Chemometrics and Intelligent Laboratory Systems 206 (2020) 104149

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Tutorial article

Novel multivariate *q*-sigma rule focusing on process variation for incipient fault detection in dynamic processes

Bo Chen, Xiong-Lin Luo [*]

Department of Automation, China University of Petroleum Beijing, China

ARTICLE INFO

Keywords:
Process monitoring and fault diagnosis
Process incipient fault detection
Process variation
Multivariate q-sigma
Minimum control limit

ABSTRACT

When incipient faults occur in chemical processes, some variables will slightly deviate from original trajectories, and process residuals will gradually be continuously biased toward one side of their mean values, i.e., process variation will occur. Traditional indices are inadequately sensitive to this situation or achieve it at the cost of a high false alarm rate (FAR). To address this situation and explore methods with low FAR in dynamic processes, canonical variate residuals (CVRs) are generated. Then, a novel multivariate *q*-sigma (M*q*-sigma) rule is proposed to monitor CVRs. It considers the process variation mentioned above in a window and sets the control limit for each variable. When tested on a simulated process, the M*q*-sigma is highly sensitive to process variations and can detect incipient faults earlier than other methods, i.e., it has the lowest detection delay and FAR.

1. Introduction

The most important aspect in chemical processes is to ensure the stable operation of systems. However, faults are inevitable due to disturbances and equipment aging. Most faults have an evolutionary process from incipient faults to serious faults. Therefore, faults should be detected early to ensure the stable operation of systems. Methods for incipient fault detection in processes are accordingly discussed in this paper.

1.1. Literature review

Process monitoring and fault detection (PM-FD) is particularly important to maintain high-quality products and process safety. Research on PM-FD technologies has received considerable attention in recent years [1,2]. In modern industries, lots of process data can be available; thus, data-driven methods, i.e., multivariate statistical process monitoring (MSPM) techniques, are widely used [3,4]. As typical MSPM methods, principal component analysis (PCA) [5–7], partial least squares [8,9], Fisher discriminant analysis [10], independent component analysis (ICA) [11,12] and canonical variate analysis (CVA) [13,14] have drawn increasing attention. Unlike model-based methods [15], they do not need priori process knowledge. To detect faults, these MSPM methods usually need to establish a threshold and a statistical model by training data off-line. Then, online data are used to calculate statistics through comparing the statistics with previous statistical models to judge whether faults have occurred.

Many abnormal events or serious faults usually evolve from incipient faults that have small magnitudes, e.g., the explosion of a nuclear power plant in Fukushima in 2011 was caused by a small aging problem [16], and the Tianjin port explosion in 2015 was caused by a small leak of nitrocellulose [17]. Numerous methods based on basic data-driven methods have been improved to enhance detection performance [18–22], but incipient faults are always neglected because they are usually covered by noise and process trend, causing minor changes in systems. Traditional methods mentioned and improved methods, such as dynamic PCA and kernel PCA, are insensitive to them and cannot detect them effectively. Thus, some novel methods and improvements for detecting incipient faults have been presented. The classical exponentially weighted moving average and cumulative sum charts and their multivariate extension solved the limitation of the lack of sensitivity and can detect small shifts in data [23–25]. Bakshi proposed the multiscale PCA which combines the wavelet analysis to extract deterministic features and approximately decorrelate autocorrelated measurements [26]. Also based on PCA, Yoon and MacGregor applied multiresolution analysis by wavelet transformations to decompose the cumulative effects of multiscale data [27]. Reis and Saraiva proposed a multiscale statistical process control method, which fully integrated the data of different resolutions [28]. A method based on a library of basis functions provided by wavelet packets is presented for conducting multiscale statistical process control by Reis et al. [29]. Grasso and Colosimo proposed an

[*] Corresponding author.
E-mail address: luoxl@cup.edu.cn (X.-L. Luo).

https://doi.org/10.1016/j.chemolab.2020.104149
Received 15 June 2020; Received in revised form 25 August 2020; Accepted 31 August 2020
Available online 12 September 2020

automated approach to enhance multiscale signal monitoring [30]. Rato and Reis proposed dynamic PCA with decorrelated residuals, it presents low auto-correlation levels and is very sensitive to incipient faults [31]. The Kullback–Leibler divergence using PCA based on a probability distribution measure was proposed by Harmouche et al. [32]. Ji et al. proposed representative smoothing techniques and a generic fault detection index to detect incipient faults [33]. An extension of CVA, canonical variate dissimilarity analysis (CVDA), was proposed by Pilario and Cao to detect incipient faults in nonlinear dynamic processes under varying operating conditions [34]. Then, they developed the CVDA into mixed kernel CVDA [35]. Ge et al. proposed the wavelet analysis method, which showed good performance, and combined it with residual evaluation [36].

These methods perform efficiently in many industrial applications, but their indices are always based on the Mahalanobis or Euclidean distance, i.e., T^2 and Q statistics. They consider only the control limits for the last statistics and focus on the information of each moment. Determining whether several points that exceed the control limits belong to the same variable is impossible. That is, these methods do not focus on the continuous change in measuring points of the same variable. When process information is transformed into distance information, the information among all variables is considered, whereas the information of a single variable is ignored.

When a system is stable, state variables will fluctuate around their steady-state or mean values; when faults occur, process variation will appear [37], i.e., some state variables will be influenced by them and then deviate from their original trajectories, process residuals will deviate from zero to one side (long or short term), i.e., absolute deviations (the difference between measured and mean values) are greater or less than zero. In an actual process, when several points are continuously biased toward one side of the mean value and exceed the threshold, which means the condition is abnormal, the system should shut down to avoid further expansion of anomalies.

A univariate control chart was first proposed by Shewhart [37] to control product quality; it is effective in process monitoring and can set a control limit for each variable individually, but it does not consider the information among variables. For variables subject to normal distribution, if their observed values fall within the upper control limit (UCL) and lower control limit (LCL) (i.e., μ-3σ $\leq x_i \leq$ μ+3σ, where x_i denotes the ith observed value, μ denotes the mean value, and σ denotes the standard deviation), then they are under control. The three-sigma rule [38] can be used to judge whether product quality is out of control. Nevertheless, in an actual process, we should not only observe whether these points are beyond the control limit but also pay attention to process variations. Faults may occur even if the control limit is not exceeded. Especially for incipient faults, the control limit 3σ does not necessarily indicate abnormal changes.

In this work, CVDA, which is an effective tool for dynamic processes, is used to generate canonical variate residuals (CVRs). Then, a multivariate q-sigma rule is proposed to monitor continuous abnormal changes in each CVR, i.e., process variation. It sets the UCL and LCL for each CVR and considers the information among all CVRs. The distribution of incipient and serious faults is monitored and discussed through changing the scope of the two control limits.

1.2. Problem statement and motivation

Statement 1: Low false alarm rate (FAR) with high detection delay (DD). Sensitivity, promptness, and robustness are highly concerned in fault detection [39]. Robustness is determined using FAR, sensitivity is determined using fault detection rate (FDR), and promptness is determined using DD. FAR measures the probability of false alarms, and a false alarm is an indication of a fault when a fault has not occurred. FDR measures the probability of successful fault detection, and a successful fault detection is an indication of a fault when a fault has occurred. DD is the time period between the start of a fault and the time of the detection time (DT) and DT is the first time after several consecutive alarms are raised. Robustness always contradicts sensitivity and promptness, i.e., FAR, FDR, and DD are difficult to consider simultaneously. Traditional methods always consider FAR first, i.e., a low FAR. Nonetheless, a low FAR will cause low FDR, then DD will increase. High DD is unconducive to incipient fault detection. The hypothetic relationship between FAR and DD is shown in Fig. 1. We always make the confidence level larger than 95%, i.e., making FAR less than 5%. If we want to detect the fault t hours after it occurs, methods 1, 2, and 3 cannot detect it within 5% FAR or they achieve it at the cost of high FAR. Thus, one of the objectives of this work is to explore methods with low DD and FAR, such as method 4.

