Statistical Approaches to Robust Identification of Multi-modal Processes

by

Yaojie Lu

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 exhibit certain forms of time-varying behaviour as well as nonlinearity. Suitable production policies might also drive a plant to switch among various operating modes leading to the multiple-mode behaviour. Moreover, industrial data are often noisy and contaminated with outliers, making the process identification challenging. Thus, reliable identification of the multi-modal process plays a key role in efficient process control and operations.

In this thesis, time-varying behaviour, nonlinearity and switching dynamics are generally treated as multi-modal behaviour. Two multi-model modelling techniques, *i.e.*, the linear parameter varying (LPV) technique and the switched modelling technique, are investigated to model the multi-modal processes. The robustness of proposed algorithms is enhanced by modelling the noise as t distributions.

The identification of multi-modal processes is formulated under statistical frameworks. The expectation-maximization (EM) algorithm is a powerful statistical approach to parameter estimation. A novel identification algorithm is proposed to minimize the adverse influence of outliers by integrating t distributions with the EM algorithm. During the iterative estimation procedure, outlying observations are downweighted by a latent variable of the t distribution automatically, so their adverse influence on identification is minimized. Furthermore, a full-Bayesian estimation method named as variational Bayesian algorithm is investigated to identify multi-modal processes. A novel iterative optimization algorithm is proposed to infer the number of operating modes from the training data. Hence the model structure (the number of modes) and model parameters are obtained simultaneously.

The proposed algorithms are verified by simulations and experiments. Finally, soft sensors based on proposed algorithms are designed to effectively estimate the steam-quality for the once-through steam generators used in the oil sands industry.

Preface

Chapter 2 of this thesis has been published as Y. Lu and B. Huang, *Robust multiple-model LPV approach to nonlinear process identification using mixture t distributions*, Journal of Process Control (in press, 2014). I was responsible for the algorithm development and manuscript composition. Dr. Biao Huang was the supervisory author and was involved with manuscript composition.

Acknowledgements

Time flew during the days of deriving equations, writing and revising my work, and seeking for potential applications. Through this journey, I not only gained solid experience in scientific research and industrial practice, but also improved interpersonal skills and communication skills.

It gives me great pleasure in acknowledging the support and help from my supervisor Dr. Biao Huang who gave me the opportunity to explore my way in the excellent Computer Process Control (CPC) group and has guided me through his patience and encouraging suggestions. This thesis could be low-quality without guidance and comments of my supervisor. I am really grateful to his patience, encouragement, and guidance during my graduate study.

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 Swanand Khare and Tianhong Pan for their suggestions and discussions about feature extraction and cluster analysis in the Bio-informatics project. I would like to thank Shima Khatibisepehr for her help and suggestions when I felt confused during my research. Help from Fei Qi (APC engineer at Suncor Energy Inc.) and Fangwei Xu (PC&A Engineer at Syncrude Canada Ltd.) during the industrial soft sensor projects 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: Tianbo Liu, Xianqiang Yang, Zhankun Xi, Aditya Tulsyan, Lei Chen, Kangkang Zhang, Ming Ma, Da Zheng, Li Xie, Nima Sammaknejad, Alireza Fatehi, Mohammad Rashed, Jiusun Zeng, Weili Xiong, Elham Naghoosi, Abhinandhan Raghu, Anahita Sadeghian, Fadi Ibrahim and many others from CPC group. I would also thank my friends for their accompanying and encouragement: Shuning Li, Mulang Chen, Ruben Gonzalez, Xiongtan Yang, Lei Sun, Chen Li, Xing Jin, Ouguan Xu, Shunyi Zhao, Xiaodong Xu, Yuan Yuan, Liu Liu, Ran Li,

Jing Zhang, Ouyang Wu, Ruomu Tan, Yuyu Yao, Cong Jin, Wei Han, Tong Chen, Yang Zhou, Bin Xia, Yang Wang and many else.

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 and my uncle Yao.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Advanced process control (APC) strategies have been developing rapidly during recent decades to meet the increasing requirements of complex industrial operations under various production conditions. Most of the APC strategies are model-based, leading to the prerequisite of an accurate and compact mathematical description of the process. Hence, process modelling and identification play an important role in control and monitoring of industrial processes.

1.1 Motivation

Due to the wide operating range of industrial processes, time-varying and nonlinear behaviour commonly exist in process operations. Traditional modelling methods based on the assumption that the process is simply operated within a fixed operating region are not capable of describing the process dynamics over the entire operating range. The model of an industrial process changes with the shift of the operating modes, resulting in the multi-modal behaviour of the process.

Under some circumstance, the operating mode of the process is dependent on the known variables which are referred to as scheduling variables, and the linear parameter varying (LPV) technique [2, 3, 4] is effective in modelling and control of these time-varying/nonlinear processes. Nevertheless, the scheduling variable is not always available due to insufficient knowledge of the process. An an alternative, the switching modelling technique [5, 6, 7, 8] has no requirement for the scheduling variables. The operating modes of the process are estimated by data-driven methods instead of being determined by the scheduling variable. Both modelling methods are

able to identify multi-modal processes. The LPV technique is preferred when there is sufficient process knowledge to investigate the scheduling variable. Otherwise, the switched modelling technique is more suitable.

Industrial data are often noisy and contaminated with outliers, which may be corrupted observations or genuine samples from a heavy-tailed distribution [9]. Transmission errors, process disturbances, and instrument degradation are common reasons that can cause outliers in recorded data. Data-driven modelling methods are usually sensitive to outliers. Statistical analysis of process data contaminated with outliers may lead to biased parameter estimation and plant-model mismatch [10]. Therefore, the problem of process identification in the presence of outliers has received great attention, and various robust identification methods have been proposed to deal with outliers [11, 12, 13].

The modelling of the noise distribution is essential to parameter estimation under statistical frameworks. Conventional Gaussian distributions are sensitive to outliers. A t distribution has the capability of varying continuously from a very heavy-tailed distribution to a Gaussian distribution by adjusting its degrees of freedom. The effect of outliers on modelling can be diminished by assigning higher probability densities to the tails [14]. In addition, a t distribution can be represented by an infinite mixture of scaled Gaussian distributions [15], which is an important property in statistical modelling.

This thesis focuses on robust identification of multi-modal processes under statistical frameworks. Two approaches to identification of multi-modal processes with sufficient robustness are proposed, and soft sensors based on proposed algorithms are designed to estimate the steam-quality for once-through steam generators in oil sands industry.

1.2 Thesis contributions

The main contribution of this thesis is the development of multi-modal process identification methods with sufficient robustness to address the difficulty of modelling time-varying behaviour, nonlinearity, and switching dynamics of industrial processes. The proposed algorithms are resistant to outliers and result in improved accuracy

and reliability of process modelling and prediction. Specifically, the contributions of this thesis are summarized as follows:

1. Modelled the time-varying/nonlinear processes using the multiple-model LPV approach, and t distributions are used to describe noise characteristic with outliers.

2. Integrated t distributions with the expectation-maximization (EM) algorithm, and made the algorithm down-weight outlying observations automatically.

3. Developed a robust variational Bayesian approach to identification of switched models. The distributions of parameters were estimated by the full-Bayesian approach, and the uncertainty of parameters was taken into account.

4. Proposed a method to infer the number of operating modes of the process from the training data-set automatically.

5. Designed multi-model soft sensors to estimate steam-quality of once-through steam generators used in oil sands industry. Evaluated the performance of various methods of steam-quality measurements.

1.3 Thesis outline

This thesis is organized in the paper format. The literature review is distributed in each chapter.

The rest of this thesis is organised as follows:

Chapter 2 develops a robust multiple-model linear parameter varying (LPV) approach to identification of the nonlinear process contaminated with outliers. The identification problem is formulated and solved under the EM framework.

Chapter 3 deals with such practical issues as unknown number of local-models, noisy operational data, and unknown switching mechanism of the switched models. A full-Bayesian process identification approach is developed in this chapter.

Chapter 4 is the design of adaptive soft sensors of steam-quality measurement for once-through steam generators used in industry.

Chapter 5 concludes the thesis.

Chapter 2

Robust Multiple-model LPV Approach to Nonlinear Process Identification Using Mixture *t* Distributions*

In this chapter, we propose a robust multiple-model linear parameter varying (LPV) approach to identification of the nonlinear process contaminated with outliers. The identification problem is formulated and solved under the EM framework. Instead of assuming that the measurement noise comes from the Gaussian distribution like conventional LPV approaches, the proposed robust algorithm formulates the LPV solution using mixture t distributions and thus naturally addresses the robust identification problem. By modulating the distribution tails through degrees of freedom, the proposed algorithm can handle various outliers. Two simulated examples and an experiment are studied to verify the effectiveness of the proposed approach.

2.1 Introduction

Processes in industry are commonly operated in a wide operating range and tend to exhibit parameter varying nature as well as nonlinearity. The dynamic behaviour of the complex nonlinear process cannot be predicted well by traditional linear models. Diverse nonlinear modelling approaches have been developed by researchers. Among

*This chapter has been published in: Y. Lu and B. Huang, *Robust multiple-model LPV approach to nonlinear process identification using mixture t distributions*, Journal of Process Control (in press, 2014).

them, nonlinear ARX models [16] and Wiener-Hammerstein models [17] are widely applied, though the robustness and prediction performance need improvement. Other techniques such as neural networks [18], self-organizing maps [19], and least-squares support vector machine (LS-SVM) [20] have been applied to identification of nonlinear processes by taking practical issues into account. Though the robustness of nonlinear process identification methods have been improved using these elaborate approaches, the identified model tends to be complicated. In literature, the linear parameter varying (LPV) approach has often been used to address the nonlinear process identification problem. This approach utilizes a linear description for the nonlinear process that involves a suitable set of scheduling variables [21]. The model complexity of the nonlinear process can be reduced significantly using the LPV approach.

The existing LPV identification approaches are commonly formulated in discrete-time framework assuming a dependence on the scheduling variable [22]. Based on the model structure, they can be characterized into three types: input-output (IO) LPV [23], state-space (SS) LPV [24] and orthogonal basis functions (OBF) based LPV [25]. In this chapter we explore the LPV identification approach based on the IO model. Most IO LPV methods existing in the literature are parameter interpolation based [23, 26]. The limitation of the parameter interpolation is the assumption that the scheduling variable should vary continuously, which may not always be the case in practice. To address this issue, the model interpolation based LPV approach [4] has been developed, although a similar idea can be traced back to an earlier study [3]. The model interpolation LPV approach is a mixture modelling technique that resembles the data clustering method, both of which assign each collected data point to one of local models and then identify local models using assigned data points. Multiple-model methods based on cluster analysis [6, 8] have been successfully applied to identification of piece-wise affine (PWA) systems, in which the transition from one local model to another is abrupt. However, for many chemical processes, the dynamic behaviour tends to vary gradually during the transition period. Thus, models identified by methods based on cluster analysis may not capture the dynamics of transition. On contrast, the transition dynamics can be approximated by a combination of local models through a smooth validity function in model interpolation LPV approaches. These approaches conform to the operation of chemical processes in

practice; namely the process is operated around several desired operating points, and dynamic transitions from one operating point to another operating point is permitted.

Although the model interpolation based LPV identification approaches proposed in the literature are relevant to the operation of the process, to the best of authors' knowledge, none of these methods has considered the influence of outliers and robustness which may greatly deteriorate the reliability of the identified model. Conventional multiple-model approaches, for example, the algorithm illustrated in [27], make use of Gaussian mixture models to approximate the complex process where the prediction error of each local model is assumed as Gaussian distributed, a main limitation of which is the lack of robustness to outliers. 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 [15]. Thus, models identified by conventional approaches may be unreliable owing to the influence of outliers. Considering the common existence of outliers in industry data sets, possibly brought by the malfunction of sensors, signal interference, incorrect recording by technicians and other unknown disturbance, LPV modelling methods should take outliers into account.

Simple approaches, like trimming and winsorizing [28], are aimed at screening outliers. Though intuitive and simple, one of the common major drawbacks is the simple discarding of data. This practice can lead to biased estimation [14]. Several advanced approaches have been proposed to deal with outliers as alternatives to simple screening. Methods are commonly designed to reduce the influence of outliers in regression. Robust regression methods, such as the M-estimator introduced by Huber [29], can iteratively down-weight the outlying data. Another way to cope with potential outliers is to use the two-component Gaussian distribution, or the contaminated Gaussian distribution [30, 31]. The outliers are taken into account when modelling the noise, where a Gaussian component with large variance (or covariance matrix) is utilized to model the outliers. Jin et al. have used the contaminated Gaussian distribution to make their algorithm robust to outliers when identifying the piecewise/switching process [8]. Although they have demonstrated the advantage of their algorithm in dealing with outliers, the solution is limited to a special type of outliers with two components of Gaussian mixture distribution. A more general

approach to resist the influence of outliers is to use t distributions. The t distribution can have longer tails than a Gaussian distribution through adjustable degrees of freedom (ν). The effect of outliers on modelling can be diminished by assigning higher probability densities to outliers (i.e. tails) [14]. Several researchers have taken advantages of t distributions to make methods robust to outliers in their applications, such as cluster analysis [32], image processing [33] and latent variable modelling [14].

However, in nonlinear process identification, to the best of authors' knowledge, no such method has been explored. In this chapter, we will develop a novel approach which makes the identification of nonlinear processes robust to outliers using the mixture t distributions under the framework of the multiple-model LPV approach. Owing to the flexibility of t distributions, the proposed method not only resists the adverse effect of outliers but also acts as an indicator of the quality of the data. The degree of freedom of the t distribution is self-adaptive to the quality of the data set. A large degree of freedom indicates a good quality of the data set while a small one corresponds to a poor data quality. When the degree of freedom is taken as a parameter to be estimated from data, the developed algorithm is capable of adapting itself to the data with various quality. Two simulation examples and an experiment verification demonstrate that the proposed method is capable of adapting to the various data quality and can provide more reliable identification results.

We illustrate the proposed approach in detail in the remainder of this chapter. Section 2 revisits t distributions, illustrating the relation between Gaussian distributions and t distributions. Then the identification problem is formulated under the multiple-model LPV framework with mixture t distributions in Section 3. Considering the difficulty of estimating parameters by maximizing the likelihood function directly, a robust identification method based on the EM algorithm is developed in Section 4. Two simulation examples and an experiment performed on a *balls-in-tubes* apparatus are utilized to verify the proposed method in Section 5. Section 6 concludes this chapter.

2.2 Revisit of t distributions

Though Gaussian distributions have nice analytical property and often yield tractable algorithms for linear models, their sensitivity to outliers is widely known. As an alternative, t distributions provide a way to broaden the Gaussian distribution for potential outliers. The probability density function of a univariate t distribution with mean μ, variance related variable σ^2, and degrees of freedom ν is [32]

$$t(y_k|\mu,\sigma^2,\nu) = \frac{\Gamma(\frac{\nu+d}{2})|\sigma^2|^{-1/2}}{(\pi\nu)^{d/2}\Gamma(\frac{\nu}{2})\{1+\delta(y_k|\mu,\sigma^2)/\nu\}^{\frac{\nu+d}{2}}}, \tag{2.1}$$

where d is the dimension of the observation y_k ($d = 1$ here), and $\Gamma(t)$ is the Gamma function with the expression $\Gamma(t) = \int_0^\infty z^{t-1}e^{-z}dz$. In addition, $\delta(y_k|\mu,\sigma^2)$ represents the squared one-dimensional Mahalanobis distance between y_k and μ:

$$\delta(y_k|\mu,\sigma^2) = (y_k-\mu)^2/\sigma^2. \tag{2.2}$$

As displayed in Fig. 2.1, t distributions can have longer tails than a Gaussian distribution owing to adjustable degrees of freedom ν. The effect of outliers on modelling can be diminished by assigning higher probability densities to outliers [14].

Moreover, the t distribution can be represented by an infinite mixture of scaled Gaussian distributions [15], so it reserves good analytical property of the Gaussian distribution. Consider a linear model given by Eq. (2.3),

$$y_k = x_k^T\theta + e_k,\ k = 1, 2, \cdots, N \tag{2.3}$$

with fully observed predictor variables; namely the regressor x_k is fully observed. θ is the regression coefficient, and e_k is the residual distributed as a t distribution with mean 0, variance related variable σ^2, and degrees of freedom ν, i.e.,

$$e_k \sim t(0,\sigma^2,\nu), \tag{2.4}$$

from which we have $y_k|(x_k,\theta,\sigma^2,\nu) \sim t(x_k^T\theta,\sigma^2,\nu)$. Essentially, the t distribution can be decomposed into scaled Gaussian distributions where the variance scale r is a Gamma distributed latent variable which depends on degrees of freedom ν [15], i.e.,

$$f(y_k|x_k,\theta,\sigma^2,\nu) = \int f(y_k|x_k,\theta,\sigma^2,\nu,r)f(r|x_k,\theta,\sigma^2,\nu)dr, \tag{2.5}$$

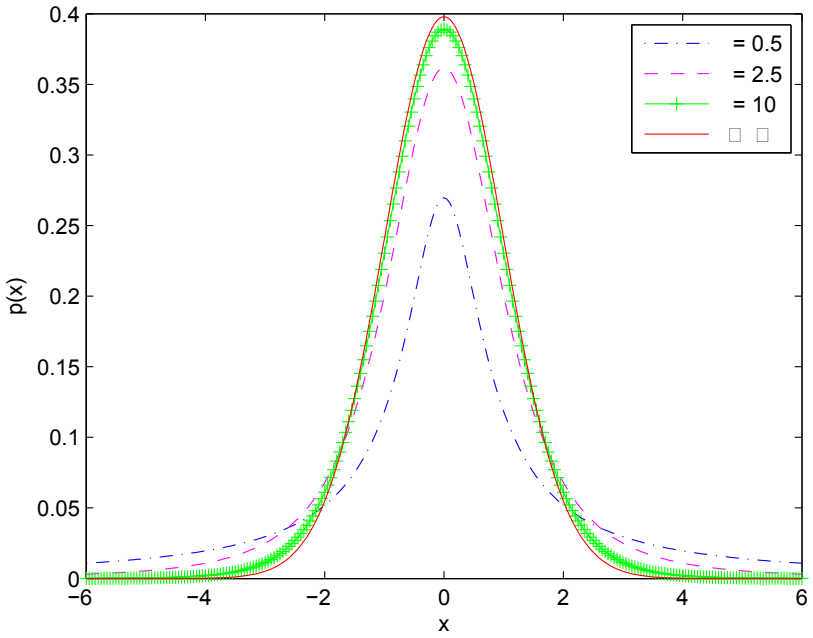

Figure 2.1: Tails of t distributions with fixed mean for various degree of freedom ν, where $\nu \to \infty$ corresponds to a Gaussian distribution.

where

$$y_k|(x_k, \theta, \sigma^2, \nu, r) = y_k|(x_k, \theta, \sigma^2, r) \sim \mathcal{N}(x_k^T\theta, \sigma^2/r), \tag{2.6}$$

$$r|(x_k, \theta, \sigma^2, \nu) = r|\nu \sim \mathcal{G}(\frac{1}{2}\nu, \frac{1}{2}\nu), \tag{2.7}$$

and the Gamma density function has the form:

$$\mathcal{G}(r|\alpha, \beta) = \frac{\beta^\alpha r^{\alpha-1} e^{-\beta r}}{\Gamma(\alpha)}, \ r > 0, \alpha > 0, \beta > 0. \tag{2.8}$$

According to the property of Gamma distributions, it is easy to understand that $r \to 1$ with probability 1 as $\nu \to \infty$, and y_k becomes marginally Gaussian distributed [34], i.e, $\mathcal{N}(x_k^T\theta, \sigma^2)$. Therefore, the family of t distributions not only reserve good analytical property of Gaussian distributions, but also provide a heavy-tailed alternative to the Gaussian family [32].

2.3 Problem formulation using mixture t distributions

Often chemical processes transit among several operating points during a normal operation. Owing to the nonlinearity of the process, local models around different

operating points are different from each other. Therefore, a single linear model usually fails to represent the whole process. Besides, the nonlinear model structure of the process is usually unknown, so representing the process by a nonlinear model structure directly is not feasible in most cases. To conquer this difficulty, the multiple-model LPV approach has been proposed in literature. In this approach, each local model around the operating point is approximated by a linear model. Among many linear model structures, the autoregressive eXogenous (ARX) model structure is often used [4, 27], while the Box-Jenkins (BJ) model structure has also been considered [22]. Since a high-order ARX model is capable of approximating any linear dynamic process [35], and the parameter estimation of the ARX model has a closed form solution under the EM framework, the ARX model structure, as expressed by Eq. (2.9), will be adopted as the local model throughout this chapter.

$$y_k = x_k^T \theta_{I_k} + e_k,\ k = 1, 2, \cdots, N \tag{2.9}$$

where $y_k \in R$ is the measured output, and $x_k \in R^n$ denotes the regressor of the process at k-th sampling instant. The regressor x_k can be expressed as:

$$x_k = [y_{k-1}\ y_{k-2} \cdots y_{k-na}\ u_{k-1}^T\ u_{k-2}^T \cdots\ u_{k-nb}^T]^T \tag{2.10}$$

where the orders of the output and input polynomial are na and nb with the relationship $n = na + m \times nb$, and $u \in R^m$ represents the input. Conventionally the prediction error ($e_k = y_k - x_k^T \theta_{I_k}$) is assumed to follow the Gaussian distribution when formulating the identification problem. As discussed in the introduction section, directly maximizing the Gaussian likelihood function lacks robustness. In order to deal with potential outliers or data with longer-than-Gaussian tails, we consider that e_k follows t distributions. A good property of the t distribution is that it can be decomposed into scaled Gaussian distributions and a Gamma distribution. In addition to the model identity I_k that is used to indicate the local model identity of the k-th sampling data point, another hidden variable which is called variance scale R_k, needs to be introduced, so that

$$y_k | (x_k, \theta_{I_k}, \sigma_{I_k}^2, R_k, I_k = i) \sim \mathcal{N}(x_k^T \theta_i, \sigma_i^2 / R_k), \tag{2.11}$$

for $k = 1, \cdots, N$, and

$$R_k | (\nu_{I_k}, I_k = i) \sim \mathcal{G}(\frac{\nu_i}{2}, \frac{\nu_i}{2}). \tag{2.12}$$

Thus, according to Section 2, given $x_k, \theta_{I_k}, \sigma^2_{I_k}, \nu_{I_k}$, and the model identity $I_k = i$, y_k follows a t distribution, i.e.,

$$y_k|(x_k, \theta_{I_k}, \sigma^2_{I_k}, \nu_{I_k}, I_k = i) \sim t(x_k^T\theta_i, \sigma_i^2, \nu_i). \tag{2.13}$$

In the proposed approach, an ARX model is adopted to capture the process dynamics around each operating point. The probability density at the current output y_k given the past input $\{u_{k-1}, \cdots, u_1\}$, the past output $\{y_{k-1}, \cdots, y_1\}$, the scheduling variable $\{T_k, \cdots, T_1\}$, and model parameters $\Phi = \{\Phi_1, \Phi_2, \cdots, \Phi_M\}$ where the parameter of i-th local model is denoted as $\Phi_i = \{\theta_i, \sigma_i^2, \nu_i\}$ $(i = 1, 2, \cdots, M)$, can be derived as:

$$\begin{aligned}
&P(y_k|y_{k-1}, \cdots, y_1, u_{k-1}, \cdots, u_1, T_k, \cdots, T_1, \Phi) \\
&= \sum_{i=1}^{M} P(y_k, I_k = i|x_k, T_k, \cdots, T_1, \Phi) \\
&= \sum_{i=1}^{M} \pi_{ik} P(y_k|x_k, T_k, \cdots, T_1, \Phi, I_k = i) \\
&= \sum_{i=1}^{M} \pi_{ik} P(y_k|x_k, \Phi_i),
\end{aligned} \tag{2.14}$$

where $x_k = [y_{k-1}, \cdots, y_1, u_{k-1}, \cdots, u_1]^T$. Given the local model identify at k-th sampling instant, the output y_k is independent of the scheduling variable T and other local model parameters $\Phi_{I_k \neq i}$, so that $P(y_k|x_k, T_k, \cdots, T_1, \Phi, I_k = i) = P(y_k|x_k, \Phi_i)$ is obtained. In addition, π_{ik} denotes the probability that the i-th model takes effect, also known as the mixing coefficient, which has the following expression:

$$\pi_{ik} = P(I_k = i|x_k, T_k, \cdots, T_1, \Phi). \tag{2.15}$$

Since the model identity at k-th sampling instant does not depend on the past input and output, past scheduling variables, and model parameters Φ when given the current scheduling variable, Eq. (2.15) can be simplified as $\pi_{ik} = P(I_k = i|T_k)$ with a constraint $\sum_{i=1}^{M} \pi_{ik} = 1$. Up to now, the probability density of output y_k given all the observed information has been formulated as the mixture t distributions, as derived in Eq. (2.14), where the distribution of each local model is formulated as a t component with the density $P(y_k|x_k, \Phi_i)$, which is same as $P(y_k|x_k, \theta_{I_k}, \sigma^2_{I_k}, \nu_{I_k}, I_k = i)$, and the mixing coefficient π_{ik}.

Once linear models around fixed operating points have been identified, the global model is obtained by model interpolation with the assistant of proper weighting func-

tions, and the estimated output at k-th instant can be expressed as a combination of estimation given by each local model, which is

$$y_k = \sum_{i=1}^{M} \alpha_i(T_k)\hat{y}_{ik} = \sum_{i=1}^{M} \alpha_i(T_k)x_k^T\hat{\theta}_i, \tag{2.16}$$

where $\alpha_i(T_k)$ is the weight of i-th local model at k-th sampling instant. In literature, several weighting functions are available, such as the cubic spline function [36], the piecewise function [37], and the exponential function [27]. Among them, the exponential function has only one parameter (validity width o) to determine when used as the weighting function, and can provide a smooth combination of local models. The exponential weighting function for i-th local model with respect to the scheduling variable at k-th sampling instant is expressed by the following equation:

$$w_{ik} = \exp\left(\frac{-(T_k - T_i)^2}{2(o_i)^2}\right), \tag{2.17}$$

where T_k denotes the measured or inferred scheduling variable at k-th sampling instant; T_i denotes the fixed operating points around which the desired products are produced, and they are assumed to be known a prior. The validity width of the i-th local model is denoted as o_i. The mixing coefficient π_{ik} is defined to be equivalent to the normalized weight, i.e.,

$$\pi_{ik} \triangleq \alpha_i(T_k) = \frac{w_{ik}}{\sum_{i=1}^{M} w_{ik}}, \quad \sum_{i=1}^{M} \pi_{ik} = 1. \tag{2.18}$$

Therefore, the parameters that need to be estimated are $\Theta = \{\Theta_1, \Theta_2, \cdots, \Theta_M\}$, where $\Theta_i = \{\Phi_i, o_i\}, i = 1, 2, \cdots, M$. The parameters can be estimated through the maximum likelihood approach, i.e.,

$$\begin{aligned}\Theta &= \arg\max_{\Theta} P(D_{obs}|\Theta) \\ &= \arg\max_{\Theta} P(y_N \ldots y_1, u_N \ldots u_1, T_N \ldots T_1|\Theta),\end{aligned} \tag{2.19}$$

where the joint probability density $P(y_N \ldots y_1, u_N \ldots u_1, T_N \ldots T_1|\Theta)$ can be expanded according to the chain rule,

$$P(y_N \ldots y_1, u_N \ldots u_1, T_N \ldots T_1|\Theta)$$

$$
\begin{aligned}
&\triangleq P(y_{N:1}, u_{N:1}, T_{N:1}|\Theta) \\
&= P(y_N|y_{N-1:1}, u_{N:1}, T_{N:1}, \Theta) \cdot P(y_{N-1:1}, u_{N:1}, T_{N:1}|\Theta) \\
&= P(y_N|y_{N-1:1}, u_{N:1}, T_{N:1}, \Theta) \cdot P(y_{N-1}|y_{N-2:1}, u_{N:1}, T_{N:1}, \Theta) \cdot P(y_{N-2:1}, u_{N:1}, T_{N:1}|\Theta) \\
&\quad \vdots \\
&= P(y_N|y_{N-1:1}, u_{N:1}, T_{N:1}, \Theta) \cdots P(y_2|y_1, u_{N:1}, T_{N:1}, \Theta) \cdot P(y_1, u_{N:1}, T_{N:1}|\Theta) \\
&= \prod_{k=2}^{N} P(y_k|y_{k-1:1}, u_{N:1}, T_{N:1}, \Theta) \cdot P(y_1|u_{N:1}, T_{N:1}, \Theta)C, \qquad (2.20)
\end{aligned}
$$

where C is the conditional probability density $P(u_{N:1}, T_{N:1}|\Theta)$, and it is a constant. Since the output at k-th instant, namely y_k, is not affected by current and future inputs $u_{N:k}$, and future scheduling variables $T_{N:k+1}$, Eq. (2.20) can be further simplified as

$$
P(y_N \dots y_1, u_N \dots u_1, T_N \dots T_1|\Theta) = P(y_1|T_1, \Theta)C \cdot \prod_{k=2}^{N} P(y_k|y_{k-1:1}, u_{k-1:1}, T_{k:1}, \Theta). \qquad (2.21)
$$

According to Eq. (2.14), $P(y_k|y_{k-1:1}, u_{k-1:1}, T_{k:1}, \Theta) = \sum_{i=1}^{M} \pi_{ik} P(y_k|x_k, \Theta_i)$. Thus, Eq. (2.19) can be finally written as,

$$
\Theta = \arg\max_{\Theta} \left\{ \sum_{i=1}^{M} \pi_{i1} P(y_1|\Theta_i)C \cdot \prod_{k=2}^{N} \sum_{i=1}^{M} \pi_{ik} P(y_k|x_k, \Theta_i) \right\}. \qquad (2.22)
$$

2.4 Robust LPV model identification using the EM algorithm

Estimating parameters by maximizing the likelihood function directly, as Eq. (2.22) does, is difficult. The expectation-maximization (EM) algorithm [38] can make the maximum-likelihood estimation of parameters feasible by iteratively maximizing a lower bound of the log-likelihood when the data has missing values or hidden variables.

In the EM algorithm, we consider that the observed data D_{obs} is incomplete, and the hidden variable D_{hid} is introduced to compose the complete data set $\{D_{obs}, D_{hid}\}$. The algorithm first performs the E-step where the conditional expectation of the log-likelihood of the complete data is derived. The expectation is referred to as the Q

function:

$$Q(\Theta|\Theta^{(n)}) = E_{D_{hid}|D_{obs},\Theta^{(n)}}\left\{\log P\left(D_{obs}, D_{hid}|\Theta\right)\right\}. \tag{2.23}$$

The second step, called the M-step, is to update the parameter Θ by maximizing the Q function. That is to find:

$$\Theta^{(n+1)} = \arg\max_{\Theta} Q(\Theta|\Theta^{(n)}). \tag{2.24}$$

These two steps are iteratively calculated. The log-likelihood function of the observed data is guaranteed to increase in each iteration, and the algorithm is guaranteed to converge to at least a local maximum of the likelihood function [39].

