Research Article

Identification of LTI Time-Delay Systems with Missing Output Data Using GEM Algorithm

Xianqiang Yang[1] and Hamid Reza Karimi[2]

[1] *Research Institute of Intelligent Control and Systems, Harbin Institute of Technology, Harbin, Heilongjiang 150080, China*
[2] *Department of Engineering, Faculty of Engineering and Science, University of Agder, 4898 Grimstad, Norway*

Correspondence should be addressed to Xianqiang Yang; xianqiangyang@gmail.com

This paper considers the parameter estimation for linear time-invariant (LTI) systems in an input-output setting with output error (OE) time-delay model structure. The problem of missing data is commonly experienced in industry due to irregular sampling, sensor failure, data deletion in data preprocessing, network transmission fault, and so forth; to deal with the identification of LTI systems with time-delay in incomplete-data problem, the generalized expectation-maximization (GEM) algorithm is adopted to estimate the model parameters and the time-delay simultaneously. Numerical examples are provided to demonstrate the effectiveness of the proposed method.

1. Introduction

The advanced process control theories have enjoyed rapid development in the past several decades to meet the growing demands of closed-loop system performances, such as improved process safety and efficiency of plant operation, consistent product quality, and economic optimization [1]. These control strategies have improved the process automation and stability through providing control solutions for the process operated under the abnormal working conditions, such as process fault [2–5], network transmission delay [6], data packet dropouts [7], and modeling error. Generally, the implementation of these control strategies relies on the understanding of the process dynamics and the availability of an accurate mathematical model of the process. In view of the difficulties and complexities imposed by modeling using first principle method, the data-driven modeling method, in which the process model is retrieved from the process data, has become a main modeling method.

Typically, the process data used in process modeling are generated by performing an identification experiment, in which a testing signal is designed and utilized to excite the process. Most of the conventional parameter estimation methods, such as prediction error method (PEM),

instrumental variable (IV) method, and subspace method, assume that the identification data are sampled regularly and recorded properly. However, this is not always true in practical industry. For example, in the development of an inferential model for the sulfur content in the gas oil product, the sulfur concentration cannot be measured directly and the lab analysis is required which takes a long time. The process variable can be sampled in every minute, but the sulfur concentration is only available in every twelve hours. Another example is the industrial process with data transmission through the network. The recorded process data are corrupted by many network-induced problems, such as transmission delay and packet dropout or missing. Therefore, parameter estimation with irregular data has not been extensively investigated in the literature.

Time-delays are commonly encountered in various engineering systems, such as chemical processes, mechanical systems, network control systems, transmission line, and economic systems [8]. Since the existence of time-delay usually causes performance degradation of the inferential model and is frequently a source of instability of the closed-loop system, it should be handled carefully in the modeling process. Common methods to estimate the time-delays are

nonparametric methods (e.g., step test or correlation analysis) and grid searching method. For example, Wang and Zhang [9] considered the robust identification problems of linear continuous time-delay systems from step responses. A linear regression equation was derived from the solution of the output time response and its various-order integrals and solved by using IV-least squares (LS) method. The parameters of the transfer function were then recovered from the LS solution. Weyer [10] considered to build a model for the open water channel. The model parameters were estimated using the grid searching method in which one model was established for each time-delay in a range. The final model was selected as the model which gave the best prediction performance for a validation data set. The time-delay estimation methods mentioned above are to estimate the time-delay and model parameters in a separate way.

Missing data problem is very common in process industry. A special example is the irregularly sampling system. Many critical parameters, such as the product concentration, steam quality, and 90% boiling point, cannot be measured directly by using the sensors. These parameters are measured through lab analysis, so only the slow rate data are available. However, the process variables, such as the temperature, pressure, and flow rate, can be measured on-line in fast rate by using the sensors. Therefore, we can treat the data samples between the slow rate data as missing data. Another example is the network control system in which data transmit via the wireless network or internet. Data missing occurs due to data packet dropout or missing. Other reasons for missing data are sensor fault, data recording system malfunction, and so forth. Several methods have been reported in the literature to handle missing data problem in system identification. For example, F. Ding and J. Ding [11] proposed an auxiliary model-based approach to cope with the problems of parameter estimation and output estimation with irregularly missing output data using the PEM method. The outputs of the auxiliary model were used in the identification process. Zhu et al. [12] considered the identification of systems with slowly and irregularly sampled output data. The output error method was employed to estimate the fast rate model based on the fast input and slow output data. However, the methods mentioned above just used part of the process data, which may lead to information missing. Moreover, the statistical properties of the model parameters and the process noise cannot be given in these methods.

