

Mixture Discriminant Monitoring: A Hybrid Method for Statistical Process Monitoring and Fault Diagnosis/Isolation

Chien-Ching Huang ^a, Tao Chen ^b and Yuan Yao ^{a*}

^a *Department of Chemical Engineering, National Tsing Hua University, Hsinchu, 31003, Taiwan, R.O.C*

^b *Department of Chemical and Process Engineering, University of Surrey, Guildford GU2 7XH, UK*

Abstract

In order to better utilize historical process data from faulty operations, supervised learning methods, such as Fisher discriminant analysis (FDA), have been adopted in process monitoring. However, such methods can only separate known faults from normal operations, and they have no means to deal with unknown faults. In addition, most of these methods are not designed for handling non-Gaussian distributed data; however, non-Gaussianity is frequently observed in industrial processes. In this paper, a hybrid multivariate approach named mixture discriminant monitoring (MDM) was proposed, in which supervised learning and statistical process control (SPC) charting techniques are integrated. MDM is capable of solving both of the above problems simultaneously during online process monitoring. Then, for known faults, a root-cause diagnosis can be automatically achieved, while for unknown faults, abnormal variables can be isolated through missing variable analysis. MDM was used on the benchmark Tennessee Eastman (TE) process, and the results showed the capability of the proposed approach.

Key words: Multivariate statistical process control; Process monitoring; Fault diagnosis; Non-Gaussian; Supervised learning.

1. Introduction

In manufacturing processes, product quality and process safety rely upon stable operations under normal conditions. In order to achieve efficient process monitoring and fault diagnosis, data-based multivariate statistical process control (MSPC) methods have been intensively researched in recent years.¹⁻³ Most of these methods conduct process modeling based on a fault-free training dataset; however, real process data always include a certain number of faulty samples due to various types of sensor faults and process abnormalities. Hence, it is natural to think about developing monitoring methods that are based on both normal-operation and fault-related information.

In recent years, Fisher discriminant analysis (FDA),⁴ which is a popular supervised learning method, has been applied to process monitoring for more efficient fault detection and diagnosis.⁵⁻⁷ FDA was originally developed as a linear dimensionality reduction technique, which determines a set of linear transformation vectors and seeks the optimal separating directions between different classes. When applied to process monitoring, FDA-based methods utilize both normal and faulty data for process modeling, where historical data are classified to different classes associated with the process status. During online monitoring, known faults can then be separated from normal operations based on

the classification model. Thus, fault detection is achieved, and diagnosis results are obtained at the same time.

However, these methods rely only on data classification, and they do not provide the control charts that conventional statistical process control (SPC) can provide. As a result, they are not capable of dealing with the unknown faults that are often observed in industrial processes. Even for known faults, monitoring solely based on classification results may lead to delayed detection. Therefore, the lack of a control chart is a major disadvantage of the existing monitoring methods based on supervised learning.

In addition, the success of FDA classification relies on the assumption of within-class Gaussianity. However, non-Gaussian distribution of process data is common, especially in regards to multiple operation modes. In multimode processes, although the normal operation data collected in each mode will approximate a normal distribution, the overall data distribution is usually non-Gaussian. Similarly, when the same type of fault occurs in different modes, the overall distribution of the faulty data will also deviate from a Gaussian distribution. To address this issue, Choi et al. proposed the use of principal component analysis (PCA), Gaussian mixture model (GMM) and optimal discriminant analysis for non-Gaussian process fault detection and diagnosis.⁸ However, their method ignores the multimodality in the classes of faults. Most recently, Yu introduced localized Fisher discriminant analysis (LFDA) to process monitoring.⁹ Although LFDA works well in non-Gaussian data classification, it does not provide a control chart for efficient fault detection.

Other methods for non-Gaussian process monitoring include various extensions of the conventional MSPC methods,¹⁰⁻¹⁴ independent component analysis (ICA),¹⁵⁻¹⁸ GMM,¹⁹⁻²¹ multiset PCA (MsPCA),²² and so on. Nevertheless, none of these methods utilize the information contained in historical faulty data, since they are all unsupervised methods. In addition, a possibilistic unsupervised clustering method was proposed to detect both known and unknown faults.²³ While such a clustering method is useful when the correct label of the fault is not available, it differs from supervised classification, as it does not utilize the known fault types. In the meantime, the possibilistic clustering does not address non-Gaussian distributions.

To overcome the above problems, this paper proposed a hybrid multivariate approach known as mixture discriminant monitoring (MDM). This approach integrates mixture discriminant analysis (MDA)^{24, 25} and SPC charting. The primary contributions of this study are as follows. The proposed MDM method inherits the advantages of both supervised learning and SPC charting techniques. It not only can deal with non-Gaussian process data but also has the ability to detect both known and unknown faults efficiently. Furthermore, for known faults, a root-cause diagnosis can be achieved, while for unknown faults in which the exact types are unavailable, fault isolation is carried on using missing data analysis, in order to identify the most contributing variables.

The rest of the paper is organized as follows. In section 2, the proposed MDM approach is presented, including supervised modeling using MDA and the development of the SPC control chart. The entire procedure of process modeling and online fault detection based on MDM is introduced in section 3. Section 4 describes the method for the root-cause diagnosis of known faults, as well as the algorithm of unknown faults isolation. In section 5, the case studies on the benchmark Tennessee Eastman (TE) process verify the capability of the proposed approach. Finally, the conclusions are drawn in section 6.

2. Mixture discriminant monitoring

2.1 Step 1: classification model building for non-Gaussian process data

In the first step of MDM, mixture discriminant analysis (MDA)^{24, 25} is adopted to build a classification model for the non-Gaussian process data. MDA is a supervised learning method for multimodal data

classification in which the class densities are modeled by Gaussian mixtures of several subclass models, where each subclass has a multivariate normal distribution. Such properties make MDA a better method than FDA for non-Gaussian process data classification.

As mentioned previously, in industry, historical process data not only are collected under normal operating conditions but also contain fault-related information that is helpful to improve fault diagnosis. Associated with process status, measurements can be partitioned into different classes, with each data class corresponding to either normal operation or a particular type of abnormality. Meanwhile, depending on the process characteristics, within-class non-Gaussianity may exist, especially when the process has multiple operation modes.

Here, assume that the training dataset $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n]^T$ has the dimensions of $n \times m$, where m is the number of process variables and n is the number of samples. The training sample vector \mathbf{x}_i ($i = 1, 2, \dots, n$) is known to belong to a class j , where $j = 1, 2, \dots, J$. J is the total number of classes, including one normal operation class and $(J-1)$ faulty operation classes that corresponds to different types of known faults. To describe the within-class non-Gaussianity, each class j may be further divided into R_j subclasses, denoted by c_{jr} , $r = 1, \dots, R_j$. The total number of subclasses is $R = \sum_{j=1}^J R_j$. Within each subclass, the data will have a multivariate normal distribution with a mean vector $\boldsymbol{\mu}_{jr}$ and a covariance matrix $\boldsymbol{\Sigma}_{jr}$. Therefore, the non-Gaussian distribution in class j can be modeled by a mixture of R_j Gaussian distributions. Especially in multimode processes, R_1 is usually equal to the number of modes, supposing class 1 is the normal operation class. The values of R_j may vary for different j , because various types of faults may be observed in different operation modes.

The prior probability for class j is denoted by Π_j , which is usually known or can be estimated from the training data. The mixing probability for the r th subclass within class j is an unknown parameter, π_{jr} , and $\sum_{r=1}^{R_j} \pi_{jr} = 1$. The mixture density for class j is:

$$m_j(\mathbf{x}) = p(\mathbf{x} | j) = \sum_{r=1}^{R_j} \pi_{jr} (2\pi)^{-m/2} |\boldsymbol{\Sigma}_{jr}|^{-1/2} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu}_{jr})^T \boldsymbol{\Sigma}_{jr}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{jr})}{2}\right). \quad (1)$$

The parameters, $\boldsymbol{\mu}_{jr}$, $\boldsymbol{\Sigma}_{jr}$ and π_{jr} , can be estimated by maximizing the following likelihood function:

$$l^{mix}(\theta) = \prod_{i=1}^n p(\mathbf{x}_i | \theta), \quad (2)$$

where the parameter $\theta = \{\pi_{jr}, \boldsymbol{\mu}_{jr}, \boldsymbol{\Sigma}_{jr}; r = 1, 2, \dots, R_j, j = 1, 2, \dots, J\}$. This can be achieved by the iterative expectation-maximization (EM) algorithm. In each iteration, provided that \mathbf{x} is a training sample in the j th class, the posterior probability of \mathbf{x} belonging to the r th subclass of class j is:

$$P(c_{jr} | \mathbf{x}, j) = \frac{\pi_{jr} (2\pi)^{-m/2} |\boldsymbol{\Sigma}_{jr}|^{-1/2} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu}_{jr})^T \boldsymbol{\Sigma}_{jr}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{jr})}{2}\right)}{\sum_{k=1}^{R_j} \pi_{jk} (2\pi)^{-m/2} |\boldsymbol{\Sigma}_{jk}|^{-1/2} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu}_{jk})^T \boldsymbol{\Sigma}_{jk}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{jk})}{2}\right)}. \quad (3)$$

The above equation is known as the expectation step (E-step), which is followed by the maximization step (M-step):

$$\pi_{jr} \propto \sum_{\mathbf{x}_i \in \text{class } j} P(c_{jr} | \mathbf{x}_i, j), \quad \sum_{r=1}^{R_j} \pi_{jr} = 1, \quad (i = 1, \dots, n) \quad (4)$$

$$\boldsymbol{\mu}_{jr} = \sum_{\mathbf{x}_i \in \text{class } j} \mathbf{x}_i P(c_{jr} | \mathbf{x}_i, j) / \sum_{\mathbf{x}_i \in \text{class } j} P(c_{jr} | \mathbf{x}_i, j), \quad (5)$$

$$\boldsymbol{\Sigma}_{jr} = \frac{\sum_{\mathbf{x}_i \in \text{class } j} P(c_{jr} | \mathbf{x}_i, j) (\mathbf{x}_i - \boldsymbol{\mu}_{jr})(\mathbf{x}_i - \boldsymbol{\mu}_{jr})^T}{\sum_{\mathbf{x}_i \in \text{class } j} P(c_{jr} | \mathbf{x}_i, j)}. \quad (6)$$

The E-step and the M-step are repeated iteratively until convergence. The posterior class probabilities of data sample \mathbf{x} are then calculated using Bayes' theorem:

$$P(j | \mathbf{x}) \propto \prod_j \sum_{r=1}^{R_j} \pi_{jr} (2\pi)^{-m/2} |\boldsymbol{\Sigma}_{jr}|^{-1/2} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu}_{jr})^T \boldsymbol{\Sigma}_{jr}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{jr})}{2}\right). \quad (7)$$

For data classification, when a new sample, \mathbf{x}_{new} , is measured, its posterior class probabilities can be calculated using (7). Then, \mathbf{x}_{new} is classified into class j corresponding to the largest $P(j | \mathbf{x}_{\text{new}})$, where $j = 1, 2, \dots, J$.

