Received: 20 November 2017 | Revised: 20 March 2018 | Accepted: 1 May 2018

DOI: 10.1002/cem.3048

RESEARCH ARTICLE

WILEY Journal of CHEMOMETRICS

Multimode complex process monitoring using double-level local information based local outlier factor method

Lei Wang | Xiaogang Deng | Yuping Cao

College of Information and Control Engineering, China University of Petroleum, Qingdao 266580, China

Correspondence
Xiaogang Deng, College of Information and Control Engineering, China University of Petroleum, Qingdao 266580, China.
Email: dengxiaogang@upc.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 21606256 and 61403418; Postgraduate Innovation Project of China University of Petroleum, Grant/Award Number: YCX2017058; Fundamental Research Funds for the Central Universities, Grant/Award Number: 17CX02054; Natural Science Foundation of Shandong Province, China, Grant/Award Numbers: ZR2014FL016, ZR2016FQ21 and ZR2016BQ14; Shandong Provincial Key Programme of Research & Development, Grant/Award Number: 2018GGX101025

Abstract

Industrial processes typically have multiple operating modes with complex data distribution and locality faults, which challenges the traditional multivariate statistical process monitoring methods. To address this problem, a double-level local information-based local outlier factor (LOF) method is proposed in this work for multimode complex process monitoring. First, to handle the multimodality, the local neighborhood standardization strategy is adopted to utilize the statistical information of local data structure. Second, the variable LOF method is proposed to determine reasonable boundary for complex data distribution and simultaneously reflect local variable behaviors. For better online monitoring, a weighting strategy is applied to emphasize the local variable information, and the Bayesian inference is employed to integrate the LOF value of each variable. To isolate the fault variables, a contribution plot is designed. Finally, a numerical example and the benchmark Tennessee Eastman process are used to demonstrate the effectiveness and superiority of the proposed method.

KEYWORDS

Bayesian inference, contribution plot, double-level local information, local outlier factor, multimode process monitoring

1 | INTRODUCTION

During the last decades, industrial process monitoring approaches have developed rapidly to fit the increasingly urgent demands of ensuring processes safety and improving product quality.[1-6] Due to the extensive applications of distributed control systems and sensor networks in modern industrial processes, a huge amount of process data are available and data-driven process monitoring is attracting more and more research interests. Traditional multivariate statistical process monitoring (MSPM) methods, such as principal component analysis (PCA),[7] partial least squares,[8] independent component analysis,[9] and Fisher discriminant analysis,[10] have shown significant success in many cases. However, there are still some valuable problems deserving further investigation. The main problems limiting the application of MSPM methods are often caused by 2 issues of unimodal assumption and Gaussian distribution.

In practical industrial processes, operating conditions often change because of various factors, including diverse manufacturing operating strategies, different product specifications, changing market demands, and different external environment. For different operation modes, the statistical characteristics in mean and covariance of collected dataset are significantly different. Therefore, traditional MSPM methods, which are established based on one operating mode assumption, would not well accommodate multimode process monitoring. To address the multimodality of data

distribution, the multiple modeling strategies were first investigated. Zhao et al[11] studied multiple modeling strategy-based PCA method, which divides multimode training dataset into different groups through clustering technique and then builds individual model for each group separately. Ge et al[12] utilized Bayesian inference to synthesize the monitoring results of all modes to construct a global monitoring index. Recently, Zhao et al[13] proposed a between-mode relative analysis algorithm to investigate the between-mode relationship, where each mode is divided into 3 different systematic subspaces and 1 residual subspace. The mixture modeling strategy is another effective method, which can represent the data sources driven by several operating modes.[14-16] Yu et al[17] proposed a finite Gaussian mixture model approach and constructed probabilistic monitoring statistics based on Bayesian inference. Yu et al[18] also developed Bayesian inference-based Gaussian mixture contribution plot to indentify the faulty variables. Different from multiple modeling or mixture modeling strategies, local modeling approaches are also studied by researchers. Ma et al[19] proposed a novel local neighborhood standardization strategy, while Wang et al[20] provided a weighted k neighborhood standardization method to transform multimode data to be approximately unimodal or Gaussian distribution. Deng et al[21] employed the PCA similarity factor between the current data window and its local neighborhood data window to monitor process status. Several different multimode monitoring approaches are also proposed and obtain satisfactory monitoring performances.[22-26]

However, most aforementioned multimode process monitoring algorithms are based on the assumption that the process variables within each individual operating mode should follow a multivariate Gaussian distribution. In fact, the process variables cannot conform to the strict Gaussian distribution but follow the non-Gaussian distribution or the mixture distribution of Gaussian and non-Gaussian. In other words, the distribution of practical chemical process data is always uncertain and yet difficult to accurately determine. To address this issue, a series of local outlier factor (LOF) statistics-based monitoring methods, which can cope with multimodality and uncertainty of data distribution simultaneously, have been proposed in different cases. Ma et al[27] proposed a NSLOF method for better monitoring performance. Ma et al[28] developed an adaptive LOF method to handle time-varying and multimode characteristics. Song et al[29] combined a LOF-based clustering strategy and LOF-based statistics for multimode process monitoring. Besides, Song et al[30] developed a recursive LOF algorithm for multimode identification to consider both stable and transitional modes. Even though these extensions of LOF-based methods have shown their superiorities, there are still 2 valuable issues deserving further investigation. One problem is the influence of variable scales. If the LOF approach is directly applied to the original data, the neighborhood of a sample would be mainly determined by the variables with large scales, which would mask the information of small-scale variables. The other problem is the information mining of local variable behaviors. For practical industrial processes, a fault would only affect some specific variables. That is to say, the process faults would show locality. Traditionally, the LOF value of a sample is computed with its neighbors in normal training dataset, where the Euclidean distance between data pair reflects the distance of 2 sample points. Therefore, the local variable information may be weakened within conventional LOF computation.

To develop an efficient multimode complex process monitoring method, a novel double-level local information-based LOF (DLI-LOF) approach is proposed in this paper. First, the first-level local statistical information (ie, mean and variance of local data structure) is performed by the local neighborhood standardization (LNS) strategy, which can be regarded as z-score formulation in local subregion. Thus, the multimode data can be transformed into an approximately unimodal distribution with zero mean and 1 variance. Then, the second-level local variable behaviors are considered, which means local variable information within LOF calculation. Different from traditional LOF computation, the Euclidean distance between data pairs is decomposed to express the distance relationship of single variable in the VLOF approach. Besides, a weighting strategy and Bayesian inference are applied to highlight the local variable behaviors. To isolate the fault sources, a contribution plot corresponding to DLI-LOF method is designed to analyze the fault variables. Finally, a numerical simulation and the benchmark Tennessee Eastman (TE) process are applied to evaluate the proposed method.

The remainder of this paper is organized as follows. In section 2, traditional LOF is reviewed briefly and followed by a motivation analysis of the proposed method. Section 3 gives a detailed explanation about double-level local information-based method. Section 4 describes the fault detection and fault isolation procedures. Then, 2 simulation examples are presented to verify the proposed method in section 5. Section 6 gives some discussions about the proposed method. Finally, some conclusions are drawn in section 7.

2 | PRELIMINARIES

This section provides a brief review of LOF computation and then presents the motivation analysis of this work.

2.1 | Local outlier factor

Local outlier factor is a famous unsupervised data mining technique,[31,32] which can locate the anomalous points of the given dataset with free specific data distribution assumption. And a summary of LOF calculation is presented briefly. Given training dataset $\boldsymbol{X} \in \boldsymbol{R}^{n \times m}$ and a test sample $\boldsymbol{x} \in \boldsymbol{R}^{m}$, the LOF computing procedure is listed as follows[27]:

(1) For the sample $\boldsymbol{x}$, find its K nearest neighbors in $\boldsymbol{X}$ by using Euclidean distance, denoting as $N(\boldsymbol{x}) = [\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^K]$, where a prior parameter K represents the size of $N(\boldsymbol{x})$.
(2) For each neighbor sample $\boldsymbol{x}^f (1 \le f \le K)$, determine its K distance $k\_\text{distance}(\boldsymbol{x}^f)$, which represents the Euclidean distance between $\boldsymbol{x}^f$ and its Kth nearest neighbor.
(3) Obtain the reachability distance of the sample $\boldsymbol{x}$, defined as

$$\text{reach\_}d(\mathbf{x}, \mathbf{x}^f) = \max\{k\_\text{distance}(\mathbf{x}^f), d(\mathbf{x}, \mathbf{x}^f)\}, \tag{1}$$

where $d(\boldsymbol{x}, \boldsymbol{x}^f)$ represents the Euclidean distance between $\boldsymbol{x}$ and its fth neighbor.
(4) Compute the local reachability density (lrd) for the sample $\boldsymbol{x}$, which is expressed as

$$\text{lrd}(\mathbf{x}) = \frac{K}{\sum_{f=1}^{K} \text{reach\_}d(\mathbf{x}, \mathbf{x}^f)}. \tag{2}$$

(5) Compute the LOF of sample $\boldsymbol{x}$, formulated by

$$\text{LOF}(\mathbf{x}) = \frac{1}{K} \sum_{f=1}^{K} \frac{\text{lrd}(\mathbf{x}^f)}{\text{lrd}(\mathbf{x})}. \tag{3}$$

From the above procedure of LOF calculation, it can be seem that LOF indicates the degree of how isolated a sample is with respect to its surrounding neighbors. If sample $\boldsymbol{x}$ is not an outlier, the LOF value would be approximately equal to 1. If $\boldsymbol{x}$ is an outlier, the LOF value would be larger than 1, because the lrd of $\boldsymbol{x}$ would be smaller than that of its neighbors. Therefore, LOF can identify whether some sample is an outlier without any data distribution assumption.