2.4.1 E-step

Under the assumption made in Section 3, the complete data set in this scenario is $\{D_{obs}, D_{hid}\} = \{(Y, U, T), (R, I)\}$. Thus, the Q function is written as

$$Q(\Theta|\Theta^{(n)}) = E_{R,I|\Theta^{(n)},Y,U,T}\{\log P(Y, U, T, R, I|\Theta)\}. \tag{2.25}$$

The log-likelihood of the complete data can be factorized into a summation of related probability densities, as derived in Eq. (2.26).

$$\begin{aligned}
\log P(Y, U, T, R, I|\Theta) &= \log P(Y|U, T, R, I, \Theta)P(R|U, T, I, \Theta)P(I|U, T, \Theta)P(U, T|\Theta) \\
&= \log \prod_{k=1}^{N} P(y_k|x_k, R_k, \Theta_{I_k})P(R_k|I_k, \nu_{I_k})P(I_k|T_k, \Theta_{I_k})C \\
&= \sum_{k=1}^{N}\{\log P(y_k|x_k, R_k, \Theta_{I_k}) + \log P(R_k|I_k, \nu_{I_k}) \\
&\quad + \log P(I_k|T_k, \Theta_{I_k})\} + \log C,
\end{aligned} \tag{2.26}$$

where C is the conditional probability $P(U, T|\Theta)$ that is a constant value. Therefore, the Q function, namely Eq. (2.25), can be expanded as:

$$\begin{aligned}
Q(\Theta|\Theta^{(n)}) = E_{R,I|\Theta^{(n)},Y,U,T}\{&\sum_{k=1}^{N} \log P(y_k|x_k, R_k, \Theta_{I_k}) + \sum_{k=1}^{N} \log P(R_k|I_k, \nu_{I_k}) \\
&+ \sum_{k=1}^{N} \log P(I_k|T_k, \Theta_{I_k}) + \log C\}.
\end{aligned} \tag{2.27}$$

Recall that $y_k|(x_k, R_k, \Theta_{I_k}) \sim \mathcal{N}(x_k^T\theta_{I_k}, \sigma_{I_k}^2/R_k)$ and $R_k|(I_k, \nu_{I_k}) \sim \mathcal{G}(\frac{\nu_{I_k}}{2}, \frac{\nu_{I_k}}{2})$, so that

$$\begin{aligned}\log P(y_k|x_k, R_k, \Theta_{I_k}) &= \log \frac{1}{\sqrt{(2\pi)\sigma_{I_k}^2/R_k}} e^{-\frac{R_k(y_k - x_k^T\theta_{I_k})^2}{2\sigma_{I_k}^2}} \\ &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma_{I_k}^2 + \frac{1}{2}\log R_k - \frac{R_k}{2{\sigma_{I_k}}^2}(y_k - x_k^T\theta_{I_k})^2,\end{aligned} \tag{2.28}$$

$$\begin{aligned}\log P(R_k|I_k, \nu_{I_k}) &= \log \frac{(\nu_{I_k}/2)^{\frac{\nu_{I_k}}{2}} {R_k}^{\frac{\nu_{I_k}}{2}-1}}{\Gamma(\frac{\nu_{I_k}}{2})} e^{-\frac{\nu_{I_k}}{2}R_k} \\ &= -\log\Gamma(\frac{1}{2}\nu_{I_k}) + \frac{1}{2}\nu_{I_k}\log(\frac{1}{2}\nu_{I_k}) + \frac{1}{2}\nu_{I_k}(\log R_k - R_k) - \log R_k.\end{aligned} \tag{2.29}$$

Thus the log-likelihood of the complete data can be expressed as a linear function of R_k, $\log R_k$ and I_k, and the above expectation calculation can be implemented by taking the expectation of each term over R_k and $\log R_k$, $k = 1, 2, \cdots, N$, conditional on $I_k = i$, as well as D_{obs} and $\Theta^{(n)}$, and then over I_k given D_{obs} and $\Theta^{(n)}$. To summarize, the expectation calculations boil down to

$$\begin{aligned} E(R_k|I_k = i, \Theta^{(n)}, Y, U, T), \\ E(\log R_k|I_k = i, \Theta^{(n)}, Y, U, T), \\ E(I_k|\Theta^{(n)}, Y, U, T). \end{aligned} \tag{2.30}$$

The posterior probability that the i-th component of the mixture generates the data point y_k, given the observed data set D_{obs} and the estimated parameters in the previous M-step $\Theta^{(n)}$, can be calculated based on the Bayes' rule:

$$\begin{aligned} P(I_k = i|\Theta^{(n)}, Y, U, T) &= P(I_k = i|\Theta^{(n)}, y_k, \cdots, y_1, u_{k-1}, \cdots, u_1, T_k, \cdots, T_1) \\ &= P(I_k = i|\Theta^{(n)}, y_k, x_k, T_k) \\ &= \frac{P(y_k|\Theta^{(n)}, x_k, T_k, I_k = i)P(I_k = i|\Theta^{(n)}, x_k, T_k)}{P(y_k|\Theta^{(n)}, x_k, T_k)} \\ &= \frac{\pi_{ik}^{(n)} P(y_k|{\Theta_i}^{(n)}, x_k, T_k, I_k = i)}{\sum_{i=1}^{M} \pi_{ik}^{(n)} P(y_k|{\Theta_i}^{(n)}, x_k, T_k, I_k = i)} \end{aligned}$$

$$= \frac{\pi_{ik}^{(n)} P(y_k|x_k, \sigma_i^{2(n)}, \nu_i^{(n)}, \theta_i^{(n)})}{\sum_{i=1}^{M} \pi_{ik}^{(n)} P(y_k|x_k, \sigma_i^{2(n)}, \nu_i^{(n)}, \theta_i^{(n)})} \triangleq \tau_{ik}^{(n)}. \quad (2.31)$$

The Gamma distribution is the conjugate prior distribution over R_k, and hence the conditional posterior distribution over R_k, namely, its distribution given $\{I_k = i, \Theta^{(n)}, Y, U, T\}$, follows a Gamma distribution as well [34]:

$$R_k|I_k = i, \Theta^{(n)}, Y, U, T \sim gamma\left(\frac{\nu_i^{(n)} + 1}{2}, \frac{\nu_i^{(n)} + \delta(y_k|{x_k}^T\theta_i^{(n)}, \sigma_i^{2(n)})}{2}\right). \quad (2.32)$$

From Eq. (2.32) we can get the expectation of the conditional posterior distribution over R_k and $\log R_k$ according to the property of Gamma distributions [32]:

$$E(R_k|I_k = i, \Theta^{(n)}, Y, U, T) = \frac{\nu_i^{(n)} + 1}{\nu_i^{(n)} + \delta(y_k|{x_k}^T\theta_i^{(n)}, \sigma_i^{2(n)})} \triangleq r_{ik}^{(n)}, \quad (2.33)$$

$$\begin{aligned} E(\log R_k|I_k = i, \Theta^{(n)}, Y, U, T) &= \psi(\frac{\nu_i^{(n)} + 1}{2}) - \log(\frac{\nu_i^{(n)} + \delta(y_k|{x_k}^T\theta_i^{(n)}, \sigma_i^{2(n)})}{2}) \\ &= \log r_{ik}^{(n)} + \{\psi(\frac{{\nu_i}^{(n)} + 1}{2}) - \log\frac{{\nu_i}^{(n)} + 1}{2}\}, \end{aligned} \quad (2.34)$$

where $\psi(\nu)$ is the digamma function that is equivalent to the first order derivative of $\Gamma(\nu)$ divided by $\Gamma(\nu)$, i.e.,

$$\psi(\nu) = \frac{\partial\Gamma(\nu)/\partial\nu}{\Gamma(\nu)}. \quad (2.35)$$

By using the results obtained in Eq. (2.31), Eq. (2.33), and Eq. (2.34), the Q function presented in Eq. (2.27), can be written as

$$Q(\Theta|\Theta^{(n)}) = \sum_{k=1}^{N}\sum_{i=1}^{M} \tau_{ik}^{(n)}\{Q_1(o_i) + Q_2(\nu_i) + Q_3(\theta_i, {\sigma_i}^2)\} + \log C \quad (2.36)$$

where

$$Q_1(o_i) = \left(\frac{e^{-\frac{(T_k - T_i)^2}{2o_i^2}}}{\sum_{i=1}^{M} e^{-\frac{(T_k - T_i)^2}{2o_i^2}}}\right), \quad (2.37)$$

$$Q_2(\nu_i) = -\log\Gamma(\frac{1}{2}\nu_i) + \frac{1}{2}\nu_i\log(\frac{1}{2}\nu_i) - \left(\log r_{ik}^{(n)} + \{\psi(\frac{{\nu_i}^{(n)} + 1}{2}) - \log\frac{{\nu_i}^{(n)} + 1}{2}\}\right)$$

$$+\frac{1}{2}\nu_i\left(\log r_{ik}^{(n)}+\{\psi(\frac{\nu_i^{(n)}+1}{2})-\log\frac{\nu_i^{(n)}+1}{2}\}-r_{ik}^{(n)}\right), \tag{2.38}$$

and

$$\begin{aligned}Q_3(\theta_i,{\sigma_i}^2) &= -\frac{1}{2}\log 2\pi-\log\sigma_i-\frac{r_{ik}^{(n)}}{2{\sigma_i}^2}(y_k-x_k^T\theta_i)^2\\ &\quad+\frac{1}{2}\left(\log r_{ik}^{(n)}+\{\psi(\frac{\nu_i^{(n)}+1}{2})-\log\frac{\nu_i^{(n)}+1}{2}\}\right).\end{aligned} \tag{2.39}$$

In short, we can see that the main task of the E-step is to determine $\tau_{ik}^{(n)}$ and $r_{ik}^{(n)}$. They are the key components of the Q function along with the observed data, the estimated parameters from the previous M-step, and the unknown parameters that need to be estimated in the next M-step.

2.4.2 M-step

The M step is to update parameters by maximizing the Q function derived as Eq. (2.36). Through the maximization, we can get analytical solutions for $\theta_i^{(n+1)}$ and ${\sigma_i^2}^{(n+1)}$. As for $o_i^{(n+1)}$ and $\nu_i^{(n+1)}$, we need to use numerical methods to search for the optimum. The updated parameters can be expressed as follows:

$$\theta_i^{(n+1)}=\frac{\sum_{k=1}^{N}\tau_{ik}^{(n)}r_{ik}^{(n)}x_k y_k}{\sum_{k=1}^{N}\tau_{ik}^{(n)}r_{ik}^{(n)}x_k x_k^T}, \tag{2.40}$$

$${\sigma_i^2}^{(n+1)}=\frac{\sum_{k=1}^{N}\tau_{ik}^{(n)}r_{ik}^{(n)}(y_k-x_k^T\theta_i^{(n+1)})^2}{\sum_{k=1}^{N}\tau_{ik}^{(n)}}, \tag{2.41}$$

$$\begin{aligned}o_i^{(n+1)} &= argmax_{o_i,i=1,2,\cdots,M}\left\{\sum_{k=1}^{N}\sum_{i=1}^{M}\tau_{ik}^{(n)}Q_1(o_i)\right\},\\ &s.t.\ o_{\min}\le o_i\le o_{\max}\end{aligned} \tag{2.42}$$

$$v_i^{(n+1)}: \ \frac{\partial}{\partial \nu_i} \sum_{k=1}^{N} \sum_{i=1}^{M} \tau_{ik}^{(n)} Q_2(\nu_i) = 0 \Leftrightarrow$$

$$-\psi(\tfrac{1}{2}\nu_i) + \log(\tfrac{1}{2}\nu_i) + 1 + \{\psi(\tfrac{\nu_i^{(n)}+1}{2}) - \log \tfrac{\nu_i^{(n)}+1}{2}\} + \frac{1}{\sum_{k=1}^{N} \tau_{ik}^{(n)}} \sum_{k=1}^{N} \tau_{ik}^{(n)} (\log r_{ik}^{(n)} - r_{ik}^{(n)}) = 0. \tag{2.43}$$

To reduce the computation load for searching the numerical optimum, we let M local t components have equal degrees of freedom; that is to say, let $\nu_1 = \nu_2 = \cdots = \nu_M = \nu$. This is reasonable as the degree of freedom represents the data quality. Thus the maximizing step for ν can be derived as

$$-\psi(\frac{1}{2}\nu)+\log(\frac{1}{2}\nu)+1+\{\psi(\frac{\nu^{(n)}+1}{2})-\log\frac{\nu^{(n)}+1}{2}\}+\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{M}\tau_{ik}^{(n)}(\log r_{ik}^{(n)} - r_{ik}^{(n)}) = 0. \tag{2.44}$$

2.4.3 Algorithm summary and discussion

The E-step and M-step are updated iteratively until the convergence of the algorithm. The log-likelihood of the observed data, which is calculated through Eq. (2.45) referring to Eq. (2.22), acts as the indicator of convergence:

$$L(\Theta^{(n+1)}) = \sum_{k=1}^{N} \log(\sum_{i=1}^{M} \pi_{ik}^{(n+1)} P(y_k|x_k, \Theta_i^{(n+1)})) + C. \tag{2.45}$$

When $|L(\Theta^{(n+1)}) - L(\Theta^{(n)})| \leq \epsilon$, model parameters converge to their true value, and the iteration stops. ϵ is the threshold of the stop criterion. In summary, the proposed robust algorithm is executed as Table 2.1 shows.

Table 2.1: Procedure of the proposed robust algorithm

1.	*Initialization.* Set $n = 0$. Assign random values to parameters $\Theta^{(0)}$.
2.	*E-step.* Evaluate $\tau_{ik}^{(n)}$ by Eq. (2.31) and $r_{ik}^{(n)}$ by Eq. (2.33).
3.	*M-step.* Update parameters $\theta_i^{(n+1)}$ by Eq. (2.40), $\sigma_i^{2(n+1)}$ by Eq. (2.41), $o_i^{(n+1)}$ by Eq. (2.42), and $\nu^{(n+1)}$ by Eq. (2.44).
4.	*Evaluate* $L(\Theta^{(n+1)})$. Calculate the new value of the log-likelihood of the observed data.
5.	*Check stop criterion.* If $\|L(\Theta^{(n+1)}) - L(\Theta^{(n)})\| \leq \epsilon$, stop. Otherwise, set $n = n + 1$, and go to step 2.

The equation for calculating $\theta_i^{(n+1)}$ has a similar form as the one in [8] which introduces a scaling weight to reduce the adverse influence of the outliers when solving the identification problem of switching process dynamics. However, unlike the scaling weight which is determined manually according to the priori knowledge as shown in [8], the weight $r_{ik}^{(n)}$ in this work is completely determined by the data set itself.

When the collected data set is contaminated with outliers, the estimated degrees of freedom $\nu^{(n)}$ will have a small value. Suppose that the current output y_k is generated around the i-th operating point. If it is an outlier, the distance between y_k and $x_k^T\theta_i^{(n)}$, namely $\delta(y_k|{x_k}^T\theta_i^{(n)}, \sigma_i^{2(n)})$, is large, and according to Eq. (2.33), $r_{ik}^{(n)}$ will be small. Hence, the k-th sampling data point will be down-weighted when estimating the model parameters using Eq. (2.40) and Eq. (2.41). On the other hand, If the sample is normal, $r_{ik}^{(n)}$ is close to 1, and thus the data point takes full effect in parameter estimation.

Besides, if the collected data has a good quality, i.e., the noise follows a Gaussian distribution, the estimated degrees of freedom $\nu^{(n)}$ will be large and the distance $\delta(y_k|{x_k}^T\theta_i^{(n)}, \sigma_i^{2(n)})$ will be small, and hence $r_{ik}^{(n)}$ is always close to 1. All data points collected around the i-th operating point take full effect in estimating parameters of the i-th local model, so the proposed robust approach is consistent with the conventional approach when the noise is Gaussian distributed. Therefore, the proposed approach is an adaptive approach to dealing with outliers.

2.5 Simulation & experiment

2.5.1 Numerical example

The first simulation example is adopted from [4]. Consider a first order continuous-time process with the following transfer function,

$$G(s,p) = \frac{K(p)}{\tau(p)s+1}, \tag{2.46}$$

where

$$K(p) = 0.6 + p^2,\ \tau(p) = 3 + 0.5p^3,\ p \in [1,\ 4]. \tag{2.47}$$

Both time constant τ and gain K vary in the operation range $p \in [1,\ 4]$. Therefore, a single linear model cannot capture the process behaviour in the whole operation

trajectory [4].

Consider first the three operating points, i.e., $p = 1,\ p = 2.25$ and $p = 4$. The true values of local model parameters, namely $\theta_i,\ i = 1, 2, 3$, are obtained by discretizing the continuous-time transfer function at the corresponding operating point. The scheduling variable p varies as follows (shown in Fig. 2.2):

- First period: 100 seconds, at the operating point $p = 1$;
- Second period: 300 seconds, the scheduling variable p varies linearly in time from 1 to 2.25;
- Third period: 150 seconds, at the operating point $p = 2.25$;
- Fourth period: 200 seconds, the scheduling variable p varies linearly in time from 2.25 to 4;
- Fifth period: 150 seconds, at the operating point $p = 4$.

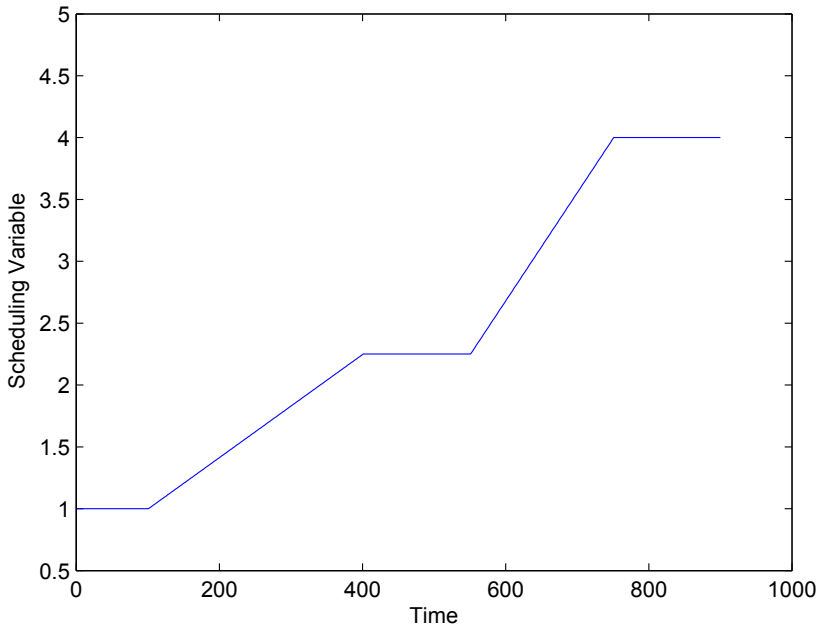

Figure 2.2: Variation of the noise-free scheduling variable p.

To identify the nonlinear process, a random binary sequence (RBS) signal is designed for persistently exciting the process around each operating point. The regular

LPV approach proposed in [27] only considers Gaussian distributed noise, the performance of which will be deteriorated by outliers. The proposed robust approach can eliminate the adverse effect of outliers owing to the adaptive degrees of freedom. How severe contamination the proposed approach can deal with, or what the breakdown point of the proposed approach is, is of interest to investigate. A set of Gaussian noise with mean 0 and variance 0.015 is generated as the process noise. To find the breakdown point of the proposed algorithm, we substitute part of the process noise by uniformly distributed outliers which are located in the range [-3.5, -3]. The percentage of outliers ranges from 0 to 50 % with a 1% incremental size. Fig. 2.3 shows the variation of degrees of freedom (ν) with the percentage of outliers, from which we can see that the breakdown point is about 38% outliers in this simulated case study. After 38%, the estimated degrees of freedom tend to be unstable and may be far away from the true value. Therefore, within a broad range of corruption, the proposed approach is capable of adapting to the various quality of data, that is to say, adaptively using large degrees of freedom to identify the model from data sets having good quality while using small degrees of freedom to deal with poor quality data sets. We also compare the results of parameter estimation obtained by the regular approach and the proposed robust approach. The relative error, as defined by Eq. (2.48), is used to indicate the accuracy of estimation. Fig. 2.4 shows the average relative error of 6 local-model parameters. The relative error given by the proposed robust approach is around 5%, and it is consistently smaller than the one given by the regular approach when the percentage of outliers ranges from 0% to 38%. In fact, the relative error from the regular approach is fairly large and unpredictable and Fig. 2.4 only shows the best case of the regular approach. After 38 % outliers, the proposed approach still outperforms the regular one, even though the estimated degrees of freedom are not accurate. It can be explained by Eq. (2.33). Although the estimated $\nu^{(n)}$ is much larger than the true value, it will not influence the weight $r_{ik}^{(n)}$ significantly. For the normal data point, the distance $\delta(y_k|{x_k}^T\theta_i^{(n)}, \sigma_i^{2(n)})$ is small, so $r_{ik}^{(n)}$ is close to 1. For the outlying data point, the distance is large, so $r_{ik}^{(n)}$ is smaller than 1. Therefore, even if $\nu^{(n)}$ is not well estimated, normal data points still take effect in parameter estimation while outliers are down-weighted.

$$Relative\ Error = \frac{|\hat{\theta} - \theta_{True}|}{\theta_{True}} \times 100\%. \tag{2.48}$$

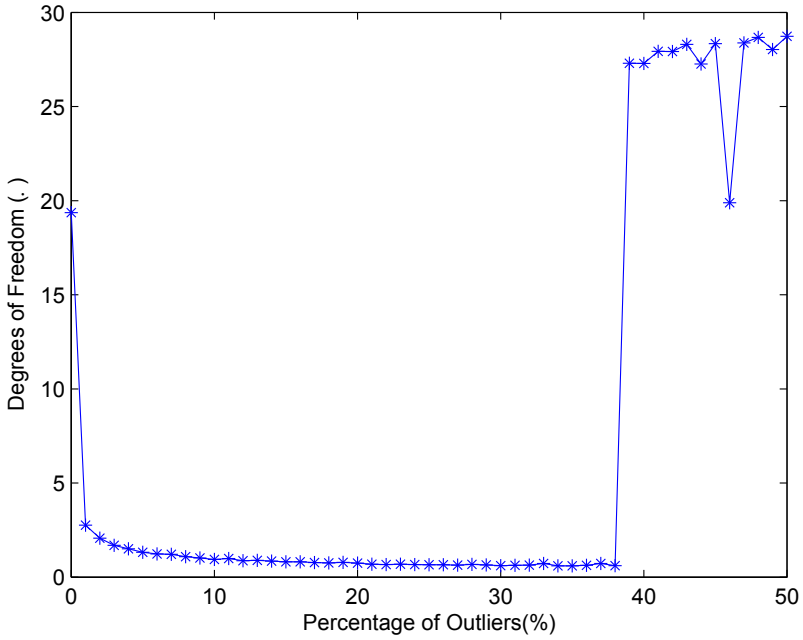

Figure 2.3: Variation of ν with the increasing degree of corruption.

To fairly compare the regular approach and the proposed robust approach, Monte Carlo simulations (50 runs) are performed in two cases:

Case I. the process noise is Gaussian distributed.

Case II. the process noise is contaminated with 10% outliers.

We list the parameters estimated by two methods in Table 2.2 for comparison. *True* denotes the true parameter values of local models at the operating point $p = 1$, $p = 2.25$, and $p = 4$, respectively. *Regular* means the estimated parameter by the regular LPV approach as proposed in [27]. *Robust* stands for the estimated parameter by the proposed robust LPV approach using mixture t distributions. The mean and one standard deviation of the estimation are presented in Table 2.2. In Case I, these two identification algorithms have a comparable performance, while in Case II, the comparison demonstrates that the proposed approach outperforms the regular approach in dealing with outliers. The estimated parameters by the proposed approach are closer to true values, and the standard deviations are smaller.

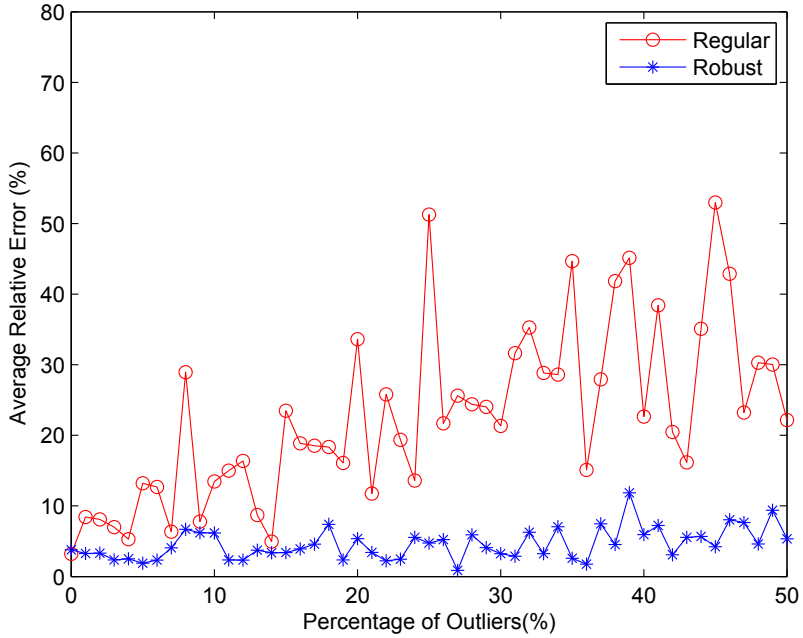

Figure 2.4: Comparison of average relative error of 6 local-model parameters from two approaches.

In addition, the proposed approach will have more significant advantage over the regular one when the data set is contaminated more severely according to Fig. 2.4.

In order to inspect the prediction performance of the identified LPV model by the proposed approach, both self-validation and cross-validation are considered. In self-validation, instead of using the training data set which contains a lot of outliers, we generate an outlier-free data set randomly around original operating points ($p = 1$, $p = 2.25$, and $p = 4$), and in cross-validation, we make use of a new data set that is generated around new operating points. To generate the data set for cross-validation, three new operating points, i.e., $p = 1.1$, $p = 2.15$, and $p = 3.9$, are considered. In addition to the regular approach and the proposed robust approach, a robust nonlinear process identification method is also compared. Here, the LS-SVM [20] which is a kernel-based learning method is considered. Under the situation where the noise is contaminated with 10% outliers, identification results from one random example of simulations are used for model validation, as shown in Fig. 2.5(a) and Fig. 2.5(b). The prediction of LPV model identified by the proposed approach captures the trend of the measurement in both self-validation and cross-validation, so does the

Table 2.2: Comparison of estimated parameters by two methods.

	θ_1	θ_2	θ_3
True	0.751	0.891	0.972
	0.398	0.615	0.468
Case I: No outliers in the process noise			
Regular	0.791($\pm$0.0158)	0.881($\pm$0.0080)	0.973($\pm$0.0038)
	0.389($\pm$0.0273)	0.624($\pm$0.0180)	0.481($\pm$0.0171)
Robust	0.793($\pm$0.0172)	0.880($\pm$0.0086)	0.973($\pm$0.0041)
	0.383($\pm$0.0302)	0.626($\pm$0.0186)	0.483($\pm$0.0167)
Case II: 10% outliers in the process noise			
Regular	0.839($\pm$0.0169)	0.936($\pm$0.0093)	0.993($\pm$0.0036)
	0.409($\pm$0.1247)	0.575($\pm$0.1394)	0.482($\pm$0.1344)
Robust	0.771($\pm$0.0106)	0.887($\pm$0.0050)	0.969($\pm$0.0016)
	0.438($\pm$0.0286)	0.599($\pm$0.0241)	0.476($\pm$0.0235)

prediction of the model fitted by the robust LS-SVM, indicating that both the LPV model and the LS-SVM model are capable of representing the nonlinear process. However, the LPV model identified by the proposed approach has a much simpler model structure than the kernel-based model given by LS-SVM. The performance of the regular approach is not good compared with other two approaches. The root mean square errors of self-validation (SV RMSE) and cross-validation (CV RMSE) are compared in Table 2.3. As shown in Table 2.3, the proposed robust approach has a better prediction performance than other two methods.

Table 2.3: Comparison of prediction performance of identified models.

