AIChE

Robust Identification of Piecewise/Switching Autoregressive Exogenous Process

Xing Jin and Biao Huang
Dept. of Chemical and Materials Engineering, University of Alberta, Edmonton, AB, Canada, T6G 2G6

DOI 10.1002/aic.12112
Published online November 9, 2009 in Wiley InterScience (www.interscience.wiley.com).

A robust identification approach for a class of switching processes named PWARX (piecewise autoregressive exogenous) processes is developed in this article. It is proposed that the identification problem can be formulated and solved within the EM (expectation-maximization) algorithm framework. However, unlike the regular EM algorithm in which the objective function of the maximization step is built upon the assumption that the noise comes from a single distribution, contaminated Gaussian distribution is utilized in the process of constructing the objective function, which effectively makes the revised EM algorithm robust to the latent outliers. Issues associated with the EM algorithm in the PWARX system identification such as sensitivity to its starting point as well as inability to accurately classify ''un-decidable'' data points are examined and a solution strategy is proposed. Data sets with/without outliers are both considered and the performance is compared between the robust EM algorithm and regular EM algorithm in terms of their parameter estimation performance. Finally, a modified version of MRLP (multi-category robust linear programming) region partition method is proposed by assigning different weights to different data points. In this way, negative influence caused by outliers could be minimized in region partitioning of PWARX systems. Simulation as well as application on a pilot-scale switched process control system are used to verify the efficiency of the proposed identification algorithm.
 AIChE J, 56: 1829–1844, 2010
Keywords: process control, statistical analysis', system identification, robust EM, PWARX/switched system, hybrid system

Introduction

Hybrid systems are gaining increasing attention in process control research.[1–3] Hybrid systems are dynamic systems in which both continuous and discrete valued dynamics exist simultaneously. A number of examples on hybrid systems can be found in fields such as process control, embedded system, electrical circuits, biological process, and so on.[4] In chemical process industry, due to the inborn complexity of the chemical process as well as wide application of automated control systems, discrete behavior of the chemical process control system is commonly experienced.[5,6] The intricate interactions between continuous behavior (driven by the underlying physical laws such as mass and energy conversation) and the discrete events pose great challenges for both academic researchers and industrial practitioners. Motivated by the economic, safety, and environmental considerations, relevant researches on hybrid chemical process modeling.[7] optimization,[8,9] and control have been conducted.[5,6]

As an important subclass of hybrid systems, PWA (Piecewise affine) system is receiving a growing interest owing to its capability of describing a large number of processes by switching among different affine subsystems when state/input comes to a different region. Furthermore, equivalence of PWA systems to other forms of hybrid system models, such as mixed logical dynamical system, linear complementary systems, and max–min-plus-scaling systems has been proved

Correspondence concerning this article should be addressed to B. Huang at biao.huang@ualberta.ca.

under mild conditions,[10,11] which further stimulates the interest in the research of PWA systems. Provided that each affine subsystem is a linear ARX model, the PWA system can be considered as a PWARX system.

In the past few years, a number of methods for PWARX systems identification have been put forward. A data clustering based identification algorithm[12] has been proposed in which data clustering, linear identification, and region partition are performed together to identify PWA subsystems as well as valid region for each subsystem from input–output data. Juloski et al.[13] use a particle filter method to estimate local ARX model parameters by sequentially processing the input–output data. A bounded error method is developed to solve the identification problem.[14] After roughly classifying the data, a further refinement step is taken to get rid of "undecidable" and infeasible data points. Nakada et al.[15] apply a statistical clustering strategy to classify each data point to its relevant regions. An algebraic geometric approach is introduced for the identification of switched linear hybrid systems.[16] In this identification algorithm, the data classification procedure is set to be independent of the sub-model parameter estimation procedure by using hybrid decoupling constraint. Ragot et al.[17] suggest an adaptive weighting method for iteratively classifying the data points and estimation of local ARX sub-models. An optimization technique is introduced by Roll et al.[18] By considering hybrid system identification as a prediction error minimization problem, they use mixed integer programming to search for globally optimal identification solution, which becomes computationally infeasible when facing a large data set.

It can be seen from the analysis above that for the PWARX system identification, the main issue involved is how to identify each local ARX model when partition of regression space is controlled by unobservable variable and no relevant priori knowledge is at hand. It needs to be pointed out that Nakada et al.[15] have used the regular *EM* algorithm to cluster the data points by iteratively calculating the center of different clusters which, to some extent, is similar to the data classification process presented in Ferrari-Trecate et al.[12] by using "K-means" like algorithm. The distinguished features of the method developed in this article are (1) By treating the unobservable data point identity as "missing variable" and formulating the PWARX system identification within the *EM* algorithm framework, not only are the data points from different sub-models classified, but also the local ARX model parameters are estimated in the same time. This enables us to make full use of the convergence property of the *EM* algorithm within finite iteration steps. (2) Robustness of the *EM* algorithm is considered and a novel strategy for achieving this is proposed. It is not rare to find outliers in practical applications and the performance of the regular *EM* algorithm could deteriorate significantly in the presence of the statistical outliers.[19] Therefore, being robust to outliers is desirable as well as necessary for the *EM* algorithm.

In real applications, contamination of measurements by noise may lead to misclassification of data points which lie far away from region of the intersection area. For such misclassified data points, they can be viewed as outliers to the misattributed sub-models and accuracy of parameter estimation can be greatly deteriorated. For the other data points which are classified to the correct sub-model, once they are polluted with abnormal measurement or process noise, they still need to be removed from data sets for better parameter estimation. A robust parameter estimation method is applied by weighting every data point so that in local parameter estimation, smaller or even zero weights are given to outliers whereas normal data points are granted with much higher weight.

After finishing the data classification and sub-model identification, a robust linear programming for multi-category discrimination of polyhedral regions[20] is applied. This polyhedral region discrimination method has already been used in Juloski et al.[13] and different data points are weighted based on their likelihood to be undecidable. However, only considering those undecidable data points may not be sufficient as we have found that misclassification could also happen for data points which lie far away from intersection area. To reduce the effect of the outliers that may be present in data clusters, weights obtained in robust parameter estimation are used for each data point in the region partition stage.

When PWARX systems are considered, we assume that the switching of the system is triggered by different operating regions of the input and output although we do not know in advance how the region of the input and output is partitioned. However, if the switching mechanism of the system can not be represented by regressor space partition, the switched system would be regarded as switching along the time. This provides a more general way of treating the switched system. In this article, we also apply the proposed PWARX system identification algorithm to a simulated continuous fermenter as well as an experimental switched control system upon which two controllers with different characteristics operate alternatively. The identification results verify the validity of the proposed identification algorithm and show the potential of its usage in identification of various kinds of switched linear system.

The main contributions of this article are: (1) Formulate and solve the PWARX system identification problem under the *EM* algorithm framework. (2) A robust strategy is proposed for the *EM* algorithm. In the literature, some *EM* algorithms with different robustness strategies have already been suggested to handel the outliers.[19,21,22] However, the proposed strategy is based on the rigorous contaminated Gaussian distribution to describe the outliers, leading to an explicit weighted least square solution. The effectiveness of the robust procedure in resisting the abnormal data/outliers is demonstrated through comparison with the existing benchmarking method. (3) A computationally cost-efficient method is developed for the classification of un-decidable data points. (4) Evaluation of the proposed identification method is performed on a simulated continuous fermenter as well as a pilot-scale switched control system. The capability of the proposed PWARX system identification algorithm in handling switched linear systems is demonstrated.

The remainder of the article is organized as follows: (1) The PWARX system is formulated and several fundamental issues regarding the identification are explained through an example. (2) An overall introduction to the *EM* algorithm and formalizations for the identification of PWARX systems are obtained based on an improved version of the *EM*

algorithm. (3) An approach to initialize the *EM* algorithm together with a data classification refinement procedure is given. (4) Derivation of robust parameter estimation for each sub-model parameters with reclassified data points. (5) Effectiveness of the proposed algorithm through an illustrative simulation example. (6) Potential application of the proposed algorithm in chemical process operation monitoring and process modeling. (7) How a process control experiment is performed and the way the proposed PWARX system identification method is used in the analysis of the experimental data. (8) Conclusion.

Problem Statement

As an important subclass of PWA systems, a PWARX system is formulated as:[12,13,18]

$$y_k = \begin{cases} \theta_1^T \begin{bmatrix} x_k \\ 1 \end{bmatrix} + e_k, & x_k \in \chi_1 \\ \vdots & , \quad k = 1,2\ldots N \\ \theta_M^T \begin{bmatrix} x_k \\ 1 \end{bmatrix} + e_k, & x_k \in \chi_M \end{cases} \tag{1}$$

where N, M represent number of data points collected and number of sub-models respectively, $y_k \in R$ is the output, $x_k \in R^n$ is the regressor, which consists of past input and output,

$$x_k = \begin{bmatrix} y_{k-1} & y_{k-2} \cdots y_{k-na} & u_{k-1}^T & u_{k-2}^T \cdots u_{k-nb}^T \end{bmatrix}^T \tag{2}$$

where na and nb are orders of the output and input, $u \in R^m$ is the input and $n = na + m \cdot nb$. $e_k \in R$ is Gaussian distributed noise with zero mean and variance σ^2. $\theta_i \in R^{n+1}$ is the parameter vector of the ith sub-model and $\{\chi_i\}_{i=1}^M$ is the polyhedral region of the input–output space. Given the number of sub-model M, the PWARX system identification problem can be stated as:

Problem: *Assigning each data point in data set* (x_k, y_k),$k = 1,2\ldots N$ *to one of M sub ARX models and identifying the parameter vector* $\{\theta_i\}_{i=1}^M$ *along with polyhedral region* $\{\chi_i\}_{i=1}^M$ *for each local ARX model.*

It is assumed that the number of sub-models M and the order of each sub-model (ARX) are given a priori. In the case that the number of sub-models is unknown, there also exist methods to estimate it.[12,15] On the other hand, if the orders of the sub-ARX models are unknown, fixed high-order ARX models can be used since a sufficiently high-order ARX model can approximate any linear dynamic system.[23] This article will, however, focus on the robust identification problem for the PWARX models given the assumptions stated above.

Example 1. Consider the following bimodal PWARX example which is used in Juloski et al.,[13]

$$y_k = \begin{cases} \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} x_k \\ 1 \end{bmatrix} + e_k, & x_k \in [-2.5 \quad 0] \\ & , \quad k = 1,2\ldots N \\ \begin{bmatrix} -1 & 2 \end{bmatrix} \begin{bmatrix} x_k \\ 1 \end{bmatrix} + e_k, & x_k \in [0 \quad 2.5] \end{cases} \tag{3}$$

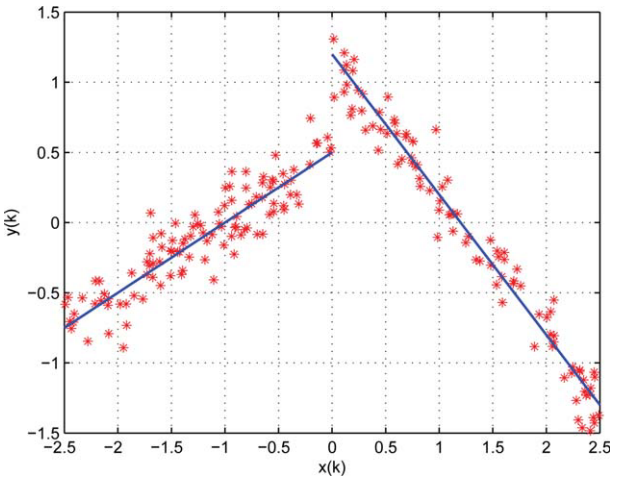

Figure 1. Data set generated by PWARX system (3).

