CHEMICAL ENGINEERING TRANSACTIONS

VOL. 76, 2019

A publication of

AIDIC

The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Guest Editors: Petar S. Varbanov, Timothy G. Walmsley, Jiří J. Klemeš, Panos Seferlis
Copyright © 2019, AIDIC Servizi S.r.l.
ISBN 978-88-95608-73-0; **ISSN** 2283-9216

DOI: 10.3303/CET1976115

Analysis of the Informative Data of Industrial Data-Based Modelling

Li-Kun Yuan, Bao-Chang Xu*, Zhi-Shan Liang

Department of Automation, China University of Petroleum Beijing, 102249, China
xbcyl@cup.edu.cn

Modelling plays an important role in continuously operated plants as in chemical, metallurgical and aerospace fields, directly affecting subsequent researches of industrial processes. Rapid industrial development requires a high-quality model. Generally, external excitation is applied to the industrial process, ensuring the informative data for system identification. However, these processes are regularly operated to achieve a secure operation and to meet production objectives. For the safety and avoid violating product quality, external excitation can be forbidden or limited in the plant. Historical records are logged, which is natural to use them for analysis. These data are available at less, even no cost for identification. In this paper, based on the definition of information matrix and attenuating excitation, a detailed standard is aimed to propose to help extract informative data with respect to the chosen model structure to support system identification. The assessment of this standard is evaluated in a case, and the benefits are demonstrated, which ensures identification information enough, decreases information waste and achieves less, even no impact or cost of industrial processes.

1. Introduction

Industrial processes tend towards to complexity with the rapid development. The precision of the process model is significant for the development of control systems (Oravec et al., 2018). Tests with external excitation aimed at system identification are usually prohibitive to ensure processes running properly. Even allowed, the data set is limited. With the widespread use of the Internet and real time database (RTDB), such as a distributed control system (DCS) or safety instrumented system (SIS), historical data have been recorded. Based on existing industrial strategies, the data are available for fault diagnosis and equipment maintenance, which already achieved commercialization, seldom used for modelling. It’s obvious that historical data are available at no cost for analysis. As set points seldom changed in large continuously operated plants, collected data are mostly stationary, with little information of the system dynamics. Nevertheless, transient changes occurred and might excite the process, such transient regions contain information about the process dynamics. It was concluded that only parts (almost 1.5 %) of historically recorded data from a continuously operated chemical plant contained useful information for modelling (Bittencourt et al., 2015). It was proved that parameter biases could be decreased when an informative data set extracted from the entire data set (Carrette et al., 1996). It is important to isolate historical data, which intends to yield an informative data set suitable for system identification.
The algorithm can extract information to consistently estimate the system if (Arengas and Kroll, 2017) the data is informative, and the model set contains the true system. The model structure is based on prior knowledge, such as system controllability, observability (Leitolda et al., 2018). Considering continuously operated plants, such as chemical industry, ARX or auto-regressive moving average model (ARMAX) is the first choice of model structure ((Zhang et al., 2017). Besides the identifiable model structure, the key to identification is the informative data. Several kinds of literature discussed this. For autoregressive exogenous single-input single-output (ARX SISO) model structure, the demand for identification input excitation is persistently exciting (PE) of $2n$ (Ljung, 1999), n is the order of the model. Identifiability has been studied in (Gevers et al., 2009), stating the degree of input excitation required for typical SISO model structures. These theories are widely used, but too strict, affecting industrial processes operation and resulting in identification information redundancy. Attenuating excitation is proposed in (Ding, 2011); the effect of attenuating excitation is temporary, which is more suitable

Paper Received: 15/03/2019; Revised: 10/06/2019; Accepted: 02/07/2019
Please cite this article as: Yuan L.-K., Xu B.-C., Liang Z.-S., 2019, Analysis of the Informative Data of Industrial Data-Based Modelling, Chemical Engineering Transactions, 76, 685-690 DOI:10.3303/CET1976115

for the condition that long-term input excitation is not allowed in identification. Thus, based on information matrix and attenuating excitation, this paper provides a new standard, which is possible to achieve less, even no impact or cost of industrial processes, more suitable for industrial identification. Simulation proves that data set isolated or designed by this standard can be used for identification to meet the requiring accuracy, which improves the data efficiency and reduces information waste.

