Bayesian method for multimode non-Gaussian process monitoring

Zhiqiang Ge, Zhihuan Song*, Muguang Zhang, Ruowei Fu, and Zhibo Zhu

Abstract—Non-Gaussian processes monitoring has recently caught much attentions in this area, with several methods successfully developed, such as non-parameter estimation, independent component analysis (ICA), support vector data description (SVDD), and etc. However, most of current research works are under the assumption that the process is operated in a single mode. This paper proposed a novel method for monitoring multimode non-Gaussian processes, which is based on Bayesian inference. To improve the comprehension of the process for the operation engineer, a corresponding mode localization approach is also given. A case study on the Tennessee Eastman (TE) benchmark process shows the feasibility and efficiency of the proposed method.

I. INTRODUCTION

Data-based methods for process monitoring have been a hot research spot since the last decade, especially the multivariate statistical process control (MSPC) method [1-3]. Nowadays, non-Gaussian process monitoring has caught a lot of attentions in this area, which can be traced to the last decade [4-6]. As an efficient non-Gaussian information extraction technology, independent component analysis (ICA) has been introduced and intensively researched for process monitoring in recent years [7-10]. However, most of the proposed methods are restricted to single mode non-Gaussian processes, only few research studies have been carried out for monitoring multimode non-Gaussian processes. In our opinion, this new kind of process widely exists in modern industrial areas. Therefore, monitoring method development for this special kind of process should be regarded as a new contribution to this area.

In contrast, multimode monitoring for traditional processes has been studied for many years, and several different methods have been proposed, such as recursive approaches, multiple-model methods, and etc. [11-17]. Particularly, the Gaussian Mixture Model (GMM) based method was considered as an efficient way for monitoring multimode processes, which has also been intensively researched in recent years [15-17]. However, GMM has an assumption that each single mode dataset should be Gaussian distributed, thus can be modeled by a Gaussian distribution model. For those processes whose dataset is not Gaussian distributed in each single mode, the condition of GMM would be violated. In this case, a new monitoring viewpoint should be proposed.

This paper aims to develop a new method for monitoring multimode non-Gaussian processes, and simultaneously improve the process comprehension for the process engineer. First, to carry out the method through unsupervised manner, data clustering methods such as C-means clustering are employed for partitioning the process dataset into several different sub-groups. Second, a recent developed ICA and support vector data description (SVDD) model structure is used for non-Gaussian modeling in each sub-group dataset. Then the Bayesian inference strategy is introduced for online calculation of mode posterior probability for the current data sample, which is followed by the final monitoring decision development step. Besides, one may intend to know which operation mode the current monitored data sample belongs to. Basically, the mode information will improve the comprehension of the operation engineer for the process, which we call “mode localization” in this paper. The rest parts of the paper are organized as follows. In the next section, the proposed monitoring and mode localization method is detailedly demonstrated, which is followed by a case study on the TE benchmark process. Finally, some conclusions are made.

II. METHOD DEVELOPMENT AND ANALYSIS

As been discussed, the process dataset should be partitioned into different sub-groups in the first step. Common data clustering methods such as K-means, C-means or other related technologies can be employed. Suppose the original process dataset is denoted as $X \in R^{m \times n}$, where $m$ is the number of process variables, and $n$ is the number of data samples. After data clustering and partitioning, the sub-group datasets can be represented as

$$X = [\mathbf{X}_1 \mid \mathbf{X}_2 \mid \cdots \mathbf{X}_c \mid \cdots \mathbf{X}_C]$$

(1)

where $\mathbf{X}_c \in R^{m \times n_c}$ ($c = 1, 2, \cdots, C$) is the $c$-th sub-group dataset, with $n_c$ data samples, $C$ is the number of operation modes (clusters). Then the monitoring model can be developed in each operation mode, and the Bayesian-based multimode process monitoring and mode localization approach can be constructed, which are demonstrated in the following subsections.
A. Monitoring model and statistic development in each operation mode

In this paper, ICA is employed for non-Gaussian information extraction and dimension reduction in each sub-group dataset, which is given as follow [7]

\[ X_c = A_c \hat{S}_c + E_c \]

\[ \hat{S}_c = W_c X_c \]

\[ E_c = X_c - A_c \hat{S}_c \]

(2)

where \( A_c \) and \( W_c \) is the mixing and demixing matrix, \( E_c \) is the residual matrix after non-Gaussian components have been extracted, which can be considered as Gaussian distribution dataset.