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Let $N = 200$ and the data set (x_k, y_k), $k = 1,2\ldots200$ is generated by the bimodal PWARX system expressed by Eq. 3. $e_k \sim N(0,0.025)$ and input x_k follows a uniform distribution between -2.5 and 2.5. Figure 1 shows the data set.

From Figure 1, it can be seen that the polyhedral boundary that separates two regions is the y-axis given by $\{x = 0\}$ and the regressor x is a scalar in this specific case. We will use this simple example to illustrate the PWARX identification problem.

EM Algorithm

EM *algorithm revist*

Assume that a complete data set C consists of two parts: $\{C_{\text{obs}}, C_{\text{mis}}\}$, C_{obs} is the data collected from the process and it is called incomplete data set. C_{mis} needs to be estimated from C_{obs} and is called missing data set. The objective is to maximize the likelihood of the data collected in the incomplete data set C_{obs}[24]

$$L(C_{\text{obs}}, \Theta) = \int f(C|\Theta) dC_{\text{mis}} \tag{4}$$

with respect to parameter Θ and $f(\cdot)$ is the probability distribution function. To compute the maximum likelihood of Eq. 4, the *EM* algorithm takes two steps, *E*-step and *M*-step.

E-step uses the parameter estimation of previous iteration to compute the expectation of the complete data likelihood[24]

$$Q(\Theta|\Theta^{\text{old}}) = E_{C_{\text{mis}}|(\Theta^{\text{old}}, C_{\text{obs}})}\{\log L(C, \Theta)\} \tag{5}$$

where the conditional expectation is defined as $E_{A|B}\{g(A)\} \triangleq \int g(A) f(A|B) dA$ if A is a continuous random variable; if A is a discrete random variable the integration is replaced by summation. *M*-step maximizes the expectation shown in Eq. 5 with respect to Θ so as to ensure that the newly found Θ^{New} makes the log likelihood of the complete data set C non-decreasing, which equally means that

$$Q\left(\Theta^{\text{New}}|\Theta^{\text{old}}\right) \geq Q\left(\Theta|\Theta^{\text{old}}\right), \forall\Theta \tag{6}$$

Formulation of the PWARX system identification problem based on the EM *algorithm*

Regular EM *Algorithm.* Define $Z_k = \{x_k,\ y_k\}$, $k = 1,2\ldots N$ as the observed data set generated from a PWARX system. Therefore, for data Z_k, its conditional probability equals[25]

$$P(Zk|Z_{k-1\ldots}Z_1) = \sum_{i=1}^{M} \alpha_i P(Z_k|\theta_i, A_{k-1\ldots}Z_1) \tag{7}$$

where α_i is the probability that ith sub-model takes effect. As C_{obs} in the PWARX system is the observed dataset $Z_k = \{x_k, y_k\}$, $k = 1,2\ldots N$, the maximum likelihood equation for system parameter estimation is

$$\begin{aligned}\max_{\Theta} L(C_{\text{obs}}, \Theta) &= \max_{\Theta} P(C_{\text{obs}}|\Theta) \\ &= \max_{\theta_i, i=1\ldots M} \prod_{k=1}^{N} \sum_{i=1}^{M} \alpha_i P(Z_k|\theta_i, Z_{k-1\ldots}Z_1)\end{aligned} \tag{8}$$

To simplify the problem, rather than maximizing the likelihood function directly, one usually maximizes the log likelihood function,

$$\begin{aligned}\max_{\Theta} \log L(C_{\text{obs}}, \Theta) &= \max_{\Theta} \log P(C_{\text{obs}}|\Theta) \\ &= \max_{\theta_i, i=1\ldots M} \prod_{k=1}^{N} \log \sum_{i=1}^{M} \alpha_i P(Z_k|\theta_i, Z_{k-1}\cdots Z_1)\end{aligned} \tag{9}$$

The parameters may be estimated from Eq. 9 by brute force maximization, but this optimization is still difficult.

To make the problem tractable and solve the maximum likelihood estimation problem, we introduce $I = \{I_1, I_2,\ldots,I_N\}$ as a "missing variable" to denote the sub-model identity of each data point. Following Eq. 5, the expression for expectation of complete data $C = \{C_{\text{obs}}, I\}$ is:

$$\begin{aligned}&Q(\Theta|\Theta_{\text{old}}) \\ &\quad = E_{I|(\Theta^{\text{old}}, C_{\text{obs}})}\{\log P(C_{\text{obs}}, I|\Theta)\} \\ &\quad = E_{I|(\Theta^{\text{old}}, C_{\text{obs}})}\{\log P(Z_N, Z_{N-1}\ldots Z_1, I_N\ldots I_1|\Theta)\} \\ &\quad = E_{I|(\Theta^{\text{old}}, C_{\text{obs}})}\left\{\log \prod_{k=1}^{N} P(Z_k, I_k|Z_{k-1}, \ldots Z_1, I_{k-1}, \ldots I_1, \Theta)\right\} \\ &\quad = E_{I|(\Theta^{\text{old}}, C_{\text{obs}})}\left\{\log \prod_{k=1}^{N} P(Z_k|Z_{k-1}, \ldots Z_1, I_k, \ldots I_1, \Theta)P(I_k)\right\} \\ &\quad = E_{I(\Theta^{\text{old}}, C_{\text{obs}})}\left\{\sum_{k=1}^{N} \log\left[\alpha_{I_k} P\left(Z_k|\theta_{I_k}, A_{k-1}\ldots Z_1\right)\right]\right\}\end{aligned} \tag{10}$$

where α_{Ik} represents the probability that Z_k comes from the I_kth sub-model. Here, $I_k \in \{1,2\ldots M\}$ represents the true sub-model that Z_k comes from. In deriving Eq. 10, we have used the fact that $P(Z_k|Z_{k-1},\ldots,Z_1, I_k,\ldots,I_1,\Theta) = P(Z_k|Z_{k-1},\ldots,Z_1, I_k,\Theta)$ since I_k completely determines which sub-model that Z_k belongs to. We have also used the equation $P(I_k \mid Z_{k-1},\ldots Z_1, I_{k-1},\ldots I_1,\Theta) = P(I_k)$; namely switching between the sub-models is completely random and does not depend on which sub-model the system takes in previous instants.

By moving the Expectation operator inside the summation, Eq. 10 becomes

$$\begin{aligned}Q\left(\Theta|\Theta^{\text{old}}\right) &= \sum_{k=1}^{N} E_{I|(\Theta^{\text{old}}, C_{\text{obs}})}\{\log \alpha_{I_k} + \log P(Z_k|\theta_{I_k}, Z_{k-1}\ldots Z_1)\} \\ &= \sum_{k=1}^{N}\sum_{i=1}^{M} P\left(I_k = i|\Theta^{\text{old}}, C_{\text{obs}}\right) \log \alpha_i \\ &\quad + \sum_{k=1}^{N}\sum_{i=1}^{M} P\left(I_k = i|\Theta^{\text{old}}, C_{\text{obs}}\right) \log P(Z_k|\theta_i, Z_{k-1}\ldots Z_1)\end{aligned} \tag{11}$$

where $P(I_k = i \mid \Theta^{\text{old}}, C_{\text{obs}})$ in Eq. 11 denotes the probability that the kth data point comes from the ith sub-model, and can be derived following the Bayes rule as

$$\begin{aligned}&P\left(I_k = i|\Theta^{\text{old}}, C_{\text{obs}}\right) = P\left(I_k = i|\Theta^{\text{old}}, Z_k, Z_{k-1}, \ldots, Z_1\right) \\ &= \frac{P\left(Z_k|I_k = i, \Theta^{\text{old}}, Z_{k-1}, \ldots, Z_1\right)P\left(I_k = i|\Theta^{\text{old}}, Z_{k-1}, \ldots, Z_1\right)}{\sum_{i=1}^{M} P(Z_k|I_k = i, \Theta^{\text{old}}, Z_{k-1}, \ldots, Z_1)P(I_k = i|\Theta^{\text{old}}, Z_{k-1}, \ldots, Z_1)} \\ &= \frac{P\left(Z_k|\theta_i^{\text{old}}, Z_{k-1}, \ldots, Z_1\right)P(I_k = i)}{\sum_{i=1}^{M} P\left(Z_k|\theta_i^{\text{old}}, Z_{k-1}, \ldots, Z_1\right)P(I_k = i)} \\ &= \frac{\alpha_i P\left(Z_k|\theta_i^{\text{old}}, Z_{k-1}, \ldots Z_1\right)}{\sum_{i=1}^{M} P\left(Z_k|\theta_i^{\text{old}}, Z_{k-1}\ldots Z_1\right)\alpha_i}\end{aligned} \tag{12}$$

For notational simplicity, we will use $P_{k,i}$, namely the probability that the kth data point comes from the ith sub-model, to denote $P(I_k = i|\Theta^{\text{old}}, C_{\text{obs}})$ in the remainder of the article.

Let $\overline{x_k} = \begin{bmatrix} x_k \\ 1 \end{bmatrix}$, then for the PWARX system, $\log P(Z_k|\theta_i, Z_{k-1}\ldots Z_1)$ in Eq. 11 equals

$$\begin{aligned}&\log P(Z_k|\theta_i, Z_{k-1}\ldots Z_1) \\ &\quad = \log \frac{1}{\sqrt{2\pi}\sigma} - \exp^{-\frac{1}{2\sigma^2}\left(y_k - \theta_i^T\bar{x}\right)^T\left(y_k - \theta_i^T\bar{x}\right)} \\ &\quad = -\log\sqrt{2\pi}\sigma - \frac{1}{2\sigma^2}\left(y_k - \theta_i^T\bar{x}\right)^T\left(y_k - \theta_i^T\bar{x}\right)\end{aligned} \tag{13}$$

Therefore, substituting Eq. 13 of $\log P(Z_k|\theta_i, Z_{k-1}\ldots Z_1)$ in the second term on the right hand side of (11), after computing Eq. 12 by using previous estimation of θ_i^{old} and α_i^{old}, and then taking it into the right hand side of Eq. 11, we get the expectation of complete data $Q(\Theta|\Theta^{\text{old}})$.

