Estimation of ARX models with time varying time delays

by

Yujia Zhao

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Process Control

Department of Chemical and Materials Engineering
University of Alberta

Abstract

Processes in industry usually encounter time varying time delays as well as outliers in measurement data. These make identification of the process a challenging problem. Thus, a reliable estimation of the time delay and a correct estimation of the noise to include outliers are essential to efficient process identification.

In this thesis, time-varying delay is modeled by a Markov chain in order to reflect the correlation between any consecutive delay values. To deal with this problem, two approaches are considered: off-line parameter estimation (batch estimation) and on-line adaptive parameter estimation (recursive estimation). Two statistical frameworks, i.e., the expectation-maximization (EM) algorithm and a full-Bayesian estimation method named as variational Bayesian (VB), are investigated to model the time delay processes. Normally distributed measurement noise is modeled by the Gaussian distribution in the proposed method, while in the presence of large random noises, the robustness of the proposed algorithms is enhanced by modeling the noise as t-distributions. During the iterative estimation procedure, outlying observations are down-weighted by a latent variable of the t-distribution automatically, and hence, minimizing their adverse influence on identification.

The proposed algorithms are verified by simulations and experiments. Finally, models based on the proposed algorithms are identified to effectively predict the production rate for the time-delay extraction process used in the oil sands industry.

Acknowledgements

This thesis would not have been possible without the help and support of the kind people around me. I would like to thank all those who have contributed to this thesis and because of whom my graduate experience has been one that I will cherish forever.

Above all, I acknowledge the support and help from my Professor Dr. Biao Huang for his constant guidance and patience which enabled me at all times to explore my way in the excellent Computer Process Control (CPC) group. His constant encouragement helped me to carry on when I encountered difficulties. His rigorous attitude towards academic work inspired me. I am also thankful to him for giving me opportunities to realize my ideas as well as gain valuable industrial project experience. My special thanks will also go to Dr. Alireza Fatehi who has supported me in both my research and project. Without Dr. Alireza's careful instructions and helpful advice, I could have hardly completed this thesis. In addition, I appreciate the help from Shekhar Sharma who spent much time and energy to improve the manuscript of the thesis.

I have met many exceptional members in our CPC group where we can ask questions, share ideas and discuss solutions. I would like to thank Yaojie Lu, Ming Ma, Ruomu Tan, Ouyang Wu and Rishik Ranjan for their help and suggestions when I felt confused during my research. Support from Elom Domlan (APC engineer at Suncor Energy Inc.) during the industrial project is greatly appreciated. I would also give my sincere thanks to my colleagues who helped me to broaden my horizon and stimulate ideas, they are: Jiusun Zeng, Weili Xiong, Ouguan Xu, Ruben Gonzalez, Nima Sammaknejad, Mohammad Rashed, Elham Naghoosi, Anahita Sadeghian, Fadi Ibrahim, Rahul Raveendran, Shabnam Sedghi, and many others from the CPC group. I would also thank my friends for their accompanying and encouragement: Xuanyuan Yin, Su Liu, Ya Gu, Yanjun Liu, Ruoxia Li, Wenhan Shen, Zheyuan Liu, Yanjun Ma, Xiaodong Xu, Yuan Yuan, Liu Liu, Jing Zhang, Chao Shang, Qie Liu, Xiaofeng Yuan, Qingchao Jiang and many others.

This thesis would not have been possible without the financial support from NSERC of Canada and Alberta Innovates Technology Futures. Finally, I owe my deepest gratitude to

my dear parents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Recently, advanced process control (APC) strategies have been developing rapidly to meet the increasing requirements of complex industrial operations. Most of the APC strategies are model-based, resulting in the prerequisite of an accurate and compact mathematical description of the process. Hence, system identification plays an important role in industrial process control.

Industrial processes usually have time delay, fixed or varying, which should be considered in the modeling. Traditional methods assume the delay is a fixed value and it can be treated as a parameter in the identification problem [1, 2]. Hence, the delay estimation is often solved by maximum a posteriori (MAP) [3] or maximum likelihood estimation (MLE) [4, 5] along with the model parameters. However, the delay is usually associated with some process variable transmission (e.g. liquid flow rate). Higher flow rate results in smaller time delay while lower flow rate results in longer time delay. Thus, varying delay is more reasonable in most situations. When considering a varying delay, a separate distribution or model should be used to describe it. Xie et al. [6] assumed a uniform distribution for the delay, which means that the delay value varies randomly among some presumed values with the same probability. In this thesis, it is proposed that the switching mechanism of time delay follows a Markov chain. The transition of delay from one value to another is governed by a probability.

Industrial processes are usually time varying, because of aging, switch between operating points, changes of the raw material composition or the requested product material properties. A recursive parameter estimation algorithm is required when the process is time varying to capture the trend of the change in model parameters [7, 8]. In this thesis, we propose a recursive version of the EM algorithm which can update the model parameter

estimates for the time delay problem.

Industrial data is often noisy or contaminated by outliers. Common reasons that can cause outliers in recorded data include transmission errors, process disturbances, and instrument degradation [9, 10]. Data-driven modeling methods are usually sensitive to outliers and resulting models may lead to biased parameter estimation and plant-model mismatch. Therefore, modeling of the noise distribution is essential to parameter estimation. T-distribution has the capability of tuning continuously from a very heavy-tailed distribution to a Gaussian distribution by adjusting its degrees of freedom [11, 12]. The effect of outliers on modeling can be diminished by assigning higher probability densities to the tails. In addition, a t-distribution can be represented by an infinite mixture of scaled Gaussian distributions, which is an important property in statistical modeling.

This thesis focuses on recursive and robust estimation of time delay processes under two statistical frameworks (EM and VB). The proposed algorithms are validated by simulation examples and pilot scale experiments. Models based on the proposed algorithms are designed to predict the production rate in an oil sands industrial case study.

1.2 Thesis contributions

The main contribution of this thesis is the development of time delay process identification methods with robustness. The proposed algorithms are resistant to outliers and result in improved accuracy and reliability of process modeling and prediction. Specifically, the contributions of this thesis are summarized as follows:

1. Modeled the time-delay processes using the Hidden Markov Model (HMM).

2. Developed a recursive EM method to update the model parameters.

3. Integrated t-distributions with the expectation-maximization (EM) algorithm and variational Bayesian (VB) approach, and made the algorithm down-weight outlying observations automatically.

4. The distributions of parameters were estimated by the Bayesian approach, and the uncertainty of parameters was taken into account.

5. Used designed time delay models to estimate the model parameters to predict production rate in an oil sands industrial case study.

1.3 Thesis outline

This thesis is organized as follows:

In Chapter 2, we develop a batch estimation algorithm for the parameter invariant time delay process and a recursive estimation algorithm for the parameter variant time delay process. The identification problem is formulated and solved under the EM framework.

In Chapter 3, we deal with time delay process in the presence of noisy operational data under the EM framework.

In Chapter 4, we still deal with robust estimation of the time delay process. A variational Bayesian identification approach is developed in this chapter.

In Chapter 5, we construct a model with variable time delay to predict production rate in an oil sands industrial case study.

Chapter 6 summarizes the main results of this thesis and discusses future research directions.

Chapter 2

ARX Model Estimation with Time Varying Time Delays

Output time delay is often encountered in industrial processes. In this chapter, we consider a class of output time delays that can change at every sample. The mechanism of the varying time delays is modeled by Markov Chain. Both time invariant and time variant model parameters are considered. The former is solved by expectation maximization algorithm (EM) while the latter is solved by recursive EM algorithm. The proposed identification is demonstrated by simulation examples as well as by pilot-scale experiments.

2.1 Introduction

Time varying properties of industrial processes pose a challenge for system identification. Both, the model structure and parameters can vary with time. Delay variation is among well-known structural time variations in the process plants. Almost all industrial processes involve transportation of materials. Since the transportation speed varies frequently according to changing flow rates, varying time delay is an inherent characteristic of these processes.

Physical processes can be modeled through first principles. However, this approach requires detailed understanding and is often difficult due to the complexity of the industrial systems. An alternative approach is to construct data-driven models, such as AutoRegressive eXogenous (ARX) models [13, 14], which do not require in depth process know-how. Maximum likelihood estimation(MLE) [15] and maximum a posteriori estimation (MAP) [16] are commonly applied to solve parameter estimation problems. However, when part of the data is not available (e.g. time delays), one often resorts to the expectation-maximization (EM) algorithm [17].

System identification with time delay has been extensively studied. In literature, both constant and varying, but unknown time delay have been studied [18]. Zhang and Li proposed an identification method based on steepest descent algorithm to address varying time delay problems in [19]. Xie et al. applied EM algorithm to identify an FIR model with time varying delays in [6]. In this case, delay was considered as a hidden variable and followed a uniform distribution. However, the delay was treated as a sequence of randomly switching values with no relationship between any two consecutive values. The assumption of a random delay sequence can result in model over estimation.

Markov chain model has been applied for estimation of time-series, identification of time varying systems and many other applications. In [14], Jin and Huang solved a switching system identification problem, in which they proposed that certain behaviors exhibited in the switching dynamics followed a Markov chain. Shengyuan et al. proposed a control method for Markov jump systems with time varying delays [20]. In [21], Kim applied an HMM to the modeling of econometric time-series. Bar-Shalom and Li described the targets tracking problem where the target motions switched via Markov jump systems [22].

Most of the research on time delay identification is limited to time invariant systems. In real-world applications, such as adaptive control, filtering, and prediction, it is essential to address the time varying properties of the processes [23, 24]. In this situation, the batch EM algorithm for parameter estimation is suitable for time invariant systems only. Recursive Least Squares (RLS) method [25] is suitable for modeling linear processes with time varying coefficients, but it cannot deal with the hidden variable problem. In [26], Lang et al. proposed a moving window EM algorithm strategy for parameter estimation in order to reduce sensitivity to possibly unreliable initial parameter values. This method can be used to recursively update model parameters by shifting the fixed size window one sample forward. However, it is computationally expensive because the EM algorithm is run every time a new observation becomes available.

In [27], Titterington proposed recursive EM (REM), which is a stochastic approximation method to update model parameters in the presence of hidden variables. Chung et al. tested Titterington's recursive technique on both constant and time varying parameter models [28]. In [29], Cappe et al. proposed a novel recursive algorithm that is similar to the batch EM algorithm. Ozkan applied this recursive version of the EM algorithm for joint state and mixture measurement noise estimation [30]. However, Cappe's technique is not an iterative method, and hence, cannot make the best use of every sample.

In this chapter, we study the parameter estimation of both time variant and time in-

variant processes in the presence of varying time delay. Up to the authors knowledge, the identification problem with Markov chain as the time delay switching mechanism, has not been studied. In this study, we consider the delay to follow a HMM variation and use ARX as the process model for system identification. Since the exact value of delay is unknown, the parameters of the model and the HMM is estimated using EM algorithm for LTI process and REM algorithm for LTV process. HMM assumption for the varying and unknown time delay will be compared with the fixed delay assumption and the independent delay assumption described in [6]. The proposed iterative version of the recursive EM algorithm is an extension of Cappe's technique with an additional iterative stage to improve the parameter estimation. It is compared with the batch EM, moving window EM and the recursive EM without iteration described in [29]. Comparisons between the proposed methods and the most relevant existing algorithms for the identification of ARX models in terms of their identification performance are conducted using pilot-scale experiments.

The remainder of this chapter is organized as follows. A detailed description of the ARX model identification problem in the presence of time varying time delay is presented in the Section 2. The following Section 3 and Section 4 apply EM Algorithm and recursive EM algorithm to solve the time invariant and variant identification problems, respectively. Section 5 gives two numerical examples to validate the proposed identification algorithm for time invariant system identification and a recursive version of the proposed algorithm for time variant system. In Section 6, experiments are conducted to validate the proposed methods followed by the conclusion in Section 7.

2.2 Problem Statement

In industries, certain variables are determined by laboratory analysis. Since the lab analysis generally has a lower frequency than the online measurement, the process is often dual-rate with both fast and slow rate variables. Meanwhile, time delay depends on factors such as liquid flow rate, which are not constant. Thus, delay is also time varying. Consider the following dual rate ARX model with varying delay:

$$\begin{aligned} y_{T_k} &= \psi_{T_k-\lambda_k}\theta + \nu_{T_k}, \\ \psi_{T_k-\lambda_k} &= \left[\begin{array}{cccccc} y_{T_{k-1}} & \cdots & y_{T_{k-na}} & u_{T_k-\lambda_k} & \cdots & u_{T_k-nb-\lambda_k} \end{array}\right] \in \mathbb{R}^{1\times(na+nb+1)}, \end{aligned} \tag{2.1}$$

where $\{y_{T_k}, k = 1, 2, ..., N\}$ is the slow rate output variable, while $\{u_t, t = 1, 2, ..., L\}$ is the fast rate input variable. The slow rate sampling time is Δ times that of the fast rate ($L/N = \Delta$). na and nb are the output and input orders respectively. $\theta \in \mathbb{R}^{(na+nb+1)\times 1}$ is

the regression parameter vector where $nb+\lambda_k < \Delta$. ν_t is associated measurement noise, and is assumed to follow an i.i.d. Gaussian distribution with zero mean and unknown variance σ^2.

The actual value of the time delay is unknown. However, the time delay sequence $\{\lambda_k, k = 1, 2, ..., N\}$ is modeled by a hidden Markov chain. The Markov property means that the kth instant time delay is only dependent on the $(k-1)th$ instant time delay:

$$P(\lambda_k|\lambda_{k-1}, ..., \lambda_1) = P(\lambda_k|\lambda_{k-1}). \tag{2.2}$$

The transition probability and initial distribution of time delay are denoted by the following two parameters:

$$\begin{array}{l} \alpha_{ij} = P(\lambda_k = i|\lambda_{k-1} = j), k = 2, 3...N, 1 \leq i, j \leq d, \\ \pi_i = P(\lambda_1 = i), 1 \leq i \leq d. \end{array} \tag{2.3}$$

Since the delay is unknown, the identification of the system in Equation 2.1 can be carried out by the EM algorithm. The observed variables, missing variables and parameters to be estimated are denoted as:

$$\begin{array}{l} C_{obs} = \{Y, U\} = \left\{y_{T_N}, y_{T_{N-1}}, \cdots, y_{T_1}, u_{T_N}, u_{T_N-1}, \cdots, u_1\right\}, \\ C_{mis} = \Lambda = \{\lambda_N, \lambda_{N-1}, \cdots, \lambda_1\}, \\ \Theta = \left\{\theta, \sigma^2, \alpha_{ij}, \pi_i\right\}, 1 \leq i, j \leq d. \end{array} \tag{2.4}$$

2.3 Off-line estimation method for the time invariant system

2.3.1 Constructing the Q-function

The EM algorithm constructs the conditional expectation of the complete data likelihood with respect to the missing data (Q function) and maximizes it iteratively using the past parameter estimation, resulting in Maximum Likelihood Estimation (MLE) of the parameters of interest. The Q function is defined as:

$$Q\left(\Theta|\Theta^h\right) = E_{C_{mis}|C_{obs},\Theta^h}\left\{logP\left(C_{obs}, C_{mis}|\Theta\right)\right\}, \tag{2.5}$$

where Θ^h is the parameter estimate from the previous iteration step and E denotes the expectation value. Substituting Equation 2.4 into Equation 2.5 and according to chain rule (general product rule), the Q function is rewritten as:

$$\begin{array}{l} Q\left(\Theta|\Theta^h\right) \\ = E_{\Lambda|Y,U,\Theta^h}\left\{logP\left(Y, U, \Lambda|\Theta\right)\right\} \\ = E_{\Lambda|Y,U,\Theta^h}\left\{log\left[P\left(Y|U, \Lambda, \Theta\right)P\left(\Lambda|U, \Theta\right)C\right]\right\}, \end{array} \tag{2.6}$$

where $C \triangleq P\left(U|\Theta\right)$ is a constant value because the input is deterministic. Given the regressor vector, the output does not depend on future information and is independent of

each other. In addition, time delay only depends on the latest delay according to the HMM property in Equation 2.2. Hence, the log-likelihood of the complete data can be further decomposed to

$$\begin{aligned}
&logP\left(C_{obs}, C_{mis}|\Theta\right) \\
&= log\left[\prod_{k=1}^{N} P\left(y_{T_k}|\psi_{T_k-\lambda_k}, \lambda_k, \Theta\right) \times \prod_{k=2}^{N} P\left(\lambda_k|\lambda_{k-1}, \Theta\right) \times P\left(\lambda_1|\Theta\right) \times C\right] \\
&= \sum_{k=1}^{N} logP\left(y_{T_k}|\psi_{T_k-\lambda_k}, \lambda_k, \Theta\right) + \sum_{k=2}^{N} logP\left(\lambda_k|\lambda_{k-1}, \Theta\right) + logP\left(\lambda_1|\Theta\right) + logC.
\end{aligned} \tag{2.7}$$

To implement the expectation for the Q function, the posterior conditional probability of the delay is utilized. As a result, the final expression for the Q function is

$$\begin{aligned}
&Q(\Theta|\Theta^h) \\
&= \sum_{k=1}^{N}\sum_{i=1}^{d} P\left(\lambda_k = i|y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) logP\left(y_{T_k}|\psi_{T_k-\lambda_k}, \lambda_k = i, \Theta\right) \\
&+ \sum_{k=2}^{N}\sum_{i=1}^{d}\sum_{j=1}^{d} P\left(\lambda_k = i, \lambda_{k-1} = j|y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) logP\left(\lambda_k = i|\lambda_{k-1} = j, \Theta\right) \\
&+ \sum_{i=1}^{d} P\left(\lambda_1 = i|y_{T_1}, u_{T_1:1}, \Theta^h\right) logP\left(\lambda_1 = i|\Theta\right) + logC,
\end{aligned} \tag{2.8}$$

where $P\left(y_{T_k}|\psi_{T_k-\lambda_k}, \lambda_k = i, \Theta\right)$ is calculated by

$$P\left(y_{T_k}|\psi_{T_k-\lambda_k}, \lambda_k = i, \Theta\right) = \frac{1}{\sqrt{2\pi\sigma^2}} exp^{-\frac{1}{2\sigma^2}\left(y_{T_k}-\psi_{T_k-i}\theta\right)^2}, \tag{2.9}$$

Substituting Equations 2.3 and 2.9 into Equation 2.8, the Q function can be decomposed into three parts as follows,

$$Q(\Theta|\Theta^h) = Q1\left(\theta, \sigma^2\right) + Q2\left(\alpha_{ij}\right) + Q3\left(\pi_i\right) + logC, \tag{2.10}$$

where each part corresponds to different parameters as shown below,

$$\begin{aligned}
Q1\left(\theta, \sigma^2\right) &= \sum_{k=1}^{N}\sum_{i=1}^{d} P\left(\lambda_k = i|y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)\left[-log\sqrt{2\pi\sigma^2} - \tfrac{1}{2\sigma^2}\left(y_{T_k} - \psi_{T_k-i}\theta\right)^2\right], \\
Q2\left(\alpha_{ij}\right) &= \sum_{k=2}^{N}\sum_{i=1}^{d}\sum_{j=1}^{d} P\left(\lambda_k = i, \lambda_{k-1} = j|y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \log\left(\alpha_{ij}\right), \\
Q3\left(\pi_i\right) &= \sum_{i=1}^{d} P\left(\lambda_1 = i|y_{T_1}, u_{T_1:1}, \Theta^h\right) \log\left(\pi_i\right).
\end{aligned} \tag{2.11}$$

To calculate $Q(\Theta|\Theta^h)$, the following posterior conditional probabilities are calculated.

1. In $Q3\left(\pi_i\right)$, the posterior probability for the $1st$ time delay

$$P\left(\lambda_1 = i|y_{T_1}, u_{T_1:1}, \Theta^h\right) = \frac{P\left(y_{T_1}|\psi_{T_1-\lambda_1}, \lambda_1 = i, \Theta^h\right)\left(\pi_i\right)^h}{\sum_{m=1}^{d} P\left(y_{T_1}|\psi_{T_1-\lambda_1}, \lambda_1 = m, \Theta^h\right)\left(\pi_m\right)^h}. \tag{2.12}$$

2. In $Q2\left(\alpha_{ij}\right)$, the posterior joint probability for the kth and $(k-1)th$ time delay

$$\begin{aligned} & P\left(\lambda_k = i, \lambda_{k-1} = j | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \\ & = \frac{P\left(y_{T_k} | \psi_{T_k - \lambda_k}, \lambda_k = i, \Theta^h\right) (\alpha_{ij})^h P\left(\lambda_{k-1} = j | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \Theta^h\right)}{\sum_{m=1}^{d} \sum_{n=1}^{d} P\left(y_{T_k} | \psi_{T_k - \lambda_k}, \lambda_k = m, \Theta^h\right) (\alpha_{mn})^h P\left(\lambda_{k-1} = n | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \Theta^h\right)}, \end{aligned} \tag{2.13}$$

where $P\left(\lambda_{k-1} = j | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \Theta^h\right)$ is obtained through the discrete valued state propagation of Markov chain starting from the initial estimation of $P\left(\lambda_1 | y_{T_1}, u_{T_1:1}, \Theta^h\right)$.

3. In $Q1\left(\theta, \sigma^2\right)$, the posterior probability for the kth time delay

$$P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) = \sum_{j=1}^{d} P\left(\lambda_k = i, \lambda_{k-1} = j | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right). \tag{2.14}$$

Details of derivation of Equations 2.12 and 2.13 are presented in Appendix A.

2.3.2 Maximizing the Q-function

A new estimate of the parameters $\Theta = \{\theta, \sigma^2, \alpha_{ij}, \pi_i\}$ is calculated by taking the derivative of the Q function over the corresponding parameters and equating it to zero. Therefore, by taking the derivative of $Q1\left(\theta, \sigma^2\right)$ with respect to θ and σ^2, we obtain

$$\begin{aligned} \theta^{h+1} = & \left\{ \sum_{k=1}^{N} \sum_{i=1}^{d} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) {\psi_{T_k - i}}^T \psi_{T_k - i} \right\}^{-1} \\ & \left\{ \sum_{k=1}^{N} \sum_{i=1}^{d} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) {\psi_{T_k - i}}^T y_{T_k} \right\} \end{aligned} \tag{2.15}$$

and

$$\begin{aligned} \left(\sigma^2\right)^{h+1} & = \frac{\sum_{k=1}^{N} \sum_{i=1}^{d} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \left(y_{T_k} - \psi_{T_k - i} \theta^{h+1}\right)^2}{\sum_{k=1}^{N} \sum_{i=1}^{d} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)} \\ & = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{d} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \left(y_{T_k} - \psi_{T_k - i} \theta^{h+1}\right)^2. \end{aligned} \tag{2.16}$$

When conducting the computation of α_{ij} and π_i, we need to consider the following constraints

$$\sum_{j=1}^{d} \alpha_{ij} = 1, \sum_{i=1}^{d} \pi_i = 1. \tag{2.17}$$

Introducing Lagrange multipliers L_α and L_π, and taking the derivative of $\left\{ Q2\left(\alpha_{ij}\right) + L_\alpha \left(\sum_{j=1}^{d} \alpha_{ij} - 1 \right) \right\}$ with respect to α_{ij} and L_α, and $\left\{ Q3\left(\pi_i\right) + L_\pi \left(\sum_{i=1}^{d} \pi_i - 1 \right) \right\}$ with respect to π_i and L_π,

Table 2.1: Expectation and Maximization steps

Initialization. Set $h = 0$. Assign random values to Θ^h.
Do{
E-step: Evaluate $P\left(\lambda_1 = i | y_{T_1}, u_{T_1:1}, \Theta^h\right)$ by Eqn.2.12;
For $k = 2 : N$, {
Evaluate $P\left(\lambda_k = i, \lambda_{k-1} = j | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)$ by Eqn.2.13;
Evaluate $P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)$ by Eqn.2.14}.
M-step: Evaluate $\Theta^{h+1} = \{\theta, \sigma^2, \alpha_{ij}, \pi_i\}$ by Eqn.s 2.15, 2.16, 2.18, 2.19;
} while $\left(\frac{\|\Theta^{h+1} - \Theta^h\|^2 - \|\Theta^h\|^2}{\|\Theta^h\|^2} < 0.001\right)$.

we obtain

$$(\alpha_{ij})^{h+1} = \frac{\sum_{k=2}^{N} P\left(\lambda_k = i, \lambda_{k-1} = j | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)}{\sum_{k=2}^{N} \sum_{j=1}^{d} P\left(\lambda_k = i, \lambda_{k-1} = j | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)} \tag{2.18}$$

and

$$(\pi_i)^{h+1} = \frac{P\left(\lambda_1 = i | y_{T_1}, u_{T_1:1}, \Theta^h\right)}{\sum_{m=1}^{d} P\left(\lambda_1 = m | y_{T_1}, u_{T_1:1}, \Theta^h\right)} = P\left(\lambda_1 = i | y_{T_1}, u_{T_1:1}, \Theta^h\right). \tag{2.19}$$

The expectation and maximization steps continue until parameter convergence, resulting in the ML estimate of the parameters. The calculation procedure is shown in Table 2.1.

2.4 On-line estimation method for the time variant system

Industrial processes are usually time variant, and hence the model parameters are also functions of time. Therefore, a recursive parameter estimation algorithm should be developed to re-estimate parameters when new data is available.

The problem of recursive parameter estimation under missing data was first studied in [27]. Titterington proposed the following stochastic approximation:

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma_{n+1} I^{-1}\left(\hat{\theta}_n\right) \nabla_\theta log\, f\left(Y_{n+1}; \hat{\theta}_n\right), \tag{2.20}$$

where $\{\gamma_n\}$ is a decreasing sequence of positive step size, $I(\theta)$ is the Fisher Information Matrix (FIM), ∇_θ is the first order derivative operator over θ and $\hat{\theta}_{n+1}$ is the new estimate using the new observation Y_{n+1}. However, $I(\theta)$ is not always guaranteed to be positive definite [29] and makes the algorithm unreliable.

Recently, Cappe and Moulines [29] proposed a recursive EM algorithm for latent data models with independent observations. The main advantage of this approach to recursive

parameter estimation in latent data models is its analogy with the standard batch EM algorithm, which makes the recursive algorithm easy to implement. Based on this algorithm, we propose an iterative version of the recursive EM algorithm, which makes better use of every data point and has more accurate parameter estimation results.

2.4.1 Identification using recursive EM algorithm

The basic idea of recursive EM algorithm [29] is to replace the expectation step by a stochastic approximation step, while keeping the maximization step unchanged. Consider the recursive Q-function

$$\hat{Q}_{n+1}(\theta) = \hat{Q}_n(\theta) + \gamma_{n+1}\left(E_{\hat{\theta}_n}\left[log\, f(X_{n+1};\theta)\,|Y_{n+1}\right] - \hat{Q}_n(\theta)\right), \tag{2.21}$$

where Y_{n+1} is the observation at the $(n+1)th$ time instant and X_{n+1} is the complete data including both observed and unobserved data. In this formula, the expectation of log distribution of a new data point is with respect to the hidden variables given $\hat{\theta}_n$, where $\hat{\theta}_n$ is the updated parameter for observation Y_n. Thus, it is known and unchanged at time step $n+1$.

Applying this algorithm to the time delay problem described in Section 2, we have

$$\hat{Q}_{n+1}(\Theta) = \hat{Q}_n(\Theta) + \gamma_{n+1}\left(E_{\hat{\Theta}_n}\left[\log P\left(y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}, \lambda_{n+1}, \lambda_n; \Theta\right)\left|y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}\right.\right] - \hat{Q}_n(\Theta)\right), \tag{2.22}$$

where the posterior distribution of new missing data is computed using $\hat{\Theta}_n$, which is the latest parameter estimate.

The Q function can be further derived as

$$\begin{aligned}
&\hat{Q}_{n+1}(\Theta)\\
&= (1-\gamma_{n+1})\,\hat{Q}_n(\Theta) + \gamma_{n+1}E_{\hat{\Theta}_n}\left[\log P\left(y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}, \lambda_{n+1}, \lambda_n; \Theta\right)\left|y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}\right.\right]\\
&\vdots\\
&= \prod_{p=2}^{n+1}(1-\gamma_p)E_{\Theta_0}\left[\log P\left(y_{T_1}, \psi_{T_1-\lambda_1}, \lambda_1; \Theta\right)|y_{T_1}, \psi_{T_1-\lambda_1}\right]\\
&+ \sum_{k=2}^{n}\prod_{p=k+1}^{n+1}(1-\gamma_p)\gamma_k E_{\hat{\Theta}_{k-1}}\left[\log P\left(y_{T_k}, \psi_{T_k-\lambda_k}, \lambda_k, \lambda_{k-1}; \Theta\right)|y_{T_k}, \psi_{T_k-\lambda_k}\right]\\
&+\gamma_{n+1}E_{\hat{\Theta}_n}\left[\log P\left(y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}, \lambda_{n+1}, \lambda_n; \Theta\right)\left|y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}\right.\right].
\end{aligned} \tag{2.23}$$

The three log-likelihood terms in the Q function can be decomposed by the chain rule to be

$$\begin{aligned}
&\log P\left(y_{T_1}, \psi_{T_1-\lambda_1}, \lambda_1; \Theta\right) = \log P\left(y_{T_1}|\psi_{T_1-\lambda_1}, \lambda_1; \Theta\right) + \log P\left(\lambda_1; \Theta\right) + C_1,\\
&\log P\left(y_{T_k}, \psi_{T_k-\lambda_k}, \lambda_k, \lambda_{k-1}; \Theta\right) = \log P\left(y_{T_k}|\psi_{T_k-\lambda_k}, \lambda_k; \Theta\right) + \log P\left(\lambda_k|\lambda_{k-1}; \Theta\right) + C_k,\\
&\log P\left(y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}, \lambda_{n+1}, \lambda_n; \Theta\right) = \log P\left(y_{T_{n+1}}|\psi_{T_{n+1}-\lambda_{n+1}}, \lambda_{n+1}; \Theta\right) + \log P\left(\lambda_{n+1}|\lambda_n; \Theta\right) + C_{n+1}.
\end{aligned} \tag{2.24}$$

where we use the property that λ_{k-1} does not change at time k, i.e. $P(\lambda_{k-1};\Theta)$ is independent of Θ and can be considered to be constant. The final expression for the Q function is divided into three parts, each corresponding to different parameters, as shown below,

$$\hat{Q}_{n+1}(\Theta) = \hat{Q}4_{n+1}\left(\theta,\sigma^2\right) + \hat{Q}5_{n+1}(\alpha_{ij}) + \hat{Q}6_{n+1}(\pi_i) + C_\Theta, \tag{2.25}$$

where $\hat{Q}4_{n+1}\left(\theta,\sigma^2\right), \hat{Q}5_{n+1}(\alpha_{ij}), \hat{Q}6_{n+1}(\pi_i)$ and C_Θ are as follows,

1. The term related to θ, σ^2

$$\begin{aligned}
&\hat{Q}4_{n+1}\left(\theta,\sigma^2\right)\\
&= \prod_{p=2}^{n+1}(1-\gamma_p)\sum_{i=1}^{d} P(\lambda_1 = i|y_{T_1},\psi_{T_1-\lambda_1};\Theta_0)\log P\left(y_{T_1}|\psi_{T_1-\lambda_1},\lambda_1 = i;\Theta\right)\\
&+\sum_{k=2}^{n}\prod_{p=k+1}^{n+1}(1-\gamma_p)\gamma_k\sum_{i=1}^{d} P(\lambda_k = i|y_{T_k},\psi_{T_k-\lambda_k};\hat{\Theta}_{k-1})\log P\left(y_{T_k}|\psi_{T_k-\lambda_k},\lambda_k = i;\Theta\right)\\
&+\gamma_{n+1}\sum_{i=1}^{d} P\left(\lambda_{n+1} = i|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};\hat{\Theta}_n\right)\log P\left(y_{T_{n+1}}|\psi_{T_{n+1}-\lambda_{n+1}},\lambda_{n+1} = i;\Theta\right),
\end{aligned} \tag{2.26}$$

where

$$logP\left(y_{T_k}|\psi_{T_k-\lambda_k},\lambda_k = i;\Theta\right) = -log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}[y_{T_k} - \psi_{T_k-i}\theta]^2$$

2. The term related to α_{ij}

$$\begin{aligned}
&\hat{Q}5_{n+1}(\alpha_{ij})\\
&= \sum_{k=2}^{n}\prod_{p=k+1}^{n+1}(1-\gamma_p)\gamma_k\sum_{i=1}^{d}\sum_{j=1}^{d} P\left(\lambda_k = i,\lambda_{k-1} = j|y_{T_k},\psi_{T_k-\lambda_k};\hat{\Theta}_{k-1}\right) logP\left(\lambda_k = i|\lambda_{k-1} = j;\Theta\right)\\
&+\gamma_{n+1}\sum_{i=1}^{d}\sum_{j=1}^{d} P\left(\lambda_{n+1} = i,\lambda_n = j|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};\hat{\Theta}_n\right) logP\left(\lambda_{n+1} = i|\lambda_n = j;\Theta\right),
\end{aligned} \tag{2.27}$$

where

$$logP\left(\lambda_k = i|\lambda_{k-1} = j;\Theta\right) = log\alpha_{ij}$$

3. The term related to π_i

$$\hat{Q}6_{n+1}(\pi_i) = \prod_{p=2}^{n+1}(1-\gamma_p)\sum_{i=1}^{d} P\left(\lambda_1 = i|y_{T_1},\psi_{T_1-\lambda_1};\Theta_0\right) logP\left(\lambda_1 = i;\Theta\right), \tag{2.28}$$

where

$$logP\left(\lambda_1 = i;\Theta\right) = log\pi_i$$

4. C_Θ is a constant

For each parameter in $\hat{\Theta}_{n+1}$, the derivative can be formulated separately and set equal to zero. The detailed derivation is provided in Appendix B and the estimation results are as follows,

1. For the regressor parameter, by solving the first order derivative of $\hat{Q}4_{n+1}\left(\theta,\sigma^2\right)$ with respect to θ, we can write the regression parameter estimate, as follows

$$\hat{\theta}_{n+1} = \left(\hat{\theta}_{n+1}\right)_{.den}^{-1}\left(\hat{\theta}_{n+1}\right)_{.num}, \tag{2.29}$$

where the numerator vector and denominator matrix are

$$\begin{aligned}\left(\hat{\theta}_{n+1}\right)_{.den} &= (1-\gamma_{n+1})\left(\hat{\theta}_{n}\right)_{.den} + \gamma_{n+1}\sum_{i=1}^{d} P\left(\lambda_{n+1}=i|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};\hat{\Theta}_n\right)\psi_{T_{n+1}-i}{}^{T}\psi_{T_{n+1}-i},\\ \left(\hat{\theta}_{n+1}\right)_{.num} &= (1-\gamma_{n+1})\left(\hat{\theta}_{n}\right)_{.num} + \gamma_{n+1}\sum_{i=1}^{d} P\left(\lambda_{n+1}=i|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};\hat{\Theta}_n\right)\psi_{T_{n+1}-i}{}^{T}y_{T_{n+1}}.\end{aligned} \tag{2.30}$$

The $\left(\hat{\theta}_{n}\right)_{.num}$ and $\left(\hat{\theta}_{n}\right)_{.den}$ are the numerator and denominator used to obtain $\hat{\theta}_n$ in the previous step.