Method	SV RMSE	CV RMSE
Regular	1.7097	1.2734
Robust	0.3794	0.3885
LS-SVM[20]	0.9561	1.1789

2.5.2 Continuous stirred tank reactor

The continuous stirred tank reactor (CSTR) is a widely used production unit in chemical and petrochemical processes. The CSTR studied here is an exothermic process with irreversible reaction $A \rightarrow B$ that has been utilized in literatures for nonlinear system state estimation and model predictive control [40, 41, 42]. The first principle model of the process can be written as follows according to the mass balance

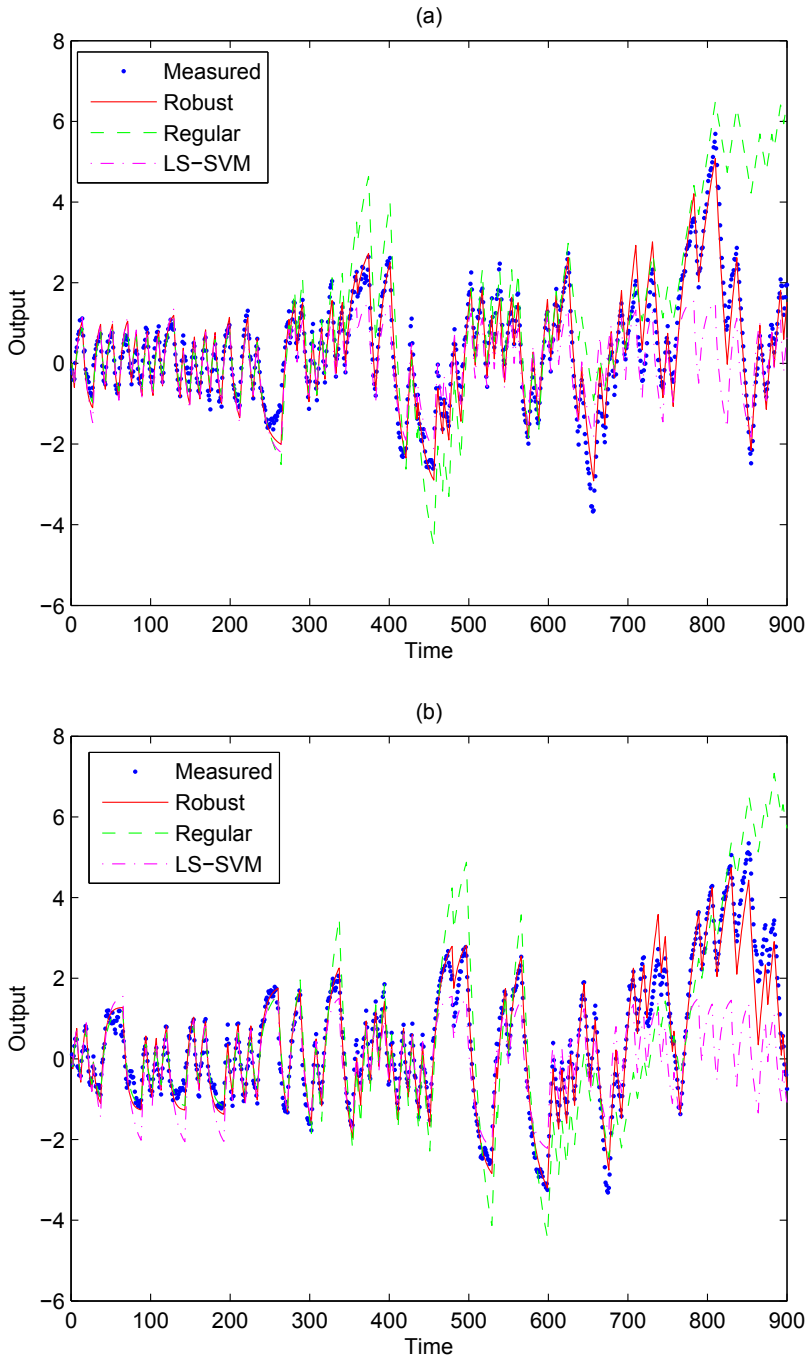

Figure 2.5: (a)Self-validation of identified models; (b)cross-validation of identified models. The data sets used in validation have no synthetic outliers.

and heat balance of the process [40]:

$$\frac{dC_A(t)}{dt} = \frac{q(t)}{V}(C_{A0}(t) - C_A(t)) - k_0 C_A(t) \exp\left(\frac{-E}{RT(t)}\right) \tag{2.49}$$

$$\begin{aligned}\frac{dT(t)}{dt} = {} & \frac{q(t)}{V}(T_0(t) - T(t)) - \frac{(-\Delta H)k_0 C_A(t)}{\rho C_p} \exp\left(\frac{-E}{RT(t)}\right) \\ & + \frac{\rho_c C_{pc}}{\rho C_p V} q_c(t) \left\{1 - \exp\left(\frac{-hA}{q_c(t)\rho C_p}\right)\right\} (T_{c0}(t) - T(t))\end{aligned} \tag{2.50}$$

The variables and their corresponding steady state values in the above two equations are listed in Table 2.4 [40].

Table 2.4: Variables and their steady state values of the CSTR process.

Parameter	Value	Unit
Product concentration of component A, C_A	$output_1$	mol/L
Temperature of the reactor, T	$output_2$	K
Coolant flow rate, q_c	$input$	L/min
Process flow rate, q	100	L/min
Feed concentration of component A, C_{A0}	1	mol/L
Feed temperature, T_0	350.0	K
Inlet coolant temperature, T_{c0}	350.0	K
Specific heats, $C_p,\ C_pc$	1	cal/(g K)
Heat transfer term, hA	7×10^5	cal/(min K)
Reactor volume, V	100	L
Liquid density, $\rho,\ \rho_c$	1×10^3	g/L
Heat of reaction, $-\Delta H$	-2×10^5	cal/mol
Activation energy term, E/R	1×10^4	K
Reaction rate constant, k_0	7.2×10^{10}	min^{-1}

In this simulation, the relationship between the coolant flow rate (q_c) and the product concentration of component A (C_A) is considered for building a single input single output (SISO) model. The input-output data set is generated from the simulated process. To verify the proposed method, we add measurement noise to the simulated output. The magnitude of the added white noise is 0.5% of the magnitude of the noise-free output, and 5% of the noise is randomly selected and replaced by outliers which are uniformly distributed in the range of [-0.03, 0.02]. The input-output data of the process contaminated by outliers is presented in Fig. 2.6. Since the coolant flow rate (q_c) influences the model significantly, it is chosen as the scheduling variable that varies from 97 L/min to 109 L/min [27]. Three operating points are taken into account, with q_c being 97 L/min, 103 L/min, and 109 L/min, respectively.

A LPV model is identified for the CSTR process by the proposed robust approach. For each local model, the normalized weight varies with different coolant flow rate, as shown in Fig. 2.7. When $q_c = 97L/min$, the weight of the first local model is larger than 0.95 while all other weights are close to 0. It means that the first local model is the dominant one when the scheduling variable is around 97 L/min. As q_c increases, the weight of the first model decreases quickly while that of the second one increases until q_c reaches the range dominated by the second local model, and so on. The effective range of the first local model is quite small because of a small validity width of this local model (o_1). The second local model has a bigger validity width, leading to a larger effective range, so does the third local model. Thus, the overall prediction of the identified model can be effectively represented by a weighted combination of prediction given by these three local models. The identified LPV model is much simpler than the original nonlinear process, but it approximates the complex CSTR process as demonstrated through the model validation.

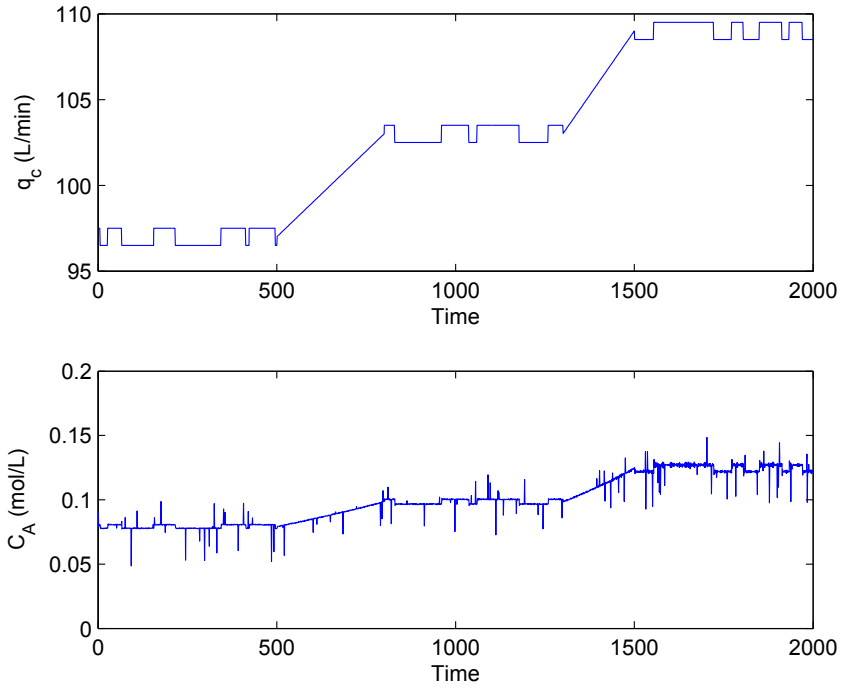

Figure 2.6: Input-output data set of the CSTR process.

Results of self-validation based on the training data ($q_c \in \{97, 103, 109\}$ L/min)

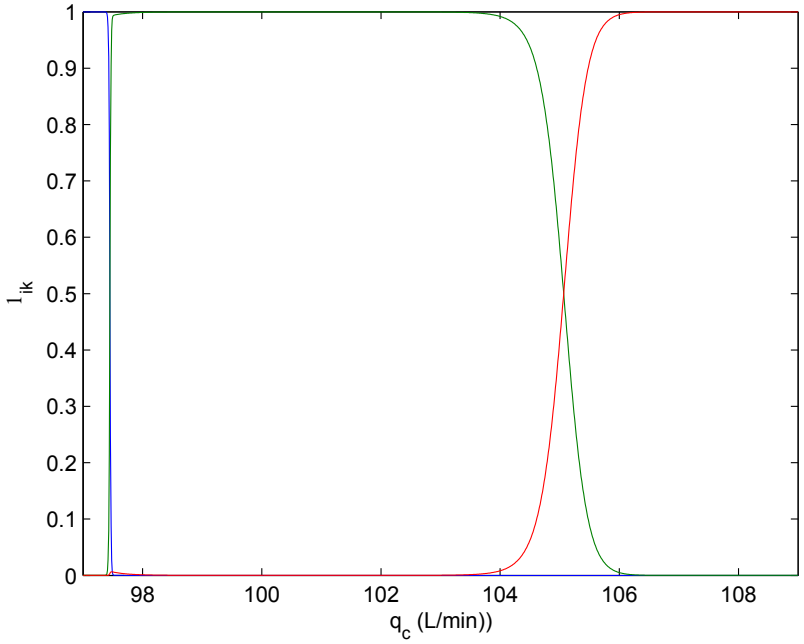

Figure 2.7: Normalized weight of each local model under different operating conditions (coolant flow rates).

are shown in Fig. 2.8. Cross-validation results around new operating points ($q_c \in \{98, 103, 107\}$ L/min) are presented in Fig. 2.9. In both figures, the predicted output of the model identified by the proposed method can capture the dynamics of the true process well, so does that of the model fitted by LS-SVM. The LPV model identified by the regular method is not as reliable as models from the other two methods. Table 2.5 shows the RMSE of both self-validation and cross-validation. It is clear that the performance of the proposed method is comparable with that of the LS-SVM method. However, compared with the nonlinear model fitted by LS-SVM, the LPV model identified by the proposed method is much simpler and more intuitive since it is a weighted combination of three local linear models.

Table 2.5: Comparison of prediction performance of identified CSTR models.

Method	SV RMSE	CV RMSE
Regular	0.00373	0.00237
Robust	0.00352	0.00050
LS-SVM[20]	0.00351	0.00038

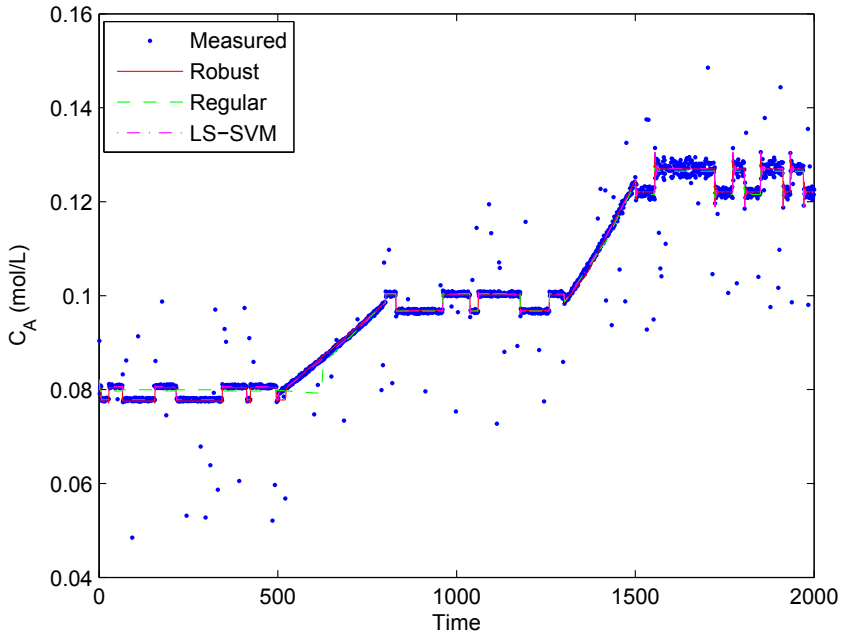

Figure 2.8: Self-validation of identified models for CSTR. Predicted output of the LPV model identified by the proposed method capture the dynamics of the true process well, so does that of the model fitted by LS-SVM.

2.5.3 Experiment evaluation

An experiment using a pilot-scale system called *balls-in-tubes* is conducted to further demonstrate the validity of the proposed approach. The experimental apparatus is presented in Fig. 2.10. The system has four modules. Each of them has a tube with a ping-pong ball inside, an ultrasonic sensor at the top to measure the distance between the sensor and the ball, and a DC fan at the bottom of the tube to lift the ball. The box under each tube can be viewed as a still balloon blown up by the fan, and its outlet blows air into the tube. These four tubes are connected at the fans' inlets by means of an input manifold that has a fan at the inlet on the left and an adjustable outlet on the right. This fan acts as the main fan of the system, sucking air into the system from the outside. An output manifold which has an outlet also connects tubes on the top. More detailed description of the experimental apparatus can be found in [43].

The manifolds play a key role in the experiment. The input manifold necessitates the sharing of air, and the output manifold restricts the air flow. They cause a

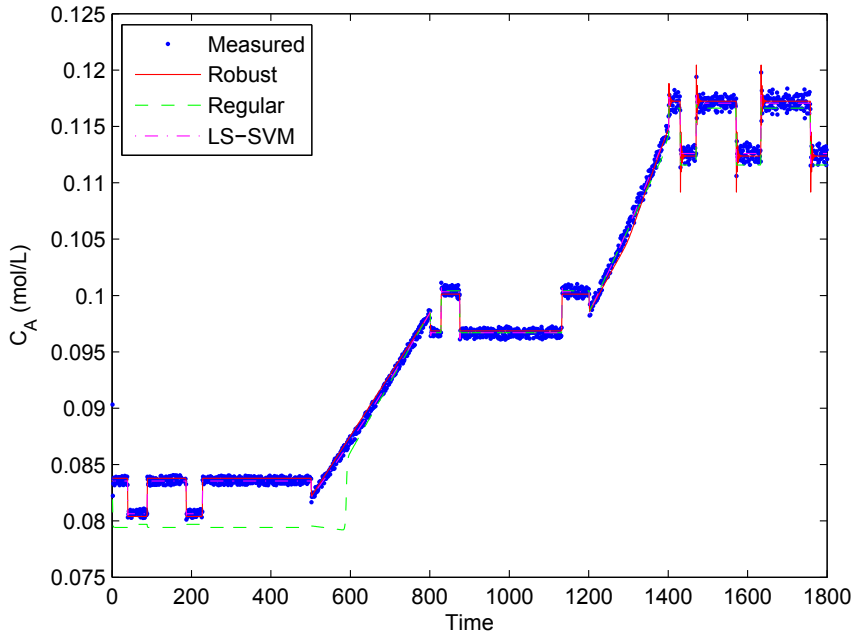

Figure 2.9: Cross-validation of identified models for CSTR. The LPV model identified by the regular method fails to predict the dynamics of the process when the coolant flow rate is low. Both the LPV model identified by the proposed method and the model fitted by LS-SVM can predict the dynamics of the true process.

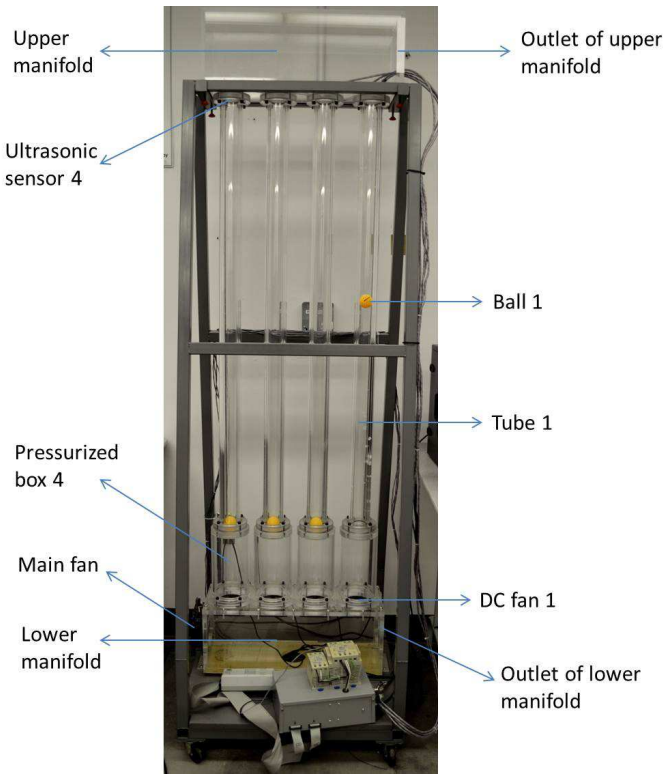

Figure 2.10: Experimental apparatus of the *balls-in-tubes* system.

significant coupling among the four tubes [43]. The speed of the main fan at the input manifold is fixed, so the amount of the air that can be allocated by DC fans is viewed as a constant. Therefore, if a significant amount of air is allocated to some tubes, then the fans for other tubes have less air that can be allocated. In our experiment, tube 1 is the module that we are interested in. The process model between the input, namely the fan speed, and the output, namely the ball's height, varies owing to the limited air in the manifold. When speeds of other fans increase, more air is allocated into other tubes, so that the fan for tube 1 has less air that can be allocated, resulting in the dropping of the ball if the fan speed remains constant. As a result, the model parameters of tube 1 will be influenced by the fan speed of other tubes. Therefore, the fan speed of other tubes can be considered as the scheduling variable for tube 1 because of their significant influence. For simplicity, only the fan for tube 4 is utilized while the fans for tube 2 and tube 3 respectively are not run. In other words, only the speed of the fan for tube 4 takes the role of the scheduling variable in this experiment. Meanwhile, a PID controller is used for maintaining the

height of the ball in tube 1 around the set point. Random binary sequence signal is designed as the set point signal to persistently excite the process. The controller's signal and the ball's height are sampled as the input and output data of the process. Fig. 2.11 shows the influence of the scheduling variable on the input of tube 1, and Fig. 2.12 presents the affected output of tube 1. These two figures indicate that it is appropriate to take the speed of fan 4 as the scheduling variable.

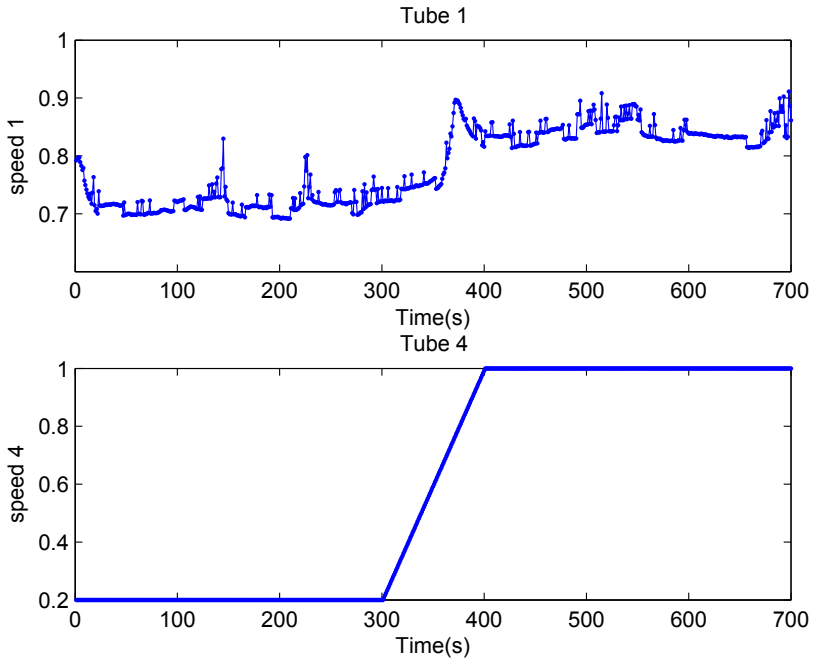

Figure 2.11: Influence of the scheduling variable to the input. When the speed of fan 4 increase, more air is allocated into tube 4, meaning the fan for tube 1 has less air that can be allocated. Fan 1 has to increase the speed to make the ball recover to the previous height.

In the experiment, the operating points are selected at 10%, 50%, and 100% of the maximum speed of fan 4 respectively, and the scheduling variable varies from one operating point to another with gradual transition, as shown in Fig. 2.13. The collected input and output data (training data) are presented in Fig. 2.14. The data set collected from the ultrasonic sensors is noisy, with some measurement spikes corresponding to errors greater than 10 cm [43], naturally representing the outliers. ARX models are adopted as local models around operating points. Utilizing the proposed robust algorithm, a LPV model is identified for representing this parameter

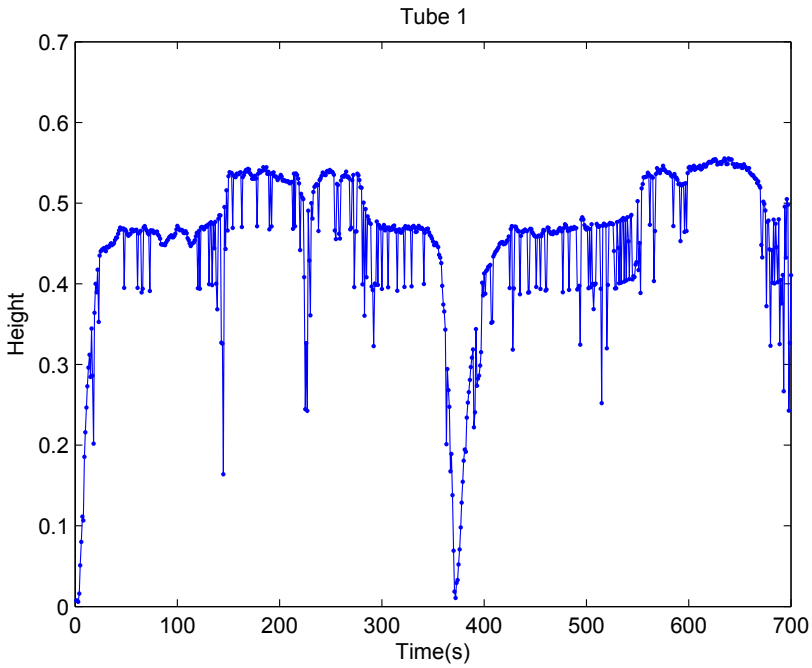

Figure 2.12: Influence of the scheduling variable to the output. The ball drops suddenly when the scheduling variable increases from 0.2 to 1. After the fan speed increases, the ball's height recovers to the set point.

varying process.

The prediction performance of identified models by different nonlinear process identification approaches is compared through model validation. The self-validation is first performed using the collected training data. Fig. 2.15 displays the validation result. Note that the input signal of the process is actually the control signal determined by the PID controller, so this is a closed-loop identification problem. When a measured output is an outlier, the control signal given by the PID controller changes sharply because of the large difference between the measurement and the set point, so the estimated output contains spikes too, corresponding to outliers in the measurement. To display the outliers in the measurement, the one-step ahead predictor is utilized. The residual between the measured output and one-step ahead prediction is presented in Fig. 2.16, from which we can see that quite a few measured data points are outliers. To further verify the identified LPV model, a data set collected around new operating points is used for cross-validation. The new operating points are at 30% and 80% of the maximum speed of fan 4, and the scheduling variable varies

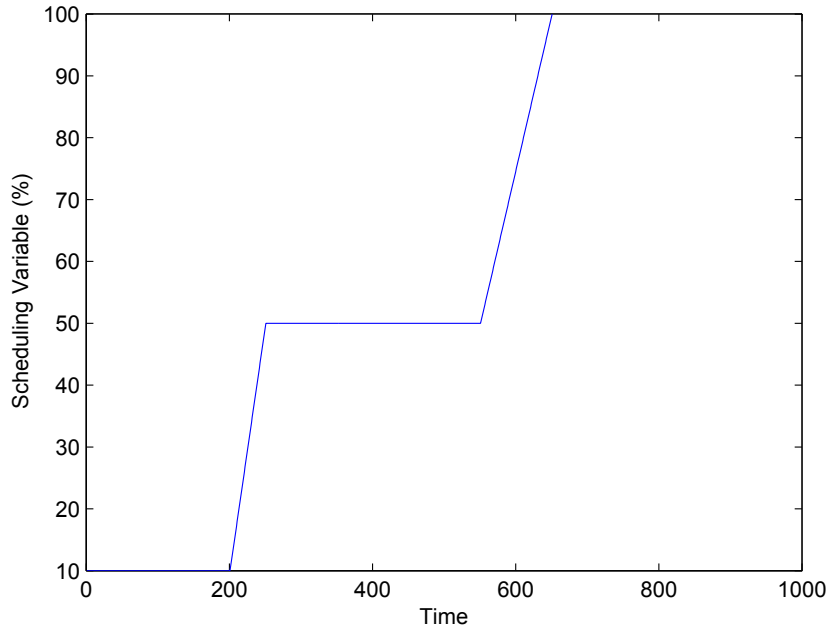

Figure 2.13: The operating points are at 10%, 50%, and 100% of the maximum speed respectively, and the scheduling variable varies from one operating point to another operating point with gradual transition.

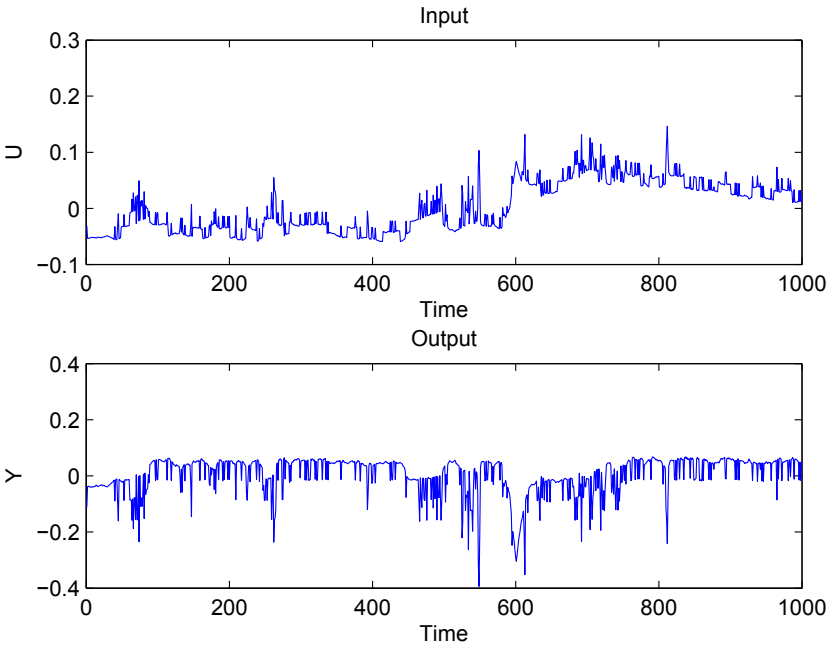

Figure 2.14: The collected input and output data (training data).

from 30% to 80% gradually. The result of cross-validation is shown in Fig. 2.17. The dynamic predictions of models identified by all three methods are comparable by graphic visualization. To have a closer look at the prediction performance, we list the RMSE of model validation in Table 2.6. Although the regular LPV identification approach has a smaller RMSE in self-validation, the prediction performance of the regular approach around new operating points is poorer than the proposed robust approach. The prediction performance of the model identified by LS-SVM is also poorer than the proposed even though it has a more complex model structure.

Table 2.6: Comparison of prediction performance.

Method	SV RMSE	CV RMSE
Regular	0.0315	0.0444
Robust	0.0355	0.0369
LS-SVM[20]	0.0385	0.0727

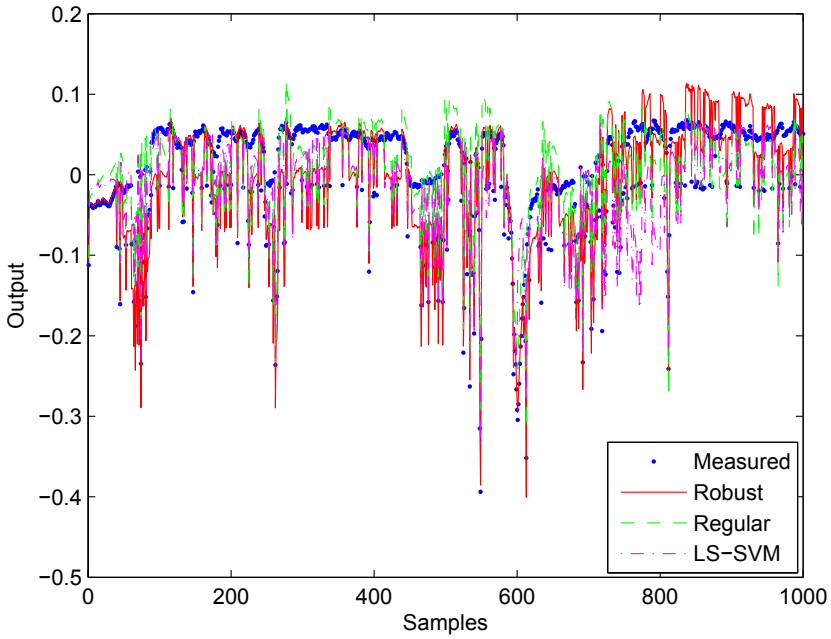

Figure 2.15: Self-validation of identified models. The infinite-step ahead predictions given by different models are compared with the measurement.

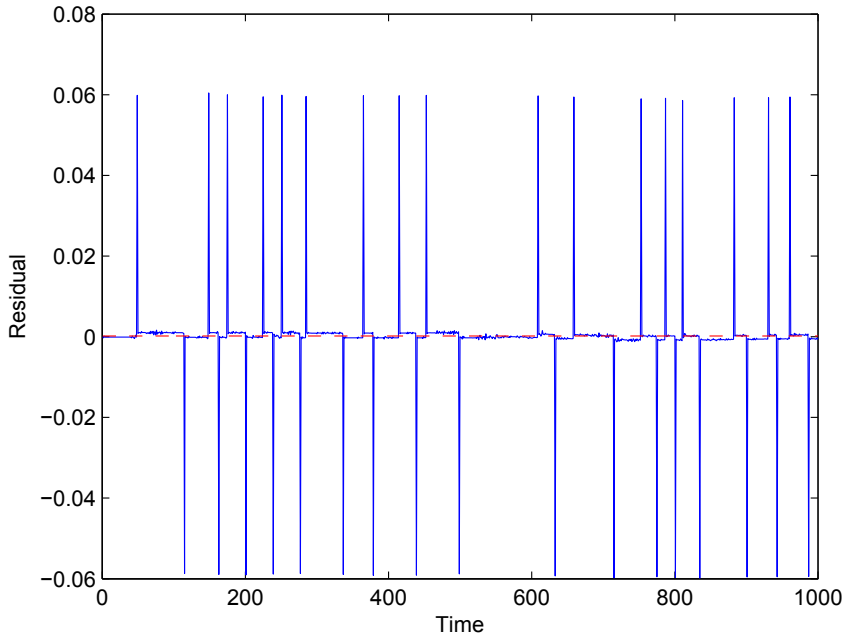

Figure 2.16: Residual between the one-step ahead prediction of the model identified by the proposed method and the measured output.

