A New Distance Correlation Metric and Bagging Method for NARX Model Estimation

J. R. A. Solares, H. L. Wei
Department of Automatic Control and System Engineering, Faculty of Engineering, University of Sheffield.

Abstract

System identification is a challenging and interesting engineering problem that has been studied for decades. In particular, the NARMAX methodology has been extensively used with interesting results. Such methodology identifies a deterministic parsimonious model by ranking a set of candidate terms using a linear dependency metric with respect to the output. Other metrics have been used that identify nonlinear dependencies, like the mutual information, but they are hard to interpret. In this work, the distance correlation metric is implemented together with the bagging method. These two implementations enhance the performance of the NARMAX methodology providing interpretability of nonlinear dependencies and uncertainty measures in the model identified. A comparison of the new BOFR-dCor (B*agging* O*rthogonal* F*orward* R*egression* using d*istance* Cor*relation*) algorithm is done with respect to the traditional OFR (O*rthogonal* F*orward* R*egression*) algorithm and the OFR-MI (O*rthogonal* F*orward* R*egression* using M*utual* I*nformation*) algorithm showing interesting results that improve interpretability and uncertainty analysis.

Keywords Bagging; Bootstrap; Distance Correlation; NARX Models; System Identification

1. INTRODUCTION

System identification consists on identifying a mathematical model that describes the behaviour of a system based on recorded input-output data. One of the most popular approaches is the NARMAX (N*onlinear* A*uto*R*egressive* M*oving* A*verage* with *eXogenous* inputs) methodology [1]. This approach ranks a set of candidate terms based on their non-centralised squared correlation with the output data and identifies a deterministic parsimonious model. The non-centralised squared correlation only identifies linear dependencies therefore, new metrics, like the mutual information [2] have been implemented recently to identify nonlinear dependencies. The mutual information is hard to interpret and there is still a need to extend the deterministic notion of the NARMAX methodology to deal with uncertainties. In this work, the distance correlation metric [3] is implemented together with the bagging (b*ootstrap* agg*regat*ing) method [4]. These two implementations enhance the performance of the NARMAX methodology providing interpretability of nonlinear dependencies and uncertainty measures in the model identified.

This work is organised as follows. In section 2 a brief summary of nonlinear system identification, that includes the NARX model and Orthogonal Forward Regression algorithm, is discussed. Section 3 reviews the bootstrap and bagging method. In section 4 the distance correlation metric is described. Our new BOFR-dCor (B*agging* O*rthogonal* F*orward* R*egression* using d*istance* Cor*relation*) algorithm is proposed in section 5. A comparison with the traditional OFR (O*rthogonal* F*orward* R*egression*) algorithm and the OFR-MI (O*rthogonal* F*orward* R*egression* using M*utual* I*nformation*) algorithm is presented in section 6. The work is concluded in section 7 and section 8 mentions the acknowledgements.

2. NONLINEAR SYSTEM IDENTIFICATION

System Identification is an experimental approach where a mathematical equation is identified based on recorded data obtained from the system of study [5]. Since 1940s, the identification of nonlinear systems has been developed considerably. In particular, the NARMAX has been used in a diverse set of scenarios [1]. In general, the process of system identification requires three steps [5]:

a) Model Structure Selection
b) Parameter Estimation
c) Model Validation

2.1. THE NARX MODEL

The NARX (N*onlinear* A*uto*R*egressive* with *eXogenous* inputs) model is a nonlinear recursive difference equation with the following general form: $y(t) = f\left(y(t-1), \dots, y(t-n_y), u(t-1), \dots, u(t-n_u)\right) + e(t)$, where $f(\cdot)$ represents a unknown nonlinear function, $y(t)$, $u(t)$ and $e(t)$ are the output, input and prediction error signals, n_y and n_u are the maximum lags for the output and input signals [2]. If the function $f(\cdot)$ is a polynomial model, then the general form can be expressed in a L*inear*-I*n*-T*he*-P*arameters* (LITP) form: $y(t) = \sum_{m=1}^{M} \theta_m \phi_m(\boldsymbol{\varphi}(t))$, where θ_m are the coefficients of the polynomial, $\phi_m(\boldsymbol{\varphi}(t))$ are the multivariable polynomial terms that are function of the regressor vector $\boldsymbol{\varphi}(t) = [y(t-1), \dots, y(t-n_y), u(t-1), \dots, u(t-n_u)]^T$ of past outputs and inputs, and M is the number of polynomial terms [1], [2].