2. For the noise variance, by solving the first order derivative of $\hat{Q}4_{n+1}\left(\theta,\sigma^2\right)$ with respect to σ^2, we obtain

$$\hat{\sigma}^2_{n+1} = (1-\gamma_{n+1})\,\hat{\sigma}^2_n + \gamma_{n+1}\sum_{i=1}^{d} P\left(\lambda_{n+1}=i|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};\hat{\Theta}_n\right)\left[y_{T_{n+1}}-\psi_{T_{n+1}-i}\hat{\theta}_{n+1}\right]^2. \tag{2.31}$$

3. For the transition probability, by solving the first order derivative of $\hat{Q}5_{n+1}\left(\alpha_{ij}\right)$ with respect to α_{ij} and considering the constraint $\left\{\sum_{j=1}^{d}\alpha_{ij}=1\right\}$, we obtain

$$\left(\hat{\alpha}_{ij}\right)_{n+1} = \frac{\left(\hat{\alpha}_{ij}\right)_{n+1.num}}{\left(\hat{\alpha}_{ij}\right)_{n+1.den}}, \tag{2.32}$$

where the numerator and denominator are

$$\begin{aligned}\left(\hat{\alpha}_{ij}\right)_{n+1.num} &= (1-\gamma_{n+1})\left(\hat{\alpha}_{ij}\right)_{n.num} + \gamma_{n+1}P\left(\lambda_{n+1}=i,\lambda_n=j|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};\hat{\Theta}_n\right),\\ \left(\hat{\alpha}_{ij}\right)_{n+1.den} &= (1-\gamma_{n+1})\left(\hat{\alpha}_{ij}\right)_{n.den} + \gamma_{n+1}\sum_{j=1}^{d}\left(\lambda_{n+1}=i,\lambda_n=j|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};\hat{\Theta}_n\right).\end{aligned} \tag{2.33}$$

The $\left(\hat{\alpha}_{ij}\right)_{n.num}$ and $\left(\hat{\alpha}_{ij}\right)_{n.den}$ are the numerator and denominator used to obtain $\left(\hat{\alpha}_{ij}\right)_{n}$.

4. Since the estimate for π_i plays no role in the recursive algorithm, there is no need to derive the recursive formula for this parameter.

To evaluate the above parameter estimation, the following two posterior terms should be determined,

1. The joint posterior distribution required to estimate α_{ij} in Equation 2.32 is calculated as follows,

$$P\left(\lambda_{n+1}=i,\lambda_n=j|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};\hat{\Theta}_n\right)=\frac{P\left(y_{T_{n+1}}|\psi_{T_{n+1}-\lambda_{n+1}},\lambda_{n+1}=i;\hat{\Theta}_n\right)\left(\hat{\alpha}_{ij}\right)_n\pi_j^n}{\sum_{m=1}^{d}\sum_{l=1}^{d}P\left(y_{T_{n+1}}|\psi_{T_{n+1}-\lambda_{n+1}},\lambda_{n+1}=m;\hat{\Theta}_n\right)\left(\hat{\alpha}_{ml}\right)_n\pi_l^n} \tag{2.34}$$

   where

$$\pi_j^n=P\left(\lambda_n=j;\hat{\Theta}_n\right)=P\left(\lambda_n=j|y_{T_n},\psi_{T_n-\lambda_n};\hat{\Theta}_{n-1}\right)$$

   is the posterior distribution of missing time delay λ_n, calculated in the previous recursion. The parameters of this distribution do not change at time $n+1$, so π_j^n is fixed for the current time instant.

2. The posterior distribution required to estimate θ, σ^2 in Equations 2.29, 2.31 is calculated as follows,

$$P\left(\lambda_{n+1}=i|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};\hat{\Theta}_n\right)=\sum_{j=1}^{d}P\left(\lambda_{n+1}=i,\lambda_n=j|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};\hat{\Theta}_n\right). \tag{2.35}$$

From the overall derivation, it is noticed that the recursive estimation of parameters can decrease the computational complexity. This is because, in the E step, we do not need to evaluate the Q function using historical data, and in the M step, the obtained estimator is a simple updating of the previous estimate. Another advantage is that, while parameters change, REM responds faster because it "forgets" the effect of old information by a factor of γ_{n+1}.

2.4.2 Iterative recursive EM algorithm

In Cappe & Moulines' recursive formula [29], the Q function is an approximate lower bound of the log-likelihood, and maximizing it with respect to the parameters can achieve parameter estimate. We can set ${\Theta_{n+1}}^0=\hat{\Theta}_n$, and iteratively maximize $\hat{Q}_{n+1}\left(\Theta\right)$ to have better parameter estimation. That is, at each sample time, we iteratively update the conditional expectation to obtain a better parameter estimation. Applying this algorithm to the time

delay problem described in Section 2, we have

$$\begin{aligned}
&\hat{Q}_{n+1}\left(\Theta|{\Theta_{n+1}}^h\right)\\
&=(1-\gamma_{n+1})\,\hat{Q}_n\left(\Theta\right)+\gamma_{n+1}E_{{\Theta_{n+1}}^h}\left[\log P\left(y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}},\lambda_{n+1},\lambda_n;\Theta\right)\left|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}}\right.\right]\\
&\vdots\\
&=\prod_{p=2}^{n+1}(1-\gamma_p)E_{\Theta_0}\left[\log P\left(y_{T_1},\psi_{T_1-\lambda_1},\lambda_1;\Theta\right)|y_{T_1},\psi_{T_1-\lambda_1}\right]\\
&+\sum_{k=2}^{n}\prod_{p=k+1}^{n+1}(1-\gamma_p)\gamma_k E_{\hat{\Theta}_{k-1}}\left[\log P\left(y_{T_k},\psi_{T_k-\lambda_k},\lambda_k,\lambda_{k-1};\Theta\right)|y_{T_k},\psi_{T_k-\lambda_k}\right]\\
&+\gamma_{n+1}E_{{\Theta_{n+1}}^h}\left[\log P\left(y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}},\lambda_{n+1},\lambda_n;\Theta\right)\left|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}}\right.\right].
\end{aligned}\tag{2.36}$$

For each parameter in Θ_{n+1}, the same procedure as given in Section 4.1 can be used to find the new estimate. As a result, we can obtain the following formulas for each parameter:

1. The regressor parameter at $(h+1)th$ iteration for the $(n+1)th$ sample,

$$\left({\theta_{n+1}}^{h+1}\right)=\left({\theta_{n+1}}^{h+1}\right)_{.den}^{-1}\left({\theta_{n+1}}^{h+1}\right)_{.num},\tag{2.37}$$

where

$$\begin{aligned}
\left({\theta_{n+1}}^{h+1}\right)_{.den}&=(1-\gamma_{n+1})\left(\hat{\theta}_n\right)_{.den}+\gamma_{n+1}\sum_{i=1}^{d}P\left(\lambda_{n+1}=i|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};{\Theta_{n+1}}^h\right)\psi_{T_{n+1}-i}{}^T\psi_{T_{n+}}\\
\left({\theta_{n+1}}^{h+1}\right)_{.num}&=(1-\gamma_{n+1})\left(\hat{\theta}_n\right)_{.num}+\gamma_{n+1}\sum_{i=1}^{d}P\left(\lambda_{n+1}=i|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};{\Theta_{n+1}}^h\right)\psi_{T_{n+1}-i}{}^T y_{T_n}
\end{aligned}\tag{2.38}$$

2. The noise variance at $(h+1)th$ iteration for the $(n+1)th$ sample,

$$\sigma_{n+1}^{2h+1}=(1-\gamma_{n+1})\,\hat{\sigma}_n^2+\gamma_{n+1}\sum_{i=1}^{d}P\left(\lambda_{n+1}=i|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};{\Theta_{n+1}}^h\right)\left[y_{T_{n+1}}-\psi_{T_{n+1}-i}{\theta_{n+1}}^{h+1}\right]^2.\tag{2.39}$$

3. The transition probability at $(h+1)th$ iteration for the $(n+1)th$ sample,

$$(\alpha_{ij})_{n+1}^{h+1}=\frac{(\alpha_{ij})_{n+1.num}^{h+1}}{(\alpha_{ij})_{n+1.den}^{h+1}},\tag{2.40}$$

where

$$\begin{aligned}
(\alpha_{ij})_{n+1.num}^{h+1}&=(1-\gamma_{n+1})\,(\hat{\alpha}_{ij})_{n.num}+\gamma_{n+1}P\left(\lambda_{n+1}=i,\lambda_n=j|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};{\Theta_{n+1}}^h\right),\\
(\alpha_{ij})_{n+1.den}^{h+1}&=(1-\gamma_{n+1})\,(\hat{\alpha}_{ij})_{n.den}+\gamma_{n+1}\sum_{j=1}^{d}\left(\lambda_{n+1}=i,\lambda_n=j|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};{\Theta_{n+1}}^h\right).
\end{aligned}\tag{2.41}$$

Similarly, the posterior distribution of delay is calculated based on the parameter estimate and delay distribution at the previous hth iteration:

$$P\left(\lambda_{n+1}=i,\lambda_n=j|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};{\Theta_{n+1}}^h\right)=\frac{P\left(y_{T_{n+1}}|\psi_{T_{n+1}-\lambda_{n+1}},\lambda_{n+1}=i;{\Theta_{n+1}}^h\right)(\alpha_{ij})_{n+1}^{h+1}\pi_j^n}{\sum_{m=1}^{d}\sum_{l=1}^{d}P\left(y_{T_{n+1}}|\psi_{T_{n+1}-\lambda_{n+1}},\lambda_{n+1}=m;\hat{\Theta}_n\right)(\alpha_{ml})_{n+1}^{h+1}\pi_l^n} \tag{2.42}$$

and

$$P\left(\lambda_{n+1}=i|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};{\Theta_{n+1}}^h\right)=\sum_{j=1}^{d}P\left(\lambda_{n+1}=i,\lambda_n=j|y_{T_{n+1}},\psi_{T_{n+1}-\lambda_{n+1}};{\Theta_{n+1}}^h\right). \tag{2.43}$$

At the $(n+1)th$ recursion, there is iteration between delay distribution calculation and parameter estimation. After convergence of the iterative parameter estimation, we denote their values as $\hat{\Theta}_{n+1}=\left\{\hat{\theta}_{n+1},\hat{\sigma}_{n+1}^2,(\hat{\alpha}_{ij})_{n+1}\right\}$. Moreover, the following values are stored:

$$\left\{\left(\hat{\theta}_{n+1}\right)_{.num},\left(\hat{\theta}_{n+1}\right)_{.den},\hat{\sigma}_{n+1}^2,(\hat{\alpha}_{ij})_{n+1.num},(\hat{\alpha}_{ij})_{n+1.den},\pi_i^{n+1}\right\}.$$

When the next data point is available, the proposed iterative algorithm is applied to update the parameters. It is noted that the proposed iterative recursive estimation algorithm is an iterative process under the EM framework. Therefore, it is more similar to the batch EM algorithm than the original recursive EM algorithm. Because of the iteration at every recursion, the best use of each data point is made.

2.5 Simulation Studies

In this section, numerical examples are given to show the advantages of the proposed algorithm for both time invariant and time variant cases.

2.5.1 Identification of time invariant system

Consider the following first order system:

$$\begin{aligned} y_{T_k} &= 0.95y_{T_{k-1}}+0.4u_{T_k-\lambda_k}+\upsilon_{T_k} \\ u &\sim N(0,1) \\ \upsilon &\sim N(0,0.01) \\ \lambda_k &\in \{1,2,3,4\} \end{aligned} \tag{2.44}$$

Input signal u is a normally distributed random variable with zero mean and unit variance. Measurement noise v is a normally distributed random variable with zero mean and 0.01

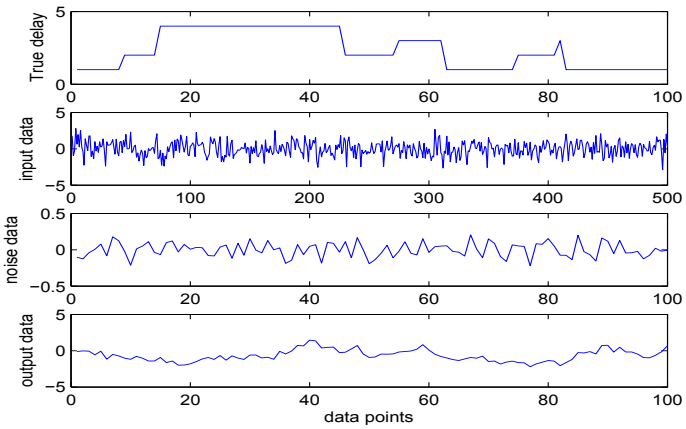

Figure 2.1: Simulation data for the time invariant process

variance. Delay varies among {1,2,3,4}, in the form of a Markov chain. The true transition matrix governing the delay switching mechanism is

$$A = \{\alpha_{ij}\} = \begin{pmatrix} 0.90 & 0.06 & 0.03 & 0.01 \\ 0.02 & 0.90 & 0.06 & 0.02 \\ 0.01 & 0.06 & 0.90 & 0.02 \\ 0.01 & 0.03 & 0.06 & 0.90 \end{pmatrix}. \tag{2.45}$$

In simulation, $\Delta = 5$, therefore, $L = 500$ fast-rate input data points and $N = 100$ slow-rate output data points are collected for system identification. Figure 2.1 contains the fast-rate input data, true delay generated by the transition probability, measurement noise, and output data. As illustrated, the delay does not change frequently and may remain constant for some prolonged periods of time. This agrees with the high probability of diagonal elements in transition matrix A.

The first 50 slow rate output samples and 250 fast rate input samples are selected as the training data set and the rest form the test data set. Applying the proposed algorithm of Section 3, parameter estimation converges within 10 iterations of the EM algorithm (Figure 2.2). The HMM transition probability is estimated as:

$$\hat{A} = \{\hat{\alpha}_{ij}\} = \begin{pmatrix} 0.8660 & 0.0685 & 0.0427 & 0.0228 \\ 0.0552 & 0.9221 & 0.0223 & 0.0003 \\ 0.0003 & 0.0275 & 0.9513 & 0.0208 \\ 0.0011 & 0.0243 & 0.0454 & 0.9291 \end{pmatrix}, \tag{2.46}$$

which is close to the real one given in Equation 2.45. Using this transition matrix, the estimated delay of the training data set can be obtained by

$$\hat{\lambda}_k = arg\max_i \quad P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \hat{\Theta}\right), \tag{2.47}$$

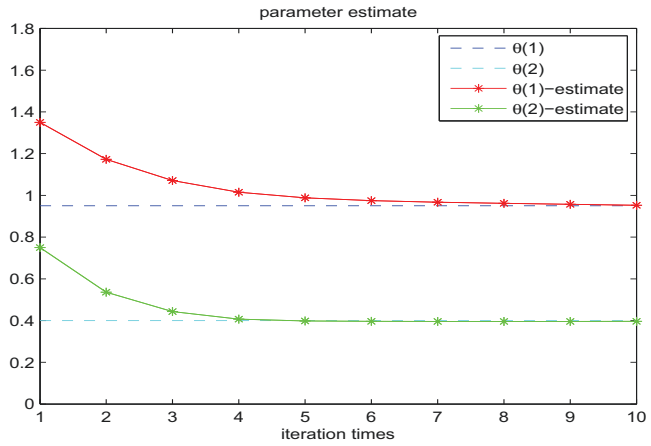

Figure 2.2: Parameters estimation for the time invariant process

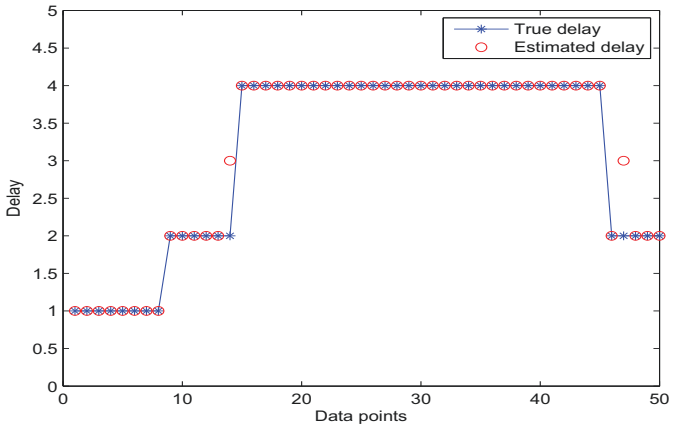

Figure 2.3: Delay estimation for the time invariant process

where $\hat{\Theta} = \{\hat{\theta}, \hat{\sigma}^2, \hat{\alpha}_{ij}, \hat{\pi}_i\}$ are the estimated parameters. The delay estimation result is illustrated in Figure 2.3, which agrees with the true delay value with an accuracy of 92%.

For the test data set, we can predict the most probable value of time delay, predict the simulation output, and then compare the output prediction with the measurement. To predict the time delay, the prior distribution of the delay $\lambda_k = i$ up to the observation at time $k-1$ in the test data set is calculated by Equation 2.48

$$\begin{aligned} & P\left(\lambda_k = i | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \hat{\Theta}\right) \\ &= \sum_{j=1}^{d} P\left(\lambda_k = i, \lambda_{k-1} = j | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \hat{\Theta}\right) \\ &= \sum_{j=1}^{d} \hat{\alpha}_{ij} P\left(\lambda_{k-1} = j | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \hat{\Theta}\right), \end{aligned} \tag{2.48}$$

where the posterior delay distribution $\lambda_{k-1} = j$ up to the observation at time $k-1$ is

calculated by

$$\begin{aligned} & P\left(\lambda_{k-1}=j|y_{T_{k-1}:T_1},u_{T_{k-1}:1},\hat{\Theta}\right) \\ &= \sum_{l=1}^{d} P\left(\lambda_{k-1}=j,\lambda_{k-2}=l|y_{T_{k-1}:T_1},u_{T_{k-1}:1},\hat{\Theta}\right) \\ &= \sum_{l=1}^{d} \frac{P\left(y_{T_{k-1}}|\lambda_{k-1}=j,\psi_{T_{k-1}-\lambda_{k-1}},\hat{\Theta}\right)\hat{\alpha}_{jl}P\left(\lambda_{k-2}=l|y_{T_{k-2}:T_1},u_{T_{k-2}:1},\hat{\Theta}\right)}{\sum_{m=1}^{d}\sum_{n=1}^{d} P\left(y_{T_{k-1}}|\lambda_{k-1}=m,\psi_{T_{k-1}-\lambda_{k-1}},\hat{\Theta}\right)\hat{\alpha}_{mn}P\left(\lambda_{k-2}=n|y_{T_{k-2}:T_1},u_{T_{k-2}:1},\hat{\Theta}\right)}. \end{aligned} \tag{2.49}$$

Equation 2.49 is the discrete state propagation of Markov chain starting from the posterior distribution of last delay in the training data set. This is because the test data set is continuous with the training data set. Delay value prediction can be obtained by the calculated prior delay distribution in Equation 2.48. Therefore, simulation prediction of test data set is computed based on Equation 2.1, given $\hat{\theta}$ and

$$\hat{\lambda}_k = arg\max_i \quad P\left(\lambda_k=i|y_{T_{k-1}:T_1},u_{T_{k-1}:1},\hat{\Theta}\right). \tag{2.50}$$

In order to further test effectiveness of using Markov chain to model delay correlation, we consider a more realistic way to generate the time delays in the simulation. Consider delay is caused by transportation of materials in a pipe. The output measurement has varying time delay because of the varying flow rate. The flow rate is generated by passing a white noise sequence through a low-pass filter. The values of the transportation time delay are inversely proportional to the flow rates because of the fixed intersection area of the pipe. The delay values are rounded to the nearest integers {1,2,3,4} for discrete time system simulation. Using new simulated data, we apply the proposed method and compare it with two other alternatives. In the first method, which is called the independent delay estimation, delay is considered to change randomly with a uniform distribution instead of a Markov chain. In the second method, which is called the fixed delay estimation, delay is considered to be constant. In the implementation of independent delay estimation, prior distribution of delay is considered to follow a uniform distribution among {1,2,3,4}, while the parameters are estimated using EM algorithm. In the fixed delay estimation, time delay is considered to be the same for all samples. The unknown value is uniformly distributed among {1,2,3,4}, and the identification uses the EM algorithm as well.

The performance of the three different methods for both training and test data sets is listed in Table 4.2. In all cases, the accuracy of the proposed hidden Markov model based delay estimation is higher than both the independent delay and the fixed delay estimation methods and its RMSE is smallest. This simulation result demonstrates the effectiveness of using the Markov chain model to describe practical correlated delays caused by transportations.

Table 2.2: A Summary of the RMSE for the Time Invariant Process

	Self validation		Cross validation	
	Accuracy*	RMSE	Accuracy*	RMSE
Markov delay	92%	0.0821	90%	0.0983
Independent delay	85%	0.1226	79%	0.1630
Fixed delay	55%	0.3672	42%	0.4313

* Rate of accuracy of delay estimation.

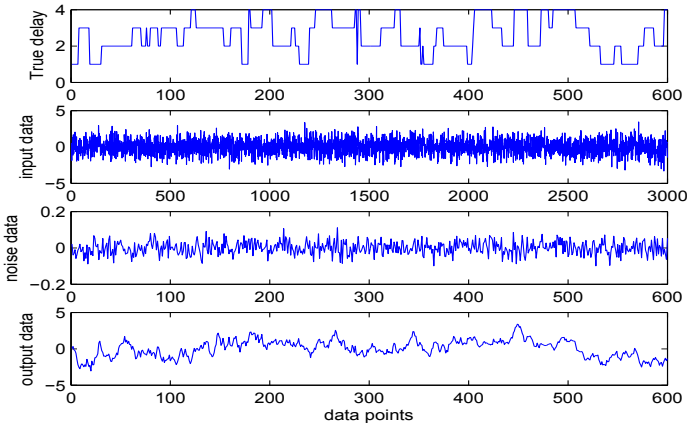

Figure 2.4: Simulation data for the time variant process

2.5.2 Identification of time variant system

Consider the following time variant process:

$$\begin{aligned} y_{T_k} &= 0.90 y_{T_{k-1}} + 0.5 u_{T_k - \lambda_k} + \nu_{T_k}, k = 1, \ldots, 200 \\ y_{T_k} &= 0.94 y_{T_{k-1}} + 0.4 u_{T_k - \lambda_k} + \nu_{T_k}, k = 201, \ldots, 400 \\ y_{T_k} &= 0.98 y_{T_{k-1}} + 0.3 u_{T_k - \lambda_k} + \nu_{T_k}, k = 401, \ldots, 600 \\ u &\sim N(0, 1) \\ \upsilon &\sim N(0, 0.01) \\ \lambda_k &\in \{1, 2, 3, 4\} \end{aligned} \tag{2.51}$$

This process is an extension of the previous time invariant process because it contains three different stages, with 200 output data points for each stage. The regression parameters for the three stages are different, while all other parameters are the same as the previous one. Figure 2.4 contains input data, true delay, measurement noise, and output data.

From a practical point of view, we start iterative recursive EM after several data points are available [29]. Batch EM (Section 3) is applied when 20 slow rate output data and 100

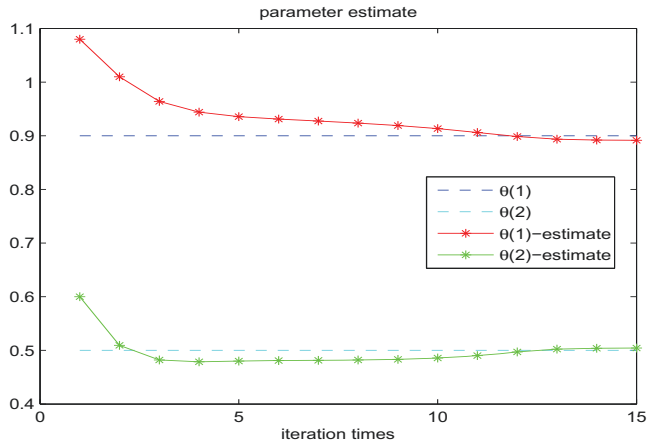

Figure 2.5: Parameter estimation by batch EM for the first 20 sample data of output

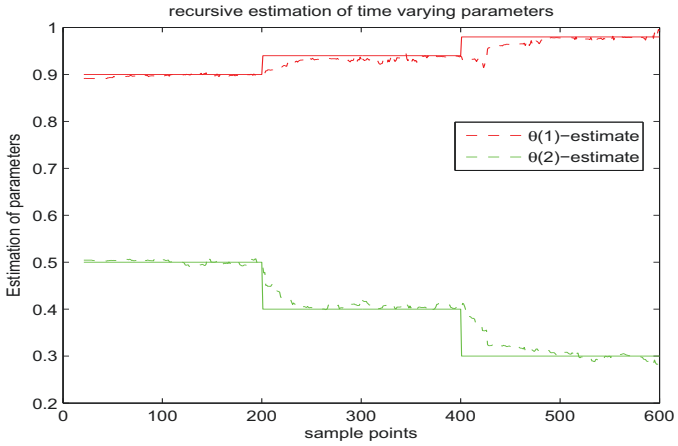

Figure 2.6: Parameter estimation for the time variant process by iterative REM

fast rate input data are available. Next, the iterative recursive EM (Section 4.2) is applied, so that parameters can be updated at each data point. Moreover, we select $\gamma_{n+1} = 5\%$ as a fixed step size for this time variant process.

Figure 2.5 shows the parameter estimation of batch EM on 20 slow rate data points, which converges within 15 iterations. Figure 2.6 shows the parameter estimation for the remaining 580 slow rate data points sequentially. As shown in the figure, recursive estimation converges to the true parameter value within around 39 data points.

The delay estimation result in Figure 2.7 agrees with the true delay value with an accuracy of 93%. The comparison of the real and the predicted value of the output shows the estimation matches the real output with an RMSE value of 0.1023.

In order to illustrate the advantage of dealing with time variant property using the

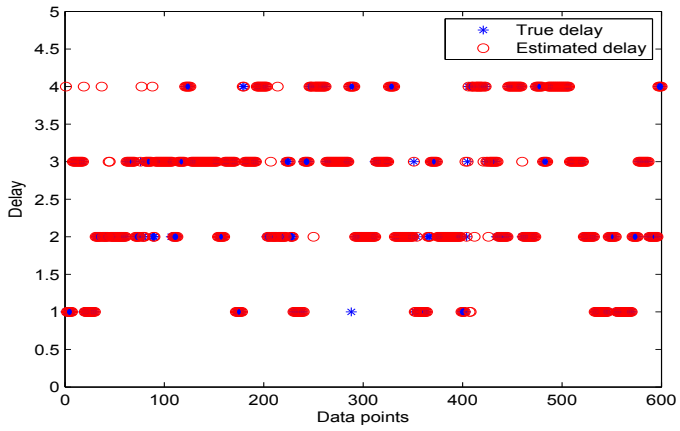

Figure 2.7: Delay estimation for the time variant process

Table 2.3: A Summary of the RMSE for the Time Variant Process

	RMSE	Accuracy*	Computation time	Convergence time
Iterative REM	0.1023	93%	8.0 s	39 samples
REM without iteration	0.1264	89%	6.4 s	46 samples
Batch EM	0.2369	82%	4.9 s	N/A
Moving window EM**	0.1048	85%	300.2 s	78 samples

* Rate of accuracy of delay estimation.
** 100 sample window based.

method in Section 4.2, we compare it with three other methods. In the first method, batch EM estimation is applied (Section 3). In the second method, we apply moving window EM while the third method is the recursive version without iteration (Section 4.1).

In the implementation of batch estimation, first half of data is used for training while the second half is used for testing. In the implementation of moving window EM, a fixed window of 100 past data points is sequentially used to estimate parameters. Therefore, we do not have an estimate for the first 99 samples.

As listed in Table 2.3, the proposed iterative recursive EM algorithm has the smallest RMSE and the rate of accuracy of the delay estimation is the highest. It is noted that moving window EM and the proposed method have similar RMSE. However, the moving window EM has more computational cost and requires more sample time to converge to the new parameters. The convergence time is the time taken for the estimate to reach within 5% of the true value after a parameter change occurs.

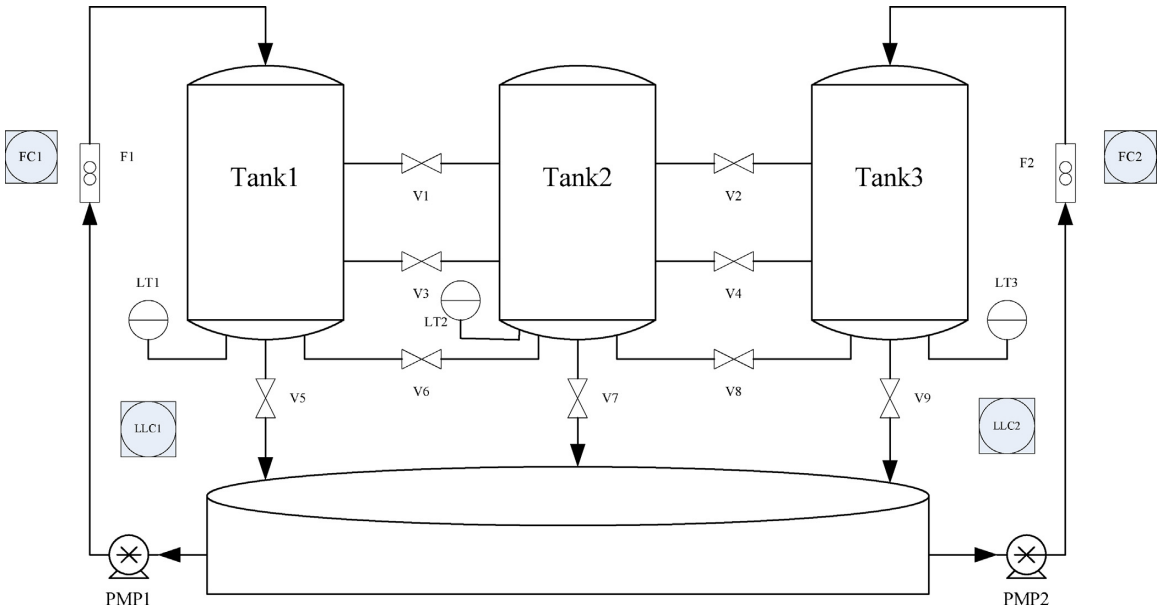

Figure 2.8: Schematic diagram of the hybrid tank system

2.6 Experimental Evaluation

In this section, experimental examples are given to show the advantage of the proposed algorithm for both time invariant and time variant cases. The system identification experiment is designed and performed on a pilot-scale hybrid tank system. The schematic diagram is displayed in Figure 4.7.