For the M-step of the EM algorithm, derivatives are taken with respect to α_i, θ_i, and noise variance σ^2 in an effort to maximize the likelihood of the parameters given the observed data. After several steps of algebraic manipulations, expressions for θ_i, α_i, and σ^2 at each iteration can be derived as:

$$\theta_i^{\text{New}} = \frac{\sum_{k=1}^{N} P_{k,i}\bar{x}_k y_k}{\sum_{k=1}^{N} P_{k,i}\bar{x}_k\bar{x}_k^T} \tag{14}$$

$$\alpha_i^{\text{New}} = \frac{\sum_{k=1}^{N} P_{k,i}}{N} \tag{15}$$

$$\left(\sigma^{\text{New}}\right)^2 = \frac{\sum_{k=1}^{N}\sum_{i=1}^{M} P_{k,i}\left(y_k - \left(\theta_i^{\text{New}}\right)^T\bar{x}_k\right)^T\left(y_k - \left(\theta_i^{\text{New}}\right)^T\bar{x}_k\right)}{\sum_{k=1}^{N}\sum_{i=1}^{M} P_{k,i}} \tag{16}$$

$P_{k,i}$ is updated using value of $(\theta_i)^{\text{New}}$, $(\alpha_i)^{\text{New}}$, and $(\sigma^{\text{New}})^2$, and then the updated $P_{k,i}$ can be used for the calculation of *E*-step in the next iteration.

If we take a look back at Eq. 14, it can be found that this is a typical quadratic minimization problem and the method of weighted least squares has been used unconsciously to find the parameters $(\theta_i)^{\text{New}}$ for each local ARX model. As a matter of fact, Eq. 14 could be transferred into classic results in weighted least squares,

$$(\theta_i)^{\text{New}} = \left(\bar{X}^T W\bar{X}\right)^{-1}\bar{X}^T WY \tag{17}$$

where $\bar{X} = \begin{bmatrix} x_1 & \dots & x_N \\ 1 & \dots & 1 \end{bmatrix}_{(n+1,N)}$, $Y = \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}_{(N,1)}$,

$$W = \begin{bmatrix} P_{1,i} & 0 & \dots & 0 \\ 0 & P_{2,i} & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & P_{N,i} \end{bmatrix}_{(N,N)}$$

is weighting matrix for each data point with dimension (*N*,*N*). Therefore, for PWARX systems, the *EM* algorithm works as a combination of weight updating (which is *E*-step) and weighted least squares (which is *M*-step).

Robust EM *Algorithm*. It is noticed that in the maximization step of the regular *EM* algorithm for PWARX systems identification, an assumption that the residual errors follow the single normal distribution is made. The resulted maximization step of the regular *EM* algorithm is essentially an ordinary least squares procedure as shown in Eq. 17. However, if the data set contains outliers, which is not uncommon is real applications, ordinary least squares could fail and parameter estimation results obtained from it can be misleading.[26]

To offset the negative influence brought by outliers, the maximum likelihood objective function for parameter estimation is built based on a mixture distribution function[26–28] instead of single normal distribution as being used in the maximization step of the regular *EM* algorithm. This mixture distribution function is also known as a contaminated Gaussian distribution.

As a result, $P(Z_k|\theta_i, Z_{k-1}\dots Z_1)$ in Eq. 13 can be written as:

$$\begin{aligned} P((x_k,y_k)|\theta_i,(x_{k-1},y_{k-1})\dots(x_1,y_1)) &= P(e_k) \\ &= mP\left(e_k^{\text{regular}}\right) + (1-m)P\left(e_k^{\text{outlier}}\right) \end{aligned} \tag{18}$$

In Eq. 18, error e consists of two parts: error introduced by regular noise e_{regular} and error caused by irregular noise or outlier e_{outlier}, and m is the probability that the noise is the regular noise.

Setting that the ratio of noise variance between e_{outlier} and e_{regular} equals d^2 and $d \gg 1$, Eq. 18 can be further written as:

$$P(e_k) = m\frac{1}{\sqrt{2\pi}\sigma}\exp^{-0.5\frac{e_k^T e_k}{\sigma^2}} + \frac{1-m}{d}\frac{1}{\sqrt{2\pi}\sigma}\exp^{\frac{-0.5e_k^T e_k}{d^2\sigma^2}} \tag{19}$$

Take the logarithm of Eq. 19 and substitute it into the right hand side of Eq. 11. Again, for the *M*-step of the robust *EM* algorithm, derivatives are taken over θ_i, α_i, and noise variance σ^2 so as to ensure that Eq. 6 is always satisfied. After several steps of mathematical manipulation, the equation for calculating new θ_i is:

$$\theta_i = \frac{\sum_{k=1}^{N} P_{k,i}w_k\begin{bmatrix} x_k \\ 1 \end{bmatrix}y_k}{\sum_{k=1}^{N} w_k\begin{bmatrix} x_k \\ 1 \end{bmatrix}\begin{bmatrix} x_k \\ 1 \end{bmatrix}^T} = \left(X^T WX\right)^{-1}X^T WY \tag{20}$$

where in Eq. 20 $w_k = \frac{m\frac{P(e_k^{\text{regular}})}{\sigma^2} + (1-m)\frac{P(e_k^{\text{outlier}})}{d^2\sigma^2}}{mP(e_k^{\text{regular}}) + (1-m)P(e_k^{\text{outlier}})}$,

$$W = \begin{bmatrix} P_{1,i}w_1 & 0 & \dots & 0 \\ 0 & P_{2,i}w_2 & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & P_{N,i}w_N \end{bmatrix}_{(N,N)},$$

$$X = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}_{N,2} \text{ and } Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}_{N,1}.$$

From Eq. 20, it can be seen that for different data points, different weights are given based on their measurement noise as well as the probability of coming from the *i*th sub-model. Starting from an initial guess of the system parameters $\theta_i^{\text{initial}}, i = 1,2\dots M$, the robust *EM* algorithm iterates and usually converges within finite iterations.

The priori knowledge of m, d, and the covariance of data set σ^2 is not required since m and d can normally be set to between 0.7–0.9 and 10–15, respectively[27] and σ^2 has already been estimated in the *EM* algorithm using Eq. 16. Furthermore, Farris et al.[29] pointed out that the robust parameter estimation procedure is not very sensitive to the values of m and d.

Initialization of *EM* Algorithm and Refinement of Data Classification

Initialization of EM *algorithm*

Even though it is guaranteed that *EM* algorithm can converge to a stationary point after steps of iteration, there is no guarantee that this stationary point is a global or even local maxima. It may turn out to be a saddle point, and starting point plays an important role in determining what the convergence point of the *EM* algorithm would be.[25,30]

Table 1. Comparison of Identified Results with Different Staring Points

Starting Point	Identified Parameters
$\theta_{\text{ini1}} = \begin{bmatrix} 0.800 & -0.839 \\ -0.116 & -0.156 \end{bmatrix}$	$\theta_{\text{iden1}} = \begin{bmatrix} -0.998 & 1.999 \\ 0.517 & 0.521 \end{bmatrix}$
$\theta_{\text{ini2}} = \begin{bmatrix} 0.385 & 0.922 \\ -0.919 & 0.059 \end{bmatrix}$	$\theta_{\text{iden2}} = \begin{bmatrix} 0.770 & 1.013 \\ 0.117 & 0.032 \end{bmatrix}$

To demonstrate the importance of the starting point for the *EM* algorithm, by using equations obtained in the next section, Example 1 is identified using the *EM* algorithm with different starting points.

The true parameters of the PWARX system are: $\theta_{\text{true}} = \begin{bmatrix} \theta_1^T \\ \theta_2^T \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 0.5 & 0.5 \end{bmatrix}$. It is shown in Table 1 that the *EM* algorithm could converge to totally different results under different starting points. This uncertainty in the performance of the *EM* algorithm introduced by the starting point has to be resolved as dependency on the initial conditions is undesirable for an identification algorithm.

Among various initializing strategies, pre-running of the *EM* algorithm for several times with randomly selected starting points before entering the main *EM* step is most widely used owing to its simplicity and efficiency. After taking h trials for pre-running, the one with the largest likelihood would be chosen as the initial condition for the main *EM* algorithm. To balance the computation time spent on the initialization step and goodness of the initial condition obtained, the stopping condition has to be appropriately defined:[30]

$$\frac{L(C,\Theta)^m - L(C,\Theta)^{m-1}}{L(C,\Theta)^m - L(C,\Theta)^1} \le l \tag{21}$$

where $L(C,\Theta)$ denotes the complete data likelihood and l is a stopping value which can be tuned.

Refinement of data classification result

A classification refinement procedure was proposed to deal with "un-decidable" data points by Bempoard et al.[14] It uses bounded error along with spatial location information of "un-decidable" data points to achieve data reclassification. However, given tens of thousands of data points, we found that it is computationally formidable by searching through every data point instead of focusing on only those "un-decidable" data points. Hence, in this article, "certainty" of each data point is defined based on the probability value $P_{k,i}$, $k = 1,2\ldots N$, $i = 1,2\ldots M$ obtained in the *EM* algorithm and "un-decidable" data points are determined in terms of their "certainty" level.

Data clustering in the *EM* algorithm is realized by comparing $P_{k,i}$, $i = 1,2\ldots M$ for the kth data and finding the largest $P_{k,i}$ among all possible modes.

$$\text{mod}\, e_k = \arg\max_i P_{k,i}, k = 1,2,\ldots N, i = 1,2,\ldots M \tag{22}$$

Through Eq. 22, the mode that each data point belongs to can be determined.

However, the data point such as those displayed in Figure 2, which lie in the grayed zone, may be classified to either of the two sub-models, and most of time, the classification result can be quite random as it is greatly influenced by the noise level. Hence, for the data points lying in the interaction area among several sub-models, they are normally "un-decidable" and misclassification can happen with a great chance.

Therefore, a refinement is needed to deal with those "un-decidable" data points. Here, we propose to use "certainty" level of each data point to denote the probability that a data point is "un-decidable." The lower the "certainty" is, the more likely the data point is "un-decidable." The certainty of the kth data point is defined as,

$$\text{Certainty}(Z_k) = \left|\log\frac{P_{k,i_{\min}}}{P_{k,j_{\min}}}\right| \tag{23}$$

$$(i_{\min}, j_{\min}) = \arg\min_{i,j}\left|\log\frac{P_{k,i}}{P_{k,j}}\right|, i = 1,2\ldots M-1 \quad j = i+1,\ldots M \tag{24}$$

In Eq. 24, the probability of a data point belonging to each sub-model is compared and a pair of sub-models $\{i_{\min}, j_{\min}\}$ with closest probabilities are picked. When probability between this couple of sub-models is equal, which implies that the data can be attributed to either sub-model, the "certainty" value obtained through Eq. 23 will be zero. Hence, in this way, "un-decidable" data points can be sorted out effectively using this "certainty" metric so that in the following classification refinement procedure, only those "un-decidable" data with low "certainty" levels are considered. The procedure for the refinement of "un-decidable" data points classification result is:

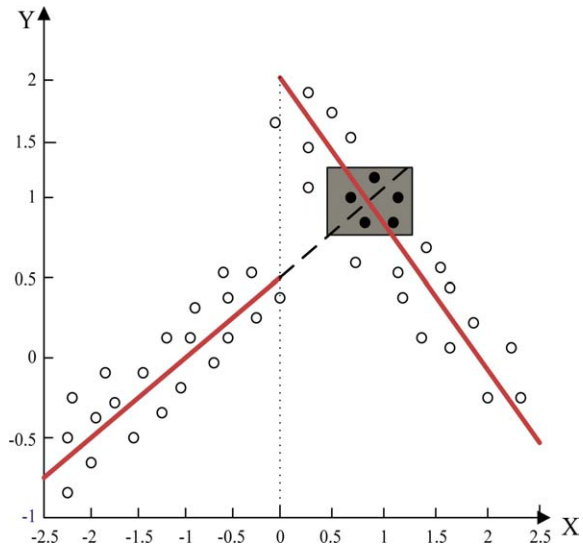

Figure 2. An explanation figure for "un-decidable" data points generated by the bimodal PWARX system in Example 1.[31]