2. Preliminaries

2.1 The prediction error identification setup

Considering a linear time-invariant discrete-time single-input single-output process $\mathcal{S}$

$$\mathcal{S}: y(t) = G_0(z)u(t) + H_0(z)e(t) \tag{1}$$

In (1), z is forward-shift operator, $G_0(z)$ and $H_0(z)$ are the process transfer functions, $u(t)$ and $y(t)$ are separately the system input and output, $e(t)$ is a zero-mean white noise with variance ${\sigma_e}^2$. The "true" system can be expressed in a compact form by $\mathcal{S} \triangleq [G_0(z)\ H_0(z)\ ]$.
(1) is identified using a model structure $\mathrm{M}(\theta)$ parametrized be a vector $\theta \in \boldsymbol{\mathcal{R}}^d$

$$\mathrm{M}(\theta): y(t) = G(z,\theta)u(t) + H(z,\theta)e(t) \tag{2}$$

Assuming that the transfer function has a non-zero delay, both for $G_0(z)$ and for $G(z,\theta)$. The set of models $\mathrm{M}(\theta)$, for all θ in some set $D_\theta \in \boldsymbol{\mathcal{R}}^d$, defines the model set $\mathcal{M} \triangleq \{M(\theta)|\theta \in D_\theta\}$. The true system meets the situation $\mathcal{S} \in \mathcal{M}$, if there is a θ_0 such that $\mathrm{M}(\theta_0) = \mathcal{S}$.The one-step-ahead predictor of $y(t)$ is defined as:

$$\hat{y}(t|t-1\,,\theta) = W_u(z,\theta)u(t) + W_y(z,\theta)y(t) \tag{3}$$

where $W_u(z,\theta) = H^{-1}(z,\theta)G(z,\theta), W_y(z,\theta) = [1 - H^{-1}(z,\theta)]$, $z(t) \triangleq [\mathrm{u(t)\ y(t)}]^T$.
Thus, (3) can be expressed as

$$\hat{y}(t|t-1\,,\theta) = W(z,\theta)z(t) \tag{4}$$

where $W(z,\theta) \triangleq [W_u(z,\theta) \quad W_y(z,\theta)]$.
Then the one-step-ahead prediction error $\varepsilon(t,\theta)$ is expressed as:

$$\varepsilon(t,\theta) \triangleq y(t) - \hat{y}(t,\theta) \triangleq H^{-1}(z,\theta)[y(t) - G(z,\theta)u(t)] \tag{5}$$

Using a data set of length N and the prediction error method (PEM) yields the estimate $\hat{\theta}_N$ (Ljung, 1999)

$$\hat{\theta}_N = arg \min_{\theta \in D_\theta} \frac{1}{N} \sum_{t=1}^{N} \varepsilon^2(t,\theta) \tag{6}$$

If $\mathcal{S} \in \mathcal{M}$ and $\hat{\theta}_N \xrightarrow{N\to\infty} \theta_0$, the parameter error converges to a Gaussian random variable: $\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{N\to\infty} N(0,P_\theta)$, with

$$P_\theta = [I(\theta)]^{-1}|_{\theta=\theta_0} \tag{7}$$

$$I(\theta) = \frac{1}{{\sigma_e}^2} E[\psi(t,\theta)\psi(t,\theta)^T] \tag{8}$$

$$\psi(t,\theta) = -\frac{\partial\varepsilon(t,\theta)}{\partial\theta} = \frac{\partial\hat{y}(t,\ \theta)}{\partial\theta} = \nabla_\theta W(z,\theta)z(t) \tag{9}$$

where $\nabla_\theta W(z,\theta) = \frac{\partial W(z,\ \theta)}{\partial\theta}$, $I(\theta)$ is called an information matrix.