2.2 | Motivation analysis

In chemical processes, fault samples would show different characteristics compared with the normal data. Thus, abnormal samples can be regarded as outliers and thereby LOF-based statistic would perform as an efficient monitoring index. Unlike the conventional distance-based monitoring statistic, the density-based LOF statistic is distribution-free and more accurate. However, there are some valuable problems deserving further investigation, with 2 questions of how to better handle multimodality of data distribution and how to mine local variable information.

> *Statement* 1. The diverse scales of variables may degrade monitoring performance of traditional LOF method.

Because LOF is calculated by using a sample's neighbors in normal dataset and these neighbors are usually within one operating mode, it can be regarded as a local modeling strategy for multimode process monitoring. However, within LOF calculation, the Euclidean distance between data pairs is strongly influenced by the scales of variables. Some variables with large scales would show heavy proportion to represent the distance relationship among 2 samples. Thereby, if 1 fault occurs to a variable with small scale, the fault information would be submerged by those large-scale variables. Although some efficient data preprocessing technique like z-score can transform original data into normalized distribution, it would be insufficient to handle multimode data with the situation that variables under different operating modes would also show significantly various scales. Therefore, it is better for LOF monitoring method if the multimode data can be efficiently normalized. Under the guild of this expectation, the LNS strategy[19] is employed in this work. With the novel data reprocessing technique, the statistical characteristics of mean and variance of local data structures can be used to transform multimode data into an approximately unimodal distribution.

Statement 2. The local variable behaviors should be reflected within LOF calculation.

The LOF method employs the Euclidean distance among sample pairs to finally express the anomaly of a testing data. However, it should be pointed out that tradition LOF calculation is not the best choice for process monitoring because some fault only involves specific variables, while tradition LOF puts all variables as a whole and may weaken the influence of local fault variables. Thus, it is necessary to emphasize the effects of different variables. Considering a simple numerical example with variable $x_1 \sim N(2,0.6)$ and variable $x_2 \sim N(2,0.4)$, respectively, 400 samples are generated as normal training dataset and a testing sample $\boldsymbol{x}_{\text{new}}$ is set as fault data, as shown in Figure 1. In this simulation, 20 neighbors are located for a sample and 99% confidence limit is set to determine the normal boundary. As shown in Figure 2, the fault sample $\boldsymbol{x}_{\text{new}}$ is treated as in normal operation under traditional LOF monitoring due to that its LOF value is within the confidence limit. This is because the lrd of sample $\boldsymbol{x}_{\text{new}}$ is quite similar to that of its neighbors. For intuitional explanation, the first neighbor $\boldsymbol{x}_{\text{new}}^{1}$ (marked as rectangle) of $\boldsymbol{x}_{\text{new}}$ and the Kth neighbor $\boldsymbol{x}_{\text{new}}^{1,K}$ (marked as diamond) of $\boldsymbol{x}_{\text{new}}^{1}$ are located in Figure 1, respectively. Clearly, the distance between $\boldsymbol{x}_{\text{new}}$ and $\boldsymbol{x}_{\text{new}}^{1}$ is approximate to that between $\boldsymbol{x}_{\text{new}}^{1}$ and $\boldsymbol{x}_{\text{new}}^{1,K}$. And this situation would be found in most neighbors of sample $\boldsymbol{x}_{\text{new}}$. Therefore, the LOF value of $\boldsymbol{x}_{\text{new}}$ would be approximately equal to those of normal training data.

Remark 1. In this numerical simulation, a 2-dimension process is designed for analysis of statement 2 so that a fault sample is set insignificantly to profit the following explanation. Also, the original data are used

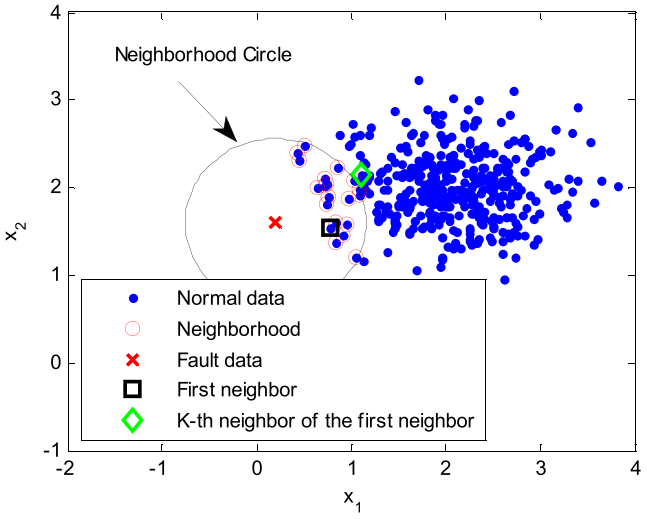

FIGURE 1 Neighborhood distribution of fault data

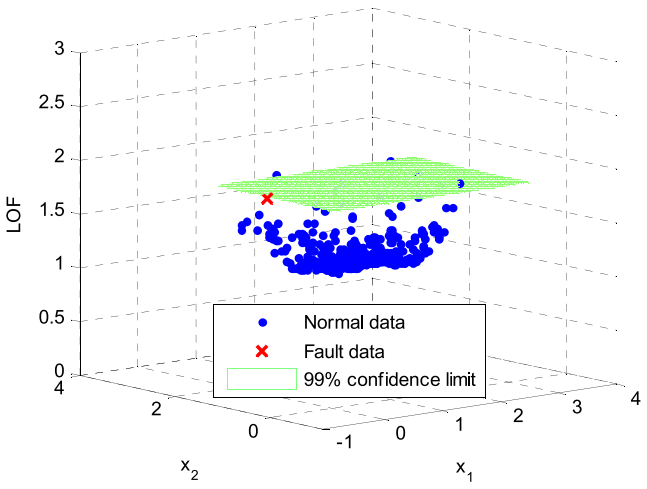

FIGURE 2 Local outlier factor values of fault data and normal data

for LOF computing because the 2 variables are with the similar scales. The following procedure is to illustrate the effects of local variables.

According to Figure 1, we can also find that $\boldsymbol{x}_{\text{new}}$ sample reflects a little fault characteristic with its variable x_1, whereas the variable x_2 shows normal status. Clearly, this information cannot be reflected by traditional LOF method because all variables are treated as a whole to express the distance between data pairs. To investigate the local variable behaviors, another toy illustration is designed as Figure 3, where point A is a testing sample and B is its neighbor in normal condition, and C and D are neighbors of B and C, respectively. The dashed lines denote the normal boundaries of variables x_1 and x_2. Simply, the LOF value of sample A can be calculated as 1 with a condition $d_1 < d_2$, which shows that A is within normal status. However, we can find that variable x_1 of A is abnormal, which cannot be reflected by traditional LOF. If we transform the Euclidean distance between data pairs into a vector form (for example, distance vector between A and B is $d(A,B)_{\text{vec}} = [|A_{x_1} - B_{x_1}|, |A_{x_2} - B_{x_2}|]^{\text{T}}$), then the reachability distance of a single variable can be determined separately, like $\text{reach\_}d(A,B)_{\text{vec}} = [|A_{x_1} - B_{x_1}|, |B_{x_2} - C_{x_2}|]^{\text{T}}$. In this case, the LOF expression of sample A can be finally obtained as $\text{LOF}(A)_{\text{vec}} = \left[\frac{|A_{x_1} - B_{x_1}|}{|C_{x_1} - D_{x_1}|}, \frac{|B_{x_2} - C_{x_2}|}{|B_{x_2} - C_{x_2}|}\right]^{\text{T}}$. Clearly, the first element of $\text{LOF}(A)_{\text{vec}}$ is larger than 1, while the second element is equal to 1, which indicates that variable x_1 of A is in abnormal condition and variable x_2 of A is still in normal state. The results are in accordance with the reality. Under this guide, the LOF values of variable x_1 and variable x_2 for the fault sample in Figure 1 are presented in Figure 4. According to this figure, the LOF value of variable x_1 exceeds its confidence limit while variable x_2 is still indicated in normal condition, which satisfies the reality.

Based on above analysis, information of fault variables would be weakened by those fault-free variables. Thus, local variable behaviors should be considered within LOF calculation. In this work, a novel LOF computation approach, referred to as variable LOF (VLOF), is proposed to highlight the local variable information.