To conduct the EM iteration above, the number of total classes J and the cluster sizes R_j ($j = 1, 2, \dots, J$), i.e., the number of subclasses in each class, should be determined. Usually, the knowledge of class number J is available. In multimode processes, the values of R_j may also be known. If this is not the case, k-means or learning vector quantization (LVQ) techniques can be adopted for the estimation of R_j . For more details about the selection of cluster sizes, please refer to the cited paper²⁴.

2.2 Step 2: development of control chart

Supervised learning techniques, including MDA, were originally proposed for classifying data into known classes according to certain probability indexes or similarity indexes. Most of these techniques have no means to deal with data belonging to unknown classes. In industry, it is nearly impossible to know all types of faults. On the contrary, historical databases often only contain normal operating data and several frequently observed types of faults, while other faults are unknown. As illustrated in Fig. 1, unknown faults may be classified into either normal operation class or classes corresponding to certain known faults, causing missing alarms or incorrect diagnosis. Hence, it is necessary to integrate SPC control charts with supervised learning to improve both fault detection and diagnosis.

To construct a control chart, two key components must be determined, which are the monitoring statistic and the control limits. In the second step of MDM-based process modeling, a monitoring statistic is derived from the probability density function (*pdf*) of the data. If a collected sample is classified into class j , the corresponding value of the monitoring statistic is equal to $-\log(m_j(\mathbf{x}))$, where $j = 1, 2, \dots, J$. Then, using a confidence level of $100\beta\%$, the *pdf* confidence bound of subclass k in class j is defined using a threshold h_{jk} that satisfies the following:

$$\int_{\mathbf{x}:m_j(\mathbf{x})>h_{jk}} m_j(\mathbf{x})d\mathbf{x} = \beta, \quad (8)$$

where \mathbf{x} belongs to this subclass. The corresponding subclass control limit is $-\log(h_{jk})$. There are several ways to estimate h_{jk} . When the number of historical samples in class j is sufficiently large, the likelihood of all these samples can be calculated and h_{jk} can be directly identified¹⁹. This approach is usually applicable to the normal operation class, which could contain thousands of nominal data points for model training. However, it may be unreliable to use such an approach to calculate h_{jk} for faulty operation classes, since the historical sample number of each type of fault is possibly limited. To solve this problem, numerical Monte Carlo (MC) simulations^{26, 27} can be conducted to approximate the integral in (8). This paper omitted the technical details of Monte Carlo simulations, in order to keep the paper within a reasonable length and to maintain readability. Interested readers can refer to the cited references.

Using the monitoring statistic and the control limits, a control chart can be plotted for online fault detection. The detailed procedure will be introduced in the next section. The control limits introduced above are particularly suitable for monitoring multimode processes in which the data are approximately normally distributed within each mode. Under such a situation, a control chart that has the subclass control limit $-\log(h_{jk})$ can reveal the local characteristics of each operation mode. An alternative selection is to use the class control limit $-\log(h_j)$ instead of $-\log(h_{jk})$ in the control chart, where h_j is the *pdf* confidence bound of class j derived in a similar way as in (8). Such a control chart is suggested when monitoring single-mode non-Gaussian processes.

It is worthy to notice that, in conventional MSPC, the process model is always assumed to exactly represent the nominal operation conditions (NOC), and thus any deviation from this model raises alarms. In this paper, a similar assumption is followed, i.e. the MDM model is assumed to well represent both the NOC and the operation conditions of known faults. However, in reality, even if the operation conditions are stationary and the model structure is within the correct family of the true process, the model cannot be perfectly obtained with limited data. Instead, there exists uncertainty in the model parameters. The significance of the results depends not only on the size but also on the distribution of the training data. To set a clear focus of this paper, the assessment of model uncertainty is not conducted here, and will be addressed in the follow-on study.

Another issue in industrial applications is the changing behaviors of processes caused by aging factors, changes in raw materials, weather conditions, etc. Such changing behaviors are out of the scope of this paper and not considered in the above modeling procedure, but some discussions are provided as follows. In MSPC, the changing behaviors have been well addressed by using the recursive models or moving windows.²⁸⁻³¹ Similar idea can be utilized to improve the performance of MDM in such situations. However, the frequency of model updating (or the size of each moving window) is still an open question in recursive modeling, which relates to the process characteristics and is case-dependent. Process knowledge is very helpful in the determination of the model updating frequency and the window size. Alternatively, one can also update the model whenever a new sample is available. There is a tradeoff between maximizing the use of available data and minimizing the computation time.

3. Procedure of modeling and online fault detection

The modeling procedure of the proposed MDM method is shown in Fig. 2(a) and summarized below:

1. Build an MDA classification model based on the historical process data, including both normal and faulty operation information.
2. Calculate the control limits for different subclasses, based on the *pdf* values achieved in the MDA modeling step.
3. Store the model parameters, including $\boldsymbol{\mu}_{jr}$, $\boldsymbol{\Sigma}_{jr}$, π_{jr} ($r = 1, 2, \dots, R_j$; $j = 1, 2, \dots, J$) and the control limits.

The built model can then be utilized in online fault detection. The flowchart is plotted in Fig. 2(b).

1. When new process measurements, \mathbf{x}_{new} , are collected, the classification model is utilized to assign the data into a known class, according to the largest value of $P(j|\mathbf{x}_{\text{new}})$, where $j = 1, 2, \dots, J$.
2. Calculate the value of the monitoring statistic $-\log(m_a(\mathbf{x}_{\text{new}}))$, where $\mathbf{x}_{\text{new}} \in \text{class } a$.
3. Determine the subclasses that \mathbf{x}_{new} belong to by finding the maximum value of $P(c_{ar}|\mathbf{x}, a)$, where $r = 1, \dots, R_j$.
4. Supposing $\mathbf{x}_{\text{new}} \in \text{subclass } g \text{ in class } a$, plot both the monitoring statistic $-\log(m_a(\mathbf{x}_{\text{new}}))$ and the corresponding control limit $-\log(h_{ag})$ on the control chart and compare them.
 - (a) If the monitoring statistic is below the control limit and class a is the normal operation class, the process status is normal.
 - (b) If the monitoring statistic is below the control limit and class a is a faulty operation class, a known fault is detect.
 - (c) If the monitoring statistic is beyond the control limit, an unknown fault is captured.

Note that the common assumptions of conventional SPC methods should also be held for utilizing MDM in fault detection. Specifically, the training data should well characterize the both normal operation and the known faulty operation of the process, while the faults to be detected should be reflected by the process measurements.

4. Fault diagnosis and isolation

When a known fault is detected, diagnosis is required to understand the root cause of the abnormality, i.e., to find out the fault type. In MDM, this is a quite easy job. Since the new measurements are classified into a certain faulty operation class, the diagnosis results are automatically achieved.

If a new fault with an unknown type is detected, fault isolation should be performed to identify the most responsible variables. Contribution plots are widely used tools in MSPM for fault isolation;^{32, 33} however, such methods often suffer from the smearing effect and provide misleading results.³⁴ To overcome such a problem, the idea of fault reconstruction based on missing variable analysis was adopted in this paper. The original idea of fault reconstruction was to deal with the situation in which there is only one process variable contributing to the faults at each time.³⁵ During fault isolation, each variable is treated as if it was missing and the monitoring statistic is re-estimated. Consequently, the variable that most significantly reduces the monitoring statistic is considered to be the cause of the alarm. A similar idea has been extended to analyze the joint effect of several variables.

Inspired by the previous research, this study developed an algorithm for MDM-based fault isolation. Without loss of generality, the normal operation class is labeled as class 1. The procedure is as follows:

1. Set $d = 1$.
2. Select d variables as missing variables and re-estimate the monitoring statistic $-\log(m_1(\mathbf{x}_{\text{new}}))$ corresponding to the “normal operation class”. Identify the d variables that minimize $-\log(m_1(\mathbf{x}_{\text{new}}))$.
3. If the minimum value of the re-estimated statistic is below the confidence bound, then the corresponding d variables are isolated as the cause of the fault, and the algorithm is terminated. Otherwise, set $d = d + 1$, and return to step 2.

When applying the above algorithm, there is an important question that must be answered: how to estimate the monitoring statistic $-\log(m_1(\mathbf{x}_{\text{new}}))$ when several variables are missing?