2.6.1 Identification of time invariant system

In this experiment, only the right tank, Tank3 and the middle tank, Tank2 are used. Therefore, the valves V7-V9 are open, and valves V1-V6 are closed. The inlet flow from right pump 2 and Tank3 level are considered as input and output, respectively. Initially, a constant input value of 5.5 is introduced to the process. As a result, the output turns to steady state after a period of time. Next, a filtered random binary signal (RBS) with level $[-0.7, 0.7]$ is added to the input signal to stimulate the system and generate experimental data. Input and output data are shown in Figure 2.9. The input is measured every 16 seconds, while the output is measured every 48 seconds. The Markov chain time delay sequence is manually imposed with two values, 16 seconds and 32 seconds, with following transition probability:

$$A = \{\alpha_{ij}\} = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix}$$

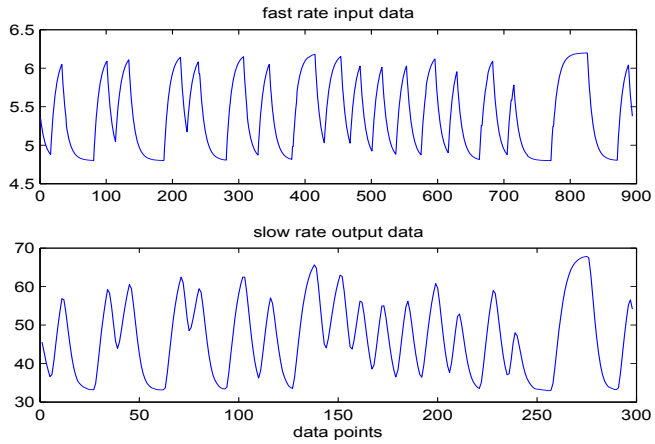

Figure 2.9: Input and output data for the time invariant process

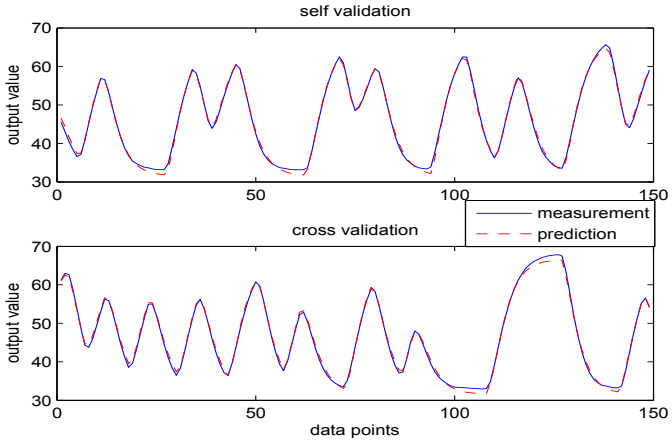

Figure 2.10: Self validation and cross validation for the time invariant process

The data is divided into two halves; the first half forms the training set, and the second half forms the test set. Consider the model structure for this plant is a first order ARX model given as

$$y_{T_k} = a y_{T_{k-1}} + b u_{T_k - \lambda_k}.$$

Applying the proposed EM algorithm of Section 3 to the normalized training set, we obtain the estimates $a = 0.6271$, $b = 0.4445$. The transition probability matrix is

$$\hat{A} = \{\hat{\alpha}_{ij}\} = \begin{pmatrix} 0.9656 & 0.0344 \\ 0.0575 & 0.9425 \end{pmatrix}.$$

From the self and cross validation results of the infinite step ahead prediction in Figure 2.10, we can see that the proposed method gives good estimation results.

Next, we regenerate time delays through transportation delay and compare the proposed method with the two alternatives described in Section 5.1. The performances of the three

Table 2.4: A Summary of the RMSE for the Time Invariant Experiment

	Self validation	Cross validation
Markov delay	0.5935	0.6478
Independent delay	1.1256	1.3652
Fixed delay	3.6352	4.6298

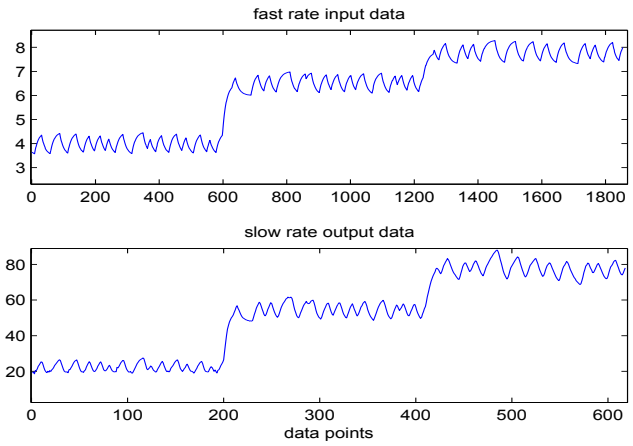

Figure 2.11: Input and output data for the time variant process

different methods are listed in Table 4.4. We can see that the RMSE of Markov delay based method is the smallest.

2.6.2 Identification of time variant system

In this experiment, the valves V2, V4, V7, V8 and V9 are open, and valves V1, V3, V5 and V6 are closed. The operating point of the plant is changed at different points in time. Figure 2.11 contains 1854 fast rate input and 618 slow rate output data for all three operating points, where the output is also generated with manually imposed time delay.

The first 20 slow rate data points are used to estimate the parameters together with time delay and transition probability using batch EM (Section 3). Next, the proposed iterative recursive algorithm (Section 4.2) is applied to deal with the time varying issue. Figure 2.12 shows the plot of the target and of the prediction value of water level. It is clear that they are adequately matched with an RMSE equal to 1.0217. Similarly, we can compare the result of the proposed recursive estimation with the three methods introduced in Section 5.2, as listed in Table 2.5. Based on the RMSE values, it is clear that the proposed recursive estimation algorithm has the best performance.

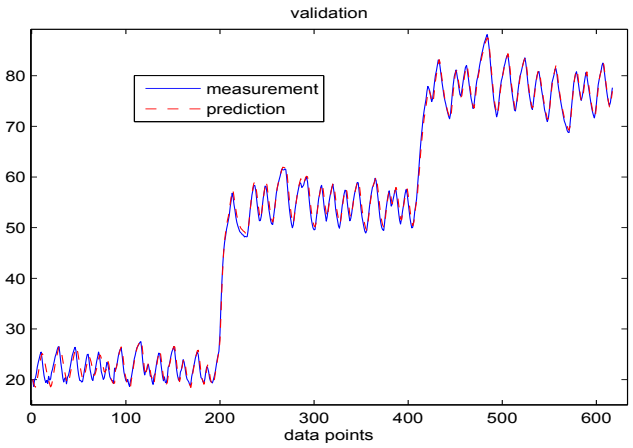

Figure 2.12: Validation result for the time variant process by iterative REM

Table 2.5: A Summary of the RMSE for the Time Variant Experiment

	RMSE
Iterative REM	1.0217
REM without iteration	1.6528
Batch EM	3.2365
Moving window EM**	1.9239

** 100 sample window based.

2.7 Conclusions

This chapter considers identification of both time invariant and time variant ARX models with time-varying time delays. The proposed algorithms can simultaneously estimate the distribution of time delay along with the parameters. Moreover, given the estimated transition probability of time delay, we can estimate the most probable value of the delay for every sampling instant.

For the time invariant system, it is shown by a numerical simulation example and a multitank system that assuming a Markov chain for the delay sequence is an effective approach to capture correlation of time delays.

For the time variant system, the proposed iterative recursive algorithm also achieves better performance and is computationally more efficient, compared with the recursive EM without iteration, the moving window EM and the batch EM algorithms. The performance is validated by a numerical simulation example and an experimental hybrid tank system.

2.A Appendix A

The posterior distribution $P\left(\lambda_1 = i|y_{T_1}, u_{T_1:1}, \Theta^h\right)$ in $Q3\left(\pi_i\right)$ and Equation 2.12 is calculated by Bayesian rules as follows:

$$\begin{aligned} & P\left(\lambda_1 = i|y_{T_1}, u_{T_1:1}, \Theta^h\right) \\ &= \frac{P\left(y_{T_1}, \lambda_1 = i|u_{T_1:1}, \Theta^h\right)}{\sum_{m=1}^{d} P\left(y_{T_1}, \lambda_1 = m|u_{T_1:1}, \Theta^h\right)} \\ &= \frac{P\left(y_{T_1}|u_{T_1:1}, \lambda_1 = i, \Theta^h\right) P\left(\lambda_1 = i|u_{T_1:1}, \Theta^h\right)}{\sum_{m=1}^{d} P\left(y_{T_1}|u_{T_1:1}, \lambda_1 = m, \Theta^h\right) P\left(\lambda_1 = m|u_{T_1:1}, \Theta^h\right)}, \end{aligned} \tag{2.52}$$

where

$$\begin{aligned} & P\left(y_{T_1}|u_{T_1:1}, \lambda_1 = i, \Theta^h\right) = P\left(y_{T_1}|\psi_{T_1-\lambda_1}, \lambda_1 = i, \Theta^h\right), \\ & P\left(\lambda_1 = i|u_{T_1:1}, \Theta^h\right) = \left(\pi_i\right)^h, \end{aligned} \tag{2.53}$$

and therefore Equation 2.12 is obtained.

The posterior distribution $P\left(\lambda_k = i, \lambda_{k-1} = j|y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)$ in $Q1\left(\theta, \sigma^2\right)$ and Equation 2.13 is calculated as follows:

$$\begin{aligned} & P\left(\lambda_k = i, \lambda_{k-1} = j|y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \\ &= \frac{P\left(y_{T_k}, \lambda_k = i, \lambda_{k-1} = j|y_{T_{k-1}:T_1}, u_{T_k:1}, \Theta^h\right)}{P\left(y_{T_k}|y_{T_{k-1}:T_1}, u_{T_k:1}, \Theta^h\right)} \\ &= \frac{P\left(y_{T_k}, \lambda_k = i, \lambda_{k-1} = j|y_{T_{k-1}:T_1}, u_{T_k:1}, \Theta^h\right)}{\sum_{m=1}^{d}\sum_{n=1}^{d} P\left(y_{T_k}, \lambda_k = m, \lambda_{k-1} = n|y_{T_{k-1}:T_1}, u_{T_k:1}, \Theta^h\right)} \\ &= \frac{P\left(y_{T_k}|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k = i, \lambda_{k-1} = j, \Theta^h\right) P\left(\lambda_k = i|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_{k-1} = j, \Theta^h\right) P\left(\lambda_{k-1} = j|y_{T_{k-1}:T_1}, u_{T_k:1}, \Theta^h\right)}{\sum_{m=1}^{d}\sum_{n=1}^{d} P\left(y_{T_k}|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k = m, \lambda_{k-1} = j, \Theta^h\right) P\left(\lambda_k = m|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_{k-1} = n, \Theta^h\right) P\left(\lambda_{k-1} = n|y_{T_{k-1}:T_1}, u_{T_k:1}, \Theta^h\right)}, \end{aligned} \tag{2.54}$$

where the three terms in both numerator and denominator can be simplified by omitting irrelevant variables

$$\begin{aligned}
&P\left(y_{T_k}|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k = i, \lambda_{k-1} = j, \Theta^h\right) = P\left(y_{T_k}|\psi_{T_k-\lambda_k}, \lambda_k = i, \Theta^h\right), \\
&P\left(\lambda_k = i|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_{k-1} = j, \Theta^h\right) = (\alpha_{ij})^h, \\
&P\left(\lambda_{k-1} = j|y_{T_{k-1}:T_1}, u_{T_k:1}, \Theta^h\right) = P\left(\lambda_{k-1} = j|y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \Theta^h\right),
\end{aligned} \tag{2.55}$$

and finally Equation 2.13 is obtained.

2.B Appendix B

For the regressor parameter estimation obtained in Equation 2.29, the derivations for $\left(\hat{\theta}_{n+1}\right)_{.den}$ and $\left(\hat{\theta}_{n+1}\right)_{.num}$ are as follows,

$$\begin{aligned}
&\left(\hat{\theta}_{n+1}\right)_{.den} \\
&= \prod_{p=2}^{n+1}(1-\gamma_p)\sum_{i=1}^{d} P(\lambda_1 = i|y_{T_1}, \psi_{T_1-\lambda_1}; \Theta_0)\psi_{T_1-i}{}^T\psi_{T_1-i} \\
&+ \sum_{k=2}^{n}\prod_{p=k+1}^{n+1}(1-\gamma_p)\,\gamma_k\sum_{i=1}^{d} P(\lambda_k = i|y_{T_k}, \psi_{T_k-\lambda_k}; \hat{\Theta}_{k-1})\psi_{T_k-i}{}^T\psi_{T_k-i} \\
&+\gamma_{n+1}\sum_{i=1}^{d} P\left(\lambda_{n+1} = i|y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}; \hat{\Theta}_n\right)\psi_{T_{n+1}-i}{}^T\psi_{T_{n+1}-i},
\end{aligned} \tag{2.56}$$

and

$$\begin{aligned}
&\left(\hat{\theta}_{n+1}\right)_{.num} \\
&= \prod_{p=2}^{n+1}(1-\gamma_p)\sum_{i=1}^{d} P(\lambda_1 = i|y_{T_1}, \psi_{T_1-\lambda_1}; \Theta_0)\psi_{T_1-i}{}^T y_{T_1} \\
&+ \sum_{k=2}^{n}\prod_{p=k+1}^{n+1}(1-\gamma_p)\,\gamma_k\sum_{i=1}^{d} P(\lambda_k = i|y_{T_k}, \psi_{T_k-\lambda_k}; \hat{\Theta}_{k-1})\psi_{T_k-i}{}^T y_{T_k} \\
&+\gamma_{n+1}\sum_{i=1}^{d} P\left(\lambda_{n+1} = i|y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}; \hat{\Theta}_n\right)\psi_{T_{n+1}-i}{}^T y_{T_{n+1}}.
\end{aligned} \tag{2.57}$$

Similarly, the formulas for $\left(\hat{\theta}_n\right)_{.den}$ and $\left(\hat{\theta}_n\right)_{.num}$ in

$$\hat{\theta}_n = \left(\hat{\theta}_n\right)_{.den}^{-1}\left(\hat{\theta}_n\right)_{.num} \tag{2.58}$$

are as follows,

$$\begin{aligned}
&\left(\hat{\theta}_n\right)_{.den} \\
&= \prod_{p=2}^{n}(1-\gamma_p)\sum_{i=1}^{d} P(\lambda_1 = i|y_{T_1}, \psi_{T_1-\lambda_1}; \Theta_0)\psi_{T_1-i}{}^T\psi_{T_1-i} \\
&+ \sum_{k=2}^{n-1}\prod_{p=k+1}^{n}(1-\gamma_p)\,\gamma_k\sum_{i=1}^{d} P(\lambda_k = i|y_{T_k}, \psi_{T_k-\lambda_k}; \hat{\Theta}_{k-1})\psi_{T_k-i}{}^T\psi_{T_k-i} \\
&+\gamma_n\sum_{i=1}^{d} P\left(\lambda_n = i|y_{T_n}, \psi_{T_n-\lambda_n}; \hat{\Theta}_n\right)\psi_{T_n-i}{}^T\psi_{T_n-i},
\end{aligned} \tag{2.59}$$

and

$$\begin{aligned}
&\left(\hat{\theta}_n\right)_{.num} \\
&= \prod_{p=2}^{n} (1-\gamma_p) \sum_{i=1}^{d} P(\lambda_1 = i | y_{T_1}, \psi_{T_1-\lambda_1}; \Theta_0) \psi_{T_1-i}{}^T y_{T_1} \\
&+ \sum_{k=2}^{n-1} \prod_{p=k+1}^{n} (1-\gamma_p) \gamma_k \sum_{i=1}^{d} P(\lambda_k = i | y_{T_k}, \psi_{T_k-\lambda_k}; \hat{\Theta}_{k-1}) \psi_{T_k-i}{}^T y_{T_k} \\
&+ \gamma_n \sum_{i=1}^{d} P\left(\lambda_n = i | y_{T_n}, \psi_{T_n-\lambda_n}; \hat{\Theta}_n\right) \psi_{T_n-i}{}^T y_{T_n}.
\end{aligned} \tag{2.60}$$

Next, substitute Equations 2.59 and 2.60, which form the parameter estimate for $\hat{\theta}_n$ at previous data point, to Equations 2.56 and 2.57, and it provides the following recursive parameter updating equation:

$$\begin{aligned}
&\hat{\theta}_{n+1} = \left(\hat{\theta}_{n+1}\right)_{.den}^{-1} \left(\hat{\theta}_{n+1}\right)_{.num} \\
&= \left\{ (1-\gamma_{n+1}) \left(\hat{\theta}_n\right)_{.den} + \gamma_{n+1} \sum_{i=1}^{d} P\left(\lambda_{n+1} = i | y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}; \hat{\Theta}_n\right) \psi_{T_{n+1}-i}{}^T \psi_{T_{n+1}-i} \right\}^{-1} \\
&\left\{ (1-\gamma_{n+1}) \left(\hat{\theta}_n\right)_{.num} + \gamma_{n+1} \sum_{i=1}^{d} P\left(\lambda_{n+1} = i | y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}; \hat{\Theta}_n\right) \psi_{T_{n+1}-i}{}^T y_{T_{n+1}} \right\}.
\end{aligned} \tag{2.61}$$

For the noise variance estimation obtained in Equation 2.31, by solving $\frac{\partial \hat{Q}4_{n+1}(\theta,\sigma^2)}{\partial \sigma^2} = 0$ and substituting previous estimate, we can obtain the updated noise variance from previous estimation,

$$\hat{\sigma}^2_{n+1} = \frac{(1-\gamma_{n+1}) \left(\hat{\sigma}^2_n\right)_{.num} + \gamma_{n+1} \sum_{i=1}^{d} P\left(\lambda_{n+1} = i | y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}; \hat{\Theta}_n\right) \left[y_{T_{n+1}} - \psi_{T_{n+1}-i} \hat{\theta}_{n+1}\right]^2}{(1-\gamma_{n+1}) \left(\hat{\sigma}^2_n\right)_{.den} + \gamma_{n+1} \sum_{i=1}^{d} P\left(\lambda_{n+1} = i | y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}; \hat{\Theta}_n\right)}, \tag{2.62}$$

where $\left(\hat{\sigma}^2_n\right)_{.num}$ and $\left(\hat{\sigma}^2_n\right)_{.den}$ are the numerator and denominator of the formula used to obtain $\hat{\sigma}^2_n$. Since that the value of denominator can be simplified to be 1, as shown below,

$$\begin{aligned}
&\left(\hat{\sigma}^2_{n+1}\right)_{.den} \\
&= (1-\gamma_{n+1}) \left(\hat{\sigma}^2_n\right)_{.den} + \gamma_{n+1} \sum_{i=1}^{d} P\left(\lambda_{n+1} = i | y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}; \hat{\Theta}_n\right) \\
&= (1-\gamma_{n+1}) \left(\hat{\sigma}^2_n\right)_{.den} + \gamma_{n+1} \\
&= (1-\gamma_{n+1}) \left((1-\gamma_n) \left(\hat{\sigma}^2_{n-1}\right)_{.den} + \gamma_n\right) + \gamma_{n+1} \\
&\vdots \\
&= (1-\gamma_{n+1}) \left((1-\gamma_n) \left(\cdots (1-\gamma_3) \left((1-\gamma_2) + \gamma_2\right) + \gamma_3 \cdots\right) + \gamma_n\right) + \gamma_{n+1} \\
&= 1,
\end{aligned} \tag{2.63}$$

the estimate for noise variance is simplified to

$$\hat{\sigma}^2_{n+1} = (1-\gamma_{n+1}) \hat{\sigma}^2_n + \gamma_{n+1} \sum_{i=1}^{d} P\left(\lambda_{n+1} = i | y_{T_{n+1}}, \psi_{T_{n+1}-\lambda_{n+1}}; \hat{\Theta}_n\right) \left[y_{T_{n+1}} - \psi_{T_{n+1}-i} \hat{\theta}_{n+1}\right]^2. \tag{2.64}$$

For the noise variance estimation obtained in Equation 2.32, similarly substituting previous estimate of $(\hat{\alpha}_{ij})_{n.num}$ and $(\hat{\alpha}_{ij})_{n.den}$ into current estimate of $(\hat{\alpha}_{ij})_{n+1.num}$ and $(\hat{\alpha}_{ij})_{n+1.den}$, we can obtain the updated noise variance from previous estimation, as shown in Equation 2.32.

Chapter 3

Robust Estimation of ARX Models with Time Varying Delays Using EM Algorithm

This chapter is concerned with the identification of time-delay processes. Time delay occurs in almost all industrial processes and can vary in various fashions, for example, continuous changing or switching. The switching mechanism is not purely random but often governed by some switching mechanism that may be described by stochastic models such as Markov chain. Measured data are often contaminated by outliers. Gaussian distribution is not sufficient to describe the actual disturbance contaminated by outliers. Instead, other probabilistic distributions such as t-distribution should be considered, thus diminishing the effect of outliers. In the presence of unknown time delay, the Expectation Maximization (EM) algorithm is applied to estimate both the parameter and time delay. The proposed algorithms are verified by numerical examples and a pilot-scale tank experiment.

3.1 Introduction

Time delay is a common phenomena in almost all processes, and it makes the system identification challenging when the delay is time varying [31]. The existence of measurement outliers is an issue as well, especially when the outliers have no fixed distribution. The combination of these two issues can deteriorate the identification significantly and the resulting estimation becomes unreliable. Thus, they should be paid attention to during system identification.

Most researches deal with constant time delays [32, 33, 34, 35]. However, time delay is often introduced in the plant due to the transportation speed. Thus, when the transportation speed changes, for instance, due to the change in the flow of the liquid in a pipe, the

delay is also changed. As a result, the delay in each sample time follows a dynamic behavior which can be expressed by dynamic correlation models such as Markov chain [18].

The property of Markov chain says that the switching delay value evolves in a Markovian fashion. This means that the delay value at the current time only depends on its immediate past and it has only finite number of discrete values. It may or may not switch to another value. In the formulation of Markov chain, the switching mechanism of time delay is governed by a transition probability. The transition matrix may assign larger probability for delay to stay unchanged and assigns lower probability to switch current delay value to distant values, which is representative of the real situation. When the switching delay value is unknown and cannot be measured directly, the switching dynamics described by a latent variable is normally referred to as a Hidden Markov Model (HMM). As one of the most important statistical models, HMM has been applied to various areas like fault diagnosis for gearbox [36], Bayesian model selection [37], model reduction [38], anomaly detection in electronic systems [39], and blind categorical de-convolution [40].

Outliers happen in industry as well, which are usually caused by occasional interruption and disturbance. Conventional approaches make use of Gaussian models to approximate the noise in the complex processes [41]. The major limitation of this method is the lack of robustness in the presence of outliers. This is because under the assumed Gaussian distribution, maximizing the likelihood function is equivalent to finding the least square solution, which is well known for the lack of robustness [42]. Thus, models identified by this approach may be unreliable in the presence of outliers.

A more general approach to model the measurement noise with outliers is to use the t-distribution [43]. The t-distribution can have long tails through adjustable degrees of freedom. This gives the ability to adjust in order to improve the modeling of noise and outliers simultaneously. A a result, the affect of outliers on modeling can be diminished by assigning proper probability densities to outliers. Christmas et al. [44] assumed a Student-t distributed excitation noise in the Bayesian AR model, where the parameter estimation performs well against Gaussian data and modeling with Student-t assumption is much more robust to outliers than both Gaussian and Gaussian mixture models. Lange et al. [45] used the Student-t distribution for robust statistical inference in both linear and nonlinear regression, and they also show how this distribution allows them to achieve more robust models by controlling the degree to down weigh the outliers. Moreover, many other researchers have employed the Student-t distribution for robust parameter estimation, like generalized component analysis [46], signal filtering and prediction [47], and image

segmentation [48, 49].

In this chapter, we will develop a novel identification approach which is robust to outliers using the t-distribution for processes with time varying time delays. The delay is supposed to follow a hidden Markov model (HMM) and its parameters are also estimated together with the parameters of the process model. A simulation example and an experimental implementation verify that the proposed method can provide more reliable identification results.

The remainder of this chapter is organized as follows. A detailed problem description of the ARX model identification in the presence of outliers and time varying time delays is presented in the next section. The following section applies EM Algorithm to solve this robust estimation problem. Then Section 4 gives a numerical example to demonstrate the proposed method. Then in Section 5, a pilot-scale experiment is conducted to further verify the proposed method. Finally, the conclusion is given in the last section.

3.2 Problem Statement

Consider the following dual rate ARX model:

$$\begin{aligned} y_{T_k} &= \psi_{T_k-\lambda_k}\theta + e_{T_k} \\ \psi_{T_k-\lambda_k} &= \left[\begin{array}{cccccc} y_{T_{k-1}} & \cdots & y_{T_{k-na}} & u_{T_k-\lambda_k} & \cdots & u_{T_k-nb-\lambda_k} \end{array}\right] \in \mathbb{R}^{1\times(na+nb+1)}. \end{aligned} \tag{3.1}$$

where $\{y_{T_k}, k = 1, 2, ..., N\}$ is the slow rate output variable. $\{u_t, t = 1, 2, ..., L\}$ is the fast rate input variable. The slow rate sampling time is Δ times that of the fast rate ($L = \Delta * N$). $\{\lambda_k, k = 1, 2, ..., N\}$ is time varying delay. $\theta \in \mathbb{R}^{(na+nb+1)\times 1}$ is the regression parameter vector where $nb + \lambda_k < \Delta$. e_t is associated measurement noise, which is considered to follow a t-distribution, i.e. $e_t \sim t(0, \sigma^2, v)$ with unknown scaling parameter σ^2 and degrees of freedom v.

From the distribution of measurement noise, $y_{T_k} \sim t(\psi_{T_k-\lambda_k}\theta, \sigma^2, v)$ is calculated by

$$P\left(y_{T_k}|\psi_{T_k-\lambda_k}\theta, \sigma^2, v\right) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{\pi v\sigma^2}}\left\{1 + \frac{\left[y_{T_k} - \psi_{T_k-\lambda_k}\theta\right]^2}{\sigma^2 v}\right\}^{-\frac{v+1}{2}}. \tag{3.2}$$

Essentially, the t-distribution can be decomposed into scaled Gaussian distributions, where the variance scale r_k is a Gamma distributed latent variable which depends on the degree of freedom v:

$$f\left(y_{T_k}|\psi_{T_k-\lambda_k}\theta, \sigma^2, v\right) = \int f\left(y_{T_k}|\psi_{T_k-\lambda_k}\theta, \sigma^2, v, r_k\right) f\left(r_k|\psi_{T_k-\lambda_k}\theta, \sigma^2, v\right) dr_k, \tag{3.3}$$

where

$$\begin{array}{l} y_{T_k}|\left(\psi_{T_k-\lambda_k}\theta,\sigma^2,v,r_k\right)=y_{T_k}|\left(\psi_{T_k-\lambda_k}\theta,\sigma^2,r_k\right)\sim N\left(\psi_{T_k-\lambda_k}\theta,\frac{\sigma^2}{r_k}\right), \\ r_k|\left(\psi_{T_k-\lambda_k}\theta,\sigma^2,v\right)=r_k|v\sim gamma\left(\frac{1}{2}v,\frac{1}{2}v\right). \end{array} \tag{3.4}$$

The time delay sequence is described by a Markov chain. The Markov property means that the kth instant time delay is only dependent on the $(k-1)th$ instant time delay:

$$P(\lambda_k|\lambda_{k-1},...,\lambda_1)=P(\lambda_k|\lambda_{k-1}). \tag{3.5}$$

The hidden Markov chain is governed by a transition probability,

$$\alpha_{ij}=P(\lambda_k=i|\lambda_{k-1}=j), k=2,3...N, 1\leq i,j\leq d, \tag{3.6}$$

while the distribution of the initial time delay is

$$\pi_i=P(\lambda_1=i), 1\leq i\leq d. \tag{3.7}$$

3.3 Time-varying time delayed ARX Model Identification using the EM Algorithm

3.3.1 Formulation under EM algorithm

The actual value of the time delay λ_k, and variance scale r_k, are unknown. Hence, the EM algorithm [17] is employed to identify the system in Equation 1. For this purpose, the observed variables, missing variables and the parameters to be estimated are denoted as:

$$\begin{array}{l} C_{obs}=\{Y,U\}=\left\{y_{T_N},y_{T_{N-1}},\cdots,y_{T_1},u_{T_N},u_{T_N-1},\cdots,u_1\right\}, \\ C_{mis}=\{\Lambda,R\}=\{\lambda_N,\lambda_{N-1},\cdots,\lambda_1,r_N,r_{N-1},\cdots,r_1\}, \\ \Theta=\{\theta,\sigma^2,v,\alpha_{ij},\pi_i\}. \end{array} \tag{3.8}$$

The EM algorithm calculates the conditional expectation of the complete data likelihood and maximizes the expectation (Q function) with respect to the parameters iteratively, resulting in maximum likelihood estimation (MLE) of the parameters of interest. The mathematical formulation of the Q function can be derived as:

$$Q\left(\Theta|\Theta^h\right)=E_{\Lambda,R|Y,U,\Theta^h}\left\{logP\left(Y,U,\Lambda,R|\Theta\right)\right\}, \tag{3.9}$$

where Θ^h is the previous estimate of the parameters from the earlier iteration step and E denotes the expectation value. Using the chain rule, the complete data likelihood can be separated as follows:

$$Q(\Theta|\Theta^h)=E_{\Lambda,R|Y,U,\Theta^h}\left\{log\left[P\left(Y|\Lambda,R,U,\Theta\right)P\left(R|\Lambda,U,\Theta\right)P\left(\Lambda|U,\Theta\right)P\left(U|\Theta\right)\right]\right\}. \tag{3.10}$$

Observations do not depend on future information and are conditionally independent of each other given the historical data. Time delay only depends on the latest delay. In addition, the input is deterministic. Therefore, the four probability terms in Equation 3.10 cab be simplified as:

$$\begin{aligned}
&P\left(Y|\Lambda,R,U,\Theta\right)=P\left(y_{T_N:T_1}|\lambda_{N:1},r_{N:1},u_{N:1},\Theta\right)\\
&\qquad\qquad=\prod_{k=1}^{N}P\left(y_{T_k}|y_{T_{k-1}:T_1},u_{T_k:1},r_k,\lambda_k,\Theta\right)=\prod_{k=1}^{N}P\left(y_{T_k}|\psi_{T_k-\lambda_k},r_k,\lambda_k,\Theta\right),\\
&P\left(R|\Lambda,U,\Theta\right)=P\left(r_{N:1}|\lambda_{N:1},u_{N:1},\Theta\right)=\prod_{k=1}^{N}P\left(r_k|\Theta\right),\\
&P\left(\Lambda|U,\Theta\right)=P\left(\lambda_{N:1}|u_{N:1},\Theta\right)=\prod_{k=2}^{N}P\left(\lambda_k|\lambda_{k-1},\Theta\right)\times P\left(\lambda_1|\Theta\right),\\
&P\left(U|\Theta\right)=C.
\end{aligned}\tag{3.11}$$

Thus, the joint probability can be replaced by the multiplication of separate conditional probabilities:

$$Q\left(\Theta|\Theta^h\right)=E_{\Lambda,R|Y,U,\Theta^h}log\left[\begin{array}{l}\prod_{k=1}^{N}P\left(y_{T_k}|\psi_{T_k-\lambda_k},r_k,\lambda_k,\Theta\right)\times\prod_{k=1}^{N}P\left(r_k|\Theta\right)\\ \times\prod_{k=2}^{N}P\left(\lambda_k|\lambda_{k-1},\Theta\right)\times P\left(\lambda_1|\Theta\right)\times C\end{array}\right].\tag{3.12}$$

Logarithm helps multiplication to be simplified into summation,

$$Q\left(\Theta|\Theta^h\right)=E_{\Lambda,R|C_{obs},\Theta^h}\left\{\begin{array}{l}\sum_{k=1}^{N}logP\left(y_{T_k}|\psi_{T_k-\lambda_k},r_k,\lambda_k,\Theta\right)+\sum_{k=1}^{N}logP\left(r_k|\Theta\right)\\ +\sum_{k=2}^{N}logP\left(\lambda_k|\lambda_{k-1},\Theta\right)+logP\left(\lambda_1|\Theta\right)+logC\end{array}\right\}.\tag{3.13}$$