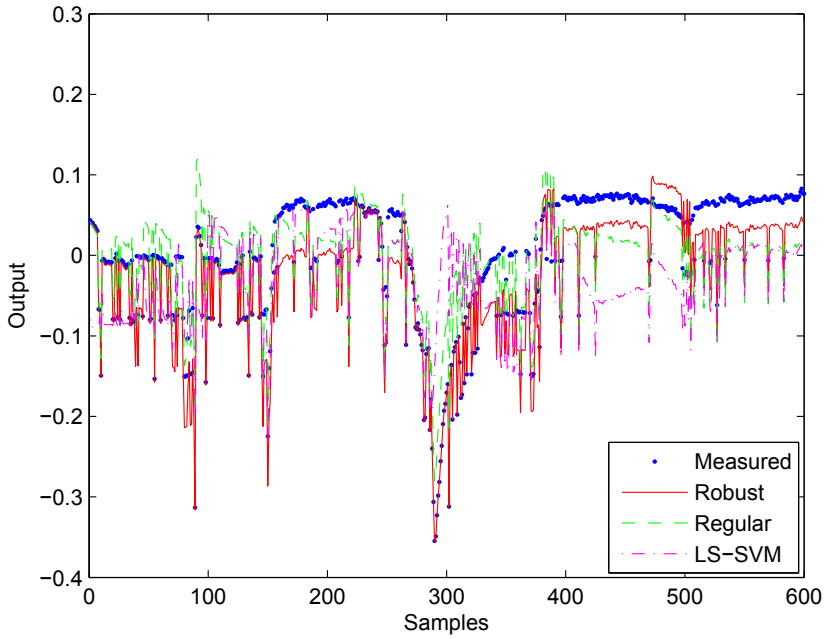

Figure 2.17: Cross-validation of identified models. The infinite-step ahead predictions given by different models are compared with the measurement around new operating points.

2.6 Conclusions

In this chapter we proposed a robust multiple-model LPV approach to identify the nonlinear process subject to outliers using mixture t distributions. The multiple-model LPV approach uses a weighted combination of local models around fixed operating points to approximate the nonlinear process across the whole operating range. In the proposed method, outliers are handled by using the mixture t distributions. Owing to the flexibility of degrees of freedom of t distributions, the proposed method shows good resistance to the influence of outliers and adaptability to data sets with different quality, and meanwhile it can also provide indication of the data quality. An effective algorithm for parameter estimation when identifying the LPV model was derived under the framework of EM algorithm. A numerical example and a simulated chemical process have demonstrated the advantage of the proposed method. The *balls-in-tubes* experiment has been conducted to further verify the effectiveness of proposed method.

Chapter 3

A Variational Bayesian Approach to Robust Identification of Switched ARX Models *

A variational Bayesian (VB) approach to robust identification of switched Auto-Regressive eXogenous (SARX) models is developed in this chapter. By formulating the problem of interest under a full Bayesian identification framework, the number of local models can be determined automatically, while accounting for the uncertainty of parameter estimates in the overall identification procedure. A set of significance coefficients is used to assign proper importance weights to the local models. The optimal values of significance coefficients are obtained by maximizing the marginal likelihood of the identification data. In this way, insignificant local models will be suppressed and the optimal number of local models can be determined. Considering the fact that the identification data may be contaminated with outliers, t distributions with adjustable tails are utilized to model the contaminating noise. The t distribution with smaller degrees of freedom can assign higher probability density to the longer tails so that the negative influence of outliers on the identification procedure can be reduced. Besides, given any new data, outliers can be detected by evaluating the predictive probability distribution. The effectiveness of the proposed Bayesian approach is demonstrated through simulated examples.

*This chapter is an extended version of the paper: Y. Lu, S. Khatibisepehr, and B. Huang, *A Variational Bayesian Approach to Identification of Switched ARX Models*, Proceedings of the 53rd IEEE Conference on Decision and Control (accepted, 2014).

3.1 Introduction

Industrial processes usually exhibit certain forms of non-linear behaviour. Suitable production policies might drive a chemical plant to switch among various operating conditions resulting in multiple modes or regimes of behaviour. Historical process data collected around different operating points may be generated by different dynamics. Thus, a single linear model may fail to capture the dynamics of a complex chemical process. In such a situation, hybrid models with multiple-model structures may be used to describe both continuous-state and discrete-state dynamics of the underlying non-linear processes [6]. Owing to its multiple-model nature, a hybrid model can capture different dynamics simultaneously. To represent a multi-modal process or a hybrid system, a switched local ARX (SARX) model is often used to approximate the non-linear dynamic behaviour of a process with various operating points [7, 8], due to its simplicity and effectiveness.

The problem of identifying SARX models has been considered widely and several approaches have been proposed such as clustering based techniques [44], particle filtering based algorithms [7], recursive least square methods [45], prediction error minimization methods [5] and expectation-maximization based algorithms [8]. Successful implementation of these methods often requires high-quality identification data as well as prior knowledge of the number of local-models, and most of them only deal with a special type of SARX models, called piecewise affine models in which the switching mechanism of the process is represented by regressor space partition. If the switching mechanism is regarded as switching along the time instead of regressor space partition, a general type of the switched process can be described[8], but the number of local-models is needed a priori.

However, there is often no prior information available about the number of local-models, so the condition required by most methods that the number of local-models is known *a priori* is inconsistent with real situation. Nakada *et al.* [6] utilized the information criteria such as the consistent Akaike's information criterion (CAIC) and the minimum description length (MDL) to determine the number of local-models. The authors performed comparison among a set of identified SARX models with different number of local-models, and selected the SARX model which had the smallest value of

information criteria as the final model with the optimal number of local-models. Vidal [46] proposed a recursive approach to estimate model parameters and the number of local-models simultaneously. However, both methods ignored the potential adverse influence of outliers. In practice, a random error caused by such issues as instrument degradation, transmission faults and process disturbances usually results in an outlier in process data [47]. Noisy measurements and outlying observations can significantly deteriorate the performance of conventional identification methods. Jin *et al.* [8] proposed an identification approach that could resist the adverse effect of outlying data, but their method addressed the problem in an *ad hoc* way, and it required a known number of local-models. Another shortcoming of these methods is that they merely obtain a set of single-valued parameter estimates. As a result, the uncertainty of parameters cannot be provided. To address the aforementioned issues, a more advanced and robust method needs to be developed in order to identify the SARX models from noisy training data with less need of subjective knowledge.

In this chapter, a new approach is developed for robust identification of the SARX models with the variational Bayesian (VB) learning method [48, 49]. Practical issues such as adverse influence of outliers, no or little prior knowledge of local-model numbers, and uncertainty of estimated parameters, are taken into account simultaneously. To handle the outliers, t distributions with adjustable degrees of freedom are utilized to model the contaminating noise. The t distribution with smaller degrees of freedom can assign higher probability density to the tails so that the negative influence of outliers on the identification procedure can be diminished [14]. t distributions are commonly utilized in robust modelling [15, 50]. Moreover, the t distribution has shown advantages in image processing and computer vision [33, 51]. An analytical expression for the approximated posterior probability density function over the model parameters is obtained by decomposing the t distribution into scaled Normal distributions and a Gamma distribution[15, 32]. Owing to the attractive property of the t distribution, the proposed approach can effectively adapt to various quality of identification data sets in order to make more reliable estimation. To determine the number of local-model, a set of significance coefficients is used to assign proper significance weights to local-models. The optimal values of significance coefficients are obtained by maximizing the marginal likelihood of the identification data [52]. In this way, local-

models that have insignificant dynamics will be suppressed and the optimal number of the local-models can be determined. In the proposed approach, the VB algorithm iteratively maximizes a lower bound of the marginal data likelihood, from which a set of approximating posterior distributions over model parameters can be obtained after the lower bound converges. Thus, the uncertainty of estimated parameters is also taken into account when determining the optimal number of local-models.

The proposed approach is discussed in detail in the remainder of this chapter. Section 2 briefly revisits the variational Bayesian algorithm. The illustration of the SARX identification problem is presented in Section 3. Section 4 is the mathematical formulation of the proposed approach under the variational Bayesian framework. In Section 5, identification of a simulated SARX process and a non-linear model is used to verify the proposed approach. Section 6 concludes this chapter.

3.2 Revisit of variational Bayesian approach

Let Y be a set of observed data, X be the hidden variable and Ω be the model structure. In Bayesian model identification, the marginal likelihood or evidence $p(Y|\Omega)$ is the key quantity to be evaluated. Usually the following integral has to be solved:

$$p(Y|\Omega) = \int p(X, \theta|\Omega) p(Y|X, \theta, \Omega) dX d\theta, \tag{3.1}$$

where θ is the parameter vector and $p(Y|X, \theta, \Omega)$ is the data likelihood. The integral is often intractable as a result of introducing the distribution of parameters. The VB approach makes the integration tractable by introducing a free joint distribution $q(X, \theta)$. That is,

$$\ln p(Y|\Omega) = \ln \int q(X, \theta) \frac{p(Y, X, \theta|\Omega)}{q(X, \theta)} dX d\theta \tag{3.2a}$$

$$\geq \int q(X, \theta) \ln \frac{p(Y, X, \theta|\Omega)}{q(X, \theta)} dX d\theta. \tag{3.2b}$$

Due to the concavity of the logarithm function, (3.2b) is obtained from (3.2a) by employing Jensen's inequality. Under the VB framework, the free joint distribution can be factorized [53], namely $q(X, \theta) \approx q(X)q(\theta)$. Then we have

$$\ln p(Y|\Omega) \geq \int q(X)q(\theta) \ln \frac{p(Y, X, \theta|\Omega)}{q(X)q(\theta)} dX d\theta \triangleq F_\Omega[q(X), q(\theta)]. \tag{3.3}$$

The difference between $\ln p(Y|\Omega)$ and the lower bound $F_\Omega[q(X), q(\theta)]$ can be evaluated by the following equation:

$$\begin{aligned}
&\ln p(Y|\Omega) - F_\Omega[q(X), q(\theta)] \\
&= \int q(X)q(\theta) \left\{ \ln \frac{p(Y, X, \theta|\Omega)}{p(X, \theta|Y, \Omega)} - \ln \frac{p(Y, X, \theta|\Omega)}{q(X)q(\theta)} \right\} dX d\theta \\
&= \int q(X)q(\theta) \ln \frac{q(X)q(\theta)}{p(X, \theta|Y, \Omega)} dX d\theta \triangleq KL(q||p). \qquad (3.4)
\end{aligned}$$

$KL(q||p)$ denotes the Kullback-Leibler (KL) divergence between the free distribution $q(X, \theta)$ and the true joint posterior distribution of hidden variables and parameters, namely $p(\theta, X|Y, \Omega)$. The VB algorithm optimizes the free distributions $q(X)$ and $q(\theta)$ iteratively by maximizing the lower bound so that the difference between $q(X, \theta)$ and $p(\theta, X|Y, \Omega)$ is minimized. By taking the functional derivatives of the lower bound with respect to $q(X)$ and $q(\theta)$ and equating them to zeros, the following update equations can be obtained [53]:

$$q(X)^{(t+1)} \propto \exp \left[\int \ln p(X, Y|\theta, \Omega) q(\theta)^{(t)} d\theta \right], \qquad (3.5a)$$

$$q(\theta)^{(t+1)} \propto p(\theta|\Omega) \exp \left[\int \ln p(X, Y|\theta, \Omega) q(X)^{(t+1)} dX \right]. \qquad (3.5b)$$

where superscript (t) denotes the number of iterations. These two iterative updates will not stop until convergence. They resemble the E-step and M-step of the EM algorithm. Thus, (3.5a) and (3.5b) are often referred to as variational Bayesian E-step and M-step, respectively. Since the VB approach estimates the posterior distribution over hidden variables and parameters, the uncertainty of parameter estimates is obtained after the algorithm converges.

3.3 Problem statement

Industrial processes are often non-linear and may run under various operating conditions. Variations in the feed flow-rate, catalyst type, and operating temperature are some of the factors that may change the operating modes. Thus one single linear model usually fails to capture the complete dynamics of an industrial process. Hybrid models such as SARX models are often used to address this issue. Although the

problem of SARX model identification has been studied by many researchers, several challenging issues in this area are open for further investigation, such as robust identification and determining number of local models. In this work, we will address the identification of SARX models by embedding t distributions into the variational Bayesian framework. The robustness and ability to determine local-model numbers of the proposed algorithm will be elaborated. Consider a switched linear ARX model defined by

$$y_k = x_k^T \theta_i + \epsilon_i(k), \tag{3.6}$$

where $x_k \in R^h$ is the regressor, and $k = \{1, 2, \cdots, N\}$ denotes the sampling instant. $\theta_i \in R^h$ is i-th local-model parameters, and $i = \{1, 2, \cdots, m\}$ is the local-model identity that indicates the random model switching from one operating mode to another operating mode. The noise of the local-model is denoted as $\epsilon_i(k)$. The regression vector has the following form:

$$x_k = [y_{k-1}\ y_{k-2}\ \cdots\ y_{k-n_a}\ u_{k-1}^T\ u_{k-2}^T\ \cdots\ u_{k-n_b}^T\ 1]^T, \tag{3.7}$$

where $u_k \in R^d$ and $y_k \in R$ are the input and output of the process with orders n_a and n_b, and $n_a + d \cdot n_b + 1 = h$ where n_a and n_b are assumed to be known.

Conventionally, the noise $\epsilon_i(k)$ is assumed to be zero-mean Normal distributed with precision (inverse variance) δ_i, where the identification results may be greatly influenced by outliers. Compared with a Normal distribution, a t distribution, with adjustable longer-than-Normal tails, can resist the adverse effect of outliers better. Therefore, in order to robustly identify the SARX model, the noise $\epsilon_i(k)$ is considered to follow the t distribution. Given the regression vector, local-model parameters, and degrees of freedom of the t distribution (ν_i), the output also follows a t distribution,*i.e.*,

$$Y_k | \{x_k, \theta_i, \delta_i, \nu_i\} \sim t(x_k^T \theta_i, \delta_i, \nu_i), \tag{3.8}$$

where the t distribution has the following expression:

$$t(y_k | x_k^T \theta_i, \delta_i, \nu_i) = \frac{\Gamma(\frac{\nu_i+1}{2}) \delta_i^{1/2}}{(\pi\nu_i)^{1/2} \Gamma(\frac{\nu_i}{2}) \{1 + \delta_i (y_k - x_k^T \theta_i)^2 / \nu_i\}^{\frac{\nu_i+1}{2}}}, \tag{3.9}$$

and $\Gamma(\nu)$ is the Gamma function which is $\Gamma(\nu) = \int_0^\infty z^{\nu-1} e^{-z} dz$.

A property of the t distribution is that it can be expressed by an infinite mixture of scaled Normal distributions[15], which makes it feasible to obtain analytical solutions in the derivation. A hidden variable R_k representing the scale of the noise at the k-th sampling instant is introduced. Thereby, the t distribution with degrees of freedom ν_i can be decomposed into scaled Normal distributions and a Gamma distribution [15], *i.e.*,

$$t(y_k|x_k^T\theta_i, \delta_i, \nu_i) = \int_0^\infty \mathcal{N}(y_k|x_k^T\theta_i, R_k, \delta_i)\mathcal{G}(R_k|\frac{\nu_i}{2}, \frac{\nu_i}{2})dR_k, \tag{3.10}$$

where

$$\mathcal{N}(y_k|x_k^T\theta_i, R_k, \delta_i) = \frac{(R_k\delta_i)^{1/2}}{\sqrt{2\pi}} e^{-\frac{R_k\delta_i(y_k - x_k^T\theta_i)^2}{2}}, \tag{3.11a}$$

$$\mathcal{G}(R_k|\frac{\nu_i}{2}, \frac{\nu_i}{2}) = \frac{1}{\Gamma(\frac{\nu_i}{2})}(\frac{\nu_i}{2})^{\frac{\nu_i}{2}} R_k^{\frac{\nu_i}{2}} e^{-\frac{\nu_i}{2}R_k}. \tag{3.11b}$$

The local-model identity I and the noise scale R are taken as the hidden variable H, *i.e.*, $H_k = \{I_k, R_k\}$, $I_k = i$ and $k = 1, 2, ..., N$, where N is the number of collected data points. The parameter set of the SARX model, denoted as Θ, includes local-model parameters θ, parameter precision β, noise precision δ, and degrees of freedom ν. For instance, if a SARX process has m local-models where m needs to be inferred, the parameters that need to be estimated are $\{\Theta_1, \Theta_2, \cdots, \Theta_m\}$, and Θ_i includes $\Phi_i = \{\theta_i, \beta_i, \delta_i\}$ and ν_i, $i = i, 2, ..., m$.

In reality, the true number of local-models is unknown in SARX model identification process. By taking advantages of the variational Bayesian approach, the number of local-models m can be inferred from a given data set automatically instead of being pre-assigned as the traditional approaches do. Suppose that M is an upper bound of the number of local-models, where some local-models are insignificant. We use a set of significance coefficients $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_M\}$ to indicate the significance of each local-model with a constraint that

$$\sum_{i=1}^{M} \alpha_i = 1, \ \alpha_i \in [0, 1], \tag{3.12}$$

where $\alpha_i = 0$ means that the corresponding local-model is redundant, and $\alpha_i = 1$ indicates that the switched model is only composed of the i-th local-model. Meanwhile, the prior probability of $I_k = i$ is defined as $p(I_k = i|\alpha) = \alpha_i$. The marginal likelihood (evidence) is conditioned on the significance coefficients α, namely $p(Y, U|\alpha)$. We are

intended to optimize the values of α by maximizing the conditional evidence $p(Y, U|\alpha)$. The significance coefficient α_i corresponding to the insignificant local-models tend to converge to zero.

3.4 Robust identification of SARX models

3.4.1 Prior

The joint prior distribution over model parameters Φ can be expressed as

$$p(\Phi|\alpha) = \prod_{i=i}^{M} p(\theta_i|\beta_i)p(\beta_i)p(\delta_i), \tag{3.13}$$

and, as corresponding conjugates, the Normal distribution and the Gamma distribution are adopted as prior distributions respectively:

$$p(\theta_i|\beta_i) = \mathcal{N}(0, \beta_i^{-1} I_{h\times h}), \tag{3.14a}$$

$$p(\beta_i|a_0, b_0) = \mathcal{G}(a_0, b_0), \tag{3.14b}$$

$$p(\delta_i|c_0, d_0) = \mathcal{G}(c_0, d_0), \tag{3.14c}$$

where $I_{h\times h}$ is an identity matrix, and a_0, b_0, c_0 and d_0 are constant hyper-parameters. The optimization of the degrees of freedom ν_i is treated separately from other parameters, and it will be considered in the variational Bayesian M-step.

3.4.2 Variational Bayesian E-step

In the variational Bayesian E-step, we maximize the lower bound $F[q(R, I), q(\Phi)]$ with respect to $q(R, I)$ by fixing $q(\Phi)$ and ν. The lower bound $F[q(R, I), q(\Phi)]$ can be formulated as

$$F[q(R, I), q(\Phi)] = \sum_{I} \int q(R, I)q(\Phi) \ln \frac{p(Y, U, R, I, \Phi, \nu|\alpha)}{q(R, I)q(\Phi)} dR d\Phi \tag{3.15}$$

By solving the following optimization problem:

$$\max_{q(R,I)} \{F[q(R, I), q(\Phi)]\}, \ \textit{subject to} \sum_{I} \int q(R, I)dR = 1, \tag{3.16}$$

we get:

$$q(R, I) = \frac{1}{Z_{R,I}} e^{\langle \ln p(Y,U,R,I|\Phi,\nu,\alpha)\rangle_{q(\Phi)}}, \tag{3.17a}$$

$$Z_{R,I} = \sum_{I} \int e^{\langle \ln p(Y,U,R,I|\Phi,\nu,\alpha)\rangle_{q(\Phi)}} dR, \tag{3.17b}$$

where $Z_{R,I}$ is a normalizing constant, and $\langle\cdot\rangle_{q(\Phi)}$ denotes the expectation operation.

The above derived equation is intractable, so further simplification should be conducted. The log-likelihood of complete data can be expanded as follows:

$$\begin{aligned}
\ln p(Y,U,R,I|\Phi,\nu,\alpha) &= \ln\{p(Y|U,R,I,\Phi,\nu,\alpha)p(R|I,U,\Phi,\nu,\alpha)p(I|U,\Phi,\nu,\alpha)p(U|\Phi,\nu,\alpha)\} \\
&= \ln\left\{\prod_{k=1}^{N} p(y_k|x_k,R_k,I_k,\Phi)p(R_k|I_k,\nu_{I_k})p(I_k|\alpha)p(U|\Phi,\nu,\alpha)\right\} \\
&= \sum_{k=1}^{N} \ln\{p(y_k|x_k,R_k,I_k,\Phi)p(R_k|I_k,\nu_{I_k})p(I_k|\alpha)c\}.
\end{aligned} \tag{3.18}$$

$Z_{R,I}$ can then be expressed as

$$\begin{aligned}
Z_{R,I} &= \sum_{I} \int e^{\langle \ln p(Y,U,R,I|\Phi,\nu,\alpha)\rangle_{q(\Phi)}} dR \\
&= \prod_{k=1}^{N} \sum_{I_k} \int e^{\left\langle \ln\left\{p(y_k|x_k,R_k,I_k,\Phi)p(R_k|I_k,\nu_{I_k})p(I_k|\alpha)c\right\}\right\rangle_{q(\Phi)}} dR_k.
\end{aligned} \tag{3.19}$$

Define

$$Z_{R_k,I_k} = \sum_{I_k} \int e^{\left\langle \ln\left\{p(y_k|x_k,R_k,I_k,\Phi)p(R_k|I_k,\nu_{I_k})p(I_k|\alpha)c\right\}\right\rangle_{q(\Phi)}} dR_k, \tag{3.20}$$

and then we have $Z_{R,I} = \prod_{k=1}^{N} Z_{R_k,I_k}$. As a result, the joint distribution $q(R,I)$ can be written as

$$\begin{aligned}
q(R,I) &= \frac{1}{\prod_{k=1}^{N} Z_{R_k,I_k}} \prod_{k=1}^{N} e^{\left\langle \ln\left\{p(y_k|x_k,R_k,I_k,\Phi)p(R_k|I_k,\nu_{I_k})p(I_k|\alpha)c\right\}\right\rangle_{q(\Phi)}} \\
&= \prod_{k=1}^{N} \frac{1}{Z_{R_k,I_k}} e^{\left\langle \ln\left\{p(y_k|x_k,R_k,I_k,\Phi)p(R_k|I_k,\nu_{I_k})p(I_k|\alpha)c\right\}\right\rangle_{q(\Phi)}}.
\end{aligned} \tag{3.21}$$

On the other hand, based on the independent identical distribution ($i.i.d$) assumption of hidden variables at each sampling instant, $q(R,I)$ can be decomposed as follows:

$$q(R,I) = \prod_{k=1}^{N} q(R_k,I_k). \tag{3.22}$$

Therefore, the joint density $q(R_k, I_k)$ can be calculated by the following equation:

$$q(R_k, I_k) = \frac{1}{Z_{R_k,I_k}} e^{\left\langle \ln\left\{p(y_k|x_k,R_k,I_k,\Phi)p(R_k|I_k,\nu_{I_k})p(I_k|\alpha)c\right\}\right\rangle_{q(\Phi)}}. \tag{3.23}$$

To simplify the expression, define

$$A_{ki} = e^{\langle \ln\{p(y_k|x_k,R_k,I_k=i,\Phi)p(R_k|I_k=i,\nu_i)p(I_k=i|\alpha)c\}\rangle_{q(\Phi)}}, \tag{3.24}$$

and let $B_{ki} = \int A_{ki} dR_k$. Then we have

$$q(R_k, I_k = i) = \frac{A_{ki}}{\sum_{i=1}^{M} B_{ki}}. \tag{3.25}$$

By integrating R_k out of the joint density, we get the probability of $I_k = i$, *i.e.*,

$$q(I_k = i) = \int q(R_k, I_k = i) dR_k = \frac{B_{ki}}{\sum_{i=1}^{M} B_{ki}}. \tag{3.26}$$

Then, the conditional density $q(R_k|I_k = i)$ can be obtained as

$$q(R_k|I_k = i) = \frac{q(R_k, I_k = i)}{q(I_k = i)} = \frac{A_{ki}}{B_{ki}}. \tag{3.27}$$

Clearly, the key point of the variational Bayesian E-step is to obtain the expression of A_{ki} and B_{ki}. Recall Eq. (3.11a), Eq. (3.11b), and $p(I_k = i|\alpha) = \alpha_i$. Then the log-likelihood of the complete data can be rewritten as

$$\begin{aligned}
&\ln\{p(y_k|x_k, R_k, I_k = i, \Phi)p(R_k|I_k = i, \nu_i)p(I_k = i|\alpha)c\} \\
=&\left\{\frac{1}{2}(-\ln 2\pi + \ln R_k + \ln \delta_i) - \frac{R_k\delta_i}{2}(y_k - x_k^T\theta_i)^2\right\} \\
+&\left\{-\frac{\nu_i}{2}R_k + \frac{\nu_i}{2}\ln\frac{\nu_i}{2} + (\frac{\nu_i}{2} - 1)\ln R_k - \ln\Gamma(\frac{\nu_i}{2})\right\} + \ln\alpha_i + \ln c.
\end{aligned} \tag{3.28}$$

Consequently, the expected log-likelihood in Eq. (3.24) is expressed as

$$\begin{aligned}
&\langle \ln\{p(y_k|x_k, R_k, I_k = i, \Phi)p(R_k|I_k = i, \nu_i)p(I_k = i|\alpha)c\}\rangle_{q(\Phi)} \\
=&\ln(\omega_{ki}R_k^{(f_{ki}-1)}) - g_{ki}R_k,
\end{aligned} \tag{3.29}$$

with

$$\omega_{ki} = \frac{\alpha_i\widehat{\delta}_i^{1/2}(\frac{\nu_i}{2})^{\frac{\nu_i}{2}}c}{\sqrt{2\pi}\Gamma(\frac{\nu_i}{2})}, \tag{3.30a}$$

$$f_{ki} = \frac{\nu_i + 1}{2}, \tag{3.30b}$$

$$g_{ki} = \frac{1}{2}\left\{\nu_i + \overline{\delta}_i\left(y_k^2 - 2y_k x_k^T \overline{\theta}_i + x_k^T \left\langle \theta_i \theta_i^T \right\rangle_{q(\theta)} x_k\right)\right\}, \tag{3.30c}$$

where $\widehat{\delta}_i = e^{\langle \ln \delta_i \rangle_{q(\delta)}}$, $\overline{\delta}_i = \langle \delta_i \rangle_{q(\delta)}$, and $\overline{\theta}_i = \langle \theta_i \rangle_{q(\theta)}$. Finally, we can obtain the expression of A_{ki} and B_{ki} as follows,

$$\begin{aligned} A_{ki} &= e^{\langle \ln\{p(y_k|x_k,R_k,I_k=i,\Phi)p(R_k|I_k=i,\nu_i)p(I_k=i|\alpha)c\}\rangle_{q(\Phi)}} \\ &= e^{\ln(\omega_{ki}R_k^{(f_{ki}-1)}) - g_{ki}R_k} \\ &= \omega_{ki}R_k^{(f_{ki}-1)}e^{-g_{ki}R_k}. \end{aligned} \tag{3.31}$$

$$\begin{aligned} B_{ki} &= \int A_{ki} dR_k \\ &= \int \omega_{ki}R_k^{(f_{ki}-1)}e^{-g_{ki}R_k} dR_k \\ &= \omega_{ki}g_{ki}^{-f_{ki}} \int (g_{ki}R_k)^{(f_{ki}-1)}e^{-g_{ki}R_k} d(g_{ki}R_k) \\ &= \omega_{ki}g_{ki}^{-f_{ki}}\Gamma(f_{ki}). \end{aligned} \tag{3.32}$$

The approximated posterior probability of $I_k = i$ can be calculated by the following equation:

$$q(I_k = i) = \frac{B_{ki}}{\sum_{i=1}^{M} B_{ki}} = \frac{\omega_{ki}g_{ki}^{-f_{ki}}\Gamma(f_{ki})}{\sum_{i=1}^{M} \omega_{ki}g_{ki}^{-f_{ki}}\Gamma(f_{ki})} \triangleq q_{ki}, \tag{3.33}$$

and the conditional density $q(R_k|I_k = i)$ has the expression as

$$q(R_k|I_k = i) = \frac{A_{ki}}{B_{ki}} = \frac{R_k^{(f_{ki}-1)}g_{ki}^{f_{ki}}}{\Gamma(f_{ki})}e^{-g_{ki}R_k}. \tag{3.34}$$

Clearly, $R_k|I_k = i \sim \mathcal{G}(f_{ki}, g_{ki})$, and the expected value of R_k is

$$r_{ki} = \langle R_k \rangle_{q(R_k|I_k=i)} = \frac{f_{ki}}{g_{ki}}. \tag{3.35}$$

3.4.3 Variational Bayesian M-step

In the variational Bayesian M-step, we maximize the lower bound $F[q(R, I), q(\Phi)]$ with respect to $q(\Phi)$ and ν by fixing $q(R, I)$. Considering that $q(\Phi) = \prod_{i=1}^{M} q(\theta_i)q(\beta_i)q(\delta_i)$

with the prior density $p(\Phi) = \prod_{i=1}^{M} p(\theta_i|\beta_i)p(\beta_i)p(\delta_i)$, we can maximize the lower bound with respect to each density composing $q(\Phi)$. $F[q(R,I), q(\Phi)]$ can be rewritten as

$$F[q(R,I), q(\Phi)] = \int q(\Phi)\left\langle \ln \frac{p(Y,U,R,I|\Phi,\nu,\alpha)}{q(R,I)} \right\rangle_{q(R,I)} d\Phi + \int q(\Phi) \ln \frac{p(\Phi,\nu|\alpha)}{q(\Phi)} d\Phi, \tag{3.36}$$

and $\langle \ln p(Y,U,R,I|\Phi,\nu,\alpha)\rangle_{q(R,I)}$ can be expressed as follows:

$$\begin{aligned}
&\langle \ln p(Y,U,R,I|\Phi,\nu,\alpha)\rangle_{q(R,I)} \\
=&\langle \sum_{k=1}^{N} \ln \{p(y_k|x_k,R_k,I_k,\Phi)p(R_k|I_k,\nu_{I_k})p(I_k|\alpha)c\}\rangle_{q(R,I)} \\
=&\sum_{k=1}^{N}\sum_{i=1}^{M} q_{ki}\{\frac{1}{2}(\ln \widetilde{r}_{ki} + \ln \delta_i) - \frac{r_{ki}\delta_i}{2}(y_k - x_k^T\theta_i)^2 - \frac{\nu_i}{2}r_{ki} + \frac{\nu_i}{2}\ln\frac{\nu_i}{2} \\
&+(\frac{\nu_i}{2} - 1)\ln \widetilde{r}_{ki} - \ln\Gamma(\frac{\nu_i}{2}) + \ln\alpha_i\} + C,
\end{aligned} \tag{3.37}$$

where $\ln \widetilde{r}_{ki} = \langle \ln R_k \rangle_{q(R_k|I_k=i)}$ and C is a constant.