2.2. ORTHOGONAL FORWARD REGRESSION ALGORITHM

The OFR (O*rthogonal* F*orward* R*egression*) algorithm was developed in the late 1980s by Billings, *et al.* [1]. It is a greedy algorithm that performs parameter estimation of NARMAX models that can be expressed in a LITP form [2].

3. BOOTSTRAP AND BAGGING

The bootstrap is a computer-based method that computes measures of accuracy to statistical estimates. Considering that observations at a given time may depend on previously measured observations, the data set is divided in overlapping blocks of fixed length, B. The first and last observations appear in fewer blocks than the rest; therefore the data set is wrapped around a circle to make all data

points participate equally. Then the blocks are sampled with replacement until a new data set is created with the same length as the original one. This procedure is repeated R times; therefore, R outputs are generated and all of them can be used to predict a numerical value via averaging (for regression problems) or via voting (for classification problems). This procedure is known as bagging (bootstrap aggregating) [4].

4. Distance Correlation

The distance correlation provides a new approach to measure all types of nonlinear dependences between random vectors with finite first moments and arbitrary, not necessarily equal dimension. The distance correlation requires the computation of centred pairwise distance matrices. The procedure is described in [3].

5. The New BOFR-dCor Algorithm

The bagging method and sample distance correlation are combined with the OFR algorithm to produce the BOFR-dCor (B*agging* O*rthogonal* F*orward* R*egression* using d*istance* Co*rrelation*) algorithm. The main steps of the algorithm are the following:

a) Orthogonalise all the regressors in a model so that the correlations between all the terms are removed.
b) Determine significant terms using the distance correlation metric between each regressor and the system output.
c) Estimate the corresponding parameters of the selected terms.
d) Repeat R times.

6. Comparison with OFR and OFR-MI

The following model is taken from [2]:

$$y(t) = -0.5y(t-2) + 0.7y(t-1)u(t-1) + 0.6u^2(t-2) + 0.2y^3(t-1) - 0.7y(t-2)u^2(t-2) + e(t) \quad (3)$$

where the input $u(t)$ follows a uniform distribution $\mathcal{U}(-1,1)$ and the error $e(t)$ follows a normal distribution $\mathcal{N}(0, 0.02^2)$. The parameter values proposed in [2] are used as well in this work. The maximum lags for the input and output are $n_y = n_u = 4$ and the nonlinear degree is $\ell = 3$. The stop criterion for the algorithms is when the error-to-signal ratio (ESR) is less than 0.05. A total of 500 input-output data points were generated. The best model found by the OFR algorithm is: $y(t) = 0.33y(t-4)u^2(t-2) + 0.50u^2(t-2) - 0.64y(t-2) + 0.70y(t-1)u(t-1) + 0.19y^3(t-1)$. The best model found by the OFR-MI algorithm is: $y(t) = -0.49y(t-2) + 0.62u^2(t-2) + 0.62y(t-1)u(t-1) - 0.64y(t-2)u^2(t-2)$. The BOFR-dCor algorithm is applied using a total of 1000 bootstrap samples and a block length of 5. Three top model structures are identified with 775, 150 and 21 votes, respectively. The second most-voted model is $y(t) = -0.50y(t-2) + 0.60u^2(t-2) + 0.71y(t-1)u(t-1) - 0.69y(t-2)u^2(t-2) + 0.20y^3(t-1)$. The method also provides the standard deviation of each parameter. It can be seen that the new algorithm's solution outperforms its predecessors.

7. Conclusions

A new algorithm under the NARMAX methodology that provides interpretability of nonlinear dependencies and uncertainty measures in the model identified is proposed. The algorithm produces results that outperform its predecessors. Extensions and enhancements are being investigated.

Acknowledgements

We acknowledge the support for J. R. Ayala Solares from a University of Sheffield Full Departmental Fee Scholarship and a scholarship from the Mexican National Council of Science and Technology (CONACYT).

References

1. Billings S. Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains: Wiley; 2013.
2. Wei HL, Billings S. Model Structure Selection Using an Integrated Forward Orthogonal Search Algorithm Assisted by Squared Correlation and Mutual Information. Int. J. of Modelling, Identification and Control. 2008; 3(4): p. 341-356.
3. Székely G, Rizzo M, Bakirov N. Measuring and Testing Dependence by Correlation of Distances. The Annals of Statistics. 2007; 35(6): p. 2769-2794.
4. Efron B, Tibshirani R. An Introduction to the Bootstrap: Springer; 1993.
5. Söderström T, Stoica P. System Identification: Prentice Hall; 1989.