The work introduced in this paper aims at handling the identification problem of the LTI systems with missing output data in the presence of time-delay. The identification problem is formulated under the scheme of the generalized expectation-maximization (GEM) algorithm and the time-delay and missing output data are handled simultaneously. The GEM algorithm consists of expectation step (E-step) and maximization step (M-step). In the M-step, the maximization problem is transformed into an equivalent minimization problem and this problem is solved by using a general numerical optimization algorithm.

The rest of this paper is organized as follows. The problem statement is presented in Section 2. A brief revisit of the GEM algorithm and the mathematical formulation of the identification of LTI time-delay systems with incomplete data set are given in Section 3. Numerical examples are presented in Section 4 to show the effectiveness of the proposed method. The conclusions are given in Section 5.

2. Problem Statement

Consider the LTI system described by the following output error (OE) time-delay model:

$$y_t = G\left(z^{-1}\right) u_{t-\tau} + e_t, \tag{1}$$

where τ is the time-delay which is assumed to be integer multiples of the sampling period, e_t is the Gaussian white noise with zero mean and variance σ^2, and y_t and u_t are the output and input, respectively. The transfer function $G(z^{-1})$ has the following form:

$$G\left(z^{-1}\right) = \frac{\sum_{i=1}^{n_b} b_i z^{-i}}{1 + \sum_{i=1}^{n_a} a_i z^{-i}}. \tag{2}$$

Here, we assume that the model orders n_a and n_b are known a priori and the time-delay τ is uniformly distributed in a known range of $[d_1, d_2]$.

The identification data $\{y_t, u_t\}_{t=1,2,\ldots,N}$ are collected. We denote $\{y_t\}_{t=1,2,\ldots,N}$ as Y and $\{u_t\}_{t=1,2,\ldots,N}$ as U. Since part of the output data are missing completely at random (MCAR), the output data set Y can be divided into $Y_{\text{obs}} = \{y_t\}_{t=t_1,\ldots,t_\alpha}$ and $Y_{\text{mis}} = \{y_t\}_{t=m_1,\ldots,m_\beta}$. Therefore, the identification problem is to estimate the parameters $\theta = \{a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}\}$, the noise variance σ^2, and the time-delay τ based on the identification data Y_{obs} and U.

3. Parameter Estimation Using the GEM Algorithm

3.1. GEM Algorithm Revisit. The GEM algorithm is a general-purpose iterative optimization algorithm to derive the maximum likelihood (ML) estimate and it has attracted great attentions of the researcher due to its flexibility in handling the missing data or hidden state [13]. Denote the missing data set by C_{mis} and the observed data set by C_{obs}. The main idea of the GEM algorithm is that, instead of optimizing the likelihood of the observed Y_{obs}, the conditional expectation of the complete data likelihood function with respect to the missing data set is calculated in the E-step and the maximization problem is solved in the M-step. The procedures of the GEM algorithm to calculate the ML estimate can be described as follows [13]:

E-step: given the C_{obs} and the parameter estimate Θ^s in previous iteration, the Q-function can be calculated by

$$Q\left(\Theta \mid \Theta^s\right) = E_{C_{\text{mis}} \mid C_{\text{obs}}, \Theta^s} \left\{\log p\left(C_{\text{mis}}, C_{\text{obs}} \mid \Theta\right)\right\}, \tag{3}$$

M-step: find the Θ^{s+1} to increase $Q(\Theta \mid \Theta^s)$ over its value at Θ^s; that is,

$$Q\left(\Theta^{s+1} \mid \Theta^s\right) \geqslant Q\left(\Theta^s \mid \Theta^s\right). \tag{4}$$

The E-step and M-step alternate until the relative change of the parameter estimate between neighboring iterations is smaller than a prespecified arbitrary small constant or the maximal iteration number is achieved.

3.2. LTI Time-Delay System Identification with Missing Output Using GEM Algorithm. Here, we treat the time-delay τ as a hidden state variable. The observed data set C_{obs} is constructed as $C_{obs} = \{Y_{obs}, U\}$ and the missing data set C_{mis} is constructed as $C_{mis} = \{Y_{mis}, \tau\}$. The parameter vector Θ is constructed as $\Theta = \{\theta, \sigma^2\}$.