2.2 Identifiability, Informative data and Persisting exciting

Several concepts have been proposed in the scientific literature. Here some definitions are as follows.
Definition 1 (Ljung, 1999) (Identifiability) A parametric model structure $M(\theta)$ is locally identifiable at a value θ_1 if $\exists\delta > 0$, such that, for all θ in $||\theta - \theta_1|| \leq \delta$

$$[W(z,\theta) \quad W(z,\theta)] = [W_u(z,\theta_1) \quad W_y(z,\theta_1)] \quad \forall z \Rightarrow \theta = \theta_1 \tag{10}$$

If $\delta \to \infty$, The model structure is globally identifiable at θ_1 .
Identifiability Gramian $\Gamma(\theta) \in \boldsymbol{\mathcal{R}}^{d\times d}$ is proposed as (Ljung, 1999)

$$\Gamma(\theta) \triangleq \int_{-\pi}^{\pi} \nabla_\theta W(e^{j\omega},\theta)\,\nabla_\theta W^H(e^{j\omega},\theta) \tag{11}$$

where $\nabla_\theta W(e^{j\omega},\theta) = \frac{\partial W(e^{j\omega},\theta)}{\partial\theta}$, the notation $\mathrm{M}^H(e^{j\omega}) = \mathrm{M}^T(e^{-j\omega})$, ω is angular velocity.

Definition 2 (Ljung, 1999) (Informative Data) A quasistationary data set $z(t)$ is informative with respect to a parametric model set $\{M(\theta), \theta \in D_\theta\}$ if, for any two models $W(z,\theta_1)$ and $W(z,\theta_2)$ in that set

$$E\{[W(z,\theta_1) - W(z,\theta_2)]z(t)\}^2 = 0 \Rightarrow\ W(e^{j\omega},\theta_1) =\ W(e^{j\omega},\theta_2) \text{ at almost all } \omega \tag{12}$$

Combining the information matrix (8)(9) and using Parseval's Theorem yields

$$I(\theta) = \frac{1}{2\pi{\sigma_e}^2}\int_{-\pi}^{\pi} \nabla_\theta W(e^{j\omega},\theta)\varphi_z(\omega)\,W^H(e^{j\omega},\theta)d\omega \tag{13}$$

where $\phi_z(\omega)$ is the power spectrum of the data $z(t)$.The matrix $I(\theta)$ is semi-definite by construction and will play a central role in this paper. The information matrix shows the combination of model structure identifiability and the richness of the data (through $\phi_z(\omega)$).

Combining informative data (12) and information matrix (13) and using Taylor Theorem then

$$E\{[W(z,\theta_1) - W(z,\theta_2)]z(t)\}^2 = (\theta_1-\theta_2)^T I(\theta_1)(\theta_1-\theta_2) + \rho(|\theta_1-\theta_2|^2) \tag{14}$$

Where $\lim_{\theta_1\to\theta_2}(\rho(|\theta_1-\theta_2|^2)/\theta_1-\theta_2) = 0$

The information matrix $I(\theta)$ should be positive definite that ensures $E\{[W(z,\theta_1) - W(z,\theta_2)]z(t)\}^2 = 0$, which implies $\theta_1 = \theta_2$.The positive definiteness of $I(\theta_1)$ depends on the data set through $\phi_z(\omega)$, can be used to analyse the informativity of the data set.

Definition 3 (Ljung, 1999) (Persisting Exciting) A quasistationary data $\mathrm{u(t)}$ is persisting exciting of order n, if power spectrum $\phi_u(\omega)$ is different from zero on at least n frequency points in the interval $(-\pi,\pi]$, where $\phi_u(\omega) = \lim_{N\to\infty}\frac{|u_N(\omega)|^2}{N}$, N is the length of the data set.

Obviously, step data is persisting exciting of order 1, n different sinusoids data is persisting exciting of order $2\mathrm{n}$, white noise is persisting exciting of order ∞, which is widely used for identification, but is not fit in this paper.

2.3 Attenuating Excitation

For the continuously operated plants, tests with external excitation are usually prohibitive. Even allowed, also limited. The attenuating excitation with a temporary impact on the system is proposed, more meaningful in practice.

There are serious forms of attenuating excitation. This paper only provides the following form:

$$\mathrm{u(t)} = e^{-at}\mathrm{u_0(t)}, a > 0 \tag{15}$$

where $\mathrm{u_0(t)}$ is persisting exciting satisfied $I(\theta) > 0$, a is defined as attenuating index.

Obviously, the greater the attenuating index a is the faster-attenuating excitation decays, which means less information provided. Bet there would be an extreme attenuating index a which ensures the informative data about the identification, guarantees the required model accuracy. From the separation of signal property, the attenuating excitation $\mathrm{u(t)}$ is energy data, should be analysed by energy spectrum $\varepsilon(\omega)$, where $\varepsilon(\omega) = |u(\omega)|^2$.