Statement 2: Misdiagnosis. T^2 and Q statistics focus on distance information, i.e., judging whether the distance exceeds the allowable distance in space to determine whether a fault occurs. However, the distance information is determined using all variables. That is, several consecutive statistics exceeding the allowable distance may not be caused by the same variable, as shown in Fig. 2. Fig. 2(a) and (b) show that the fourth point of x and the third point of y are out of control. Only one point exceeds its control limit, which is usually considered a false alarm because a certain degree of false alarm is allowed in an actual industrial process. Nevertheless, when distance information is used to represent process information, misdiagnosis may occur. As shown in Fig. 2(c), the distance information presents that the blue dot in the box indicates that it is under control, and the red dot outside the box indicates that it is out of control. The third and fourth points continuously exceed the control limit; because the probability of multiple points exceeding the control limit is considerably less than that of one point, we usually think it is a fault rather than a false alarm. In fact, the third and fourth points are contributed by y and x, respectively. From Fig. 2(a) and (b), the situation should be a false alarm, not a fault. One of the objectives of this work is to prevent this misdiagnosis.

Statement 3: Process variation. When several points are continuously biased toward one side of the mean value, but these points do not exceed the control limit, judging whether a fault occurs in accordance with the distance information is difficult, as shown in Fig. 3. In this figure, the blue and red dots represent the normal samples and faults, respectively. Incipient faults are always sufficiently small to be covered by noise and disturbance. Consequently, they are difficult to be detected using the control limit obtained through normal condition before they evolve gradually to be serious faults that exceed the allowed range. That is, incipient faults are difficult to detect by using distance information. According to Shewhart [37], when faults occur, process variations will

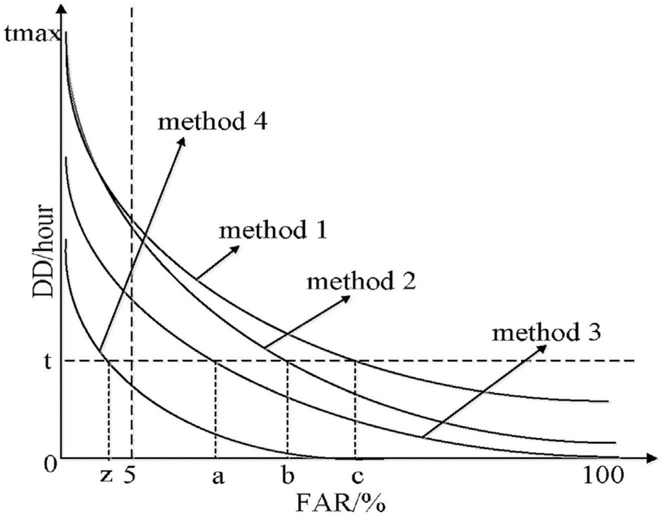

Fig. 1. Hypothetic relationship between FAR and DD. t is the DD we want; tmax denotes the maximum DD; a, b, c, and z are FARs of the four methods corresponding to t.

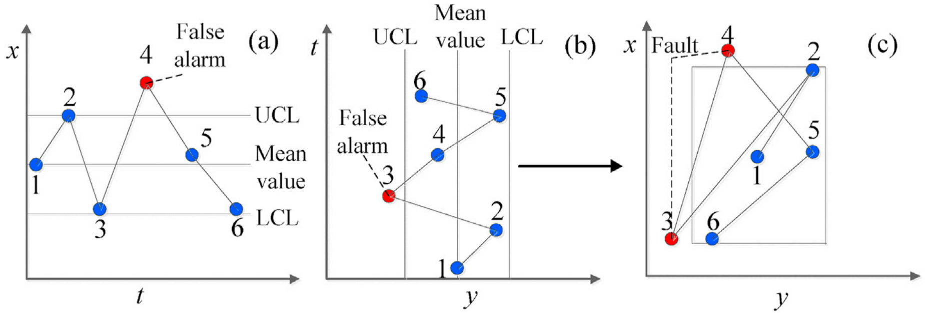

Fig. 2. Example of process information of two variables: (a) change in variable x, (b) change in variable y, (c) distance information of process (red dot: out of control; blue dot: under control). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

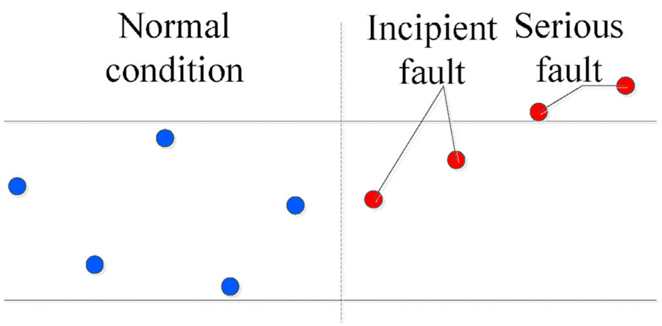

Fig. 3. Simple example of process variation.

occur; hence, incipient faults can be detected by monitoring process variations. A process variation means that when the condition is normal, state variables will fluctuate around their steady-state values; when faults occur, state variables will be continuously biased toward one side of their steady-state values. Fig. 3 shows an example of process variation.

To trigger an alarm, the point must exceed the control limit, i.e., incipient faults can be detected by reducing the control limit. Nonetheless, the traditional square statistical method monitors the distance information at every moment. When the control limit is reduced to detect incipient faults, the false alarm will increase. Thus, the current work uses a window as a unit to monitor the variation information of each variable. When the alarm is triggered, a process variation occurs in the window, rather than a point is out of control. Misdiagnosis and false alarm will be reduced. The minimum control limit ensuring that no false alarm occurs can be obtained by reducing control limit to detect process variations in fault condition as early as possible.

The detailed contributions are as follows. In Section 2, the details of traditional CVA monitoring and CVR are revisited. Section 3 focuses on the situation in which absolute deviations are greater or less than zero in long or short term. To monitor process variation of each variable and explore methods with low FAR, a multivariate q-sigma rule is proposed, and CVR combined with a multivariate q-sigma rule index (CVR-Mq-sigma) is formed. The motivation behind the index and the methodology are introduced. Section 4 contains the description of the case, results, and discussion. The results show that the proposed method is superior to CVA, CVDA, and generalized canonical correlation analysis (GCCA) [40] in incipient fault detection; in particular, CVR-Mq-sigma can detect faults at the earliest with the lowest FAR. Therefore, this method is further analyzed, such as the distributions of incipient and serious process variations. The conclusion of this paper and the intended future work are indicated in Section 5.

2. CVA revisited

Similar to PCA, CVA is a linear dimension reduction technique, but it maximizes the correlation between two data sets. It is widely used in industrial processes. CVDA is based on CVA, and CVR is generated via CVDA. The details about process monitoring are as follows.

2.1. CVA training

For the observation vector y containing m variables

$$y \in \mathrm{R}^{m} \tag{1}$$

It is expanded at time k, and the past data vectors$p_k$and future data vectors$f_k$are expressed as

$$p_k = \left[y_{k-1}, y_{k-2}, \cdots, y_{k-p}\right]^{\mathrm{T}} \in \mathrm{R}^{mp} \tag{2}$$

$$f_k = \left[y_k, y_{k+1}, \cdots y_{k+f-1}\right]^{\mathrm{T}} \in \mathrm{R}^{mf} \tag{3}$$

where, p and f are the numbers of lags considered in the past and future windows of data, respectively. $p_k$and$f_k$are then normalized to zero mean and unit variance.