Suppose each sample \mathbf{x} can be divided into two parts, $\hat{\mathbf{x}}$ and $\tilde{\mathbf{x}}$, where $\mathbf{x} = \begin{bmatrix} \hat{\mathbf{x}} \\ \tilde{\mathbf{x}} \end{bmatrix}$. For subclass r of class 1 (the normal operation class), the joint probability density of $\tilde{\mathbf{x}}$ and $\hat{\mathbf{x}}$ is described as:

$$p(\hat{\mathbf{x}}, \tilde{\mathbf{x}} | 1, r) = (2\pi)^{-m/2} |\Sigma_{1,r}|^{-1/2} \exp\left(-\frac{(\mathbf{x} - \boldsymbol{\mu}_{1,r})^T \Sigma_{1,r}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{1,r})}{2}\right). \quad (9)$$

Based on the parameters of the joint distribution model, it is straightforward to compute the expectations of $\tilde{\mathbf{x}}$ in class 1 given $\hat{\mathbf{x}}$ by conditioning the joint distribution on $\hat{\mathbf{x}}$ and taking the expected value.³⁶ In each subclass r , denote the means corresponding to $\tilde{\mathbf{x}}$ and $\hat{\mathbf{x}}$ as $\tilde{\boldsymbol{\mu}}_{1,r}$ and $\hat{\boldsymbol{\mu}}_{1,r}$, while the two variance matrices and the covariance matrix are $\hat{\Sigma}_{1,r}$, $\tilde{\Sigma}_{1,r}$ and $\check{\Sigma}_{1,r}$, respectively. The mean vector and the covariance matrix estimated from the joint density can then be separated as $\boldsymbol{\mu}_{1,r} = [\hat{\boldsymbol{\mu}}_{1,r}, \tilde{\boldsymbol{\mu}}_{1,r}]$ and $\Sigma_{1,r} = \begin{bmatrix} \hat{\Sigma}_{1,r} & \check{\Sigma}_{1,r} \\ \check{\Sigma}_{1,r} & \tilde{\Sigma}_{1,r} \end{bmatrix}$. Thus, supposing $\tilde{\mathbf{x}}$ is missing, its conditional expectation in subclass r is derived as:

$$\hat{\tilde{\mathbf{x}}}_r = \tilde{\boldsymbol{\mu}}_{1,r} + \check{\Sigma}_{1,r} (\hat{\Sigma}_{1,r})^{-1} (\hat{\mathbf{x}} - \hat{\boldsymbol{\mu}}_{1,r}). \quad (10)$$

Consequently, for the mixture of all subclasses, the overall conditional expectation is:

$$\hat{\tilde{\mathbf{x}}} = \sum_{r=1}^{R_1} P(c_{1,r} | \hat{\mathbf{x}}) \hat{\tilde{\mathbf{x}}}_r, \quad (11)$$

where $P(c_{1,r} | \hat{\mathbf{x}})$ is the probability of missing variables $\tilde{\mathbf{x}}$ belonging to the r th subclass, prior to observing $\tilde{\mathbf{x}}$. $P(c_{1,r} | \hat{\mathbf{x}})$ is calculated as:

$$P(c_{1,r} | \hat{\mathbf{x}}) = \frac{\pi_{1,r} p(\hat{\mathbf{x}} | c_{1,r})}{\sum_{v=1}^{R_1} \pi_{1,v} p(\hat{\mathbf{x}} | c_{1,v})}, \quad (12)$$

$$p(\hat{\mathbf{x}} | c_{1,r}) = \frac{\exp[-\frac{1}{2}(\hat{\mathbf{x}} - \hat{\boldsymbol{\mu}}_{1,r})^T \hat{\boldsymbol{\Sigma}}_{1,r}^{-1}(\hat{\mathbf{x}} - \hat{\boldsymbol{\mu}}_{1,r})]}{(2\pi)^{\frac{m}{2}} |\hat{\boldsymbol{\Sigma}}_{1,r}|^{\frac{1}{2}}}. \quad (13)$$

After the missing variables $\tilde{\mathbf{x}}$ is reconstructed, $-\log(m_1(\mathbf{x}_{new}))$ can be calculated based on (7).

Notice that when the number of variables is large, the recently proposed branch and bound (BAB) method³⁷ can be adopted to reduce the computation burden of the proposed algorithm and find the minimum set of variables affected by the fault in a short time.

Some discussions about the fault isolation procedure are provided as follows. First, when there exist several different combinations of variables that are regarded as the potential contributors to the detected fault, the above procedure is able to prioritize the combinations according to the reconstructed monitoring statistic. In other words, since the fault isolation is an optimization problem, it always finds the combination that gives the best reconstructed monitoring statistic, i.e. the minimum value of $-\log(m_1(\mathbf{x}_{new}))$. Second, the essence of multivariate contribution analysis is to look at the "multivariate" contribution, i.e. to identify the most contributing combination of variables. Usually, it is not necessary to assign contribution of each individual variable to the fault. If really needed, one can re-run univariate contribution analysis on the isolated variables, and obtain the related information from the reconstructed values of the monitoring statistic. Third, when there are many correlated variables identified as the contributors, it may be desired to find out which variables are the root cause of the fault. Currently, multivariate contribution analysis does not explicitly address this problem. Usually, such situation is difficult to handle without process knowledge. Nevertheless, causality analysis may shed some light on root-cause diagnosis. Recently, Yang and Xiao gave a review on the related subject.³⁸

5. Case studies

5.1 Tennessee Eastman process

The TE benchmark process,³⁹ which has been widely accepted in performance tests of control algorithms and monitoring approaches, was utilized to verify the effectiveness of the proposed hybrid method. This process has five main units, which are the reactor, condenser, separator, stripper and compressor, respectively. The streams of the plant consist of eight components, among which components A, B and C are the gaseous reactants fed to the reactor, while G and H are the products. A brief flow chart of the TE process is shown in Fig. 3. Since the TE plant is open-loop unstable, the decentralized control strategy⁴⁰ was implemented for the case studies in this paper. This process can be operated in six different modes, two of which (mode 1 and mode 3) were selected to simulate the multimode operation and generate non-Gaussian distributed data. A total of 52 process variables were measured for process monitoring, as listed in Table 1 in the Supporting Information. The sampling interval was 0.05 hours. For illustration of the effectiveness of the proposed MDM method, historical measurements from normal operation and three different types of known faults in both modes were collected for process modeling. These three types of faults are faults 4, 12 and 13 in the original paper³⁹, which are caused by a step change in the reactor cooling water inlet temperature, random variations in the condenser cooling water inlet temperature, and the slow drift of process reaction kinetics, respectively. These data will be utilized in subsections 5.2 - 5.5. Then, an overall comparison between MDM and FDA will be conducted in subsection 5.6, based on all 20 types of faults listed in Table 2 in the Supporting Information.

5.2 Process modeling and model verification

For process monitoring, an MDM model was built following the procedure presented in section 3, which was based on the training data consisting of 5000 normal operation samples collected from each mode and 500 faulty samples collected from each mode for each type of fault. The number of normal operation data was much larger than the number of historical faulty data, which was consistent with the reality of industrial processes. The training data were assigned into four classes, in which the normal data were labeled as class 1 and classes 2-4 consisted of different types of faulty data. Each class had two subclasses that had a clear engineering meaning: subclass 1 corresponded to operation mode 1, while subclass 2 corresponded to mode 3. Then, the control limits were calculated for the different subclasses, as described in subsection 2.2. For comparison, an FDA model was also built based on the same training dataset. Another 500 normal samples were collected from each mode for model verification, together with 500 faulty samples collected from each mode for each type of fault. Note that during the modeling and verification step, the transitional data from normal operation to faulty operations were not included in either the training or the testing set, since knowledge about such data may not be sufficient in real industry, making it difficult to label them in supervised learning.

The two process models resulting from MDM and FDA⁵ were compared as shown in Fig. 4, which shows the typical monitoring results of the training data. In Fig. 4(a), FDA misclassified many of the training samples, because it did not consider the non-Gaussianity of the data. Especially, it could not correctly differentiate fault 12 from normal operation in mode 1, and the modeling results of fault 13 were poor as well. In contrast, as shown in Fig. 4(b), the MDM monitoring statistic nearly always stayed below the 99% control limit of each corresponding subclass, indicating that MDM could properly model the training data and correctly reflect the underlying process status. The monitoring results of the verification data, which are not shown here, were similar to those shown in Fig. 4. The comparison revealed the effectiveness of MDM in modeling non-Gaussian distributed data.

5.3 Online detection and diagnosis of known faults

To examine the efficiency of online fault detection and diagnosis using the proposed approach, two testing scenarios with different kinds of known faults were designed.

The first scenario was based on running at operation mode 1. After a normal operation period with 50 sampling intervals, fault 4 occurred in the process and lasted for 350 sampling intervals, after which the process returned to normal operation. Another fault (fault 12) happened at the 451th sampling interval and lasted until the 800th sampling interval at which time the process returned to normal. Finally, fault 13 affected the process between the 851th and the 1200th sampling intervals. Different from the training and verification data, the transitional samples between normal and faulty operations were also collected for online monitoring.

The monitoring results based on FDA are shown in Fig. 5(a). Although the classification results of normal data and fault 4 were satisfactory, faults 12 and 13 could not be correctly identified. Most data from fault 12 were regarded as being normal, indicating that there were missing alarms in FDA-based fault detection. The detection results of fault 13 were better, although misclassifications still occurred. However, a large detection delay could be observed, since the fault was not detected before the 887th sampling interval. As introduced previously, fault 13 refers to the slow drift of the process reaction kinetics, which is relatively difficult to detect.