3. Case Study

In this paper, considering the following ARX system as the "true" system to describe the liquid level system in industrial process, where $\mathrm{u(t)}$ is flow rate (L/s) as the input, $\mathrm{y(t)}$ is liquid level (mm) as the output.

$$\mathrm{A(z)y(t) = B(z)u(t) + e(t)} \tag{16}$$

with $\mathrm{A(z) = 1 - 1.5z^{-1} + 0.7z^{-2}}$, $\mathrm{B(z) = z^{-1} + 0.5z^{-2}}$, $\mathrm{e(t)}$ is independent zero-mean white noise with variance ${\sigma_e}^2 = 0.05$.

The system can be described as

$$\mathrm{y(t)} = \mathbf{h}^T\mathrm{(t)}\theta + \mathrm{e(t)} \tag{17}$$

where for an ARX model the following expressions are derived

$$\mathbf{h}\mathrm{(t)} = [\mathrm{-y(t-1), -y(t-2), u(t-1), u(t-2)}\ ]^T,\ \theta = [\mathrm{a_1, a_2, b_1, b_2}\ ]^T$$

Let N be the length of data set, then

$\mathbf{y}_{\boldsymbol{N}} = [\mathrm{y}(1), \mathrm{y}(2), \dots, \mathrm{y}(\mathrm{N})\ ]^{T}$, $\mathbf{e}_{\boldsymbol{N}} = [\mathrm{e}(1), \mathrm{e}(2), \dots, \mathrm{e}(\mathrm{N})\ ]^{T}$, $\mathbf{H}_{\boldsymbol{N}} = [\mathbf{h}^{\boldsymbol{T}}(1), \mathbf{h}^{\boldsymbol{T}}(2), \dots, \mathbf{h}^{\boldsymbol{T}}(\mathrm{N})]^{\boldsymbol{T}}$**,**

In the simulation, white noise sequence with zero mean and variance $\sigma_u{}^2 = 1$, $\mathrm{u(t)} = \sin\left(\frac{1}{2}\pi\mathrm{t}\right)$, and $\mathrm{u(t)} = e^{-at}\sin\left(\frac{1}{2}\pi\mathrm{t}\right), \mathrm{a} > 0$ are used as the input data at specified operating point, respectively. The output data for the identification procedure are obtained from the simulations. These conditions can simulate industrial processes. Recursive least squares (RLS) as the algorithm is applied to estimate the parameters, performed in MATLAB. By analysing the identification results with different input, the worst condition proposed through attenuating excitation, which ensures informative data for the chosen model structure has been verified. To ensure the reliability of the simulation, 100 repetitions of independent experiments with data length $\mathrm{N} = 400$ were carried out. The relative error of the estimated parameters (δ) was used as the evaluation criterion.

Table 1: The parameter values with different inputs

Input data $u(t)$ [L/s]	a_1 [1]	a_2 [1]	b_1 $[s/m^2]$	b_2 $[s/m^2]$	$\delta(\lvert\frac{\hat{\theta}-\theta}{\theta}\rvert)$ [%]
True value	-1.5	0.7	1	0.5	0
White noise with variance $\sigma_u{}^2 = 1$	-1.5126	0.7150	1.0300	0.5141	1.80
$\sin\left(\frac{1}{2}\pi\mathrm{t}\right)$	-1.4771	0.6747	0.9760	0.5164	2.72
$e^{-at}\sin\left(\frac{1}{2}\pi\mathrm{t}\right), \mathrm{a} = 0.1$	-1.4756	0.7260	1.0326	0.5261	3.38