For a training data set with N number of samples, the past and future Hankel matrices are formed by$p_k$andf_k, $k \in [p+1, p+S]$.

$$Y_p = \left[p_{p+1}, p_{p+2}, \cdots, p_{p+S}\right] \in \mathrm{R}^{mp\times S} \tag{4}$$

$$Y_f = \left[f_{p+1}, f_{p+2}, \cdots, f_{p+S}\right] \in \mathrm{R}^{mf\times S} \tag{5}$$

where, $S = N - p - f + 1$.

Then, the sample covariance and cross covariance of $Y_p$and $Y_f$can be obtained.

$$\Sigma_{pp} = \frac{1}{S-1} Y_p Y_p^{\mathrm{T}} \in \mathrm{R}^{mp\times mp} \tag{6}$$

$$\Sigma_{ff} = \frac{1}{S-1} Y_f Y_f^{\mathrm{T}} \in \mathrm{R}^{mf\times mf} \tag{7}$$

$$\Sigma_{fp} = \frac{1}{S-1} Y_f Y_p^{\mathrm{T}} \in \mathrm{R}^{mf\times mp} \tag{8}$$

A singular value decomposition is used to maximize the correlation of $Y_p$andY_f.

$$\Sigma_{ff}^{-1/2} \Sigma_{fp} \Sigma_{pp}^{-1/2} = U \Sigma V^{\mathrm{T}}, \tag{9}$$

where, U and V are singular vectors, and$\Sigma$is the diagonal matrix of descending singular values. Then, CVs, i.e., $X_p$and$X_f$with a maximizing correlation, can be obtained using the projection matrices J and L from$Y_p$andY_f.

$$J = V^{\mathrm{T}} \Sigma_{pp}^{-1/2} \tag{10}$$

$$L = U^{\mathrm{T}} \Sigma_{ff}^{-1/2} \tag{11}$$

$$X_p = JY_p \in \mathrm{R}^{mp \times S} \tag{12}$$

$$X_f = LY_f \in \mathrm{R}^{mf \times S} \tag{13}$$

2.2. CVA monitoring

When CVA is used for online monitoring, T^2 and Q statistics are two widely used indices. In fact, most system dynamic behavior is explained using only n strongly correlated CVs [34]. At time k, the index T^2 is expressed as

$$T_k^2 = p_k^{\mathrm{T}} J_n^{\mathrm{T}} J_n p_k, \tag{14}$$

where, J_n contains the first n rows of J.

The residual vector index Q at time k is expressed as

$$Q_k = p_k^{\mathrm{T}} F^{\mathrm{T}} F p_k, \tag{15}$$

where, $F = (\mathrm{I} - V_n V_n^{\mathrm{T}}) \Sigma_{pp}^{-1/2}$ ($V_n$contains the first n columns of V).

The kernel density estimation (KDE) is used to estimate the UCLs of the above mentioned two indices. The radial basis function is chosen as the kernel in this paper

$$\mathrm{K}(\mathrm{g}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mathrm{g}^2}{2}\right), \tag{16}$$

Given a specific confidence levelα, UCLs (T^2_{UCL}andQ_{UCL}) can be obtained using $\mathrm{P}(T^2 < T^2_{\mathrm{UCL}}) = \alpha$ and$\mathrm{P}(Q < Q_{\mathrm{UCL}}) = \alpha$, with

$$\mathrm{P}(x < \mathrm{b}) = \int_{-\infty}^{\mathrm{b}} \frac{1}{S\mathrm{h}} \sum_{k=1}^{S} \mathrm{K}\left(\frac{x - x_k}{\mathrm{h}}\right) \mathrm{d}x, \tag{17}$$

wherex_k, $k = 1, 2, \ldots, S$ are the samples of x, and h is the kernel bandwidth. Additional details on KDE can be found in Ref. [13].

The indices are calculated using Eqs. (14) and (15) from real-time data and compared with corresponding UCLs to detect whether faults occur.

2.3. CVR

In traditional CVA, past and future-projected CVs are generated in the monitoring process. However, only past-projected CVs are used for process monitoring, whereas future-projected CVs are always ignored. Thus, the concept in which the dissimilarity between past- and future-projected CVs is applied for indicating process health is introduced. Larimore suggested a statistical index that quantifies model residuals between the past- and future-projected CVs in the CVA state subspace [41]. The CVR, z is shown as follows

$$z_k = L_n f_k - \Sigma_n J_n p_k \in \mathrm{R}^n, \tag{18}$$

where, z_k denotes the CVR at time k, L_n contains the first n rows of L, and Σ_n contains the first n rows and n columns of Σ. When no fault occurs, the expectation of CVR is:

$$\mathrm{E}(z_k) = L_n \mathrm{E}(f_k) - \Sigma_n J_n \mathrm{E}(p_k) = 0 \tag{19}$$

In the opinion of Pilario and Cao [34], CVRs can be regarded as the dissimilarity features that could measure the departure of past-projected CVs from future-projected CVs. On this basis, the CVDA index is formed and tested in a continuous stirred tank reactor. The results show that CVR is an effective feature for dynamic process monitoring. Thus, it is used to handle dynamic process and generate residual in this work.

3. Multivariate *q*-sigma rule-based monitoring strategy

The three-sigma rule has been used to monitor a single variable subject to normal distribution. When the condition is normal, almost all samples of the variable are within UCL (μ+3σ) and LCL (μ-3σ); this condition can be expressed as

$$\begin{aligned} &\mathrm{P}(\mu - \sigma \le x \le \mu + \sigma) \approx 0.6827 \\ &\mathrm{P}(\mu - 2\sigma \le x \le \mu + 2\sigma) \approx 0.9545 \\ &\mathrm{P}(\mu - 3\sigma \le x \le \mu + 3\sigma) \approx 0.9973, \end{aligned} \tag{20}$$

where, x denotes the observed value, μ denotes the mean value, and σ denotes the standard deviation.

Eq. (20) shows that approximately 68.27% of the samples fall within the first control limit; 95.45% of the samples fall within the second control limit; and 99.73% of the samples fall within the third control limit. The third control limit is typically used to monitor product quality. However, the three-sigma rule will be inadequately sensitive to it because incipient faults always have small magnitudes. DT is critical to incipient fault detection [34]; under the condition of ensuring FAR, the shorter the DT is, the better. To decrease DT, the control limit should be narrowed to make incipient fault points trigger the alarm, such as the two-sigma rule, which can be explained by Fig. 4. From Fig. 4, when control limits are narrowed, increasing points (whether normal or fault) are out of control and trigger the alarm. However, only the point under fault condition will trigger the alarm continuously, i.e., when continuous alarm is given, a process variation occurs, it is also the biggest difference between fault condition and normal condition. On this basis, several consecutive points in a window are monitored. If they all trigger the alarm, then a fault has occurred; otherwise, the condition is normal. Meanwhile, a window is used as a unit to observe the process variation will also reduce FAR. The detail of the proposed method is as follows.

Traditional methods set only the control limit for the last statistics. On the contrary, control limits are considered for each variable here. When a sample exceeds the control limit, the sample is out of control.

In math, when $x > 0$, it equals its absolute value, then we have

$$x - |x| = 0 \tag{21}$$

Thus, for UCL, when one point is out of control, the following equation must exist:

$$(z_{ij} - (\mu_i + q\sigma_i)) - |z_{ij} - (\mu_i + q\sigma_i)| = 0, \tag{22}$$

where, z_{ij} is the jth sample of the ith CVR; μ_i is the mean value of the ith CVR, which equals zero here; σ_i is the standard deviation of the ith CVR; and $0 \le q \le 3$.