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

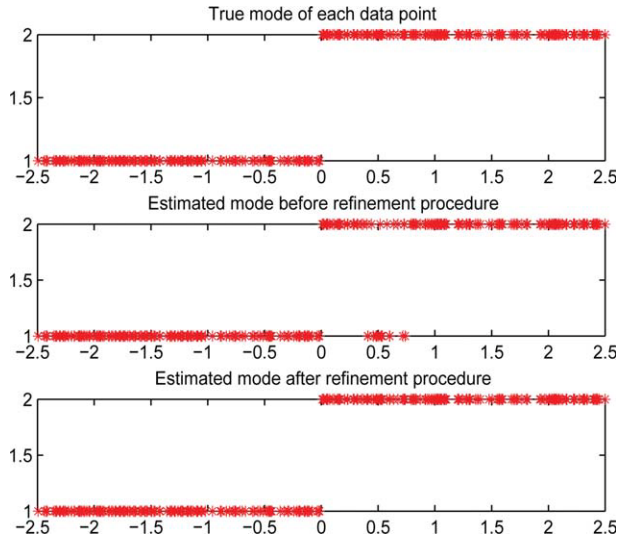

Figure 3. Comparison of data classification results before and after refinement procedure for Example 1.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Step 1. Check "certainty" level of the kth data point, if it is lower than some specified level lb, then go to step 2; otherwise check $(k + 1)$th data until it reaches N.

Step 2. Collect p data points that are closest to the kth data point in terms of the Euclidean distance.

Step 3. For these $(p + 1)$ data points, label them as $q = 1,2\ldots p + 1$, the probabilities $P_{q,i}$ of those data points classified to the same ith sub-model are added up, and the sub-model with the highest sum is taken as the new sub-model identity for this kth data point. This is based on the similarity principle. Namely, a data point that has low certainty value is classified to the same sub-model as its closest decidable neighbors tend to belong.

Step 4. If $k < N$, go back to step 1; otherwise stop and exit the refinement procedure.

As an illustration, set $lb = 2$ and $p = 0.1 * N$, the refinement procedure is applied to Example 1 to show its effectiveness in correcting the misclassified "un-decidable" data.

As we can see from Figure 3, some misclassifications have occurred in the intersection area between two sub-models, and the refinement procedure can reclassify those misclassified data to the correct mode by exploring their spatial information.

Robust Parameter Re-Estimation of Local Models and Region Partition

Robust parameter re-estimation of each local ARX model

The reclassification procedure introduced above can effectively reduce the chance of misclassification for the "un-decidable" data points located in the intersection areas of different sub-models by exploring the spatial information of those data points. This provides us an opportunity that by applying linear regression techniques such as ordinary least squares to each reclassified cluster, more accurate parameter estimation results could be obtained. However, if the data set of a local ARX model contains outliers, as pointed out earlier in the robust *EM* algorithm section, ordinary least squares can suffer. These outliers mainly come from two sources:

1. Measurements of process variables may be corrupted by occasional outliers such as noise spikes that come along with normal noise. These data may be clustered in a correct group but they deviate significantly from the rest of data within the same group.

2. Occurrence of misclassification which mistakenly assigns data points to sub-models other than the one they truly belong to.

Therefore, it is necessary to use another robust parameter estimation method so as to immunize the parameter estimation procedure from the outliers in each clustered data set. Here, once again, contaminated Gaussian distribution is utilized as the assumption for the residual errors and it is expected that by granting small weights to the abnormal data points, the identified local ARX models can better represent the dynamics of the system.

Let $Z_{n,i} = \{x_{n,i}, y_{n,i}\}$, $i = 1,2\ldots M$, $n = 1,2\ldots N_i$, where N_i represents the number of data points assigned to the ith sub-model. Assume that the sampling instant of $Z_{n,i}$ is m_n, following the same approach as being introduced in the other section, the likelihood function would be:

$$J_i = \prod_{n=1}^{N_i} P(Z_{m_n}|\theta_i, Z_{m_n-1}, Z_{M_n-2}\ldots Z_1) = \prod_{n=1}^{N_i} P(x_{m_n}, y_{m_n})|\theta_i,(x_{m_n-1}, y_{m_n-1})\ldots(x_1, y_1)) \quad (25)$$

For an ARX model, we have

$$y_{n,i} = \theta_i^T \begin{bmatrix} x_{n,i} \\ 1 \end{bmatrix} + e_{n,i}, i = 1,2\ldots M, n = 1,2\ldots N_i \quad (26)$$

Considering that the noise $e_{n,i}$ follows the contaminated Gaussian distribution which consists of regular noise e_{regular} and irregular noise e_{outlier}, following the same way as having been done in the M-step of the robust *EM* algorithm, the expression for θ_i is:

$$\theta_i = \frac{\sum_{n=1}^{N_i} w_{n,i} \begin{bmatrix} x_{n,i} \\ 1 \end{bmatrix} y_{n,i}}{\sum_{n=1}^{N_i} w_{n-i} \begin{bmatrix} x_{n,i} \\ 1 \end{bmatrix} \begin{bmatrix} x_{n,i} \\ 1 \end{bmatrix}^T} = \left(X_i^T W_i X_i\right)^{-1} X_i^T W_i Y_i \quad (27)$$

where

$w_{n,i} = \frac{m\frac{P(e_{n,i}^{\text{regular}})}{\sigma^2} + (1-m)\frac{P(e_{n,i}^{\text{outlier}})}{d^2\sigma^2}}{mP(e_{n,i}^{\text{regular}}) + (1-m)P(e_{n,i}^{\text{outlier}})}$, $X_i = \begin{bmatrix} x_{1,i} & 1 \\ \vdots & \vdots \\ x_{N_i,1} & 1 \end{bmatrix}_{N_i,2}$,

$Y_i = \begin{bmatrix} y_{1,i} \\ \vdots \\ y_{N_i,i} \end{bmatrix}_{N_i,1}$, $W_i = \begin{bmatrix} w_{1,i} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{N_i}, i \end{bmatrix}_{N_i,N_i}$. If error $e_{n,i}$ satisfies[26]

$$mP\left(e_{n,i}^{\text{regular}}\right) > (1-m)P\left(e_{n,i}^{\text{outlier}}\right) \quad (28)$$

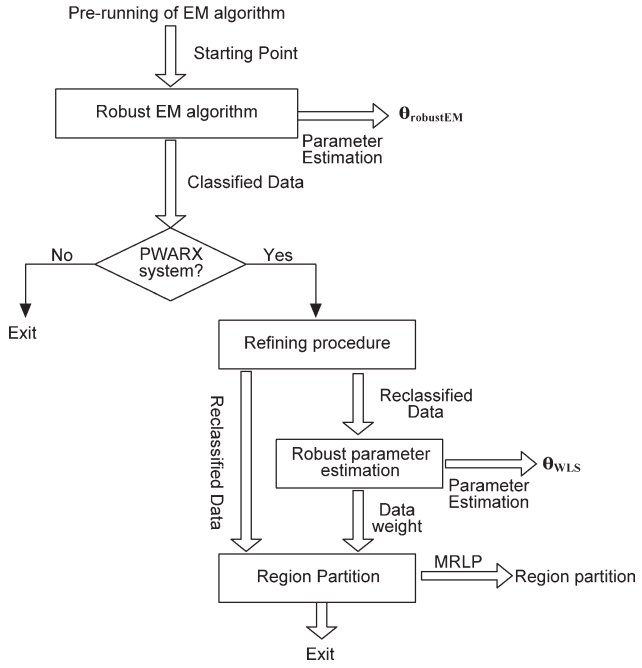

Figure 4. Framework of robust PWARX system identification using *EM* algorithm.

$Z_{n,i}$ can be treated as a normal data point and a relatively high weight $w_{n,i}$ is assigned to it automatically.

Region partition

In addition to data clustering and local ARX model parameter estimation, polyhedral region $\{\chi\}_{i=1}^{M}$ in which each sub-model takes effect should also be estimated to complete the identification of a PWARX system. Here, we use the MRLP algorithm which is introduced by Bennett et al.[20] The rationale behind this method is that for the data point $Z_{n,i} = \{x_{n,i}, y_{n,i}\}$, $n = 1,2,\ldots N_i$ that belongs to region i, the following Equation always holds:[20]

$$x_{n,i}(\omega_i - \omega_j) > \gamma_i - \gamma_j, \quad i,j = 1.2\ldots M, \quad i \neq j \tag{29}$$

or equivalently

$$x_{n,i}(\omega_i - \omega_j) \geq \gamma_i - \gamma_j + 1, \quad i,j = 1.2\ldots M, \quad i \neq j \tag{30}$$

where ω_i, ω_j, γ_i, and γ_j are parameters of hyperplane that separates region χ_i and χ_j. The hyperplane that separates χ_i and χ_j satisfies $x_{n,i}(\omega_i - \omega_j) = \gamma_i - \gamma_j$. Define

$$G_n^{i,j} = \max\left(-x_{n,i}(\omega_i - \omega_j) + (\gamma_i - \gamma_j) + 1, 0\right) \tag{31}$$

Then, a linear objective function can be defined as:

$$J_{rp} = \min_{\omega_i,\gamma_i} \sum_{i=1}^{M} \sum_{j=1,i\neq j}^{M} \sum_{n=1}^{N_i} G_n^{i,j} \tag{32}$$

where N_i denotes the number of data points classified to the ith sub-model. Normally, $G_n^{i,j}$ should always be zero when Eq. 32 holds. However, if misclassification of data occurs, Eq. 31 or 32 will be violated, which means that $G_n^{i,j}$ is larger than zero. As a result, the accuracy of region partition results from linear minimization Eq. 32 could be reduced because of those misclassified data points.

As discussed in the robust parameter estimation section, weights for data points that are suspicious of being outliers are small or even equal to 0. These weights can be used to weight each data point in Eq. 32. Therefore, a new objective equation would be:

$$J_{rp} = \min_{\omega_i,\gamma_i} \sum_{i=1}^{M} \sum_{j=1,i\neq j}^{M} \sum_{n=1}^{N_i} w_{n,i} G_n^{i,j} \tag{33}$$

To summarize, the complete framework of the algorithm introduced in this article is shown as Figure 4.

Returning to Example 1, the initial variance σ_{ini}^2 is arbitrarily chosen as 0.04. On the basis of the recommendation given in Biernacki et al.,[30] the stopping condition l and pre-running times h are set to be 0.01 and 10, respectively. For lb and p, we find that, in general, the un-decidable data points can be effectively detected and classified with the choice of $lb = 2$ and $p = 0.1 * N$. Applying the proposed algorithm to Example 1, we obtain the results shown in Table 2.