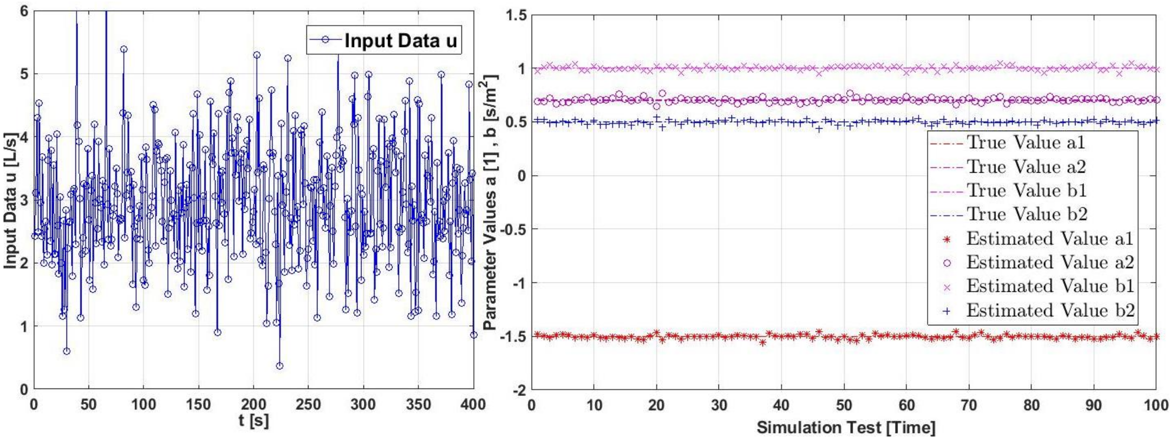

Figure 1: Curves of parameter values with input white noise (variance $\sigma_u{}^2 = 1$)

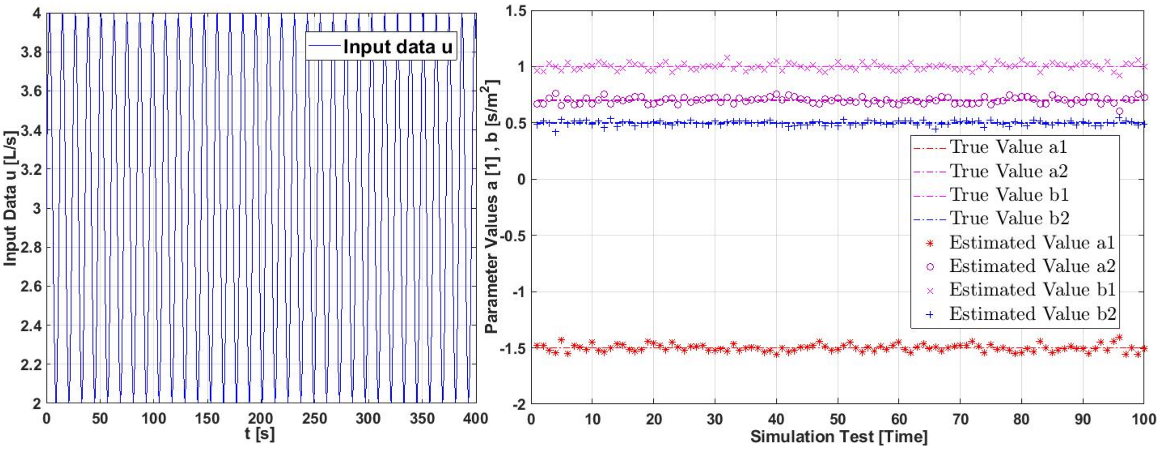

Figure 2: Curves of parameter values with input $sin\left(\frac{1}{2}\pi t\right)$

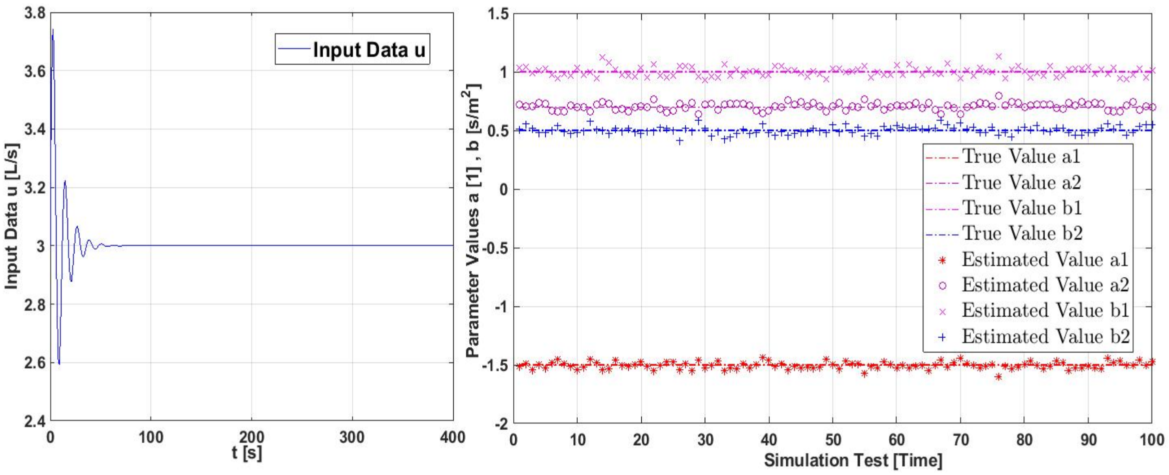

Figure 3: Curves of parameter values with input $e^{-0.1t}\, sin\left(\frac{1}{2}\pi t\right)$