For each variable, at time k, w samples in a window exceed the UCL; we have

$$\sum_{j=k}^{w+k-1} \left((z_{ij} - (\mu_i + q\sigma_i)) - |z_{ij} - (\mu_i + q\sigma_i)| \right) = 0 \tag{23}$$

For MSPM, the information of each variable should be linked. Given any variable out of control, the index is equal to zero, which indicates that process variations occur. Thus, for UCL, at time k, the index M_1 of n variables is

$$M_{1k} = \prod_{i=1}^{n} \sum_{j=k}^{w+k-1} \left((z_{ij} - (\mu_i + q\sigma_i)) - |z_{ij} - (\mu_i + q\sigma_i)| \right) \tag{24}$$

For LCL, at time k, index M_2 can be obtained.

$$M_{2k} = \prod_{i=1}^{n} \sum_{j=k}^{w+k-1} \left((z_{ij} - (\mu_i - q\sigma_i)) + |z_{ij} - (\mu_i - q\sigma_i)| \right) \tag{25}$$

Two indices are combined into one index M.

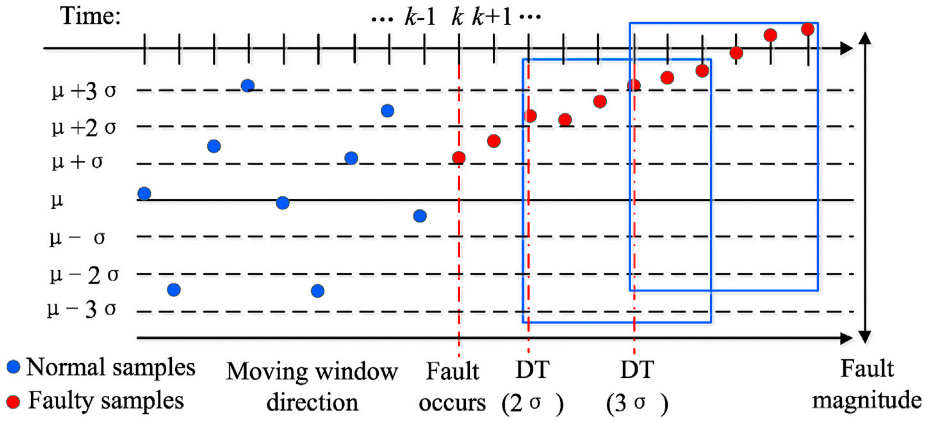

Fig. 4. Detection of an incipient fault by using μ±*q*σ with a moving window, 0 ≤ *q* ≤ 3.

$$M_k = M_{1k} M_{2k} \tag{26}$$

When M_k is equal to zero, *w* consecutive samples of at least one variable exceed the UCL or LCL, which indicates that a process variation occurs at time *k*; thus, the alarm value of *M* is zero. For traditional indices, when several consecutive points exceed the UCL, determining whether they belong to the same variable is difficult. For example, if six points are considered, and the first three points that exceed the UCL may be contributed by the first variable, but the last three that exceed the UCL may be contributed by the second variable; thus, a fault may not occur, as described in Fig. 2.

To apply the CVR-M*q*-sigma to process monitoring, the mean value (zero here) and the standard deviation of each CVR should be computed off-line. The online monitoring computes M_k in real time to check whether the system is normal or faulty. The CVR-M*q*-sigma procedure for incipient fault detection is provided in Fig. 5.

In this study, the monitoring performance of CVR-M*q*-sigma is evaluated in accordance with DD, FAR, and FDR. DD is the time period

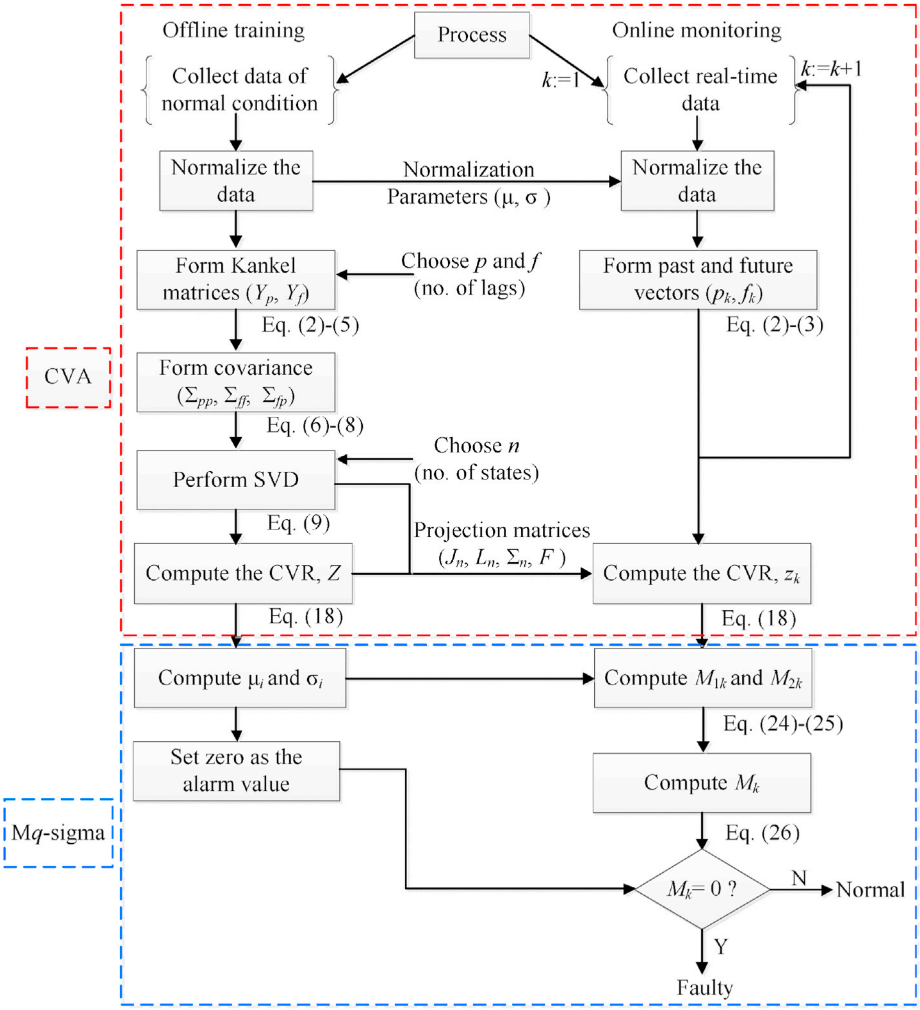
Offline training
Process
Online monitoring
Collect data of normal condition
k:=1
Collect real-time data
k:=k+1
Normalize the data
Normalization Parameters (μ, σ)
Normalize the data
Form Kankel matrices (Y_p, Y_f)
Choose p and f (no. of lags)
Form past and future vectors (p_k, f_k)
Eq. (2)-(5)
Eq. (2)-(3)
CVA
Form covariance (Σ_{pp}, Σ_{ff}, Σ_{fp})
Eq. (6)-(8)
Choose n (no. of states)
Perform SVD
Eq. (9)
Projection matrices (J_n, L_n, Σ_n, F)
Compute the CVR, Z
Compute the CVR, z_k
Eq. (18)
Eq. (18)
Compute μ_i and σ_i
Compute M_{1k} and M_{2k}
Eq. (24)-(25)
Set zero as the alarm value
Compute M_k
Mq-sigma
Eq. (26)
M_k= 0 ?
N
Normal
Y
Faulty

Fig. 5. CVR-M*q*-sigma procedure for incipient fault detection.

between the start of a fault and DT. Here, DT is the first time that index M equals zero (Fig. 4). This method considers w consecutive samples not one, its FAR and FDR will be low. FAR and FDR are computed as follows:

$$\mathrm{FAR} = \frac{\text{no. of samples } (M = 0 \mid \text{fault} - \text{free})}{\text{total samples } (\text{fault} - \text{free})} \times 100\%, \tag{27}$$

$$\mathrm{FDR} = \frac{\text{no. of samples } (M = 0 \mid \text{fault})}{\text{total samples } (\text{fault})} \times 100\% \tag{28}$$