The expectation can be replaced by the multiplication of the conditional probability of the missing variables to the corresponding likelihood functions. Then the final Q function expression is composed of four terms corresponding to different parameters

$$Q(\Theta|\Theta^h)=Q_1(\theta,\sigma^2)+Q_2(v)+Q_3(\alpha_{ij})+Q_4(\pi_i)+logC.\tag{3.14}$$

where

$$\begin{aligned}
&Q_1(\theta,\sigma^2)=\sum_{k=1}^{N}\int_0^{\infty}\sum_{i=1}^{d}\left\{\begin{array}{l}P\left(r_k|y_{T_k:T_1},u_{T_k:1},\lambda_k=i,\Theta^h\right)P\left(\lambda_k=i|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)\\ logP\left(y_{T_k}|\psi_{T_k-\lambda_k},r_k,\lambda_k=i,\Theta\right)\end{array}\right\}dr_k,\\
&Q_2(v)=\sum_{k=1}^{N}\int_0^{\infty}\sum_{i=1}^{d}\left\{\begin{array}{l}P\left(r_k|y_{T_k:T_1},u_{T_k:1},\lambda_k=i,\Theta^h\right)P\left(\lambda_k=i|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)\\ logP\left(r_k|\Theta\right)\end{array}\right\}dr_k,\\
&Q_3(\alpha_{ij})=\sum_{k=2}^{N}\sum_{i=1}^{d}\sum_{j=1}^{d}P\left(\lambda_k=i,\lambda_{k-1}=j|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)logP\left(\lambda_k=i|\lambda_{k-1}=j,\Theta\right),\\
&Q_4(\pi_i)=\sum_{i=1}^{d}P\left(\lambda_1=i|y_{T_1},u_{T_1:1},\Theta^h\right)logP\left(\lambda_1=i|\Theta\right).
\end{aligned}\tag{3.15}$$

In $Q_1(\theta, \sigma^2)$ and $Q_2(v)$, the $P\left(y_{T_k}|\psi_{T_k-\lambda_k}, r_k, \lambda_k = i, \Theta\right)$ and $P(r_k|\Theta)$ are calculated by

$$\begin{aligned} P\left(y_{T_k}|\psi_{T_k-\lambda_k}, r_k, \lambda_k = i, \Theta\right) &= \frac{1}{\sqrt{2\pi\sigma^2/r_k}} \exp\left(-\frac{\left[y_{T_k}-\psi_{T_k-i}\theta\right]^2}{2\sigma^2/r_k}\right), \\ P\left(r_k|\Theta\right) &= \frac{(v/2)^{\frac{v}{2}}(r_k)^{\frac{v}{2}-1}}{\Gamma(v/2)} \exp\left(-\frac{v}{2}r_k\right). \end{aligned} \tag{3.16}$$

Accordingly,

$$\begin{aligned} &\log P\left(y_{T_k}|\psi_{T_k-\lambda_k}, r_k, \lambda_k = i, \Theta\right) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2}\log r_k - \frac{r_k}{2\sigma^2}(y_{T_k} - \psi_{T_k-i}\theta)^2, \\ &\log P\left(r_k|\Theta\right) = -\log\Gamma\left(v/2\right) + \frac{v}{2}\log\left(v/2\right) + \left(\frac{v}{2}-1\right)\log r_k - \frac{v}{2}r_k. \end{aligned} \tag{3.17}$$

In $Q_3(\alpha_{ij})$ and $Q_4(\pi_i)$, $P(\lambda_k = i|\lambda_{k-1} = j, \Theta)$ and $P(\lambda_1 = i|\Theta)$ are the transition probability and delay distribution at the initial time instant:

$$\begin{aligned} P(\lambda_k = i|\lambda_{k-1} = j, \Theta) &= \alpha_{ij}, \\ P(\lambda_1 = i|\Theta) &= \pi_i. \end{aligned} \tag{3.18}$$

3.3.2 Expectation step

To conduct the expectation calculation for the four Q function terms in Eqn.3.15, the following posteriors should be determined,

1. $P\left(r_k|y_{T_k:T_1}, u_{T_k:1}, \lambda_k = i, \Theta^h\right), 1 \le k \le N$
2. $P\left(\lambda_k = i|y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right), 2 \le k \le N$
3. $P\left(\lambda_k = i|y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right), 2 \le k \le N$
4. $P\left(\lambda_1 = i|y_{T_1}, u_{T_1:1}, \Theta^h\right)$

Considering Equation 3.17, $\log P\left(y_{T_k}|\psi_{T_k-\lambda_k}, r_k, \lambda_k = i, \Theta\right)$ is a linear function of r_k and $logr_k$. Therefore, the integration with respect to r_k in Eqn.3.15 can be implemented by taking the expectation over r_k and $logr_k$. This means that the integration computation converts to terms including $E\left(r_k|y_{T_k:T_1}, u_{T_k:1}, \lambda_k = i, \Theta^h\right)$ and $E(logr_k|y_{T_k:T_1}, u_{T_k:1}, \lambda_k = i, \Theta^h)$.

The Gamma distribution is the conjugate prior distribution over r_k, and hence the conditional posterior distribution of r_k follows a Gamma distribution as well,

$$r_k|\left(y_{T_k:T_1}, u_{T_k:1}, \lambda_k = i, \Theta^h\right) \sim gamma\left(\frac{v^h+1}{2}, \frac{v^h + \frac{1}{(\sigma^2)^h}\left(y_{T_k} - \psi_{T_k-i}\theta^h\right)^2}{2}\right). \tag{3.19}$$

Detailed derivation is explained in Appendix A. Therefore, we can get the expectation of the conditional posterior distribution over r_k and $logr_k$ according to the property of Gamma

distributions

$$
\begin{aligned}
&E\left(r_k|y_{T_k:T_1},u_{T_k:1},\lambda_k=i,\Theta^h\right)=\frac{v^h+1}{v^h+\frac{1}{(\sigma^2)^h}\left(y_{T_k}-\psi_{T_k-i}\theta^h\right)^2}\triangleq\bar{r}_{ki}^h\\
&E\left(\log r_k|y_{T_k:T_1},u_{T_k:1},\lambda_k=i,\Theta^h\right)\\
&=\Psi\left(\frac{v^h+1}{2}\right)-\log\left(\frac{v^h+\frac{1}{(\sigma^2)^h}\left(y_{T_k}-\psi_{T_k-i}\theta^h\right)^2}{2}\right)\\
&=\Psi\left(\frac{v^h+1}{2}\right)-\log\left(\frac{v^h+1}{2\bar{r}_{ki}^h}\right),
\end{aligned}
\tag{3.20}
$$

where $\Psi(v)$ is the derivative of the logarithm of the gamma function., i.e., $\Psi(v)=\frac{\partial\Gamma(v)}{\partial v}\frac{1}{\Gamma(v)}$.

The joint probability, $P\left(\lambda_k=i,\lambda_{k-1}=j|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)$ when $k\geq 2$, can be obtained by

$$
\begin{aligned}
&P\left(\lambda_k=i,\lambda_{k-1}=j|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)\\
&=\frac{\left\{\begin{array}{c}P\left(y_{T_k}|\lambda_k=i,\psi_{T_k-\lambda_k},\Theta^h\right)\times{\alpha_{ij}}^h\\ \times P\left(\lambda_{k-1}=j|y_{T_{k-1}:T_1},u_{T_{k-1}:1},\Theta^h\right)\end{array}\right\}}{\sum_{m=1}^{d}\sum_{n=1}^{d}\left\{\begin{array}{c}P\left(y_{T_k}|\lambda_k=m,\psi_{T_k-\lambda_k},\Theta^h\right)\times{\alpha_{mn}}^h\\ \times P\left(\lambda_{k-1}=n|y_{T_{k-1}:T_1},u_{T_{k-1}:1},\Theta^h\right)\end{array}\right\}},
\end{aligned}
\tag{3.21}
$$

Therefore, the conditional probability of λ_k when $k\geq 2$ is obtained as:

$$
P\left(\lambda_k=i|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)=\sum_{j=1}^{d}P\left(\lambda_k=i,\lambda_{k-1}=j|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)
\tag{3.22}
$$

and the conditional probability of λ_k when $k=1$ is obtained as:

$$
P\left(\lambda_1=i|y_{T_1},u_{T_1:1},\Theta^h\right)=\frac{P\left(y_{T_1}|\lambda_1=i,\psi_{T_1-\lambda_1},\Theta^h\right){\pi_i}^h}{\sum_{m=1}^{d}P\left(Z_1|\lambda_1=m,\psi_{T_1-\lambda_1},\Theta^h\right){\pi_m}^h}.
\tag{3.23}
$$

Detailed derivations of $P\left(\lambda_k=i,\lambda_{k-1}=j|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)$ when $k\geq 2$ and $P\left(\lambda_1=i|y_{T_1},u_{T_1:1},\Theta^h\right)$ are presented in Appendix B.

Considering the above derivations, the four terms $Q_1(\theta,\sigma^2)$, $Q_2(v)$, $Q_3(\alpha_{ij})$ and $Q_4(\pi_i)$ can be written as:

1. The term corresponding to the estimation of θ,σ^2

$$
Q_1(\theta,\sigma^2)=\sum_{k=1}^{N}\sum_{i=1}^{d}\left\{\begin{array}{l}P\left(\lambda_k=i|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)\\ \left[\begin{array}{l}-\log\sqrt{2\pi}-\frac{1}{2}\log\sigma^2+\frac{1}{2}\Psi\left(\frac{v^h+1}{2}\right)\\ -\frac{1}{2}\log\left(\frac{v^h+1}{2\bar{r}_{ki}^h}\right)-\frac{\bar{r}_{ki}^h}{2\sigma^2}(y_{T_k}-\psi_{T_k-i}\theta)^2\end{array}\right]\end{array}\right\}
\tag{3.24}
$$

2. The term corresponding to the estimation of v

$$
Q_2(v)=\sum_{k=1}^{N}\sum_{i=1}^{d}\left\{\begin{array}{l}P\left(\lambda_k=i|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)\\ \left[\begin{array}{l}-\log\Gamma(v/2)+\frac{v}{2}\log(v/2)-\frac{v}{2}\bar{r}_{ki}^h\\ +\left(\frac{v}{2}-1\right)\left\{\Psi\left(\frac{v^h+1}{2}\right)-\log\frac{v^h+1}{2\bar{r}_{ki}^h}\right\}\end{array}\right]\end{array}\right\}
\tag{3.25}
$$

3. The term corresponding to the estimation of α_{ij}

$$Q_3(\alpha_{ij}) = \sum_{k=2}^{N}\sum_{i=1}^{d}\sum_{j=1}^{d} P\left(\lambda_k = i, \lambda_{k-1} = j | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \tag{3.26}$$

4. The term corresponding to the estimation of π_i

$$Q_4(\pi_i) = \sum_{i=1}^{d} P\left(\lambda_1 = i | y_{T_1}, u_{T_1:1}, \Theta^h\right) \log(\pi_i) \tag{3.27}$$

3.3.3 Maximization step

A new estimate of the parameters is obtained by taking the derivative of the Q function over the parameters and equating them to zero. The Q function is divided into terms such that each consists of one set of the parameters. Therefore, the derivative can be taken in each individual term with respect to that set of parameters.

1. Model parameters:

$$\begin{aligned}
&\tfrac{\partial Q_1\left(\theta,\sigma^2\right)}{\partial\theta} = 0 \\
&\Rightarrow \theta^{h+1} = \left\{\sum_{k=1}^{N}\sum_{i=1}^{d} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \bar{r}_{ki}^h \psi_{T_k-i}{}^T \psi_{T_k-i}\right\}^{-1} \\
&\left\{\sum_{k=1}^{N}\sum_{i=1}^{d} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \bar{r}_{ki}^h \psi_{T_k-i}{}^T y_{T_k}\right\}
\end{aligned} \tag{3.28}$$

2. Scaling parameter:

$$\begin{aligned}
&\tfrac{\partial Q_1\left(\theta,\sigma^2\right)}{\partial\sigma^2} = 0 \\
&\Rightarrow \left(\sigma^2\right)^{h+1} = \frac{\sum_{k=1}^{N}\sum_{i=1}^{d} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \bar{r}_{ki}^h \left[y_{T_k} - \psi_{T_k-i}\theta\right]^2}{\sum_{k=1}^{N}\sum_{i=1}^{d} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)} \\
&= \tfrac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{d} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \bar{r}_{ki}^h \left[y_{T_k} - \psi_{T_k-i}\theta^{h+1}\right]^2
\end{aligned} \tag{3.29}$$

3. Degree of freedom

$$\begin{aligned}
&\tfrac{\partial Q_2(v)}{\partial v} = 0 \\
&\Rightarrow \sum_{k=1}^{N}\sum_{i=1}^{d} \left\{ \begin{array}{l} P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right) \\ \left[\begin{array}{l} -\Psi(v/2) + \log(v/2) + 1 - \bar{r}_{ki}^h \\ +\Psi\left(\frac{v^h+1}{2}\right) - \log\frac{v^h+1}{2\bar{r}_{ki}^h} \end{array}\right] \end{array} \right\} = 0
\end{aligned} \tag{3.30}$$

4. Transition probability of HMM:

When conducting the computation of α_{ij}, we need to consider the constraint that

Table 3.1: Procedure of expectation and maximization steps

Initialization. Set $h = 0$. Assign random values to Θ^h.
Do
E-step: Calculate $\bar{r}_{1i}^h$ and $P\left(\lambda_1 = i \middle| y_{T_1}, u_{T_1:1}, \Theta^h\right)$;
for $k = 2 : N$
Calculate $\bar{r}_{ki}^h$
and $P\left(\lambda_k = i, \lambda_{k-1} = j \middle| y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)$
and $P\left(\lambda_k = i \middle| y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)$;
M-step: Calculate $\left\{\theta, \sigma^2, v, \alpha_{ij}, \pi_i\right\} \stackrel{\Delta}{=} \Theta^{h+1}$;
until $\frac{\|\Theta^{h+1} - \Theta^h\|^2 - \|\Theta^h\|^2}{\|\Theta^h\|^2} < 0.001$.

$\sum_{j=1}^{d} \alpha_{ij} = 1$, thus we need to introduce Lagrange multiplier L_α,

$$\begin{aligned} &\frac{\partial}{\partial \alpha_{ij}} \left\{ Q_3\left(\alpha_{ij}\right) + L_\alpha \left(\sum_{j=1}^{d} \alpha_{ij} - 1 \right) \right\} = 0 \\ &\Rightarrow \left(\alpha_{ij}\right)^{h+1} = \frac{\sum_{k=2}^{N} P\left(\lambda_k = i, \lambda_{k-1} = j \middle| y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)}{\sum_{k=2}^{N} \sum_{j=1}^{d} P\left(\lambda_k = i, \lambda_{k-1} = j \middle| y_{T_k:T_1}, u_{T_k:1}, \Theta^h\right)} \end{aligned} \tag{3.31}$$

5. Initial probability of hidden delay:

   Considering the constraint that $\sum_{i=1}^{d} \pi_i = 1$, we introduce Lagrange multiplier L_π,

$$\begin{aligned} &\frac{\partial}{\partial \pi_i} \left\{ Q_4\left(\pi_i\right) + L_\pi \left(\sum_{i=1}^{q} \pi_i - 1 \right) \right\} = 0 \\ &\Rightarrow \left(\pi_i\right)^{h+1} = \frac{P\left(\lambda_1 = i \middle| y_{T_1}, u_{T_1:1}, \Theta^h\right)}{\sum_{i=1}^{d} P\left(\lambda_1 = i \middle| y_{T_1}, u_{T_1:1}, \Theta^h\right)} = P\left(\lambda_1 = i \middle| y_{T_1}, u_{T_1:1}, \Theta^h\right) \end{aligned} \tag{3.32}$$

Using Equations 3.28 to 3.32, the new estimate of the parameters is obtained and replaced as the initial parameter estimate in the next iteration of the EM algorithm. This procedure is continued until the parameters estimation converges to the ML solution. The calculation procedure is carried out as shown in Table 3.1.

3.4 Simulation Study

In this section, a numerical example is given to show the effectiveness of the proposed algorithm.

Consider the following dual rate system:

$$\begin{aligned} y_{T_k} &= 0.9y_{T_{k-1}} + 0.4u_{T_k-\lambda_k} + \upsilon_{T_k} \\ u &\sim N(0,1) \\ \upsilon &\sim N\left(0,\sigma^2\right) \\ \lambda_k &\in \{1,2,3,4\} \end{aligned} \tag{3.33}$$

where u, y and v are the input, output and measurement noise, respectively. $T_k = k\Delta$ is the sample time of output which is Δ times slower than the sample time of the input signal, u. In this example $\Delta = 5$. The input signal u is a normally distributed random variable with zero mean and unit variance. Measurement noise v follows a normal distribution with zero mean and σ^2 variance. Then we substitute part of the measurement noise by drift values between -5 and 5 in order to simulate the outliers. Delay is varying among {1,2,3,4}, in the form of a Markov chain. The true transition matrix governing the switching of delay is

$$A = \{\alpha_{ij}\} = \begin{pmatrix} 0.90 & 0.06 & 0.03 & 0.01 \\ 0.02 & 0.90 & 0.06 & 0.02 \\ 0.01 & 0.06 & 0.90 & 0.02 \\ 0.01 & 0.03 & 0.06 & 0.90 \end{pmatrix} \tag{3.34}$$

In simulation, $L = 500$ fast-rate inputs and $N = 100$ slow-rate outputs are collected for system identification. Figure 3.1 contains the fast-rate input data, true delay generated according to the transition probability, measurement noise, and output data. As we can see, the measurement noise has some drifting values, which introduces some outliers into the output measurement. The first 50 slow rate output samples and 250 fast rate input samples are used as the training data to estimate the parameters of the model and the rest are used to test the estimation.

Applying the proposed algorithm of Section 3 to the training data set for $\sigma^2 = 0.01$, the estimated parameters converge within 5 iterations of the EM algorithm as shown in Figure 3.2. The HMM transition probability is estimated as:

$$\hat{A} = \{\hat{\alpha}_{ij}\} = \begin{pmatrix} 0.8741 & 0.0634 & 0.0625 & 0.0001 \\ 0.0639 & 0.8176 & 0.0845 & 0.0340 \\ 0.0104 & 0.0849 & 0.8514 & 0.0533 \\ 0.0035 & 0.0606 & 0.1627 & 0.7732 \end{pmatrix}. \tag{3.35}$$

Using this transition matrix, the estimated delay of the training data set can be obtained by

$$\hat{\lambda}_k = arg\,\max_i \;\; P\left(\lambda_k = i | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \hat{\Theta}\right), \qquad \tfrac{N}{2} + 1 \le k \le N \;. \tag{3.36}$$

where $\hat{\Theta} = \{\hat{\theta}, \hat{\sigma}^2, \hat{v}, \hat{\alpha}_{ij}, \hat{\pi}_i\}$. The estimated delays from training data are illustrated by the first half of Figure 3.3, which also coincide with the true delay with an accuracy of 91%.

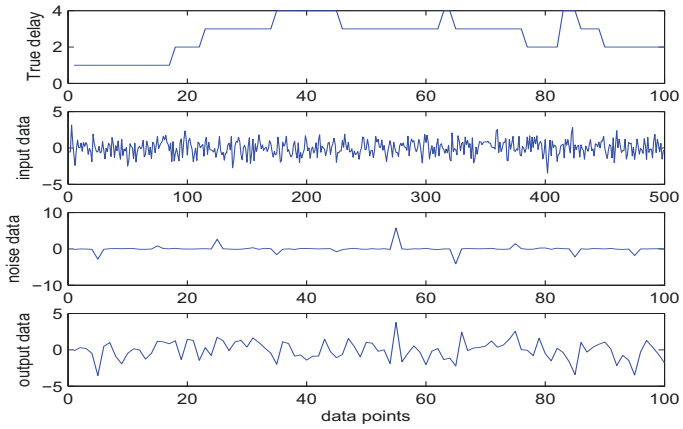

Figure 3.1: Simulation data

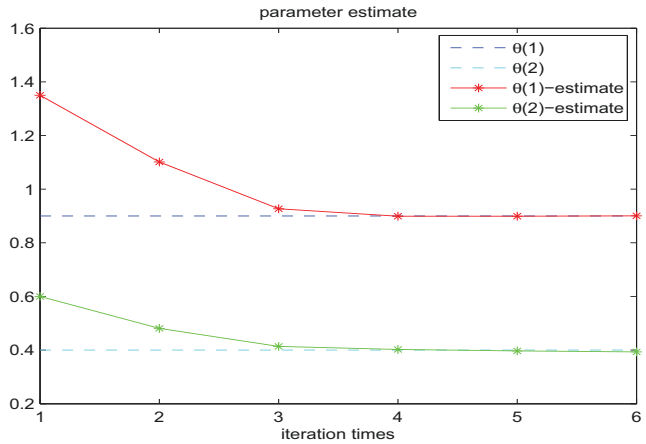

Figure 3.2: Parameters estimation

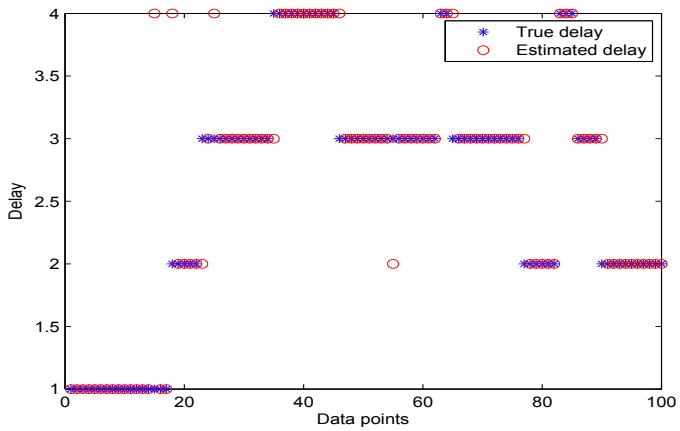

Figure 3.3: Delay estimation for the simulation system 3.33 (First 50 data points of the figure is from training data set, while last 50 data points is test data set)

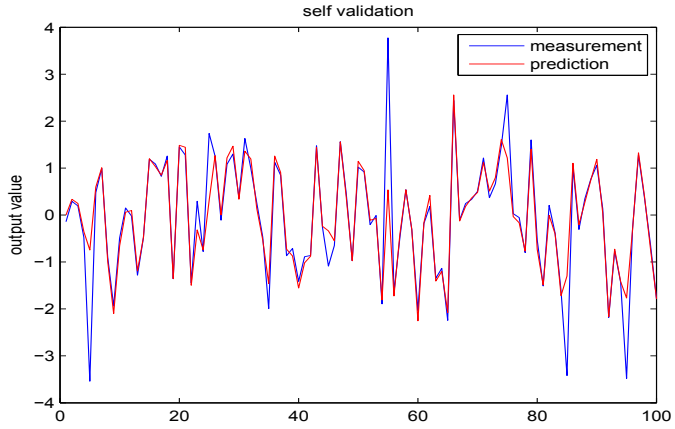

Figure 3.4: Output prediction of the system 3.33 (First 50 data points of the figure are from training data set, while last 50 data points are from the test data set)

The self validation result is illustrated in Figure 3.4, from which we can see that the outliers are successfully rejected.

The test data set is used to validate the performance of the estimation by predicting

$$\hat{\lambda}_k = arg\max_i \quad P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \hat{\Theta}\right), \qquad 1 \le k \le \frac{N}{2}, \tag{3.37}$$

given $\hat{\theta}$ and $\hat{\lambda}_k$. In order to have the prediction of time delay value, we need to calculate

the prior distribution of $\lambda_k = i$ up to the observation at time $k-1$,

$$\begin{aligned} & P\left(\lambda_k = i | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \hat{\Theta}\right) \\ &= \sum_{j=1}^{d} P\left(\lambda_k = i, \lambda_{k-1} = j | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \hat{\Theta}\right) \\ &= \sum_{j=1}^{d} \hat{\alpha}_{ij} P\left(\lambda_{k-1} = j | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \hat{\Theta}\right). \end{aligned} \tag{3.38}$$

The prior distribution is used to estimate the delay value at the kth time instant,

$$\hat{\lambda}_k = arg\, \underset{i}{max} \;\; P\left(\lambda_k = i | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \hat{\Theta}\right), \qquad \tfrac{N}{2} + 1 \le k \le N \;. \tag{3.39}$$

and then the prediction of $\hat{y}_{T_k}$ is computed. The prediction of the next observation needs the prior distribution of λ_{k+1}. It is computed based on the posterior distribution of $\lambda_k = i$ and expectation value of r_k given $\lambda_k = i$,

$$\begin{aligned} & P\left(\lambda_k = i | y_{T_k:T_1}, u_{T_k:1}, \hat{\Theta}\right) \\ &= \sum_{j=1}^{d} P\left(\lambda_k = i, \lambda_{k-1} = j | y_{T_k:T_1}, u_{T_k:1}, \hat{\Theta}\right) \\ &= \sum_{j=1}^{d} \frac{P\left(y_{T_k} | \lambda_k = i, y_{T_{k-1}:T_1}, u_{T_k:1}, \hat{\Theta}\right) \hat{\alpha}_{ij} P\left(\lambda_{k-1} = j | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \hat{\Theta}\right)}{\sum_{m=1}^{d} \sum_{n=1}^{d} P\left(y_{T_k} | \lambda_k = m, y_{T_{k-1}:T_1}, u_{T_k:1_k}, \hat{\Theta}\right) \hat{\alpha}_{mn} P\left(\lambda_{k-1} = n | y_{T_{k-1}:T_1}, u_{T_{k-1}:1}, \hat{\Theta}\right)} \end{aligned} \tag{3.40}$$

where the mean value of r_k given $\lambda_k = i$ is as follows,

$$E\left(r_k | y_{T_k:T_1}, u_{T_k:1}, \lambda_k = i, \hat{\Theta}\right) = \frac{\hat{v} + 1}{\hat{v} + \frac{1}{\hat{\sigma}^2}\left[y_{T_k} - \psi_{T_k - i}\hat{\theta}\right]^2} \triangleq \bar{r}_{ki} \tag{3.41}$$

The estimated time delay for the test data set is displayed in the second half of Figure 3.3. The cross validation result is illustrated in the second half of Figure 3.4.

In order to further test effectiveness of using Markov chain to model delay correlation, we consider a more realistic way to generate the time delays in the simulation. Consider delay is caused by transportation of materials in a pipe. The output measurement has varying time delay because of the varying flow rate. The flow rate is generated by passing a white noise sequence through a low-pass filter. The values of the transportation time delay are inversely proportional to the flow rates because of the fixed intersection area of the pipe. The delay values are rounded to the nearest integers {1,2,3,4} for discrete time system simulation. Using new simulated data, we apply the proposed method and compare it with three alternative methods. In the first method, the delay is supposed to follow a Markov chain but the measurement noise is considered to follow a Gaussian distribution, which is called regular Markov delay estimation. In the second method, the delay is supposed to change randomly but not follow a Markov chain, which is called independent delay

Table 3.2: A Summary of the Robust EM Estimation Performance (training set)

	$\sigma^2 = 0.01$		$\sigma^2 = 0.04$		$\sigma^2 = 0.09$	
	Accuracy*	RMSE	Accuracy*	RMSE	Accuracy*	RMSE
Markov delay (Robust)	91%	0.060	80%	0.153	68%	0.223
Markov delay (Regular)	80%	0.086	70%	0.192	58%	0.420
Independent delay (Regular)	70%	0.102	55%	0.318	47%	0.451
Fixed delay (Regular)	N/A	0.302	N/A	0.420	N/A	0.537

* Accuracy of delay estimation.

Table 3.3: A Summary of the Robust EM Estimation Performance (test set)

	$\sigma^2 = 0.01$		$\sigma^2 = 0.04$		$\sigma^2 = 0.09$	
	Accuracy*	RMSE	Accuracy*	RMSE	Accuracy*	RMSE
Markov delay (Robust)	90%	0.076	75%	0.189	65%	0.264
Markov delay (Regular)	76%	0.096	65%	0.226	55%	0.480
Independent delay (Regular)	68%	0.126	50%	0.360	40%	0.491
Fixed delay (Regular)	N/A	0.362	N/A	0.433	N/A	0.590

* Accuracy of delay estimation.

estimation. In the third method, the delay is supposed to be constant, which is called fixed delay estimation. In the implementation of independent delay estimation, the prior distribution of delay is considered to be uniformly distributed among {1,2,3,4} for each sample, while the EM algorithm is used for identification. In the fixed delay estimation, we consider that the delay for all samples is the same, and that the unknown value is uniformly distributed among {1,2,3,4}, and the identification is still by EM algorithm as well.

The performance of the three different methods for the training data set is given in Table 3.2, and for the test data set is listed in Table 3.3. The performance is compared under three different noise levels. In each level, the accuracy of the proposed hidden Markov model delay estimation is higher than both the independent delay and the fixed delay estimation methods, and the $RMSE$ is the smallest. When noise variance becomes large, the performance of all three methods degrades, however, the performance of the proposed robust HMM method is always better than the others.

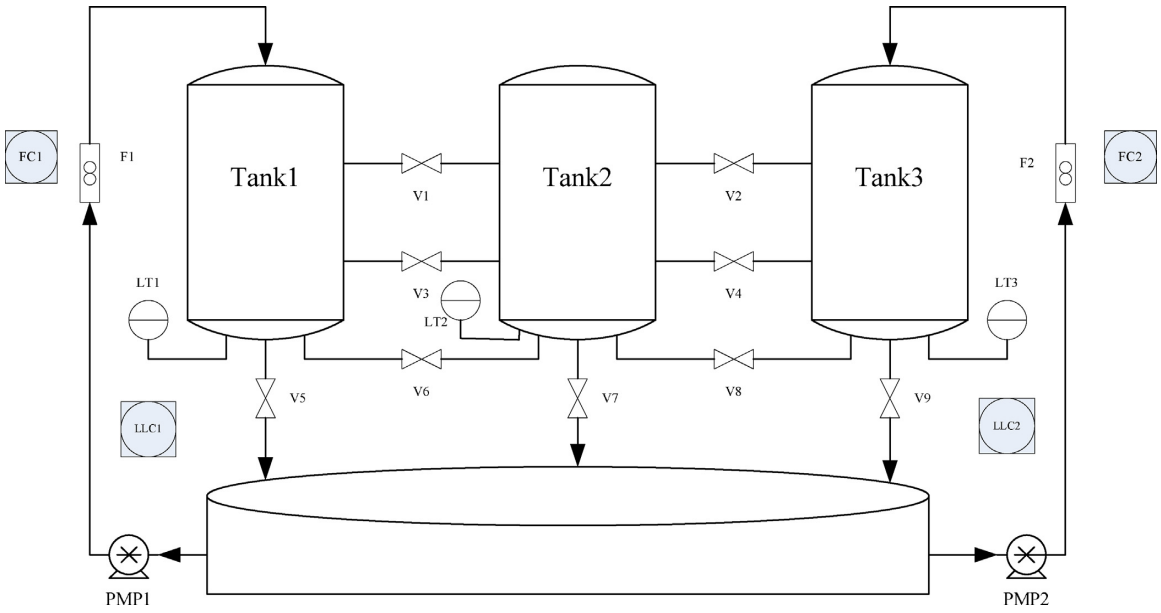

Figure 3.5: Schematic diagram of the hybrid tank system

3.5 Experimental Evaluation

A system identification experiment is designed and performed on a pilot-scale hybrid tank system. The schematic diagram of the plant is displayed in Figure 4.7. In the experiment, the right hand side tank, Tank3, and the middle tank, Tank2, are used. Therefore, the valves V1, V3, V5, V6 are close, and valves V2, V4, V7, V8, V9 are open. A basis input with amplitude $u = 5.5$ is introduced to the right hand side pump, PMP2. As a result, Tank3 water level turns to steady state at $y = 46$ after a period of time. Then a filtered random binary signal (RBS) with level $[-0.7, 0.7]$ is added to the input to stimulate the system and generate experimental data.

Input and output data are shown in Figure 3.6. The input is measured with a sampling rate of 16 seconds, while the output is measured with a sampling rate of 48 seconds associated with a manually imposed time delay, which is randomly varying between 16 seconds and 32 seconds. The switching of time delay follows a Markov chain with transition probability:

$$A = \{\alpha_{ij}\} = \begin{pmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{pmatrix}.$$

As shown in Figure 3.6, we also manually imposed some outliers to the output.

Considering that the delay follows a sequence of Markov chain, to validate the proposed algorithm in the identification of time invariant system, the data are normalized and divided into two halves; the first half being the training data set, and the second half being the test data set.