From Table 1 and Figure 1-3, the decrease of excitation degree, which means less informative data provided, has a certain impact on the identification results. But the relative error fluctuation is very small; that is, the model parameters can still be estimated effectively.

From the definition of attenuating excitation, it's obvious that the greater the attenuating index a is, the faster-attenuating excitation decays, the less informative data provides. But there would be an extreme value of the attenuation index that guarantees informativity of experiments to meet reasonable accuracy. Considering this case, the extreme value of the attenuation index can be calculated by the theory mentioned in this paper, the simulation results are shown in Table 2 and Figure 4-5.

Table 2: The parameter values with different attenuating excitations

Input data $u(t) = e^{-at}\sin\left(\frac{1}{2}\pi t\right)$ [L/s]	a_1 [1]	a_2 [1]	b_1 $[s/m^2]$	b_2 $[s/m^2]$	$\delta(\left\lvert\frac{\hat{\theta}-\theta}{\theta}\right\rvert)$ [%]
True value	-1.5	0.7	1	0.5	0
$a = 0.1$	-1.4824	0.7260	1.0326	0.5261	3.38
$a = 1$	-1.4476	0.6236	1.0531	0.4559	8.71
$a = 2$	-1.5172	0.7669	1.1984	0.6463	23.22
$a = 3$	-1.4507	0.6492	1.8831	0.1609	69.20

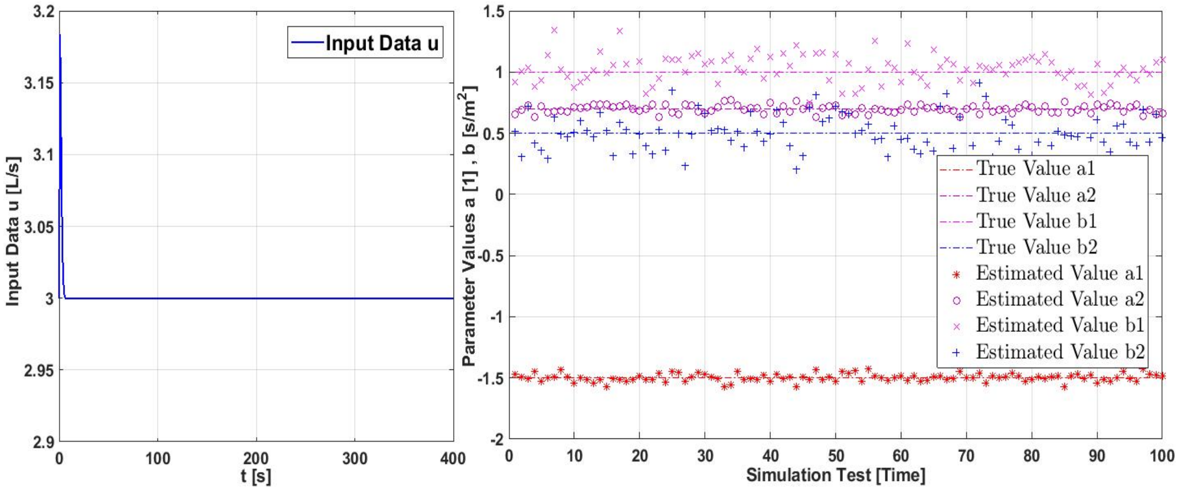

Figure 4: Curves of parameter values with input $e^{-t}\, sin\left(\frac{1}{2}\pi t\right)$