4. Case study

In this section, the performance of the proposed CVR-Mq-sigma method is illustrated with a simulated process (i.e., Tennessee Eastman [TE] process). We also test the proposed method on a real industrial process (i.e., multiphase flow process), but due to length limitations, this case study can be seen in the supporting information, the results also support our point of view. The CVA, CVDA, and GCCA methods are compared with the proposed method. Their performance are analyzed and discussed. For CVA, CVDA, and GCCA, DD is the time period between the start of a fault and DT. DT is the first time after w consecutive alarms are raised, and its objective is to avoid short false alarms produced by an indicator before the actual detection of the fault. The FAR and FDR of CVA, CVDA, and GCCA differ from those of CVR-Mq-sigma and are computed as follows.

$$\mathrm{FAR} = \frac{\text{no. of samples } (I > I_{\mathrm{UCL}} \mid \text{fault} - \text{free})}{\text{total samples } (\text{fault} - \text{free})} \times 100\%, \tag{29}$$

$$\mathrm{FDR} = \frac{\text{no. of samples } (I > I_{\mathrm{UCL}} \mid \text{fault})}{\text{total samples } (\text{fault})} \times 100\%, \tag{30}$$

where, $I = \{T^2, Q, D\}$.

4.1. Process description

TE process was proposed in accordance with an actual chemical plant of Eastman Company by Downs and Vogel [42] in 1993; it has been widely used as a test problem for process monitoring, fault diagnosis, and process control technology. The process includes five major units: a reactor, a condenser, a separator, a recycle compressor, and a product stripper. Two products are created from four exothermic reactants. TE has 41 measured variables and 12 manipulated variables. Readers can refer to Reference [42] for further details. In this work, only 50 variables are used to reflect the state of TE process because the compressor recycle valve, stripper steam valve, and agitator speed are constant under the control strategy. The first 13 faults of TE process, including step change faults (faults 1–7), random variation faults (faults 8–12), and slow drift faults (fault 13), are used to evaluate the monitoring performance of proposed index. Original fault data of TE process are used in most papers, and many methods have good monitoring performance. One of the features of incipient faults is small magnitude. Thus the magnitudes of these faults are reduced to 10% of original values in this work. Then, the data are generated. For process monitoring, each run for each fault lasts 48 h, and the sampling period is 0.05 h. A fault is introduced after 8 h of simulation, i.e., 960 samples exist for each fault. The first 160 samples are normal data, and the last 800 samples are fault data. They are considered the test data. For training data, 960 samples are also available, but no faults are introduced. The MATLAB code can be downloaded from the following website: http://depts.washington.edu/control/LARRY/TE/download.html.

5. Results and discussion

The performance of CVA, CVDA, GCCA, and CVR-Mq-sigma is compared. According to Rato [43,44], contrary to the FDR, the FAR should not be dependent upon the method; all methods should have the same FAR to start with. Only then the detections results can be compared. Because the proposed method must satisfy that there is no process variation in normal condition, i.e., its FAR is 0. Thus, for fair comparison, all UCLs of CVA, CVDA, and GCCA are adjusted to make their FARs equal 0 exactly. In accordance with the research on TE in Reference [45], the window width is 6, i.e., six consecutive points are considered.

The number of past and future lags (p and f) and the number of states (n) must be chosen because the proposed index is based on CVA. The lags are determined using the autocorrelation function of the summed squares of all measurements [13]; p and f are both set to 5, which is the maximum number of significant lags based on the autocorrelation analysis on training data. Then, n is determined using FAR, which minimizes FAR of CVA [46], thus, n is set to 15. The Akaike information criterion [45] can be also used to select n. For comparison, the parameters of GCCA are used as defined in Reference [40], and CVA-based methods have the same p, f, and n.

The Jarque–Bera (JB) test can be used to check the Gaussianity [47]. It is defined as follows:

$$\mathrm{JB} = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right), \tag{31}$$

where, n is the sample size, s is the sample skewness, and k is the sample kurtosis. In this case study, the logical values of CVR JB test equal 0 at the 5% significance level, which indicates that all CVRs of training set obey a normal distribution. Therefore, the q-sigma rule can be used here.

When the condition is normal, the CVRs are distributed around zero mean. When faults occur, some of them will change, as shown in Fig. 6. Fault 1 occurs, the fourth CVR changes seriously (long term), and Fault 3 occurs, and changes slightly (short term, difficult to see by humans). This work focuses on these process variations.

The performance comparison of Fault 3 is shown in Fig. 7. The UCL and alarm value are marked; if the fault is detected, the DT is also marked, and the first point of six consecutive alarms is marked using black dot. The fault is introduced at 8 h. Fig. 7 shows that the four methods hardly detect this incipient fault. Some points exceed the control limit, but they are not continuous; thus, DT is large. In terms of DT, T^2 of CVA detects the fault at 36.85 h, T^2_{r1} of GCCA detects the fault at 34.95 h, and the Q of CVA, CVDA, and Mq-sigma ($q = 3$) fail to detect Fault 3 (no six consecutive points). Fig. 7(e) and (f) show that for CVR-Mq-sigma, when $q = 3$, the index is far away from the alarm value 0, and Fault 3 is almost the same as the normal condition; when q is reduced to 1, the fault is detected, i.e., six consecutive points of at least one variable exceeding 1σ exist, and a process variation is detected at 10.2 h. No any false alarm occurs for CVR-Mq-sigma, which indicates that even in the range of 1σ to 3σ, no process variation occurs in normal condition.

Horizontal dashed line: UCL or alarm value; vertical dashed line: start of fault; solid line: detection index.

Table 1 shows the monitoring performance of the four methods. Although the fault magnitudes are reduced to 10% of original values, Faults 1, 2, 6, 7, 8, and 13 still have strong influences on the system, they can be regarded as serious faults. The four methods can easily detect them and have high FDR. For the remaining faults, some methods fail to detect them. They can be regarded as incipient faults, such as Fault 5, and only T^2 of CVA, T^2_{r1} of GCCA, and CVR-Mq-sigma ($q = 1$) are effective. In terms of FDR of some faults, GCCA can detect more points exceeding the UCL than others, but it is not the index that detects faults first. The proposed method has better performance than others in terms of DD. Table 1 also indicates the influence of q on the DD and FDR of CVR-Mq-sigma, i.e., when q is smaller, DD is short and FDR is higher. CVR-Mq-sigma has the best performance, especially in DD, when q equals 1. However, regardless of whether q is 3 or 1, its FAR always equals 0 because six consecutive points must be beyond the UCL or LCL to consider a fault. No six consecutive points exceed the control limit under normal condition, although q equals 1, i.e., no process variation occurs

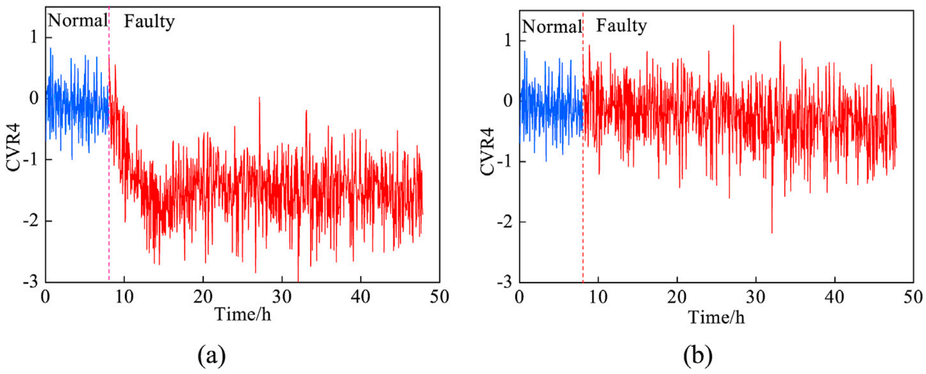

Fig. 6. Change in the fourth CVR when (a) Fault 1 occurs and (b) Fault 3 occurs.