To Maximize $F[q(R,I), q(\Phi)]$ w.r.t $q(\theta_i)$, we need to calculate the first order functional derivative w.r.t. $q(\theta_i)$, *i.e.*,

$$\frac{\partial\left\{F[q(R,I),q(\Phi)] + \lambda\left(\int q(\theta_i)d\theta_i - 1\right)\right\}}{\partial q(\theta_i)} = 0, \tag{3.38}$$

where λ is the Lagrangian multiplier. By arranging terms relevant to $q(\theta_i)$, we can obtain the expression of $q(\theta_i)$, *i.e.*,

$$\begin{aligned}
q(\theta_i) &\propto p(\theta_i|\overline{\beta}_i)\exp\{\sum_{k=1}^{N} q_{ki}(-\frac{r_{ki}\overline{\delta}_i}{2})(y_k - x_k^T\theta_i)^2\} \\
&\propto \exp\{-\frac{1}{2}\theta_i^T(\overline{\beta}_i I)\theta_i\}\exp\{\sum_{k=1}^{N} q_{ki}(-\frac{r_{ki}\overline{\delta}_i}{2})(y_k - x_k^T\theta_i)^2\} \\
&\propto \exp\{-\frac{1}{2}(\theta_i^T(\overline{\beta}_i I + \sum_{k=1}^{N} q_{ki}r_{ki}\overline{\delta}_i x_k x_k^T)\theta_i + \sum_{k=1}^{N} q_{ki}r_{ki}\overline{\delta}_i(y_k^2 - 2\theta_i^T y_k x_k))\}.
\end{aligned} \tag{3.39}$$

The expression given in the right-hand side of Eq. (3.39) indicates that $q(\theta_i)$ is a Gaussian density function, *i.e.*, $q(\theta_i) = \mathcal{N}(\theta_i|\widetilde{\mu}_i, \widetilde{\Lambda}_i^{-1})$ with

$$\widetilde{\Lambda}_i = \overline{\beta}_i I + \sum_{k=1}^{N} q_{ki}r_{ki}\overline{\delta}_i x_k x_k^T, \tag{3.40a}$$

$$\widetilde{\mu}_i = \widetilde{\Lambda}_i^{-1} \sum_{k=1}^{N} q_{ki} r_{ki} \overline{\delta}_i y_k x_k. \tag{3.40b}$$

Therefore, the expected value of θ_i is $\overline{\theta}_i = \langle \theta_i \rangle_{q(\theta)} = \widetilde{\mu}_i$.

The same procedure is followed to maximize $F[q(I), q(\Phi)]$ with respect to $q(\beta_i)$, and the expression of $\ln q(\beta_i)$ is

$$\begin{aligned}
\ln q(\beta_i) &= \langle \ln p(\theta_i|\beta_i) + \ln p(\beta_i) - \ln q(\theta_i) \rangle_{q(\theta_i)} + C_i \\
&= -\frac{1}{2} \ln \left| \beta_i^{-1} I \right| - \frac{1}{2} \left\langle \theta_i^T (\beta_i I) \theta_i \right\rangle_{q(\theta_i)} + (a_0 - 1) \ln \beta_i - b_0 \beta_i + a_0 \ln b_0 + C_i \\
&= (a_0 + \frac{h}{2} - 1) \ln \beta_i - (b_0 + \frac{1}{2} \left\langle \theta_i^T \theta_i \right\rangle_{q(\theta_i)}) \beta_i + C_i,
\end{aligned} \tag{3.41}$$

where the dimension of the identity matrix I is $h \times h$ and C_i is the constant that is not related to β_i. The expression obtained for $q(\beta_i)$ indicates that the approximated posterior over β_i is a Gamma density function, *i.e.* $q(\beta_i) = \mathcal{G}(\widetilde{a}_i, \widetilde{b}_i)$ with

$$\widetilde{a}_i = a_0 + \frac{h}{2}, \tag{3.42a}$$

$$\widetilde{b}_i = b_0 + \frac{1}{2} \left\langle \theta_i^T \theta_i \right\rangle_{q(\theta_i)} = b_0 + \frac{1}{2} tr(\widetilde{\Lambda}_i^{-1} + \widetilde{\mu}_i \widetilde{\mu}_i^T), \tag{3.42b}$$

from which we get the expected value of β_i by $\overline{\beta}_i = \langle \beta_i \rangle_{q(\beta)} = \widetilde{a}_i / \widetilde{b}_i$.

Similarly, the expression obtained for $q(\delta_i)$ indicates that the approximated posterior over δ_i is also a Gamma density function, *i.e.* $q(\delta_i) = \mathcal{G}(\widetilde{c}_i, \widetilde{d}_i)$ with

$$\widetilde{c}_i = c_0 + \frac{1}{2} \sum_{k=1}^{N} q_{ki}, \tag{3.43a}$$

$$\widetilde{d}_i = d_0 + \frac{1}{2} \sum_{k=1}^{N} q_{ki} r_{ki} \left(y_k^2 - 2 \overline{\theta}_i^T y_k x_k + x_k^T (\widetilde{\Lambda}_i^{-1} + \widetilde{\mu}_i \widetilde{\mu}_i^T) x_k \right). \tag{3.43b}$$

The expected value of δ_i is $\overline{\delta}_i = \langle \delta_i \rangle_{q(\delta)} = \widetilde{c}_i / \widetilde{d}_i$.

In addition, the lower bound $F[q(I), q(\Phi)]$ is maximized with respect to ν. The prior distribution over ν is $p(\nu|\alpha) = \prod_{i=1}^{M} p(\nu_i)$. Consider that the prior distribution over ν_i is an exponential distribution [54], *i.e.*, $p(\nu_i) = e_0 exp(-e_0 \nu_i)$. The posterior density over ν_i, denoted as $q(\nu_i)$, is proportional to the production of the prior and the likelihood over ν_i, *i.e.*,

$$\ln q(\nu_i) = \ln p(\nu_i) + \sum_{k=1}^{N} q_{ki} \left\{ \frac{\nu_i}{2} \ln \frac{\nu_i}{2} - \ln \Gamma(\frac{\nu_i}{2}) + (\frac{\nu_i}{2} - 1) \ln \widetilde{r}_{ki} - \frac{\nu_i}{2} r_{ki} \right\} + C_{v_i}$$

$$= -(e_0 + \frac{1}{2}\sum_{k=1}^{N} q_{ki}(\ln \widetilde{r}_{ki} - r_{ki}))\nu_i + (\frac{\nu_i}{2}\ln\frac{\nu_i}{2} - \ln\Gamma(\frac{\nu_i}{2}))\sum_{k=1}^{N} q_{ki} + C_{\nu_i}, \tag{3.44}$$

where C_{ν_i} is the constant that is not relevant to ν_i. According to Eq. (3.36) and Eq. (3.37), maximizing the lower bound w.r.t ν_i is equivalent to maximizing $\ln q(\nu_i)$. Hence, we can resort to MAP approach through numerical search to update the value of ν_i:

$$\nu_i = \arg\max_{\nu_i > 0} \{\ln q(\nu_i)\}. \tag{3.45}$$

3.4.4 Significance coefficients optimization

Given the significance coefficients α, the VBE-step and VBM-step are iteratively computed until convergence of parameters. Upon the convergence, we will obtain a lower bound $F[q(I), q(\Phi)]$ that is approaching the marginal log-likelihood $\ln p(Y, U|\alpha)$. The optimal values of significance coefficients can be updated by maximizing this lower bound with respect to α subject to $\sum_{i=1}^{M} \alpha_i = 1$ [52]. By solving the following equation

$$\frac{\partial\left\{F[q(I), q(\Phi)] + \lambda(\sum_{i=1}^{M}\alpha_i - 1)\right\}}{\partial\alpha_i} = 0, \tag{3.46}$$

the update equation for α_i is obtained as

$$\alpha_i = \frac{\sum_{k=1}^{N} q_{ki}}{N}. \tag{3.47}$$

During the optimization procedure, the significance coefficients of redundant local-models will converge to zero quickly. This would provide an automated mechanism to eliminate the insignificant local-models and determine the number of local-models from the identification process.

3.4.5 Lower bound evaluation

The lower bound is evaluated after each iteration of the update equations as it is an indicator of convergence. Having updated the estimates of $q(I)$, $q(\Phi)$, ν and α, the

value of lower bound can be calculated as follows:

$$F[q(R,I),q(\Phi)] = \sum_I \int q(R,I)q(\Phi)\ln\frac{p(Y,U,R,I|\Phi,\nu,\alpha)}{q(R,I)}dRd\Phi + \int q(\Phi)\ln\frac{p(\Phi,\nu|\alpha)}{q(\Phi)}d\Phi. \tag{3.48}$$

However, calculating multiple integration is computationally expansive. As pointed out by Takekawa *et al.* [54], the integration can be avoided through mathematical operation. The first term of Eq. (3.48) can be transformed into

$$\sum_{k=1}^{N}\ln\sum_{i=1}^{M}B_{ki}. \tag{3.49}$$

The derivation is given in Appendix. The second term of Eq. (3.48) is equivalent to

$$-KL\left[q(\Phi)||p(\Phi|\alpha)\right] + \ln p(\nu|\alpha). \tag{3.50}$$

Therefore, the lower bound can be evaluated by

$$F[q(R,I),q(\Phi)] = \sum_{k=1}^{N}\ln\sum_{i=1}^{M}B_{ki} - KL\left[q(\Phi)||p(\Phi|\alpha)\right] + \ln p(\nu|\alpha). \tag{3.51}$$

3.4.6 Outlier detection through predictive density

In the Bayesian learning, one goal is to perform density estimation with respect to the identified model [55]. Inspired by the method of density estimation proposed in [55], we propose an outlier detection approach based on the identified model. The density of a new output y_{n+1} given the training data and current input u_n is

$$\begin{aligned} p(y_{n+1}|Y,U,u_n) &= \sum_I \int p(y_{n+1},R,I,\Phi|Y,U,u_n)dRd\Phi \\ &= \sum_I \int p(y_{n+1}|Y,U,u_n,R,I,\Phi)p(R,I,\Phi|Y,U,u_n)dRd\Phi. \end{aligned} \tag{3.52}$$

The likelihood $p(y_{n+1}|Y,U,u_n,R,I,\Phi)$ can be simplified as $p(y_{n+1}|x_{n+1},R_{n+1},I_{n+1},\Phi)$. Under the variational Bayesian framework, the true posterior density is approximated by the variational posterior, namely,

$$p(R,I,\Phi|Y,U,u_n) \approx q(R,I,\Phi) = q(R,I)q(\Phi). \tag{3.53}$$

Therefore, the predictive density can be approximated by

$$p(y_{n+1}|Y,U,u_n) \approx \sum_I \int p(y_{n+1}|x_{n+1},R_{n+1},I_{n+1},\Phi)q(R,I)q(\Phi)dRd\Phi. \tag{3.54}$$

The log predictive density can be derived as

$$\begin{aligned}\ln p(y_{n+1}|Y,U,u_n) \approx & \ln \sum_I \int p(y_{n+1}|x_{n+1},R_{n+1},I_{n+1},\Phi)q(R,I)q(\Phi)dRd\Phi \\ \geq & \sum_I \int q(R,I)q(\Phi)\ln p(y_{n+1}|x_{n+1},R_{n+1},I_{n+1},\Phi)dRd\Phi \\ = & \sum_{i=1}^{m} q_{n+1,i}\{\frac{1}{2}(\ln(\frac{\widehat{\delta}_i}{2\pi}) + \ln \widetilde{r}_{n+1,i}) \\ & -\frac{r_{n+1,i}\overline{\delta}_i}{2}(y_{n+1}^2 - 2y_{n+1}x_{n+1}^T\overline{\theta}_i + x_{n+1}^T\langle\theta_i\theta_i^T\rangle_{q(\theta)}x_{n+1})\} \end{aligned} \tag{3.55}$$

Based on the predictive density, outliers in the coming data can be detected. Normal data points have a large value of the approximated density while an outlying data point has a much smaller value.

3.4.7 Algorithm summary

We summarize the execution procedure of the proposed approach in Fig. 3.1. Given a set of identification data, the posterior density of the model identity and the noise scale at each sampling instant is calculated through q_{ki} and r_{ki} in the VB E-step. The value of r_{ik} will be vary small for outliers, which automatically leads to significant down-weighting of outlying data. In the VB M-step, the posterior density of each parameter is updated until convergence, from which we can get the expected value of parameters. The number of local-models is obtained by optimizing the significance coefficients, where redundant local-models will be eliminated. Based on the identified model, outlier detection can be performed resorting to the approximate predictive density, which is not shown in Fig. 3.1.

3.5 Simulation examples

3.5.1 A numerical example

Consider a simulated switched ARX process with m local-models described as follows:

$$y_k = x_k^T\theta_i + \epsilon_i(k),\ k \in \{1,2,\cdots,N\},\ i \in \{1,2,\cdots,m\}, \tag{3.56}$$

where the regression vector is

$$x_k = [y_{k-1}, y_{k-2}, u_{k-1}, u_{k-2}, 1]^T. \tag{3.57}$$

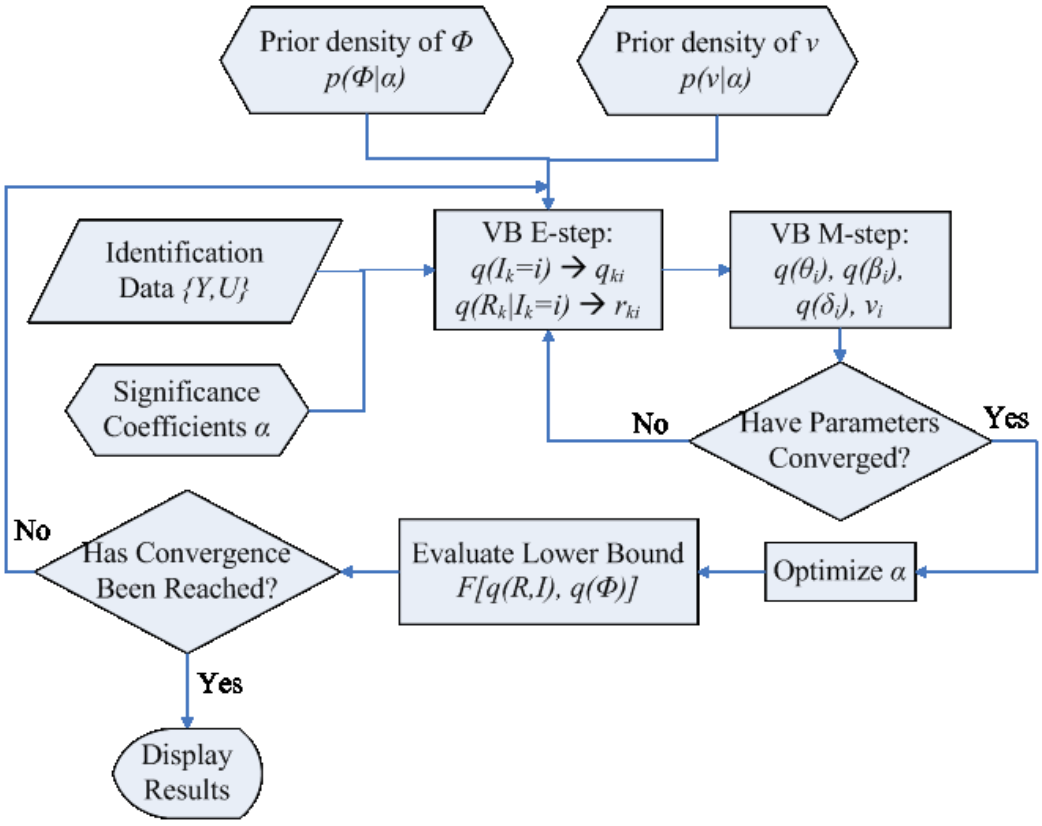

Figure 3.1: Flowchart of the robust identification of SARX models by the proposed approach.

To generate simulated data, the number of identification data points is set to be $N = 600$ and the true number of local-models is $m = 3$ which needs to be inferred from the data. The added disturbing noise is denoted by $\epsilon_i(k)$. A sequential number generator is designed to randomly generate a sequence of operating points for the underlying process. The operating point changes randomly every 100 sampling instants. The true parameters of local-models at three operating points are

$$\theta_1 = [1.21 \ -0.49 \ -0.30 \ 0.90 \ 0.5]^T,$$
$$\theta_2 = [1.39 \ -0.50 \ 0.20 \ -0.45 \ 0]^T,$$
$$\theta_3 = [-1.20 \ -0.72 \ 0.60 \ -0.70 \ 2.00]^T.$$

Two kinds of input signal are designed to persistently excite the simulated process, *i.e.*, uniformly distributed white noise over the interval $[-1, \ 1]$ and a random binary sequence (RBS) with level $[-1, \ 1]$. The choice of input signal depends on the comparison methods. If clustering based methods (comparing methods) fail to cluster the collected data when the input signal is a RBS, then uniformly distributed input signal will be adopted for the sake of comparison. Output data sets are generated by the simulated process. In the case that ϵ_i is Normal distributed noise with mean 0 and variance 0.025, a snapshot of the random switching operating point as well as the simulated input-output data is shown in Fig. 3.2. To simulate the situation when the noise is contaminated with outliers, we randomly replace part of the noise by uniformly distributed outliers over the range $[-3.5, -3]$. Since the number of local-models ($\hat{m}$) is unknown, in addition to parameter estimation, we need to infer $\hat{m}$ from the identification data corrupted by outliers. When the identification data set is of good quality, the method proposed by Nakata *et al.* [6] (denoted as *Nakata*) is able to determine $\hat{m}$ by selecting the model that has the smallest value of information criteria. The consistent Akaike's information criterion (CAIC) is adopted as the information criterion in comparison. Classical clustering methods such as the *K-means* algorithm can also estimate the number of local-models (clusters) by combining indexes that will optimize the clustering structure. The data vector at k-th sampling instant is composed of y_k and x_k (the term related to non-zero mean in x_k is not included). Here a widely used clustering method, *i.e.*, *K-means* plus Davies-Bouldin index which is denoted by *K-DB*, is considered. Implementation of these conven-

tional methods involves three distinct steps: 1. Clustering the identification data when given the number of clusters, 2. Determining the number of clusters through information criteria or DB index, 3. Estimating parameters of local-models based on clustering results. On the contrary, the proposed approach (denoted by *Robust*) is able to determine the number of local-models as well as estimate the model parameters simultaneously. Moreover, the proposed approach takes noise and outliers into account while conventional methods ignore this practical issue. In addition, the regular approach under the VB framework (denoted by *Regular*), where the noise ϵ_i is modelled as a Normal distribution, is compared as well.

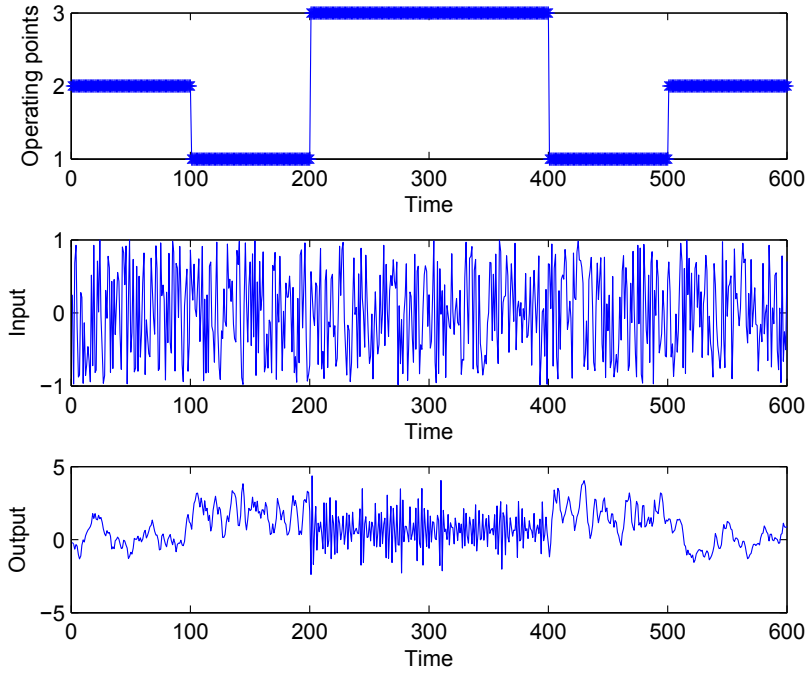

Figure 3.2: An example of generated data by the simulated process.

A. Determine the number of local-models

The number of local-models can be estimated by aforementioned methods. Their performances are accessed under two scenarios:

- Case I: the noise follows Normal distribution;
- Case II: the noise is contaminated with 5% outliers.

In each case, a Monte Carlo simulation (50 runs) is performed to estimate $\hat{m}$. The upper bound of $\hat{m}$ is set to be $M = 5$ for all investigated methods. Fig. 3.3 and Fig. 3.4 show the results of the estimated number of local-models by compared methods in Case I and Case II, respectively. From Fig. 3.3, we can see that the performance of *Nakata*, *Regular*, and *Robust* are comparable when the identification data have a good quality. It occurs 47 times (*Nakata*), 50 times (*Regular*), and 50 times (*Robust*) respectively when the estimated number of local-models equals to the true number which is $m = 3$. The performance of *K-DB* is poor in this simulation. It might be caused by the non-spherical clusters, since the *K-DB* performs well only for spherical clusters. In Case II where the data are contaminated with outliers, however, the capabilities of investigated methods deteriorate significantly except for the proposed robust approach. The *Nakata* and *Regular* tend to give more local-models under the influence of outliers, and the *K-DB* tends to use fewer local-models, while the *Robust* is capable of resisting outliers and estimating $\hat{m}$ correctly.

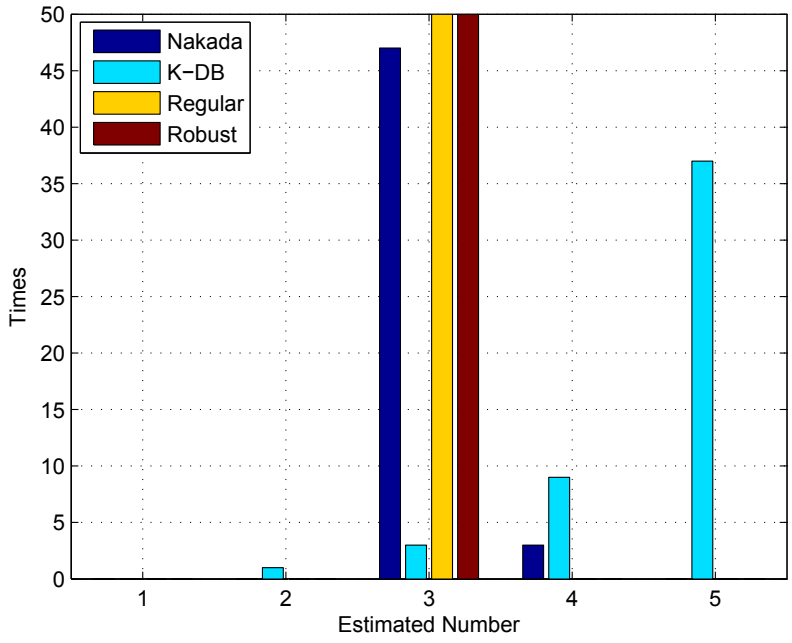

Figure 3.3: Bar chart of the estimated number of local-models in Case I.

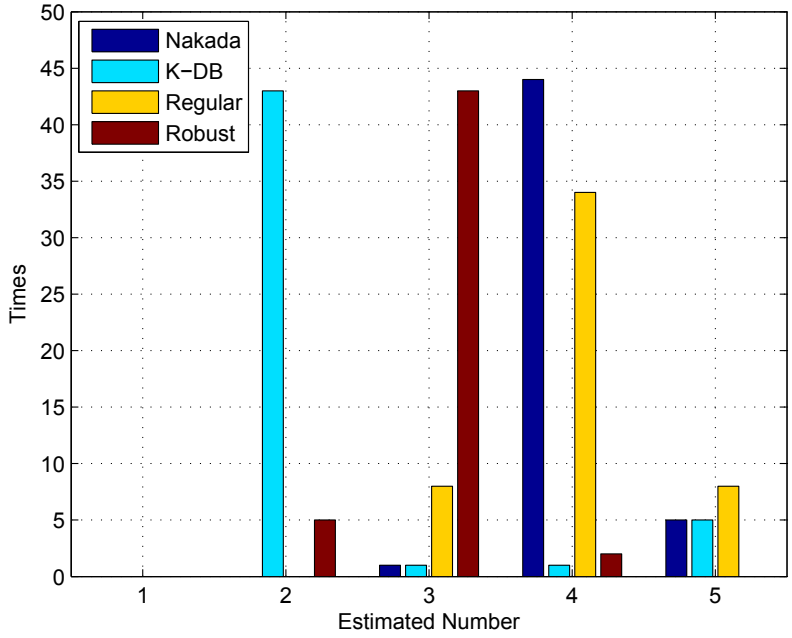

Figure 3.4: Bar chart of the estimated number of local-models in Case II..

B. Estimate parameters of local-models

The performance of these methods in parameter estimation is compared as well. The *K-DB* is not considered due to the poor performance of dealing with non-spherical clusters. The aforementioned two cases are investigated again. The number of local-models is assumed to be known in order to have a fair comparison of their performances in parameter estimation. Monte Carlo simulations with 50 runs each are constructed. In Case I, the identification data sets have a good quality, and results from these methods are fairly satisfactory, as presented in Fig. 3.5. For each local-model, the coefficient vector, denoted by θ_i, $i = 1, 2, 3$, consists of five elements, namely $[P_1\ P_2\ P_3\ P_4\ P_5]$. The mean values of estimation from these methods are close to the true value, and the standard deviations are small. However, as Fig. 3.6 shows, when the data sets are contaminated with 5% outliers, the *Nakata* and the *Regular* fail to estimate the parameters accurately and reliably, while the performance of the proposed approach is consistently satisfactory.

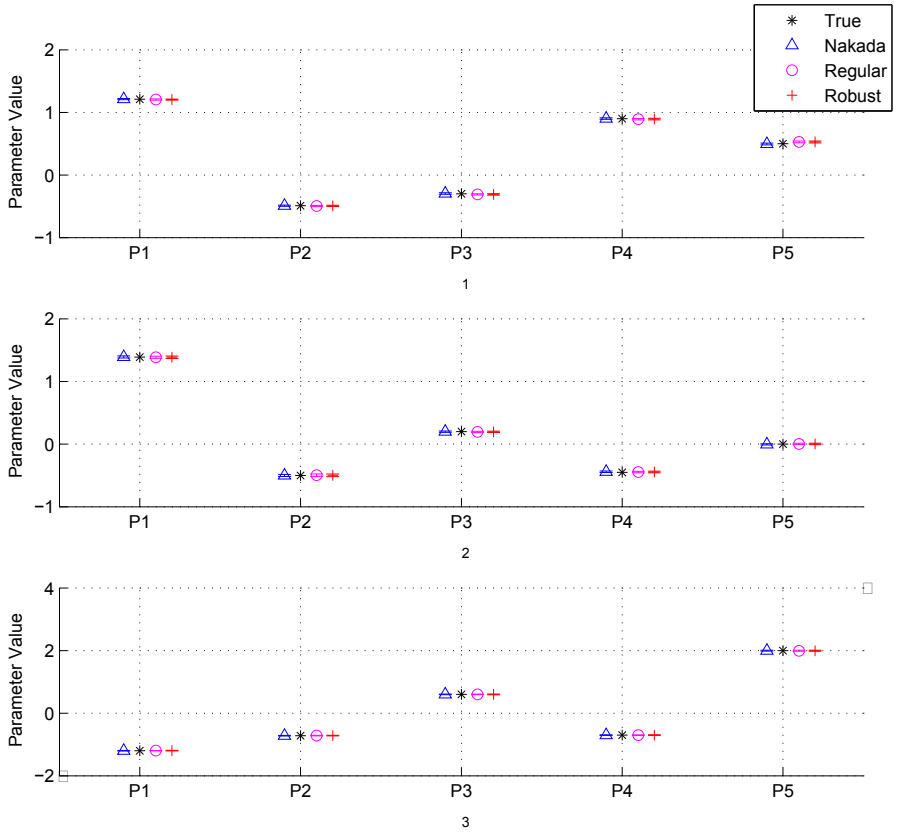

Figure 3.5: Case I: comparison of the mean and standard deviation given by three methods.

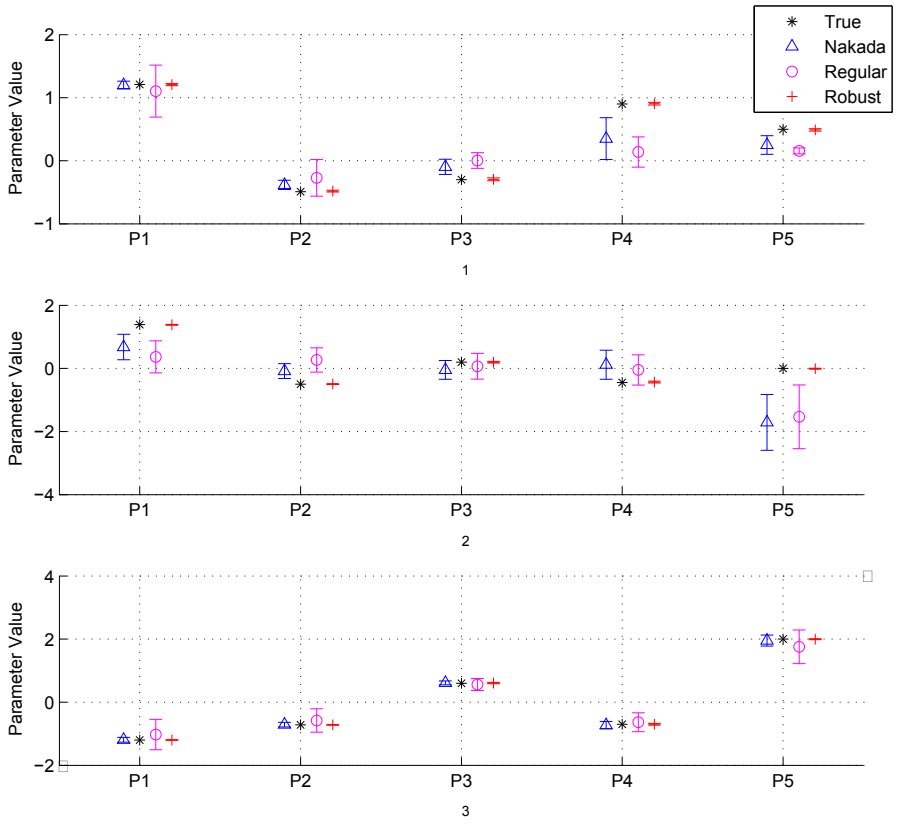

Figure 3.6: Case II: comparison of the mean and standard deviation given by three methods.