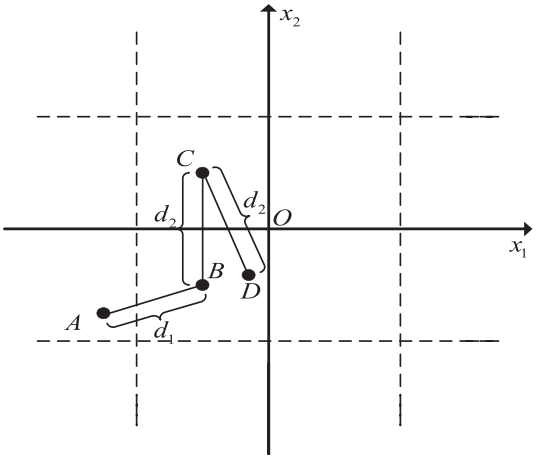

FIGURE 3 Illustration of local variable behaviors

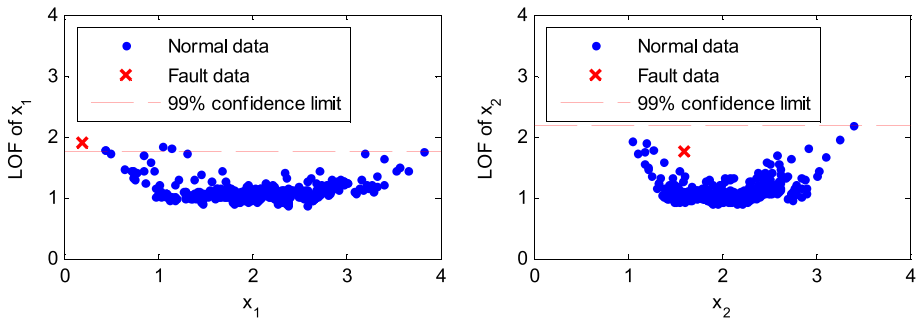

FIGURE 4 Local outlier factor values of variable x_1 and variable x_2

3 | THE PROPOSED DLI-LOF SCHEME

From the problem statements mentioned in section 2.2, a sufficient data preprocessing technique should be employed and the local variable behaviors should be emphasized within LOF computation. To achieve an efficient multimode complex process monitoring method, a novel double-level local information-based LOF method is proposed in this paper. First, the LNS strategy is employed to extract the local statistical information of data structure. Then, a novel LOF calculation approach is proposed to mine local variable behaviors. And a weighting strategy is used to highlight the local variable information, while Bayesian inference is adopted to obtain a global monitoring statistic. In the end of this section, a contribution plot is designed to isolate the fault variables.

3.1 | Local neighborhood standardization

With the LNS strategy, the original multimode data can be transformed into an approximately unimodal distribution. And a brief introduction of LNS is presented as follows.

For any training sample $\boldsymbol{x} \in \boldsymbol{R}^m$, its K nearest neighbors $N(\boldsymbol{x})$ in the training data $\boldsymbol{Y} \in \boldsymbol{R}^{n \times m}$ are expressed by

$$N(\boldsymbol{x}) = \left[\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^K\right], \tag{4}$$

where $\boldsymbol{x}^j$ $(1 \le j \le K)$ is the jth nearest sample of the sample $\boldsymbol{x}$ and is determined by using the Euclidean distance. According to the literature,[19] sample $\boldsymbol{x}$ can be normalized as follows:

$$\overline{\mathbf{x}} = \frac{\mathbf{x} - \text{mean}(N(\mathbf{x}^1))}{\text{std}(N(\mathbf{x}^1))}, \tag{5}$$

where mean($N(\boldsymbol{x}^1)$) and std($N(\boldsymbol{x}^1)$) represent the mean and standard deviation of the K nearest neighbors of the sample $\boldsymbol{x}^1$. According to Equation 5, the LNS strategy can be regarded as z-score formulation in local subregion. Because a sample is normalized by using the mean and variance of its local neighbors, usually, these neighbors are under the same operating mode. This novel data preprocessing method is to use local mean and variance information to approximately represent the global mean and variance of single operating mode. Therefore, the multimodality of data distribution can be efficiently eliminated.

3.2 | Variable local outlier factor method

According to the calculation of LOF, it can be found that traditional LOF expression may weaken the local variable behaviors. This is because the distance measurement between data pairs is to reflect the distance relationship among all measured variables. Hence, the influences of faulty variables are inevitably weakened by nonfaulty variables. Therefore, information of local variables should be considered and emphasized. As mentioned in section 2.2, traditional Euclidean distance can be decomposed to reflect the distance of single variable. In this case, the squared Euclidean distance between sample $\boldsymbol{x}$ and its neighbor $\boldsymbol{x}^f$ can be reformulated as

$$d^2\left(\mathbf{x}, \mathbf{x}^f\right) = \sum_{i=1}^{m} \left[\boldsymbol{\xi}_i^{\mathrm{T}}\left(\mathbf{x} - \mathbf{x}^f\right)\right]^2, \tag{6}$$

where $\boldsymbol{\xi}_i^{\mathrm{T}}$ is the unit vector of the ith variable x_i, expressed as

$$\boldsymbol{\xi}_i = \left[\underbrace{0, \cdots, 0,}_{1 \to i-1} \underbrace{1,}_{i} \underbrace{0, \cdots, 0}_{i+1 \to m}\right]^{\mathrm{T}}, \tag{7}$$

Therefore, the distance expression for variable x_i can be obtained as

$$d\left(x_i, x_i^f\right) = \left|\boldsymbol{\xi}_i^{\mathrm{T}}\left(\mathbf{x} - \mathbf{x}^f\right)\right|. \tag{8}$$

Based on Equation 8, the reachability distance of variable x_i is defined as

$$\text{reach\_}d\left(x_i, x_i^f\right) = \max\left\{k\text{\_distance}\left(x_i^f\right), d\left(x_i, x_i^f\right)\right\}, \tag{9}$$

where k_distance(x_i^f) denotes the distance between sample $\boldsymbol{x}^f$ and its Kth nearest neighbor $\boldsymbol{x}^{f,\,K}$ in variable x_i, formulated as

$$k\text{\_distance}\left(x_i^f\right) = \left|\boldsymbol{\xi}_i^{\mathrm{T}}\left(\mathbf{x}^f - \mathbf{x}^{f,K}\right)\right|. \tag{10}$$

Then, the lrd and LOF of variable x_i can be computed as the following equations:

$$\text{lrd}(x_i) = \frac{K}{\sum_{f=1}^{K} \text{reach\_}d\left(x_i, x_i^f\right)}, \tag{11}$$

$$\text{LOF}(x_i) = \frac{1}{K}\sum_{f=1}^{K}\frac{\text{lrd}\left(x_i^f\right)}{\text{lrd}(x_i)}. \tag{12}$$

Given sufficient training data, the confidence limit of LOF(x_i) can be obtained by using the kernel density estimation method.[33,34] With the typically Gaussian kernel function, the probability density function of statistic y is defined as

$$\widehat{f}(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}h}\exp\left[-\frac{(y-y_i)}{2h^2}\right], \tag{13}$$

where h is the kernel window width. Then, the value $\widetilde{y}$, which is estimated as

$$\int_{-\infty}^{\widetilde{y}}\widehat{f}(y)dy = 1-\alpha, \tag{14}$$

can provide the confidence limit, where α is the significance level.

According to Equation 12, the local variable behavior can be reflected by using the value of LOF(x_i), and thereby, the effect of information smearing would not exist in the proposed LOF calculation approach, which is named as VLOF method.

3.3 | Weighting strategy and Bayesian inference

To emphasize the information of fault variables, an online weighting strategy is designed to distinguish the roles of different variables. For LOF(x_i), its boundary to judge whether it should be weighted can be determined as

$$\text{LOF}(x_i)^{\max} = \frac{\text{LOF}(x_i)^{\max 1} + \text{LOF}(x_i)^{\max 2}}{2}(1 \le i \le m), \tag{15}$$

where LOF$(x_i)^{\max 1}$ and LOF$(x_i)^{\max 2}$ are the first 2 maximum values of LOF(x_i) in the normal training dataset. For a test data $\boldsymbol{x}(h)$ at the hth sample instant, if 1 variable's LOF value exceeds the boundary LOF$(x_i)^{\max}$, it can be regarded as the abnormal variable and should be given a large weight. Otherwise, the weight should be relatively small. For the LOF value of variable x_i, a real-time weighting coefficient $w_i(h)$ can be designed as

$$w_i(h) = \begin{cases} \left(\dfrac{\text{LOF}(x_i,h)}{\text{LOF}(x_i)^{\max}}\right)^2 & \text{if } \text{LOF}(x_i,h) > \text{LOF}(x_i)^{\max} \\ & \&\text{LOF}(x_i,h-1) > \text{LOF}(x_i)^{\max} \\ 1 & \text{otherwise} \end{cases}, \tag{16}$$

Bayesian inference[12,35] is then adopted to integrate information from each variable. Firstly, statistic LOF(x_i, h) of variable x_i is transformed into probability forms, defined as

$$P(x_i, h|N) = \exp\left(-\frac{w_i(h)\mathrm{LOF}(x_i, h)}{\mathrm{LOF}(x_i)_{\mathrm{lim}}}\right), \tag{17}$$

$$P(x_i, h|F) = \exp\left(-\frac{\mathrm{LOF}(x_i)_{\mathrm{lim}}}{w_i(h)\mathrm{LOF}(x_i, h)}\right), \tag{18}$$

where N denotes the normal condition; while F identifies the faulty condition, $\mathrm{LOF}(x_i)_{\mathrm{lim}}$ is the confidence limit corresponding to LOF(x_i). Then, the fault occurrence probability of variable x_i is calculated as

$$P(F|x_i, h) = \frac{P(x_i, h|F)P(F)}{P(x_i, h|N)P(N) + P(x_i, h|F)P(F)}, \tag{19}$$

where $P(F)$ is the prior fault probability set as the significance level α, while $P(N)$ represents the prior normal probability set as $1 - \alpha$. The final fusion statistic index can be determined as

$$\mathrm{BIC}_{\mathrm{VLOF}}(h) = \sum_{i=1}^{m}\left\{P(F|x_i, h)\frac{P(x_i, h|F)}{\sum_{l=1}^{m} P(x_l, h|F)}\right\}. \tag{20}$$

Remark 2. The computation efficiency of the proposed method mainly depends on 2 kinds of neighbors searching. The first one is to normalize new sample, and the second one is to calculate LOF value. And the Bayesian inference used in DLI-LOF method just contains several mathematical steps, which spends little computation time. Therefore, the neighbor-searching contributes most to the computational complexity. However, in online monitoring stage, we only need to find the first neighbor of each new sample, and all the neighbors of each sample in training dataset have been located in historical modeling stage. Thus, the online computation efficiency is acceptable for process monitoring.