After the information extraction step, we are in the position to construct monitoring statistics for both of the non-Gaussian component and residual subspaces. To improved the deficiency of the traditional \( I^2 \) monitoring statistic developed by Lee, et. al [8], SVDD method was recently introduced for constructing more efficient non-Gaussian component monitoring statistic [10]. In the present paper, the ICA-SVDD model and monitoring structure is used in each sub-group dataset, which can be described as

\[
\min_{R_c, a_c, \xi_c} R_c^2 + C \sum_{i=1}^{n} \xi_{c,i,j} \]

\[
st. \left\| \Phi(\hat{s}_{c,i,j}) - a_c \right\|^2 \leq R_c^2 + \xi_{c,i,j}, \xi_{c,i,j} \geq 0 \]

(3)

where \( c = 1,2,\cdots,C \), \( \hat{s}_{c,i,j} \) is the component vector of i-the data sample in operation mode \( c \), \( a_c \) is the center of the hypersphere, \( C \) gives the trade-off between the volume of the hypersphere and the number of errors. \( \xi_{c,i,j} \) represents the slack variable which allows the probability that some of the training samples can be wrongly classified. Suppose \( K(\cdot) \) is the kernel function which is often selected as the Gaussian kernel function, the center \( a_c \) and the radius \( R_c \) of SVDD model for operation mode \( c \) can be determined by [10]

\[
a_c = \sum_{j=1}^{N} \alpha_j \Phi(\hat{s}_{c,j}) \]

\[
R_c = \sqrt{1 - 2 \sum_{i=1}^{d} \alpha_i K(\hat{s}_{c,0}, \hat{s}_{c,i}) + \sum_{i=1}^{d} \sum_{j=1}^{d} \alpha_i \alpha_j K(\hat{s}_{c,i}, \hat{s}_{c,j})} \]

(4)

where \( \hat{s}_{c,0} \) is referred to a support vector of this SVDD model.

Then the new non-Gaussian monitoring statistic can be defined as

\[
NGS_{c,i,j} = d^2(\Phi(\hat{s}_{c,i,j})) - \left\| \Phi(\hat{s}_{c,i,j}) - a_c \right\|^2 \leq NGS_{c,i,j}^2 = R_c^2 \]

(5)

For the residual information, the traditional \( SPE \) monitoring statistic can be constructed [1]

\[
SPE_i = e^T c e_i \leq SPE_{c,\text{lim}} = g_c X_{\text{lim}}^2 \]

(6)

where \( c = 1,2,\cdots,C \), \( \alpha \) is the selected significance level, and \( g_c = v_c / (2m_c) \), \( h_c = 2m_c^2 / v_c \), in which \( m_c \) and \( v_c \) are the mean and variance values of \( SPE \) within operation mode \( c \).

B. Multimode monitoring based on Bayesian inference

For online monitoring, denote the new data sample as \( x_{\text{new}} \), the new monitoring strategy can be described as follows. First, \( x_{\text{new}} \) is auto-scaled and monitored by each sub-group model, thus

\[
\hat{s}_{c,\text{new}} = W_c x_{\text{new}} \]

\[
NGS_{c,\text{new}} = d^2(\Phi(\hat{s}_{c,\text{new}})) - \left\| \Phi(\hat{s}_{c,\text{new}}) - a_c \right\|^2 \]

\[
e_{c,\text{new}} = x_{\text{new}} - A_c \hat{s}_{c,\text{new}} \]

\[
SPE_{c,\text{new}} = e^T_{c,\text{new}} e_{c,\text{new}} \]

(7)

(8)

where \( c = 1,2,\cdots,C \). Before the introduction of Bayesian inference strategy for posterior probability calculation for the new data sample in different operation modes, a transformation from the monitoring statistical value to the probability distribution value should be made to facilitate the Bayesian inference, which is given as follows

\[
P_{NGS}(x_{\text{new}} | c) = \exp\left\{ -\frac{NGS_{c,\text{new}}}{NGS_{c,\text{lim}}} \right\} \]

(9)

\[
P_{SPE}(x_{\text{new}} | c) = \exp\left\{ -\frac{SPE_{c,\text{new}}}{SPE_{c,\text{lim}}} \right\} \]

(10)

Then the posterior probability of the new data sample belongs to each operation mode can be calculated as [18]

\[
P_{NGS}(c | x_{\text{new}}) = \frac{P_{NGS}(c, x_{\text{new}})}{\sum_{c=1}^{C} P_{NGS}(c, x_{\text{new}}) P(c)} = \frac{P_{NGS}(x_{\text{new}} | c) P(c)}{\sum_{c=1}^{C} P_{NGS}(x_{\text{new}} | c) P(c)} \]