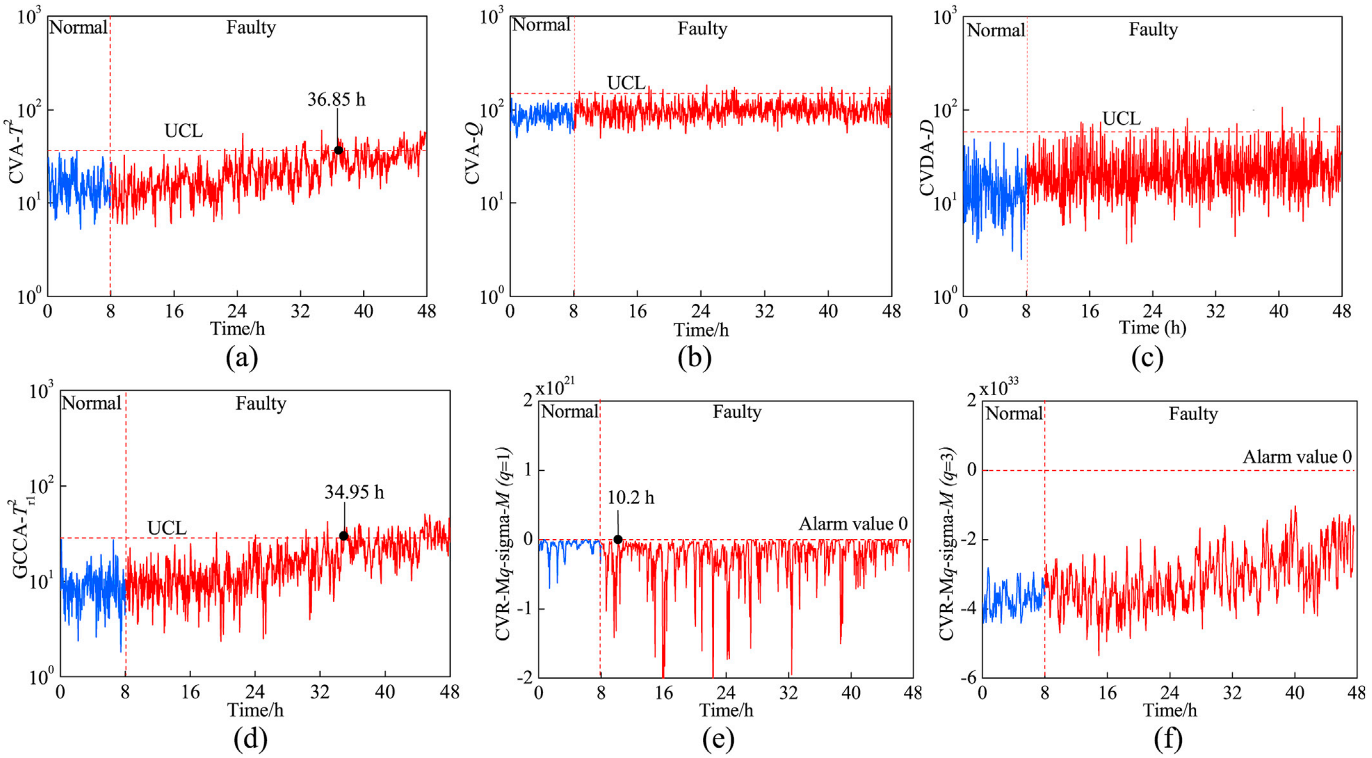

Fig. 7. Monitoring charts for Fault 3 by using.
(a) and (b) CVA, (c) CVDA, (d)T_{r1}^2 of GCCA, (e) and (f) CVR-Mq-sigma, in which $q = 1$ and $q = 3$.

under normal condition. On this basis, q can be further reduced. Fig. 8 shows the relationship between its FAR and q under normal condition.

As shown in Fig. 8, a false alarm occurs when q is less than 0.82. Thus the q is reduced to 0.82, and it represents the minimum control limit under the condition that no process variation occurs under normal condition. The performance is shown in the last column of Table 1. The results show that DD is further reduced for Fault 3. For other faults, FDRs are also increased, but the performance improvement is generally inconsiderable. In an actual industrial process, faults should be detected as early as possible to prevent them from becoming serious accidents, i.e., DD is particularly important for incipient fault detection. FDR is suitable for assessing the sensitivity of a detection index off-line not in real time. When faults are detected, the system should be shut down, instead of waiting for FDR to reach a certain value. Determining how many fault points exist in total in real time is also difficult. In general, CVR-Mq-sigma is a highly effective tool that can detect faults at an early stage from the perspective of process variation.

Table 1 also presents that for incipient faults (Faults 3, 4, 5, 9, 11, 12), the change of q has strong influence on the detection performance, by reducing q, it can even detect the fault that is difficult to detect, and 3σ is ineffective. For other serious faults, even if q is reduced from 3 to 1, there is no significant change in the detection performance, especially in DD; process variation can be detected when q equals 3. The process variations for incipient faults are mainly distributed between 1σ and 3σ, and 1σ should be used to detect them; the process variations for serious faults are distributed outside 3σ, and q equals to 3 will be fine.

From another point of view, the confidence level for traditional fault detection methods for setting a threshold is excessively large because their first consideration is low FAR. As Chen et al. indicated in their work [40], the main objective for introducing a threshold is to reduce FAR to an acceptable level, and a threshold that leads to zero FAR is the best. A large threshold will usually cause considerable points to be below the threshold; thus, low FAR will also lead to low FDR. Then, DD will increase, which will result in insensitivity to incipient faults, and faults cannot be detected at the early stage. Fault 3 is regarded as an example, and the confidence level and q are reduced. The relationship between

Table 1
Monitoring performance for TE incipient faults.

Fault	Index	CVA		CVDA	GCCA		Proposed CVR-Mq-sigma			
		T^2	Q	D	T^2_{r1}	T^2_{r2}	$q = 3$	$q = 2$	$q= 1$	$q= 0.82$
							M			
Free	FAR	0	0	0	0	0	0	0	0	0
1	DD	0.55	0.45	0.55	0.55	0.55	0.65	0.20	**0.15**	0.15
	FDR	98.62	99.12	98.62	98.62	98.62	97.60	99.49	**99.62**	99.62
2	DD	4.75	3.95	14.20	5.85	6.35	2.00	0.65	**0.20**	0.20
	FDR	86.95	90.34	53.32	88.62	83.63	18.06	63.01	**95.33**	97.60
3	DD	28.85	ND	ND	26.95	ND	ND	ND	**2.20**	2.15
	FDR	11.54	5.90	4.89	**12.25**	0	0	0	1.64	3.66
4	DD	36.75	16.40	ND	26.95	ND	ND	ND	**2.15**	2.15
	FDR	11.29	27.60	5.77	**29.00**	0.25	0	0	1.39	5.43
5	DD	36.75	ND	ND	36.85	ND	ND	ND	**2.15**	2.15
	FDR	**10.92**	4.64	5.27	6.00	0	0	0	1.52	3.91
6	DD	0	0	0.35	0	0	0.30	0	0	0
	FDR	100	100	95.61	100	100	15.91	93.60	100	100
7	DD	0	0	0	0	0	0	0	0	0
	FDR	100	100	100	100	100	100	100	100	100
8	DD	1.65	1.75	2.55	1.90	2.65	2.45	1.50	**0.90**	0.90
	FDR	**79.17**	67.25	28.11	63.00	45.12	7.95	25.63	64.39	69.70
9	DD	28.65	ND	ND	36.85	ND	ND	ND	**2.15**	2.15
	FDR	**11.04**	4.89	5.14	6.12	0	0	0	1.64	4.17
10	DD	14.95	1.05	14.80	24.35	1.20	0.85	0.80	**0.75**	0.75
	FDR	20.45	43.16	42.41	13.13	31.62	14.77	34.34	**55.05**	60.48
11	DD	28.65	11.05	ND	24.65	ND	5.15	3.75	**0.40**	0.40
	FDR	12.55	**33.75**	12.92	28.0**0**	2.13	0.63	3.41	12.25	16.92
12	DD	ND	ND	ND	36.85	ND	ND	ND	**2.15**	2.15
	FDR	**10.79**	5.02	5.27	6.50	0	0	0	1.64	4.17
13	DD	8.85	8.05	8.50	8.75	11.50	3.60	3.10	**2.15**	2.15
	FDR	70.51	73.15	54.58	70.37	51.50	49.62	58.33	**73.23**	77.02

DD: detection delay (hours); FDR: fault detection rate (%); FAR: false alarm rate (%); ND: not detected (consistently for six consecutive sampling times).

