The Effect of Nonlinearity Measure on Model Order Selection in Identification of Chemical Processes

Mohammad Rezaee Yazdi
Automation and Instrumentation Department
Petroleum University of Technology
Ahwaz, Iran

Hooshang Jazayerirad
Automation and Instrumentation Department
Petroleum University of Technology
Ahwaz, Iran

ABSTRACT
Most processes in chemical industry reveal nonlinear behavior. A key requirement for many advanced process control is the availability of an accurate dynamic model, with lowest order, for process. This paper handles the problem of model order selection in non-linearity chemical processes identification procedure. In this respect a 'nonlinearity test' method and four model order selection criteria known as the Aikaike Information Criterion (AIC), the Minimum Description Length (MDL), the Exponentially Embedded Family (EEF) and Unmodeled Output Variation (UOV) are considered. The abilities of these criteria in determining the order of the model subjected to different levels of nonlinearity are compared. For this purpose, two chemical processes: a two-tank system and a continuous stirred tank reactor (CSTR) with different levels of nonlinearity are employed. It has been shown that in a system with high level of nonlinearity, the UOV criterion is able to select the lowest order model compared to the other criteria.

Keywords
Chemical process, model order selection, nonlinearity test, system identification, input/output lags, CSTR

1. INTRODUCTION
Numerous information currently exist describing nonlinear behavior of processes such as CSTRs, distillation columns, evaporators and biotechnological processes. Nonlinearity in a chemical process may arise from a variety of sources. It may be due to characteristics of the process such as temperature dependence of reaction rates. It may also result from process limitations such as valve limits, leading to input saturation (i.e., flow rate manipulation) or from physical constraints on output variables (e.g., mole fractions of chemical species) [1].
Process nonlinearity is one of the most relevant factors in system identification problems that plays significant role in identifying a model for a system from measured input/output data, without necessarily knowing anything about the physical laws controlling the system. A model, in general, is any qualitative description of a system, taking

into account the most important factors that affect the system. Many models are derived from fundamental physical laws. A model will never be complete, but good approximations may be possible [2].

Model order selection is also a basic problem in system identification. The model order selection criterion should be able to detect under modeling (too simple model) as well as over modeling (too complex model) [3]. It is proved that, in non-linear systems, the model order selection and its criteria are affected by the level of nonlinearity.
In this paper, at first we test the level of nonlinearity of several differential equations and the system dynamical equations of two application examples. Then we apply the above criteria to determine the model order of these systems.

1. NONLINEARITY TEST
One of the first steps in system identification is to consider a linear model for the system. But in many cases, which level of nonlinearity is high, a linear model cannot be a good approximation of the system. To avoid an unnecessary time-waste, it is desirable to determine level of nonlinearity by employing simple tests. Nonlinearity tests are well known in the field of system identification [4, 5]. Among the tests listed, in these papers, one of them, known as 'superposition checks' can be described as follows.
Two main characteristics of linear systems which are superposition

$$Y(t) = f[\varphi_1(t) + \varphi_2(t)] = f[\varphi_1(t)] + f[\varphi_2(t)] \quad (1)$$

and homogeneity

$$Y(t) = f[\alpha\varphi(t)] = \alpha f[\varphi(t)] \quad (2)$$

These characteristics are not satisfied in nonlinear systems; at least not over their entire operating range. If no disturbance affects the system, these conditions are often quite easy to check. It is necessary to assume that the system is at rest for (1) and (2). If the system is stable one can also wait until the transient has disappeared before checking the conditions.

For testing a system, one can try the following procedure:

1-Apply a zero input signal and wait for the steady state to occur. Investigate if there is a DC offset (D).

2- Apply two different input signals, u_1 and u_2, obeying

$$u_2(t) = cu_1(t) \quad (3)$$

3- Calculate the below ratio

$$r(t) = \frac{y_2(t) - D}{y_1(t) - D} \quad (4)$$

4- Calculate the 'nonlinearity index', '' v'', for the system

$$v = \max \left| \frac{r(t) - c}{c} \right| \quad (5)$$

where '' v'' is between 0 and 1. So the closer "v" is to one, the higher the nonlinearity level of the system will be.

2. MODEL ORDER SELECTION

The modeling problem being considered is the prediction of a data value from a set of realizations of measured data. Model selection criteria are relative. No absolute measure of model fit exists, and if we do not include the 'correct' model in the set that we consider, then we will certainly make the wrong choice. It is generally implicitly assumed that any criterion that will select the correct model over all others given a large amount of data will select the 'most appropriate' model from a set of wrong but approximately correct models. Exactly what the most appropriate model is, depends on the situation, but in many cases we are interested in minimizing the prediction errors when new data are presented [6].