(11)

\[
P_{SPE}(c | x_{\text{new}}) = \frac{P_{SPE}(c, x_{\text{new}})}{\sum_{c=1}^{C} P_{SPE}(c, x_{\text{new}}) P(c)} = \frac{P_{SPE}(x_{\text{new}} | c) P(c)}{\sum_{c=1}^{C} P_{SPE}(x_{\text{new}} | c) P(c)} \]

(12)

where \( P(c), c = 1,2,\cdots,C \) are prior probabilities for each operation mode, which can be simply determined as

\[
P(c) = \frac{n_c}{n} \]

(13)

After the posterior probabilities of the new data sample \( x_{\text{new}} \) have been obtained, we should decide its fault probability in each operation mode, which can be calculated as [17]
where \( \mathbf{x}_{c,t} \) is the train sample in each operation mode. Fault probability values in eqs. (14) and (15) can be simply determined by measuring the number of training samples whose statistic values are smaller than that of the new data sample. Precisely, they can also be determined by kernel density estimation with appropriate significant level.

Finally, the monitoring results in each operation mode can be combined together, thus two new Bayesian based monitoring statistics can be constructed as

\[
BMS_{NGS}(\mathbf{x}_{\text{new}}) = \sum_{c=1}^{C} [P_{NGS}(c|\mathbf{x}_{\text{new}})P_{f,NGS}^{c}(\mathbf{x}_{\text{new}})]
\]

and

\[
BMS_{SPE}(\mathbf{x}_{\text{new}}) = \sum_{c=1}^{C} [P_{SPE}(c|\mathbf{x}_{\text{new}})P_{f,SPE}^{c}(\mathbf{x}_{\text{new}})]
\]

It is noticed that the values of \( P_{f,NGS}(\mathbf{x}_{\text{new}}) \), and \( P_{f,SPE}(\mathbf{x}_{\text{new}}) \) are both ranged from 0 to 1, and the posterior probabilities \( P_{NGS}(c|\mathbf{x}_{\text{new}}) \) and \( P_{SPE}(c|\mathbf{x}_{\text{new}}) \) are restricted as

\[
\sum_{c=1}^{C} P_{NGS}(c|\mathbf{x}_{\text{new}}) = 1 \quad \text{and} \quad \sum_{c=1}^{C} P_{SPE}(c|\mathbf{x}_{\text{new}}) = 1
\]

Therefore, the combined statistics \( BMS_{NGS}(\mathbf{x}_{\text{new}}) \) and \( BMS_{SPE}(\mathbf{x}_{\text{new}}) \) are also between 0 and 1. If we set the confidence level as \( 1 - \alpha \), then the process will be considered to be abnormal when either of the monitoring statistic value exceeds \( 1 - \alpha \).

### C. Mode localization

Compared to traditional methods, this new proposed monitoring strategy does not need to know which operation mode the current data sample belongs to. Hence, no process knowledge need to be incorporated to determine which monitoring model should be used. Instead, all monitoring models in different operation modes are employed. The new method treats the monitoring task in a probabilistic manner, which can also be regarded as a soft monitoring strategy. Although the mode information of the monitored data sample is not necessary for process monitoring, this is very important for operation and further comprehension of the process. Depending on the posterior probabilities obtained in eqs. (11) and (12), the mode information of the new data sample can be determined. However, based on the normalization of the posterior probability, the new data sample will be assigned to the operation mode which has the largest posterior probability value. A possible pitfall of this method is that, if the new data sample comes from an unknown operation mode, it will also be assigned to one of the known operation mode that has been identified before. Hence, if we use the posterior probability for mode identification, there will be inevitable false identifications and misunderstandings of the process which may cause further risks.

To solve this problem, the joint probability of the monitored data sample and each operation mode can be utilized, which are given as

\[
P_{NGS}(\mathbf{x}_{\text{new}}, c) = P_{NGS}(\mathbf{x}_{\text{new}}|c)P(c)
\]

and

\[
P_{SPE}(\mathbf{x}_{\text{new}}, c) = P_{SPE}(\mathbf{x}_{\text{new}}|c)P(c)
\]

Different from the posterior probability, the joint probability can not only work well when the new data sample belongs to some known operation mode, but also it can successfully identify the change of the process. Hence, if the new data sample belongs to a new operation mode, the joint probabilities of the monitored data sample with all known operation modes will approach to zero. Similarly, when a fault happens, the joint probabilities of the monitored faulty sample with all known operation modes will also become very small values.