The MDM-based monitoring results are plotted in Fig. 5(b). Since all normal operation data were correctly classified into the first class, the monitoring statistic $-\log(m_1(\mathbf{x}))$ was used for online monitoring of the data. Obviously, during the periods of normal operation, the values of $-\log(m_1(\mathbf{x}))$

were well below corresponding the control limit $-\log(h_{1,1})$, where the subscript “1,1” refers to subclass 1 in class 1 and indicates no false alarm. Similarly, the samples collected between intervals 51-400 and 451-800 were classified into the second class (the class of fault 4) and the third class (the class of fault 12), respectively. Therefore, the statistic $-\log(m_2(\mathbf{x}))$ and the corresponding control limit $-\log(h_{2,1})$ were utilized for process monitoring between intervals 51-400, while $-\log(m_3(\mathbf{x}))$ and $-\log(h_{3,1})$ were adopted between sampling intervals 451-800. The monitoring results confirmed the classification results at most points, indicating that these two periods suffered from the known faults 4 and 12. Thus, the fault diagnosis was achieved at the same time of online fault detection. Since fault 13 was a slow drift, the online samples were not classified into class 4 (the class of fault 13) until sampling interval 865. However, the control chart shows that the monitoring statistic was outside the control limit of the normal operation class at sampling interval 863, meaning that the exact detection delay of fault 13 was only twelve sampling intervals. This result was much better than that using FDA. After another two sampling intervals, the monitoring statistic and the control limit switched to $-\log(m_4(\mathbf{x}))$ and $-\log(h_{4,1})$ according to the classification results. As seen in the control chart, for most of the time between the 865th and 1200th sampling intervals, $-\log(m_4(\mathbf{x})) \leq -\log(h_{4,1})$, indicating that the fault was a known fault, i.e. fault 13.

In the second scenario, the same types of faults occurred in the process at the same sampling intervals found in scenario 1. The only difference was that the process operated in mode 3.

As observed in Fig. 6, FDA identified faults 4 and 12 in this mode, but the monitoring results of fault 13 were not sufficient. Besides a large detection delay, many samples of fault 13 were misidentified as either fault 12 or normal data. MDM did a much better job of identifying the faults. The accuracy of the monitoring and diagnosis for normal operation, fault 4 and fault 12 was similar to those found in the first scenario. For fault 13, the control chart detected the fault at the 865th sampling interval and had an acceptable delay of 14 samples, because the monitoring statistic went beyond the control limit at that point. Then, at sampling interval 871, the fault was correctly diagnosed, since the points after 871 were classified into class 4 and the monitoring results provided by the control chart confirmed the classification.

5.4 Online detection and isolation of unknown faults

In this subsection, another two scenarios were designed to illustrate the online detection and isolation of unknown types of faults that are not included in the historical training dataset.

In scenario 3, samples 1-50 were collected from normal operation in mode 1. Then, the process operation switched to mode 3. Fault 5 in the cited paper³⁹, which is a step change in condenser cooling water inlet temperature, arose at sampling interval 101 and lasted for 350 samples.

As shown in Fig. 7(a), the FDA model raised a number of false alarms in mode 1. A larger problem was that it classified the data from fault 5 into the group of fault 12. Such a mistake is easy to understand. Since FDA did not take fault 5 into consideration during model building, the model had no means to diagnose such a fault correctly. In fact, fault 5 was regarded as fault 12 in the FDA classification, since these two types of faults were very similar and were both about the condenser cooling water inlet temperature.

As shown in Fig. 7(b), the MDM approach showed its superiority in unknown fault detection. First, there were no false alarms when the process was under normal operation. Second, although the model classified the fault into the group of fault 12, the control chart showed that there were a number of points at which the values of the monitoring statistic $-\log(m_3(\mathbf{x}))$ were outside the control limit

$-\log(h_{3,2})$. The subscript “3,2” refers to the second subclass in the third class, corresponding to fault 12 in mode 3. The monitoring results revealed that the fault did not really belong to fault 12; instead, it was identified as an unknown fault.

After detecting an unknown fault, MDM isolated the fault using the reconstruction algorithm introduced in section 4. Since the fault was detected at sampling interval 101, the data on this point were used for fault isolation. It was found that if only one variable was treated as the missing variable, the smallest re-estimated value of $-\log(m_1(\mathbf{x}))$ would be -12.602. Such a value would be obtained when variable 22, i.e. the separator cooling water outlet temperature, was reconstructed. Although the value was still larger than the control limit $-\log(h_{1,3}) = -15.7403$, variable 22 was identified as the variable contributing most to the fault. Then, two variables were treated as missing variables at the same time. The best result was provided by reconstructing both variables 22 and 11, which was equal to -23.043 and was well below the control limit. Variable 11 was the product separator temperature. Such an isolation result was reasonable, since the separator cooling water outlet temperature was directly influenced by the change in the condenser cooling water inlet temperature, and because the fault was easy to propagate and could affect the product separator temperature in a short time. The isolation result could help the operators to find out the root-cause of the fault. Synthesizing the information provided by the MDM control chart and the fault isolation inferred that this unknown fault was similar to fault 12 but had a different pattern.

The 4th scenario was about the detection of random variations in the C feed temperature, which is fault 10 in the cited paper³⁹. In this scenario, the first 50 samples were collected from normal operation in mode 1, while the unknown fault 10 occurred at the 51th sampling interval.

As indicated in Fig. 8, the FDA model partitioned most of the faulty points into the normal operation class and could not detect the fault. In comparison, MDM detected the fault at the 58th sampling interval when the monitoring statistic $-\log(m_1(\mathbf{x}))$ drifted out of the control limit $-\log(h_{1,1})$. Then, at sampling interval 62, the monitoring statistic and the control limit were switched to $-\log(m_4(\mathbf{x}))$ and $-\log(h_{4,1})$, because the following samples were classified into the class of fault 13 in mode 1. However, since $-\log(m_4(\mathbf{x})) > -\log(h_{4,1})$ at most points after sampling interval 62, the fault was not really a known fault. Fault isolation was conducted to find out the most contributing variables. The procedure of the fault isolation was similar to that in scenario 3, and is therefore not presented here.

Please note that the use of first several data points is a design choice. When fault occurs, the first several data points are representative of moving away from the normal steady-state and carry very useful information. Also, getting the diagnosis results as soon as possible is always desired.

5.5 Online detection and isolation of multiple faults

In industrial operations, multiple faults may happen at the same time. To illustrate the capability of the proposed method on handling such issue, two simultaneous faults were generated in the first operating mode, which were fault 2 and fault 4. Fault 2 is an unknown fault caused by a step change in the B composition, while fault 4 is a known fault already included in both the FDA and MDM models. Both faults occur at the 51st sampling interval simultaneously.

As shown in Fig. 9(a), the faults were detected by the FDA model efficiently. Nevertheless, FDA divided the faulty data into the class corresponding to fault 4, which means that fault 2 was not recognized. In comparison, the monitoring results of MDM are plotted in Fig. 9(b). Using MDM, the faulty data points after the 51th sample were also successfully identified. Similar to the results of FDA, these points were classified to the class of fault 4. The control chart showed that the monitoring statistic

went outside the control limit of fault 4 quickly after the fault was detected, and returned to below the control limit after about 3 more sampling intervals. This may be caused by the dynamic characteristics of fault 4. When fault 4 occurs, the reactor temperature always increases dramatically, leading to an overshoot of the monitoring statistic. Then, after a while, the pattern of the fault becomes stable, while the values of the monitoring statistic decrease and become "in-control" (i.e. below the control limit of fault 4). Notice that similar pattern can be found in Figs. 5(b) and 6(b). However, the monitoring statistic in Fig. 9(b) drifted away from the "in-control" status again after the 57th sampling interval. Such phenomenon implied that the detected fault had a different pattern from the known one, although it looked like fault 4 to some content.

To isolate the most contributing variables to the fault, the reconstruction algorithm was conducted following the steps introduced in section 4. Take the 56th sample as an example for illustration, since it was the first sample after the control chart indicating an unknown fault pattern. During the reconstruction, it was found that the value of the monitoring statistic changed significantly after taking variable 49, the reactor cooling water flow, as the missing variable. The statistic value decreased from 588.4627 to 5.7107, but was still higher than the 99% control limit of the normal operation in mode 1, which was equal to -17.9249. Such result implied that variable 49, a potential cause to fault 4, might be critical but was not the only contributor to the detected fault. After both variable 49 and variable 21 were treated as the missing variables, the reconstructed value of the monitoring statistic further decreased to -19.3865 that is smaller than the control limit. Therefore, variable 21, the reactor cooling water outlet temperature, was inferred as the second contributor to the fault. According to the results in the cited paper³⁷, the change in such variable might be linked to fault 2. In summary, the fault isolation results suggested that the detected fault was probably a mixture of fault 2 and fault 4.

From this example, it can be found that the multiple faults happening simultaneously cannot be directly diagnosed based on the classification results. Instead, such faults are identified as unknown faults by MDM. However, the fault isolation algorithm based on the missing variable analysis can indicate the most contributing variables and imply the types of the multiple faults, no matter these faults are known or unknown.

5.6 Overall comparison based on all types of faults

In this subsection, the overall comparison between MDM and FDA will be provided. Following the suggestion from an anonymous referee of this paper, these two methods were compared based on all types of TE faults. For each fault, MDM and FDA process models were built using both the normal operation data and the historical faulty data. Then, both models were applied to monitoring and diagnosing the test data. Each set of test data contains 50 normal data points and 500 faulty points.

For comparison, three indices were utilized, including the false alarm rate, the fault detection rate, and the fault diagnosis rate. In MDM, only if the points are classified into the normal operation class and the corresponding values of the monitoring statistic are below the control limit of normal operation, these points are regarded as normal operation data. Otherwise, they are detected as faulty data. In addition, only if a detected faulty point is correctly classified, this point is diagnosed successfully. Based on such definitions, the false alarm rate is ratio between the number of the normal data points detected as a fault and the total number of the normal data points in the test data set. The fault detection rate is equal to the percentage of the faulty points that are detected, while the fault diagnosis rate is the percentage of the faulty points that are diagnosed correctly. In contrast, FDA conducts both fault detection and diagnosis using the information of classification. Therefore, its fault detection rate and fault diagnosis rate are the same.