Based on the Bayesian property, the likelihood function of the complete data set can be decomposed into

$$
\begin{aligned}
p\left(C_{mis}, C_{obs} \mid \Theta\right) &= p\left(Y, U, \tau \mid \Theta\right) \\
&= p\left(Y \mid U, \tau, \Theta\right) p\left(\tau \mid U, \Theta\right) p\left(U \mid \Theta\right).
\end{aligned}
\tag{5}
$$

The term $p(Y \mid U, \tau, \Theta)$ can be further decomposed into

$$
\begin{aligned}
p\left(Y \mid U, \tau, \Theta\right) &= p\left(y_N, \ldots, y_1 \mid u_N, \ldots, u_1, \tau, \Theta\right) \\
&= p\left(y_N \mid y_{N-1}, \ldots, y_1, u_N, \ldots, u_1, \tau, \Theta\right) \\
&\quad \times p\left(y_{N-1}, \ldots, y_1 \mid u_N, \ldots, u_1, \tau, \Theta\right) \\
&= \prod_{t=1}^{N} p\left(y_t \mid y_{t-1}, \ldots, y_1, u_N, \ldots, u_1, \tau, \Theta\right).
\end{aligned}
\tag{6}
$$

Based on (1) and (2), y_t depends only on the previous input sequence $u_{t-1:1} = \{u_{t-1}, \ldots, u_1\}$, the time-delay τ, and the parameter vector Θ. Therefore, (6) can be rewritten as

$$
p\left(Y \mid U, \tau, \Theta\right) = \prod_{t=1}^{N} p\left(y_t \mid u_{t-1:1}, \tau, \Theta\right).
\tag{7}
$$

Since the time-delay τ is uniformly distributed in the range $[d_1, d_2]$, the probability of τ taking any value in this range is a constant. Since the input U is measurable data and it is independent of the parameter vector Θ, the term $p(U \mid \Theta)$ is a constant. Therefore, the last two terms of (5) will not play a role in the following derivations. The complete data likelihood function can be further written as

$$
p\left(Y, U, \tau \mid \Theta\right) = \prod_{t=1}^{N} p\left(y_t \mid u_{t-1:1}, \tau, \Theta\right) C_1,
\tag{8}
$$

where $C_1 = p(\tau \mid U, \Theta) p(U \mid \Theta)$.

Therefore, the conditional expectation of the log complete data density $Q(\Theta \mid \Theta^s)$ in (3) can be written as

$$
\begin{aligned}
&Q\left(\Theta \mid \Theta^s\right) \\
&= E_{Y_{mis}, \tau \mid C_{obs}, \Theta^s} \left\{\log\, p\left(Y, U, \tau \mid \Theta\right)\right\} \\
&= E_{Y_{mis}, \tau \mid C_{obs}, \Theta^s} \left\{\sum_{t=1}^{N} \log p\left(y_t \mid u_{t-1:1}, \tau, \Theta\right) + \log C_1\right\} \\
&= E_{Y_{mis}, \tau \mid C_{obs}, \Theta^s} \left\{\sum_{t=m_1}^{m_\beta} \log p\left(y_t \mid u_{t-1:1}, \tau, \Theta\right)\right. \\
&\qquad\qquad \left. + \sum_{t=t_1}^{t_\alpha} \log p\left(y_t \mid u_{t-1:1}, \tau, \Theta\right) + \log C_1\right\}.
\end{aligned}
\tag{9}
$$

The expectation is firstly taken with respect to the discrete variable τ; then we have

$$
\begin{aligned}
&Q\left(\Theta \mid \Theta^s\right) \\
&= E_{Y_{mis} \mid C_{obs}, \tau, \Theta^s} \left\{\sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{obs}, \Theta^s\right)\right. \\
&\qquad\qquad \times \sum_{t=m_1}^{m_\beta} \log\, p\left(y_t \mid u_{t-1:1}, \tau = d, \Theta\right) \\
&\qquad\qquad + \sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{obs}, \Theta^s\right) \\
&\qquad\qquad \times \sum_{t=t_1}^{t_\alpha} \log p\left(y_t \mid u_{t-1:1}, \tau = d, \Theta\right) \\
&\qquad\qquad \left. + \log C_1\right\}.
\end{aligned}
\tag{10}
$$

The expectation is then taken with respect to the continuous variable Y_{miss}, so we have

$$
\begin{aligned}
Q\left(\Theta \mid \Theta^s\right) &= \sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{obs}, \Theta^s\right) \\
&\quad \times \sum_{t=m_1}^{m_\beta} \int p\left(y_t \mid C_{obs}, \tau = d, \Theta^s\right) \\
&\qquad\qquad \times \log\, p\left(y_t \mid u_{t-1:1}, \tau = d, \Theta\right) dy_t \\
&\quad + \sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{obs}, \Theta^s\right) \\
&\qquad \times \sum_{t=t_1}^{t_\alpha} \log\, p\left(y_t \mid u_{t-1:1}, \tau = d, \Theta\right) + \log C_1.
\end{aligned}
\tag{11}
$$