2.1 The Aikaike Information Criterion (AIC)

According to Akaike's theory, the most accurate model has the smallest AIC:

$$\boldsymbol{AIC} = \log\mathbf{V} + \frac{2\mathrm{d}}{\mathrm{N}} \quad (6)$$

where $(\boldsymbol{V})$ is loss function, d is the number of estimated parameters, and N is the number of values in the estimation data set.

The loss function $(\boldsymbol{V})$ is defined by the following

equation:

$$\boldsymbol{V} = det \left\{ \frac{1}{N} \sum_{1}^{N} \in (t, \theta_N) \ \in (t, \theta_N)^T \right\} \quad (7)$$

2.2 The Minimum Description Length Criterion (MDL)

Rissanen has developed his MDL criterion using the framework of coding and estimation to get the shortest possible description of the observed data [7]:

$$\boldsymbol{MDL} = \boldsymbol{V} \left\{ 1 + \left(\frac{\mathrm{d} \log{(\mathrm{N})}}{N} \right) \right\} \quad (8)$$

2.3 The Exponentially Embedded Family (EEF)

It allows the user to embed two or more probability distribution functions (pdfs) into a family of (pdfs) that are indexed by one or more parameters. This embedded family has the form of an exponential family and chooses the model order that maximizes [8]:

$$\boldsymbol{EEF} = \{V - d(\log(\frac{V}{d}) + 1)\} u(\frac{V}{d} - 1) \quad (9)$$

where u(x) is the unit step function.

2.4 The Unmodeled Variation (UOV)

The parameter (r_k^y) is defined as the percentage of variation in the output measurement that is not explained by the model:

$$\left[r_k^y = \frac{V_k}{\frac{1}{N} \sum_{t=1...n} (y(t) - \overline{y})^T (y(t) - \overline{y})} \right] \times 100 \quad (10)$$

where $\bar{y}$ is the average of the sampled output:

$$\bar{y} = \frac{1}{N} \sum_{t=1...N} y(t) \quad (11)$$

The unmodeled output variation is compared to the percentage of relative variation due to noise (r^w) when there is no movement in the inputs $(t_{ss} = 1 ... M)$:

$$r^{w} = \frac{\frac{1}{M}\sum_{t_{ss}=1\ldots M}\left(y(t)-\frac{1}{M}\sum_{t_{ss}=1\ldots M} y(t)\right)^{T}\left(y(t)-\frac{1}{M}\sum_{t_{ss}=1\ldots M} y(t)\right)}{\frac{1}{N}\sum_{t=1..n}\left(y(t)-\overline{y}\right)^{T}\left(y(t)-\overline{y}\right)} \tag{12}$$

The model order (k) is selected such that (r_k^y) is greater than (r^w):

$$r_k^y > r^w \tag{13}$$

and

$$r_{k-1}^y - r_k^y > P\left(r_k^y - r_{k+1}^y\right), P > 1 \tag{14}$$

The model order is selected at the elbow in the plot of unmodeled output variation (r_k^y) by selecting an appropriate value of *P* [7].

3. EXAMPLES

The criteria that are expressed in the previous section are used to determine the model order of the following three differential equations:

1) $\frac{d}{dt}y(t) = 4y(t) + 0.5u(t)$

2) $\frac{d}{dt}y(t) = y^2(t) - y(t).u(t)$

3) $\frac{d}{dt}y(t) = y^2(t).u(t) - u(t) + \sqrt{y(t)+5}$

Equations 1 to 3 have nonlinearity levels of: low $(v = 0.05)$, medium $(v = 0.40)$ and high $(v = 0.81)$ respectively. The model order for each system is selected by sing a Nonlinear Auto Regressive with exogenous variable (NARX) model structure and a uniformly distributed random signal applied to its input. The parameters associated with the mode order, $[n_a\ n_b\ n_k]$, where n_a is output lags, n_b is input lags and n_k is delay, and the percentage of the output variation that is explained by the model are displayed in Table 1:

Table 1. The selected model order by each criterion and their percentages of the output variation

noiretirC	ytiraenilnon fo leveL		
	woL	muideM	hgiH
CIA	[2 1 1] 100%	[2 2 1] 87.25%	[2 3 1] 73.31%
LDM	[2 1 1] 100%	[2 2 1] 87.25%	[2 3 1] 73.31%
FEE	[2 1 1] 100%	[2 3 1] 86.63%	[3 2 2] 71.74%
VOU	[2 1 1] 100%	[2 1 1] 88.05%	[2 2 1] 75.13%

As can be seen in the Table 1, all criteria resulted in the same model order for the system with low level of nonlinearity. However, by increasing the level of non-linearity, the UOV criterion has shown the best performance in comparison to the other criteria. In a highly nonlinear system, MDL and AIC criteria results in lower performance indices than the UOV criterion and the EEF criterion has the least performance index.