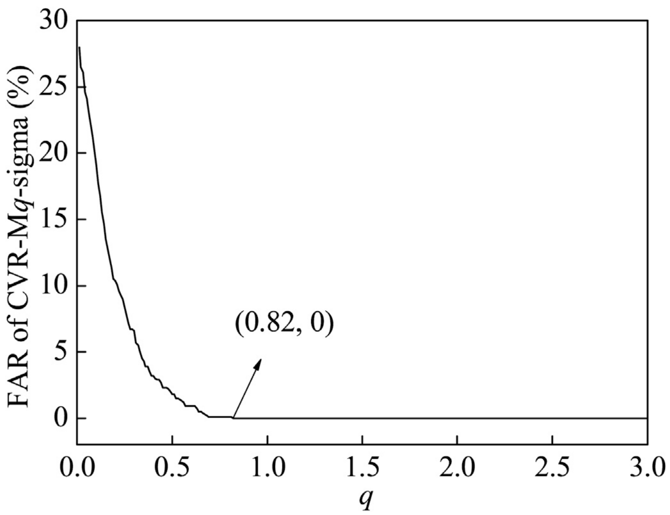

Fig. 8. Influence of q on FAR of CVR-Mq-sigma.

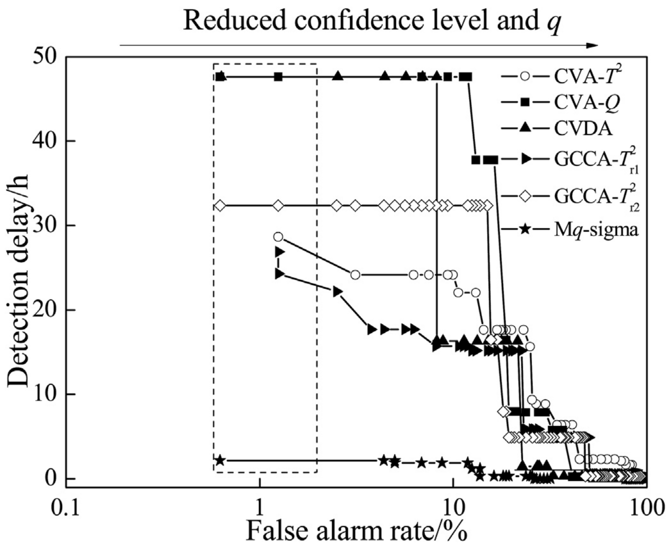

Fig. 9. Relationship between FAR and DD when confidence level and q are reduced in consideration of Fault 3 as an example. Black dashed box represents the low FAR.

FAR and DD for the four methods is shown in Fig. 9. When the confidence level and q are reduced, FARs are increased, whereas DDs are reduced (six sample points must be considered for DD; hence, DD may not be continuously reduced). For process control systems, the accuracy and rapidity of are difficult to guarantee at the same time. The aspects to consider are first stability, then accuracy, and finally rapidity. For fault detection, we always focus on low FAR, (the black dashed box in Fig. 9), resulting in a delay in fault detection. Unlike in process control systems, rapidity should be a priority in incipient fault detection because only when faults are detected and controlled in time can process control systems be stable.

Fig. 9 depicts that in terms of CVA, CVDA, and GCCA, when FAR is small, T^2_{r1} of GCCA is the first index to find the fault; when FARs are increased to approximately 10%, their DDs are suddenly reduced and indices are sensitive to the fault. The results also indicate that if the faults are aimed to be detected in time, it needs to be at the cost of high FAR. However, frequent false alarms are not what we want, and traditional methods consider a fault when several consecutive points exceed the threshold. Thus, CVR-Mq-sigma uses the concept of process variation in FAR and FDR because the probability of several consecutive points exceeding the threshold value is much smaller than that of one point. Its FAR is lower than those of traditional methods at the cost of less important FDR in real-time monitoring. The star line in Fig. 9 shows low DD and low FAR, because CVR-Mq-sigma considers control limits for each variable; it is also highly sensitive to incipient faults. CVR-Mq-sigma generally has lower DD at low FAR than traditional methods with higher DD at low FAR.

6. Conclusion

In this work, CVR-Mq-sigma is proposed for MSPM to address the issue that process variations appear when incipient faults occur. To prevent misdiagnosis, the proposed method considers control limits for each CVR. Focusing on process variations in each CVR, the proposed method observes several continuous points in a window, and the minimum control limit of each fault is obtained to detect faults as early as possible and prevent false alarms. When tested on TE process, the proposed method can detect incipient faults earliest and with the lowest false alarm, i.e., 0 false alarm, compared with CVA, CVDA, and GCCA. The results also show that (1) process variations for incipient faults are mainly distributed between 1σ and 3σ or −1σ and −3σ, and serious process variations are distributed outside 3σ or −3σ. (2) Traditional methods focus on low FAR, resulting in a large threshold, which will be insensitive to incipient faults. DD can be reduced at the cost of high FAR. (3) CVR-Mq-sigma considers control limits for each variable and the change in several continuous points. It can detect faults early with a low FAR. It is identical to method 4 in Fig. 1. Another important target for chemical processes is to eliminate faults to ensure the stable operation of systems. Thus, a self-recovery control method that will eliminate faults and does not need human intervention will be studied in the future work.

CRediT authorship contribution statement

Bo Chen: Methodology, Data curation, Software. **Xiong-Lin Luo:** Conceptualization, Supervision, Validation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (21676295).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2020.104149.

References

[1] M.S. Reis, G. Gins, Industrial process monitoring in the big data/industry 4.0 Era: from detection, to diagnosis, to prognosis, Processes 5 (3) (2017) 35.

[2] Y.A. Shardt, H. Hao, S.X. Ding, A New Soft-sensor-based process monitoring scheme incorporating infrequent KPI measurements, IEEE Trans. Ind. Electron. 62 (6) (2015) 3843–3851.

[3] S.J. Qin, Survey on data-driven industrial process monitoring and diagnosis, Annu. Rev. Contr. 36 (2) (2012) 220–234.

[4] A. Beghi, R. Brignoli, L. Cecchinato, G. Menegazzo, M. Rampazzo, F. Simmini, Data-driven fault detection and diagnosis for HVAC water chillers, Contr. Eng. Pract. 53 (2016) 79–91.

[5] Z. Ge, Z. Song, F. Gao, Review of recent research on data-based process monitoring, Ind. Eng. Chem. Res. 52 (10) (2013) 3543–3562.

[6] J. Chen, C.M. Liao, F.R.J. Lin, M.J. Lu, Principle component analysis based control charts with memory effect for process monitoring, Ind. Eng. Chem. Res. 40 (6) (2001) 1516–1527.