In order to calculate $Q(\Theta \mid \Theta^s)$, the unknown terms should be calculated firstly. Consider

$$p\left(\tau = d \mid C_{\text{obs}}, \Theta^s\right)$$

$$= p\left(\tau = d \mid Y_{\text{obs}}, U, \Theta^s\right)$$

$$= \frac{p\left(Y_{\text{obs}} \mid U, \tau = d, \Theta^s\right) p\left(\tau = d \mid U, \Theta^s\right)}{\sum_{d=d_1}^{d_2} p\left(Y_{\text{obs}} \mid U, \tau = d, \Theta^s\right) p\left(\tau = d \mid U, \Theta^s\right)}$$

$$= \frac{\prod_{t=t_1}^{t_\alpha} p\left(y_t \mid u_{t-1:1}, \tau = d, \Theta^s\right) p\left(\tau = d\right)}{\sum_{d=d_1}^{d_2} \prod_{t=t_1}^{t_\alpha} p\left(y_t \mid u_{t-1:1}, \tau = d, \Theta^s\right) p\left(\tau = d\right)}$$

$$= \frac{\prod_{t=t_1}^{t_\alpha} p\left(y_t \mid u_{t-1:1}, \tau = d, \Theta^s\right)}{\sum_{d=d_1}^{d_2} \prod_{t=t_1}^{t_\alpha} p\left(y_t \mid u_{t-1:1}, \tau = d, \Theta^s\right)},$$

$$\int p\left(y_t \mid C_{\text{obs}}, \tau = d, \Theta^s\right) \log p\left(y_t \mid u_{t-1:1}, \tau = d, \Theta\right) dy_t$$

$$= \int p\left(y_t \mid C_{\text{obs}}, \tau = d, \Theta^s\right) \log \frac{1}{\sqrt{2\pi\sigma^2}}$$

$$\times \exp\left\{-\frac{\left(y_t - G\left(z^{-1}\right) u_{t-d}\right)^2}{2\sigma^2}\right\} dy_t$$

$$= -\frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \int p\left(y_t \mid C_{\text{obs}}, \tau = d, \Theta^s\right)$$

$$\times \left(y_t - G\left(z^{-1}\right) u_{t-d}\right)^2 dy_t$$

$$= -\frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(\left(\sigma^s\right)^2 + \left(G^s\left(z^{-1}\right) u_{t-d}\right)^2\right)$$

$$+ \frac{1}{\sigma^2}\left(G\left(z^{-1}\right) u_{t-d}\right)\left(G^s\left(z^{-1}\right) u_{t-d}\right)$$

$$- \frac{1}{2\sigma^2}\left(G\left(z^{-1}\right) u_{t-d}\right)^2$$

$$= -\frac{1}{2} \log\left(2\pi\sigma^2\right)$$

$$- \frac{1}{2\sigma^2}\left(\left(\sigma^s\right)^2 + \left(G\left(z^{-1}\right) u_{t-d} - G^s\left(z^{-1}\right) u_{t-d}\right)^2\right),$$

$$p\left(y_t \mid u_{t-1:1}, \tau = d, \Theta\right)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left(y_t - G\left(z^{-1}\right) u_{t-d}\right)^2}{2\sigma^2}\right\}.$$

(12)

Therefore, the Q-function can be rewritten as

$$Q\left(\Theta \mid \Theta^s\right)$$

$$= \sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{\text{obs}}, \Theta^s\right)$$

$$\times \sum_{t=m_1}^{m_\beta} \left\{-\frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\right.$$

$$\times \left(\left(\sigma^s\right)^2 + \left(G\left(z^{-1}\right) u_{t-d} - G^s\left(z^{-1}\right) u_{t-d}\right)^2\right)\right\}$$

$$+ \sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{\text{obs}}, \Theta^s\right)$$

$$\times \sum_{t=t_1}^{t_\alpha} \left\{-\frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(y_t - G\left(z^{-1}\right) u_{t-d}\right)^2\right\}$$

$$+ \log C_1.$$

(13)

In the M-step of the GEM algorithm, the unknown parameters should be estimated to increase the Q-function by solving an optimization problem. Taking the gradient of the Q-function (13) with respect to the σ^2 and setting it to zeros, we have

$$\widehat{\sigma}^2 = \frac{1}{N}\left\{\sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{\text{obs}}, \Theta^s\right)\right.$$

$$\times \sum_{t=m_1}^{m_\beta} \left(\left(G\left(z^{-1}\right) u_{t-d} - G^s\left(z^{-1}\right) u_{t-d}\right)^2 + \left(\sigma^s\right)^2\right)$$

$$+ \sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{\text{obs}}, \Theta^s\right)$$

$$\left.\times \sum_{t=t_1}^{t_\alpha}\left(y_t - G\left(z^{-1}\right) u_{t-d}\right)^2\right\}.$$