Remark 3. It should be noted that the weighting strategy is only adopted in the online monitoring stage. To reduce the risk of high false alarm rate (FAR), we design a tolerant weighting strategy. In offline modeling stage, we determine the normal boundary of each LOF(x_i) by investigating the mean value of its 2 largest elements, as shown in Equation 15. Thus, the normal boundary can cover most normal status. Besides, the weighting strategy is applied to online LOF(x_i) if and only if 2 successive samples show abnormal conditions, as suggested in Equation 16. Therefore, the key fault variable information can be emphasized, while the noise influence can be reduced. Based on above analysis, the confidence level α is still reasonable and acceptable for online monitoring.

3.4 | Contribution plot of the proposed method

After a fault is alarmed, fault diagnosis technique then plays a necessary role to determine the responsible root causes. Contribution plot method has been successfully applied to identify fault variables in the MSPM circle.[36] As the LOF value of single variable is available in the proposed VLOF method, the identification of fault variables can be turned into the investigation of how significantly the statistic LOF(x_i) is reflected in the monitoring statistic $\mathrm{BIC}_{\mathrm{VLOF}}$. Therefore, the contribution rate of variable x_i is defined as

$$\mathrm{Cont}(x_i, h) = \frac{P(F|x_i, h)P(x_i, h)}{\mathrm{BIC}_{\mathrm{VLOF}}(h)}, \tag{21}$$

$$P(x_i, h) = \frac{P(x_i, h|F)}{\sum_{l=1}^{m} P(x_l, h|F)}. \tag{22}$$

In practice, some variables with greater contribution rates than the others are determined as the most likely root causes. In the current study, the average contribution rate of 10 consecutive samples is regarded as the final diagnosis result.

4 | THE PROCESS MONITORING PROCEDURE

This section presents the process monitoring framework based on the proposed modified DLI-LOF method, and the flowchart is shown in Figure 5. The complete procedure is summarized as follows, including offline modeling stage and online detection stage.

Offline modeling stage:

(1) Collect normal training dataset $\boldsymbol{X}$ and scale each variable by using LNS strategy.
(2) Determine the neighbors for each sample $\boldsymbol{x}$ in training dataset $\boldsymbol{X}$ by traditional Euclidean distance.
(3) Use Equation 8 to express distance relationship of variable x_i between data pairs.
(4) Compute LOF values of variable x_i according to Equations 9 to 12.
(5) Determine the confidence limit for LOF values of each variable by using kernel density estimation technique and calculate the normal boundary for LOF(x_i).

Online monitoring stage:

(1) Collect a new data vector $\boldsymbol{x}_{\text{new}}$ and standardize it by LNS strategy.
(2) Locate its neighbors in training dataset $\boldsymbol{X}$.
(3) Calculate new LOF value for each variable by using Equations 8 to 12 and determine the weighting values by using Equation 16.
(4) Compute new BIC_{VLOF} statistic by using Equations 17 to 20.
(5) If the value of BIC_{VLOF} is not beyond the confidence limit α, the system is in normal operation. Otherwise, the monitored sample is in abnormal condition, then identify the corresponding fault variables by using contribution plot.

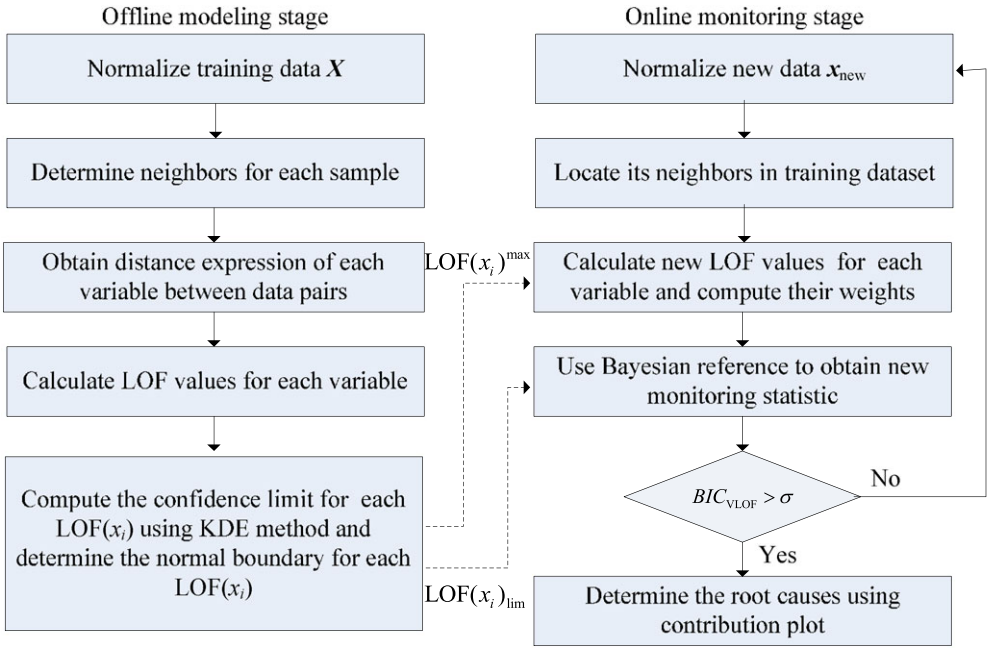

FIGURE 5 Schematic diagram of double-level local information-based local outlier factor method

5 | CASE STUDY

In this section, the proposed DLI-LOF method as well as the original LOF method is applied to monitor 2 multimode complex processes, including a numerical simulation and the benchmark TE process. Additionally, comparisons with a recently proposed method, called neighborhood standardized LOF (NSLOF),[27] are considered to investigate the practicability and the superiority of the DLI-LOF method.

5.1 | Numerical example

In this section, the performance of DLI-LOF method is demonstrated through a simple numerical example suggested by Ma et al.[27] This example consists of 5 measured variables which are generated through the mathematical model as

$$\begin{aligned} x_1 &= 0.5768s_1 + 0.3766s_2 + e_1 \\ x_2 &= 0.7382s_1^2 + 0.0566s_2 + e_2 \\ x_3 &= 0.8291s_1 + 0.4009s_2^2 + e_3 \\ x_4 &= 0.6519s_1s_2 + 0.2070s_2 + e_4 \\ x_5 &= 0.3972s_1 + 0.8045s_2 + e_5 \end{aligned} \tag{23}$$

where $e_i(1 \le i \le 5)$ is zero-mean Gaussian noise with a standard deviation of 0.01 and s_1 and $s_2$are 2 sources. In this simulation, 2 different operating modes are designed according to the following formulations

$$\begin{aligned} &\text{mode l:} \quad s_1 \quad U(-10,-7) \\ &\qquad\qquad\ \ s_2 \quad N(-15,1) \\ &\text{mode 2:} \quad s_1 \quad U(2,5) \\ &\qquad\qquad\ \ s_2 \quad N(7,1) \end{aligned} \tag{24}$$

Based on Equation 23, a total of 400 samples under normal condition are generated as training dataset which contain 200 samples from each mode. For testing, 2 fault cases are designed as follows, and each fault dataset includes 400 samples.

- Case 1: The system is initially operated under mode 1, and then a step bias of 1.4 is added to x_5 from the 201st samples.
- Case 2: The system is initially operated under mode 2, and then a ramp change of $0.01(i - 201)$ is imposed to x_1 from the 201st samples.

Traditional LOF, NSLOF, and the proposed DLI-LOF method are applied to the numerical simulation. To reduce the influence of the variable scales, z-score technique is applied in LOF method suggested by Ma et al.[27] In LNS strategy, 50 neighbors are determined for data reprocessing according to the literature.[23] And the LNS result is shown in Figure 6, from which the multimodality can be efficiently eliminated. For fair comparisons, the number of neighbors is set as 30

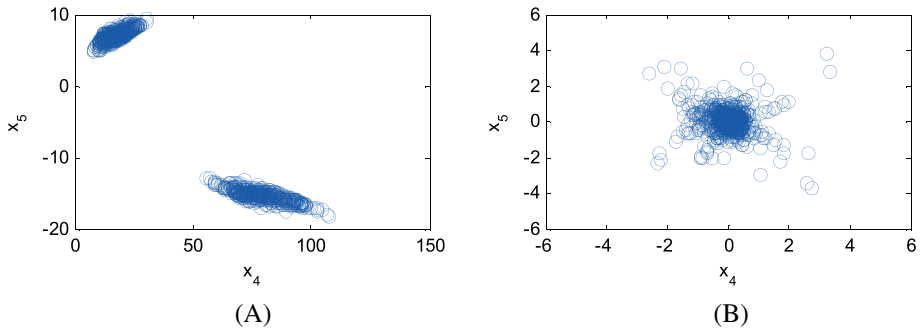

FIGURE 6 Scatter plots of the numerical example: A, original data and B, scaled by local neighborhood standardization

in the LOF calculating procedure according to the literature.[27] To determine the control limits, a 99% confidence level (ie, significance level $\alpha = 0.01$) is adopted in each monitoring approach. To evaluate the monitoring performance of different methods, 3 indices of FAR, fault detection time (FDT), and fault detection rate (FDR) are adopted in this work. FAR is the percentage of normal samples identified as fault samples over all the normal samples, FDT is the alarming point of successful fault detection, and FDR is the percentage of detected fault samples over all the fault samples.