All comparison results are listed in Table 1. In this table, the 20 types of faults are roughly divided into five groups. Faults 1, 2, 4, 6 and 7 belong to the first group, which are caused by the significant step changes of some variables and are easy to detect. Both MDM and FDA models performed well in the detection and diagnosis of such faults. The second group consists of faults 8, 12, 13, 17, 18, 19 and 20. In the monitoring and diagnosis of these faults, MDM outperformed FDA in both operation modes significantly. By investigating the detailed monitoring results that are not plotted here, it is found that large detection/diagnosis delay appeared in the FDA results, while MDM did not have such problem. This group of faults was mainly caused by slow drift or random variations in process characteristics, which were relatively insignificant in their starting periods. MDM showed its efficiency in detecting small changes. Faults 10, 11 and 14 form the third group. Again, the MDM model presented accurate detection and diagnosis results. However, FDA failed to differentiate the normal operation data and the faulty data, and provided high false alarm rate and low fault detection/diagnosis rate. In such cases, FDA could not find appropriate directions for classification, since the process data were non-Gaussian distributed. On the contrary, MDM was suited to handling such data distribution. The fourth group involves fault 5. Both MDM and FDA achieved good results in operation mode 3, but performed poorly in mode 1, where MDM showed high false alarm rate while FDA showed low fault detection/diagnosis rate. The analysis on the raw measurements indicated that the difference between the normal operation and the faulty operation was not obvious in mode 1. The last group includes faults 3, 9, 15 and 16 that have been shown to be difficult to detect.⁴¹ At the first glance, it seems that FDA had higher detection/diagnosis rates than MDM. However, since FDA's false alarm rates were also much higher than those of MDM, it was difficult to tell which methods performed better in such cases.

According to the above comparisons, it can be concluded that the overall performance of MDM was much better than that of FDA.

6. Conclusions

This paper proposed a fault detection, isolation and diagnosis method named mixture discriminant monitoring. As a hybrid method, MDM integrates supervised learning with the SPC charting technique and inherits the advantages of both techniques. It makes use of both normal and faulty data in process modeling. As a result, not only known types of faults but also new faults can be detected. For known types of faults, the diagnosis results are achieved automatically, while for unknown faults, the variables most responsible for the faults can be identified and can provide information to infer the root-causes. The effectiveness of the proposed method was verified using the benchmark Tennessee Eastman process.

It may be necessary to point out that the phenomenon of fault propagation may be observed in industrial processes. When such a phenomenon occurs, the fault characteristics become time-varying,⁴² and the distribution of process data may change along the time direction and show non-Gaussian characteristics. For known faults, the proposed MDM method can model such faulty data with mixture of Gaussians. Then, the models can be utilized in process monitoring and fault diagnosis. If such faults are unknown and not modeled with the MDM model, they can still be detected in online monitoring. However, in fault isolation, MDM does not consider time-varying fault characteristics, but only makes use of the first several data points after detection. The question of how fault diagnosis can benefit from the fault propagation information will be answered in the future study.

Future research should also focus on model updating. In industry, the amount of faulty data is usually much less than the amount of normal operation data. Therefore, it may be advantageous to first build an initial MDM model that contains only a small amount of faulty historical data. Then, when

more and more operation data are collected, the model should be updated recursively. This issue is out of the scope of this paper and will be studied in the future.

Acknowledgement

This work was supported in part by the Ministry of Economic Affairs, R.O.C. under Grant No. 102-EC-17-A-09-S1-198 and in part by the National Science Council of R.O.C. under Grant No. NSC 102-2623-E-007-007-IT.

Supporting Information Available

The supporting information contains a list of measured variables and a list of process faults involved in the case studies. This information is available free of charge via the Internet at <http://pubs.acs.org/>.

References

- (1) Kourti, T. The Process Analytical Technology initiative and multivariate process analysis, monitoring and control. *Anal. Bioanal. Chem.* **2006**, *384*, 1043.
- (2) Bersimis, S.; Psarakis, S.; Panaretos, J. Multivariate statistical process control charts: an overview. *Qual. Reliab. Eng. Int.* **2007**, *23*, 517.
- (3) AlGhazzawi, A.; Lennox, B. Monitoring a complex refining process using multivariate statistics. *Control Eng. Pract.* **2008**, *16*, 294.
- (4) Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Hum. Genet.* **1936**, *7*, 179.
- (5) Chiang, L. H.; Russell, E. L.; Braatz, R. D. Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 243.
- (6) Chiang, L. H.; Kotanchek, M. E.; Kordon, A. K. Fault diagnosis based on Fisher discriminant analysis and support vector machines. *Comput. Chem. Eng.* **2004**, *28*, 1389.
- (7) He, Q. P.; Qin, S. J.; Wang, J. A new fault diagnosis method using fault directions in Fisher discriminant analysis. *AIChE J.* **2005**, *51*, 555.
- (8) Choi, S. W.; Park, J. H.; Lee, I. B. Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. *Comput. Chem. Eng.* **2004**, *28*, 1377.
- (9) Yu, J. Localized Fisher discriminant analysis based complex chemical process monitoring. *AIChE J.* **2011**, *57*, 1817.
- (10) Hwang, D.; Han, C. Real-time monitoring for a process with multiple operating modes. *Control Eng. Pract.* **1999**, *7*, 891.
- (11) Chen, J.; Liu, J. Mixture principal component analysis models for process monitoring. *Ind. Eng. Chem. Res.* **1999**, *38*, 1478.
- (12) Lane, S.; Martin, E.; Kooijmans, R.; Morris, A. Performance monitoring of a multi-product semi-batch process. *J. Process Control.* **2001**, *11*, 1.

- (13) Zhao, S.; Zhang, J.; Xu, Y. Monitoring of processes with multiple operating modes through multiple principle component analysis models. *Ind. Eng. Chem. Res.* **2004**, *43*, 7025.
- (14) Zhao, S.; Zhang, J.; Xu, Y. Performance monitoring of processes with multiple operating modes through multiple PLS models. *J. Process Control.* **2006**, *16*, 763.
- (15) Lee, J. M.; Yoo, C. K.; Lee, I. B. Statistical process monitoring with independent component analysis. *J. Process Control.* **2004**, *14*, 467.
- (16) Ge, Z.; Song, Z. Process monitoring based on independent component analysis-principal component analysis (ICA-PCA) and similarity factors. *Ind. Eng. Chem. Res.* **2007**, *46*, 2054.
- (17) Liu, X.; Xie, L.; Kruger, U.; Littler, T.; Wang, S. Statistical-based monitoring of multivariate non-Gaussian systems. *AIChE J.* **2008**, *54*, 2379.
- (18) Rashid, M. M.; Yu, J. A new dissimilarity method integrating multidimensional mutual information and independent component analysis for non-Gaussian dynamic process monitoring. *Chemom. Intell. Lab. Syst.* **2012**, *115*, 44.
- (19) Thissen, U.; Swierenga, H.; De Weijer, A.; Wehrens, R.; Melssen, W.; Buydens, L. Multivariate statistical process control using mixture modelling. *J. Chemom.* **2005**, *19*, 23.
- (20) Yu, J.; Qin, S. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE J.* **2008**, *54*, 1811.
- (21) Yu, J. A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. *Chem. Eng. Sci.* **2011**, *68*, 506.
- (22) Zhao, C.; Yao, Y.; Gao, F.; Wang, F. Statistical analysis and online monitoring for multimode processes with between-mode transitions. *Chem. Eng. Sci.* **2010**, *65*, 5961.
- (23) Detroja, K. P.; Gudi, R. D.; Patwardhan, S. C. A possibilistic clustering approach to novel fault detection and isolation. *J. Process Control.* **2006**, *16*, 1055.
- (24) Hastie, T.; Tibshirani, R. Discriminant analysis by Gaussian mixtures. *J. Roy. Statist. Soc. Ser. B.* **1996**, *58*, 155.
- (25) Fraley, C.; Raftery, A. E. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **2002**, *97*, 611.
- (26) Chen, T.; Morris, J.; Martin, E. Probability density estimation via an infinite Gaussian mixture model: application to statistical process monitoring. *J. Roy. Statist. Soc. Ser. C.* **2006**, *55*, 699.
- (27) Bishop, C. M. *Pattern recognition and machine learning*; 1 ed; Springer: New York, 2006.
- (28) Dayal, B.; MacGregor, J. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. *J. Process Control.* **1997**, *7*, 169.
- (29) Qin, S. Recursive PLS algorithms for adaptive data modeling. *Comput. Chem. Eng.* **1998**, *22*, 503.
- (30) Li, W.; Yue, H.; Valle-Cervantes, S.; Qin, S. Recursive PCA for adaptive process monitoring. *J. Process Control.* **2000**, *10*, 471.
- (31) Wang, X.; Kruger, U.; Irwin, G. W. Process monitoring approach using fast moving window PCA. *Ind. Eng. Chem. Res.* **2005**, *44*, 5691.

- (32) Miller, P.; Swanson, R.; Heckler, C. Contribution plots: a missing link in multivariate quality control. *Appl. Math. Comput. Sci.* **1998**, *8*, 775.
- (33) Conlin, A.; Martin, E.; Morris, A. Confidence limits for contribution plots. *J. Chemom.* **2000**, *14*, 725.
- (34) Qin, S. J. Statistical process monitoring: basics and beyond. *J. Chemom.* **2003**, *17*, 480.
- (35) Yue, H. H.; Qin, S. J. Reconstruction-based fault identification using a combined index. *Ind. Eng. Chem. Res.* **2001**, *40*, 4403.
- (36) Cohn, D.; Ghahramani, Z.; Jordan, M. Active learning with statistical models. *J Artif. Intell. Res.* **1996**, *4*, 129.
- (37) Kariwala, V.; Odiowei, P. E.; Cao, Y.; Chen, T. A branch and bound method for isolation of faulty variables through missing variable analysis. *J. Process Control.* **2010**, *20*, 1198.
- (38) Yang, F.; Xiao, D. Progress in root cause and fault propagation analysis of large-scale industrial processes. *J. Control Sci. Eng.* **2012**, *2012*,
- (39) Downs, J.; Vogel, E. A plant-wide industrial process control problem. *Comput. Chem. Eng.* **1993**, *17*, 245.
- (40) Ricker, L. Decentralized control of the Tennessee Eastman challenge process. *J. Process Control.* **1996**, *6*, 205.
- (41) Yin, S.; Ding, S. X.; Haghani, A.; Hao, H.; Zhang, P. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *J. Process Control.* **2012**, *22*, 1567–1581.
- (42) Zhao, C.; Sun, Y.; Gao, F. A Multiple-Time-Region (MTR)-Based Fault Subspace Decomposition and Reconstruction Modeling Strategy for Online Fault Diagnosis. *Ind. Eng. Chem. Res.* **2012**, *51*, 11207.