(14)

Substituting $\widehat{\sigma}^2$ into the Q-function (13), we get

$$Q\left(\Theta \mid \Theta^s\right)$$

$$= -\frac{N}{2} \log\left(2\pi\right) - \frac{N}{2}$$

$$- \frac{N}{2} \log\left\{\frac{1}{N}\left[\sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{\text{obs}}, \Theta^s\right)\right.\right.$$

$$\times \left(\sum_{t=m_1}^{m_\beta} \left(\left(G\left(z^{-1}\right) u_{t-d} - G^s\left(z^{-1}\right) u_{t-d}\right)^2\right.\right.$$

$$+ \left(\sigma^s\right)^2\right)$$

$$\left.\left.\left.+ \sum_{t=t_1}^{t_\alpha}\left(y_t - G\left(z^{-1}\right) u_{t-d}\right)^2\right)\right]\right\}.$$

(15)

Based on the monotonicity of the log function, the problem is transformed into optimizing the following cost function:

$$
\begin{aligned}
J\left(\theta \mid \theta^s\right) \\
= \frac{1}{N}\Bigg[& \sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{\mathrm{obs}}, \Theta^s\right) \\
& \times \Bigg(\sum_{t=m_1}^{m_\beta} \left(\left(G\left(z^{-1}\right)u_{t-d} - G^s\left(z^{-1}\right)u_{t-d}\right)^2 + \left(\sigma^s\right)^2 \right) \\
& + \sum_{t=t_1}^{t_\alpha} \left(y_t - G\left(z^{-1}\right)u_{t-d}\right)^2 \Bigg) \Bigg].
\end{aligned}
\tag{16}
$$

Here, we introduce the variable x_t^τ denoting the noise-free output with time-delay τ. Based on (1) and (2), we have

$$
x_t^\tau = \frac{\sum_{i=1}^{n_b} b_i z^{-i}}{1 + \sum_{i=1}^{n_a} a_i z^{-i}} u_{t-\tau} = \left(\phi_{t-\tau}^\tau\right)^T \theta,
\tag{17}
$$

where $\phi_{t-\tau}^\tau = [-x_{t-1}^\tau, \ldots, -x_{t-n_a}^\tau, u_{t-\tau-1}, \ldots, u_{t-\tau-n_b}]^T$. Therefore, the cost function can be rewritten as

$$
\begin{aligned}
J\left(\theta \mid \theta^s\right) \\
= \frac{1}{N}\Bigg[& \sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{\mathrm{obs}}, \Theta^s\right) \\
& \times \Bigg(\sum_{t=m_1}^{m_\beta} \left(\left(\left(\phi_{t-d}^d\right)^T \theta - \left(x_t^d\right)^s\right)^2 + \left(\sigma^s\right)^2 \right) \\
& + \sum_{t=t_1}^{t_\alpha} \left(y_t - \left(\phi_{t-d}^d\right)^T \theta\right)^2 \Bigg) \Bigg],
\end{aligned}
\tag{18}
$$

where $\left(x_t^d\right)^s = \left(\phi_{t-d}^d\right)^T \theta^s$. However, the cost function (18) cannot be optimized directly due to the unmeasurable $\{x_t^d\}_{t=1,\ldots,N, d=d_1,\ldots,d_2}$. Here, we adopt the auxiliary model principle and the auxiliary model can be constructed based on the estimates obtained in the previous iteration. That is,

$$
\widehat{x}_t^\tau = \frac{\sum_{i=1}^{n_b} \widehat{b}_i z^{-i}}{1 + \sum_{i=1}^{n_a} \widehat{a}_i z^{-i}} u_{t-\tau} = \left(\widehat{\phi}_{t-\tau}^\tau\right)^T \widehat{\theta},
\tag{19}
$$

where $\widehat{\phi}_{t-\tau}^\tau = [-\widehat{x}_{t-1}^\tau, \ldots, -\widehat{x}_{t-n_a}^\tau, u_{t-\tau-1}, \ldots, u_{t-\tau-n_b}]^T$. Therefore, the cost function (18) with $\phi_{t-\tau}^\tau$ substituted by $\widehat{\phi}_{t-\tau}^\tau$ can be optimized by using the damped Newton algorithm,

$$
\theta^{s+1} = \theta^s - [R^s]^{-1} J'\left(\theta^s \mid \theta^s\right),
\tag{20}
$$