First, case 1 with a step fault is applied for method comparison. The monitoring charts of LOF, NSLOF, and DLI-LOF methods are provided in Figure 7. Clearly, traditional LOF method can hardly detect the fault because most monitoring statistics are under the confidence limit, as shown in Figure 7A. By contrast, the NSLOF method achieves better monitoring performance, and it can give a fault signal at the 211th sample with 69% FDR. However, as shown in Figure 7B, monitoring statistics fluctuate around the confidence limit, which would doubt the operators to make a clear judgment. According to Figure 7C, the proposed method outperforms both LOF and NSLOF methods. It can alarm this fault at the 201st sample immediately with a high 98% FDR. To further determine the fault root, contribution plot at the 201st to 210th samples is presented in Figure 7D, where variable x_5 shows the largest contribution rate. And this fault diagnosis result is corresponding to the reality.

The monitoring results of 3 methods under case 2 are presented in Figure 8. As for traditional LOF method, it can detect this fault at the 354th point but with only 29% FDR, which can still hardly satisfy industrial demand. According to Figure 8B, the NSLOF method outperforms LOF method and it can alarm the fault at the 279th sample with a 62% FDR. This is because the local structure information is considered to determine the standardized Euclidean distance in NSLOF method, which the effects of variable scales can be efficiently erased. By contrast, the proposed method provides better monitoring performance, and it can give an alarming signal at the 232nd point and with a higher 89.5% FDR. The FDT and the FDR are significantly reduced compared with NSLOF method. This is because the local variable information is further considered to determine more precise process status. And according to the contribution plot at the 232nd to 241st samples in Figure 8D, variable x_1 is identified as the corresponding fault cause, which is in accordance with the reality.

The detailed monitoring results are provided in Table 1, from which the proposed DLI-LOF method achieves the best monitoring performance. Also, the average FARs of LOF, NSLOF, and DLI-LOF methods are calculated as 1.25%, 0.75%, and 1.25%, respectively, which are acceptable for the given 99% confidence limit. To sum up, the numerical simulation results validate the superiority of the proposed method.

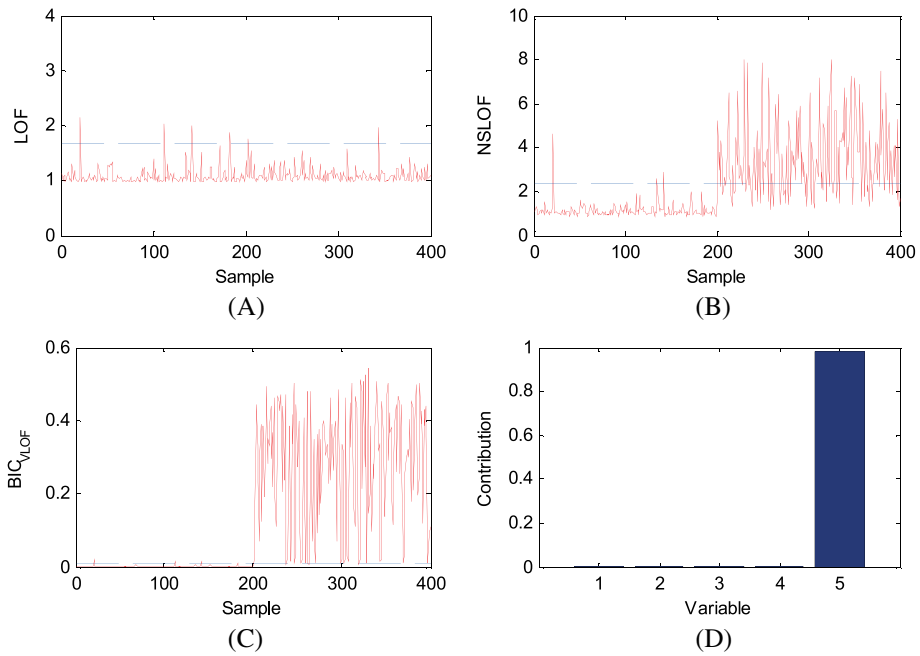

FIGURE 7 Monitoring results for case 1: A, local outlier factor (LOF), B, neighborhood standardized LOF, C, double-level local information-based LOF (DLI-LOF), and D, contribution plot of DLI-LOF

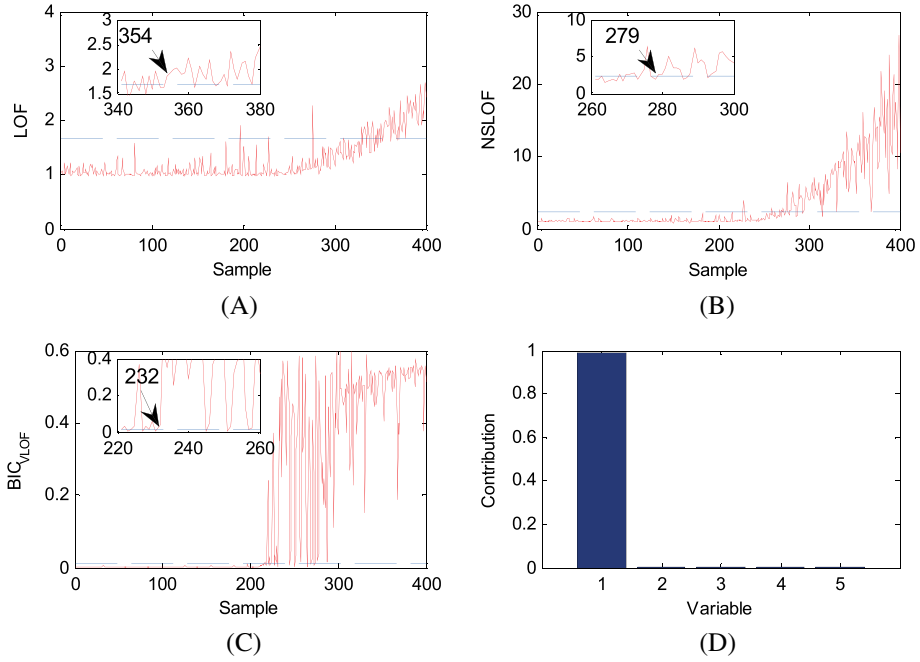

FIGURE 8 Monitoring results for case 2: A, local outlier factor (LOF), B, neighborhood standardized LOF, C, double-level local information-based LOF (DLI-LOF), and D, contribution plot of DLI-LOF

TABLE 1 Fault detection times and fault detection rates of the 3 methods for numerical example

	Fault Detection Times			Fault Detection Rates (%)		
Fault	**LOF**	**NSLOF**	**DLI-LOF**	**LOF**	**NSLOF**	**DLI-LOF**
Case 1	Failed	211	201	1	69	98
Case 2	354	279	232	29	62	89.5
Average				15	65.5	93.75

LOF, local outlier factor; NSLOF, neighborhood standardized LOF; DLI-LOF, double-level local information-based LOF.

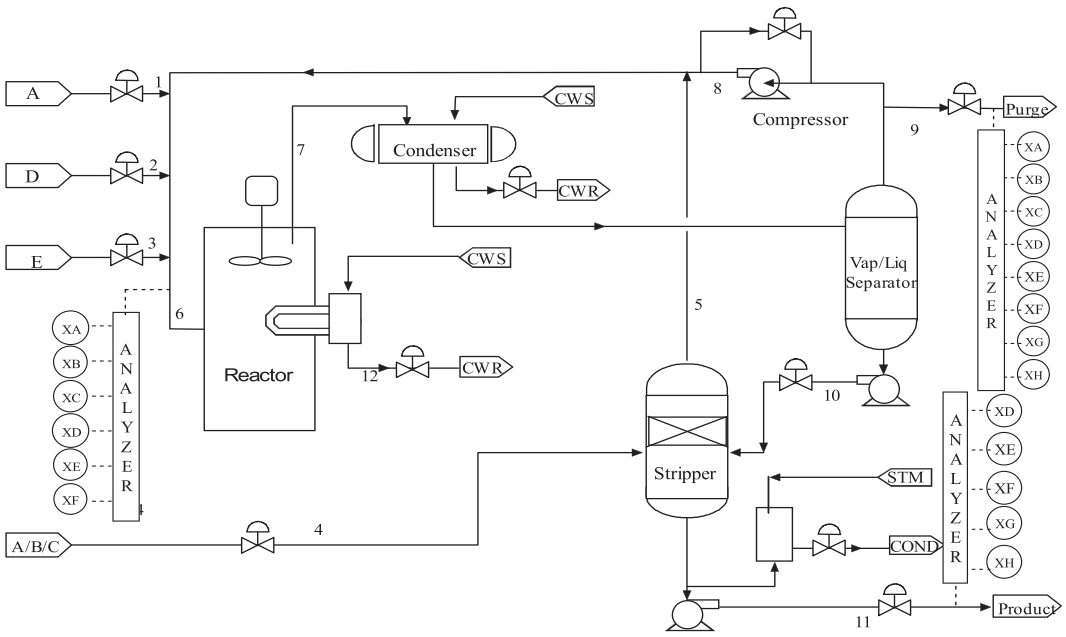

FIGURE 9 Schematic of the Tennessee Eastman process

5.2 | TE process

The TE process is a real chemical process, and its simulator is proposed by Downs and Vogel.[37] And the detailed Matlab Simulink modules are available in http://depts.washington.edu/control/LARRY/TE/download.html. In the present studies, this process has been a benchmark case widely used to evaluate different monitoring strategies.[38-42] The TE process is composed of 5 major operation units, including the product condenser, the reactor, the compressor, the separator, and the stripper, as shown in Figure 9. According to the G/H mass ratios, 6 operating modes can be generated. In this paper, modes 1 and 3 are used to obtain multimode dataset with 500 normal data from each mode. The monitored variables contain 22 continuous process variables and 9 manipulated variables, as listed in Table 2. A total of 20 programmed faults, as shown in Table 3, are considered for monitoring evaluating. Each fault dataset consists of 1000 samples, where a fault is introduced at the 201st sample.