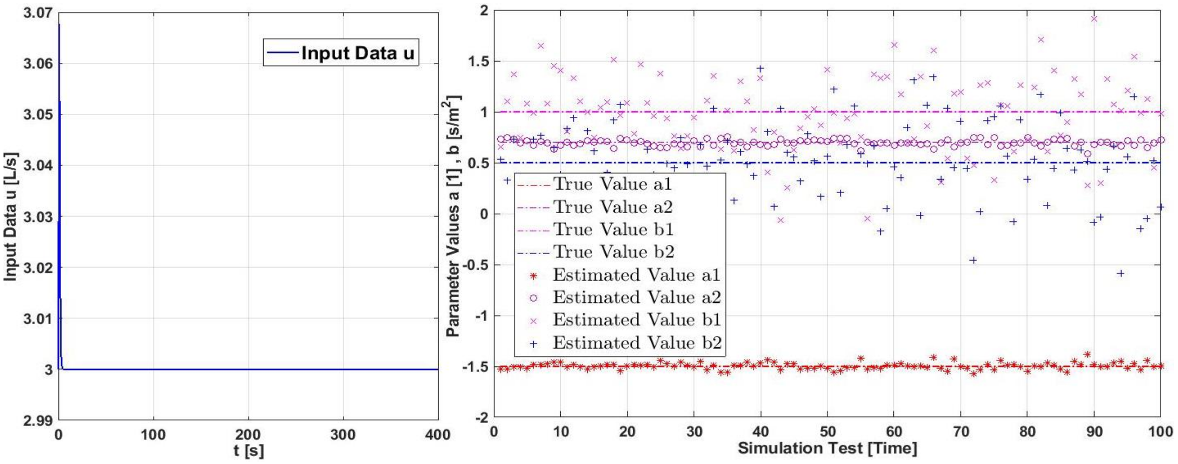

Figure 5: Curves of parameter values with input $e^{-2t}\,sin\left(\frac{1}{2}\pi t\right)$

In this case, the extreme value of the attenuating index $a = 1$ by theoretical analysis. When $a < 1$, the relative error fluctuation is very small that still guarantee the estimated accuracy. When $a > 1$, the relative error is growing rapidly, which means the parameters cannot be estimated. The simulation also proves the result.

4. Conclusions

In this paper, based on informative data and attenuating excitation, the standard of the test data which ensures the experiment informativity is proposed, analysed by a case study with respect to a chosen attenuating excitation form. The result shows the informativity can be guaranteed under a reasonable attenuating index to ensure the required model accuracy. This standard can be used for the test input design at the lowest cost with limited external excitation allowed, also used for isolating historical data. Comparing with the existing informative data requirement and the industrial historical data strategies, it can expand historical data analysis to a new industrial strategy of modelling, also improve the data utilization, reduce process information lost, which is of great significance for the low-cost of the industrial modelling.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (21676295).

References

Arengas D., Kroll A., 2017. A Search Method for Selecting Informative Data in Predominantly Stationary Historical Records for Multivariable System Identification. 21st International Conference on System Theory, Control and Computing, ICSTCC 2017, Sinaia, Romania, October 19-21, 100-105.

Bittencourt A.C., Isaksson A.J., Peretzki D., Forsman K., 2015. An Algorithm for Finding Process Identification Intervals from Normal Operating Data. Processes, 3(3), 357–383.

Carrette P., Bastin G., Genin Y. Y., Gevers M., 1996. Discarding Data May Help in System Identification. IEEE Transactions on Signal Processing, 44(9), 2300–2310.

Ding F., 2011. System Identification Part C: Identification Accuracy and Basic Problem. Journal of Nanjing University of Information Science & Technology, 3(3), 193-226.

Gevers M., Bazanella A. S., Bombois X., Miskovic L., 2009. Identification and the Information Matrix: How to Get Just Sufficiently Rich? IEEE Transactions on Automatic Control, 54(12), 2828–2840.

Oravec J., Bakošová M., Vasičkaninová A., Mészáros A., 2018. Robust Model Predictive Control of a Plate Heat Exchanger. Chemical Engineering Transactions, 70, 25-30.

Leitold D., Vathy-Fogarassy A., Abonyi J., 2018. Design-oriented Structural Controllability and Observability Analysis of Heat Exchanger Networks. Chemical Engineering Transactions, 70, 595-600.

Ljung L., 1999. System Identification: Theory for the user, 2nd ed. Prentice Hall, Upper Saddle River, NJ, United States, ISBN: 978-0136566953.

Zhang F., Zhang W., Zhang J. M., 2017. Multi-variate Identification of Crude Oil Refining Processes Using ASYM Method. Journal of East China University of Science and Technology (Natural Science Edition), 43(3), 397-403.