[7] H. Chen, B. Jiang, N. Lu, Z. Mao, Deep PCA based real-time incipient fault detection and diagnosis methodology for electrical drive in high-speed trains, IEEE Trans. Veh. Technol. 67 (2018) 4819–4830.

[8] A.A. Khan, J. Moyne, D.M. Tilbury, Virtual metrology and feedback control for semiconductor manufacturing processes using recursive partial least squares, J. Process Contr. 18 (10) (2008) 961–974.

[9] S. Yin, S.X. Ding, P. Zhang, A. Hagahni, A. Naik, Study on modifications of PLS approach for process monitoring, IFAC Proceedings Volumes 44 (1) (2011).

[10] L.H. Chiang, M. Kotanchek, A.K. Kordon, Fault diagnosis based on Fisher discriminant analysis and support vector machines, Comput. Chem. Eng. 28 (8) (2004) 1389–1401.

[11] J. Fan, Y. Wang, Fault detection and diagnosis of non-linear non-Gaussian dynamic processes using kernel dynamic independent component analysis, Inf. Sci. 59 (2014) 369–379.

[12] G. Stefatos, A.B. Hamza, Dynamic independent component analysis approach for fault detection and diagnosis, Expert Syst. Appl. 37 (12) (2010) 8606–8617.

[13] P.E.P. Odiowei, Y. Cao, Nonlinear dynamic process monitoring using canonical variate analysis and kernel density estimations, IEEE Trans. Ind. Inform. 6 (1) (2010) 36–45.

[14] R.T. Samuel, Y. Cao, Kernel canonical variate analysis for nonlinear dynamic process monitoring, IFAC-PapersOnLine 48 (8) (2015) 605–610.

[15] S.X. Ding, Model-based fault diagnosis techniques - design schemes, algorithms and tools, IFAC PapersOnLine 49 (2016) 50–56.

[16] M. Tsubokura, S. Gilmour, K. Takahashi, T. Oikawa, Y. Kanazawa, Internal radiation exposure after the fukushima nuclear power plant disaster, J. Am. Med. Assoc. 308 (7) (2012) 669.

[17] B. Sun, Tianjin port explosions, Process Saf. Prog. 34 (4) (2015), 315-315.

[18] J. Xuan, Z. Xu, Y. Sun, Selecting the number of principal components on the basis of performance optimization of fault detection and identification, Ind. Eng. Chem. Res. 54 (12) (2015) 3145–3153.

[19] H. Jian, X. Yan, Dynamic process fault detection and diagnosis based on dynamic principal component analysis, dynamic independent component analysis and Bayesian inference[J], Chemometr. Intell. Lab. Syst. 148 (2015) 115–127.

[20] S. Yin, X. Zhu, O. Kaynak, Improved PLS focused on key-performance-indicator-related fault diagnosis, IEEE Trans. Ind. Electron. 62 (3) (2015) 1651–1658.

[21] W. Ku, R. Storer, C. Georgakis, Disturbance detection and isolation by dynamic principal component analysis, Chemometr. Intell. Lab. Syst. 30 (1995) 179–196.

[22] J. Shang, M. Chen, H. Ji, D. Hua, Recursive transformed component statistical analysis for incipient fault detection, Automatica 80 (80) (2017) 313–327.

[23] R.B. Crosier, Multivariate generalizations of cumulative sum quality-control schemes, Technometrics 30 (1988) 291–303.

[24] C.A. Lowry, et al., A multivariate exponentially weighted moving average control chart, Technometrics 34 (1992) 46–53.

[25] S. Wold, Exponentially weighted moving principal components analysis and projection to latent structures, Chemometr. Intell. Lab. Syst. 23 (1994) 149–161.

[26] B.R. Bakshi, Multiscale PCA with application to multivariate statistical process control, AIChE J. 44 (7) (1998) 1596–1610.

[27] S. Yoon, J.F. MacGregor, Principal-component analysis of multiscale data for process monitoring and fault diagnosis, AIChE J. 50 (11) (2004) 2891–2903.

[28] M.S. Reis, P.M. Saraiva, Multiscale statistical process control with multiresolution data, AIChE J. 52 (6) (2006) 2107–2119.

[29] M.S. Reis, B.R. Bakshi, P.M. Saraiva, Multiscale statistical process control using wavelet packets, AIChE J. 54 (9) (2008) 2366–2378.

[30] M. Grasso, B.M. Colosimo, An automated approach to enhance multiscale signal monitoring of manufacturing processes, J. Manuf. Sci. E-T. Asme. 138 (2016), 0510031-05100316.

[31] T.J. Rato, M.S. Reis, Fault detection in the Tennessee Eastman benchmark process using dynamic principal components analysis based on decorrelated residuals (DPCA-DR)[J], Chemometr, Intell. Lab. Syst. 125 (2013) 101–108.

[32] J. Harmouche, C. Delpha, D. Diallo, Incipient fault detection and diagnosis based on Kullback–Leibler divergence using principal component analysis: Part I, Signal Process. 94 (2014) 278–287.

[33] H. Ji, X. He, J. Shang, D. Zhou, Incipient fault detection with smoothing techniques in statistical process monitoring, Contr. Eng. Pract. 62 (2017) 11–21.

[34] K.E.S. Pilario, Y. Cao, Canonical variate dissimilarity analysis for process incipient fault detection, IEEE Trans. Ind. Inform. 14 (12) (2018) 5308–5315.

[35] K.E.S. Pilario, Y. Cao, M. Shafiee, Mixed kernel canonical variate dissimilarity analysis for incipient fault monitoring in nonlinear dynamic processes, Comput. Chem. Eng. 123 (2019) 143–154.

[36] W. Ge, J. Wang, J. Zhou, H. Wu, Q. Jin, Incipient fault detection based on fault extraction and residual evaluation, Ind. Eng. Chem. Res. 54 (14) (2015) 3664–3677.

[37] W.A. Shewhart, Economic Control of Quality of Manufactured Product, Van Nostrand, 1931.

[38] F. Pukelsheim, The three sigma rule, Am. Statistician 48 (2) (1994) 88–91.

[39] L.H. Chiang, E.L. Russell, R.D. Braatz, Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis, Chemometr. Intell. Lab. Syst. 51 (1) (2000) 81–93.

[40] Z. Chen, S.X. Ding, T. Peng, C. Yang, W. Gui, Fault detection for non-Gaussian processes using generalized canonical correlation analysis and randomized algorithms, IEEE Trans. Ind. Electron. 65 (2) (2018) 1559–1567.

[41] W.E. Larimore, Optimal reduced rank modeling, prediction, monitoring and control using canonical variate analysis, IFAC Proceedings Volumes 30 (9) (1997) 61–66.

[42] J.J. Downs, E.F. Vogel, A plant-wide industrial process control problem, Comput. Chem. Eng. 17 (3) (1993) 245–255.

[43] T.J. Rato, et al., A systematic methodology for comparing batch process monitoring methods: Part I – assessing detection strength, Ind. Eng. Chem. Res. 55 (18) (2016) 5342–5358.

[44] T.J. Rato, et al., A systematic methodology for comparing batch process monitoring methods: Part II – assessing detection speed, Ind. Eng. Chem. Res. 57 (15) (2018) 5338–5350.

[45] L.H. Chiang, E.L. Russell, R.D. Braatz, Fault Detection and Diagnosis in Industrial Systems, Springer-Verlag, London, U.K, 2005.

[46] C. Ruiz-Carcel, Y. Cao, D. Mba, L. Lao, R.T. Samuel, Statistical process monitoring of a multiphase flow facility, Contr. Eng. Pract. 42 (2015) 74–885.

[47] C.M. Jarque, A.K. Bera, A test for normality of observations and regression residuals, Int. Stat. Rev. 55 (2) (1987) 163–172.