4. CASE STUDY

Two nonlinear chemical processes are considered as the application examples in this section. The first example is a *Two-Tank* system with medium nonlinearity level $(v = 0.42)$. The model of the system is given by

$$\frac{d}{dt}x_1(t) = \frac{1}{A_1}\left[\left(ku(t) - a_1\sqrt{2gx_1(t)}\ \right)\right]$$

$$\frac{d}{dt}x_2(t) = \frac{1}{A_2}\left[\left(a_1\sqrt{2gx_1(t)}\ \right) - a_2\sqrt{2gx_2(t)}\ \right)\right]$$

where The input $u(t)$ is the voltage (V) applied to a pump, which generates an inflow to the first tank, $x_1(t)$ and $x_2(t)$ denote the water level in the first and the second tank, respectively, A_i (m^2) is the cross-sectional area of tank i , 'a_i' is the cross-sectional area of the outlet hole and 'g' is the gravity constant [9].

The second system is a *Non-Adiabatic (CSTR)* with high nonlinearity level$(v = 0.86)$. Its dynamical model is given by:

$$\frac{d}{dt}C_A(t)=\frac{F}{V}(C_{Ai}(t)-C_A(t)-(K_0 e^{-E/RT(t)})C_A(t))$$

$$\frac{d}{dt}T(t)=\frac{F}{V}(T_i(t)-T(t))-(\frac{H}{C_{P\rho}})(K_0 e^{-E/RT(t)})-(\frac{UA}{\rho C_P V})(T(t)-T_j(t))$$

Where the manipulated input is $T_j(t)$ and controllable output is $C_A(t)$. For further information on this model and its derivation, see [10].

These two systems are simulated using MATLAB [9]. To generate output data, a uniformly distributed random input is applied to each of these systems. The input signals of both systems are showed in figures 1 and 2.

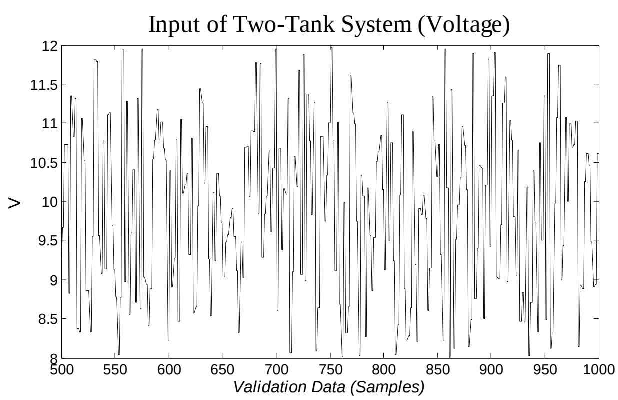

Fig 1: The input signal(uniformly distributed random) which is used for excitation of 'Two-Tank' system

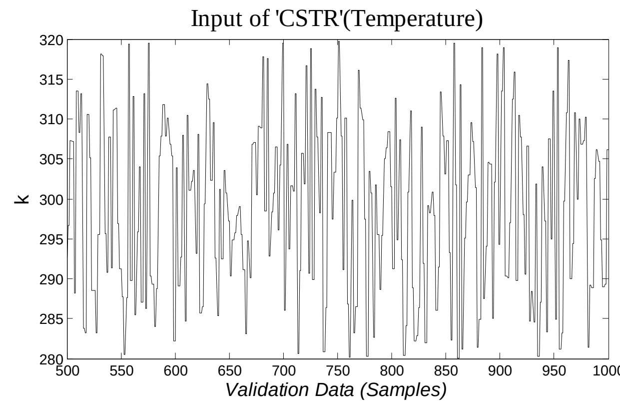

Fig 2: The input signal(uniformly distributed random) which is used for excitation of 'CSTR'system

The data set, that are obtained by exciting each model (named as the Z set), are divided into the estimation data set (Z_e) and the validation data set (Z_v). First, each system is identified using Z_e, then the resulting estimated model is validated by Z_v. Figure 3 shows the validation and the UOV-based model output for the two-tank system. Similar results are shown in Figure 4 for the CSTR.