Figure list:

Figure 1. Illustration of normal and faulty process operating regions

Figure 2. Flowchart of (a) MDM-based process modeling and (b) process monitoring.

Figure 3. A brief flow diagram of the TE process.

Figure 4. Comparison of two process models: (a) FDA monitoring results of the training data, and (b) MDM monitoring results of the training data.

Figure 5. Monitoring results of known faults in mode 1 based on (a) FDA, and (b) MDM.

Figure 6. Monitoring results of known faults in mode 3 based on (a) FDA, and (b) MDM.

Figure 7. Monitoring results of unknown fault 5 based on (a) FDA, and (b) MDM.

Figure 8. Monitoring results of unknown fault 10 based on (a) FDA, and (b) MDM.

Figure 9. Monitoring results of simultaneous faults based on (a) FDA, and (b) MDM.

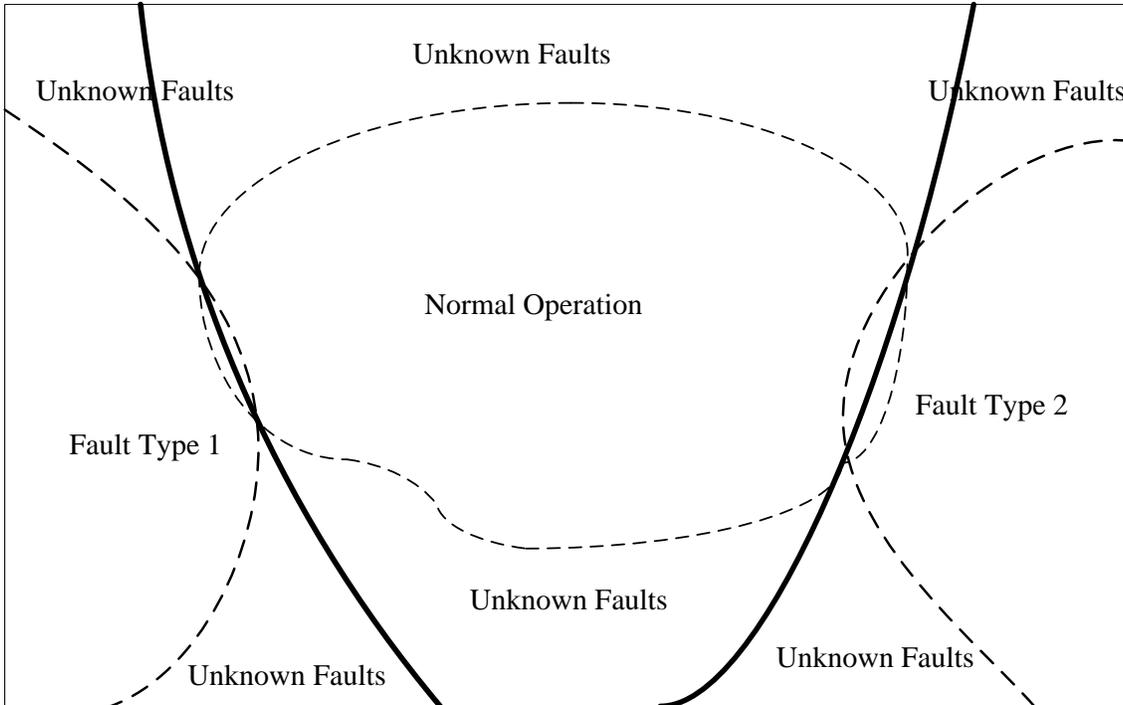


Figure 1. Illustration of normal and faulty process operating regions (the dashed curves represent the confidence bounds of known operating regions; the solid curves represent the classification boundaries).

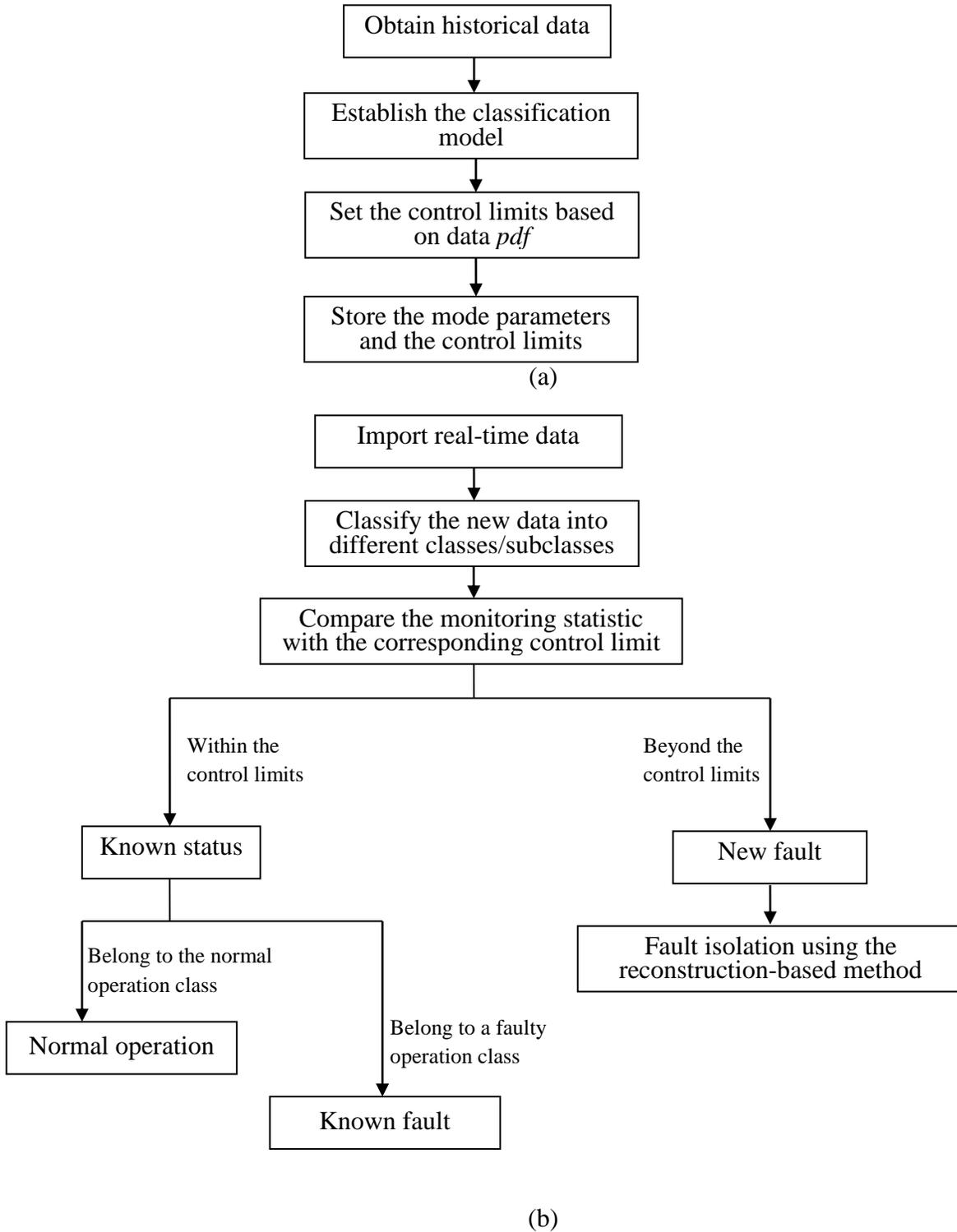


Figure 2. Flowchart of (a) MDM-based process modeling and (b) process monitoring.