where

$$
\begin{aligned}
J'\left(\theta^s \mid \theta^s\right) \\
= \frac{2}{N}\Bigg[& \sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{\mathrm{obs}}, \Theta^s\right) \\
& \times \Bigg(\sum_{t=m_1}^{m_\beta} \widehat{\phi}_{t-d}^d \left(\left(\widehat{\phi}_{t-d}^d\right)^T \theta^s - \left(x_t^d\right)^s \right) \\
& + \sum_{t=t_1}^{t_\alpha} \widehat{\phi}_{t-d}^d \left(\left(\widehat{\phi}_{t-d}^d\right)^T \theta^s - y_t \right) \Bigg) \Bigg],
\end{aligned}
$$

$$
R^s = \frac{2}{N}\Bigg[\sum_{d=d_1}^{d_2} p\left(\tau = d \mid C_{\mathrm{obs}}, \Theta^s\right) \sum_{t=1}^{N} \widehat{\phi}_{t-d}^d \left(\widehat{\phi}_{t-d}^d\right)^T \Bigg] + \lambda I,
\tag{21}
$$

where λ is a constant.

The time-delay τ can be selected as the delay in the range with maximal posterior probability. That is,

$$
\widehat{\tau} = \underset{d}{\operatorname{argmax}}\, p\left(\tau = d \mid C_{\mathrm{obs}}, \Theta^s\right).
\tag{22}
$$

The E-step and M-step alternate until the convergence condition of the GEM algorithm is met.

4. Simulation Examples

4.1. A Numerical Simulation Example. Consider the following LTI time-delay system described by the OE time-delay model:

$$
y_t = \frac{0.3z^{-1}}{1 - 0.7z^{-1}} u_{t-3} + e_t.
\tag{23}
$$

The input data and output data are generated by simulation and the noise e_t with zero mean and variance 0.01 is added to the output. The input and output data are shown in Figure 1. In the simulation, 12.5% output data are randomly missing. The parameter range of the time-delay is set to [1, 5]. The method proposed in this paper is used to estimate the parameters and the time-delay. The parameter estimate trajectories of the model parameters and the noise variance are shown in Figures 2 and 3, respectively. The estimated time-delay is 3 which is consistent with the true time-delay. To further verify the effectiveness of the proposed method, the simulations are also performed with 25% output data missing and 50% output data missing. The estimated parameters after 13 iterations are summarized in Table 1. It can be seen from these figures and the table that the proposed GEM algorithm has a good identification performance.

4.2. The Continuous Stirred Tank Reactor. The Continuous Stirred Tank Reactor (CSTR) is a benchmark example used to test the performances of different modeling and control
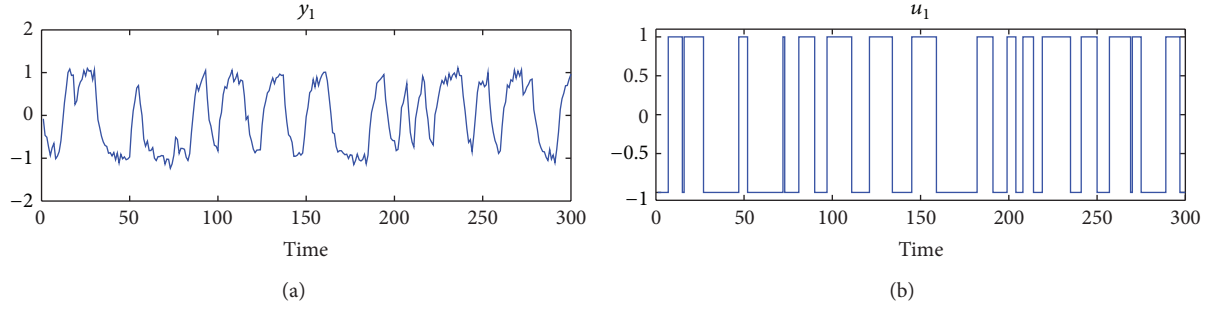
FIGURE 1: The input and output data.

TABLE 1: Estimated parameters after 13 iterations.

True value	$a = -0.7$	$b = 0.3$	$\tau = 3$	$\sigma^2 = 0.01$
Proportion of missing output	a	b	τ	σ^2
Full data set	−0.695	0.3026	3	0.0114
12.5%	−0.694	0.3042	3	0.0116
25%	−0.7012	0.2994	3	0.0121
50%	−0.6979	0.3017	3	0.0124

FIGURE 3: The estimated noise variance in each iteration.