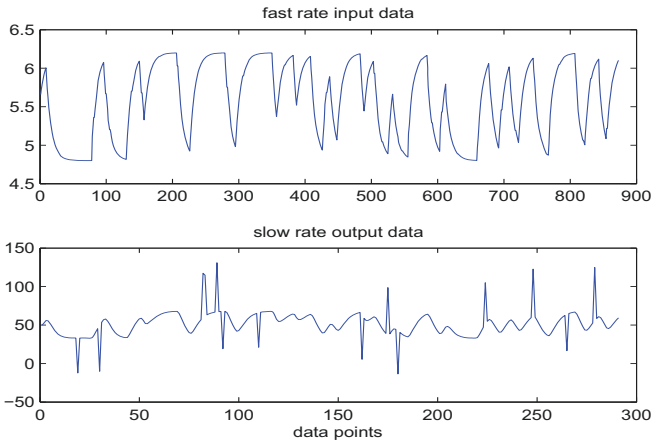

Figure 3.6: Input and output data

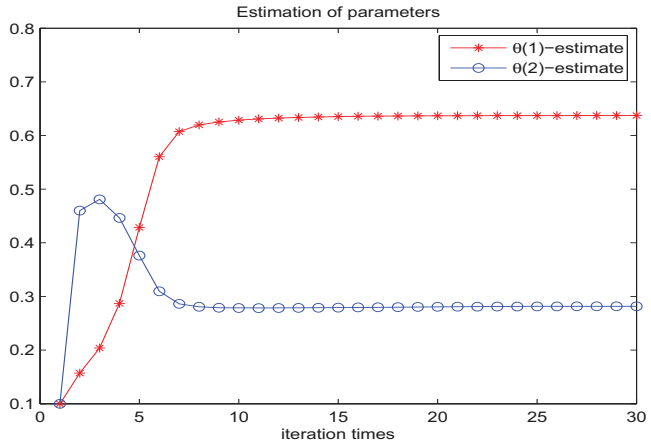

Figure 3.7: Iteration of parameter estimation

Consider the model structure for this plant to be a first order ARX model,

$$y_{T_k} = a y_{T_{k-1}} + b u_{T_k - \lambda_k}$$

Applying the proposed EM algorithm of Section 3, the regression parameters converge to $a = 0.6372$, $b = 0.2817$ after 10 iterations as shown in Figure 3.7, and the transition probability matrix is estimated as

$$\hat{A} = \{\hat{\alpha}_{ij}\} = \begin{pmatrix} 0.8560 & 0.1440 \\ 0.1550 & 0.8450 \end{pmatrix}$$

Using the procedure explained in Section 4, the delay sequence can be obtained. The delay estimation results are shown in Figure 3.8. The self and cross validation results are illustrated in Figure 3.9, which shows better accuracy of the estimation. It also shows that the model can reject the outliers.

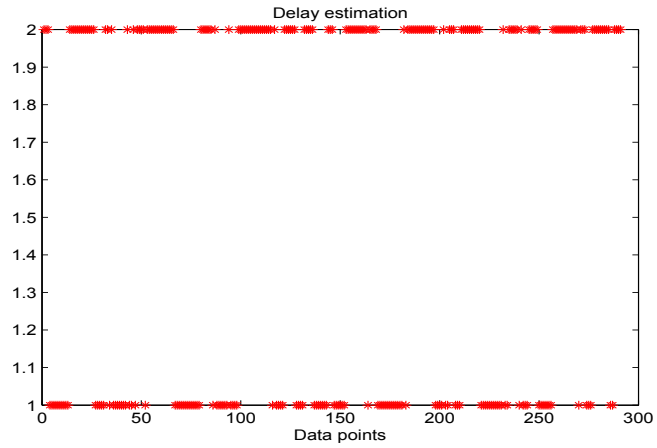

Figure 3.8: Delay estimation

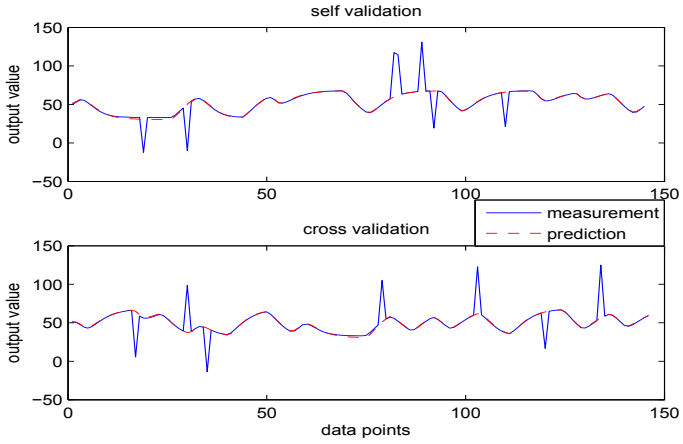

Figure 3.9: Self validation and cross validation

Table 3.4: A Summary of the RMSE in the Robust EM Estimation Experiment

	Self validation	Cross validation
Markov delay (Robust)	4.2592	5.5362
Markov delay (Regular)	7.6352	9.8651
Independent delay(Regular)	11.5352	13.5632
Fixed delay(Regular)	15.1253	17.7892

In order to show the advantages of modeling delay by a Markov chain and noise by the t-distribution, we compare the proposed method with the regular Markov delay estimation, the independent delay assumption method and fixed delay estimation method. Implementation details of these methods are illustrated in Section 4, where we regenerate time delays through correlated flow rates. The performance of the three different methods is listed in Table 4.4. We can see that the RMSE of the proposed method is the smallest compared to the other methods. This is because the structure of delay transition and noise distribution agree with the actual ones.

3.6 Conclusions

This chapter considers identification of ARX models with time varying delay in the presence of measurement outliers. Time varying delay is modeled by HMM and the measurement noise is modeled by t-distribution. In the framework of EM algorithm, the problem is solved by the proposed method. The parameters and unknown time delay are estimated, and the improved performance is demonstrated by both the numerical and experimental examples. Through comparison with three alternatives, the proposed method achieves the smallest RMSE for both the training and test data sets.

3.A Appendix A

The derivation of the conditional posterior distribution over r_k is as follows,

$$\begin{aligned}
&P\left(r_k|y_{T_k:T_1}, u_{T_k}{:}1, \lambda_k = i, \Theta^h\right) \\
&= \frac{P\Big(y_{T_k}, r_k|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k{=}i, \Theta^h\Big)}{P\Big(y_{T_k}|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k{=}i, \Theta^h\Big)} \\
&= \frac{P\Big(y_{T_k}|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k{=}i, r_k, \Theta^h\Big)P\Big(r_k|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k{=}i, \Theta^h\Big)}{\int_0^\infty P\Big(y_{T_k}|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k{=}i, r_k, \Theta^h\Big)P\Big(r_k|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k{=}i, \Theta^h\Big)dr_k} \\
&= \frac{P\Big(y_{T_k}|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k{=}i, r_k, \Theta^h\Big)P\big(r_k|\Theta^h\big)}{\int_0^\infty P\Big(y_{T_k}|y_{T_{k-1}:T_1}, u_{T_k:1}, \lambda_k{=}i, r_k, \Theta^h\Big)P\big(r_k|\Theta^h\big)dr_k}.
\end{aligned} \tag{3.42}$$

Substituting Equation 3.16 into the above equation, we have

$$\begin{aligned}
&P\left(r_k|y_{T_k:T_1},u_{T_k:1},\lambda_k=i,\Theta^h\right)\\
&=\frac{\frac{1}{\sqrt{2\pi\sigma^2/r_k}}\exp\left(-\frac{\left(y_{T_k}-\psi_{T_k-i}\theta^h\right)^2}{2(\sigma^2)^h/r_k}\right)*\frac{\left(v^h/2\right)^{\frac{v^h}{2}}(r_k)^{\frac{v^h}{2}-1}}{\Gamma(v^h/2)}\exp\left(-\frac{v^h}{2}r_k\right)}{\int_0^\infty\frac{1}{\sqrt{2\pi\sigma^2/r_k}}\exp\left(-\frac{\left(y_{T_k}-\psi_{T_k-i}\theta^h\right)^2}{2(\sigma^2)^h/r_k}\right)*\frac{(v/2)^{\frac{v^h}{2}}(r_k)^{\frac{v^h}{2}-1}}{\Gamma(v^h/2)}\exp\left(-\frac{v^h}{2}r_k\right)dr_k}.
\end{aligned}\tag{3.43}$$

To calculate the denominator, we refer to the following method,

$$\int_0^\infty(x)^{a-1}\exp\left(-bx\right)dx=\frac{1}{b^a}\int_0^\infty(bx)^{a-1}\exp\left(-bx\right)d\left(bx\right)=\frac{\Gamma\left(a\right)}{b^a},\tag{3.44}$$

so we can have

$$\begin{aligned}
&\int_0^\infty P\left(y_{T_k}|y_{T_{k-1}:T_1},u_{T_k:1},\lambda_k=i,r_k,\Theta^h\right)P\left(r_k|\Theta^h\right)dr_k\\
&=\frac{1}{\sqrt{2\pi\sigma^2}}\frac{\left(v^h/2\right)^{\frac{v^h}{2}}}{\Gamma\left(v^h/2\right)}\times\int_0^\infty(r_k)^{\frac{v^h+1}{2}-1}\exp\left(-r_k\frac{v^h+\frac{1}{(\sigma^2)^h}\left(y_{T_k}-\psi_{T_k-i}\theta^h\right)^2}{2}\right)dr_k\\
&=\frac{1}{\sqrt{2\pi\sigma^2}}\frac{\left(v^h/2\right)^{\frac{v^h}{2}}}{\Gamma\left(v^h/2\right)}\times\Gamma\left(\frac{v^h+1}{2}\right)\left(\frac{v^h+\frac{1}{(\sigma^2)^h}\left(y_{T_k}-\psi_{T_k-i}\theta^h\right)^2}{2}\right)^{-\frac{v^h+1}{2}},
\end{aligned}\tag{3.45}$$

and finally,

$$\begin{aligned}
&P\left(r_k|y_{T_k:T_1},u_{T_k:1},\lambda_k=i,\Theta^h\right)\\
&=\frac{1}{\Gamma\left(\frac{v^h+1}{2}\right)}\left(\frac{v^h+\frac{1}{\sigma^2}\left[y_{T_k}-\psi_{T_k-i}\theta\right]^2}{2}\right)^{\frac{v^h+1}{2}}(r_k)^{\frac{v+1}{2}-1}\exp\left(-r_k\frac{v^h+\frac{1}{\sigma^2}\left[y_{T_k}-\psi_{T_k-i}\theta\right]^2}{2}\right),
\end{aligned}\tag{3.46}$$

which is exactly

$$r_k|\left(y_{T_k:T_1},u_{T_k:1},\lambda_k=i,\Theta^h\right)\sim gamma\left(\frac{v^h+1}{2},\frac{v^h+\frac{1}{(\sigma^2)^h}\left(y_{T_k}-\psi_{T_k-i}\theta^h\right)^2}{2}\right).\tag{3.47}$$

3.B Appendix B

The derivation of $P\left(\lambda_k=i,\lambda_{k-1}=j|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)$ is as follows,

$$\begin{aligned}
&P\left(\lambda_k=i,\lambda_{k-1}=j|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)\\
&=\frac{P\left(y_{T_k},\lambda_k=i,\lambda_{k-1}=j|y_{T_{k-1}:T_1},u_{T_k:1},\Theta^h\right)}{\sum_{m=1}^{d}\sum_{n=1}^{d}P\left(y_{T_k},\lambda_k=m,\lambda_{k-1}=n|y_{T_{k-1}:T_1},u_{T_k:1},\Theta^h\right)}\\
&=\frac{B_{kij}}{\sum_{m=1}^{d}\sum_{n=1}^{d}B_{kmn}},
\end{aligned}\tag{3.48}$$

where, for notation simplicity,

$$B_{kij}=\left\{\begin{array}{l}P\left(y_{T_k}|\lambda_k=i,y_{T_{k-1}:T_1},u_{T_k:1},\Theta^h\right)\\P\left(\lambda_k=i|\lambda_{k-1}=j,y_{T_{k-1}:T_1},u_{T_k:1},\Theta^h\right)\\P\left(\lambda_{k-1}=j|y_{T_{k-1}:T_1},u_{T_k:1},\Theta^h\right)\end{array}\right\}.\tag{3.49}$$

The first term in B_{kij} can be further derived as

$$\begin{aligned}
&P\left(y_{T_k}|\lambda_k=i,y_{T_{k-1}:T_1},u_{T_k:1},\Theta^h\right)\\
&=P\left(y_{T_k}|\lambda_k=i,\psi_{T_k-\lambda_k},\Theta^h\right)\\
&=\int_0^\infty P\left(y_{T_k},r_k|\lambda_k=i,\psi_{T_k-\lambda_k},\Theta^h\right)dr_k\\
&=\int_0^\infty P\left(y_{T_k}|\lambda_k=i,r_k,\psi_{T_k-\lambda_k},\Theta^h\right)P\left(r_k|\Theta^h\right)dr_k,
\end{aligned} \tag{3.50}$$

which is exactly the Equation 3.45 in Appendicx B. The second term in B_{kij} is

$$P\left(\lambda_k=i|\lambda_{k-1}=j,y_{T_{k-1}:T_1},u_{T_k:1},\Theta^h\right)={\alpha_{ij}}^h. \tag{3.51}$$

Therefore, the joint posterior can be computed by

$$\begin{aligned}
&P\left(\lambda_k=i,\lambda_{k-1}=j|y_{T_k:T_1},u_{T_k:1},\Theta^h\right)\\
&=\frac{\left\{\begin{array}{c}P\left(y_{T_k}|\lambda_k=i,\psi_{T_k-\lambda_k},\Theta^h\right)\times{\alpha_{ij}}^h\\ \times P\left(\lambda_{k-1}=j|y_{T_{k-1}:T_1},u_{T_{k-1}:1},\Theta^h\right)\end{array}\right\}}{\sum_{m=1}^{d}\sum_{n=1}^{d}\left\{\begin{array}{c}P\left(y_{T_k}|\lambda_k=m,\psi_{T_k-\lambda_k},\Theta^h\right)\times{\alpha_{mn}}^h\\ \times P\left(\lambda_{k-1}=n|y_{T_{k-1}:T_1},u_{T_{k-1}:1},\Theta^h\right)\end{array}\right\}}.
\end{aligned} \tag{3.52}$$

The derivation of $P\left(\lambda_1=i|y_{T_1},u_{T_1:1},\Theta^h\right)$ is as follows,

$$\begin{aligned}
&P\left(\lambda_1=i|y_{T_1},u_{T_1:1},\Theta^h\right)\\
&=\frac{P\left(y_{T_1},\lambda_1=i|u_{T_1:1},\Theta^h\right)}{\sum_{m=1}^{d}P\left(y_{T_1}|u_{T_1:1},\Theta^h\right)}\\
&=\frac{P\left(y_{T_1}|u_{T_1:1},\lambda_1=i,\Theta^h\right)P\left(\lambda_1=i|\Theta^h\right)}{\sum_{m=1}^{d}P\left(y_{T_1},\lambda_1=m|u_{T_1:1},\Theta^h\right)P\left(\lambda_1=m|\Theta^h\right)},
\end{aligned} \tag{3.53}$$

where

$$\begin{aligned}
&P\left(y_{T_1}|u_{T_1:1},\lambda_1=i,\Theta^h\right)\\
&=P\left(y_{T_1}|\psi_{T_1-\lambda_1},\lambda_1=i,\Theta^h\right)\\
&=\int_0^\infty P\left(y_{T_1}|\psi_{T_1-\lambda_1},\lambda_1=i,r_1,\Theta^h\right)P\left(r_1|\Theta^h\right)dr_1\\
&=\frac{1}{\sqrt{2\pi\sigma^2}}\frac{(v/2)^{\frac{v^h}{2}}}{\Gamma\left(v^h/2\right)}\times\Gamma\left(\frac{v^h+1}{2}\right)\left(\frac{v^h+\frac{1}{(\sigma^2)^h}\left(y_{T_1}-\psi_{T_1-i}\theta^h\right)^2}{2}\right)^{-\frac{v^h+1}{2}},
\end{aligned} \tag{3.54}$$

and

$$P\left(\lambda_1=i|\Theta^h\right)={\pi_i}^h. \tag{3.55}$$

Finally the posterior distribution of the initial time delay is computed by

$$P\left(\lambda_1=i|y_{T_1},u_{T_1:1},\Theta^h\right)=\frac{P\left(y_{T_1}|\psi_{T_1-\lambda_1},\lambda_1=i,\Theta^h\right){\pi_i}^h}{\sum_{m=1}^{d}P\left(y_{T_1}|\psi_{T_1-\lambda_1},\lambda_1=m,\Theta^h\right){\pi_m}^h}. \tag{3.56}$$

Chapter 4

Robust Estimation of ARX Models with Time Varying Time Delays Using Variational Bayesian Approach

This work is concerned with robust identification of processes with time-varying time delays. In reality, the delay values do not simply change randomly, but there is a correlation between consecutive delays. In this work, the correlation of time delay is modeled by the transition probability of a Markov chain. Considering that the measured data are often contaminated by outliers, t-distribution is adopted to model the measurement noise. Furthermore, the variational Bayesian (VB) approach is applied to estimate the model parameters along with time delays. Compared with the classical expectation-maximization (EM) algorithm, VB approach has the advantage of capturing the uncertainty of the estimated parameter sand time delays by providing their full probabilities. The effectiveness of the proposed method is demonstrated by both a numerical example and a pilot-scale tank experiment.

4.1 Introduction

Time delay frequently occurs in many practical systems including chemical processes, transmission lines, and telecommunication. System identification in the presence of time delay has received much attention and has become one of the most active research subjects in many engineering fields. Moreover, in industrial processes, besides usual measurement noise, some collected data points may be distorted because of sensor fault or large disturbance in the data measurement. Thus, to ensure the reliability of the estimated model parameters, robust parameter estimation is essential.

Outliers occur in real process data and would affect the process identification significantly because they are different from typical process data [50]. A conventional approach to cope with potential outliers is to use the contaminated Gaussian distribution [51]. The outliers are taken into account when modeling the noise, where a Gaussian component with large variance is utilized to model the outliers. Jin et al. have used the contaminated Gaussian distribution to make their algorithm robust to outliers [52]. This solution is limited to a special type of outliers. A more general approach to model the effect of outliers is to use the t-distribution [43], which has longer tails than a Gaussian distribution. In Lu et. al's work [13], they considered the measurement noise modeled by a t-distribution, and the performance is improved compared with the noise that is modeled by a Gaussian distribution in the presence of outliers. However, their models have a single sampling rate and do not consider time-varying time delays. Time varying time delays pose a considerable challenges to identification as they introduce a hybrid identification problem.

Time delays are often due to transportation of materials in industrial processes. The estimation of time delay is an important topic in system identification. In chemical processes, time-varying delay estimation might be a more changeling problem. Techniques for time-varying delay estimation have been theoretically developed using adaptive filtering [53, 54] and quadratic convex approach [55, 56], where the delays were considered to vary between some known lower and upper bounds. Furthermore, the probability of the occurrence of time delay can be described by statistical models, such as hidden Markov model (HMM).

Multi-rate (MR) systems arise often in typical chemical processes due to the absence of online measurements for certain variables, which are usually sampled infrequently through off-line laboratory analysis, while other variables are readily measured at fast rate. The identification of MR systems with irregular output sampling in the presence of varying time delay has received increasing attention in recent years. The most frequently used technique to model MR sampled data is to down-sample the fast rate variables in accordance with the slow rate variables. However, the down-sampling technique has a critical drawback of information loss, in spite of being straightforward to implement in practice. Moreover, it leads to inaccurate models due to uncertain delay at every sampling point.

Expectation maximization (EM) algorithm [17] finds maximum likelihood (ML) estimates of parameters in an iterative method for data-driven models, where the model depends on some missing variables. A more advanced method for ML estimation in the presence of missing variable is variational Bayesian (VB). Compared with EM algorithm, the variational Bayesian (VB) approach can provide estimation of a posterior distribution

of parameters as well as the posterior distribution of latent variables [57]. Zhang et al. [58] derived a hierarchical Bayesian estimation using a variational Bayesian inference. Ma et al. [59] introduced a Bayesian estimation strategy to estimate the posterior distribution of the parameters in Dirichlet mixture model. The problem of parameter estimation in the Dirichlet mixture model is analytically intractable, due to the integral expressions of the gamma function and its corresponding derivatives. The variational Bayesian method is also widely used in many other research areas like continuous-discrete stochastic dynamic systems [60], nonlinear dynamical systems [61], and blind image de-convolution [62, 63].

In this chapter, the process is modeled by ARX model structure and the time delay is modeled by a hidden Markov model. To improve the robustness to outliers, the noise is modeled by t-distribution. The overall identification problem is formulated under the VB framework. A simulation example and an experiment verification demonstrate that the proposed method can provide more reliable identification results.

The remainder of this chapter is organized as follows. A detailed problem description on the ARX model identification in the presence of HMM for time delay is presented in the next section. Section 3 is dedicated to applying VB approach to solve this robust estimation problem. Then Section 4 gives a numerical example to validate the proposed method. In Section 5, a pilot-scale experiment is conducted to further validate the proposed method. The conclusion is given in the last section.

4.2 Problem Statement

Following is a dual rate ARX model with time varying time delays:

$$\begin{aligned} y_{T_k} &= \psi_{T_k-\lambda_k}\theta + e_{T_k} \\ \psi_{T_k-\lambda_k} &= \left[\begin{array}{cccc} y_{T_{k-1}} \cdots y_{T_{k-na}} & u_{T_k-\lambda_k} \cdots u_{T_k-nb-\lambda_k} \end{array}\right]. \end{aligned} \tag{4.1}$$

where $\{y_{T_k}, k = 1, 2, ..., N\}$ is the slow rate output variable, $\{u_t, t = 1, 2, ..., L\}$ is the fast rate input variable, and the slow rate sampling time is Δ times that of the fast rate ($L = \Delta * N$). Time delay $\{\lambda_k, k = 1, 2, ..., N\}$ is varying at every sampling instant. $\theta \in \mathbb{R}^{(na+nb+1)\times 1}$ is the parameter vector and $\psi_{T_k-\lambda_k} \in \mathbb{R}^{1\times(na+nb+1)}$ forms the regressor vector. e_t is associated measurement noise, and it is considered to follow a zero mean t-distribution, i.e. $e_t \sim t(0, \delta, v)$ with unknown variance precision, δ, and degrees of freedom, v.

The distribution of measurement noise indicates that the measured output also follows a t-distribution $y_{T_k} \sim t(\psi_{T_k-\lambda_k}\theta, \delta, v)$, which can be decomposed into scaled Gaussian distributions $N\left(\psi_{T_k-\lambda_k}\theta, \frac{1}{\delta r_k}\right)$, where r_k is an introduced latent variable that follows the Gamma

distribution $g\left(\frac{1}{2}v, \frac{1}{2}v\right)$, which is dependent on the degree of freedom v:

$$\begin{array}{l} t\left(y_{T_k}|\psi_{T_k-\lambda_k}, \theta, \delta, v\right) = \\ \int N\left(y_{T_k}|\psi_{T_k-\lambda_k}, \theta, \delta, r_k\right) g\left(r_k|v\right) dr_k. \end{array} \tag{4.2}$$

The time delay sequence is modeled by a Markov chain governed by a transition probability and a distribution of the initial time delay,

$$\begin{array}{l} \pi_i = P(\lambda_1 = i), 1 \leq i \leq d, \\ \alpha_{ij} = P(\lambda_k = i|\lambda_{k-1} = j), k = 2, 3, ..., N, 1 \leq i, j \leq d. \end{array} \tag{4.3}$$

Considering the above model, the identification problem is to estimate its parameters while the time delays are unknown. This problem is an ML problem under missing variables. The observed and missing variables are denoted as:

$$\begin{array}{l} C_{obs} = \{Z_N, Z_{N-1}, \cdots, Z_1\}, \\ C_{mis} = \{\Lambda, R\} = \{\lambda_N, \lambda_{N-1}, \cdots, \lambda_1, r_N, r_{N-1}, \cdots, r_1\}, \end{array} \tag{4.4}$$

where Z includes output and input for every time instant. The parameters to be estimated can be denoted as $\Phi = \{\Theta, v, \alpha_{ij}, \pi_i\}$, where the system parameters $\Theta = \{\theta, \delta\}$ are treated separately, because the uncertainty in these two parameters can be considered by directly assigning conjugate prior distributions and thus finding their posterior distributions. However for the hyper-parameters $\{v, \alpha_{ij}, \pi_i\}$ of hidden variables $\{\Lambda, R\}$, we will directly find the point estimates for them.

4.3 Time-varying time delayed ARX Model Identification using the VB Approach

4.3.1 Prior for the parameters

The joint prior distribution over all the model parameters Θ can be expressed as

$$P\left(\Theta\right) = P\left(\theta|b\right) P\left(\delta|c, d\right), \tag{4.5}$$

where b, c, d are constant values. The prior of regressor parameter θ is selected as a zero-mean Gaussian distribution, and we specify a Gamma prior over the precision δ:

$$\begin{array}{l} P\left(\theta|b\right) = N\left(0, bI_{na+nb+1)}\right), \\ P\left(\delta|c, d\right) = gamma\left(c, d\right), \end{array} \tag{4.6}$$

where $I_{na+nb+1}$ is an identity matrix, which has the same dimension as θ. The optimization of $\{v, \alpha_{ij}, \pi_i\}$ is treated separately from the model parameters.

4.3.2 Formulation under VB approach

The VB approach introduces free joint distributions $q(R,\Lambda)$ and $q(\Theta)$, also named as variational posterior, as an approximation of the distribution of the missing variables and parameters, to calculate the following log-likelihood:

$$\begin{aligned} &logP(C_{obs}) = \\ &\log \sum_{\Lambda} \int q(R,\Lambda)\, q(\Theta)\, \tfrac{P(C_{obs},R,\Lambda,\Theta|v,\alpha_{ij},\pi_i)}{q(R,\Lambda)q(\Theta)} dRd\Theta. \end{aligned} \tag{4.7}$$

Applying Jensen's inequality,

$$\begin{aligned} &logP(C_{obs}) \\ &\geq \sum_{\Lambda} \int q(R,\Lambda)\, q(\Theta) \log \tfrac{P(C_{obs},R,\Lambda,\Theta|v,\alpha_{ij},\pi_i)}{q(R,\Lambda)q(\Theta)} dRd\Theta \\ &\triangleq F\left[q(R,\Lambda), q(\Theta)\right]. \end{aligned} \tag{4.8}$$

Hence, we maximize the lower bound $F\left[q(R,\Lambda), q(\Theta)\right]$ instead of the original log-likelihood. Similar to regular EM algorithm, we also have VB expectation step or E-step and VB maximization step or M-step in the variational Bayesian approach. In the VB E-step, we maximize the lower bound with respect to the missing variable distribution $q(R,\Lambda)$ by fixing the parameter distribution $q(\Theta)$. In the VB M-step, we maximize the lower bound with respect to parameter distribution $q(\Theta)$ by fixing the missing variable distribution $q(R,\Lambda)$. This is an iterative procedure until the algorithm converges.

For the convenience of derivations in both VB E-step and M-step, the lower bound $F\left[q(R,\Lambda), q(\Theta)\right]$ can be further decomposed by the chain rule as follows:

$$\begin{aligned} &F\left[q(R,\Lambda), q(\Theta)\right] \\ &= \sum_{\Lambda} \int q(R,\Lambda)\, q(\Theta) \log P(C_{obs}|R,\Lambda,\Theta)\, dRd\Theta \\ &+ \sum_{\Lambda} \int q(R,\Lambda) \log P(R|v)\, dR \\ &+ \sum_{\Lambda} \int q(R,\Lambda) \log P(\Lambda|\alpha_{ij},\pi_i)\, dR \\ &+ \int q(\Theta) \log P(\Theta)\, d\Theta \\ &- \sum_{\Lambda} \int q(R,\Lambda) \log q(R,\Lambda)\, dR \\ &- \int q(\Theta) \log q(\Theta)\, d\Theta. \end{aligned} \tag{4.9}$$

The likelihood terms can be then decomposed into every time instant as

$$P(C_{obs}|R,\Lambda,\Theta) = \prod_{k=1}^{N} P(Z_k|Z_{k-1},\cdots,Z_1,r_k,\lambda_k=i,\Theta), \tag{4.10a}$$

$$P(R|v) = \prod_{k=1}^{N} P(r_k|v), \tag{4.10b}$$

$$P(\Lambda|\alpha_{ij},\pi_i) = \prod_{k=1}^{N} P(\lambda_k|\lambda_{k-1},\alpha_{ij},\pi_i), \tag{4.10c}$$

where the simplification on the condition is owing to independence of the relevant variables. The likelihood at the kth instant is:

$$\begin{aligned} &P\left(Z_k|Z_{k-1},\cdots,Z_1,r_k,\lambda_k=i,\Theta\right)= \\ &\frac{\sqrt{\delta r_k}}{\sqrt{2\pi}}\exp\left(-\frac{\delta r_k}{2}[y_{T_k}-\psi_{T_k-i}\theta]^2\right)\times C_U, \end{aligned} \tag{4.11a}$$

$$P(r_k|v)=\frac{(v/2)^{\frac{v}{2}}(r_k)^{\frac{v}{2}-1}}{\Gamma\left(v/2\right)}\exp\left(-\frac{v}{2}r_k\right), \tag{4.11b}$$

$$P\left(\lambda_k|\lambda_{k-1},\alpha_{ij},\pi_i\right)=\begin{cases}\alpha_{ij}, & k\geq 2\\ \pi_i, & k=1\end{cases}, \tag{4.11c}$$

where C_U is the profitability of the input, and it is treated as a constant value.

4.3.3 VB E-step

VB E-step maximizes the lower bound $F\left[q\left(R,\Lambda\right),q\left(\Theta\right)\right]$ with respect to $q\left(R,\Lambda\right)$ while $q\left(\Theta\right)$ is fixed in this step. The lower bound can be formulated as a function of $q\left(R,\Lambda\right)$ assuming $q\left(\Theta\right)$ is known.