C. Detect outliers by the proposed approach

As aforementioned, the proposed approach is able to detect outliers from the dynamic process through the predictive density. Given a set of testing data, the logarithmic prediction density function can be calculated approximately by Eq. (3.55). Since the statistics of the predictive density is not fully revealed, a simple but effective rule is adopted. As Fig. 3.7 presents, the approximate predictive density (ln) values of majority data points are larger than -1.5 (correspondingly, the density value is $e^{(-1.5)} \approx 0.223$), while a number of data points are located far below this value. Thus in this example -1.5 is taken as the threshold to determine whether the coming data point is an outlier or not. Fig. 3.8 shows the percentage of outliers detected in the individual runs of Monte Carlo simulation. According to the figure, The accuracy of outlier detection is around 88%. The simple rule based on the predictive density is capable of separating outliers from the testing data.

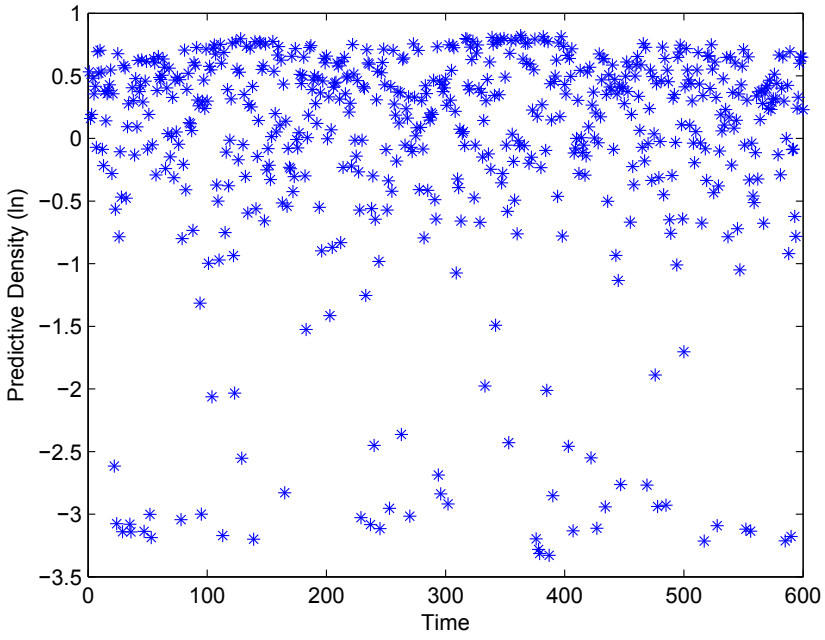

Figure 3.7: Approximate predictive density of the testing data in an individual run of Monte Carlo simulation.

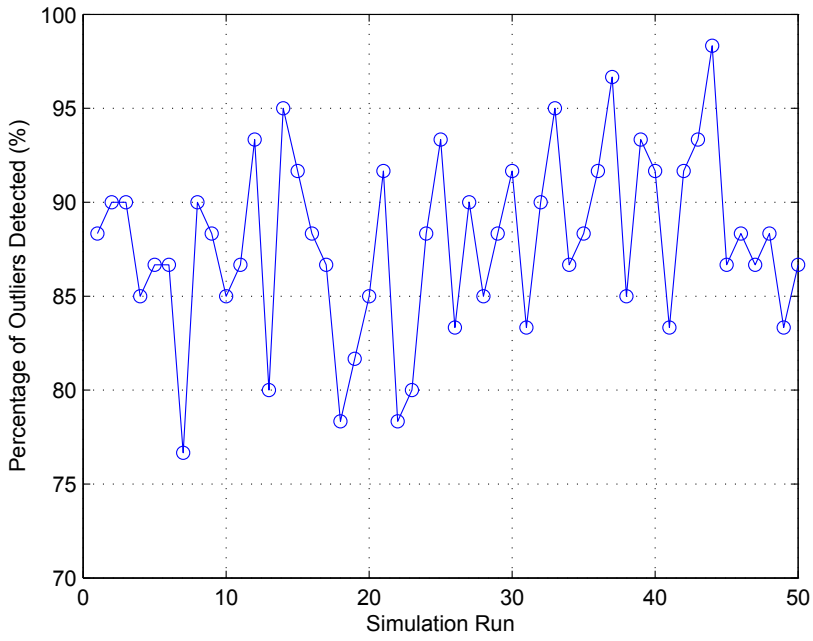

Figure 3.8: Percentage of outliers detected in the individual runs of Monte Carlo simulation.

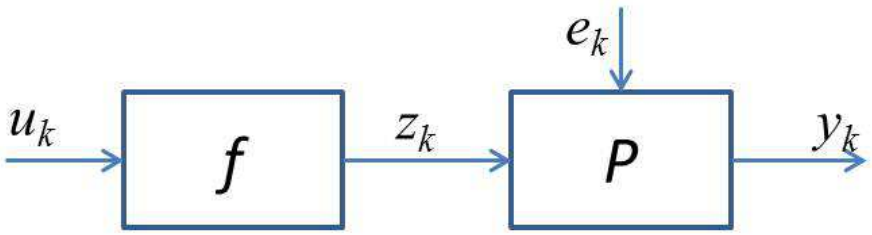

Figure 3.9: A Hammerstein model with a saturation non-linearity.

3.5.2 Application to a Hammerstein model

In non-linear process identification, a Hammerstein model shown in Fig. 3.9 is often used. It consists of a linear time-invariant (LTI) process and a static non-linear component. We consider the Hammerstein model as is investigated in [6]. It is a SISO Hammerstein model with a saturation function. The LTI plant is given by

$$y_k = -a_1 y_{k-1} - a_2 y_{k-2} + b_1 z_{k-1} + e_k, \tag{3.58}$$

where a_1, a_2, and b_1 are scalar constants, and f is a saturation function:

$$z_k = f(u_k) = \begin{cases} u_{\max}, & if \quad u_k > u_{\max} \\ u_k, & if \quad u_k \in [u_{\min}, u_{\max}] \\ u_{\min}, & if \quad u_k < u_{\min} \end{cases} \tag{3.59}$$

and u_{max} and u_{min} are scalar constants giving the upper and lower bounds of saturation, respectively.

Same as [6], when generating data sets through the Hammerstein model, we fix $a_1 = 0.5$, $a_2 = 0.1$, $b_1 = 1$, $u_{max} = 2$, and $u_{min} = -1$. The input u_k is normally distributed with mean 0 and variance 4. The noise e_k is contaminated with outliers. 95% of the noise is normally distributed with means 0 and variance 0.04, and 5% of the noise is randomly replaced by outliers which are uniformly distributed over the range $[-6, \ -4]$. The sample of data is $N = 900$.

The proposed robust approach is applied to identification of the non-linear model. Suppose we do not have any prior information about the non-linearity. The upper bound of local-model numbers is set to $M = 8$. A second order ARX model (Eq. 3.60) is adopted to approximate the local-model, so $n_a = 2$ and $n_b = 2$.

$$y_k = c_1 y_{k-1} + c_2 y_{k-2} + d_1 u_{k-1} + d_2 u_{k-2} + g + \epsilon_k, \tag{3.60}$$

where g is the mean of the local-model.

A Monte Carlo simulation with 60 trials of different noise sequences is constructed to verify the performance of the proposed approach. As Fig. 3.10 shows, in 56 individual runs of the simulation, the number of local-models is estimated to be $\hat{m} = 3$, which indicates that using three linear local-models is capable of approximating the non-linear model, although the given upper bound is $M = 8$ at the beginning of identification. Only 4 out of 60 trials, the proposed approach fails to correctly infer

m from noisy data with outliers. Based on the 56 times of correct estimation, the mean and one standard deviation of estimated parameters corresponding to three local-models are presented in Table 3.1. From Table 3.1, it is easy to figure out the parameters of the Hammerstein model, *i.e.*,

$$a_1 = -c_1 \approx 0.5,\ a_2 = -c_2 \approx 0.1,\ b_1 = d_1(Second) \approx 1,$$
$$u_{max} = \frac{g(First)}{b_1} \approx 2,\ u_{min} = \frac{g(Third)}{b_1} \approx -1. \tag{3.61}$$

The proposed approach performs fairly well in identification of the Hammerstein model, where the mean of estimation is close to the true value and the standard deviation is small. Given testing data sets, the capability of outlier detection is evaluated as well. In each individual run of the simulation, a testing data set which is contaminated with 10% outliers is assessed based on the identified model. The predictive density approach can distinguish between outliers and normal data as long as the number of local-models is estimated correctly, as Fig. 3.11 demonstrates.

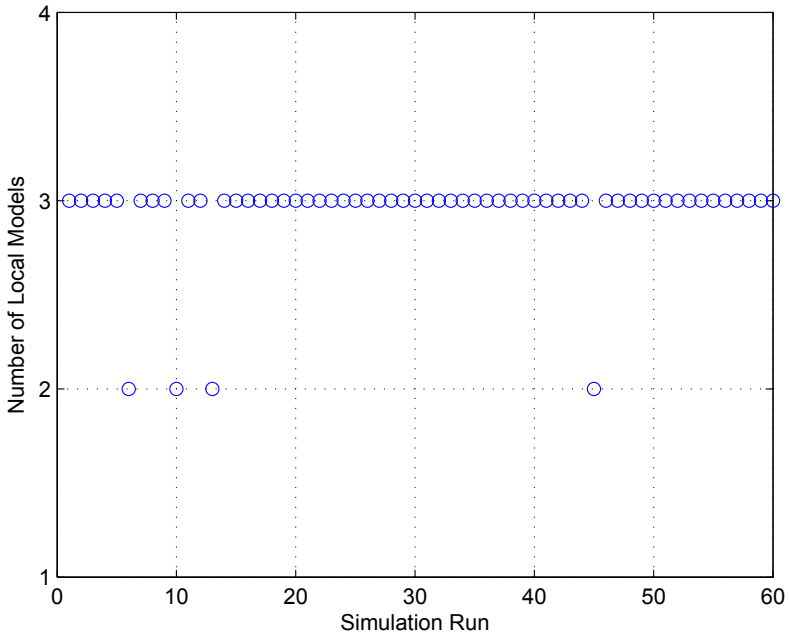

Figure 3.10: Estimated number of local-models in the simulation (60 runs).

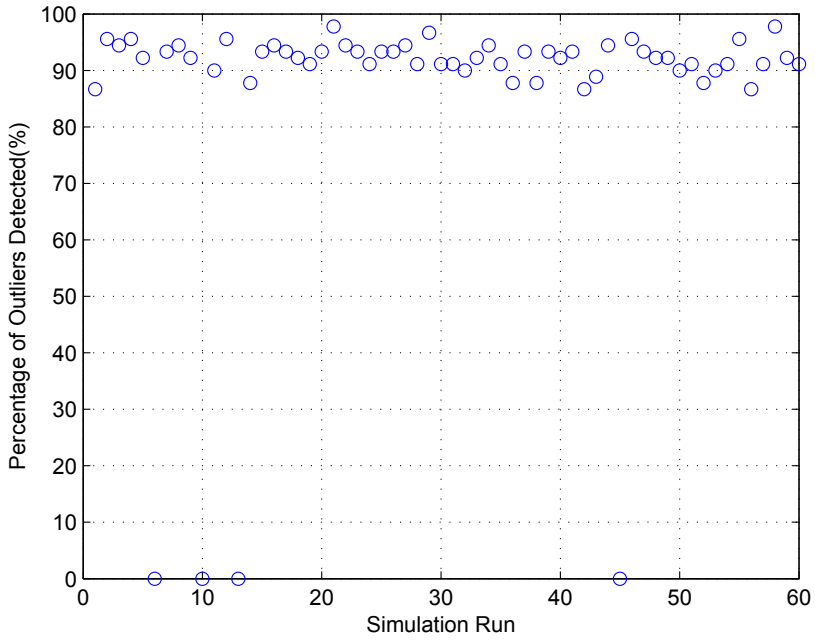

Figure 3.11: Percentage of outliers detected in the simulation (60 runs).

Table 3.1: Parameters of three local-models

	First	Second	Third
c_1	-0.4969($\pm$0.0149)	-0.5024($\pm$0.0072)	-0.4985($\pm$0.0127)
c_2	-0.1000($\pm$0.0140)	-0.0980($\pm$0.0073)	-0.0976($\pm$0.0088)
d_1	-0.0060($\pm$0.0119)	1.0004($\pm$0.0064)	-0.0002($\pm$0.0103)
d_2	-0.0028($\pm$0.0113)	-0.0008($\pm$0.0075)	-0.0004($\pm$0.0108)
g	1.9950($\pm$0.0244)	-0.0043($\pm$0.0127)	-1.0059($\pm$0.0147)

3.6 Conclusions

In this chapter, a variational Bayesian approach to identification of switched ARX models was developed. Practical issues including robustness, estimation of the number of local-models, estimation of parameter uncertainty, and outlier detection were addressed. By embedding the t distribution into the framework of the variational Bayesian approach, the outliers were down-weighted in estimation, leading to the robustness of the proposed approach. Meanwhile, a set of significance coefficients was introduced to indicate the significance of each local-model, and during the optimization, redundant or insignificant local-models were eliminated. In addition, instead of simply partitioning the data through clustering techniques, the model identity of each data point was inferred from the data during the identification process. The posterior probability density functions of parameters were estimated under the variational Bayesian framework, so the uncertainty of parameters was investigated as well. Based on the identified model, an outlier detection method through the approximate predictive density function was proposed. The effectiveness of the proposed approach for robust identification of SARX models as well as the performance of outlier detection were verified through simulated examples. The proposed approach not only resists the adverse influence of outliers, but also infers the number of local-models automatically, which makes this algorithm more applicable to the real world problems.

Appendix

The expression (3.49) can be derived as

$$\begin{aligned}
&\sum_{I} \int q(R, I) q(\Phi) \ln \frac{p(Y, U, R, I|\Phi, \nu, \alpha)}{q(R, I)} dR d\Phi \\
=&\sum_{k=1}^{N} \sum_{i=1}^{M} \int q(R_k, I_k = i) \left\{ \ln A_{ki} - \ln q(R_k, I_k = i) \right\} dR_k
\end{aligned} \tag{3.62}$$

according to Eq. (3.24). By substituting Eq. (3.25) into the above equation, we can further simplify its expression:

$$
\begin{aligned}
&\sum_{k=1}^{N}\sum_{i=1}^{M}\int q(R_k, I_k = i)\left\{\ln A_{ki} - \ln\left(A_{ki}\Big/\sum_{i=1}^{M} B_{ki}\right)\right\}dR_k \\
&= \sum_{k=1}^{N}\sum_{i=1}^{M}\int q(R_k, I_k = i)\left\{\ln \sum_{i=1}^{M} B_{ki}\right\}dR_k \\
&= \sum_{k=1}^{N}\ln\sum_{i=1}^{M} B_{ki}\sum_{i=1}^{M}\int q(R_k, I_k = i)dR_k \\
&= \sum_{k=1}^{N}\ln\sum_{i=1}^{M} B_{ki}
\end{aligned}
\tag{3.63}
$$

Chapter 4

Design of Adaptive Steam-quality Soft Sensors for Once-through Steam Generators

Advanced process control practice usually requires a reliable process model. Industrial processes are complicated, and operational data are noisy with potential outliers. These practical issues should be considered when modelling industrial processes. Once-through steam generators are commonly used in oil sands industry to provide steam for the steam assisted operations. A reliable estimation of steam-quality is a key requirement for investigating the performance of steam generators. Therefore, adaptive soft sensors for steam-quality measurement are designed to meet the requirement of industrial operations in this chapter.

4.1 Introduction

Alberta holds the world's largest reserves of bitumen which has the same order of magnitude as reserves of conventional oil in Saudi Arabia. Up to 80% of the estimated reserves could be recovered by in-situ thermal operations [56]. Conventional in-situ technologies such as steam flooding and cyclic steam stimulation (CSS) [57] have been successfully applied in Venezuela and California. New in-situ production technologies such as the steam assisted gravity drainage (SAGD) [58] and expanding solvent-SAGD (ES-SAGD) [59] are becoming dominant technologies employed for the recovery of heavy oil and oil sands in Canada. The new in-situ technologies have increased produced oil rates and reduced production costs [56].

However, whether the conventional or the new in-situ technologies, they all demand a great amount of steam in operation. In Alberta, the SAGD and its variants are widely applied in the oil sands industry. The steam generator thus plays an important role in the recovery of bitumen and heavy oil. Fig. 4.1 shows the water and steam circuit for the SAGD process. Common types of steam generators used for oil sands recovery are once-through steam generators (OTSGs) and drum boilers. Although the capital costs of drum boilers are significantly less than OTSGs and the need for vapor-liquid separators is eliminated for drum boilers, OTSGs are favoured over the drum boilers in most oil sands fields so far [1]. There are several reasons for this. OTSGs have lower heat flux than drum boilers which makes OTSGs more tolerant to the overheating caused by the scale deposition on the tubes [60]. Also, OTSGs require less maintenance as they do not have level controls and low level cuts required in drum boilers [61].

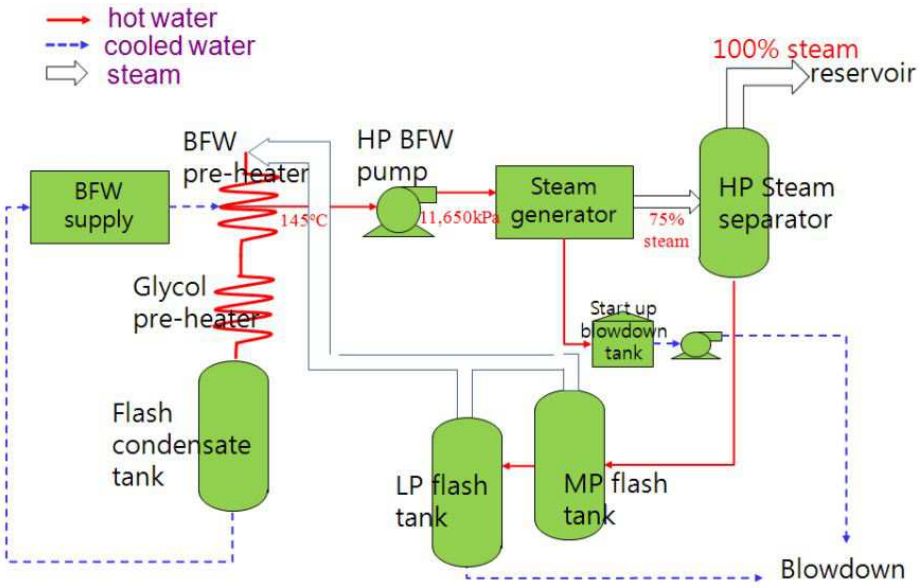

Figure 4.1: Water and steam circuit for the SAGD process [1].

The operational performance of the OTSGs can be evaluated by the steam-quality, namely the mass fraction of steam in a saturated steam/water mixture [62]. OTSGs generally should produce 75-80% quality steam. Over-high quality steam or insufficient liquid water may cause deposition of tube fouling and solids which will decrease heat transfer and increase the tube temperature. Though the OTSG has certain tolerance to the overheating, it can be damaged if the limit is exceeded. Certainly, low

quality steam means low efficiency of the OTSG, which is not economical. Therefore, the steam-quality should be monitored and maintained within a tight range in order to keep a good performance of OTSGs. Online measurement of the average steam-quality can be conducted by calculating the difference between the inlet boiler feed water (BFW) flow rate and the condensate blowdown flow rate. The online measurement is not accurate enough, so laboratory analysis of samples is usually conducted every few hours in order to more precisely monitor the steam-quality. The low accuracy of online measurement and infrequent laboratory analysis cannot meet the real-time monitoring and closed-loop control of steam-quality. To solve this problem, Xie *et. al* [62] developed soft sensors for online steam-quality measurements of OTSGs. The soft sensors have been successfully implemented online to predict steam qualities in an industrial OTSG. However, some generalization issues remain for the soft sensor to be applicable in various steam generators used across the oil sands industry.

Soft sensors, also called inferential models, can provide frequent online estimates of quality variables on the basis of the correlation with available process measurements [10]. The development of the soft sensing technology has benefited the bioprocess industry [63], chemical industry [64], steel industry [65], and oil sands industry [10]. The design of soft sensors is rooted in the process modelling. Soft sensors usually work well only for a particular operating region where the underlying identified model approximates the true process well. Therefore, the identification of the process plays a key role in the design of effective soft sensors.

Industrial processes usually exhibit certain forms of time-varying property. Also, production policies might drive a chemical plant to switch among various operating conditions, which results in multiple modes or regimes of behaviour. For example, the varying of the BFW flow rate of the OTSG can influence the steam production and steam-quality, so one single model may not be able to capture the dynamic behaviour of the process. Hence, the estimation of steam-quality given by soft sensors based on a single model is not always reliable. In such applications, multi-model soft sensors show the advantage so that both continuous dynamic behaviour and discontinuous dynamics can be described [66, 10]. Multi-model process identification has been widely studied. Quite a few approaches have been proposed to solve the

identification problem, such as expectation-maximization based method [8], clustering based method [44] and recursive identification method [46]. Despite the existence of various methods, robustness issue has not been well studied.

The t distribution is well known for robustness in statistical modelling and Bayesian inference. In literature, the t distribution has been widely utilized to deal with outlying data points [67, 68, 14]. However, this advanced technique has not been employed for the multi-model soft sensor development. The main contribution of this chapter is to introduce a statistical approach to the development of multi-model soft sensors for online measurement of steam-quality, where t distributions are integrated with the variational Bayesian (VB) [48, 52] framework in order to deal with potential outliers. First-principle models are presented first. Unknown parameters of the model are estimated by the VB approach in which the influence of outliers is taken into account. Considering the error between the first-principle model and the real process model, an adaptive bias correction term is used. The effectiveness of the adaptive multi-model soft sensors is demonstrated through prediction of steam qualities for industrial scale once-through steam generators.

4.2 Process description and models

Fig. 4.2 shows the simplified schematic diagram of an industrial steam generator (OTSG). The BFW is divided into 8 individual passes, and Passes 1 to 4 are fed into the upper deck and passes 5 to 8 are fed into the lower deck where the water is heated. The saturated steam/water mixtures flow out of the OTSG and merge into one stream, and then the stream flows into high pressure separator for the separation of dry steam from liquid water. The process variables of interest are listed in Table 4.1.

Differential pressure meters are installed at the outlet of each individual pass for steam-quality estimation. Individual-pass steam-quality (I-SQ) is continuously calculated using a combination of steam discharge pressure, pass feed water flow and the differential pressure measurement. The calculation can be adjusted through a tuning factor when the online measurement has a significant deviation from lab sample, but the exact time and magnitude of tuning are difficult to know exactly. A

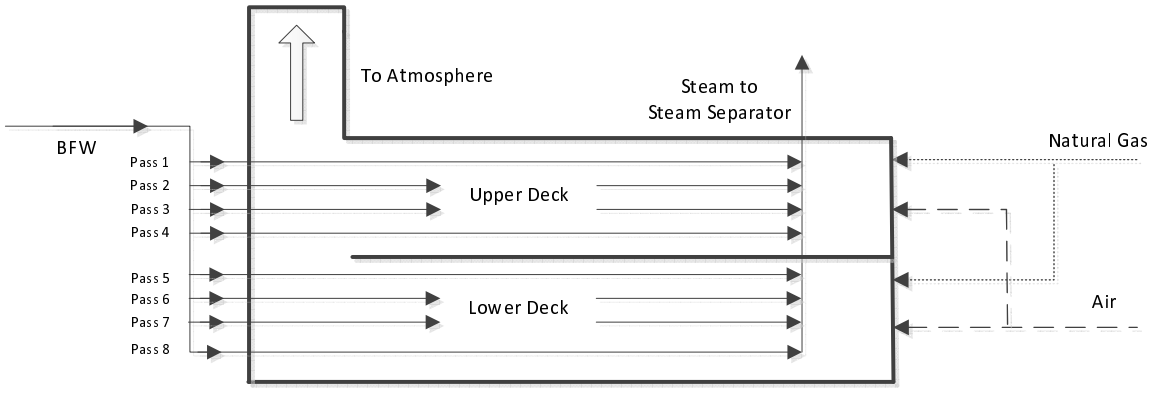

Figure 4.2: Schematic diagram of an OTSG.

Table 4.1: List of process variables

Variable	Description	Unit
Y^l	Lab analysis of I-SQ l ($l = 1, 2, \cdots, 8$)	%
X^l	Online measurement of I-SQ l ($l = 1, 2, \cdots, 8$)	%
X	Online measurement of O-SQ	%
F_f	BFW Flow rate	m^3/hr
F^l	Flow rate of Pass l	m^3/hr
T_f	Inlet temperature of BFW	°C
T^l	Outlet temperature of Pass l	°C
T_r	Temperature after Passes recombine	°C

mass balance based overall steam-quality (O-SQ) is available as well. The calculation is based on BFW flow rate, steam blowdown flow rate, and steam separator level change, but the accuracy is not guaranteed.

Although the structure of various OTSGs may be different, the heat transfer procedure of the OTSG resembles each other. The first-principle model derived by Xie *et. al* [62] is adopted as the model of the soft sensor. Hence, the model for estimating I-SQ is

$$Z^l(t) = \frac{\xi^l \rho F_f(t) \Delta H X(t) + 100\xi^l \rho F_f(t) C_{p1}[T_r(t) - T_f(t)] - 100\rho F^l(t) C_{p2}[T^l(t) - T_f(t)]}{\rho F^l(t) \Delta H}, \tag{4.1}$$

where ξ^l is the fraction of heat absorbed by l-th individual pass; ρ is the density of water; C_{p1} and C_{p2} are the heat capacity and ΔH is the enthalpy of vaporization. $X(t)$ denotes the online measurement of O-SQ. By defining $k_1 = \xi^l$, $k_2 = 100\xi^l C_{p1}/\Delta H$ and $k_3 = -100C_{p2}/\Delta H$, we can simplify the model as follows:

$$Z^l(t) = k_1 u_1(t) + k_2 u_2(t) + k_3 u_3(t), \tag{4.2}$$

where k_1, k_2 and k_3 are unknown parameters, and $u_1(t)$, $u_2(t)$ and $u_3(t)$ are treated as inputs, which can be calculated by

$$u_1(t) = \frac{F_f(t)}{F^l(t)} X(t), \tag{4.3a}$$

$$u_2(t) = \frac{F_f(t)}{F^l(t)} [T_r(t) - T_f(t)], \tag{4.3b}$$

$$u_3(t) = T^l(t) - T_f(t). \tag{4.3c}$$

The model of estimating O-SQ is the scaled online measurement of O-SQ in order to reduce the scaling error of the online measurement. That is

$$Z_O(t) = kX(t). \tag{4.4}$$

The design of multi-model soft sensors as well as single-model soft sensors for measuring I-SQ and O-SQ is based on these models. Model parameters will be estimated in the stage of soft sensor design.

4.3 Design of adaptive soft sensors

According to the structure of models, we can simply use linear regression methods to estimate model parameters, given the input data and lab samples of steam-quality (output data). However, the process may vary with time, resulting in variation of model parameters. Thus, a single linear model is not capable of predicting the steam-quality well over entire operating range. On the basis of conventional single-model based soft sensors, multi-model soft sensors are therefore designed. Furthermore, since the first-principle model is not precise owing to the omission of some factors such as heat loss, an online bias updating term, as expressed by Eq. (4.5) [69, 70], is considered when constructing the adaptive soft sensors.

$$\beta(t) = \alpha[Y(t-1) - Z(t-1)] + (1-\alpha)\beta(t-1),\ \alpha \in [0,\ 1] \tag{4.5}$$

where $Y(t-1)$ is the laboratory analysis of steam-quality and $Z(t-1)$ is the model prediction at the previous sampling instant. The bias correction term is updated only when a new laboratory data point is available.

4.3.1 Single-model based soft sensors

The single-model soft sensors of steam-quality measurement are proposed by Xie *et. al* [62], and we further improve it in this work. Since the industrial data are noisy with potential outliers, data preprocessing is necessary. A popular univariate approach to detect outliers is the 3σ rule [71],

$$|x(t) - \bar{x}| > 3\sigma \tag{4.6}$$

where $\bar{x}$ is the mean of the data sequence.

The single-model soft sensors, which is named as *Soft Sensor I*, can be constructed as

$$\hat{Y}(t) = Z(t) + \beta(t), \tag{4.7a}$$

$$Z(t) = f(\theta, U(t)) = \begin{cases} k_1u_1(t) + k_2u_2(t) + k_3u_3(t), & I - SQ \\ kX(t), & O - SQ \end{cases} \tag{4.7b}$$

$$\beta(t) = \alpha[Y(t-1) - Z(t-1)] + (1-\alpha)\beta(t-1),\ \alpha \in [0,\ 1]. \tag{4.7c}$$

$\hat{Y}(t)$ is the steam-quality prediction of the adaptive *Soft Sensor I*; θ is the model parameter, and $U(t)$ is the input of the model at time instant t. The model parameter K ($[k_1,\ k_2,\ k_3]$ for I-SQ, and k for O-SQ) and the bias parameter α are estimated simultaneously by the prediction error method (PEM). Considering the constraint of α, the projection method is used to solve the constraint optimization problem as part of the PEM algorithm. For details, please refer to [62] for the PEM algorithm and [72] for the projection method.

4.3.2 Multi-model soft sensors

The multi-model soft sensors of steam-quality are named as *Soft Sensor II*. The process of steam generation is approximated by first-principle models with multiple sets of parameters. The model parameters of each model are estimated by the VB algorithm. Considering the existence of outliers, the parameter estimation method should be able to resist the adverse influence of outliers. Though the 3σ rule is easy to implement, this procedure often fails in practice because the presence of outliers tends to inflate the variance estimation, which causes too few outliers to be detected [73]. Therefore, a more advanced method, i.e., use of the t distribution which can

eliminate the influence of outliers with the VB algorithm, is utilized for the design of multi-model soft sensors.