$\theta_{\text{regular}EM}$ and $\theta_{\text{robust}EM}$ in Table 2 are the parameter estimation results from the regular *EM* algorithm, robust *EM* algorithm respectively. θ_{WLS} is obtained by applying the robust parameter estimation procedure to the reclassified data clusters. The final estimation of variance $\hat{\sigma}^2 = 0.0278$ while true data variance equals $\sigma^2 = 0.025$. As no outliers are added in the data set, the estimation result from the regular *EM* algorithm is similar to its robust counterpart as well as robust parameter estimation procedure. The region partition result is $x = -0.00258$.

A More Complex Simulation Example

Consider the following PWARX system with three sub-models:[14,15]

$$y_k = \begin{cases} [-0.4 \quad 1 \quad 1.5]\begin{bmatrix} x_k \\ 1 \end{bmatrix} + e_k, & x_k \in \chi_1 \\ [0.5 \quad -1 \quad -0.5]\begin{bmatrix} x_k \\ 1 \end{bmatrix} + e_k, & x_k \in \chi_2, \quad k = 1,2\ldots N \\ [-0.3 \quad 0.5 \quad -1.7]\begin{bmatrix} x_k \\ 1 \end{bmatrix} + e_k, & x_k \in \chi_3 \end{cases} \tag{34}$$

where the regressor $x_k = \begin{bmatrix} y_{(k-1)} \\ u_{(k-1)} \end{bmatrix}$. Region partition is given by

Table 2. True Bimodal PWARX System Parameters and Estimated Parameters

θ_1	$\theta_{1\text{regular}EM}$	$\theta_{1\text{robust}EM}$	$\theta_{1\text{WLS}}$	θ_2	$\theta_{2\text{regular}EM}$	$\theta_{2\text{robust}EM}$	$\theta_{2\text{WLS}}$
−1	−0.9845	−0.9886	−0.9886	0.5	0.5047	0.5115	0.5115
2	2.0418	2.0392	2.0392	0.5	0.5435	0.5541	0.5540

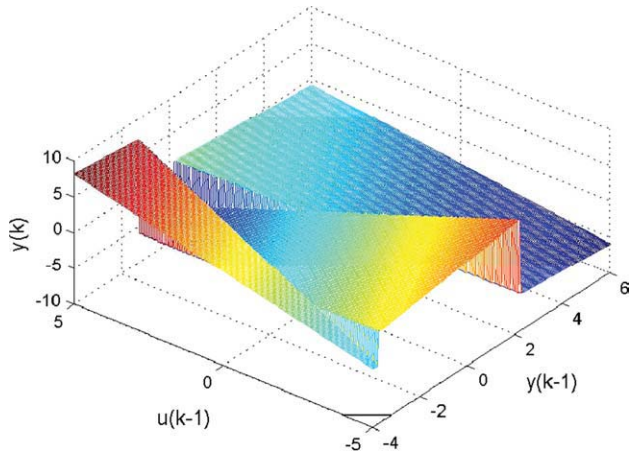

Figure 5. Hyperplane of the PWARX system.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$\chi_1 = \{[4 \quad -1]x + 10 < 0$$
$$\chi_2 = \left\{ \begin{matrix} [-4 & 1]x - 10 \le 0 \\ [5 & -1]x - 6 \le 0 \end{matrix} \right\}$$
$$\chi_3 = \{[-5 \quad -1]x + 6 \le 0\}$$

Hence, the hyperplane of the PWARX system is shown in Figure 5.

To test the capability of the algorithm in handling outliers, certain percentage of outliers are added to a data set generated by the PWARX system. Set total number of data points $N = 300$ in which around 10% are outliers. Regular Normal distributed noise $e_k \sim N(0,0.05)$ is chosen while initial guess of noise variance σ^2_{ini} is arbitrarily set 0.03. The input u_k follows a uniform distribution between -5 and 5. Let $lb = 2$, $p = 0.1 * N$, $l = 0.01$, $h = 10$, $m = 0.9$, and $d^2 = 60$, then apply the algorithm to the data set and data classification result is shown in Figure 6.

In Figure 6, we can see that before the refinement procedure, some misclassification occurs at the intersection of different sub-models (those "floating" circle points) and it is partially eliminated after the refinement procedure. Parameter estimation results from different estimation methods are listed in Table 3. where θ_i, $i = 1,2,3$ are the real parameters of its corresponding sub-model, respectively. The subscript regular *EM* stands for the parameters estimated from the regular *EM* algorithm; subscript robust *EM* represents the parameter estimation results obtained from the robust *EM* algorithm; θ_{WLS} denotes the identification results after applying the robust parameter estimation procedure to the reclassified data points. As can be seen in Table 3, the estimation from the regular *EM* algorithm for some parameters is greatly influenced by outliers while, on the other hand, the robust *EM* algorithm $\theta_{\text{robust}EM}$ and the robust parameter estimation

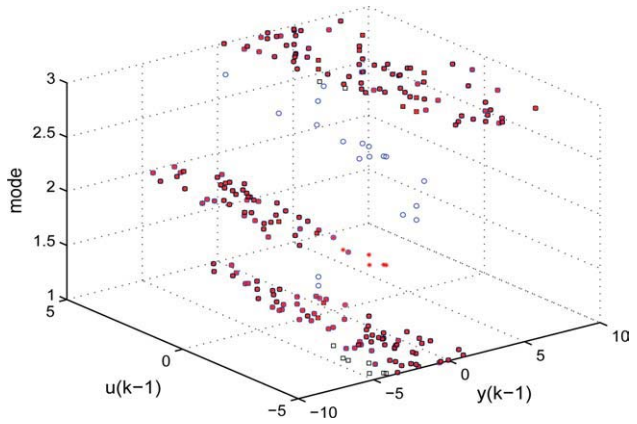

Figure 6. Comparison of data classification result before and after refinement procedure (Cross points denote the true mode of each data point, circle points represent classification results before refinement, and square ones are the results after classification refinement).

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

θ_{WLS} exhibit sufficient robustness to resist the negative influence brought by outliers. However, from the estimation results of θ_1, in which robust parameter estimation procedure outperforms the robust *EM* algorithm in terms of parameter estimation accuracy, it can be seen that the classification refinement procedure can effectively improve the performance of the parameter estimation.

Finally, transfer the weights obtained in robust parameter estimation to the region estimation procedure, the regressor partition results are obtained as:

The estimated hyperplane between χ_1 and χ_2 is $\hat{\lambda}_{12}x + 1 = 0$, $\hat{\lambda}_{12} = [0.33 \quad -0.1087]$ while χ_2 and χ_3 are separated by the hyperplane $\hat{\lambda}_{23}x + 1 = 0$, $\hat{\lambda}_{23} = [-0.823 \quad -0.152]$. It is noticed from Eq. 34 that the true hyperplane that separates χ_1 and χ_2 is $\lambda_{12}x + 1 = 0$, $\lambda_{12} = [0.4 \quad -0.1]$, and χ_2, χ_3 are separated by $\lambda_{23}x + 1 = 0$, $\lambda_{23} = [-0.83 \quad -0.167]$.

Application Example: Continuous Fermentation Reactor

Discontinuous behavior of the chemical process may originate from its inherent physicochemical discontinuities (such as phase changes, flow reversals, and shocks) or some interventions/disturbances from outside world (such as human or controller manipulation, changes in raw materials). These discrete events drive the system traveling among various operating regimes featured by different dynamics. In optimal

Table 3. Comparison of Estimated Parameters Using Different Methods with Outliers in the Data Set

θ_1	$\theta_{1\text{regular}EM}$	$\theta_{1\text{robust}EM}$	$\theta_{1\text{WLS}}$	θ_2	$\theta_{2\text{regular}EM}$	$\theta_{2\text{robust}EM}$	$\theta_{2\text{WLS}}$	θ_3	$\theta_{3\text{regular}EM}$	$\theta_{3\text{robust}EM}$	$\theta_{3\text{WLS}}$
−0.4	−0.538	−0.516	−0.405	0.5	0.3	0.442	0.454	−0.3	−0.229	−0.302	−0.296
1	1.028	1.021	1.017	−1	−0.988	−1.0	−0.997	0.5	0.508	0.486	0.498
1.5	0.904	0.972	1.397	−0.5	−0.511	−0.524	−0.514	−1.7	−2.049	−1.657	−1.687

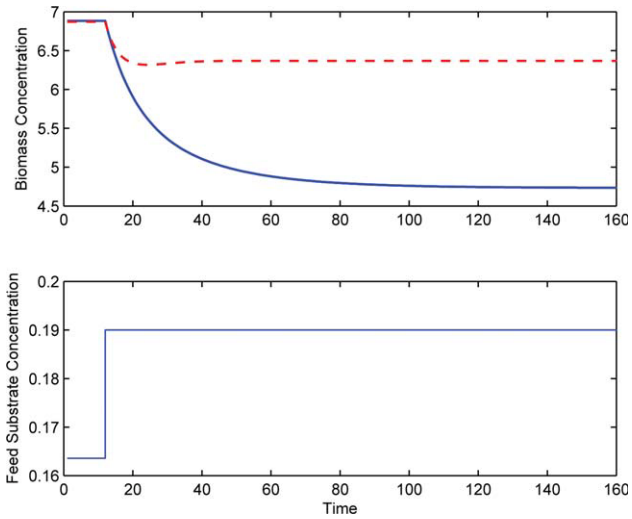

Figure 7. Input–Output data of the continuous fermenter.

Dash line represents the step response when $u_1 = 20$, solid line denotes the step response when $u_1 = 26.5$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

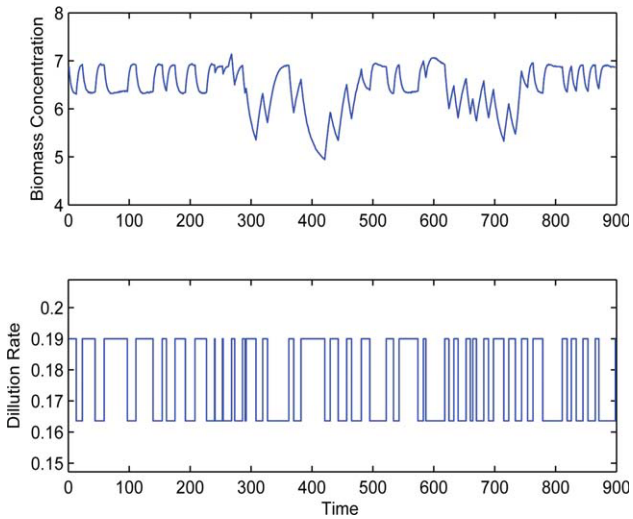

Figure 8. Input–Output data of the continuous fermenter.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

control of plant grade transition, the governing process dynamics during the plant transition period can vary dramatically and as a result, the performance of the model based controller may deteriorate greatly if only one single local model is utilized. Hence, being able to detect the changes of process dynamics and identify the relevant local models under different operating regimes could be critical in some applications.