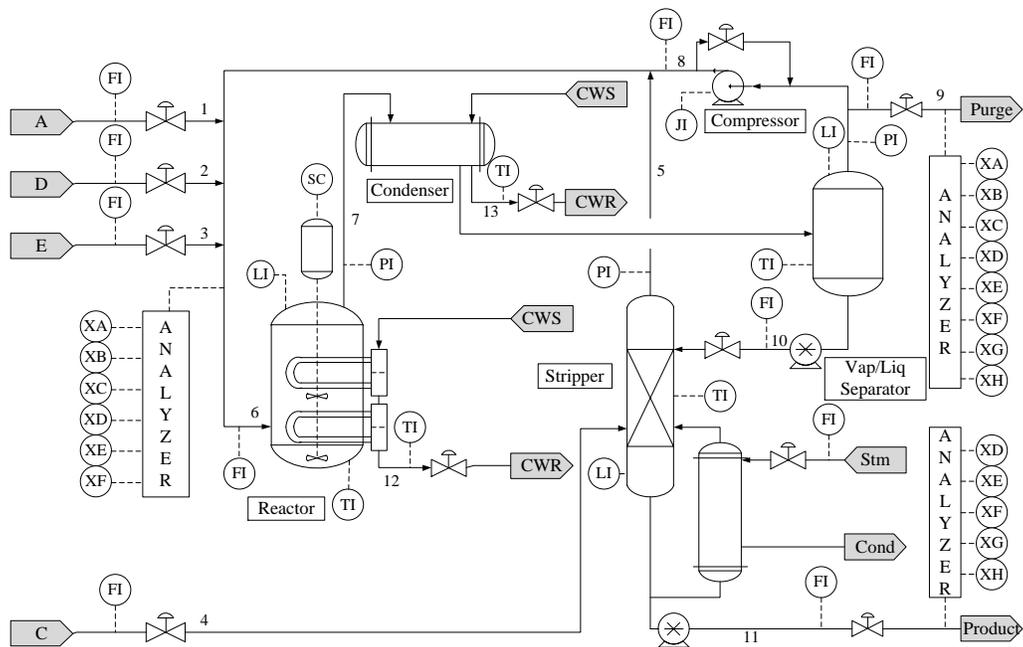
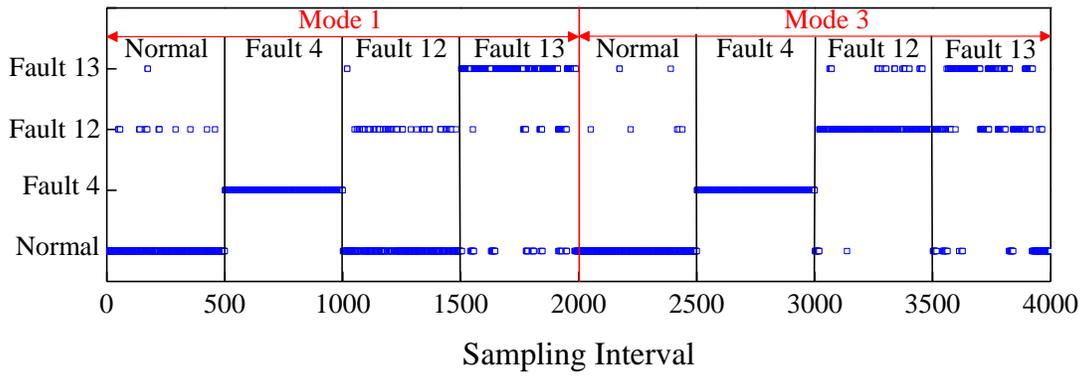
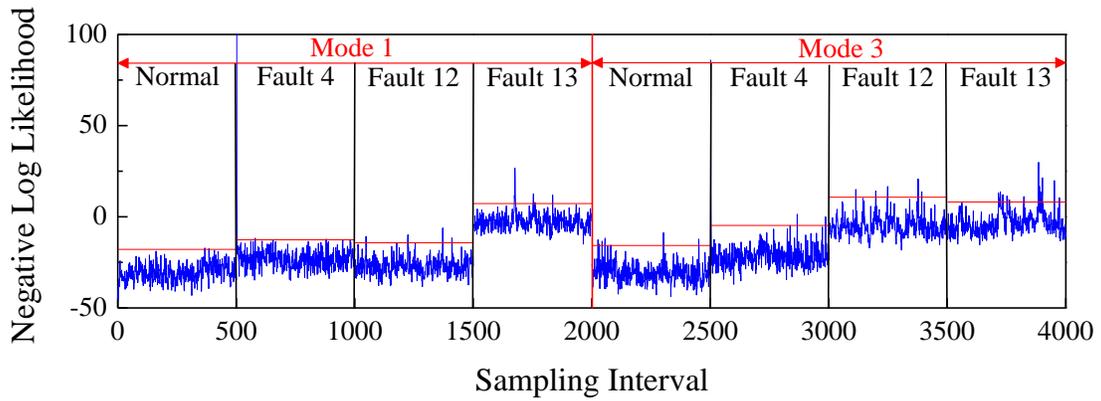


Figure 3. A brief flow diagram of the TE process.

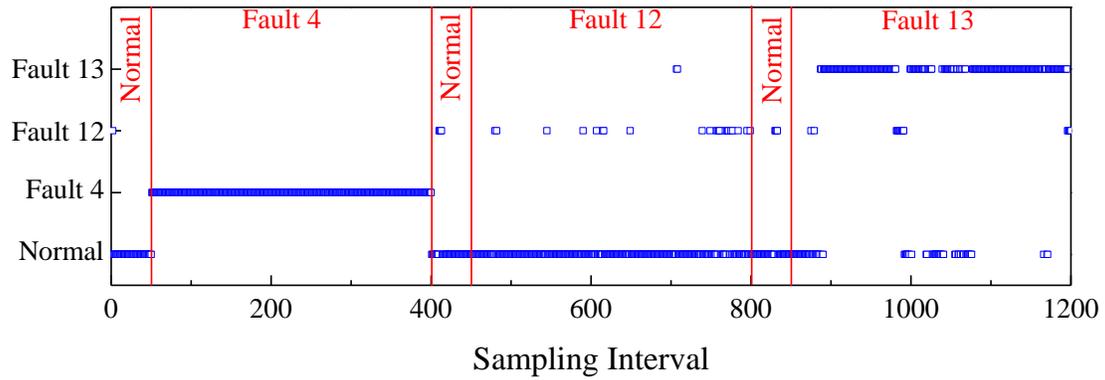


(a)

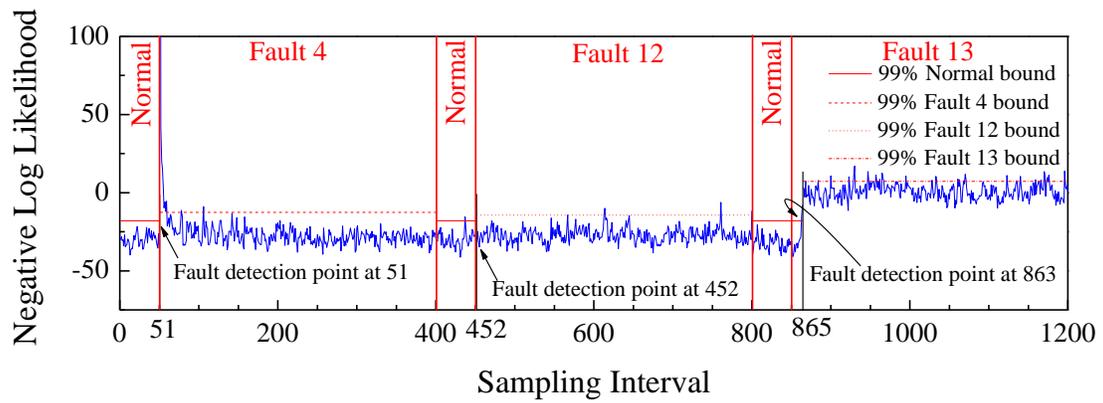


(b)

Figure 4. Comparison of two process models: (a) FDA monitoring results of the training data, and (b) MDM monitoring results of the training data.

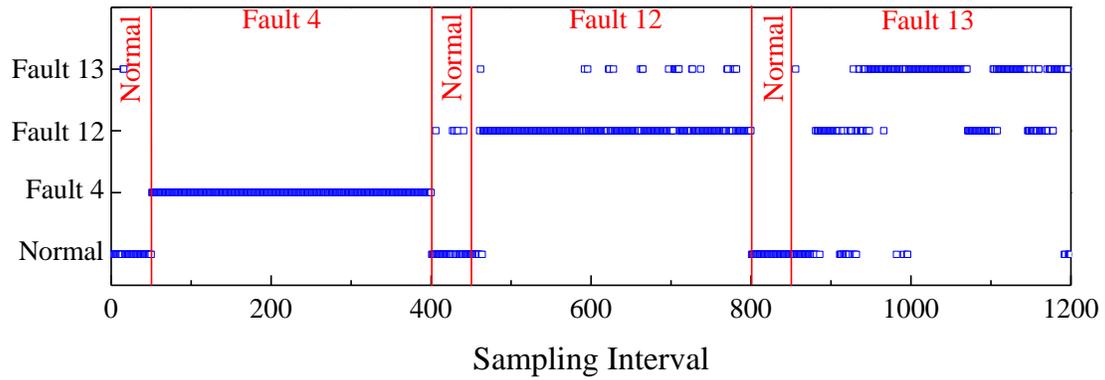


(a)

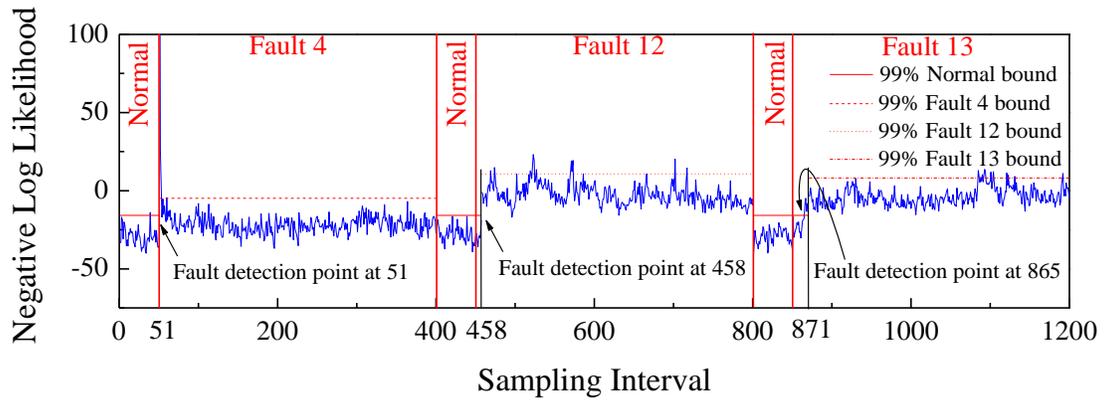


(b)

Figure 5. Monitoring results of known faults in mode 1 based on (a) FDA, and (b) MDM.