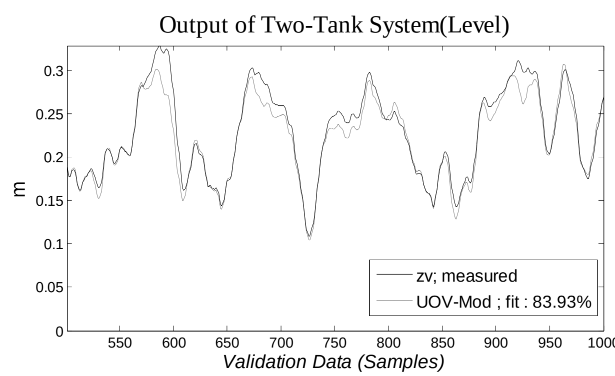

Fig 3: The measured (validation) data of the 'Two-Tank' system and the output signal of estimated model using the UOV criterion

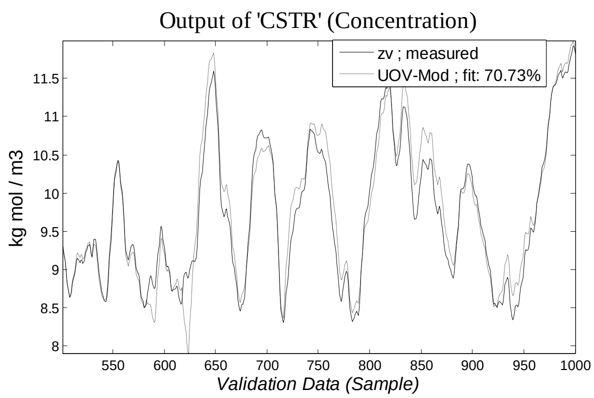

Fig 4: The measured (validation) data of the 'CSTR' system and the output signal of estimated model using the UOV criterion

Finally, the selected model orders together with the percentages of the output variation using different criteria are presented in Table 2. As shown in this table, the UOV criterion selected a more appropriate model order compared to the other criteria.

Table 2. The selected model orders for case studies and their percentages of the output variation

Criterion **Case Study**	*AIC*	*MDL*	*EEF*	*UOV*
TOW_TANK	[2 2 1] 81.29	[2 2 1] 81.29%	[2 3 1] 79.47%	[2 1 1] 83.93%
CSTR	[3 4 1] 70.04%	[3 4 1] 70.04%	[3 4 3] 69.17%	[3 2 1] 70.73%

5. CONCLUSIONS

A' Nonlinearity test' method and four model order selection criteria have been described and used for several examples and chemical processes with different levels of nonlinearity.

Nonlinearity test is first applied to determine the level of nonlinearity and the model order is then selected by utilizing the criteria. In systems with low level of nonlinearity, the model orders obtained from all criteria are identical. The model order determined by the UOV criterion in the systems with high level of nonlinearity is lower than the order determined by the other criteria. In fact, by increasing the level of non-linearity, the UOV criterion demonstrated the best performance compared to the other criteria. Among the discussed criteria, the EEF criterion resulted in the worst model selection.

6. REFERENCES

[1] Guay. M. 1996 Measurement of Nonlinearity in Chemical Process Control. Doctoral Thesis. UMI Order No. 1012., Department of Chemical Engineering, Queen's University at Kingston.

[2] Lind. I. 2006 Regressor and Structure Selection Uses of ANOVA in System Identification. Doctoral Thesis. UMI Order No.895., Department of Electrical Engineering, Linkoping University.

[3] Ridder, F.D. Pintelon, R. Schoukens, J., Gillikin, D.P., "Modified AIC and MDL Model Selection Criteria for Short Data Record", *(IEEE) TIM*, 2005.

[4] Haber, R. 1985. Nonlinearity Tests for Dynamic Processes. *Proc.IFAC Symp*.

[5] Tong, H. 1990 Nonlinear Time Series, a Dynamical System Approach. Oxford UK Clarendon Press.

[6] Stark. J.A, "Parameter-based Hypothesis Tests for Model Selection", *Signal Processing (EURASIP*), 1995.

[7] Bindlish. R, "Model order selection for process identification applied to an industrial ethylene furnace", *Jprocont*, 2003.

[8] Ding. Q, Kay. S, "Inconsistency of the MDL: On the Performance of Model Order Selection Criteria with Increasing Signal-to-Noise Ratio", *Signal Processing (IEEE)*, 2011.

[9] MALAB, "System Identification Toolbox Documentation', Mathworks Inc Natick, MA., 2010.

[10] Luyben, M.L. Luyben, W.L. 1997 Essentials of Process Control. McGraw-Hill Singapore.