The two terms $\int q\left(\Theta\right)\log P\left(\Theta\right)d\Theta$ and $\int q\left(\Theta\right)\log q\left(\Theta\right)d\Theta$ of the lower bound in Eqn. 4.9 are independent of R and Λ, so can be considered as a constant value, $C_{R,\Lambda}$, as shown

$$\begin{aligned} &F\left[q\left(R,\Lambda\right)q\left(\Theta\right)\right] \\ &=\sum_{\Lambda}\int q\left(R,\Lambda\right)\left\langle\log P\left(C_{obs}|R,\Lambda,\Theta\right)\right\rangle_{q(\Theta)}dR \\ &+\sum_{\Lambda}\int q\left(R,\Lambda\right)\log P\left(R|v\right)dR \\ &+\sum_{\Lambda}\int q\left(R,\Lambda\right)\log P\left(\Lambda|\alpha_{ij},\pi_i\right)dR \\ &-\sum_{\Lambda}\int q\left(R,\Lambda\right)\log q\left(R,\Lambda\right)dR+C_{R,\Lambda}, \end{aligned} \tag{4.12}$$

where $\left\langle .\right\rangle_{q(\Theta)}$ means the expectation operation over Θ. By solving $\max F\left[q\left(\Lambda,R\right),q\left(\Theta\right)\right]$ with respect to $q(R,\Lambda)$, such that $\sum_{\Lambda}\int q\left(R,\Lambda\right)dR=1$, we obtain:

$$q\left(R,\Lambda\right)=\frac{P\left(R|v\right)P\left(\Lambda|\alpha_{ij},\pi_i\right)\times e^B}{\sum_{\Lambda}\int P\left(R|v\right)P\left(\Lambda|\alpha_{ij},\pi_i\right)\times e^B dR}, \tag{4.13}$$

where, for notation simplicity, B is defined as

$$B=\left\langle\log P\left(C_{obs}|R,\Lambda,\Theta\right)\right\rangle_{q(\Theta)}. \tag{4.14}$$

By integrating R out of the joint density $q\left(R,\Lambda\right)$, we obtain the marginal density of time delay Λ,

$$q\left(\Lambda\right)=\frac{\int P\left(R|v\right)P\left(\Lambda|\alpha_{ij},\pi_i\right)\times e^B dR}{\sum_{\Lambda}\int P\left(R|v\right)P\left(\Lambda|\alpha_{ij},\pi_i\right)\times e^B dR}. \tag{4.15}$$

Then the conditional density $q(R|\Lambda)$ is obtained as

$$q(R|\Lambda) = \frac{q(R,\Lambda)}{q(\Lambda)} = \frac{P(R|v) \times e^B}{\int P(R|v) \times e^B dR}. \tag{4.16}$$

It is obvious that the variational posteriors depend on the log-likelihood of the observed data log-likelihood $\log P(C_{obs}|R,\Lambda,\Theta)$ in Eqn. 4.10a. Thus, e^B can be expressed as

$$\begin{aligned} & e^B \\ &= e^{\langle \log P(C_{obs}|R,\Lambda,\Theta) \rangle_{q(\Theta)}} \\ &= e^{\left\langle \sum_{k=1}^{N} \log P(Z_k|Z_{k-1},\cdots,Z_1,r_k,\lambda_k=i,\Theta) \right\rangle_{q(\Theta)}} \\ &= \prod_{k=1}^{N} e^{\langle \log P(Z_k|Z_{k-1},\cdots,Z_1,r_k,\lambda_k=i,\Theta) \rangle_{q(\Theta)}}. \end{aligned} \tag{4.17}$$

To simplify the expression, we again define

$$B_k = \langle \log P(Z_k|Z_{k-1},\cdots,Z_1,r_k,\lambda_k=i,\Theta) \rangle_{q(\Theta)}. \tag{4.18}$$

Variational posterior of R given Λ

As in Eqn. 4.16, the variational posterior of R given Λ depends on $P(R|v) \times e^B$ and $\int P(R|v) \times e^B dR$, which are expressed as

$$P(R|v) \times e^B = \prod_{k=1}^{N} P(r_k|v) \times e^{B_k}, \tag{4.19a}$$

$$\begin{aligned} & \int P(R|v) \times e^B dR = \int \prod_{k=1}^{N} P(r_k|v) \times e^{B_k} dR \\ &= \prod_{k=1}^{N} \int P(r_k|v) \times e^{B_k} dr_k, \end{aligned} \tag{4.19b}$$

and therefore

$$q(R|\Lambda) = \frac{P(R|v) \times e^B}{\int P(R|v) \times e^B dR} = \prod_{k=1}^{N} \frac{P(r_k|v) \times e^{B_k}}{\int P(r_k|v) \times e^{B_k} dr_k}. \tag{4.20}$$

On the other hand, with the *i.i.d.* assumption of R at each sampling time k, $q(R|\Lambda)$ can be decomposed as follows,

$$q(R|\Lambda) = \prod_{k=1}^{N} q(r_k|\lambda_k). \tag{4.21}$$

Now, considering the two kinds of decomposition in Eqn. 4.20 and Eqn. 4.21, we obtain

$$q(r_k|\lambda_k) = \frac{P(r_k|v) \times e^{B_k}}{\int P(r_k|v) \times e^{B_k} dr_k}, \tag{4.22}$$

which means that the variational posterior $q\left(r_k|\lambda_k\right)$ depends on the prior distribution of r_k and log-likelihood of the observed data at the kth instant:

$$\begin{aligned}&\log P\left(Z_k|Z_{k-1},\cdots,Z_1,r_k,\lambda_k=i,\Theta\right)=-\log\sqrt{2\pi}\\&+\log\sqrt{\delta}+\log\sqrt{r_k}-\tfrac{\delta r_k}{2}[y_{T_k}-\psi_{T_k-i}\theta]^2+\log C_U,\end{aligned}\tag{4.23}$$

and then the expectation of the log-likelihood at the kth instant with respect to the model parameter $\Theta=\{\theta,\delta\}$ is

$$B_k=-\log\sqrt{2\pi}+\tfrac{1}{2}\tilde{\delta}+\log\sqrt{r_k}-\tfrac{\bar{\delta}r_k}{2}g_{ki}+\log C_U,\tag{4.24}$$

where $\tilde{\delta}=\langle\log\delta\rangle_{q(\delta)}$, $\bar{\delta}=\langle\delta\rangle_{q(\delta)}$, $g_{ki}=y_{T_k}^2-2y_{T_k}\psi_{ki}\bar{\theta}+\psi_{ki}\langle\theta\theta'\rangle_{q(\theta)}{\psi_{ki}}'$ is the expectation of the quadratic term. In the expression of g_{ki}, $\bar{\theta}=\langle\theta\rangle_{q(\theta)}$, and ψ_{ki} stands for ψ_{T_k-i}.

Therefore, the variational posterior of r_k given $\lambda_k=i$ is calculated using Eqn. 4.11b and Eqn. 4.24, as shown,

$$\begin{aligned}&q\left(r_k|\lambda_k=i\right)\\&=\tfrac{P(r_k|v)\times exp(B_k)}{\int P(r_k|v)\times exp(B_k)dr_k}\\&=\tfrac{exp\left\{-r_k\frac{v+\bar{\delta}g_{ki}}{2}\right\}(r_k)^{\left(\frac{v+1}{2}-1\right)}}{\int exp\left\{-r_k\frac{v+\bar{\delta}g_{ki}}{2}\right\}(r_k)^{\left(\frac{v+1}{2}-1\right)}dr_k},\end{aligned}\tag{4.25}$$

where the terms that are irrelevant to the integration over r_k are extracted from B_k and canceled out over both the numerator and the denominator. After the computation of the integration,

$$q\left(r_k|\lambda_k=i\right)=\tfrac{1}{\Gamma\left(\frac{v+1}{2}\right)}\left(\tfrac{v+\bar{\delta}g_{ki}}{2}\right)^{\frac{v+1}{2}}(r_k)^{\frac{v+1}{2}-1}\exp\left(-r_k\tfrac{v+\bar{\delta}g_{ki}}{2}\right).\tag{4.26}$$

Clearly, the expression is in the form of a Gamma distribution,

$$r_k|\lambda_k=i\sim gamma\left(\frac{v+1}{2},\frac{v+\bar{\delta}g_{ki}}{2}\right).\tag{4.27}$$

We can get the expectation of the conditional posterior distribution over r_k and $logr_k$ according to the property of Gamma distributions,

$$\bar{r}_{ki}=\langle r_k\rangle_{q(r_k|\lambda_k=i)}=\frac{v+1}{v+\bar{\delta}g_{ki}},\tag{4.28a}$$

$$\tilde{r}_{ki}=\langle\log r_k\rangle_{q(r_k|\lambda_k=i)}=\Psi\left(\frac{v+1}{2}\right)-\log\left(\frac{v+\bar{\delta}g_{ki}}{2}\right),\tag{4.28b}$$

where $\Psi\left(v\right)$ is the derivative of the logarithm of the gamma function, i.e., $\Psi\left(v\right)=\frac{\partial\Gamma(v)}{\partial v}\frac{1}{\Gamma(v)}$.

Variational posterior of time delay Λ

Regarding the variational posterior distribution of time delay in Eqn. 4.15, different from the decomposition of $q(R|\Lambda) = \prod_{k=1}^{N} q(r_k|\lambda_k)$ in Eqn. 4.21, for the time delay Λ, because of the Markov property, the two consecutive time delays are dependent. In order to find the relationship between any two consecutive delay distributions, we need to investigate $q(\lambda_k = i, \lambda_{k-1} = j)$ and $q(\lambda_{k-1} = j)$ for $k \geq 2$ by analyzing Eqn. 4.15 as follows,

$$\begin{aligned}
&q(\lambda_k = i, \lambda_{k-1} = j) \\
&= \sum_{\lambda_{k-2:1}} q(\lambda_{k:1}) \\
&\propto \sum_{\lambda_{k-2:1}} \int P(r_{k:1}|v) P(\lambda_{k:1}) e^{B} dr_{k:1} \\
&\propto \int_{r_k} P(r_k|v) P(\lambda_k = i|\lambda_{k-1} = j) e^{B_k} dr_k \\
&\times \sum_{\lambda_{k-2:1}} \int P(r_{k-1:1}|v) P(\lambda_{k-1:1}) e^{B-B_k} dr_{k-1:1} \\
&\propto \int_{r_k} P(r_k|v) e^{B_k} dr_k \times \alpha_{ij} \times \\
&\sum_{\lambda_{k-2:1}} \int P(r_{k-1:1}|v) P(\lambda_{k-1:1}) e^{B-B_k} dr_{k-1:1},
\end{aligned} \tag{4.29a}$$

$$\begin{aligned}
&q(\lambda_{k-1} = j) \\
&= \sum_{\lambda_{k-2:1}} q(\lambda_{k-1:1}) \\
&\propto \sum_{\lambda_{k-2:1}} \int P(r_{k-1:1}|v) P(\lambda_{k-1:1}) e^{B-B_k} dr_{k-1:1}.
\end{aligned} \tag{4.29b}$$

By substituting the second expression into the first one, the relation is constructed by

$$\begin{aligned}
&q(\lambda_k = i, \lambda_{k-1} = j) \\
&\propto \int_{r_k} P(r_k|v) e^{B_k} dr_k \times \alpha_{ij} \times q(\lambda_{k-1} = j).
\end{aligned} \tag{4.30}$$

The distribution of the above equation is obtained by adding a normalizing term in the denominator, and thus,

$$\begin{aligned}
&q(\lambda_k = i, \lambda_{k-1} = j) \\
&= \frac{\int_{r_k} P(r_k|v) e^{B_k} dr_k \times \alpha_{ij} \times q(\lambda_{k-1} = j)}{\sum_{i=1}^{d} \sum_{j=1}^{d} \int_{r_k} P(r_k|v) e^{B_k} dr_k \times \alpha_{ij} \times q(\lambda_{k-1} = j)}.
\end{aligned} \tag{4.31}$$

Therefore, the delay distribution $q(\lambda_k = i)$ for $k \geq 2$ is obtained by

$$q(\lambda_k = i) = \sum_{j=1}^{d} q(\lambda_k = i, \lambda_{k-1} = j). \tag{4.32}$$

Specifically, for the initial time delay distribution,

$$\begin{aligned}
&q(\lambda_1 = i) \\
&= \frac{\int P(r_1|v) P(\lambda_1 = i) e^{B_1} dr_1}{\sum_{\lambda_1} \int P(r_1|v) P(\lambda_1 = i) e^{B_1} dr_1} \\
&= \frac{\int P(r_1|v) e^{B_1} dr_1 \times \pi_i}{\sum_{i=1}^{d} \int P(r_1|v) e^{B_1} dr_1 \times \pi_i}
\end{aligned} \tag{4.33}$$

where for any time instant k,

$$\begin{aligned}&\int P\left(r_k|v\right)e^{B_k}dr_k\\&=C_{r_k}\int\left\{-r_k\tfrac{v+\bar{\delta}g_{ki}}{2}\right\}\left(r_k\right)^{\frac{v-1}{2}}dr_k\\&=C_{r_k}\left(\tfrac{v+\bar{\delta}g_{ki}}{2}\right)^{-\frac{v+1}{2}}.\end{aligned}\tag{4.34}$$

4.3.4 VB M-step

VB M-step maximizes the lower bound $F\left[q\left(R,\Lambda\right),q\left(\Theta\right)\right]$ with respect to $q\left(\Theta\right)$ by fixing $q\left(R,\Lambda\right)$. The lower bound can be formulated as a function of $q\left(\Theta\right)$ assuming $q\left(R,\Lambda\right)$ is known a prior in this step. The lower bound in Eqn. 4.9 is rewritten as follows,

$$\begin{aligned}&F\left[q\left(R,\Lambda\right)q\left(\Theta\right)\right]\\&=\int q\left(\Theta\right)\left\langle\log P\left(C_{obs}|R,\Lambda,\Theta\right)\right\rangle_{q(R,\Lambda)}d\Theta\\&+\int q\left(\Theta\right)\log P\left(\Theta\right)d\Theta-\int q\left(\Theta\right)\log q\left(\Theta\right)d\Theta+C_{\Theta},\end{aligned}\tag{4.35}$$

where C_{Θ} represents the terms irrelevant to the calculation of $q\left(\Theta\right)$ and can be treated as a constant value. Again, for notation simplicity, in the first term of above equation, we define

$$D=\left\langle\log P\left(C_{obs}|R,\Lambda,\Theta\right)\right\rangle_{q(R,\Lambda)}.\tag{4.36}$$

By solving $\max F\left[q\left(\Lambda,R\right),q\left(\Theta\right)\right]$ with respect to $q(\Theta)$, such that $\int q\left(\Theta\right)d\Theta=1$, we obtain:

$$q\left(\Theta\right)=\frac{P\left(\Theta\right)e^{D}}{\int P\left(\Theta\right)e^{D}d\Theta},\tag{4.37}$$

where D can be computed as

$$D=\sum_{k=1}^{N}\sum_{i=1}^{d}q\left(\lambda_k=i\right)\left[\begin{array}{l}-\log\sqrt{2\pi}+\log\sqrt{\delta}+\frac{1}{2}\tilde{r}_{ki}\\-\frac{\delta\bar{r}_{ki}}{2}(y_{T_k}-\psi_{ki}\theta)^2+\log C_U\end{array}\right].\tag{4.38}$$

In order to obtain $q\left(\theta\right)$, specifically, we take first order functional derivative of $F\left[q\left(R,\Lambda\right),q\left(\Theta\right)\right]$ with respect to $q\left(\theta\right)$, with the constraint $\int q\left(\theta\right)d\theta=1$, and then get

$$q\left(\theta\right)=\frac{1}{C_{\theta}}P\left(\theta|b\right)e^{\left\langle D\right\rangle_{q(\delta)}},\tag{4.39}$$

where C_{θ} is a constant normalizing term, the prior distribution $P\left(\theta|b\right)$ and the hyper-parameter b is defined in Equation 4.6, and $\left\langle D\right\rangle_{q(\delta)}$ is the expectation of D in Eqn. 4.38

with respect to δ. Thus, we obtain the expression of $q(\theta)$,

$$\begin{aligned}
&q(\theta) \\
&= \tfrac{1}{C_\theta} P(\theta|b) \times \\
&\exp\left\{ \sum_{k=1}^{N} \sum_{i=1}^{d} q(\lambda_k = i) \left(-\tfrac{\bar{\delta}\bar{r}_{ki}}{2} \right) (y_{T_k} - \psi_{ki}\theta)^2 \right\} \\
&= \tfrac{1}{C_\theta} \exp\left\{ -\tfrac{1}{2b}\theta' I \theta \right\} \times \\
&\exp\left\{ \sum_{k=1}^{N} \sum_{i=1}^{d} q(\lambda_k = i) \left(-\tfrac{\bar{\delta}\bar{r}_{ki}}{2} \right) (y_{T_k} - \psi_{ki}\theta)^2 \right\} \\
&= \tfrac{1}{C_\theta} \exp\{ -\tfrac{1}{2}\theta' [b^{-1} I + \sum_{k=1}^{N} \sum_{i=1}^{d} q(\lambda_k = i) \times \\
&\bar{\delta}\bar{r}_{ki} \psi_{ki}{}' \psi_{ki}] \theta + \sum_{k=1}^{N} \sum_{i=1}^{d} q(\lambda_k = i)\, \bar{\delta}\bar{r}_{ki} y_{T_k} \theta' \psi_{ki}{}' \}.
\end{aligned} \tag{4.40}$$

This expression indicates that $q(\theta)$ is a Gaussian density function with mean and variance,

$$\bar{\theta} = var(\theta) \times \sum_{k=1}^{N} \sum_{i=1}^{d} q(\lambda_k = i)\, \bar{\delta}\bar{r}_{ki} y_{T_k} \psi_{ki}{}', \tag{4.41a}$$

$$var(\theta) = \left[b^{-1} I + \sum_{k=1}^{N} \sum_{i=1}^{d} q(\lambda_k = i)\, \bar{\delta}\bar{r}_{ki} \psi_{ki}{}' \psi_{ki} \right]^{-1}. \tag{4.41b}$$

Therefore,

$$\left\langle \theta\theta' \right\rangle_{q(\theta)} = var(\theta) + \bar{\theta}\bar{\theta}'. \tag{4.42}$$

Same procedure is followed to maximize $F[q(R,\Lambda), q(\Theta)]$ with respect to $q(\delta)$, with the constraint $\int q(\delta) d\delta = 1$. We obtain the expression of $q(\delta)$,

$$\begin{aligned}
&q(\delta) = \tfrac{1}{C_\delta} P(\delta|c,d)\, e^{\langle D \rangle_{q(\theta)}} \\
&= \tfrac{1}{C_\delta} P(\delta|c,d) \exp\left\{ \sum_{k=1}^{N} \sum_{i=1}^{d} q(\lambda_k = i) \left[\tfrac{1}{2}\log\delta - \tfrac{\delta\bar{r}_{ki} g_{ki}}{2} \right] \right\} \\
&= \tfrac{1}{C_\delta} \tfrac{d^c \delta^{c-1} \exp\{-d\delta\}}{\Gamma(c)} \delta^{N/2} \exp\left\{ \sum_{k=1}^{N} \sum_{i=1}^{d} q(\lambda_k = i) \tfrac{-\delta\bar{r}_{ki} g_{ki}}{2} \right\} \\
&= \tfrac{1}{C_\delta} \delta^{c + \frac{1}{2}N - 1} \exp\left\{ -\left[d + \tfrac{1}{2} \sum_{k=1}^{N} \sum_{i=1}^{d} q(\lambda_k = i)\, \bar{r}_{ki} g_{ki} \right] \delta \right\},
\end{aligned} \tag{4.43}$$

where C_δ is a constant normalizing term and the prior distribution $P(\delta|c,d)$ and c, d are defined in Equation 4.6. The expression obtained indicates that $q(\delta)$ is a Gamma density function with the following expected value,

$$\bar{\delta} = \frac{2c + N}{2d + \sum_{k=1}^{N} \sum_{i=1}^{d} q(\lambda_k = i)\, \bar{r}_{ki} g_{ki}}. \tag{4.44}$$

In order to obtain the estimate of the hyper-parameters $\{v, \alpha_{ij}, \pi_i\}$ for hidden variables

$\{\Lambda, R\}$, we rewrite the lower bound in Eqn. 4.9 as follows,

$$\begin{aligned}
&F\left[q\left(R,\Lambda\right)q\left(\Theta\right)\right] \\
&=\left\langle \log P\left(R|v\right)\right\rangle_{q(R|\Lambda)q(\Lambda)} \\
&+\left\langle \log P\left(\Lambda|\alpha_{ij},\pi_i\right)\right\rangle_{q(\Lambda)}+C_{v,\alpha_{ij},\pi_i} \\
&=\sum_{k=1}^{N}\sum_{i=1}^{d}q\left(\lambda_k=i\right)\left\{\begin{array}{l}-\log\Gamma\left(v/2\right)+\frac{v}{2}\log\left(v/2\right)\\+\left(\frac{v}{2}-1\right)\tilde{r}_{ki}-\frac{v}{2}\bar{r}_{ki}\end{array}\right\} \\
&+\sum_{k=2}^{N}\sum_{i=1}^{d}\sum_{j=1}^{d}q\left(\lambda_k=i,\lambda_{k-1}=j\right)log\alpha_{ij} \\
&+\sum_{i=1}^{d}q\left(\lambda_1=i\right)log\pi_i+C_{v,\alpha_{ij},\pi_i}.
\end{aligned} \tag{4.45}$$

For the degree of freedom v, solving the derivative of the lower bound $F\left[q\left(R,\Lambda\right)q\left(\Theta\right)\right]$ in Eqn. 4.45 with respect to v, we obtain,

$$\sum_{k=1}^{N}\sum_{i=1}^{d}q\left(\lambda_k=i\right)\left[\begin{array}{l}-\Psi\left(v/2\right)+\log\left(v/2\right)\\+1+\tilde{r}_{ki}-\bar{r}_{ki}\end{array}\right]=0. \tag{4.46}$$

From the above equation and using Matlab function, we can compute the value of v at the new iteration.

When conducting the computation of α_{ij} and π_i, we need to consider the constraint that $\sum_{j=1}^{d}\alpha_{ij}=1$ and $\sum_{i=1}^{d}\pi_i=1$, and then Lagrange multipliers L_α and L_π are introduced. Therefore, we obtain,

$$\begin{aligned}
&\frac{\partial}{\partial\alpha_{ij}}\left\{\begin{array}{l}\sum_{k=2}^{N}\sum_{i=1}^{d}\sum_{j=1}^{d}q\left(\lambda_k=i,\lambda_{k-1}=j\right)\log\left(\alpha_{ij}\right)\\+L_\alpha\left(\sum_{j=1}^{d}\alpha_{ij}-1\right)\end{array}\right\}=0 \\
&\Rightarrow\alpha_{ij}=\frac{\sum_{k=2}^{N}q(\lambda_k=i,\lambda_{k-1}=j)}{\sum_{k=2}^{N}\sum_{j=1}^{d}q(\lambda_k=i,\lambda_{k-1}=j)},
\end{aligned} \tag{4.47}$$

and

$$\begin{aligned}
&\frac{\partial}{\partial\pi_i}\left\{\sum_{i=1}^{d}q\left(\lambda_1=i\right)\log\left(\pi_i\right)+L_\pi\left(\sum_{i=1}^{q}\pi_i-1\right)\right\}=0 \\
&\Rightarrow\pi_i=\frac{q(\lambda_1=i)}{\sum_{i=1}^{d}q(\lambda_1=i)}=q\left(\lambda_1=i\right).
\end{aligned} \tag{4.48}$$

The VB E-step and VB M-step are iterated until the parameter estimation converges. To summarize, the algorithm is presented in Table 4.1.

Table 4.1: Procedure of VB E and VB M-steps

Initialization	Assign random values to b, c, d.	
	Compute	Equation
VB E-step	$q(r_k \lvert \lambda_k = i) \Rightarrow \bar{r}_{ki}, \tilde{r}_{ki}$	Eqn.4.26 $\Rightarrow$ Eqn.4.28a
	$q(\lambda_k) \rightleftarrows q(\lambda_k, \lambda_{k-1})$	Eqn.4.32 $\rightleftarrows$ Eqn.4.31
VB M-step	$q(\theta) \Rightarrow \bar{\theta}$	Eqn.4.40$\Rightarrow$ Eqn.4.41a
	$q(\delta) \Rightarrow \bar{\delta}$	Eqn.4.43$\Rightarrow$ Eqn.4.44
	v	Eqn.4.46
	α_{ij}	Eqn.4.47
	π_i	Eqn.4.48

4.4 Simulation Study

In this section, a numerical example is given to show the effectiveness of the proposed algorithm. Consider the following ARX process:

$$\begin{aligned} y_{T_k} &= 0.5y_{T_{k-1}} + 2u_{T_k-\lambda_k} + 1.5u_{T_k-1-\lambda_k} + v_{T_k} \\ u &\sim N(0,1) \\ v &\sim N\left(0,\sigma^2\right) \\ \lambda_k &\in \{1,2,3,4\} \end{aligned} \tag{4.49}$$

where u, y and v are the input, output and measurement noise, respectively. The ratio of the sampling time is $\Delta = 5$. Input signal u is a normally distributed random variable with zero mean and unit variance. Measurement noise v follows a normal distribution with zero mean and σ^2 variance. We substitute part of the measurement noise by large values between $[-20, -15] \cup [15, 20]$. Delay is varying in the form of a Markov chain. The true transition matrix governing the switching of delay is

$$A = \{\alpha_{ij}\} = \begin{pmatrix} 0.90 & 0.06 & 0.03 & 0.01 \\ 0.02 & 0.90 & 0.06 & 0.02 \\ 0.01 & 0.06 & 0.90 & 0.02 \\ 0.01 & 0.03 & 0.06 & 0.90 \end{pmatrix} \tag{4.50}$$

For the system identification, $L = 1000$ fast-rate inputs and $N = 200$ slow-rate outputs are collected, which are shown in Figure 4.1. It contains the fast-rate input data, true delay generated by the transition probability, measurement noise, and output data. As shown, output measurement has some drift values imposed intentionally to simulate outliers.

The algorithm proposed in Section 3 is applied while $\sigma^2 = 0.01$ and the constant values for b, c, d are randomly set to be positive, where b is a large value. Figure 4.2 shows the

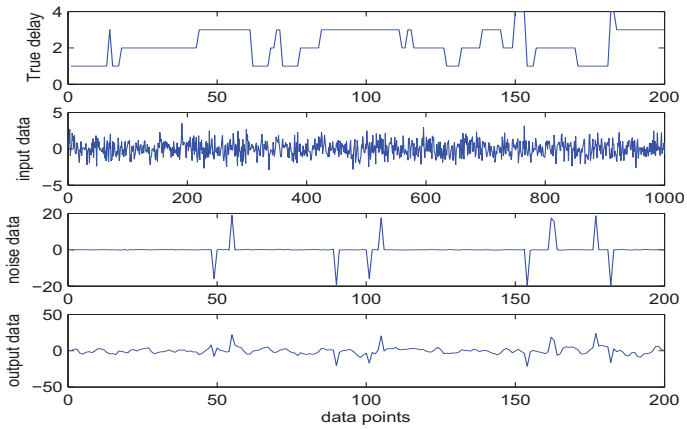

Figure 4.1: Simulation data

convergence of parameters to the true values. According to this figure, the parameters converge to their true values in around 15 iterations.

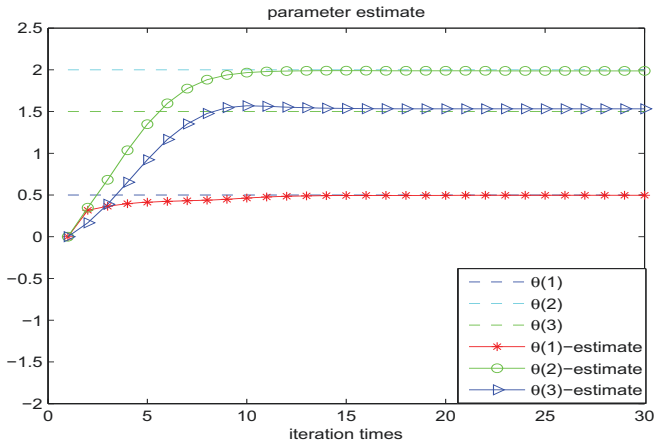

Figure 4.2: Parameters estimation

The estimated transition probability for the HMM is

$$\hat{A} = \{\hat{\alpha}_{ij}\} = \begin{pmatrix} 0.9280 & 0.0619 & 0.0100 & 0.0001 \\ 0.0271 & 0.9044 & 0.0207 & 0.0478 \\ 0.0640 & 0.1545 & 0.7806 & 0.0009 \\ 0.0004 & 0.0037 & 0.0439 & 0.9520 \end{pmatrix}, \tag{4.51}$$

which is very close to the real transition matrix. With the estimated probability distribution of time delay, the estimated time delay can be obtained through

$$\hat{\lambda}_k = \underset{i}{argmax} \;\; q(\lambda_k = i), \tag{4.52}$$

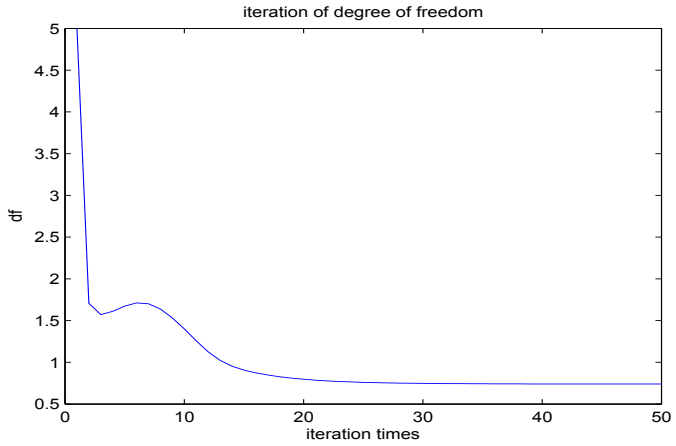

Figure 4.4: estimation of degree of freedom

The estimated delays are illustrated in Figure 4.3, which agree with the true delay with an accuracy of 92%.

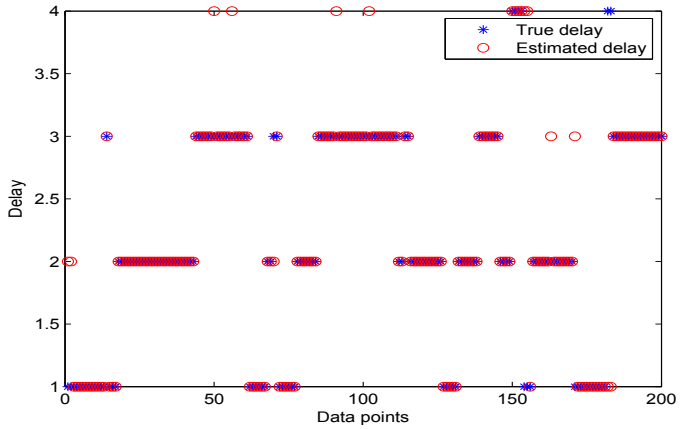

Figure 4.3: Delay estimation

The degree of freedom reflects the quality of data. If there are no outliers in the data, the degree of freedom is infinity, which means the measurement noise follows a Gaussian distribution. If there exist outliers, the more outliers there are, the smaller the degree of freedom becomes. The estimated degree of freedom is illustrated in Figure 4.4, in which the df converges to around 0.5893.

To show the advantage of adopting t-distribution to model the measurement noise, we compare the result with the approach of using Gaussian distribution to model measurement noise. In Figure 4.5, the noise data is not contaminated by outliers, thus a Gaussian distribution is sufficient to fit the noise data. It is noted that the Gaussian distribution

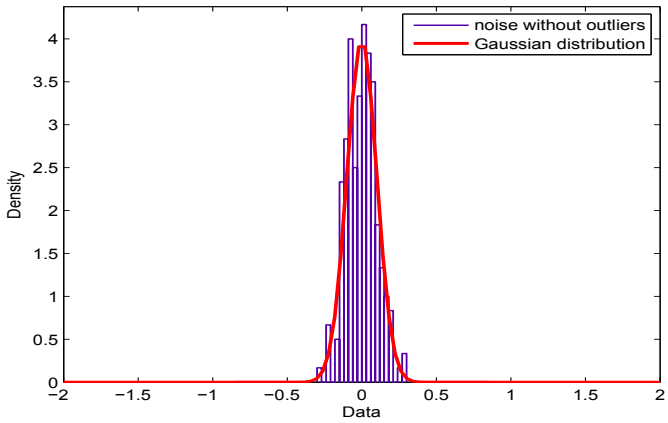

Figure 4.5: Fitting of noise without outliers

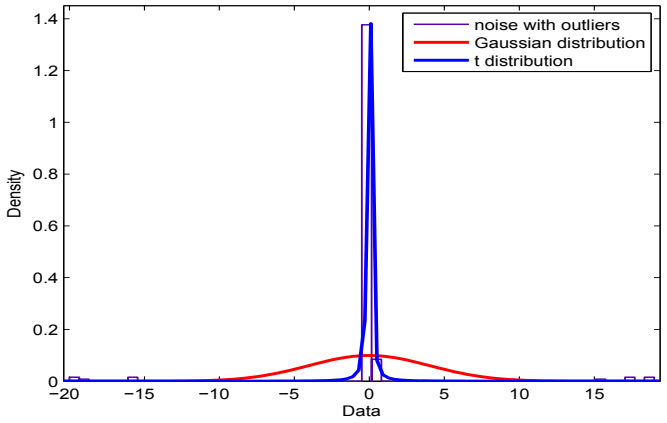

Figure 4.6: Fitting of noise with outliers

has short tails, and in this figure, the density is close to zero beyond the range $[-0.5, 0.5]$. However, if the noise contains outliers, as shown in Figure 4.6, a Gaussian distribution cannot fit the noise data while a t-distribution does very well.

In order to further test effectiveness of using Markov chain to model delay correlation, we consider a more realistic way to generate the time delays in the simulation. Consider delay is caused by transportation of materials in a pipe. The output measurement has varying time delay because of the varying flow rate. The flow rate is generated by passing a white noise sequence through a low-pass filter. The values of the transportation time delay are inversely proportional to the flow rates because of the fixed intersection area of the pipe. The delay values are rounded to the nearest integers {1,2,3,4} for discrete time system simulation. Using new simulated data, we apply the proposed method and compare

Table 4.2: A Summary of the Robust VB Estimation Performance

	$\sigma^2 = 0.01$		$\sigma^2 = 0.04$		$\sigma^2 = 0.09$	
	Accuracy*	RMSE	Accuracy*	RMSE	Accuracy*	RMSE
Markov delay (Robust)	92%	0.328	89%	0.650	85%	0.891
Markov delay (Regular)	83%	0.453	72%	0.853	61%	1.193
Independent delay (Regular)	72%	0.865	59%	1.128	46%	1.659
Fixed delay (Regular)	61%	1.023	52%	1.962	40%	2.641

* Accuracy of delay estimation.

it with three alternative methods, which are described in Section 4 of Chapter 3 (regular Markov delay estimation, independent delay estimation, and fixed delay estimation). In the implementation of independent delay estimation, the prior distribution of delay is uniformly distributed for each sample. In the fixed delay estimation, we consider that the delay for all samples is equal, and that the unknown value is uniformly distributed. The identifications are all carried out through VB approach as well.