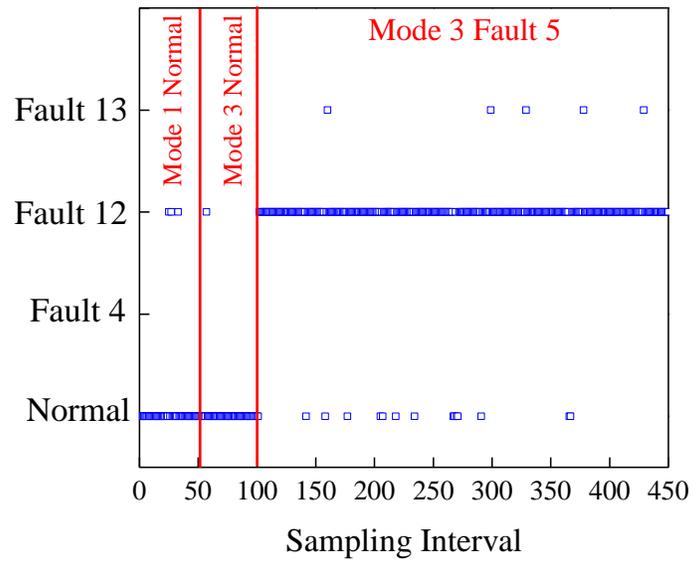


(a)

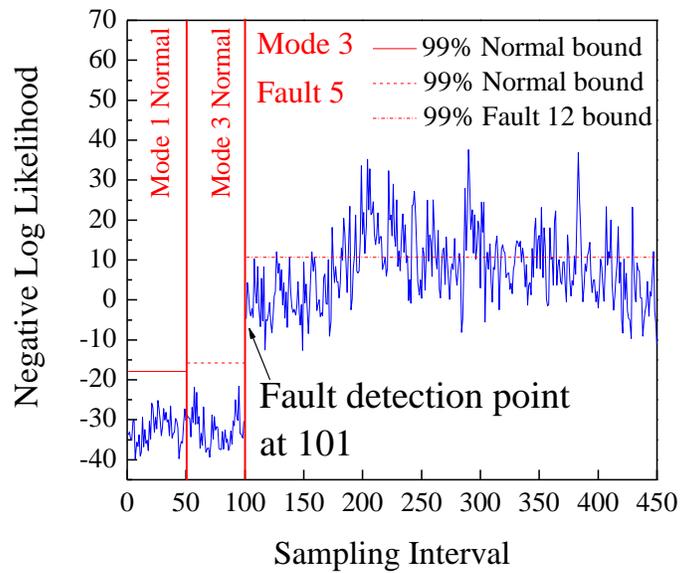


(b)

Figure 6. Monitoring results of known faults in mode 3 based on (a) FDA, and (b) MDM.

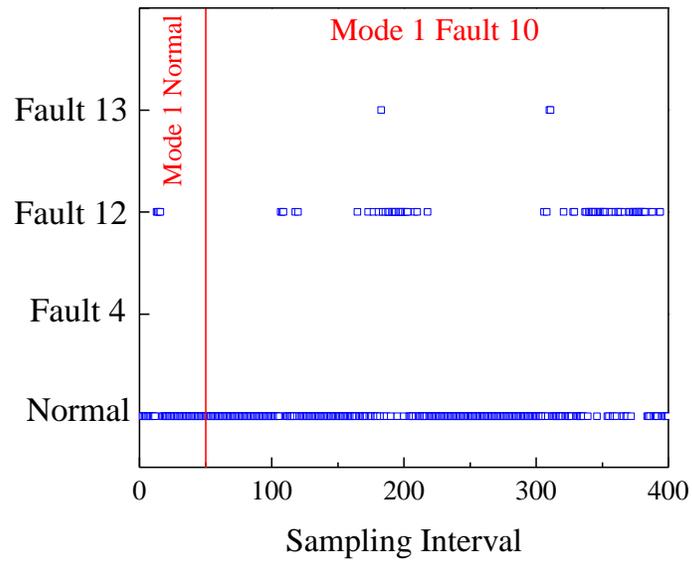


(a)

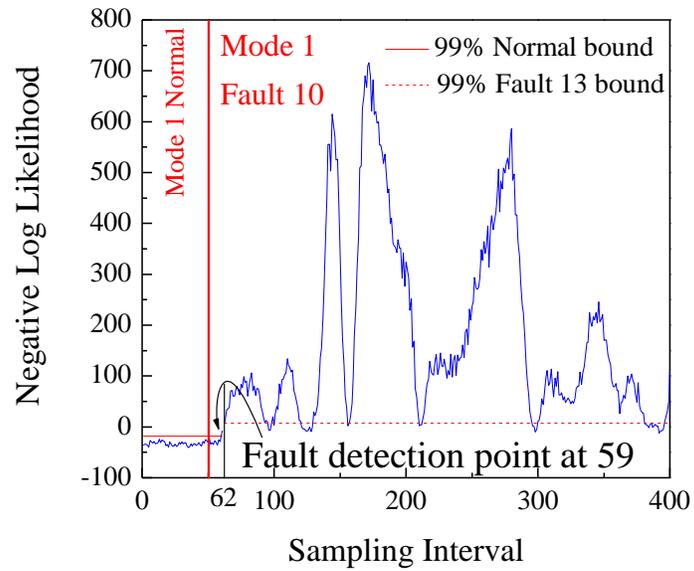


(b)

Figure 7. Monitoring results of unknown fault 5 based on (a) FDA, and (b) MDM.

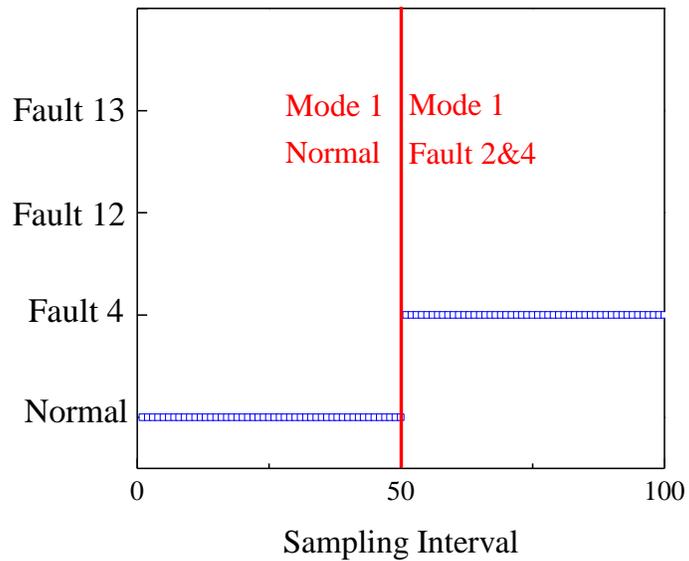


(a)

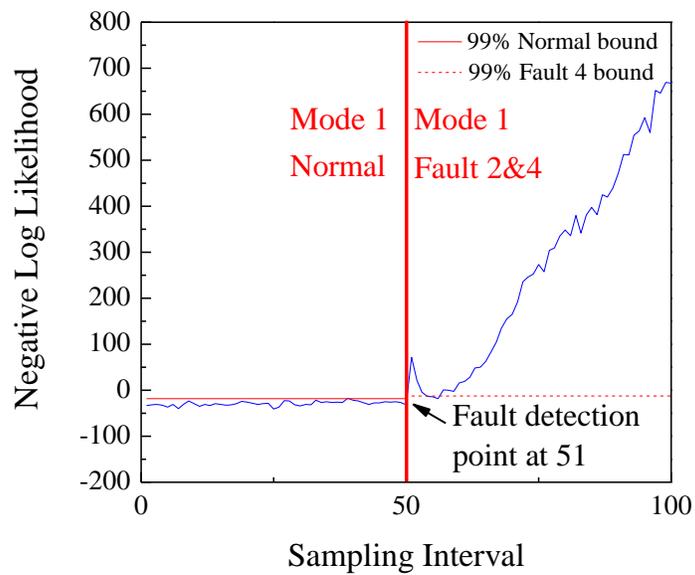


(b)

Figure 8. Monitoring results of unknown fault 10 based on (a) FDA, and (b) MDM.



(a)



(b)

Figure 9. Monitoring results of simultaneous faults based on (a) FDA, and (b) MDM.

Table 1. Overall comparison between MDM and FDA

Fault ID	MDM						FDA			
	Mode 1			Mode 3			Mode 1		Mode 3	
	False alarm rate (%)	Fault detection rate (%)	Fault diagnosis rate (%)	False alarm rate (%)	Fault detection rate (%)	Fault diagnosis rate (%)	False alarm rate (%)	Fault detection/ diagnosis rate (%)	False alarm rate (%)	Fault detection/ diagnosis rate (%)
1	0	99.8	99.4	2	99.6	99.4	0	98.8	0	98
2	0	99.4	98.8	2	97.8	97.8	0	96.8	0	96
4	0	100	100	2	100	100	0	100	0	100
6	0	100	100	0	100	100	0	100	0	100
7	0	100	100	0	100	100	0	100	0	100
8	0	95.4	95	0	96.4	96.2	0	63.4	0	85.6
12	2	98.4	98.2	0	98.6	98	0	30	2	85.8
13	0	96.6	96.2	0	95.8	94	0	77.8	0	65.6
17	0	93.8	93.6	0	91.8	91.8	0	68	0	69
18	0	87.2	85.6	2	94.4	94.2	0	55.4	0	87.2
19	4	99	99	0	97	97	0	61.2	0	74.2
20	4	95.6	95.2	0	94.4	94.4	0	90.6	0	71.8
10	0	93.4	93	2	91.4	91	26	59.2	48	59.4
11	0	99	99	0	99.2	99.2	82	59	82	66.4
14	0	99.8	99.8	0	100	100	48	69.2	68	74
5	30	51.8	51	2	100	100	0	0	0	100
3	2	20.4	18.6	0	36.2	34.4	20	75.8	16	91.8

9	12	65.4	64.2	4	62.8	60.4	64	67.8	68	66.4
15	2	8.2	7.6	2	22.8	21.8	44	77.2	56	75.8
16	6	4.8	3.6	0	25.2	23.6	66	65.6	72	69.8