FIGURE 2: The estimated model parameters in each iteration.

algorithms and the first principle model of the CSTR is described as [14]

$$\frac{dC_A}{dt} = \frac{q}{V}\left(C_{A0} - C_A\right) - k_0 C_A \exp\left(\frac{-E}{RT}\right),$$

$$\frac{dT}{dt} = \frac{q}{V}\left(T_0 - T\right) - \frac{(\Delta H)\,k_0 C_A}{\rho C_p} \exp\left(\frac{-E}{RT}\right)$$

$$+ \frac{\rho_c C_{pc}}{\rho C_p V} q_c \left\{1 - \exp\left(\frac{-hA}{q_c \rho C_p}\right)\right\}\left(T_{c0} - T\right),$$

(24)

where the product concentration C_A and the temperature T are output variables and the coolant flow rate q_c is the input

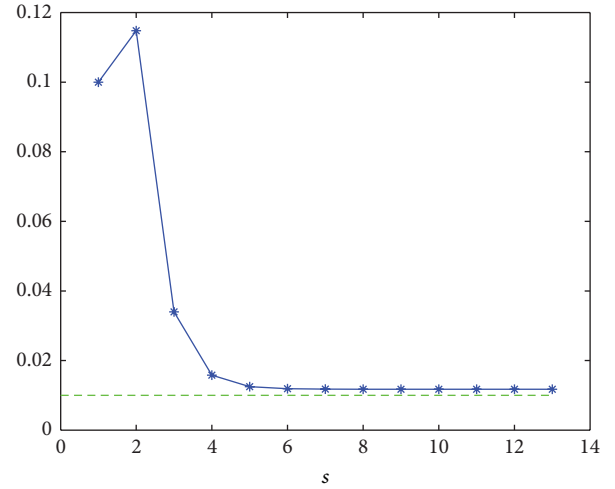variable. The steady state values of the process variables can be found in Gopaluni [14]. In this simulation, the CSTR is operated at a steady state working point which is at $q_c = 97$ L/min, $C_A = 0.0795$ mol/L, and $T = 443.4566$ K. The task here is to build a first-order model between C_A and q_c. The input and output data are generated through simulation and the noise with zero mean and variance 1×10^{-7} is added to the output data. Since time is needed to measure the concentration C_A, so the measurement delay with 1.5 minutes is also added to the output data. The input and output data is shown in Figure 4. In this simulation, 25% output data are randomly missing and the parameter range of the time-delay is set to $[0.9, 2.1]$. The proposed GEM algorithm is used to estimate the unknown parameters. The estimated parameters are $\hat{a} = -0.468$, $\hat{b} = 0.0015$, $\hat{\sigma}^2 = 1.5 \times 10^{-7}$, and $\hat{\tau} = 1.5$. The self-validation and the cross-validation results are shown in Figures 5 and 6. It can be seen from these results that the proposed method has a good identification performance and the estimated model can capture the dynamic behavior of the CSTR.
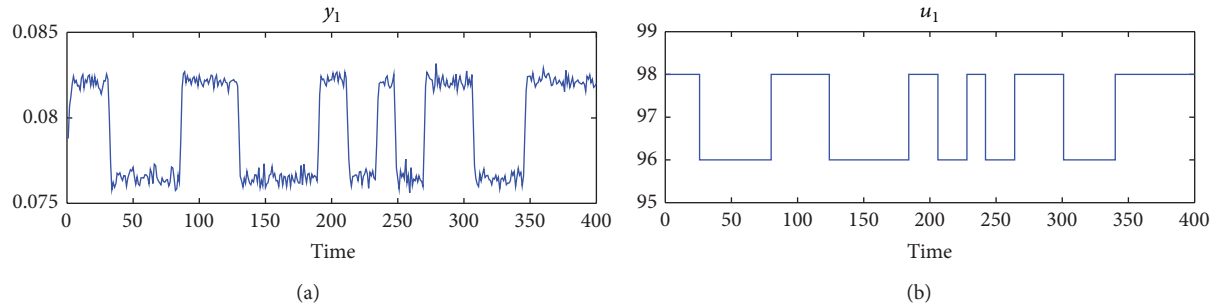
(a)

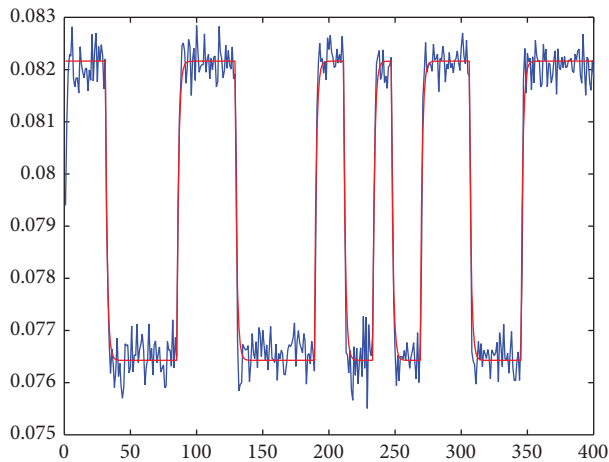(b)

Figure 4: The input and output data.