To illustrate potential applications of the proposed identification method for chemical process, a continuous fermentation reactor is given in this section.[32–34] The fermenter consists of two inputs and three outputs, namely feed substrate concentration (u_1), dilution rate (u_2), and biomass concentration (y_1), substrate concentration (y_2), production concentration (y_3). Following the same parameter setting and normal operating conditions given by Gugaliya et al.,[33] a single input, single output (SISO) model between dilution rate (u_2) and biomass concentration (y_1) is identified. To ensure the local linearity of the identified model, a random binary signal with appropriate amplitude (from 0.1636 h^{-1} to 0.19 h^{-1}) is designed for the input u_2. Assume that the feed substrate concentration u_1 fluctuates between its nominal value 20 kg/m^3 and 26.5 kg/m^3 and cannot be tracked in a timely manner, the process dynamics under different u_1 values are compared and the step responses are shown in Figure 7.

As can be seen from Figure 7, the dynamics of the model under different feed substrate concentration values exhibit different characteristics, including process gain, time constant, as well as steady state values. This indicates the necessity of detecting and identifying the switching process model to achieve satisfactory control performance. Let the feed substrate concentration change between 20 kg/m^3 and 26.5 kg/m^3 randomly, adding white noise which is about 1% of the noise-free output in power. The obtained output and input data are shown in Figue 8.

In this case, the switching frequency of the input signal is determined based on the minimum time constant of all the sub-models. Distinct dynamics coexist in the output data set shown in Figure 8 owing to the change of feed substrate concentration. Passing the data set through the proposed algorithm, the identified models as well as self-validation results are shown in Table 4 and Figure 9.

Because of the discrepancy of process behavior under different feed substrate concentration values (20 kg/m^3 and 26.5 kg/m^3 in this case), we are able to classify the data point at each instance to its clusters featured with different u_2 values. As a result, we can detect the grade change of the raw materials by analyzing and clustering the data. For the data set shown in Figure 8, 80% of the data points have been classified to the right mode, which equally means that we may be able to indirectly infer the value of feed substrate concentration correctly during 80% of the operating time.

To further validate the identified models, cross validation for both identified models is performed using newly generated data set under different u_2 values. The validation results are given in Figures 10 and 11.

Satisfactory cross validation results demonstrate the capability of the proposed identification algorithm in handling the process which exhibits various distinct dynamics owing to the influences from diverse sources. Detection of the process change not only facilitates the monitoring of the process operation by informing us with happening of dramatic change in operating condition (feed substrate concentration change in this case), but also provides us with deeper perspective on the process itself during the operation (such as

Table 4. Identified Models Under Different Feed Substrate Concentrations

Feed Substrate Concentration	Identified Model from u_2 to y_1
20 kg/m^3	$y_k = 1.6550y_{k-1} - 0.6842y_{k-2} - 6.4345u_{k-1} + 6.0230u_{k-2} + 0.2676$
26.5 kg/m^3	$y_k = 1.9010y_{k-1} - 0.9034y_{k-2} - 5.7619u_{k-1} + 5.6314u_{k-2} + 0.0374$

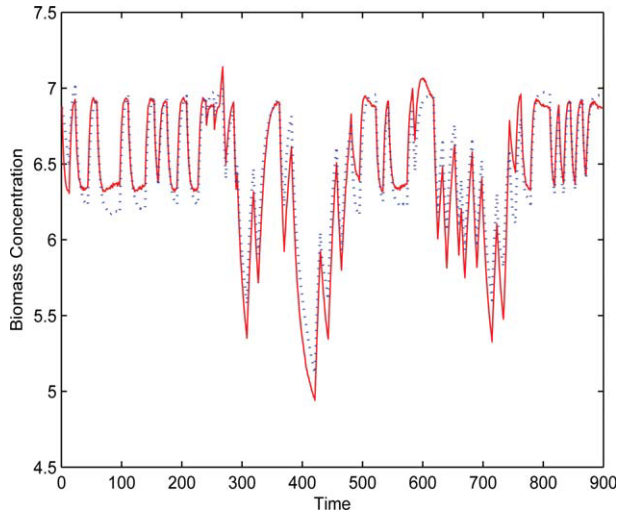

Figure 9. Comparison between model prediction and the true data, self validation MSE = 0.0155.

Dotted line represents prediction from identified models, Solid line denotes the true data from the fermenter. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

whether phase change happens). Moreover, multiple local models identified through the algorithm can also be used by model-based optimal controllers so that decent control performance could always be achieved regardless of the switching of process dynamics.

Identification of an Experimental Switched Process Control System

As a typical form of switched linear systems, the PWARX system switches because its input and output come to different domains as mathematically formulated in Eq. 1. Provided

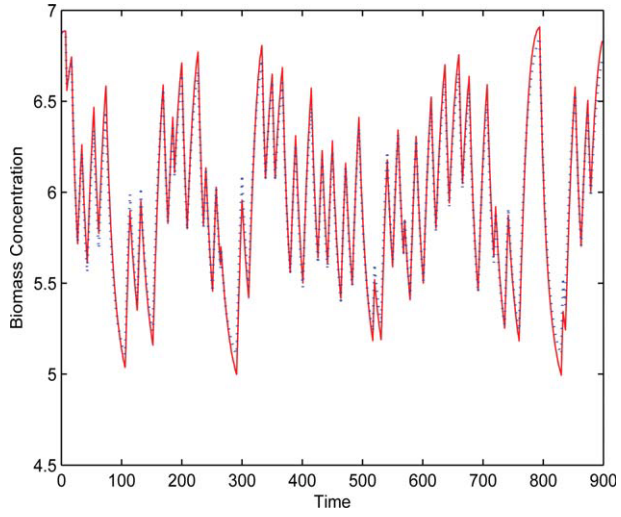

Figure 10. Comparison between model prediction and the true data when u_2 = 26.5 kg/m^3, cross validation MSE = 0.008.

Dashed line represents prediction from identified models, Solid line denotes the true data from the fermenter. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

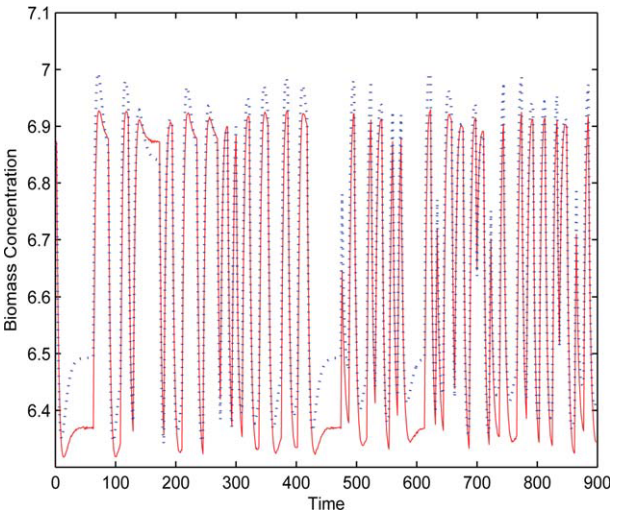

Figure 11. Comparison between model prediction and the true data when u_2 = 20 kg/m^3, cross validation MSE = 0.00366.

Dashed line represents prediction from identified models, Solid line denotes the true data from the fermenter. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

that no priori knowledge regarding the triggering signal of the switched system is at hand, time could be considered responsible for the switching of the system. This grants us a general way of treating the switched linear system and the proposed PWARX system identification algorithm can be applied to such a switched system without need of further modification.

The tank system is a pilot scale setup on which two switching level controllers with different characteristics are installed in an effort to maintain the level of the tank. The schematic diagram is shown in Figure 12.

By controlling the speed of the pump, the level controller can manipulate the inlet water flow rate of the tank so as to control its level. The closed loop step response under each of the two level controllers is given in Figure 13.

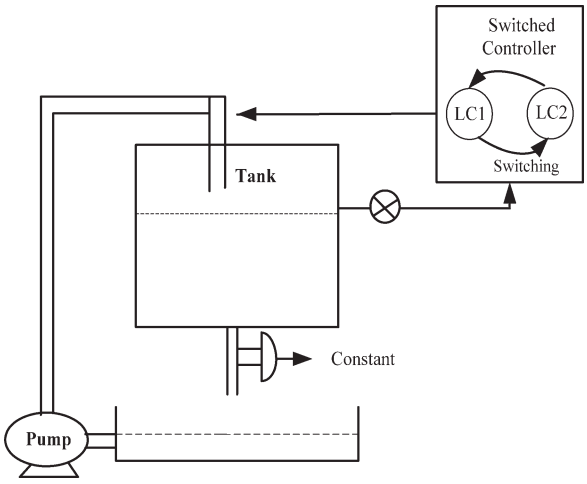

Figure 12. Schematic diagram of the tank system with switched controller.

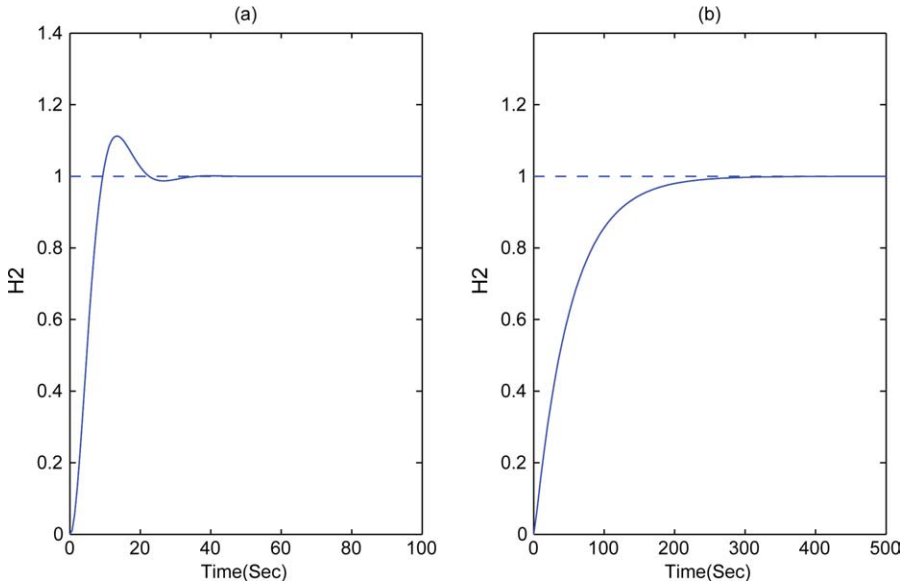

Figure 13. (a) Step response of controller 1 (b) step response of controller 2.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

It can be seen that both controllers do not have any offset in tracking the setpoint, which equally means that the steady state gains of the control loop under both controllers are one. However, apparent difference is observed in terms of their transient response to the setpoint change; in other words, the dynamics of the closed loop process under different controllers are different. Hence, the control loop operates under two modes.

Mode 1.

The loop is running on level controller 1, which enables the controller to response aggressively to any setpoint change or unknown disturbance with certain amount of overshoot.

Mode 2.

The loop is running on the sluggish level controller 2, which makes the controller act slowly in setpoint tracking or disturbance rejection. There is no overshoot under level controller 2.

For each time point, the switched controller may randomly reside on one of the two level controllers with certain probability, resulting two-mode closed-loop operations. Without knowing how the controller is switched internally, it is expected that the proposed system identification algorithm can separate the bimodal process with different dynamics and flag the identity of the controller on which the system is running at each sampling instance in the presence of the disturbances.