TABLE 2 Process monitoring variables in Tennessee Eastman process

Variable No	Process Measurements
1	A feed (stream 1)
2	D feed (stream 2)
3	E feed (stream 3)
4	Total feed
5	Recycle flow (stream 8)
6	Reactor feed rate (stream 6)
7	Reactor pressure
8	Reactor level
9	Reactor temperature
10	Purge rate (stream 9)
11	Product separator temperature
12	Product separator level
13	Product separator pressure
14	Product separator underflow (stream 10)
15	Stripper level
16	Stripper pressure
17	Stripper underflow
18	Stripper temperature
19	Stripper stream flow
20	Compressor work
21	Reactor cooling water outlet temperature
22	Separator cooling water outlet temperature
23	D feed flow valve (stream 2)
24	E feed flow valve (stream 3)
25	A feed flow valve (stream 1)
26	Total feed flow valve (stream 4)
27	Purge valve (stream 9)
28	Separator pot liquid flow valve (stream 10)
29	Stripper liquid product flow valve (stream 11)
30	Reactor cooling water flow
31	Condenser cooling water flow

TABLE 3 Process faults for Tennessee Eastman process

Fault No	Disturbance State	Type
1	A/C feed ratio, B composition constant (stream 4)	Step
2	B composition, A/C ratio constant (stream 4)	Step
3	D feed temperature (stream 2)	Step
4	Reactor cooling water inlet temperature	Step
5	Condenser cooling water inlet temperature	Step
6	A feed loss (stream 1)	Step
7	C header pressure loss-reduced availability (stream 4)	Step
8	A, B, C feed composition (stream 4)	Random variation
9	D feed temperature (stream 2)	Random variation
10	C feed temperature (stream 4)	Random variation
11	Reactor cooling water inlet temperature	Random variation
12	Condenser cooling water inlet temperature	Random variation
13	Reaction kinetics	Slow drift
14	Reactor cooling water valve	Sticking
15	Condenser cooling water valve	Sticking
16-20	Unknown	Unknown

For the TE process, the LOF, NSLOF, and the DLI-LOF methods are first applied for performance comparisons. And the parameters for LNS and LOF are the same with the numerical example. Fault 18 under mode 1 is illustrated, and the monitoring charts of 3 methods are shown in Figure 10A to C. According to Figure 10A, original LOF method cannot successfully detect this fault, and only some discrete samples give alarming signal yet with only 6.5% FDR. In contrast to LOF, the NSLOF method obtains better monitoring performance and it can detect this fault at the 288th point with 69.88% FDR, as shown in Figure 10B. Instead of using traditional Euclidean distance within original LOF approach, the NSLOF method employs standardized Euclidean distance to deal with the effects of different variable scales. By contrast, the DLI-LOF method adopts LNS strategy to erase the influence of various variable scales in multimode

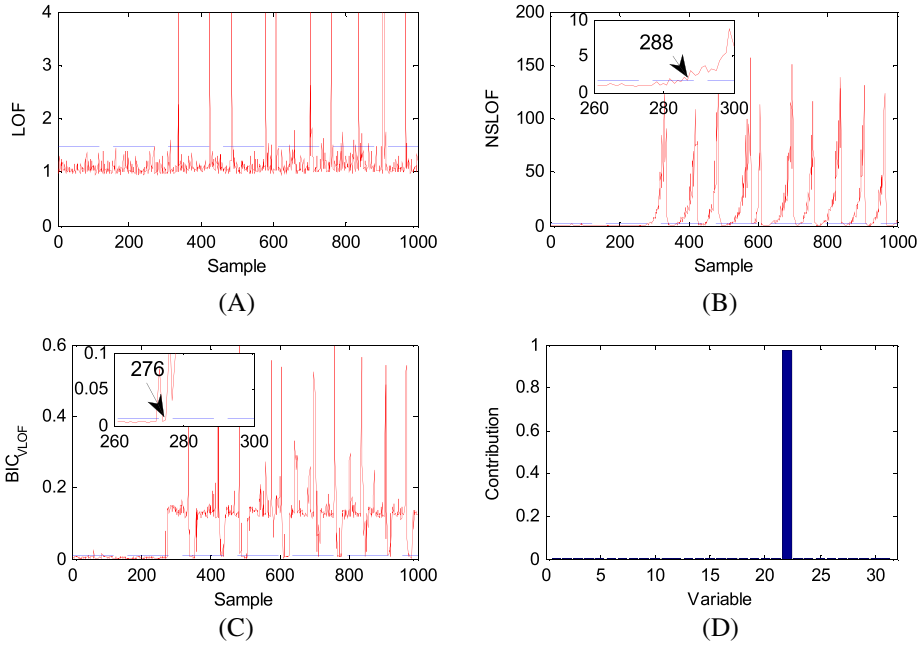

FIGURE 10 Monitoring results for fault 18 under mode 1: A, local outlier factor (LOF), B, neighborhood standardized LOF, C, double-level local information-based LOF (DLI-LOF), and D, contribution plot of DLI-LOF

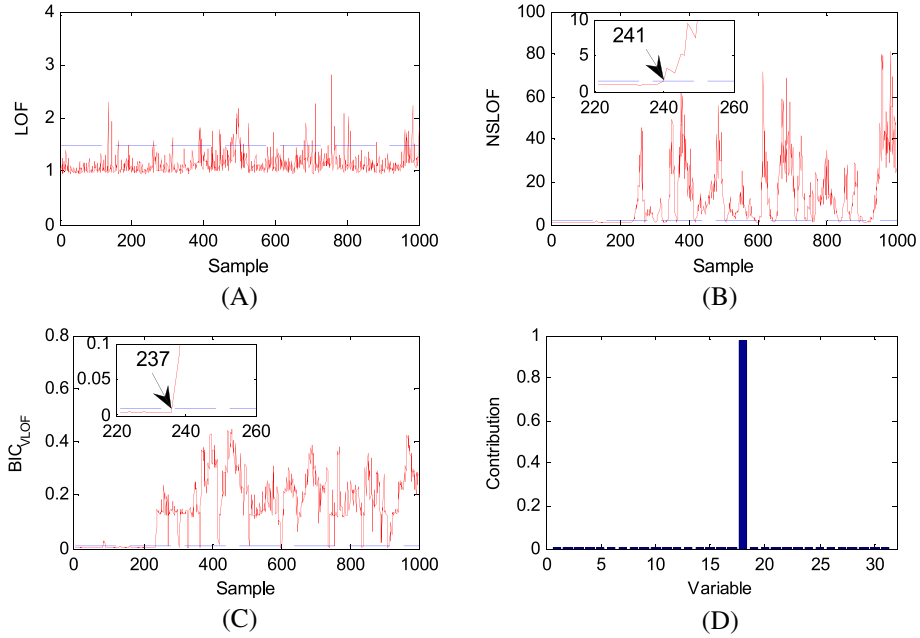

FIGURE 11 Monitoring results for fault 10 under mode 3: A, local outlier factor (LOF), B, neighborhood standardized LOF, C, double-level local information-based LOF (DLI-LOF), and D, contribution plot of DLI-LOF

TABLE 4 Fault detection times and fault detection rates for mode 1 of Tennessee Eastman process

	Fault Detection Time					Fault Detection Rate (%)				
Fault	PCA (T^2)	PCA (SPE)	LOF	NSLOF	DLI-LOF	PCA (T^2)	PCA (SPE)	LOF	NSLOF	DLI-LOF
1	203	205	202	202	202	31.38	99.5	99.88	99.88	99.88
2	210	228	207	207	207	31.25	97.25	99.25	99.25	99.25
3	Failed	Failed	Failed	Failed	350	0	0.88	1.25	2.13	7.88
4	Failed	201	201	201	201	0	100	100	100	100
5	Failed	Failed	Failed	Failed	765	0	0.88	1.13	1.38	6.13
6	202	201	201	201	201	95.1	100	100	100	100
7	201	201	201	201	201	2.75	78.63	100	100	100
8	215	214	214	212	213	76	98.25	98.38	98.63	98.5
9	Failed	Failed	Failed	Failed	367	0	1.12	1.63	4.63	13.63
10	Failed	Failed	Failed	242	239	0	4	4.63	83.13	91.5
11	764	204	204	204	204	12.5	88.75	94.75	98.75	99.0
12	Failed	Failed	372	235	230	2.5	6.38	18.38	45.0	54.75
13	224	222	221	222	221	78.88	97.38	97.5	97.5	97.5
14	Failed	217	205	205	205	0	78.12	90.13	99.5	99.5
15	Failed	Failed	Failed	Failed	764	0	0.88	1.13	2.75	5.88
16	Failed	Failed	Failed	Failed	766	0	0.88	1.13	1.5	4.75
17	739	235	234	232	232	7	77.38	86.75	96.13	96.13
18	Failed	Failed	Failed	288	276	1.5	4.88	6.5	69.88	81.13
19	Failed	Failed	760	205	205	2.75	6.5	17.0	98.88	99.25
20	253	257	252	253	250	80.25	70.88	93.75	93.5	93.88
Average						21.09	50.62	55.66	69.62	72.43

process. Further, local variable behaviors are investigated to mine more process information. Therefore, the DLI-LOF method achieves the best monitoring performance and it can detect the fault at the 276th sample with a higher 81.13% FDR. To determine the fault cause, contribution plot at the 276th to 285th samples of DLI-LOF method is provided in Figure 10D, from which the variable x_{22} (separator cooling water outlet temperature) is the most likely corresponding fault variable. According to the monitoring results for this fault, the superiority and feasibility of the proposed method can be demonstrated.