The steam generation process may vary with time owing to the variation of operating conditions, such as increase of BFW flow rate, and decrease of combustion air, and so on. Hence, we use multiple local-models to capture the time-varying property, i.e.,

$$Y(t) = Z_i(t) + e_i(t) \tag{4.8}$$

if i-th local model takes effect, and

$$Z_i(t) = f_i(\theta_i, U(t)) = \begin{cases} k_{i1}u_1(t) + k_{i2}u_2(t) + k_{i3}u_3(t), & I - SQ \\ k_i X(t), & O - SQ \end{cases} \tag{4.9}$$

where $i \in \{1, 2, \cdots, m\}$ and m is the number of local models. $e_i(t)$ is the noise whose distribution is considered to follow a t distribution with mean 0, precision (inverse variance) δ_i, and degrees of freedom ν_i. It is obvious that given the model prediction $Z_i(t)$ and distribution parameters, the output $Y(t)$ follows a t distribution as well,

$$Y(t)|\{Z(t), \delta_i, \nu_i\} \sim t(Z(t), \delta_i, \nu_i). \tag{4.10}$$

Using the expression of the t distribution directly in Bayesian estimation results in intractable problem. Hence, in Bayesian inference the t distribution is usually decomposed into scaled Normal distributions and a Gamma distribution by utilizing a hidden variable which is the scale R_t [15], i.e.,

$$t(Y(t)|Z(t), \delta_i, \nu_i) = \int_0^\infty \mathcal{N}(Y(t)|Z(t), R_t\delta_i)\mathcal{G}(R_t|\frac{\nu_i}{2}, \frac{\nu_i}{2})dR_t. \tag{4.11}$$

Since there are multiple local-models, the t-th sampling point may be generated by any local-model. In order to properly assign samples to different local-models, each sample is paired with an identity I_t indicating the corresponding local-model identity. Suppose we have N data samples; correspondingly, we have N identities, i.e., $I_t = i$, $i \in \{1, 2, \cdots, m\}$ and $t = 1, 2, \cdots, N$. The probability that $I_t = i$ should be determined. Thus, in the design of multi-model soft sensors, the hidden variables ($H = \{I,\ R\}$) are utilized to address the parameter estimation problem. The parameters of the local-model denoted by θ includes local model parameters K, parameter precision β, noise precision δ and degrees of freedom ν. Given the

identification data set $D = \{Y,\ U\}$, the parameter estimation problem is formulated under the Bayesian framework:

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{p(D)}, \tag{4.12}$$

where $p(\Theta|D)$ is the posterior probability density function and $\Theta = [\theta_1, \theta_2, \cdots, \theta_m]$. $p(D|\Theta)$ is the likelihood and $p(\Theta)$ is the joint prior distribution of parameters. $p(D)$ is the evidence which is an unknown constant. The appropriate prior distributions are assigned to each parameter, that is, K as Normal distribution, β as Gamma distribution, δ as Gamma distribution, and ν as exponential distribution.

In literature, the MAP estimation method is commonly utilized to get point-estimate of parameters based on the likelihood and the prior, and the evidence is neglected. In full-Bayesian estimation, the evidence is the key quantity to be evaluated, where the following integral is an obstacle:

$$\begin{aligned} p(D) &= \int p(D, H, \Theta) dH d\Theta \\ &= \int p(D|H, \Theta) p(H, \Theta) dH d\Theta. \end{aligned} \tag{4.13}$$

The integral is often intractable without resorting to advanced methods. A free joint density function $q(H, \Theta)$ is inserted into the integral, and the log-evidence is expressed as follows:

$$\ln p(D) = \ln \int q(H, \Theta) \frac{p(D, H, \Theta)}{q(H, \Theta)} dH d\Theta. \tag{4.14}$$

Under the VB framework, the factorization assumption ($q(H, \Theta) = q(H)q(\Theta)$) is adopted. Owing to the concavity of the logarithm function, the Jensen's inequality results in a lower bound of the log-evidence:

$$\ln p(D) \geq \int q(H)q(\Theta) \ln \frac{p(D, H, \Theta)}{q(H)q(\Theta)} dH d\Theta. \tag{4.15}$$

The lower bound is denoted as $F[q(H), q(\Theta)]$, and the lower bound equals to the log-evidence if and only if $p(H, \Theta|D) = q(H)q(\Theta)$. Hence, under the VB framework we must find the free density functions ($q(H)$ and $q(\Theta)$) that best describe the model given the observed data-set D. In this problem, the identity I is a discrete variable and the scale R is a continuous variable, so the lower bound is formulated as

$$F[q(R, I), q(\Theta)] = \sum_I \int q(R, I)q(\Theta) \ln \frac{p(D, R, I, \Theta)}{q(R, I)q(\Theta)} dR d\Theta. \tag{4.16}$$

The updating expression of $q(R, I)$ is obtained by solving the following optimization problem through the first-order functional derivative:

$$\max_{q(R,I)}\{F[q(R,I), q(\Theta)^{(n)}]\},\ \ s.t.\ \sum_{I}\int q(R,I)dR = 1, \tag{4.17}$$

where $q(\Theta)^{(n)}$ is the density function obtained in the previous iteration. Similarly, the updating expression of $q(\Theta)$ is obtained from the first-order functional derivative, i.e.,

$$\frac{\partial\{F[q(R,I)^{(n)}, q(\Theta)] + \lambda[\int q(\Theta)d\Theta - 1]\}}{\partial q(\Theta)} = 0, \tag{4.18}$$

where $q(R, I)^{(n)}$ is the updated density function and λ is the Lagrangian multiplier.

Based on the above derivations, the implementation procedure of the variational Bayesian estimation of parameters is outlined in Table 4.2.

Table 4.2: Procedure of the VB estimation method

1.	*Initialization.* Set $n = 0$. Determine prior density $p(\Theta)$ and related hyper-parameters.
2.	*Iterative updating.* Evaluate $q(R, I)^{(n+1)}$ and $q(\Theta)^{(n+1)}$.
3.	*Evaluate* $F[q(R, I)^{(n+1)}, q(\Theta)^{(n+1)}]$. Calculate the new value of the lower bound of the evidence.
4.	*Check stop criterion.* If $\lvert F[q(R, I)^{(n+1)}, q(\Theta)^{(n+1)}] - F[q(R, I)^{(n)}, q(\Theta)^{(n)}]\rvert \leq \epsilon$, stop. Otherwise, set $n = n + 1$, and go to step 2.

The parameter of each model, namely K, is estimated when the proposed algorithm converges, and the outliers in the identification data are handled by the t-distributions. The global output of the model is a weighted mixture of these basis models, i.e.,

$$Z_G(t) = \sum_{i=1}^{m}\omega_i(t)Z_i(t), \tag{4.19}$$

where the basis model is expressed by Eq. (4.9), and the mixing coefficient of each local model at t-th sampling instant is $\omega_i(t),\ i \in \{1, 2, \cdots, m\}$. The calculation of the mixing coefficient is based on the posterior probability of the identity $I_t = i$ given the t-th sampling data point and estimated parameters, i.e.,

$$\omega_i(t) = p(I_t = i|Y(t), U(t), \hat{\Theta}) \approx q(I_t = i), \tag{4.20}$$

where the posterior probability is approximated by $q(I_t = i)$ after the convergence of the algorithm. Considering the model error, the bias correction term expressed by Eq. (4.5) is utilized to construct the adaptive multi-model soft sensors, namely, *Soft Sensor II*, i.e.,

$$\hat{Y}(t) = Z_G(t) + \beta(t), \tag{4.21a}$$

$$Z_G(t) = \sum_{i=1}^{m} \omega_i(t) Z_i(t), \tag{4.21b}$$

$$\beta(t) = \alpha[Y(t-1) - Z_G(t-1)] + (1-\alpha)\beta(t-1),\ \alpha \in [0,\ 1]. \tag{4.21c}$$

Here α is not estimated simultaneously with model parameters as it was done in *Soft Sensor I*. Instead, it is treated as a tuning parameter in order to reduce the complexity.

When implementing the *Soft Sensor II* online, the steam-quality at q-th instant is predicted by the soft sensor. However, the mixing coefficient of each basis model at the new instant cannot be calculated by Eq. (4.20) since $q(I_q = i)$ is unknown. Therefore, the calculation of mixing coefficients is the key component in the real-time prediction of the steam-quality.

The identification data $\{Y, U\}$ are stored in a database. When a query sample U_q in the input space comes, the similarity s_j between U_q and each identification data point U_j is calculated based on the Euclidean distance

$$d_j = \sqrt{(U_j - U_q)^T (U_j - U_q)},\ j = 1, 2, \cdots, N, \tag{4.22}$$

where N is the number of identification data samples. In this work, the method defined by Eq. (4.23) [74] is adopted to measure the similarity:

$$s_j = exp\left(-\frac{d_j}{\sigma_{d_j}\phi}\right), \tag{4.23}$$

where σ_{d_j} is the standard deviation of $d_j (j = 1, 2, \cdots, N)$ and ϕ is a localization parameter. The similarity decreases sharply when ϕ gets small and gradually when ϕ is large.

Resorting to this method, the similarity between the query sample and every identification data point is obtained. The mixing coefficients of each identification

data point, namely $\omega_i(j)$, are stored in a data base, and the mixing coefficient of the query sample can be calculated based on the similarity as follows:

$$\omega_i(q) = \sum_{j=1}^{N} s_j \cdot \omega_i(j),\ i = 1, 2, \cdots, m. \tag{4.24}$$

Once the mixing coefficients of the query sample are calculated, the predicted steam-quality by the *Soft Sensor II* at any instant is obtained according to Eq. (4.21).

4.4 Industrial case studies

Single-model based soft sensors and multi-model soft sensors have been designed based on the first-principle model of the steam-generation process. In this section, two once-through steam generators with different structures are used to evaluate the performance of the developed *Soft Sensor I* and *Soft Sensor II*.

4.4.1 Case I: OTSG

A. Process description

The process of the once-through steam generator (OTSG) is described in Section 2. The simplified schematic diagram is shown in Fig. 4.2, and Table 4.1 lists the process variables utilized in the design of soft sensors. The real-time measurements were recorded every 10 minutes, whereas the laboratory analysis of I-SQ was logged every 6 hours. The operational and laboratory data were collected through an automated data historian. The data recorded from August 1, 2012 to August 31, 2013 are available. Missing measurements and outliers exist in the collected historical data. Simple pre-processing methods like the 3σ rule are adopted to refine the quality of the data. The identification data set consists of 664 data points, and the validation data set is composed of 332 data points.

B. Soft sensor identification

Online measurements of both I-SQ and O-SQ are available, but the accuracy is not satisfactory. Take the individual pass 1 as an example, the laboratory analysis and the online measurement of I-SQ are displayed in Fig. 4.3. Clearly, the online measurement

of steam-quality does not track the trend of the true value from 100 to 200 sampling instant.

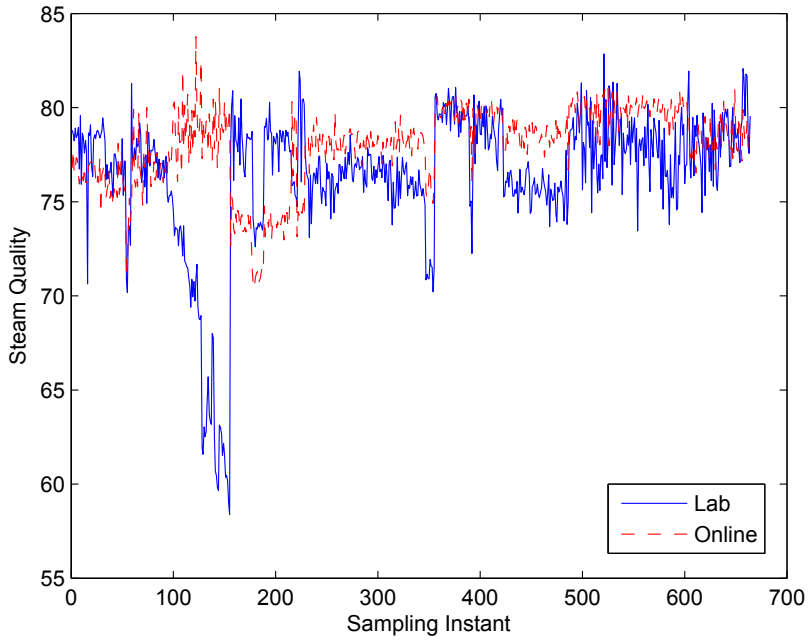

Figure 4.3: Laboratory analysis and online measurement of pass 1 steam-quality.

The identification procedure of soft sensors involves the online measurement of O-SQ, as Eq. (4.3a) shows. The trend of the online measurement can thus influence the accuracy of estimation results, especially for the single-model based soft sensors. By investigating the process data, we find that compared with the online measurement of O-SQ, the online measurements of I-SQ are much better. Therefore, the weighted average of the measured I-SQ of 8 individual passes is treated as the online measurement of O-SQ. That is to say, X is replaced by $\bar{X}$ in Eq. (4.3a), and $\bar{X}$ is

$$\bar{X}(t) = \frac{\sum_{l=1}^{8} F^l(t) X^l(t)}{\sum_{l=1}^{8} F^l(t)}. \tag{4.25}$$

In addition, there is no laboratory analysis of O-SQ, so the weighted average of the laboratory analysis of each I-SQ is calculated to be the laboratory data of O-SQ that

acts as the reference of the average steam-quality, i.e.,

$$\bar{Y}(t) = \frac{\sum_{l=1}^{8} F^l(t) Y^l(t)}{\sum_{l=1}^{8} F^l(t)}. \tag{4.26}$$

Fig. 4.4 shows the steam-quality obtained by $\bar{Y}$, $\bar{X}$ and X, respectively. Compared with the trend of the online measurement (X), the trend of $\bar{X}$ is more similar to the trend of the reference. Based on the local-model structure and collected identification data, models of soft sensors are identified. We denote the model of *Soft Sensor I* as *Model I*, and the model of *Soft Sensor II* as *Model II*. By including the bias correction term with the models, *Soft Sensor I* and *Soft Sensor II* are constructed. The performance of the soft sensors is compared with the performance of the online measurement.

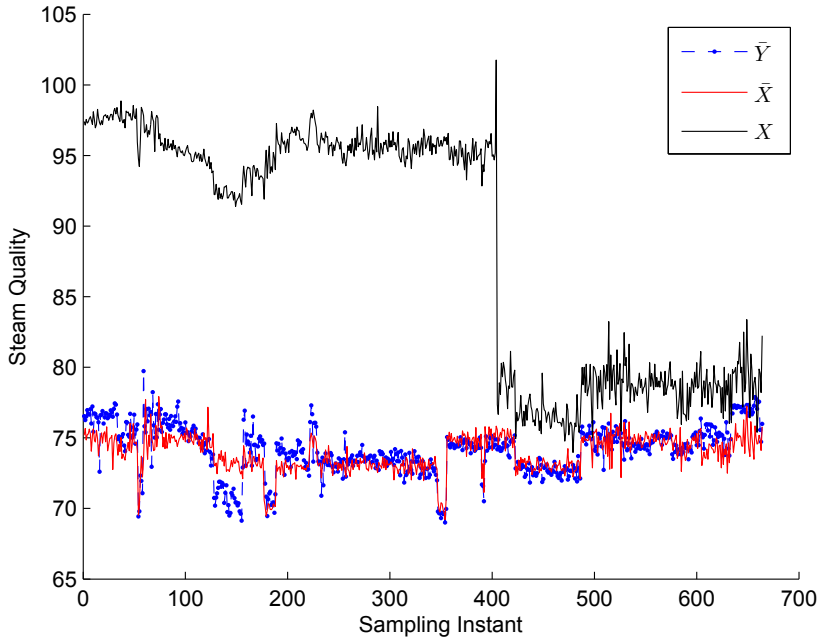

Figure 4.4: The steam-quality obtained by weighted average of lab analysis ($\bar{Y}$), weighted average of online measurements of I-SQ ($\bar{X}$), and the online measurement of O-SQ (X).

C. Soft sensor evaluation

The accuracy and reliability of the developed soft sensors should be evaluated before online implementation. The level of agreement between the predicted and reference values is evaluated by accuracy. The reliability is the degree to which the prediction errors vary. Thus, evaluating the performance of soft sensors is closely related to accessing the prediction errors (ε).

In order to evaluate the performance of soft sensors in a quantitative way, the mean absolute error (MAE), standard deviation of errors (StdE), and mean squared error (MSE) are calculated. The MAE is used to access the accuracy, and StdE is a measure of reliability. Meanwhile, the MSE reveals both accuracy and reliability of the prediction performance. The formula of MAE, StdE, and MSE are expressed as follows:

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|\varepsilon_t|, \tag{4.27}$$

$$StdE = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N}(\varepsilon_t - \bar{\varepsilon})^2}, \tag{4.28}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^2}. \tag{4.29}$$

The OTSG has 8 individual passes in total. The performance of soft sensors for each individual pass is assessed, both in self-validation and cross-validation. Fig. 4.5 shows the MAE of estimated steam-quality by the online measurement, *Model I*, *Model II*, *Soft Sensor I*, and *Soft Sensor II*, respectively. Fig. 4.6 is the StdE in self-validation and cross-validation. Fig. 4.7 presents the MSE. As for the prediction of O-SQ, the evaluation results of predictions are recorded in Table 4.3. According to the comparison, the performance of *Soft Sensor I* is better than that of the online measurement, whereas *Soft Sensor II* outperforms *Soft Sensor I*. In addition, *Model II* is more accurate and reliable than *Model I*, from which we can see the advantage of the proposed multi-model approach.

Besides, the scatter plot of the predicted values versus reference values is utilized to evaluate the prediction performance visually. Ideally, the predicted values equal to the reference values (laboratory analysis), i.e., all the data points lie on the 45 degree

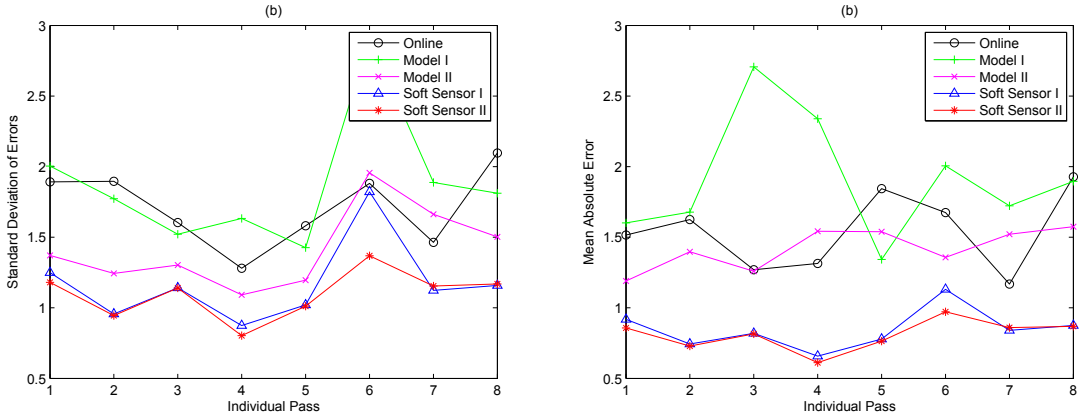

Figure 4.5: OTSG, I-SQ. (a) MAE in self-validation; (b) MAE in cross-validation

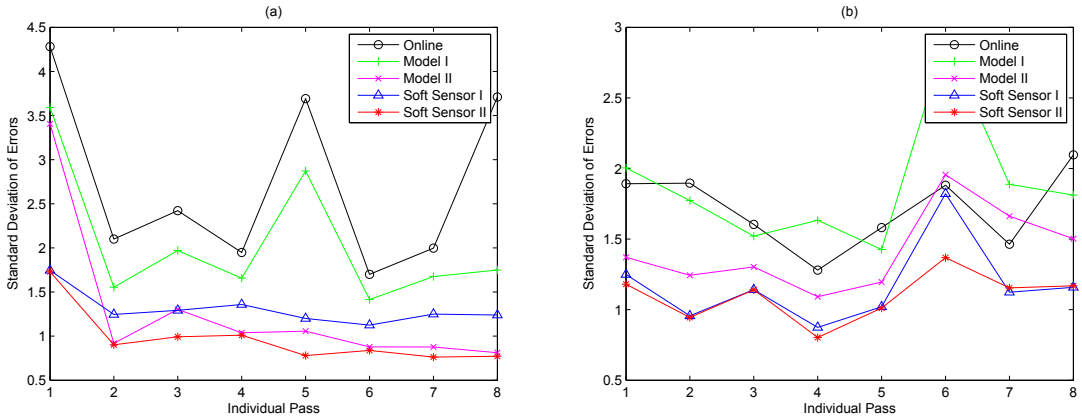

Figure 4.6: OTSG, I-SQ. (a) StdE in self-validation; (b) StdE in cross-validation

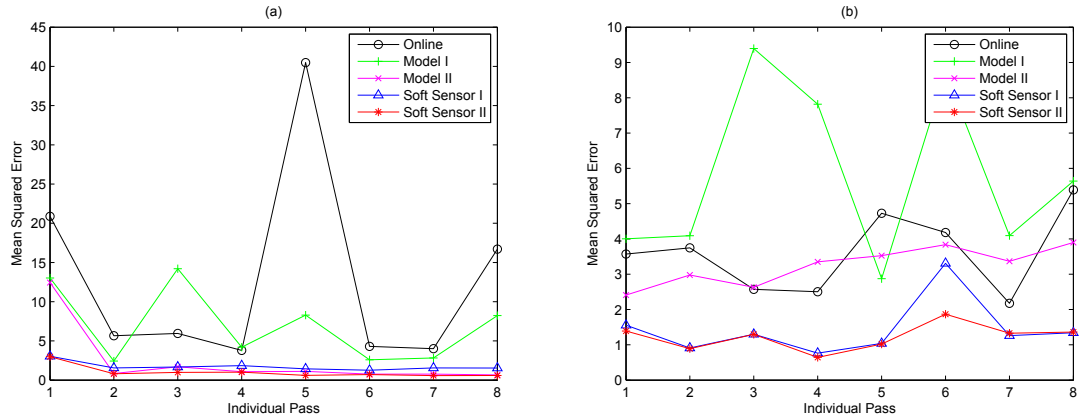

Figure 4.7: OTSG, I-SQ. (a) MSE in self-validation; (b) MSE in cross-validation

Table 4.3: A evaluation summary of the prediction performance (O-SQ).

	MAE	StdE	MSE
Self-validation			
Online measurement	0.9584	1.2383	1.5951
Model I	0.9608	1.2386	1.5717
Model II	0.5017	0.8516	0.7258
Soft Sensor I	0.6823	0.9545	0.9098
Soft Sensor II	0.4214	0.6857	0.4695
Cross-validation			
Online measurement	0.8651	1.0763	1.1596
Model I	0.9274	1.0766	1.4349
Model II	0.8850	1.0855	1.2873
Soft Sensor I	0.5911	0.7877	0.6186
Soft Sensor II	0.5640	0.7577	0.5723

line ($y = x$), indicating perfect matching between the prediction and the reference. The time-trend plot of predicted values and reference values is also widely used for visualizing the prediction performance. Take the individual pass 1 as an example, the scatter plot and time-trend plot are shown in Fig. 4.8 and Fig. 4.9. These figures indicate that the prediction performance of both soft sensors is good in this case study, and *Soft Sensor II* outperforms *Soft Sensor I* slightly.

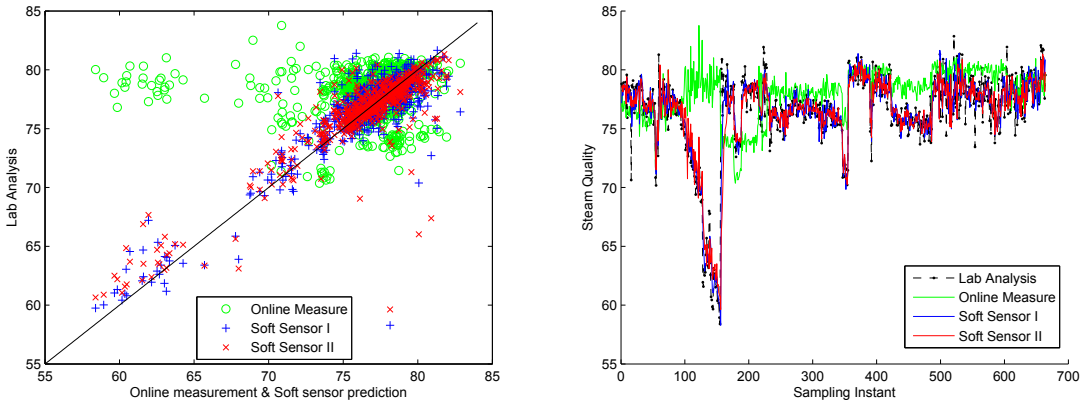

Figure 4.8: Self-validation, individual pass 1. (a) Scatter plot comparison; (b) time-trend comparison

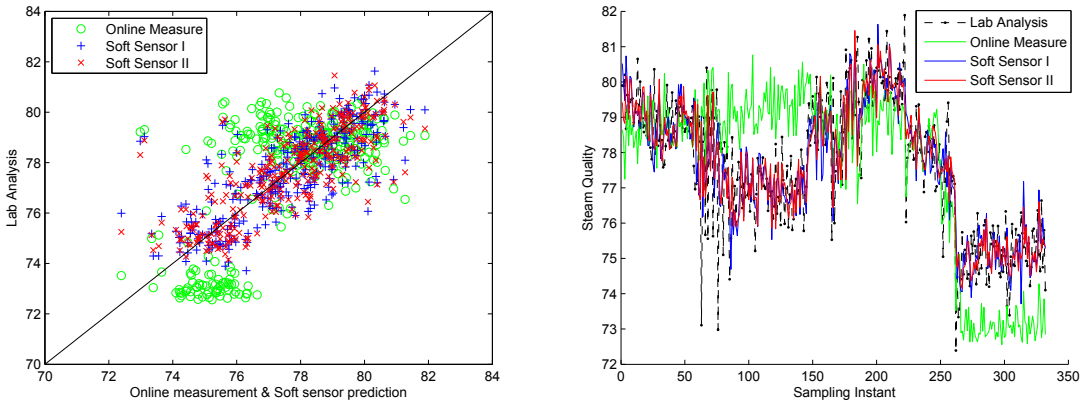

Figure 4.9: Cross-validation, individual pass 1. (a) Scatter plot comparison; (b) time-trend comparison

4.4.2 Case II: Cogen-HRSG

A. Process description

The Cogeneration consists of a Heat Recovery Steam Generator (HRSG), which provides steam for the injection into the reservoirs of SAGD systems, and a gas turbine that supplies power to the plant and sales to the grid. Fig. 4.10 shows the simplified schematic diagram of Cogen-HRSG. The fuel gas consisting of compressed air and combust is combusted in the gas turbine to generate electricity. The hot exhaust from the combustion is sent to the HRSG to provide heat for the steam generator. Since the exhaust cannot provide enough energy, extra fuel gas is burned in the duct burner to produce additional heat for the steam generator.

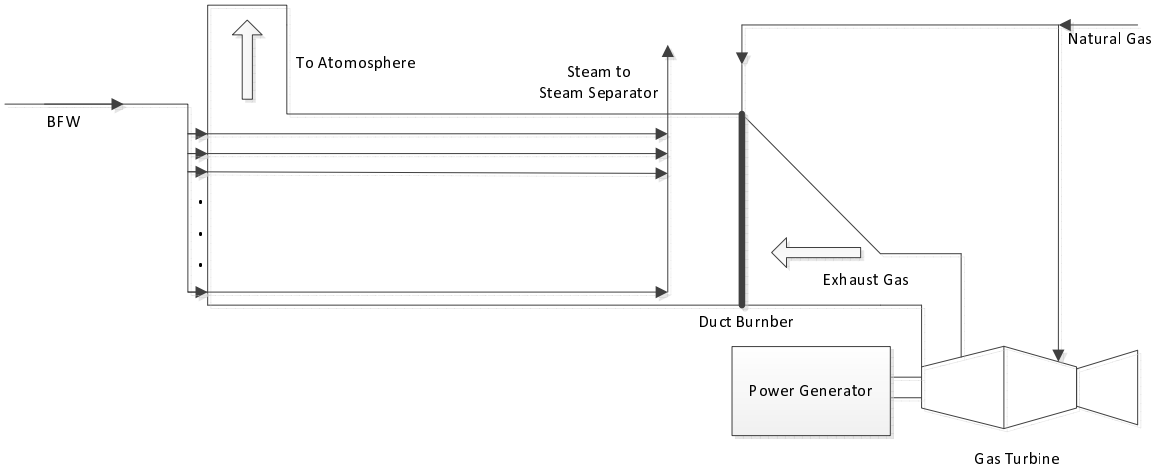

Figure 4.10: Schematic diagram of Cogen-HRSG.

The structure of Cogen-HRSG is different from the structure of OTSG, but they share the same model structure which is based on the heat transfer principle. That is to say, the process variables of interest are same, as listed in Table 4.1. The only difference is that the number of individual passes is 12 instead of 8 in this case study.

The real-time measurements of variables were recorded every 1 minute, and the laboratory analysis was logged every 6 hours. The data recorded from January 1, 2013 to July 1, 2014 are available to use. Through preprocessing, we have 1490 data points in total. The identification data set consists of the previous 1000 data points, and the rest data are used for validation.

B. Soft sensor identification and evaluation

Fig. 4.11 shows the laboratory analysis and online measurement of steam-quality in individual pass 1. The online measurement is not accurate. Following the same procedure described in Case I, models are identified first, and the soft sensors are constructed by taking the bias correction term into account.

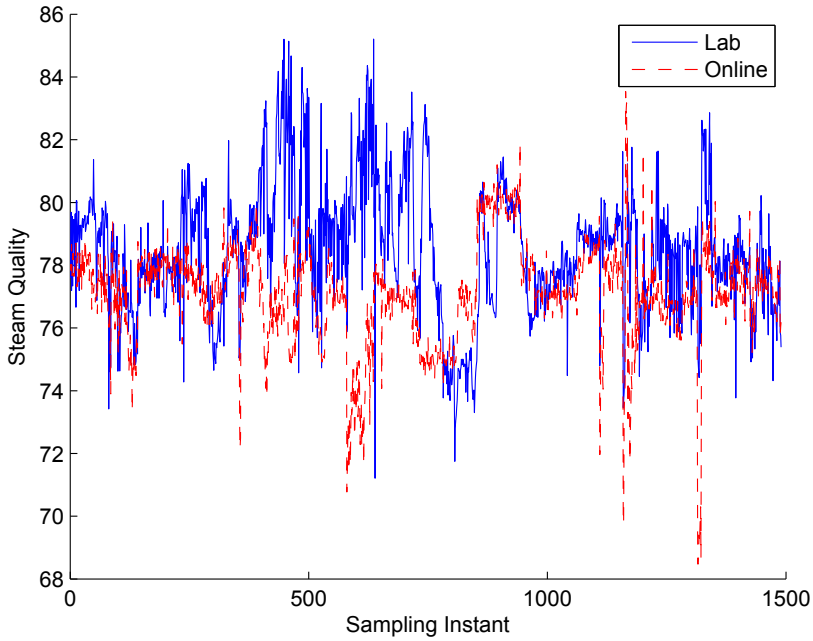

Figure 4.11: Steam-quality in individual pass 1.