As shown in Figure 14, the setpoint of the closed loop is treated as the input of the process and the water level of the tank is considered as the output, the proposed identification method can be performed after collecting the input–output data of the closed-loop process. Cases with/without outliers in the experimental data are considered. According to the algorithm framework shown in Figure 4, the robust parameter estimation for each local clustered data set θ_{WLS} is calculated only when the refinement of the classification result from the robust *EM* algorithm is achieved. Because of the fact that the (time) switched system does not have input/output dependent partition, the region partition procedure will not be performed. Only the identified models from the regular *EM* algorithm $\theta_{\mathrm{regular}EM}$ and the robust *EM* algorithm $\theta_{\mathrm{robust}EM}$ are compared to demonstrate the difference between robust and regular (non-robust) algorithm, which is the focusing point of this article.

Case 1: Experimental data without outliers

The model order for both sub-models of the bimodal closed-loop process can be deduced from the analysis of the process based on the physical knowledge of the system, the number of sub-models for the multi-model identification algorithm is known as 2. After running an experiment, the collected input–output data shown in Figure 15 are separated into two parts, one is for estimation of each local ARX model parameters whereas the other one is used for cross validating the identified models obtained from training data set.

The hidden variable that represents the identity of the controller mode on which the system is running at each sampling time point needs to be estimated and the accuracy of its estimation directly influences the quality of the identified models. Given the estimation of the hidden variable $\hat{I}_k$ for the kth sampling time point, each data point can be clustered into one of the two groups.

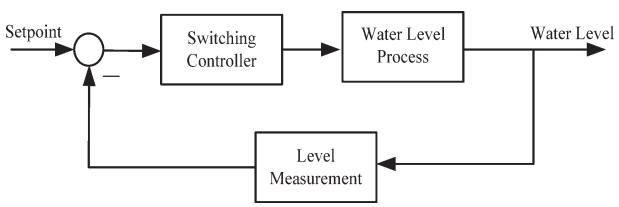

Figure 14. Block diagram of the switched control system.

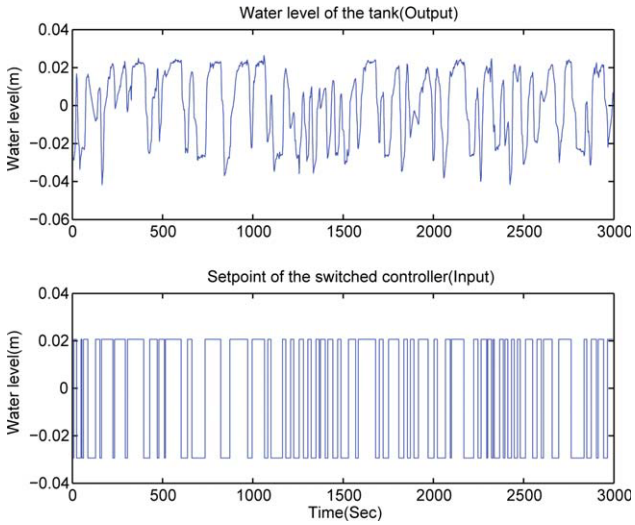

Figure 15. Input–output data of the switched control system.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

After passing the training data set through the algorithm, it is noticed that the identified switched models from the regular *EM* algorithm and its robust counterpart are similar. Figue 16 shows the estimated hidden variable $\hat{\boldsymbol{I}}_{\boldsymbol{k}}$ as well as the comparison between the simulated response of the identified models and the actual experimental data.

The identified closed-loop models under different level controllers are:

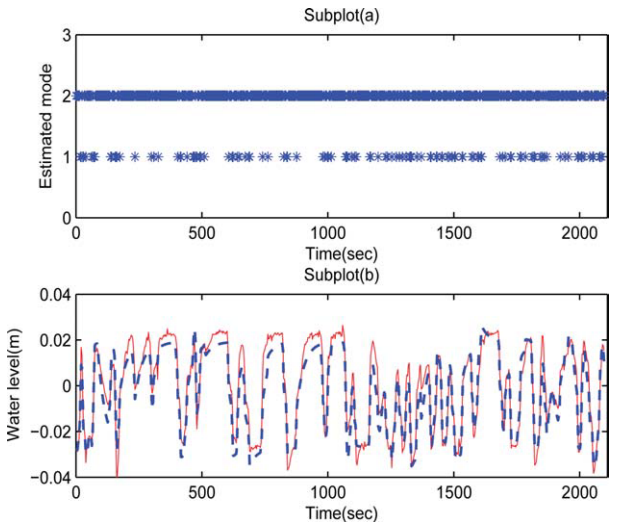

Figure 16. Subplot (a) denotes the estimated mode sequence of the switching evolution along the time for training data; subplot (b) self-validation of the identified multi-ARX modes (solid line: measured water level data, dashed line: prediction from the identified hybrid model).

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

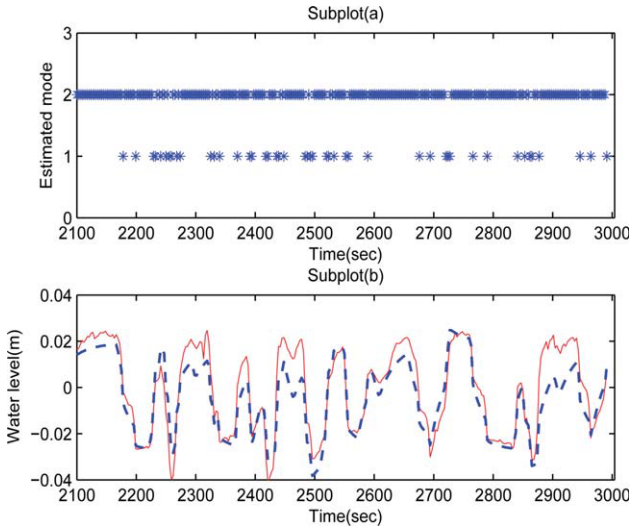

Figure 17. Subplot (a) denotes the estimated mode sequence of the switching evolution along the time for training data; subplot (b) cross-validation of the identified multi-ARX modes (solid line: measured water level data, dashed line: prediction from the identified hybrid model).

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$\textbf{Model 1}: \; y_k = 1.1430y_{k-1} - 04346y_{k-2} + 0.0572u_{k-1} + 0.2415u_{k-2} \quad (35)$$

$$\textbf{Model 2}: \; y_k = 0.9534y_{k-1} - 0.0475y_{k-2} + 0.0618u_{k-1} + 0.0336u_{k-2} \quad (36)$$

As shown in Figure 16, even though a relatively high percentage fit is achieved in self-validation, cross-validation is

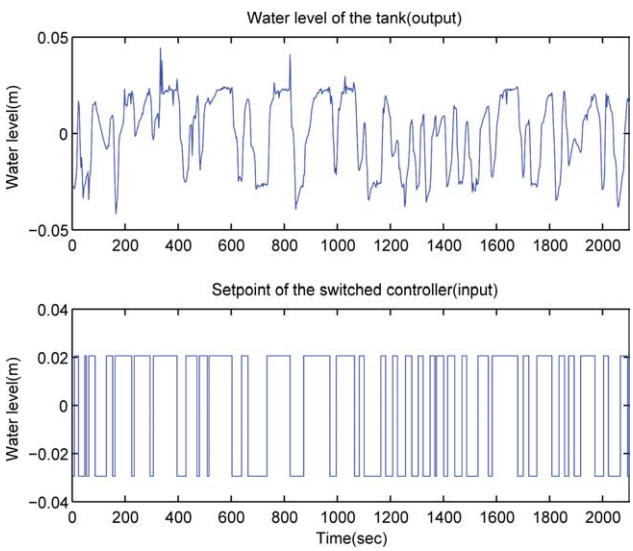

Figure 18. Input–output data of the switched control system with manually added outliers.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table 5. Comparison of Identification Results

Mode	$\theta_{reference}$	$\theta_{robustEM}$ (Mean)	$SD_{robustEM}$	$\theta_{normalEM}$ (Mean)	$SD_{normalEM}$	$\theta_{comparison}$ (Mean)	$SD_{comparison}$
Mode 1	1.1430	1.0917	0.2627	0.5034	0.4210	0.4482	0.4107
	−0.4346	−0.4153	0.2015	0.0838	0.4715	0.2229	0.4440
	0.0572	0.0924	0.1979	0.0183	0.1834	0.0173	0.1848
	0.2415	0.2297	0.1972	0.4589	0.2008	0.5404	0.1896
Mode 2	0.9534	0.9888	0.0466	0.9963	0.1442	0.9862	0.1423
	−0.0475	−0.0811	0.0541	−0.0999	0.1467	−0.0894	0.1456
	0.0618	0.0645	0.0335	0.0576	0.0619	0.0559	0.0625
	0.0336	0.0285	0.0329	0.0461	0.0606	0.0477	0.0591

SD stands for standard deviation.

still required to further test the validity of the identified models. Because of the switching time points are unknown in advance, unlike the conventional way of cross-validation, the mode identity of each data point in the cross-validation data set should be known before we are able to use the identified switched models to predict the system output. As a result, cross validation is performed through a two-step procedure. In the first step, validation data are clustered to estimate the model identity for each data point using the proposed clustering method, and then in the second step, the ARX models obtained from the identification data set are validated using the clustered data. The cross validation results of the multi-ARX models expressed in Eqs. 35 and 36 along with the validation data's estimated cluster identity are given in Figure 17.

Again, it is noticed from Figure 17 that the predictions from the identified models are close to the true system output, which implies that the identified sub-ARX models can effectively describe the dynamics of close-loop process under different controllers.

Case 2: Experimental data with outliers

No apparently abnormal data points have been observed in the data set shown in Figure 15 and the identified models have been successfully validated. To test the robustness of the proposed identification algorithm, several outliers are randomly added to the experimental training data set and Figure 18 shows the new data set after outliers are added.

With the outliers, it is expected that the performance of various identification procedures would degrade compared with the outlier-free case. As a result, the identified models described in Eqs. 35 and 36 are treated as reference models against which the new models obtained from the robust *EM* algorithm, regular *EM* algorithm, and the identification method introduced by Nakada et al.[15] Here, the reasons we chose the method[15] as the comparative method against which the proposed robust *EM* algorithm is evaluated are because both of the methods use statistical theory as a way to classify the observed data. Moreover, the structures of the parameters identified from the two algorithms are quite similar which makes them more comparable. To simplify the expression, we will refer the method put forward by Nakada et al. as the comparative method hereafter.

To sufficiently investigate the performance of different identification procedures in the presence of outliers, Monte-Carlo simulation is performed. In each run of the simulation, the percentage of outliers is fixed as 5% and the location of those outliers are randomly determined. After 100 runs, the averaged parameters along with the standard deviation of the estimated parameters from each of the identification procedures are calculated. 5 gives the identified results.