Fault 10 under mode 3 is also utilized to demonstrate, and the monitoring results are provided in Figure 11. First, traditional LOF method is applied to monitor this fault and the monitoring chart is presented in Figure 11A. Clearly, LOF method fails to detect the fault with only 6.5% FDR. According to Figure 11B, the NSLOF method achieves significantly better monitoring performance and it can detect this fault at the 241st point with 86.25%. By contrast, the proposed method can further improve the monitoring performance. It gives an alarming signal at the 237th sample as shown in Figure 11C, and the FDR is raised up to 92.25%. To isolate the fault variable(s), contribution plot at the 237th to 246th samples is provided in Figure 11D, where variable x_{18} (stripper temperature) contributes most and is identified as the most corresponding fault variable. From the monitoring results under this fault, the advantages of the proposed method can be validated again.

For all 20 programmed faulty datasets, the FDTs and FDRs of traditional LOF, NSLOF, and DLI-LOF methods are listed in Tables 4 and 5. To validate the superiority of the LOF monitoring statistics, PCA method is also applied for TE process. According to the 2 tables, LOF outperforms PCA in most fault cases. This result points out that density-based monitoring statistic is more reasonable than PCA's distance-based monitoring statistics for complex process. However, traditional LOF method cannot achieve satisfactory monitoring performance, especially for faults 10, 18, and 19 in mode 1 and faults 2 and 10 in mode 3. This means LOF method should not be directly applied for multimode process. By contrast, the NSLOF method can give significantly better monitoring results for these faults by considering the effects of variable scales. This result points to the fact that it is necessary to reduce the effect of diverse variable scales.[27] In

TABLE 5 Fault detection times and Fault detection rates for mode 3 of Tennessee Eastman process

	Fault Detection Time					Fault Detection Rate (%)				
Fault	PCA (T^2)	PCA (SPE)	LOF	NSLOF	DLI-LOF	PCA (T^2)	PCA (SPE)	LOF	NSLOF	DLI-LOF
1	330	215	215	202	202	29.75	98.63	98.38	99.88	99.75
2	455	263	257	216	213	5.88	30	49.5	97	98
3	Failed	Failed	Failed	Failed	452	2.5	2.5	1.38	2.38	11.88
4	Failed	201	201	201	201	2.13	100	100	100	100
5	204	201	201	201	201	79.75	96.75	94.13	100	100
6	202	201	201	201	201	96.77	100	100	100	100
7	201	201	201	201	201	4.5	39.88	100	100	100
8	214	216	215	213	213	82.63	98	98.13	98.5	98.5
9	851	868	852	471	357	7.5	8	7.25	11.38	25.88
10	Failed	492	481	241	237	5.5	9.63	6.5	86.25	92.25
11	568	204	204	204	204	10.13	86	89.5	97.88	98.63
12	219	218	219	208	206	88.5	96.13	96.88	99	99.25
13	253	289	248	235	235	50.88	84.88	85.5	96	95.88
14	Failed	203	202	202	202	2.25	75.25	80.63	99.88	99.88
15	Failed	Failed	Failed	Failed	866	2.25	2	1.63	1.13	2.13
16	Failed	Failed	Failed	Failed	865	2.38	2.13	1.75	1.13	2.5
17	Failed	235	234	232	232	11.25	71.5	74.63	95.75	96.13
18	310	282	282	282	284	66.25	86.75	87.38	90	89.88
19	214	218	215	208	206	46.13	67.63	73	99.25	99.38
20	257	251	251	255	254	66.13	77.25	74.0	82.38	86.88
Average						33.15	61.64	66.01	77.89	79.84

contrast to NSLOF method, the proposed DLI-LOF method further shows improved monitoring performance for almost all faulty cases. This is because the local variable information is further considered to pursue more detailed process information. Due to the utilization of double-level local information, the proposed method obtains the best monitoring performance for the TE process. Also, average FARs of PCA, LOF, NSLOF, and DLI-LOF methods for TE process are tabulated in Table 6, which are acceptable in engineering.

To intuitively demonstrate the significance of the proposed DLI-LOF approach, original LOF method with incorporation of LNS strategy, referred to as LNS-LOF method, is also applied to TE process for comprehensive comparison. Clearly, the LNS-LOF method only employs the first-level local structure statistical information and does not use the second-level local variable information. The fault detection rates of LNS-LOF and DLI-LOF methods are listed in Table 7. As can be seen from the 2 tables, the sensitivity to faults of DLI-LOF method is prompted due to the investigation of local variable behaviors, especially for faults 10 and 18 in mode 1 and faults 9 and 10 in mode 3. This is because, in traditional LOF calculating procedure, fault characteristics of some fault variables would be weakened by fault-free variables. Therefore, the second-level local information mining is applied to reflect yet emphasize the local variable information within the proposed DLI-LOF method.

TABLE 6 Average False alarm rates for mode 1 and mode 3 of Tennessee Eastman process

False Alarm Rate (%)				
PCA (T^2)	PCA (SPE)	LOF	NSLOF	DLI-LOF
1	1.25	1.51	1.28	1.28

TABLE 7 Fault detection rates (%) of local neighborhood standardization-based LOF and double-level local information-based LOF methods for Tennessee Eastman process

	Mode 1		**Mode 3**	
Fault	**LNS-LOF**	**DLI-LOF**	**LNS-LOF**	**DLI-LOF**
1	99.88	99.88	99.88	99.75
2	99.25	99.25	97.0	98
3	4.75	7.88	1.75	11.88
4	100	100	100	100
5	3.13	6.13	100	100
6	100	100	100	100
7	100	100	100	100
8	98.63	98.5	98.5	98.5
9	8.13	13.63	9.63	25.88
10	87	91.5	86.25	92.25
11	99	99.0	98.13	98.63
12	50.38	54.75	99	99.25
13	97.5	97.5	95.38	95.88
14	99.5	99.5	99.88	99.88
15	4	5.88	0.88	2.13
16	2.75	4.75	1.25	2.5
17	96.13	96.13	96	96.13
18	74.75	81.13	89.88	89.88
19	99	99.25	99.5	99.38
20	93.63	93.88	84.88	86.88
Average	70.87	72.43	77.89	79.84

6 | DISCUSSION

In this section, the selection of neighbors in LNS strategy and LOF computation are first discussed. Then, some future studies are presented.

The parameters that largely affect the performance of the developed method are 2 K values about neighbor searching. The first one is in LNS strategy, and the second one is in LOF computation. In LNS strategy, the parameter K should be large enough to estimate the local mean and variance. If K is small, the noises would strongly impact the monitoring result and lead to high FAR. In practice, K can be determined in a wide range unless K is larger than the size of one single mode dataset. In LOF calculation, the parameter K should also be large enough to cover the local information. If K is small, the statistical fluctuations of different neighbors would significantly affect the LOF value and the monitoring result may be inaccurate because of insufficient neighbors. This means high FAR and low fault detection rate. Also, a large K would lead to the increasingly computational loads. Therefore, the suggested values of the 2 parameters are listed in Table 8. Furthermore, we illustrate the influence of different K for TE process. First, we set the LNS's parameter K as 50 to test the monitoring results with different LOF's K and then set the LOF's parameter K as 30 to test the monitoring performances with different LNS's K. According to Figure 12, the average monitoring results would be insensitive to parameter changes if K is large enough, which demonstrate the above analysis.

The proposed DLI-LOF method can efficiently monitor multimode process with complex data distribution and highlight fault information hidden in local variables. However, there are 2 problems deserving further investigation. The first one is about incipient faults. This kind of fault would involve few changes on process variables, which leads to inconspicuous fault information. Therefore, the proposed DLI-LOF method, which is performed to describe the original process variables, can hardly detect these incipient faults. The second one is about transition process monitoring. In current work, the multimode training data are assumed to be collected from stable operating modes,

TABLE 8 The parameters K in double-level local information-based local outlier factor method

Parameter K	Suggested Values
K for LNS	$30 \leq K \leq 100$
K for LOF	$10 \leq K \leq 50$

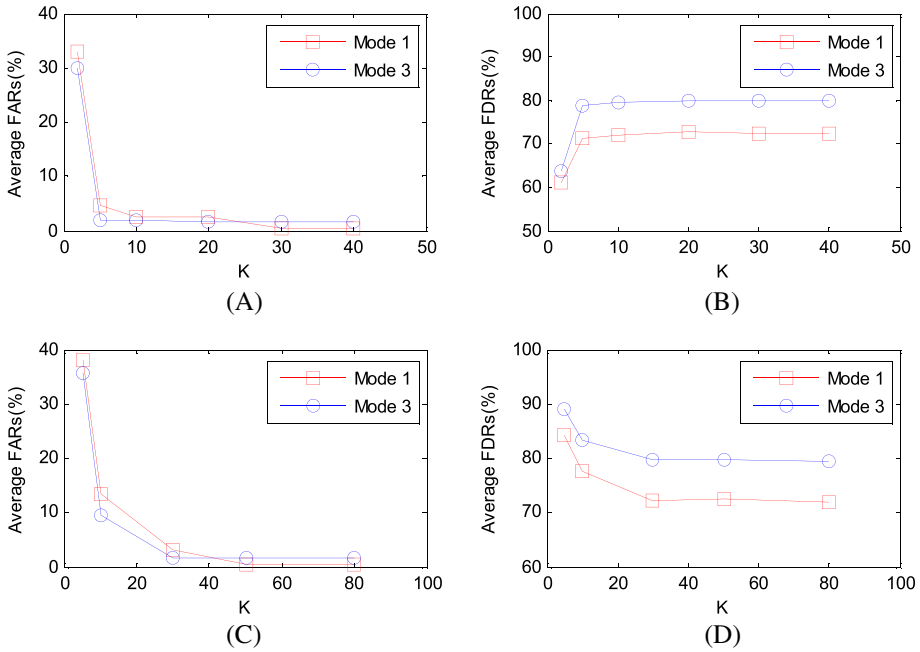

FIGURE 12 Average monitoring results in Tennessee Eastman process: A, false alarm rates (FARs) with different K in local outlier factor (LOF), B, fault detection rates (FDRs) with different K in LOF, C, FARs with different K in local neighborhood standardization (LNS), and D, FDRs with different K in LNS

ignoring the transitional samples between different modes. However, it is still important to ensure the safety of transition phase in real industrial processes. Based on the 2 issues, some related studies would be developed in our future work.