The performance of models and soft sensors for each individual pass (12 passes in total) is evaluated by the MAE, StdE and MSE, respectively. Fig. 4.12, Fig. 4.13,

and Fig. 4.14 present the comparison results of their prediction performance. We can see that the online measurement is not accurate, especially the data in self-validation. The performance of *Soft Sensor I* appears satisfactory, but the model (namely *Model I*) has a poor performance. That means the bias correction term plays a key role in the prediction of steam-quality which indicates the soft sensor prediction may not be reliable. In contrast, *Model II* is much better than *Model I*, which demonstrates the effectiveness of the multi-model approach. Moreover, the standard deviation of errors of *Soft Sensor II* is the smallest, so it is the most reliable method to predict steam-quality.

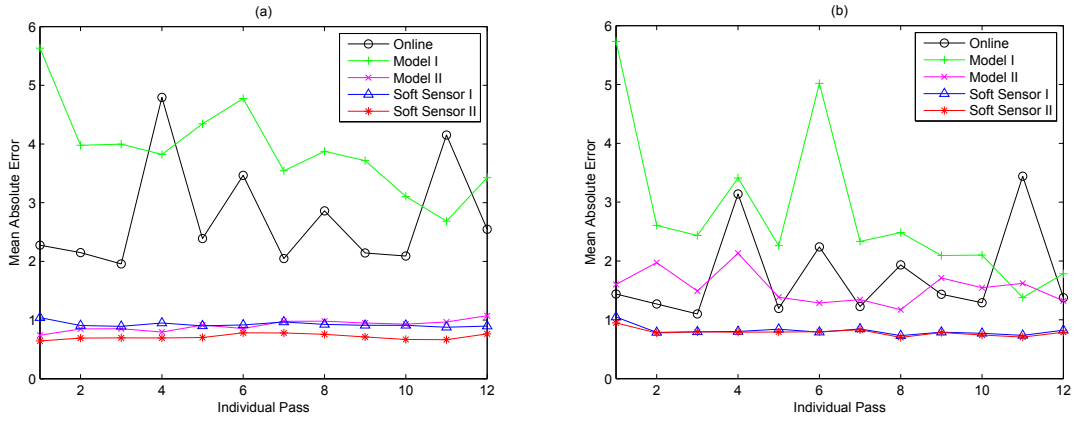

Figure 4.12: Cogen-HRSG, I-SQ. (a) MAE in self-validation; (b) MAE in cross-validation

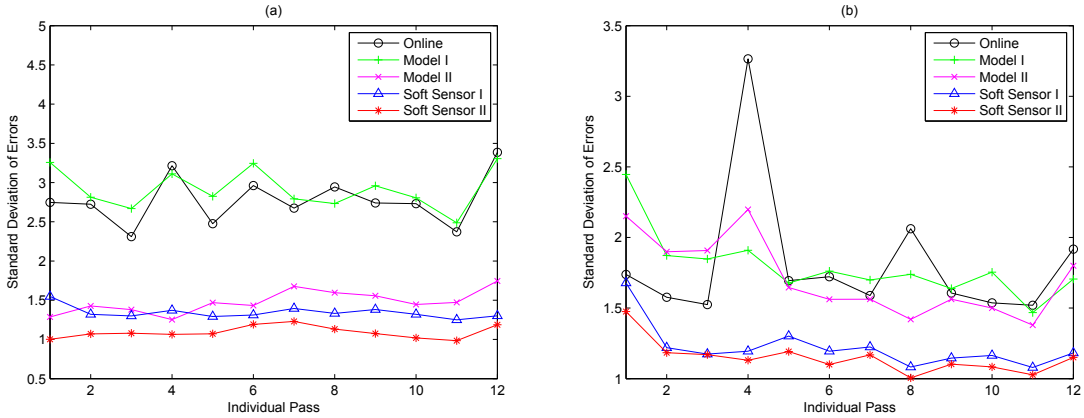

Figure 4.13: Cogen-HRSG, I-SQ. (a) StdE in self-validation; (b) StdE in cross-validation

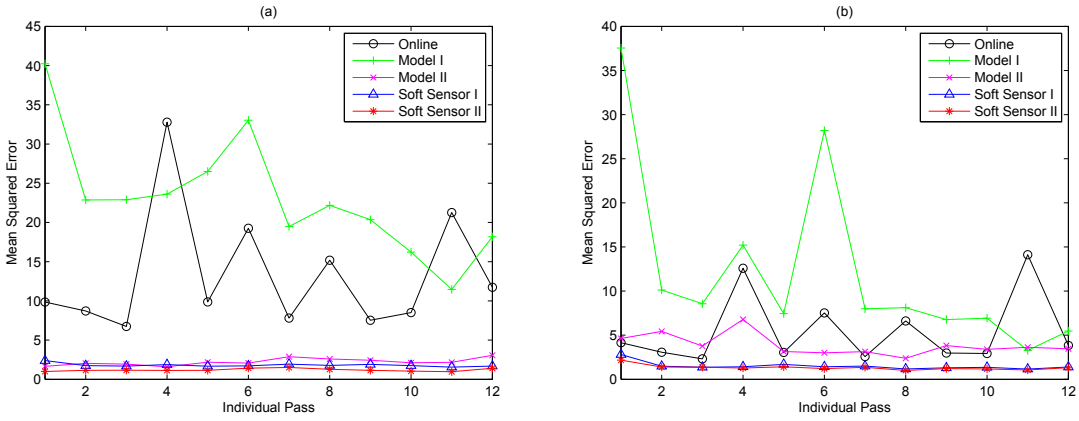

Figure 4.14: Cogen-HRSG, I-SQ. (a) MSE in self-validation; (b) MSE in cross-validation

Taking the individual pass 1 as an example, the time-trend plot and scatter plot of *Soft Sensor I* are shown in Fig. 4.15 and Fig. 4.16. The bias of the prediction given by *Model I* is obvious. After bias correction, the performance appears to improve. For comparison, the time-trend plot and scatter plot of *Soft Sensor II* are shown in Fig. 4.17 and Fig. 4.18. The performance of *Model II* is satisfactory, and after the bias correction, the performance is even better.

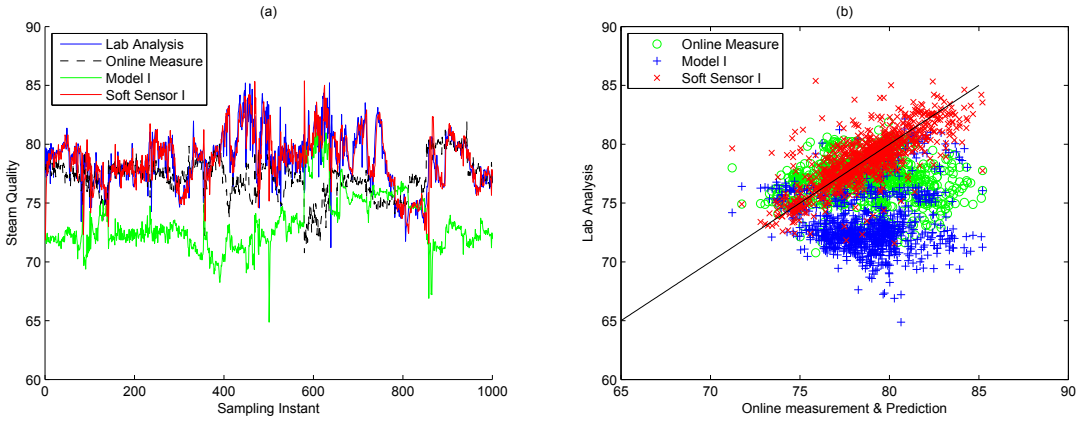

Figure 4.15: Self-validation *Soft Sensor I*, I-SQ 1. (a) Time-trend comparison; (b) scatter plot comparison

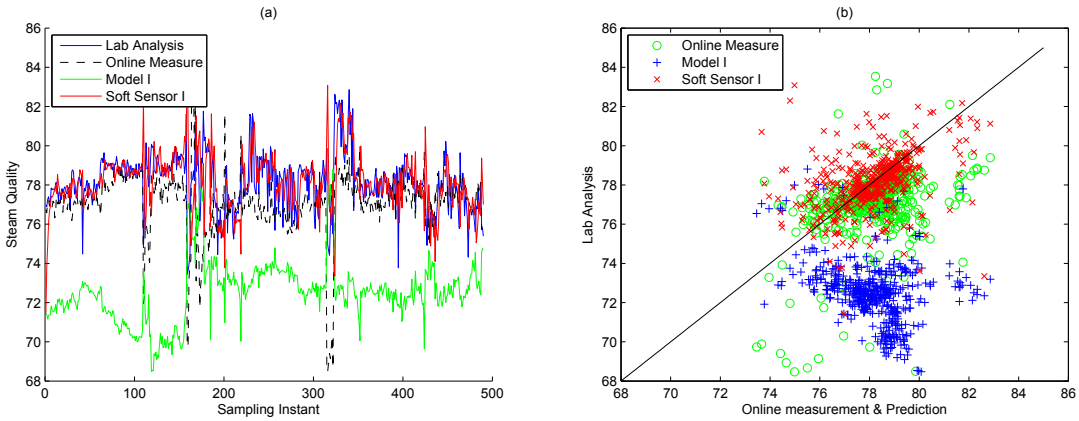

Figure 4.16: Cross-validation *Soft Sensor I*, I-SQ 1. (a) Time-trend comparison; (b) scatter plot comparison

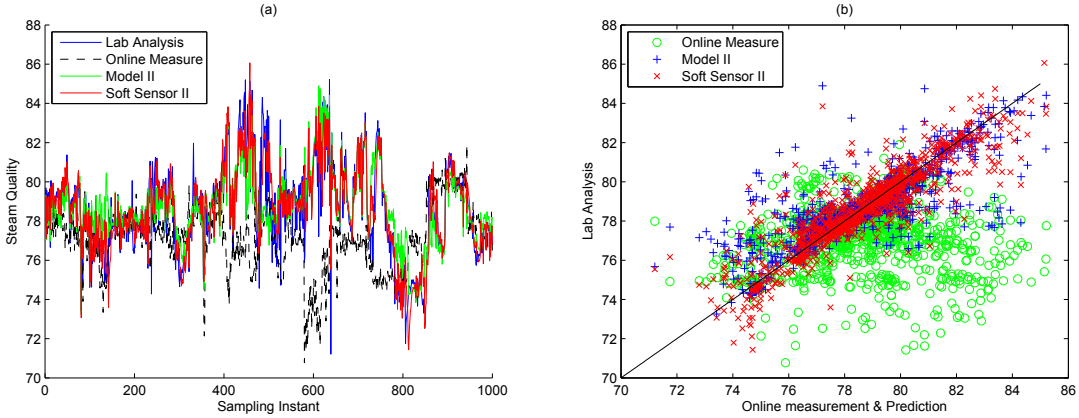

Figure 4.17: Self-validation *Soft Sensor II*, I-SQ 1. (a) Time-trend comparison; (b) scatter plot comparison

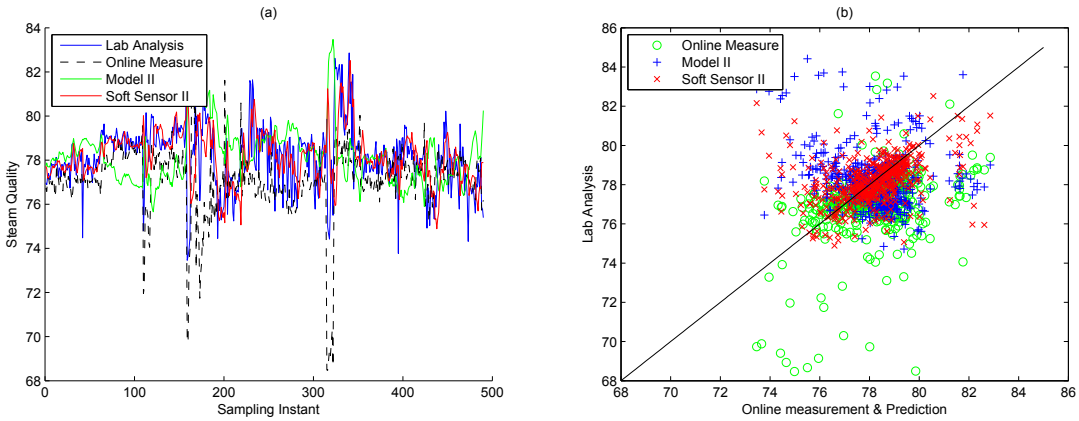

Figure 4.18: Cross-validation *Soft Sensor II*, I-SQ 1. (a) Time-trend comparison; (b) scatter plot comparison

4.5 Discussion

Industrial case studies have demonstrated the effectiveness of the developed soft sensors. *Soft Sensor I* is a single-model based inferential sensor, which is simple and easily identifiable. *Soft Sensor II* is a multi-model inferential sensor, which is more sophisticated, but more accurate and reliable.

Resorting to the online bias updating term, the soft sensors become adaptive to the operating conditions, and both soft sensors outperform the online measurement. However, we must realize that the bias updating term depends on the laboratory analysis and can only correct slow dynamics. If the laboratory analysis is not available, the prediction performance of the soft sensors can deteriorate, indicating potential problems when predicting at a point far away from lab data. Especially for *Soft Sensor I*, the prediction of the single model (*Model I*) cannot track the real steam-quality well, as we can see from figures in case studies. In contrast, *Model II*, namely the model of *Soft Sensor II*, is capable of providing acceptable predictions of the steam-quality. Furthermore, the prediction of *Model II* is more accurate and reliable than the online measurement. After bias correction, *Soft Sensor II* outperforms other fast-rate estimation of steam-quality known to this work.

When identifying *Soft Sensor I*, all parameters including the weight of the bias updating term are estimated by the PEM algorithm simultaneously. The advantage of this approach is that we do not need to tune any parameter. Besides, the computational load is not heavy. However, the reliability of this approach is in doubt. The weight (α) is a special parameter which has a constraint condition, and the convergence may be different from the convergence of the model parameter (K). If we adopt the method proposed in [62] directly, the estimation result might not be satisfactory. We have met the situation where the estimated α is larger than 1, resulting in significant oscillation and even divergence of the steam-quality prediction. By using the projection method, α is forced to be located within $[0,\ 1]$. However, $\hat{\alpha}$ may be 1 or close to 1, and the model parameters are still not well estimated. Then the bias correction term updates sharply when the new laboratory data point comes because of the large gap between the model prediction and the laboratory analysis. Therefore, estimating all parameters simultaneously using PEM may not give us the

best results.

On the contrary, the model parameters of *Soft Sensor II* are identified by the variational Bayesian method by taking the influence of outliers into account, and the weight parameter is tuned manually instead of being estimated by the algorithm in order to reduce the computational load. The variational Bayesian method is more sophisticated than the prediction error method, and thus can identify the model better, but it is computationally expensive. Fortunately, the identification procedure is off-line, so it does not matter if the estimation procedure takes longer time. During the prediction, the similarity between the query data point and historical data is calculated in order to determine the operating mode. The prediction of steam-quality by the model may still have bias, so the bias correction term is utilized to improve the prediction performance. The weight is tuned, and usually $\alpha = 0.3$ can compensate the bias between the model prediction and the laboratory analysis, and yield a smooth updating of the bias correction term.

4.6 Conclusions

The estimation of the steam-quality in the steam generator is challenging. In this research, based on the previous work of Xie *et. al*, we designed two kinds of adaptive soft sensors. The single-model based soft sensor is simple and easy to be developed. Industrial case studies verified the effectiveness of the soft sensor, but generality is still in question since the performance of the model could be worse than the performance of the online measurement under some conditions, which means that the bias correction is important in *Soft Sensor I*. A variational Bayesian approach to the development of multi-modal soft sensors is proposed. The influence of outliers is reduced by resorting to t distributions. Both the model and the soft sensor have a good performance in the case study of OTSG and Cogen-HRSG. The multi-model soft sensor is more reliable. Therefore, the advantage of *Soft Sensor II* over *Soft Sensor I* is significant if the laboratory analysis is not available for a considerable time period during the operation.

Chapter 5

Conclusions

5.1 Summary of this thesis

This thesis focuses on robust identification of industrial processes, generally viewed as multi-modal processes, with time-varying behaviour, nonlinearity, and switching dynamics. Soft sensors based on proposed algorithms for industrial applications are designed to meet the requirement of the real-time estimation of steam-quality.

The background and motivation of the robust identification of multi-modal processes were presented in Chapter 1.

Chapter 2 proposed a robust multiple-model LPV approach to identify the nonlinear process subject to outliers using mixture t distributions. The basic idea of this approach was to use a weighted combination of local-models around operating points to approximate the nonlinear process across the whole operating range. Resorting to t distributions, the outliers were down-weighted automatically during the iterative optimization.

A variational Bayesian approach to identification of switched ARX models was developed in Chapter 3. Practical issues such as robustness and outlier detection, estimation of the number of operating modes, and estimation of parameter uncertainty were addressed in this chapter. A set of mixing coefficients was introduced to indicate the significance of each local-model and redundant local-models were eliminated by the algorithm during the iterative procedure of optimization. The number of local-models and the probability distribution functions over parameters were obtained simultaneously.

The advantage of using multiple local-models in process modelling was shown in

simulations and experiment case studies. The performance of soft sensors is dependent on the derived or the identified process model. The multi-model modelling approach was applied to the design of soft sensors in Chapter 4. The previous single-model based soft sensor was simple and easy to be implemented, but the reliability can be improved further. The developed multi-model soft sensor performed well in industrial case studies of OTSG-93 and Cogen-HRSG, and the performance of the model was better than that of the online measurement of steam-quality.

5.2 Directions for future work

We worked on the modelling and identification of multi-modal processes throughout this thesis. Statistical optimization approaches including the expectation-maximization algorithm and the variational Bayesian algorithm were investigated to address the problem of process identification. Practical issues like robustness, estimation of the number of operating modes, and outlier detection were addressed in this thesis. However, some issues in the field of multi-modal process identification remain open.

On one hand, in both the LPV modelling approach and the switched modelling approach, the order of the transfer function was assumed to be known. We used the off-line cross-validation method to determine the appropriate order of the transfer function, and assumed all the local-models had the same order. In practice, the order of each local-model can be different. An advanced method need to be developed to estimate the orders along with the estimation of model parameters.

On the other hand, we assumed that the collected training data covered all operating modes of the process and thus the identified model could represent the process over the whole operating range. This assumption is not always reasonable in industry. The collected training data may not cover the whole operating range owing to the constraints of operation during a certain period. The prediction performance of the model may be unsatisfactory if the process runs in the new operating mode which is not considered in process identification. How to detect new operating modes and then update the model remains to be addressed.

In addition, the outlier detection approach through predictive density function was not fully studied. The exact calculation of predictive density function is difficult,

and it has not been addressed in literature. In Chapter 3, the predictive density was approximated by its lower bound, and a threshold of the lower bound to distinguish outliers from normal data was utilized, which is *ad-hoc*. More objective methods need to be discovered to consummate the approach to outlier detection through predictive density function.

Finally, when implementing the multi-model soft sensor online, the calculation of mixing coefficients plays a key role in the real-time prediction of the steam-quality. A method based on the similarity between the query data and each identification data point was proposed to estimate the mixing coefficients. This may not be very accurate according to the prediction performance of the model in cross-validation. Therefore, a better method to estimate the mixing coefficients needs to be explored.

Bibliography

[1] Kwan-Woong Gwak and Wisup Bae. A review of steam generation for in-situ oil sands projects. *Geosystem Engineering*, 13(3):111–118, 2010.

[2] Jeff S Shamma and Michael Athans. Analysis of gain scheduled control for nonlinear plants. *Automatic Control, IEEE Transactions on*, 35(8):898–907, 1990.

[3] Tor A Johansen and Bjarne A Foss. Operating regime based process modeling and identification. *Computers & chemical engineering*, 21(2):159–176, 1997.

[4] YC Zhu and ZH Xu. A method of LPV model identification for control. *The international Federation of Automatic Control*, pages 6–11, 2008.

[5] Jacob Roll, Alberto Bemporad, and Lennart Ljung. Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40(1):37–50, 2004.

[6] Hayato Nakada, Kiyotsugu Takaba, and Tohru Katayama. Identification of piecewise affine systems based on statistical clustering technique. *Automatica*, 41(5):905–913, 2005.

[7] Aleksandar Lj Juloski, Siep Weiland, and WPMH Heemels. A bayesian approach to identification of hybrid systems. *Automatic Control, IEEE Transactions on*, 50(10):1520–1533, 2005.

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

[9] Michael E Tipping and Neil D Lawrence. Variational inference for student-t models: Robust bayesian interpolation and generalised component analysis. *Neurocomputing*, 69(1):123–141, 2005.

[10] Shima Khatibisepehr, Biao Huang, Fangwei Xu, and Aris Espejo. A bayesian approach to design of adaptive multi-model inferential sensors with application in oil sand industry. *Journal of Process Control*, 22(10):1913–1929, 2012.

[11] B T_ Poljak and Ja Z Tsypkin. Robust identification. *Automatica*, 16(1):53–63, 1980.

[12] Wolfgang Reinelt, Andrea Garulli, and Lennart Ljung. Comparing different approaches to model error modeling in robust identification. *Automatica*, 38(5):787–803, 2002.

[13] Fernando D Bianchi and Ricardo S Sánchez-Peña. Robust identification/invalidation in an lpv framework. *International Journal of Robust and Nonlinear Control*, 20(3):301–312, 2010.

[14] Yi Fang and Myong K Jeong. Robust probabilistic multivariate calibration model. *Technometrics*, 50(3), 2008.

[15] Markus Svensén and Christopher M Bishop. Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252, 2005.

[16] Thomas Pröll and M Nazmul Karim. Model-predictive pH control using real-time NARX approach. *AIChE Journal*, 40(2):269–282, 1994.

[17] Er-Wei Bai. An optimal two-stage identification algorithm for Hammerstein–Wiener nonlinear systems. *Automatica*, 34(3):333–338, 1998.

[18] Songwu Lu and Tamer Basar. Robust nonlinear system identification using neural-network models. *Neural Networks, IEEE Transactions on*, 9(3):407–429, 1998.

[19] Amauri H de Souza Junior, Francesco Corona, and Guilherme A Barreto. Robust regional modeling for nonlinear system identification using self-organizing maps. In *Advances in Self-Organizing Maps*, pages 215–224. Springer, 2013.

[20] K De Brabanter, P Karsmakers, F Ojeda, C Alzate, J De Brabanter, K Pelckmans, B De Moor, J Vandewalle, and JAK Suykens. LS-SVMlab toolbox users guide. *ESAT-SISTA Technical Report 10-146*, pages 33–35, 2011.

[21] Wilson J Rugh and Jeff S Shamma. Research on gain scheduling. *Automatica*, 36(10):1401–1425, 2000.

[22] Vincent Laurain, Marion Gilson, Roland Tóth, and Hugues Garnier. Refined instrumental variable methods for identification of LPV Box–Jenkins models. *Automatica*, 46(6):959–967, 2010.

[23] Bassam Bamieh and Laura Giarre. Identification of linear parameter varying models. *International Journal of Robust and Nonlinear Control*, 12(9):841–853, 2002.

[24] Jan-Willem van Wingerden and Michel Verhaegen. Subspace identification of bilinear and LPV systems for open-and closed-loop data. *Automatica*, 45(2):372–381, 2009.

[25] R Tóth, PSC Heuberger, and PMJ Van den Hof. An LPV identification framework based on orthonormal basis functions. In *proceedings of the 15th IFAC Symposium on System Identification, Saint-Malo, France*, 2009.

[26] Mark Butcher and Alireza Karimi. Data-driven tuning of linear parameter-varying precompensators. *International Journal of Adaptive Control and Signal Processing*, 24(7):592–609, 2010.

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

[28] John W Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962.

[29] Peter J Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[30] IB Tjoa and LT Biegler. Simultaneous strategies for data reconciliation and gross error detection of nonlinear systems. *Computers & chemical engineering*, 15(10):679–690, 1991.

[31] Moustapha Alhaj-Dibo, Didier Maquin, and José Ragot. Data reconciliation: A robust approach using a contaminated distribution. *Control Engineering Practice*, 16(2):159–170, 2008.

[32] GJ MacLachlan and D Peel. Finite mixture models. *Wiley series in probability and statistics*, pages 221–237, 2000.

[33] Demetrios Gerogiannis, Christophoros Nikou, and Aristidis Likas. The mixtures of students t-distributions as a robust framework for rigid registration. *Image and Vision Computing*, 27(9):1285–1294, 2009.

[34] Chuanhai Liu and Donald B Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5(1):19–39, 1995.

[35] Lennart Ljung. System identification: theory for the user. 1987.

[36] Yucai Zhu. Estimation of an N–L–N Hammerstein–Wiener model. *Automatica*, 38(9):1607–1614, 2002.

[37] SA Billings and WSF Voon. Piecewise linear identification of non-linear systems. *International journal of control*, 46(1):215–235, 1987.

[38] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[39] Jeff A Bilmes et al. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4(510):126, 1998.

[40] J Duane Morningred, Bradley E Paden, Dale E Seborg, and Duncan A Mellichamp. An adaptive nonlinear predictive controller. *Chemical Engineering Science*, 47(4):755–762, 1992.

[41] R Senthil, K Janarthanan, and J Prakash. Nonlinear state estimation using fuzzy Kalman filter. *Industrial & engineering chemistry research*, 45(25):8678–8688, 2006.

[42] Zuhua Xu, Jun Zhao, Jixin Qian, and Yucai Zhu. Nonlinear MPC using an identified LPV model. *Industrial & Engineering Chemistry Research*, 48(6):3043–3051, 2009.

[43] Nicanor Quijano, Alvaro E Gil, and Kevin M Passino. Experiments for dynamic resource allocation, scheduling, and control: new challenges from information technology-enabled feedback control. *Control Systems, IEEE*, 25(1):63–79, 2005.

[44] Giancarlo Ferrari-Trecate, Marco Muselli, Diego Liberati, and Manfred Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2):205–217, 2003.

[45] Laurent Bako, Khaled Boukharouba, Eric Duviella, and Stéphane Lecoeuche. A recursive identification algorithm for switched linear/affine models. *Nonlinear Analysis: Hybrid Systems*, 5(2):242–253, 2011.

[46] René Vidal. Recursive identification of switched arx systems. *Automatica*, 44(9):2274–2287, 2008.

[47] Shima Khatibisepehr and Biao Huang. A bayesian approach to robust process identification with arx models. *AIChE Journal*, 59(3):845–859, 2013.

[48] Hagai Attias. Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 21–30. Morgan Kaufmann Publishers Inc., 1999.

[49] Matthew J. Beal and Zoubin Ghahramani. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7:453–464, 2003.

[50] David Peel and Geoffrey J McLachlan. Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348, 2000.

[51] Chad Aeschliman, Johnny Park, and Avinash C Kak. A novel parameter estimation algorithm for the multivariate t-distribution and its application to computer vision. In *Computer Vision–ECCV 2010*, pages 594–607. Springer, 2010.

[52] Adrian Corduneanu and Christopher M Bishop. Variational bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA, 2001.

[53] JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, M West, et al. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7:453–464, 2003.

[54] Takashi Takekawa and Tomoki Fukai. A novel view of the variational bayesian clustering. *Neurocomputing*, 72(13):3366–3369, 2009.

[55] Sotirios P Chatzis and Dimitrios I Kosmopoulos. A variational bayesian methodology for hidden markov models utilizing student's-t mixtures. *Pattern Recognition*, 44(2):295–306, 2011.

[56] Tawfik Noaman Nasr, Oluropo Rufus Ayodele, et al. Thermal techniques for the recovery of heavy oil and bitumen. In *SPE International Improved Oil Recovery Conference in Asia Pacific*. Society of Petroleum Engineers, 2005.

[57] SMF Ali. Innovative steam injection techniques overcome adverse reservoir conditions. *Journal of Canadian Petroleum Technology*, 41(8):14–15, 2002.

[58] NR Edmunds, JA Kovalsky, SD Gittins, ED Pennacchioli, et al. Review of phase a steam-assisted gravity-drainage test. *SPE Reservoir Engineering*, 9(02):119–124, 1994.

[59] Tawfik Noaman Nasr and Ezra Eddy Isaacs. Process for enhancing hydrocarbon mobility using a steam additive, May 15 2001. US Patent 6,230,814.

[60] Roger M Butler. Thermal recovery of oil and bitumen. 1991.

[61] JP Fanaritis, Pa Warren, JD Kimmel, et al. Review of once-through steam generators. *Journal of Petroleum Technology*, 17(04):409–416, 1965.

[62] Li Xie, Yu Zhao, Daniel Aziz, Xing Jin, Litao Geng, Errol Goberdhansingh, Fei Qi, and Biao Huang. Soft sensors for online steam quality measurements of otsgs. *Journal of Process Control*, 23(7):990–1000, 2013.

[63] Adilson José de Assis et al. Soft sensors development for on-line bioreactor state estimation. *Computers & Chemical Engineering*, 24(2):1099–1103, 2000.

[64] Sungyong Park and Chonghun Han. A nonlinear soft sensor based on multivariate smoothing procedure for quality estimation in distillation columns. *Computers & Chemical Engineering*, 24(2):871–877, 2000.

[65] Manabu Kano and Yoshiaki Nakagawa. Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry. *Computers & Chemical Engineering*, 32(1):12–24, 2008.

[66] Roderick Murray-Smith and T Johansen. *Multiple Model Approaches to Nonlinear Modelling and Control*. CRC press, 1997.

[67] Kenneth L Lange, Roderick JA Little, and Jeremy MG Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

[68] Ingemar J Cox, Joe Kilian, Tom Leighton, and Talal Shamoon. A secure, robust watermark for multimedia. In *Information Hiding*, pages 185–206. Springer, 1996.

[69] André D Quelhas and José Carlos Pinto. Soft sensor models: Bias updating revisited. In *Proceedings of the International Symposium on Advanced Control of Chemical Processes 2009*, 2009.

[70] Shengjing Mu, Yingzhi Zeng, Ruilan Liu, Ping Wu, Hongye Su, and Jian Chu. Online dual updating with recursive pls model and its application in predicting crystal size of purified terephthalic acid (pta) process. *Journal of Process Control*, 16(6):557–566, 2006.

[71] Roger Ratcliff. Methods for dealing with reaction time outliers. *Psychological bulletin*, 114(3):510, 1993.

[72] Eli M Gafni and Dimitri P Bertsekas. Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization*, 22(6):936–964, 1984.

[73] Bao Lin, Bodil Recke, Jørgen KH Knudsen, and Sten Bay Jørgensen. A systematic approach for soft sensor development. *Computers & chemical engineering*, 31(5):419–425, 2007.

[74] Sanghong Kim, Manabu Kano, Hiroshi Nakagawa, and Shinji Hasebe. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *International journal of pharmaceutics*, 421(2):269–274, 2011.