Table 4.2 shows the performance of the three different methods, which is compared under three different noise levels. The root mean square error (RMSE) of the output prediction and the accuracy of time delay estimation of the proposed method are compared with the other methods. At each level, the accuracy of the proposed hidden Markov model delay estimation is highest, and the RMSE is consequently smaller than both the independent and fixed delay estimation methods. At larger noise levels, the performance of all three methods degrade, however, the performance of the proposed robust HMM method is always better than the other three.

The VB approach is an improvement of the EM algorithm by providing the parameter distribution instead of single point estimation. In order to show the difference of EM and VB, Table 4.3 displays the parameter estimation results and the RMSE of validation results. As the simulation results suggest, the VB has better parameter estimation accuracy than the EM, and the RMSE result of VB is smaller than that of EM. From the derivation of VB and EM, it is clear that VB can provide the posterior parameter distribution $q(\theta)$, while the EM only gives a point estimation $\hat{\theta}$ for parameter. This difference plays important role in the computation of the hidden variable (X) posteriors. In the EM-E step, we simply substitute the parameter point estimation $\hat{\theta}$ into the computation of posterior probability

Table 4.3: The Performance Comparison between EM and VB

	point estimation/mean value	RMSE
EM (Robust)	$\begin{bmatrix} 0.5056 \\ 2.1052 \\ 1.4752 \end{bmatrix}$	0.395
VB (Robust)	$\begin{bmatrix} 0.5033 \\ 2.0058 \\ 1.4938 \end{bmatrix}$	0.328

of hidden variable X.

$$P\left(X|C_{obs},\hat{\theta}\right)=\frac{P\left(C_{obs}|X,\hat{\theta}\right)P\left(X\right)}{\int\limits_{X}P\left(C_{obs}|X,\hat{\theta}\right)P\left(X\right)dX}. \tag{4.53}$$

The observation is often a quadratic form,

$$P\left(C_{obs}|X,\theta\right)=\frac{1}{\sqrt{2\pi\sigma^{2}}}\exp\left(-\frac{\left(y-\varphi\theta\right)^{2}}{2\sigma^{2}}\right), \tag{4.54}$$

which includes $\theta\theta^{T}$, and it is calculated as $\hat{\theta}\hat{\theta}^{T}$. In the VB-E step, we use the parameter distribution $q\left(\theta\right)$ to evaluate the posterior probability of hidden variables X. That is

$$q\left(X\right)=\frac{P\left(X\right)\exp\left\{\int\limits_{\theta}q\left(\theta\right)\log P\left(C_{obs}|X,\theta\right)d\theta\right\}}{\int\limits_{X}P\left(X\right)\exp\left\{\int\limits_{\theta}q\left(\theta\right)\log P\left(C_{obs}|X,\theta\right)d\theta\right\}dX}. \tag{4.55}$$

Here, in this computation, the mean value of the quadratic form $\theta\theta^{T}$ is computed as $E\left(\theta\theta^{T}\right)=\bar{\theta}\bar{\theta}^{T}+var\left(\theta\right)$, instead of simple $\hat{\theta}\hat{\theta}^{T}$ as in the EM. The $\hat{\theta}$ in the EM is not same as $\bar{\theta}$ in VB, and even though $\bar{\theta}\approx\hat{\theta}$, the EM cannot provide the variance of parameter estimation, $\mathrm{var}\left(\theta\right)$, so $\hat{\theta}\hat{\theta}^{T}\neq\bar{\theta}\bar{\theta}^{T}+\mathrm{var}\left(\theta\right)$. To summarize, the parameter estimation and missing variable estimation of VB are better than EM, and the reason is the parameter distribution obtained in VB rather than a point estimation as in EM.

4.5 Experimental Evaluation

The hybrid tank experiment is designed and performed in this section. The schematic diagram of the facility is displayed in Figure 4.7. In the experiment, the valves V1, V3, V5, V6 are close, and the valves V2, V4, V7, V8, V9 are open. Therefore, only the right hand side tank, Tank3, and the middle tank, Tank2, are used. A basis input with amplitude is

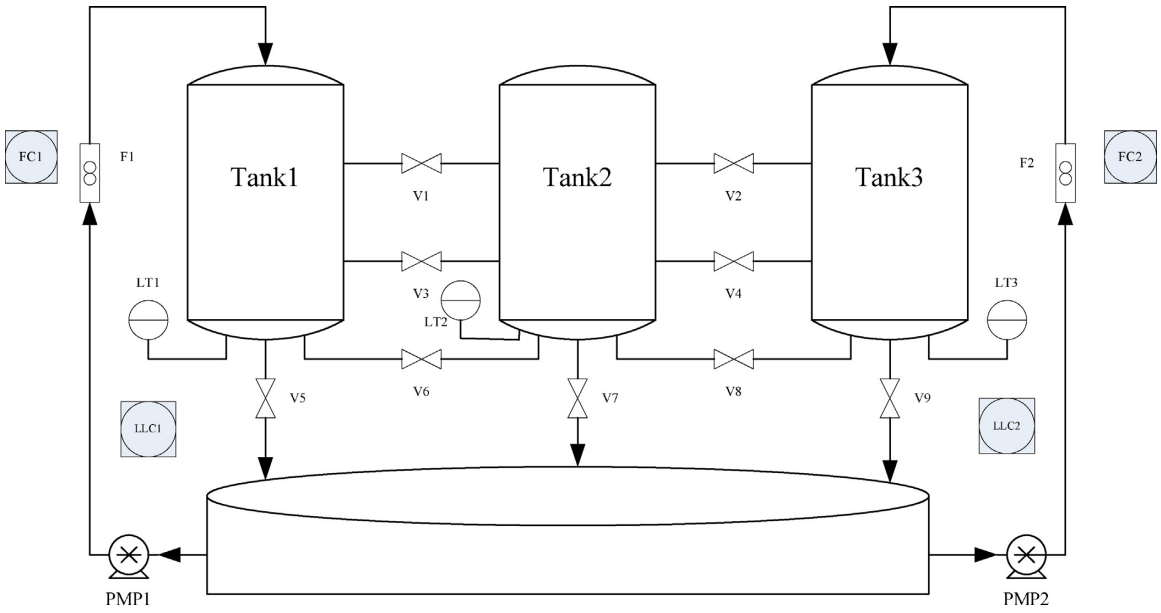

Figure 4.7: Schematic diagram of the hybrid tank system

introduced to the right hand side pump PMP2, and as a result, Tank3 water level becomes steady after a period of time. Then a filtered random binary signal (RBS) is added to the input to stimulate the system and generate experimental data.

The input and output data are shown in Figure 4.8. The input is measured with a sampling rate of 16 seconds, while the output is measured with a sample rate of 48 seconds associated with a manually imposed time delay, which is randomly varying between 16 seconds and 32 seconds. The switching of time delay follows a Markov chain with transition probability:

$$A = \{\alpha_{ij}\} = \begin{pmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{pmatrix}.$$

As shown in the Figure 4.8, we also imposed some outliers to the measurement of the output.

Modeling the delay by a Markov chain, we normalize and divide the data into two halves, the first half being the training data set, and the second half being the test data set. Consider the model structure for this plant is a first order ARX model,

$$y_{T_k} = a y_{T_{k-1}} + b u_{T_k - \lambda_k}$$

Applying the proposed VB approach of Section 3, the parameters of the model and delay transition are estimated. Figure 4.9 illustrates the convergence of the model parameters in around 15 iterations.

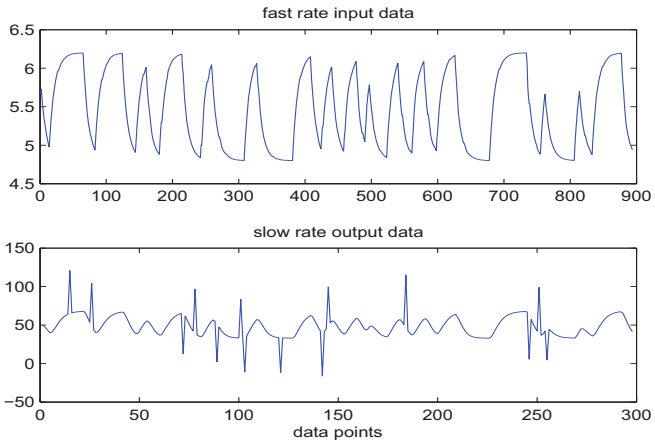

Figure 4.8: Input and output data

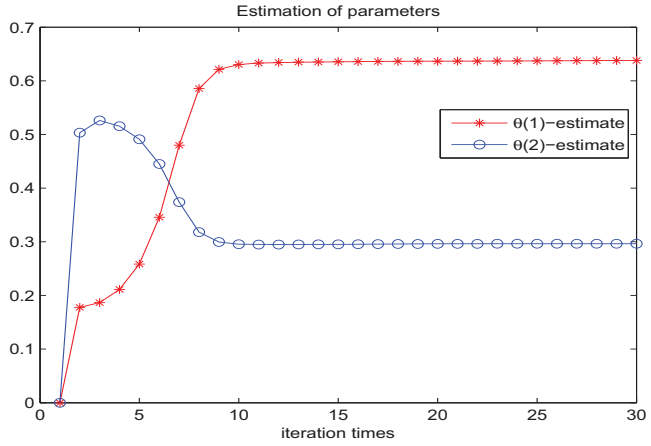

Figure 4.9: Parameters estimation for the experiment

The final estimated parameters are $a = 0.6379$, $b = 0.2963$, and the transition probability matrix is estimated as

$$\hat{A} = \{\hat{\alpha}_{ij}\} = \begin{pmatrix} 0.8513 & 0.1487 \\ 0.1452 & 0.8548 \end{pmatrix}$$

The estimated degree of freedom is illustrated in Figure 4.10, in which the df converges to around 0.7525.

Denoting the parameters estimated by the training data set as $\hat{\Phi} = \left\{\hat{\theta}, \hat{\delta}, \hat{v}, \hat{\alpha}_{ij}, \hat{\pi}_i\right\}$, we can also estimate the time delay of the test data set, predict the output of test data, and then compare it with the actual measurement. Firstly, the prior delay distribution of test data set is calculated by

$$q\left(\lambda_k = i\right) = \sum_{j=1}^{d} \hat{\alpha}_{ij} \times q\left(\lambda_{k-1} = j\right). \tag{4.56}$$

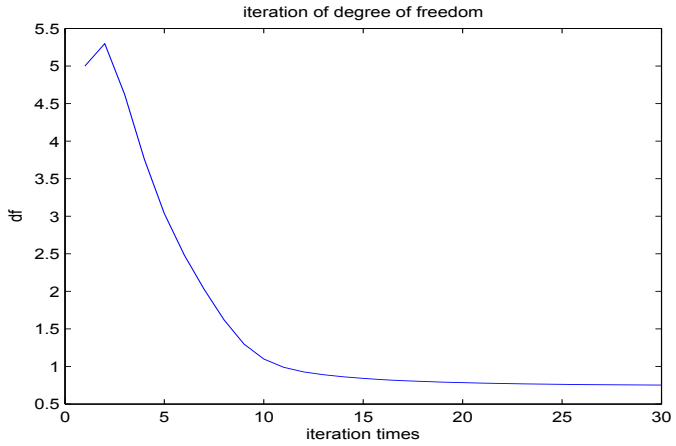

Figure 4.10: estimation of degree of freedom for the experiment

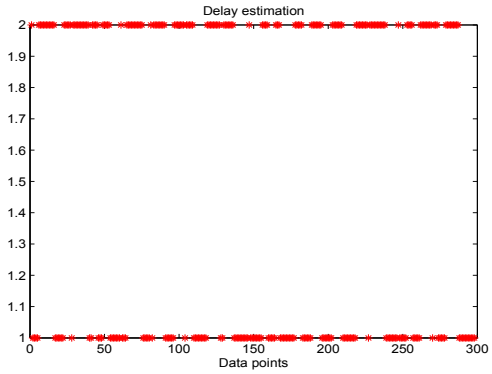

Figure 4.11: Delay estimation

This prior distribution is used to predict the delay value at the kth time instant by Eqn. 4.52. The delay sequence estimation results are shown in Figure 4.11.

In order to consider the uncertainty of time delay, the prediction of y_{T_k} is computed based on following expected prediction,

$$\hat{y}_{T_k} = \sum_{i=1}^{d} q\left(\lambda_k = i\right) \times \left[\begin{array}{cc} \hat{y}_{T_{k-1}} & u_{T_k - i} \end{array} \right] \hat{\theta}. \tag{4.57}$$

The self and cross validation results are illustrated in Figure 4.12, which show good performance of the estimation.

We also compare the proposed method with the regular Markov delay estimation, the independent delay assumption method and the fixed delay estimation method. The implementation of these three methods is illustrated in Section 4, where we regenerate time delays through correlated flow rates. The performance of the three different methods is listed in Table 4.4. We can see that the RMSE of the proposed method is the smallest.

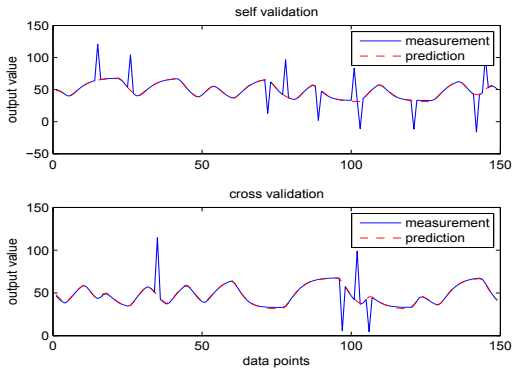

Figure 4.12: Self validation and cross validation

Table 4.4: A Summary of the RMSE in the Robust VB Estimation Experiment

	Self validation	Cross validation
Markov delay (Robust)	4.3203	5.4190
Markov delay (Regular)	7.8523	9.6985
Independent delay(Regular)	11.5284	13.1235
Fixed delay(Regular)	15.4258	17.6528

4.6 Conclusions

This chapter applies variational Bayesian approach for the identification problem with time varying delay in the presence of outliers. In the proposed method, time varying delay is modeled by HMM and the measurement noise is handled by using t-distribution. Owing to the flexibility of degrees of freedom of the t-distribution, the proposed method shows good resistance to the influence of outliers. In the framework of VB approach, an effective algorithm when identifying the ARX model was derived to estimate both distribution of time delay and model parameters. Compared with EM algorithm, the VB approach can provide parameter distribution instead of only the point estimation of the parameters. With the complete parameter distribution, we can obtain better estimation of the hidden variables' distributions. The advantage of the proposed method is demonstrated and verified by both numerical and experimental examples.

Chapter 5

Modeling of An Oil Sands Extraction Process with Time Delay

This chapter is concerned with constructing an input-output model for an oil sands extraction process. Time delay is studied in the chapter, which can be fixed or time varying. When it is fixed, we can estimate a constant value for the delay. However, when it is time varying and unknown at every sampling instant, we apply the methods developed in previous chapters to solve the hidden variable problem. The methods include Expectation Maximization (EM) algorithm, Variational Bayesian (VB) approach, and robust estimation using t-distribution.

5.1 Introduction

Knowing production rates including froth and bitumen production rates is important in the oil sands industry. Online measurements of these variables are available through hard sensors. However, it is desirable to have a process model which can compute the production rate (output) when given the input flow rate. The overall purpose of this work is to predict the froth production rate by the breaker feed rate, and then predict the bitumen production rate by the froth rate.

Simple linear regression is a common method to construct data-driven models. Using this method, time delay can only be considered as a constant value. Although unknown, it may be figured out by trying different values and selecting the one that gives the best fit. However, the real time delay could be varying, hence, other advanced methods should be considered. As in chapters 2, 3, and 4, we developed different algorithms to solve the time varying and unknown time delay problem. Chapter 2 considers that the time delay is

time varying and uses a Markov chain model to present the time delay correlation. Chapter 3 further considers the presence of measurement outliers, which is common in industrial data. Using t-distribution to model measurement noise, we add robustness in the parameter estimation. Both chapters 2 and 3 are based on the EM algorithm, which is well known for the sensitivity to initialization of parameters. When the initial guess of parameters is far from the true value, the algorithm can converge to some other local optimal value. In Chapter 4, the robust estimation algorithm is based on VB approach. VB initializes a prior distribution for the parameters, iteratively computes the posterior distribution and ultimately finds the global optimal value for the parameters. Besides, compared with EM algorithm, the VB approach can provide parameter distribution in addition to the hidden variable distribution, while in the EM algorithm, only an point estimation of the model parameter can be obtained while the estimation error or the estimation variance cannot be obtained automatically. Therefore, the VB approach considers the uncertainty of the parameter estimation.

The remainder of this chapter is organized as follows. A simplified process description for the oil sands extraction and data analysis are presented in the next section. The following section will apply algorithms developed in chapters 2, 3, and 4 to solve this process modeling problem. The conclusion is given in the final section.

5.2 Process description and data analysis

5.2.1 Process description

The main objective of the oil sands recovery process is to separate bitumen from other components, which are mainly water and solids, through a chain of extraction processes. The oil sands are first mixed with hot water in breakers and the resulting slurry is then fed into a Primary Separation Vessel (PSV) to facilitate bitumen flotation and sand settling (upper part of Figure 5.1). The impure bitumen froth floats to the top of the PSV and is further treated in the froth treatment plant to remove residual water and fine solids, after which bitumen production is obtained (lower part of Figure 5.1).

The oil sands extraction models contain the following two sub-models:

1. Predicting froth production using breaker feed rates.

2. Predicting diluted bitumen rate using froth rate.

Here, we only choose the first one as an example because another sub-model can be constructed similarly. Since that the extraction process involves long pipes, there is time delay

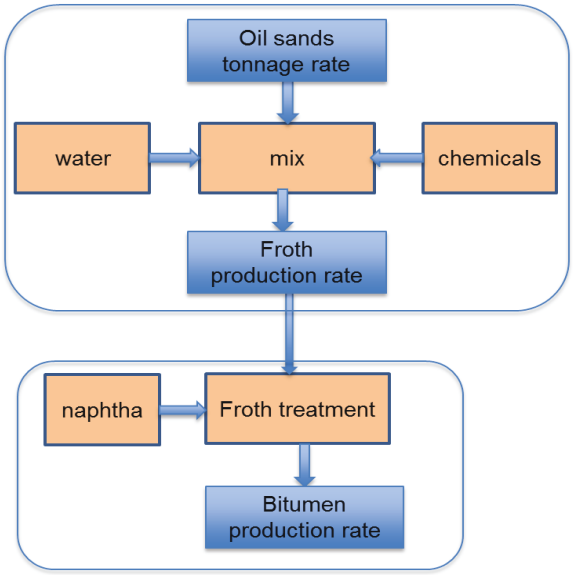

Figure 5.1: Diagram of the oil sands extraction process

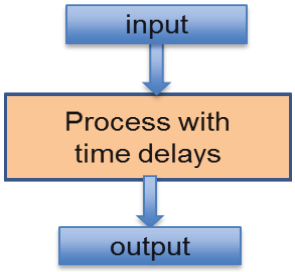

Figure 5.2: Simplified diagram of the oil sands primary extraction

from the input and output, which is shown as a simplified diagram in Figure 5.2.

5.2.2 Process Data Analysis

A list of the influential process variables for the plant is presented in Table 5.1. These variables have been identified by exploiting analytic knowledge as well as considering the availability of measuring devices. The real-time measurements are recorded every minute. For proprietary reasons, the data appearing in this chapter is normalized. The data in hand is from the first week of January, 2015, which contains 10,080 data points.

Since the three inputs $x1, x2, x3$ are not independent but processed together in separation cells, they are added together, and so are the three outputs $y1, y2, y3$. Figure 5.3 illustrates the added input x and added output y.

Table 5.1: A Summary of the Influential Process Variables

Process Variable	Symbol
$x1$	Ore feed rate to Plant
$x2$	Ore feed rate to Plant
$x3$	Ore feed rate to Plant
$y1$	Froth production rate from Plant
$y2$	Froth production rate from Plant
$y3$	Froth production rate from Plant

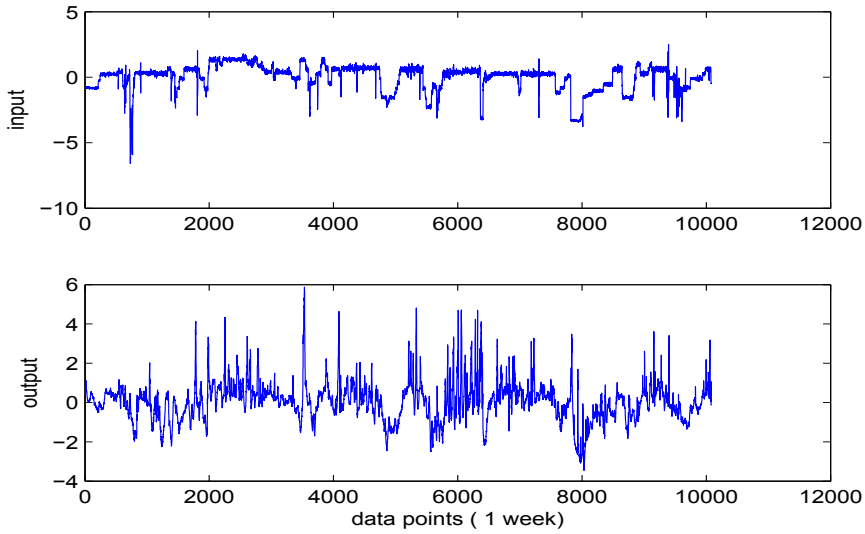

Figure 5.3: Process data (every 1 minute)

5.2.3 Performance evaluation

Generally, the model validation is executed by evaluating the accuracy of the proposed method using a separate evaluation data set. The accuracy means agreement between the predicted and target values. The prediction errors between the predicted and target values are also referred to as residuals. The graphical techniques used in analysis of residuals are listed as:

- *Scatter plot of predicted values v.s. measurement*: The ideal case would be for all the data points to lie on the $y = x$ line, indicating perfect agreement between the predicted and measured values.
- *Run-sequence plot of predicted and measured values*: The time trends of the predicted and measured values are plotted together to visually assess the accuracy and reliability of the constructed model.

The prediction can be evaluated quantitatively using root mean square error (RMSE) criteria, which indicates the overall prediction performance in terms of both accuracy and reliability:

$$\begin{aligned} & RMSE \\ &= \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(y_{T_k} - \hat{y}_{T_k}\right)^2} \\ &= \sqrt{\frac{1}{N}\sum_{k=1}^{N}\varepsilon_k^2}. \end{aligned} \tag{5.1}$$

5.3 Process modeling under different methods

The modeling of the oil sands extraction process is carried out by four different methods: the least squares regression (LSR), the regular EM proposed in Chapter 2, the robust EM proposed in Chapter 3 and the robust VB proposed in Chapter 4. The process model is selected as the first order ARX model:

$$y_{T_k} = a y_{T_{k-1}} + b u_{T_k - \lambda_k}. \tag{5.2}$$

Since that the least squares regression cannot consider the probability for time delay taking different values, the time delay is estimated by a constant value $\hat{\lambda}_k$ through the process. Therefore, the prediction is computed as given below,

$$\hat{y}_{T_k} = \frac{b}{1 - az^{-1}} u_{T_k - \hat{\lambda}_k}. \tag{5.3}$$

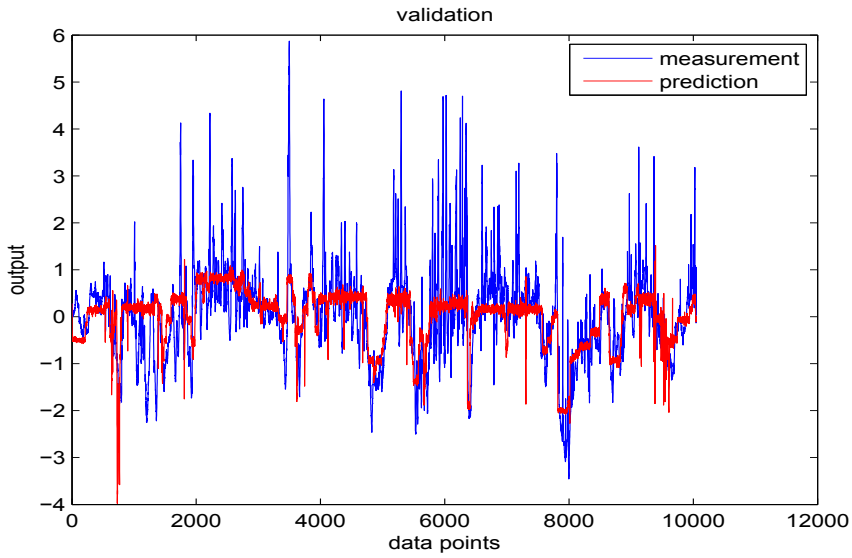

Figure 5.4: Validation results using LSR

However, in the three offline estimation methods proposed in previous chapters, we consider that time delay can take several discrete values with probability $P(\lambda_k = i)$, where k represents kth sampling instant and i is the value of delay. Therefore, for these three methods, we consider an expected prediction as shown below,

$$\hat{y}_{T_k} = \frac{b}{1 - az^{-1}} \sum_{i=1}^{d} P\left(\lambda_k = i\right) u_{T_k - i}. \tag{5.4}$$

5.3.1 Modeling using LSR

The least squares regression (LSR) is a well known modeling method. With the information vector x and output y, the parameter θ of the model $y = \theta x$, is calculated by $\theta = \left(x^T x\right)^{-1} x^T y$. In this method, time delay is considered to be a constant value. It is calculated by the Matlab function *delayest* and the value is 34 samples, which equals 34 minutes.

The validation result in Figure 5.4 shows the predicted output based on the estimated model can capture the overall trend of the measured output. The scatter plot of predicted values versus measurement (Figure 5.5) indicates that some drifting values cannot be predicted.

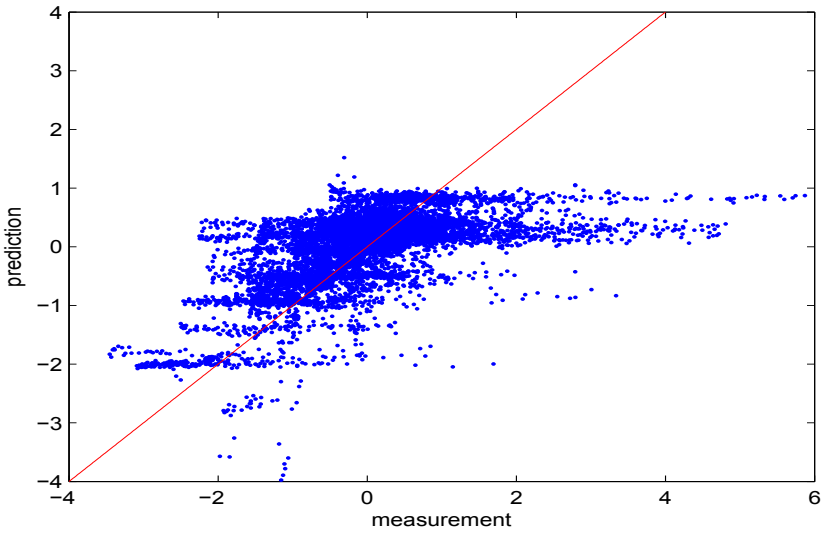

Figure 5.5: Scatter plot of predicted values v.s. measurement using LSR

5.3.2 Modeling using EM

Expectation maximization (EM) algorithm is an iterative method for finding parameter estimates in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood (Q function), and a maximization (M) step, which computes parameters maximizing the Q function. These parameters estimates are then used to determine the distribution of the latent variables in the next E step. With respect to the unknown time delay problem, the corresponding algorithm is developed in chapter 2.

Since an initial estimation of the time delay is around 34 minutes, we consider four discrete values {10,20,30,40} as the possible values for the delay. The data is processed to better fit the selected model, where both input and output are summed up every 10 minutes and then the output is downsampled every 5 samples. Figure 5.6 shows the data after processing.

Applying the algorithm of Chapter 2, the model parameters converge within 5 EM iterations (Figure 5.7) and we obtain the model parameters $a = 0.1596$, $b = 0.5837$. The time delay sequence is estimated in Figure 5.8 while the estimated transition probability

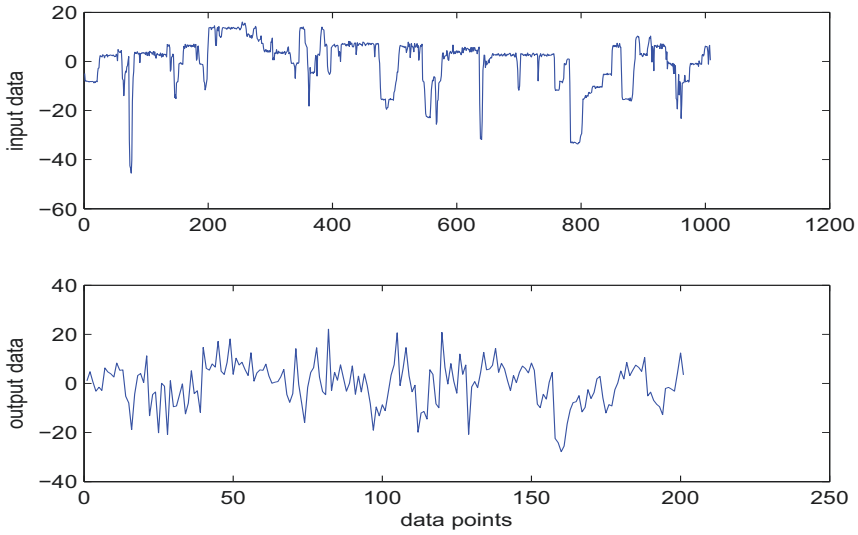

Figure 5.6: Process data (every 10 minutes)

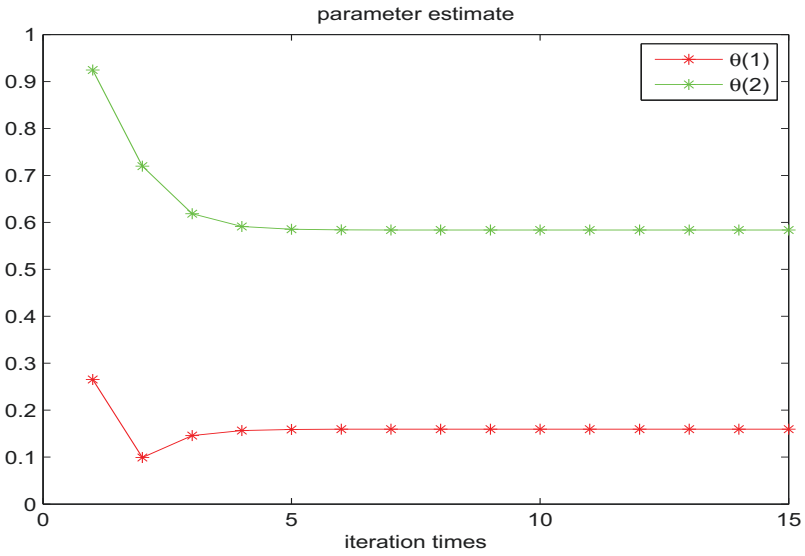

Figure 5.7: Parameter estimation using EM

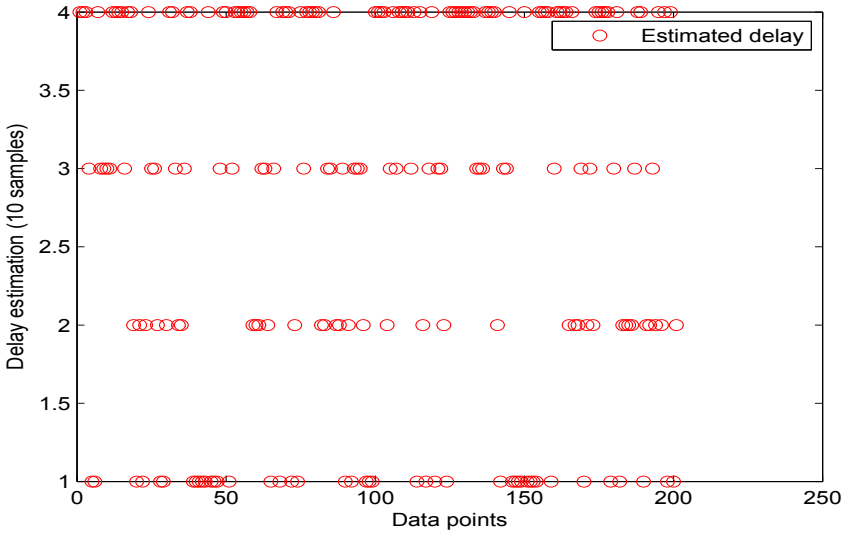

Figure 5.8: Delay estimation using EM

matrix of the HMM is

$$\hat{A} = \{\hat{\alpha}_{ij}\} = \begin{bmatrix} 0.2096 & 0.1980 & 0.2145 & 0.3779 \\ 0.1822 & 0.1698 & 0.2258 & 0.4222 \\ 0.1629 & 0.1677 & 0.2312 & 0.4383 \\ 0.1274 & 0.1256 & 0.1773 & 0.5696 \end{bmatrix} \tag{5.5}$$

The validation result in Figure 5.9 shows great match between prediction and target values. The scatter plot (Figure 5.10) agrees with this result.