7 | CONCLUSION

In this paper, a novel DLI-LOF method is proposed for multimode complex process monitoring. To sufficiently erase the effects of different variable scales in multimode dataset, the local statistical information of data structure is first used to standardize each monitored sample. Then, the local variable information is further investigated to mine more meaningful process information. To isolate fault variables, the corresponding contribution plot is designed for the proposed DLI-LOF approach. Simulation results on a numerical example and the well-known TE process have evidently validated the effectiveness and superiority of the DLI-LOF method.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (61403418, and 21606256), the Natural Science Foundation of Shandong Province, China (ZR2014FL016, ZR2016FQ21, and ZR2016BQ14), the Fundamental Research Funds for the Central Universities (17CX02054), the Postgraduate Innovation Project of China University of Petroleum (YCX2017058), the Shandong Provincial Key Programme of Research & Development (2018GGX101025).

ORCID

Lei Wang http://orcid.org/0000-0001-7011-0633

REFERENCES

1. Ge Z, Song Z, Gao F. Review of recent research on data-based process monitoring. *Ind Eng Chem Res*. 2013;52(10):3543-3562.
2. Ge Z, Song Z, Ding S, et al. Data mining and analytics in the process industry: the role of machine learning. *IEEE Access*. 2017;5:20590-20616.
3. Liu Y, Pan Y, Sun Z, Huang D. Statistical monitoring of wastewater treatment plants using variational Bayesian PCA. *Ind Eng Chem Res*. 2014;53(8):3272-3282.
4. Liu Y, Pan Y, Wang Q, Huang D. Statistical process monitoring with integration of data projection and one-class classification. *Chemom Intel Lab Syst*. 2015;149:1-11.
5. Deng X, Tian X, Chen S, Harris CJ. Fault discriminant enhanced kernel principal component analysis incorporating prior fault information for monitoring nonlinear processes. *Chemom Intel Lab Syst*. 2017;162:21-34.
6. Ge Z. Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemom Intel Lab Syst*. 2017;171:16-25.
7. Kano M, Hasebe S, Hashimoto I, Ohno H. A new multivariate statistical process monitoring method using principal component analysis. *Comput Chem Eng*. 2001;25(7–8):1103-1113.
8. Dong J, Zhang K, Huang Y, Li G, Peng K. Adaptive total PLS based quality-relevant process monitoring with application to the Tennessee Eastman process. *Neurocomputing*. 2015;154:77-85.
9. Cai L, Tian X, Chen S. A process monitoring method based on noisy independent component analysis. *Neurocomputing*. 2014;127(1):231-246.
10. Adil M, Abid M, Khan AQ, Mustafa G, Ahmed N. Exponential discriminant analysis for fault diagnosis. *Neurocomputing*. 2016;171:1344-1353.
11. 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(22):7025-7035.
12. Ge Z, Song Z. Multimode process monitoring based on Bayesian method. *J Chemometr*. 2009;23(12):636-650.
13. Zhao C, Wang W, Qin Y, Gao F. Comprehensive subspace decomposition with analysis of between-mode relative changes for multimode process monitoring. *Ind Eng Chem Res*. 2015;54(12):3154-3166.
14. Xie X, Shi H. Dynamic multimode process modeling and monitoring using adaptive Gaussian mixture models. *Ind Eng Chem Res*. 2012;51(15):5497-5505.

15. Yu J. A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. *Chem Eng Sci*. 2012;68(1):506-519.
16. Zhang S, Wang F, Tan S, Wang S, Chang Y. Novel monitoring strategy combining the advantages of the multiple modeling strategy and Gaussian mixture model for multimode processes. *Ind Eng Chem Res*. 2015;54(47):11866-11880.
17. Yu J, Qin SJ. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE J*. 2008;54(7):1811-1829.
18. Yu J. A new fault diagnosis method of multimode processes using Bayesian inference based Gaussian mixture contribution decomposition. *Eng Appl Artif Intel*. 2013;26(1):456-466.
19. Ma H, Hu Y, Shi H. A novel local neighborhood standardization strategy and its application in fault detection of multimode processes. *Chemometr Intell Lab*. 2012;118(7):287-300.
20. Wang G, Liu J, Zhang Y, Li Y. A novel multi-mode data processing method and its application in industrial process monitoring. *J Chemometr*. 2015;29(2):126-138.
21. Deng X, Tian X. Multimode process fault detection using local neighborhood similarity analysis. *Chinese J Chem Eng*. 2014;22(11):1260-1267.
22. Wang F, Tan S, Peng J, Chang Y. Process monitoring based on mode identification for multi-mode process with transitions. *Chemometr Intell Lab*. 2012;110(1):144-155.
23. Wang F, Tan S, Yang Y, Shi H. Hidden Markov model-based fault detection approach for a multimode process. *Ind Eng Chem Res*. 2016;55(16):4613-4621.
24. Yang Y, Ma Y, Song B, Shi H. An aligned mixture probabilistic principal component analysis for fault detection of multimode chemical processes. *Chinese J Chem Eng*. 2015;23(8):1357-1363.
25. Guo J, Yuan T, Li Y. Fault detection of multimode process based on local neighbor normalized matrix. *Chemometr Intell Lab*. 2016;154:162-175.
26. Zhong N, Deng X. Multimode non-Gaussian process monitoring based on local entropy independent component analysis. *Can J Chem Eng*. 2017;95(2):313-330.
27. Ma H, Hu Y, Shi H. Fault detection and identification based on the neighborhood standardized local outlier factor method. *Ind Eng Chem Res*. 2013;52(6):2389-2402.
28. Ma Y, Shi H, Ma H, et al. Dynamic process monitoring using adaptive local outlier factor. *Chemom Intell Lab*. 2013;127(18):89-101.
29. Song B, Shi H, Ma Y, Wang J. Multisubspace principal component analysis with local outlier factor for multimode process monitoring. *Ind Eng Chem Res*. 2014;53(42):16453-16464.
30. Song B, Tan S, Shi H. Key principal components with recursive local outlier factor for multimode chemical process monitoring. *J Process Contr*. 2016;47:136-149.
31. Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. *SIGMOD International Conference on Management of Data*. 2000;29(2):93-104.
32. Duan L, Xu L, Guo F, Lee J, Yan B. A local density based spatial clustering algorithm with noise. *Inform Syst*. 2007;32(7):978-986.
33. Gonzalez R, Huang B, Lau E. Process monitoring using kernel density estimation and Bayesian networking with an industrial case study. *Isa T*. 2015;58:330-347.
34. Hong X, Gao J, Chen S, Zia T. Sparse density estimation on the multinomial manifold. *IEEE T Neur Net Lear*. 2015;26(11):2972-2977.
35. Wang B, Yan X, Jiang Q, Lv Z. Generalized Dice's coefficient-based multi-block principal component analysis with Bayesian inference for plant-wide process monitoring. *J Chemometr*. 2015;29(3):165-178.
36. Kerkhof PVD, Vanlaer J, Gins G, et al. Analysis of smearing-out in contribution plot based fault isolation for statistical process control. *Chem Eng Sci*. 2013;104(50):285-293.
37. Downs JJ, Vogel EF. A plant-wide industrial process control problem. *Comput Chem Eng*. 1993;17(3):245-255.
38. Song B, Ma Y, Shi H. Improved performance of process monitoring based on selection of key principal components. *Chin J Chem Eng*. 2015;23(12):1951-1957.
39. Cai L, Tian X, Chen S. Monitoring nonlinear and non-Gaussian processes using Gaussian mixture model-based weighted kernel independent component analysis. *IEEE T Neur Net Lear*. 2017;8(1):122-135.
40. Gao X, Hou J. An improved SVM integrated GS-PCA fault diagnosis approach of Tennessee Eastman process. *Neurocomputing*. 2016;174:906-911.
41. Xu Y, Deng X. Fault detection of multimode non-Gaussian dynamic Bayesian independent component analysis. *Neurocomputing*. 2016;200:70-79.
42. Jiang Q, Yan X. Monitoring multi-mode plant-wide processes by using mutual information-based multi-block PCA, joint probability, and Bayesian inference. *Chemom Intell Lab*. 2014;136(9):121-137.

How to cite this article: Wang L, Deng X, Cao Y. Multimode complex process monitoring using double-level local information based local outlier factor method. *Journal of Chemometrics*. 2018;32:e3048. https://doi.org/10.1002/cem.3048