### III. TE BENCHMARK CASE STUDY

TE process is a benchmark process which has been widely used to test the performance of various control monitoring approaches [2]. This process consists of 41 measured variables and 12 manipulated variables, a set of 20 programmed faults can be simulated in this process. The details of the process description can be found in Downs and Vogel [19]. In the present paper, 16 continuous process variables are selected for process monitoring, which are listed in Table 1. There are 6 operation modes available for simulation case studies, two of which (mode 1 and mode 3) are selected for the simulation of multimode manner of the process, which are represented as mode 1 and mode 2 in this paper.

<table>
<thead>
<tr>
<th>No.</th>
<th>Measured variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A feed</td>
</tr>
<tr>
<td>2</td>
<td>D feed</td>
</tr>
<tr>
<td>3</td>
<td>E feed</td>
</tr>
<tr>
<td>4</td>
<td>A and C feed</td>
</tr>
<tr>
<td>5</td>
<td>Recycle flow</td>
</tr>
<tr>
<td>6</td>
<td>Reactor feed rate</td>
</tr>
<tr>
<td>7</td>
<td>Reactor temperature</td>
</tr>
<tr>
<td>8</td>
<td>Purge rate</td>
</tr>
<tr>
<td>9</td>
<td>Product separator temperature</td>
</tr>
<tr>
<td>10</td>
<td>Product separator pressure</td>
</tr>
<tr>
<td>11</td>
<td>Product separator underflow</td>
</tr>
<tr>
<td>12</td>
<td>Stripper pressure</td>
</tr>
<tr>
<td>13</td>
<td>Stripper temperature</td>
</tr>
<tr>
<td>14</td>
<td>Stripper steam flow</td>
</tr>
<tr>
<td>15</td>
<td>Reactor cooling water outlet temperature</td>
</tr>
<tr>
<td>16</td>
<td>Separator cooling water outlet temperature</td>
</tr>
</tbody>
</table>
For the monitoring model construction, 1000 training samples have been collected under each operation mode. To develop the ICA-SVDD model, 3 independent components are selected and the SVDD parameters are chosen as $C = 0.1$ and $\sigma = 5$ (width parameter of the Gaussian kernel). The confidence levels of all monitoring statistics are set as 99%. To test the feasibility and efficiency of the proposed method, a normal dataset and several faulty datasets are generated separately, all of which have a total of 1000 samples. Among the 1000 samples in the normal dataset, the first 500 samples are from mode 1, and the last 500 ones are collected under mode 2. The monitoring results of this normal dataset are given in Fig. 1. It is very clear that the process is considered to be in good condition based on these monitoring results. However, if we monitor this dataset with a single model which is developed under only one operation mode, the monitoring results are different, which are shown in Fig. 2. As seen in this figure, the last 500 data samples are judged to be abnormal by the monitoring model built in the first operation mode. On the other hand, the first 500 data samples are abnormal based on the second monitoring model. To be clear, the joint probability of these 1000 data samples with different operation modes can be calculated, which are plotted in Fig. 3. Incorporating the mode localization results with Fig. 2, one can find that the first 500 data samples are collected under the first operation mode, while the other 500 ones are from mode 2.
To evaluate the fault detection ability of the proposed method, two generated faulty datasets are employed, both of which have 1000 data samples. The first faulty dataset was generated under mode 1, with 500 samples collected in normal condition and then fault 10 was introduced. The data structure of the second faulty dataset was the same as the first one except that it was collected under the second mode. The monitoring results of these two faulty datasets are given in Fig. 4 (a) and (b), which both indicate that the fault has been detected. The advantage of the ICA-SVDD model to other methods in monitoring non-Gaussian processes has already been reported, hence, it is omitted here. However, the focus of the present paper is multimode non-Gaussian process monitoring and the mode localization. By examining the mode localization results of the first faulty datasets in Fig. 5 (a), one can easily conclude that the fault was introduced under the first operation mode. Similar conclusion can be made based on the mode localization results in Fig. 5 (b). It is worth noticed that the joint probabilities of the first faulty dataset with the second operation mode and the second faulty dataset with the first operation mode are both very small (comparable to zero). Due to the page limitation, these results are not shown.
IV. CONCLUSIONS

This paper has proposed a novel method for multimode non-Gaussian process monitoring, which is based on the integrated ICA-SVDD model and the Bayesian inference. With this new monitoring framework, the traditional statistic value is transformed to the fault probability, which can be combined more easily to form the final monitoring decision. Besides, the mode information of the monitored data sample can also be obtained under this probabilistic monitoring structure, which can greatly improve the comprehension for the process engineer and thus additionally enhance the reliability of the process. Further research studies upon the proposed monitoring framework include fault identification, fault isolation and etc.

REFERENCES