5.3.3 Robust modeling using EM

Robust estimators are useful when observations contain large values or are sampled from a heavy-tailed distribution. Student's t-distributions with small degrees of freedom have heavy tails. Therefore, maximum likelihood estimation using these distributions provides simultaneous robust estimates of location and scale. In addition, the likelihood values can be used to choose among the available t-distributions, avoiding subjective choice of an estimator. The robust estimation of unknown time delay problem using EM algorithm has been developed in chapter 3.

Applying the algorithm in chapter 3, the model parameters converge within 10 EM iterations (Figure 5.11) and we obtain $a = 0.1633$, $b = 0.6099$. The time delay sequence is

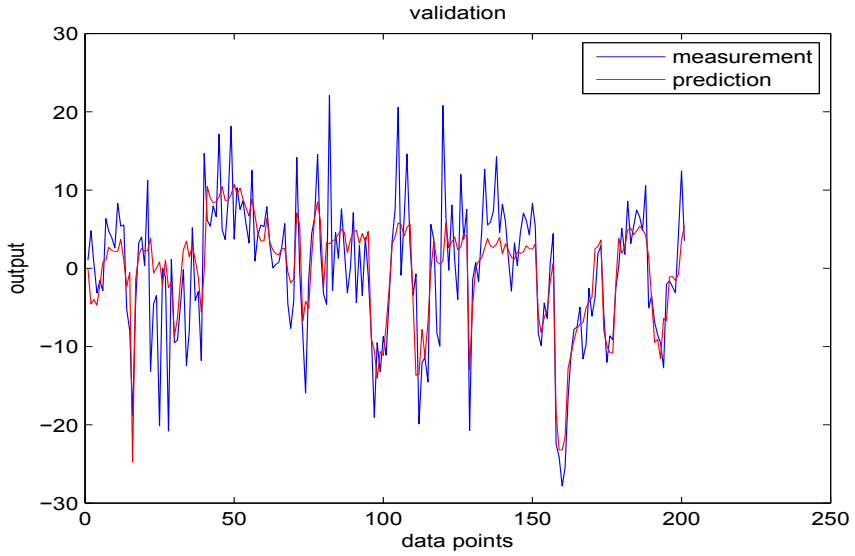

Figure 5.9: Validation results using EM

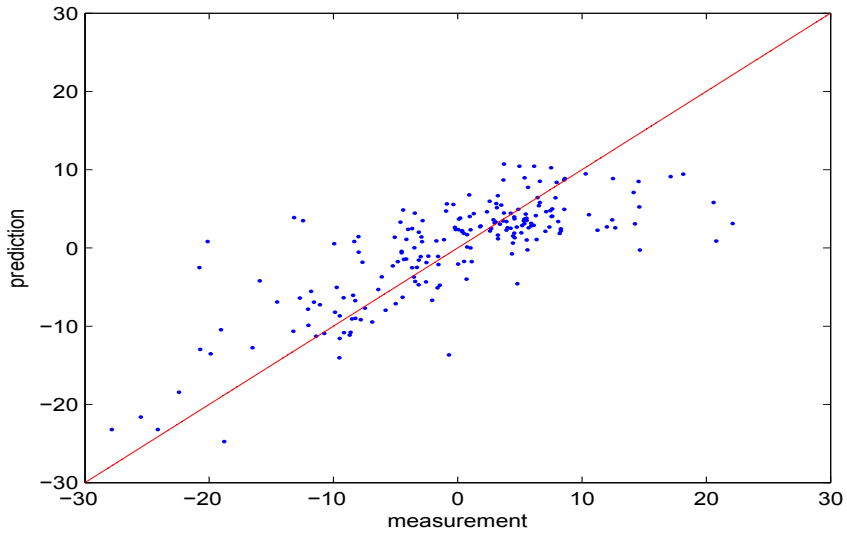

Figure 5.10: Scatter plot of predicted values v.s. measurement using EM

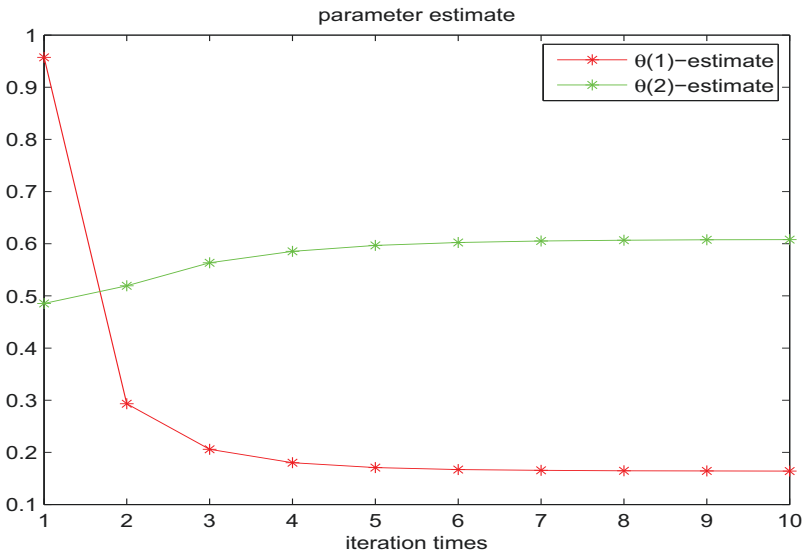

Figure 5.11: Parameter estimation using robust EM

estimated in Figure 5.12. The transition probability estimate is

$$\hat{A} = \{\hat{\alpha}_{ij}\} = \begin{bmatrix} 0.2963 & 0.2362 & 0.1227 & 0.3448 \\ 0.1924 & 0.1200 & 0.1515 & 0.5361 \\ 0.1208 & 0.1060 & 0.1615 & 0.6117 \\ 0.0861 & 0.0735 & 0.1084 & 0.7320 \end{bmatrix} \tag{5.6}$$

The validation result in Figure 5.13 shows a good match between prediction and target value. The scatter plot (Figure 5.14) agrees with the results given in the figures.

5.3.4 Robust modeling using VB

Variational Bayesian methods are for approximating intractable posterior distributions. They are typically used in complex statistical models consisting of observed variables as well as unknown parameters and latent variables. As is typical in Bayesian inference, the parameters and latent variables are grouped together as unobserved variables, which is a slightly different from EM. Another difference from EM is that VB considers the uncertainty of parameter estimations by estimating the distribution of parameters, while EM can only have point estimation. The robust estimation of unknown time delay problem using VB approach has been developed in Chapter 4.

Applying the algorithm in Chapter 4, the model parameters converge within 20 iterations (Figure 5.15) and we obtain $a = 0.1767$, $b = 0.5871$. The time delay sequence is estimated

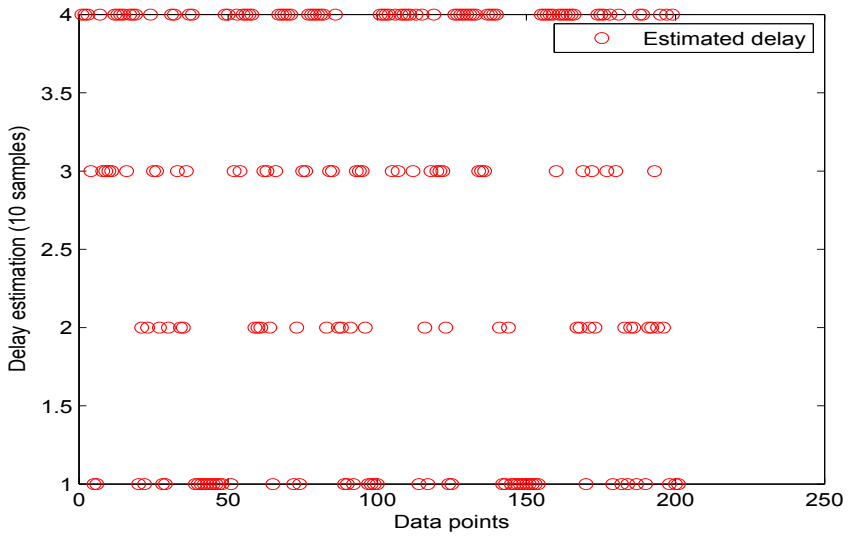

Figure 5.12: Delay estimation using robust EM

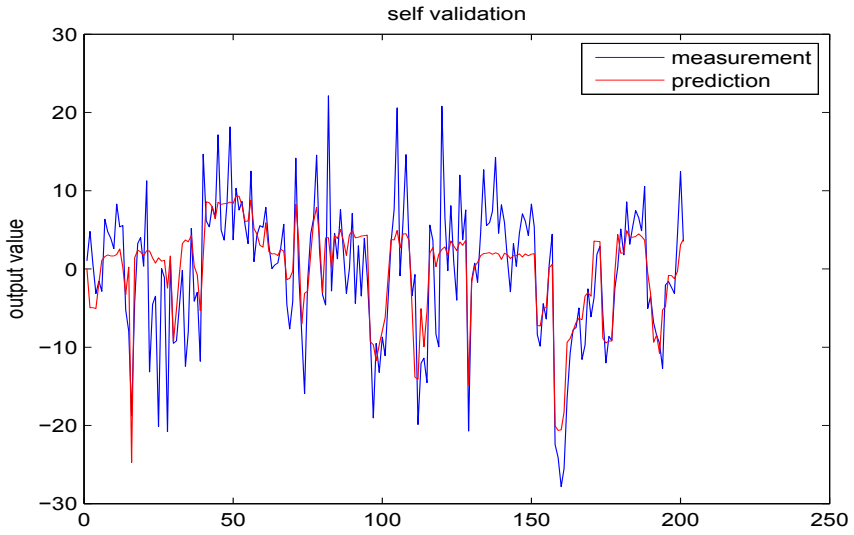

Figure 5.13: Validation results using robust EM

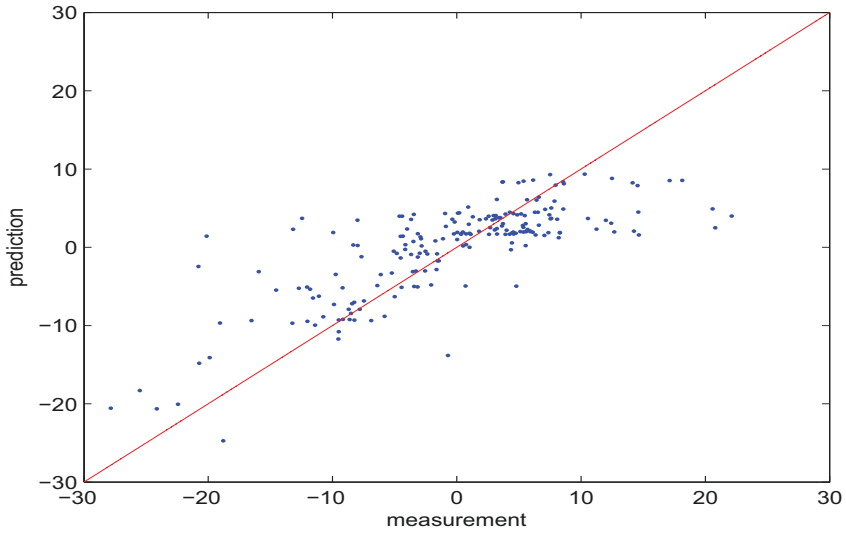

Figure 5.14: Scatter plot of predicted values v.s. measurement using robust EM

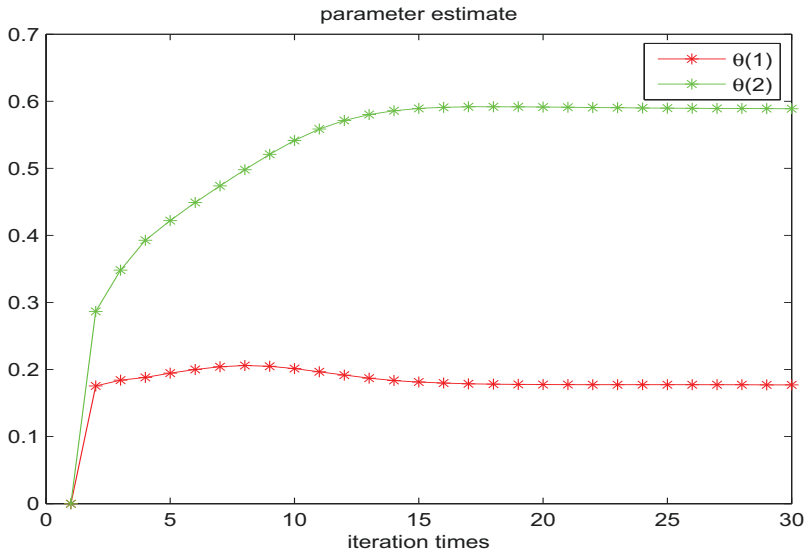

Figure 5.15: Parameter estimation using robust VB

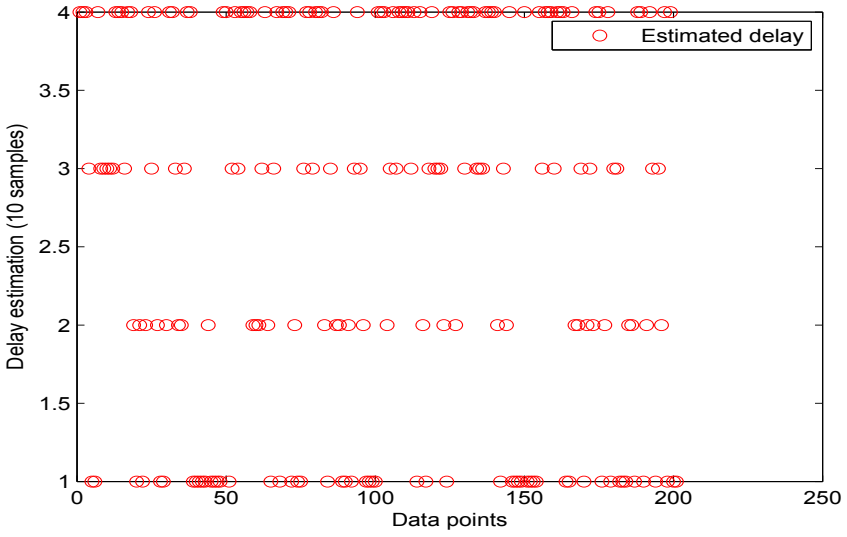

Figure 5.16: Delay estimation using robust VB

and is shown in Figure 5.16. The estimation result for the HMM transition matrix is

$$\hat{A} = \{\hat{\alpha}_{ij}\} = \begin{bmatrix} 0.1545 & 0.2122 & 0.1738 & 0.4595 \\ 0.1171 & 0.1580 & 0.1800 & 0.5450 \\ 0.0864 & 0.1410 & 0.1860 & 0.5866 \\ 0.0604 & 0.0929 & 0.1195 & 0.7272 \end{bmatrix} \tag{5.7}$$

The validation result in Figure 5.17 shows a good match between prediction and target value. The scatter plot (Figure 5.18) agrees with the above figure and conclusion.

5.3.5 Discussion

Table 5.2 presents a summary of the performance of the four different methods applied above. It is clear that the Markov chain based method achieves smaller RMSE than the constant delay based method. Among the three proposed methods, the robust estimation has smaller RMSE than the regular estimation due to the existence of outliers. The VB approach has smaller RMSE than the EM algorithm based method. This is because EM algorithm only uses the point estimation of the model parameter to obtain time delay distribution, while VB approach uses the full parameter distribution to obtain improved time delay distribution. The time delay distribution is adopted to have an expected prediction of the output, so the VB approach achieves better performance than the EM algorithm.

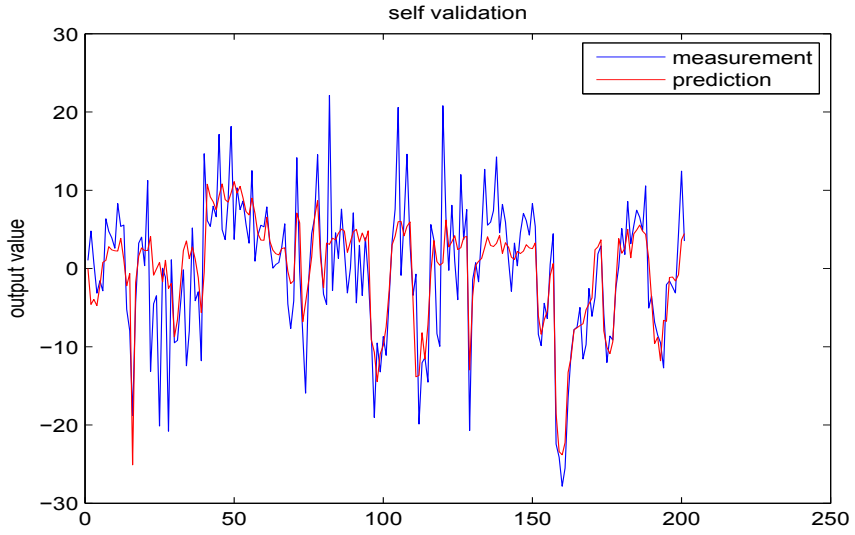

Figure 5.17: Validation results using robust VB

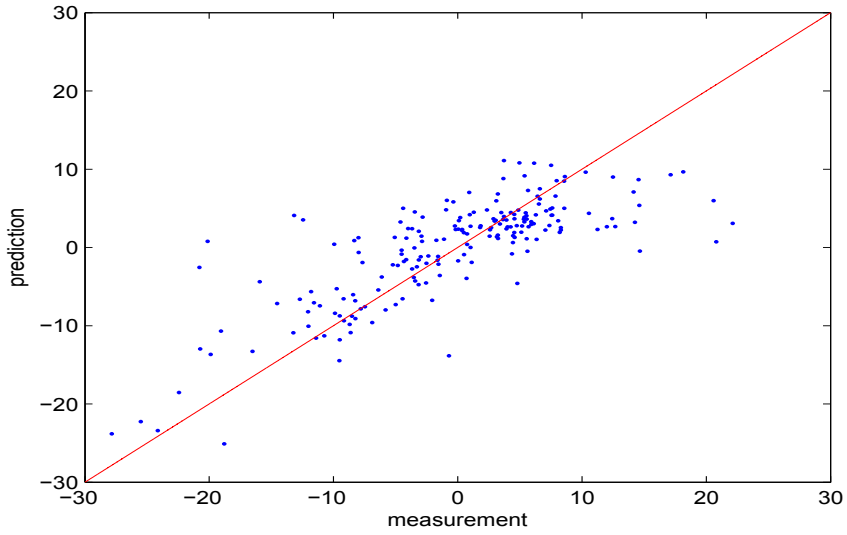

Figure 5.18: Scatter plot of predicted values v.s. measurement using robust VB

Table 5.2: A Summary on the Performance of Oil Sands Extraction Process Modeling

	Delay or delay range (minutes)	RMSE
Least Squares Regression	$\lambda = 34$	2.7985
Chapter 2: regular EM	$\lambda_t \in \{10, 20, 30, 40\}$	1.8143
Chapter 3: robust EM	$\lambda_t \in \{10, 20, 30, 40\}$	1.4659
Chapter 4: robust VB	$\lambda_t \in \{10, 20, 30, 40\}$	1.2692

5.4 Conclusion

This chapter applies the proposed methods that were developed in previous chapters to an industrial process. Considering time varying time delay and t-distributed measurement noise, the proposed methods can better fit the real industrial data. It is validated in this chapter that all the proposed methods can be useful to solve this industrial modeling problem.

Chapter 6

Conclusions

6.1 Summary of thesis

In this thesis, we have focused on system identification of dual rate processes with varying time delay. Process modeling based on proposed algorithms was tested on a product estimation for oil sands extraction process.

Chapter 1 presented the background and motivation of the system identification of dual rate industrial processes with varying time delay.

Chapter 2 proposed a batch mode EM algorithm for the parameter invariant process and a recursive version of the EM algorithm for the parameter variant process. The process was modeled as ARX and time delay was modeled as HMM. In both situations, ARX parameters, HMM parameters and the value of varying delays at every sampling time were estimated using the developed algorithms.

Chapter 3 proposed a robust approach to identify the varying delay process subject to outliers using t-distribution. The basic idea of using t distributions is to have the outliers weighted automatically during the iterative optimization process. Meanwhile, ARX model for the process and HMM model for the delay was adopted.

Chapter 4 proposed a variational Bayesian approach for the identification of the dual rate process with HMM time delay variation. Practical issues such as robustness, estimation of model parameter uncertainty and switching mechanism of time delays were addressed in this chapter.

The advantage of using hidden Markov model for time delay in process modeling was shown in simulations and experimental studies.

Chapter 5 designed models for the real-time prediction of production for oil sands extraction process based on the algorithms developed in the previous chapters.

6.2 Directions for future work

Based on the work presented in the previous chapters, it can be extended in the following aspects:

The determination of model order. Through Chapters 2-4, we considered the order of process model is known. This restricts the flexibility and accuracy of modeling because the identified model might not be suitable for the real process. A model order determination method should be developed which can find the optimal value for the process order.

The determination of delay range. The range of delay of the dimension of HMM is also considered to be known, however, in real industrial processes, the varying delay has unknown range. A faulty assumption of the delay range will lead to inaccurate delay estimation and result in failure of modeling. Thus a method to estimate the delay range should be developed.

Uncertainty of HMM parameters. In Chapter 4, the uncertainty of the ARX model parameters was considered while the HMM parameters were estimated as constant values. To obtain more accurate estimation results, the uncertainty of HMM parameters should also be considered.

Bibliography

[1] W. Ma and J. Huang. Accurate time delay estimation based on SINC filtering. In *2002 6th International Conference on Signal Processing*, volume 2, pages 1621–1624, 2002.

[2] J. Benesty, J. Chen, and Y. Huang. Time-delay estimation via linear interpolation and cross correlation. *IEEE Transactions on Speech and Audio Processing*, 12(5):509–519, 2004.

[3] Z. H. Michalopoulou and M. Picarelli. A gibbs sampling approach to maximum a posteriori time delay and amplitude estimation. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 3001–3004, 2002.

[4] K. R. Rajeswari and D. E. Rani. Time-delay estimation using MLE approach for wide-band radar systems. In *1998 Fourth International Conference on Signal Processing Proceedings*, volume 2, pages 1493–1496, 1998.

[5] P. H. Moghaddam and H. Amindavar. A new algorithm for multipath time delay estimation in low SNR using MLE method. In *1998 International Symposium on Underwater Technology*, pages 35–38, 1998.

[6] H. Yang L. Xie and B. Huang. FIR model identification of multirate processes with random delays using EM algorithm. *AIChE Journal*, 59(11):4124–4132, 2013.

[7] D. C. Huynh, M. W. Dunnigan, and S. J. Finney. On-line parameter estimation of an induction machine using a recursive least-squares algorithm with multiple time-varying forgetting factors. In *2010 IEEE International Conference on Power and Energy*, pages 444–449, 2010.

[8] G. Belforte, Y. T. Teo, and T. T. Tay. Recursive parameter estimation for time-varying systems in presence of unknown but bounded measurement noise. In *1992 Singapore International Conference on Intelligent Control and Instrumentation*, volume 2, pages 945–950, 1992.

[9] A. Taha and A. S. Hadi. A general approach for automating outliers identification in categorical data. In *2013 ACS International Conference on Computer Systems and Applications*, pages 1–8, 2013.

[10] R. Baragona and F. Battaglia. Outliers detection in multivariate time series by independent component analysis. *Neural Computation*, 19(7):1962–1984, 2007.

[11] A. Aravkin, M. Styer, Z. Moratto, A. Nefian, and M. Broxton. Student's t robust bundle adjustment algorithm. In *2012 19th IEEE International Conference on Image Processing*, pages 1757–1760, 2012.

[12] Z. Zohny and J. Chambers. Modelling interaural level and phase cues with student's t-distribution for robust clustering in MESSL. In *2014 19th International Conference on Digital Signal Processing*, pages 59–62, 2014.

[13] Y. Lu and B. Huang. Robust multiple-model LPV approach to nonlinear process identification using mixture t distributions. *Journal of Process Control*, 24(9):1472 – 1488, 2014.

[14] X. Jin and B. Huang. Identification of switched Markov autoregressive exogenous systems with hidden switching state. *Automatica*, 48(2):436–441, 2012.

[15] G. C. Chow. Maximum-likelihood estimation of misspecified models. *Economic Modelling*, 1(2):134–138, 1984.

[16] J. Gauvain and C. H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

[17] N. M. Laird A. P. Dempster and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Joutnal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[18] J. P. Richard. Time-delay systems: an overview of some recent advances and open problems. *Automatica*, 39(10):1667–1694, 2003.

[19] T. Zhang and Y. Li. A control scheme for bilateral teleoperation systems based on time-varying communication delay identification. In *1st International Symposium on Systems and Control in Aerospace and Astronautics*, pages 6–278, 2006.

[20] S. Xu, J. Lam, and X. Mao. Delay-dependent H infinite control and filtering for uncertain Markovian jump systems with time-varying delays. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 54(9):2070–2077, 2007.

[21] C. J. Kim. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(12):1–22, 1994.

[22] Y. Bar-Shalom and X. Li. Estimation and tracking- principles, techniques, and software. *IEEE Antennas and Propagation Magazine*, 38:62, 1993.

[23] S. C Rutan. Recursive parameter estimation. *Journal of chemometrics*, 4(2):103–121, 1990.

[24] F. A. G. Dumortier. Theory and practice of recursive identification: Lennart ljung and torsten sderstrm. *Automatica*, 21(4):499–501, 1985.

[25] I. H. Grant. Recursive least squares. *Teaching Statistics*, 9(1):15–18, 1987.

[26] L. Lang, B. R. Bakshi, and P. K. Goe. Parameter estimation with a moving window particle filtering for nonlinear dynamic models. *The 2007 AIChE Annual Meeting*, 2007.

[27] D. M. Titterington. Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):257–267, 1984.

[28] P. J. Chung, W. J. J. Roberts, and J. F. Bhme. Recursive K-distribution parameter estimation. *IEEE Transactions on Signal Processing*, 53(2):397–402, 2005.

[29] O. Capp and E. Moulines. On-line expectation maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

[30] E. Ozkan, C. Fritsche, and F. Gustafsson. Online EM algorithm for joint state and mixture measurement noise estimation. In *2012 15th International Conference on Information Fusion*, pages 1935–1940, 2012.

[31] J. Shalchian, A. Khaki-Sedigh, and A. Fatehi. A subspace based method for time delay estimation. In *2010 4th International Symposium on Communications, Control and Signal Processing*, pages 1–4, 2010.

[32] L. Chen, L. Han, B. Huang, and F. Liu. Parameter estimation for a dual-rate system with time delay. *ISA Transactions*, 53(5):1368–1376, 2014.

[33] X. Gao, F. Xie, and H. Hu. Enhancing the security of electro-optic delayed chaotic system with intermittent time-delay modulation and digital chaos. *Optics Communications*, 352:77–83, 2015.

[34] Q. Lin, R. Loxton, C. Xu, and K. L. Teo. Parameter estimation for nonlinear time-delay systems with noisy output measurements. *Automatica*, 60:48–56, 2015.

[35] M. M. Moser, C. H. Onder, and L. Guzzella. Recursive parameter estimation of exhaust gas oxygen sensors with input-dependent time delay and linear parameters. *Control Engineering Practice*, 41:149–163, 2015.

[36] Y. Jia, L. Sun, and H. Teng. A comparison study of hidden Markov model and particle filtering method: Application to fault diagnosis for gearbox. In *2012 IEEE Conference on Prognostics and System Health Management*, pages 1–7, 2012.

[37] M. Johansson and T. Olofsson. Bayesian model selection for Markov, hidden Markov, and multinomial models. *IEEE Signal Processing Letters*, 14(2):129–132, 2007.

[38] G. Kotsalis, A. Megretski, and M. A. Dahleh. A model reduction algorithm for hidden Markov models. In *2006 45th IEEE Conference on Decision and Control*, pages 3424–3429, 2006.

[39] E. Dorj, C. Chen, and M. Pecht. A Bayesian hidden Markov model-based approach for anomaly detection in electronic systems. In *2013 IEEE Aerospace Conference*, pages 1–10, 2013.

[40] D.V. Lindberg and H. Omre. Blind categorical deconvolution in two-level hidden Markov models. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):7435–7447, 2014.

[41] X. Jin, B. Huang, and D. S. Shook. Multiple model LPV approach to nonlinear process identification with EM algorithm. *Journal of Process Control*, 21(1):182–193, 2011.

[42] M. Svensn and C. M. Bishop. Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252, 2005. Trends in Neurocomputing: 12th European Symposium on Artificial Neural Networks.

[43] S. Shoham, M. R. Fellows, and R. A. Normann. Robust, automatic spike sorting using mixtures of multivariate t-distributions. *Journal of Neuroscience Methods*, 127(2):111–122, 2003.

[44] J. Christmas and R. Everson. Robust Autoregression: Student-t innovations using variational Bayes. *IEEE Transactions on Signal Processing*, 59(1):48–57, 2011.

[45] K. L. Lange, R. J. A. Little, and J. M. G. Taylor. Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

[46] M. E. Tipping and N. D. Lawrence. Variational inference for student-t models: Robust Bayesian interpolation and generalised component analysis. *Neurocomputing*, 69(13):123–141, 2005.

[47] G. Deng. EM algorithms for robust signal filtering and prediction. In *2004 12th European Signal Processing Conference*, pages 625–628, 2004.

[48] Q. M. J. Wu H. Zhang and T. M. Nguyen. Image segmentation by a new weighted student's t-mixture model. *IET Image Processing*, 7(3):240–251, 2013.

[49] H. Zhang, Q. M. J. Wu, and T. M. Nguyen. A robust fuzzy algorithm based on student's t-distribution and mean template for image segmentation application. *IEEE Signal Processing Letters*, 20(2):117–120, 2013.

[50] J. Zhang and Y. Leung. Robust clustering by pruning outliers. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 33(6):983–998, Dec 2003.

[51] I. Tjoa and L. Biegler. Simultaneous strategies for data reconciliation and gross error detection of nonlinear systems. *Computers and Chemical Engineering*, 15(10):679–690, 1991.

[52] X. Jin and B. Huang. Robust identification of piecewise/switching autoregressive exogenous process. *AIChE Journal*, 56(7):1829–1844, 2010.

[53] D. Boudreau and P. Kabal. Joint time-delay estimation and adaptive recursive least squares filtering. *IEEE Transactions on Signal Processing*, 41(2):592–601, 1993.

[54] D. Boudreau and P. Kabal. Joint gradient-based time delay estimation and adaptive filtering. In *IEEE International Symposium on Circuits and Systems*, pages 3165–3169, 1990.

[55] H. Zhang, F. Yang, X. Liu, and Q. Zhang. Stability analysis for neural networks with time-varying delay based on quadratic convex combination. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4):513–521, 2013.

[56] R. Sirisongkol and X. Liu. Stability analysis of recurrent neural networks with time-varying delay and disturbances via quadratic convex technique. In *2014 Fifth International Conference on Intelligent Control and Information Processing*, pages 130–137, 2014.

[57] N. Nasios and A. G. Bors. Variational learning for Gaussian mixture models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(4):849–862, 2006.

[58] G. Zhang and N. Kingsbury. Variational Bayesian image restoration with group-sparse modeling of wavelet coefficients. *Digital Signal Processing*, 2015.

[59] Z. Ma, P. K. Rana, J. Taghia, M. Flierl, and A. Leijon. Bayesian estimation of Dirichlet mixture model with variational inference. *Pattern Recognition*, 47(9):3143–3157, 2014.

[60] J. Ala-Luhtala, S. Srkk, and R. Pich. Gaussian filtering and variational approximations for bayesian smoothing in continuous-discrete stochastic dynamic systems. *Signal Processing*, 111:124–136, 2015.

[61] T. Baldacchino, E. J. Cross, K. Worden, and J. Rowson. Variational Bayesian mixture of experts models and sensitivity analysis for nonlinear dynamical systems. *Mechanical Systems and Signal Processing*, 2015.

[62] P. Ruiz, X. Zhou, J. Mateos, R. Molina, and A. K. Katsaggelos. Variational Bayesian blind image deconvolution: A review. *Digital Signal Processing*, 2015.

[63] Y. Li and G. Zhang. Blind seismic deconvolution using variational Bayesian method. *Journal of Applied Geophysics*, 110:82–89, 2014.