms. Another difficulty is that, when repeating the modeling process but just using different input–output data, a number of different models may be generated and it is hard to determine which model should be chosen as the final one.

In this paper, the stability orthogonal regression (SOR) is proposed to build a more parsimonious model by reducing the redundant model terms. A main advantage of SOR is that it can produce an improved model with fewer redundant terms than the original OFR method, and further it may provide the chance to produce an optimal model without any redundant terms. This is achieved by introducing the stability selection scheme into the OFR method. The stability selection was introduced in [8] and mainly aims to produce a stable model with minimal redundant terms. The main principle of stability selection is that it produces multiple intermediate models using sub-sampling techniques. Then the final model consists of the most frequently selected terms in the intermediate models.

This paper is organized by starting to introduce the NARX model and OFR method, then propose the SOR method and analyze its properties in theory, followed by numerical examples.

2. Basics

2.1. NARX model

The linear-in-the-parameters NARX model can be written in the matrix form given as follows:

$$\mathbf{y} = \mathbf{P}\Theta + \Xi \tag{1}$$

where $\mathbf{y} = [y(1), \ldots, y(N)]^T$ is the output vector, $\Theta = [\theta_1, \ldots, \theta_M]^T$ is the weight vector, $\Xi = [e(1), \ldots, e(N)]^T$ is the residual vector. The matrix $\mathbf{P}$ is the whole candidate terms given by $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_M]$, which is an $N \times M$ matrix with $\mathbf{p}_i = [p_i(1), \ldots, p_i(N)]^T$.

The main objective of the subset or term selection is to select the useful terms $\mathbf{P}_m = [\mathbf{p}_{i_1}, \ldots, \mathbf{p}_{i_m}]$ from the whole candidate term pool $\mathbf{P}$, where m denotes the number of selected terms and $[i_1, \ldots, i_m]$ are indexes. Then the coefficients of the selected terms can be written as $\Theta_m = [\boldsymbol{\theta}_{i_1}, \ldots, \boldsymbol{\theta}_{i_m}]$. Using orthogonal least squares (OLS) method, Eq. (1) can be factorized as

$$\mathbf{y} = \mathbf{W}\mathbf{A}\Theta + \Xi \tag{2}$$

here matrix A is an $M \times M$ unit upper triangular matrix. $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M]$ is an $N \times M$ matrix with orthogonal columns that satisfies $\mathbf{W}^T\mathbf{W} = diag[w_i^T w_i]$. For brief, Eq. (2) can be rewritten as

$$\mathbf{y} = \mathbf{W}\mathbf{A}\Theta + \Xi = \mathbf{W}\mathbf{g} + \Xi \tag{3}$$

where $\mathbf{g} = [g_1, g_2, \ldots, g_M]^T = \mathbf{A}\Theta$ is the orthogonal weight vector.

2.2. Orthogonal forward regression (OFR)

OFR is one of most well-known term selection methods and it mainly involves a series of orthogonal composition using OLS method. OFR begins with an empty model without any terms in it and then gradually builds a model by adding one term that gives the largest decrease or increase in the cost function at a time until the model performance is met under some stopping criterion. The first important task in OFR is to choose a cost function for determining which term is included to a resultant model. The error reduction ratio (ERR) is a popular criterion for term selection, and its value is derived from the sum of squares of the model output. More specifically, the sum of squares of the output variables $\mathbf{y}$ is

$$\mathbf{y}^T\mathbf{y} = \sum_{i=1}^{m} g_i^2 \mathbf{w}_i^T \mathbf{w}_i + \Xi^T\Xi. \tag{4}$$

It can be seen that $g_i^2\mathbf{w}_i^T\mathbf{w}_i$ is the contribution of the term $\mathbf{w}_i$ to the sum of squares of the output. The ERR value due to $\mathbf{w}_i$ is defined as [2]

$$[err]_i = g_i^2\mathbf{w}_i^T\mathbf{w}_i/(\mathbf{y}^T\mathbf{y}) = g_i\mathbf{w}_i^T\mathbf{y}/(\mathbf{y}^T\mathbf{y}). \tag{5}$$

The details of the OFR procedure using the ERR criterion are summarized as follows [2,9]:

At the kth step, for $1 \le i \le M$, $i \ne i_1, \ldots, i \ne i_{k-1}$ the following procedure are calculated:

$$\left.\begin{array}{l} if\ \ k = 1 \\ \quad \mathbf{w}_1^{(i)} = \mathbf{p}_i \\ else \\ \quad a_{jk}^{(i)} = \mathbf{w}_j^T\mathbf{p}_i/\mathbf{w}_j^T\mathbf{w}_j,\ 1 \le j < k \\ \quad \mathbf{w}_k^{(i)} = \mathbf{p}_i - \sum_{j=1}^{k-1} a_{jk}^{(i)}\mathbf{w}_j \end{array}\right\} \tag{6}$$

and

$$\left.\begin{array}{l} g_k^{(i)} = (\mathbf{w}_k^{(i)})^T\mathbf{y}/(\mathbf{w}_k^{(i)})^T\mathbf{w}_k^{(i)}, \\ err_k^{(i)} = g_k^{(i)}(\mathbf{w}_k^{(i)})^T\mathbf{y}/\mathbf{y}^T\mathbf{y} \end{array}\right\}. \tag{7}$$

The largest ERR value is calculated using $err_k^{(i_k)} = \max\left\{err_k^{(i)}, i \ne i_1, \ldots, i_{k-1}\right\}$ and the term related to the number i_k is