Figure 5: The self-validation results. The blue line is the real process data and the red line is the simulated output of the estimated model.
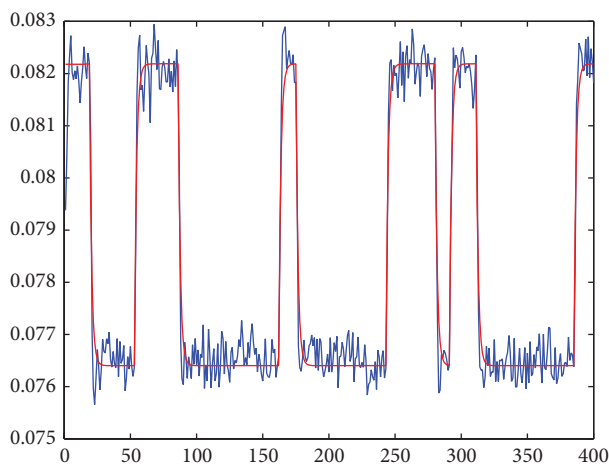

Figure 6: The cross-validation results. The blue line is the real process data and the red line is the simulated output of the estimated model.

5. Conclusion

This paper considers the identification problem of LTI systems with irregular data set. The time-delay and the missing data are commonly encountered problems in process industry and the existence of these problems makes the process modeling a challenging task. The identification problem with incomplete data set in the presence of time-delay is formulated under the scheme of the GEM algorithm and the model parameters and the time-delay are estimated simultaneously in this algorithm. Numerical examples are presented to demonstrate the efficacy of the proposed method.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] S. Yin, H. Luo, and S. X. Ding, "Real-Time implementation of fault-tolerant control systems with performance optimization," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 5, pp. 2402–2411, 2014.

[2] S. Yin, S. X. Ding, A. Haghani, and P. Zhang, "A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process," *Journal of Process Control*, vol. 22, no. 9, pp. 1567–1581, 2012.

[3] S. Yin, X. Yang, and H. R. Karimi, "Data-driven adaptive observer for fault diagnosis," *Mathematical Problems in Engineering*, vol. 2012, Article ID 832836, 21 pages, 2012.

[4] S. Yin, S. X. Ding, A. H. A. Sari, and H. Hao, "Data-driven monitoring for stochastic systems and its application on batch process," *International Journal of Systems Science*, vol. 44, no. 7, pp. 1366–1376, 2013.

[5] S. Yin, G. Wang, and H. Karimi, "Data-driven design of robust fault detection system for wind turbines," *Mechatronics*, 2013.

[6] H. Dong, Z. Wang, and H. Gao, "Distributed H_∞ filtering for a class of Markovian jump nonlinear time-delay systems over lossy sensor networks," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 10, pp. 4665–4672, 2013.

[7] H. Dong, Z. Wang, J. Lam, and H. Gao, "Fuzzy-model-based robust fault detection with stochastic mixed time delays and successive packet dropouts," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 42, no. 2, pp. 365–376, 2012.

[8] J. P. Richard, "Time-delay systems: an overview of some recent advances and open problems," *Automatica*, vol. 39, no. 10, pp. 1667–1694, 2003.

[9] Q. G. Wang and Y. Zhang, "Robust identification of continuous systems with dead-time from step responses," *Automatica*, vol. 37, no. 3, pp. 377–390, 2001.

[10] E. Weyer, "System identification of an open water channel," *Control Engineering Practice*, vol. 9, no. 12, pp. 1289–1299, 2001.

[11] F. Ding and J. Ding, "Least-squares parameter estimation for systems with irregularly missing data," *International Journal of Adaptive Control and Signal Processing*, vol. 24, no. 7, pp. 540–553, 2010.

[12] Y. C. Zhu, H. Telkamp, J. H. Wang, and Q. L. Fu, "System identification using slow and irregular output samples," *Journal of Process Control*, vol. 19, no. 1, pp. 58–67, 2009.

[13] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, New York, NY, USA, 2007.

[14] R. B. Gopaluni, "A particle filter approach to identification of nonlinear processes under missing observations," *The Canadian Journal of Chemical Engineering*, vol. 86, no. 6, pp. 1081–1092, 2008.