In comparison with the reference models, it is found that the presence of the outliers greatly skews the identification results from both the comparative method and the regular *EM* algorithm whereas, on the other hand, the robust *EM* algorithm identification procedure renders identified models closer to the reference ones by effectively spotting the outliers and diminishing their negative influence on parameter

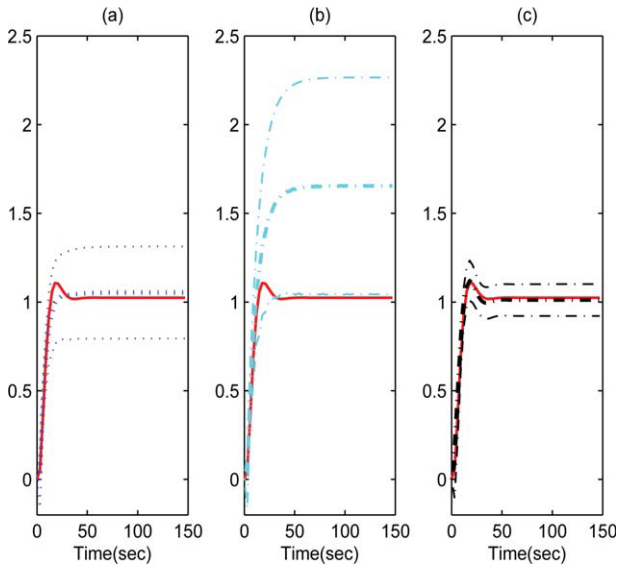

Figure 19. Comparison of the step response of the identified models for Mode 1.

(a) Step response of the regular *EM* algorithm identified model. Solid line: step response of the reference model, bold dotted line: mean step response of the identified models, dotted line: standard deviation bound of the step response from the identified models. (b) Step response of the comparative method identified model. Solid line: step response of the reference model, bold dash-dotted line: mean step response of the identified models, dash-dotted line: standard deviation bound of the step response from the identified models. (c) Step response of the robust *EM* algorithm identified model. Solid line: step response of the reference model, bold dash-dotted line: mean step response of the identified models, dash-dotted line: standard deviation bound of the step response from the identified models. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

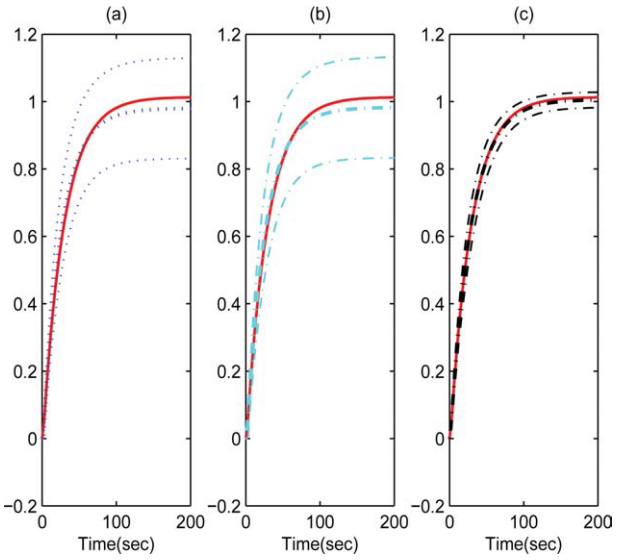

Figure 20. Comparison of the step response of the identified models for Mode 2.

(a) Step response of the regular *EM* algorithm identified model. Solid line: step response of the reference model, bold dotted line: mean step response of the identified models, dotted line: standard deviation bound of the step response from the identified models. (b) Step response of the comparative method identified model. Solid line: step response of the reference model, bold dash-dotted line: mean step response of the identified models, dash-dotted line: standard deviation bound of the step response from the identified models. (c) Step response of the robust *EM* algorithm identified model. Solid line: step response of the reference model, bold dash-dotted line: mean step response of the identified models, dash-dotted line: standard deviation bound of the step response from the identified models. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

estimation. To demonstrate the discrepancy of various identified models in a clearer way, step tests are performed, respectively, for the models listed in Table 5. It is found that 2 out of 100 sets of parameters obtained from the regular *EM* algorithm and the comparative method have poles outside of the unit circle, which can end up with unstable step response. These unstable models are removed before calculating the mean and standard deviation of the Monte-Carlo simulation for the regular *EM* algorithm and the comparative method.

The mean as well as the standard deviation of the step responses from the two sets of models identified from the *EM* algorithm with ($\theta_{\text{robust}EM}$) or without robust procedure ($\theta_{\text{regular}EM}$) together with the comparative method for Mode 1 and 2 have been calculated. Figures 19 and 20 show the comparison results.

Not surprisingly, the model identified from the robust *EM* algorithm outperforms the models rendered by the other two methods. It gives more accurate estimation of the step response on average whereas, in the mean time, enjoys much smaller standard deviation bound.

Conclusion

A new approach for identification of PWARX systems is developed in this article. The *EM* algorithm is used in data clustering as well as parameter estimation of the PWARX system. By expressing the noise distribution in a contaminated Gaussian distribution form, a robust *EM* algorithm is developed and its robustness to the outliers is demonstrated through simulations and pilot-scale experiment. It is shown that pre-running of the *EM* algorithm for several times with certain stopping condition can be an effective way to overcome its pitfall of sensitiveness to the starting point. Also, for those ''un-decidable'' data points in the data clustering, a refinement procedure can classify them to their own clusters by using the information provided by their spatially closest data points. For PWARX systems, in case that outliers exist in the reclassified local data set, robust parameter estimation of each local ARX model is realized by using the contaminated Gaussian distribution for residual errors. Region partition of PWARX systems is performed by using slightly modified MRLP algorithm in which data points are weighted based on their probability of being outliers. In this way, the influence of those abnormal data points could be minimized in the process of hyperplane determination. Finally, successful identification of simulated PWARX systems, a simulated continuous fermenter, and an experimental switched control system is achieved by using the proposed identification algorithm. The identification results confirm the potential capability of the proposed PWARX systems identification algorithm in handling a class of switched linear systems.

Acknowledgments

This work is supported in part by Natural Sciences and Engineering Research Council of Canada.

Literature Cited

1. Baneqee A, Arkun Y. Model predictive control of plant transitions using a new identification technique for interpolating nonlinear models. *J Process Control*. 1998;8:441–457.
2. Mhaskar P, El-Farra NH, Christofides PD. Predictive control of switched nonlinear systems with scheduled mode transitions. *IEEE Trans Automat Contr*. 2005;50:1670–1680.
3. Morari M. *Recent Development in the Control of Constrained Hybrid Systems*. lake Louise, Canada: CPC7, 2007.
4. Juloski A. *Ph.D. Thesis*, Netherlands: Eindhoven University of Technology, 2004.
5. El-Farra NH, Christofides PD. Coordinating feedback and switching control of hybrid nonlinear processes. *AIChE J*. 2003;49:2079–2098.
6. El-Farra NH, Mhaskar P, Christofides PD. Output feedback control of switched nonlinear systems using multiple Lyapunov functions. *Sys Control Lett*. 2005;54:1163–1182.
7. Barton PI, Pantelides CC. Modeling of combined discrete/continuous processes. *AIChE J*. 1994;40:966–979.
8. Barton PI, Banga JB, Galan S. Optimization of hybrid discrete/continuous dynamic systems. *Comput Chemi Eng*. 2000;24:2171–2182.
9. Barton PI, Lee CK. Design of process operations using hybrid dynamic optimization. *Comput Chem Eng*. 2004;28:955–969.
10. Bempoar A, Ferrari-Trecate G. Observability and controllability of piecewise affine and hybrid systems. *IEEE Trans Automat Contr*. 2000;45:1864–1876.
11. Heemels W, Schutter BD, Bemporad A. Equivalence of hybrid dynamical models. *Automatica*. 2001;37:1085–1901.
12. Ferrari-Trecate G, Muselli M, Liberati D, Morari M. A clustering technique for the identification of piecewise affine systems. *Automatica*. 2003;39:205–217.
13. Juloski A, Weiland S, Heemels WPMH. A Bayesian approach to identification of hybrid systems. *IEEE Trans Automat Contr*. 2005;50:1520–1533.

14. Bempoard A, Garull A, Paolett S, Vicino A. A bounded error approach to piecewise affine system identification. *IEEE Trans Automat Contr*. 2005;50:1567–1580.
15. Nakada H, Takaba K, Katayama T. Identification of piecewise affine systems based on statistical clustering technique. *Automatica*. 2003;39:205–217.
16. Vida R, Soatto S, Ma Y, Sastry S. *An algebraic approach to the identification of a class of linear hybrid systems*. In Proceedings of the 42nd IEEE conference on Decision and Control. Maui, HI: 2003: 167–172.
17. Ragot J, Mourot G, Maquin D. *Parameter estimation of switching piecewise linear systems*. In Proceedings of the 42nd IEEE conference on Decision and Control. Maui, HI: 2003;5783–5788.
18. Roll J, Bemporad A, ljung L. Identification of piecewise systems via mixed integer programming. *Automatica*. 2004;40:37–50.
19. Saldju T, Landgrebe A. Robust parameter estimation for mixture model. *IEEE Trans Geosci Remote Sen*. 2000;38:439–445.
20. Bennett KP, Mangasarian OL. Multicategory discrimination via linear programming. *Optimi Meth Software*. 1994;3:27–39.
21. Kalyani S, Giridhar K. *Robust statistics based expectation maximization algorithm for channel tracking in OFDM systems. Communications, 2007. ICC '07*. IEEE International Conference on 2007:3051–3056.
22. Saint-Jean C, Frelicot CM, Vachon B. *Advances in Pattern Recognition*. Berlin: Springer, 2000.
23. Ljung L. *System Identification, Theory for the User*; New Jersey: Prentice Hall; 1987.
24. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the *EM* algorithm. *J R Stat Soc Series B*. 1977;39:1–38.
25. Mclachlan GJ, Krishnan T. *The* EM *Algorithm and Its Extensions*. New York: John Willey & Sons, 1996.
26. Tjoa IB, Biegler L. Simultaneous strategy for data reconciliation and gross error detection of nonlinear systems. *Comput Chem Eng*. 1991;15:679–690.
27. Albuquerque JS, Biegler L. Data reconciliation and gross error detection for dynamic systems. *AIChE J*. 1996;42:2841–2856.
28. Ragot J, Chadli M; Maquin D. Data reconciliation: A robust approach using contaminated distribution. 16th IFAC World Congress. Prague, Czech Republic, 2005
29. Farris RH, Law VJ. An efficient computational technique for generalized application of maximum likelihood to improve correlation of experimental data. *Comput Chem Eng*. 1979;3:95–104.
30. Biernacki C, Celeux G, Govaert G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Compu Stat Data Anal*. 2003;41:561–575.
31. Paoletti S, Juloski AL, Ferrari-Trecate G, Vidal R. Identification of hybrid systems: a tutorial. *Eur J Control*. 2007;13:249–268.
32. Henson MA, Seborg DE. Nonlinear control strategies for continuous fermenters. *Chem Eng Sci*. 1992;4:821–835.
33. Gugaliya JK, Gudi RD, Lakashminarayanan S. Multi-model decomposition of nonlinear dynamics using a fuzzy-CART approach. *J Process Control*. 2005;15:417–434.
34. Venkat AN, Vijaysai P, Gudi RD. Identification of complex nonlinear process based on fuzzy decomposition of the steady state space. *J Process Control*. 2003;13:473–488.

Manuscript received Apr. 19, 2009, revision received Aug. 7, 2009, and final revision received Oct. 8, 2009.