IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 3, 2022, accepted May 31, 2022, date of publication June 8, 2022, date of current version June 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3181360

Fault Detection in Industrial Systems Using Maximized Divergence Analysis Approach

BENBEN JIANG[1], (Member, IEEE), AND QIUGANG LU[2]
[1]Department of Automation, Tsinghua University, Beijing 100084, China
[2]Department of Chemical Engineering, Texas Tech University, Lubbock, TX 79409, USA

Corresponding author: Qiugang Lu (jay.lu@ttu.edu)

The work of Benben Jiang was supported by NSFC under Grant 62192751. The work of Qiugang Lu was supported by the Texas Tech University.

ABSTRACT Dimensionality reduction techniques including partial least-squares (PLS) and principal component analysis (PCA) have been widely applied for data-driven process monitoring. However, the objectives of PCA- and PLS-based techniques are not specific for fault detection where a superior detection performance results from a large divergence (i.e., difference) between normal operating data and faulty data. In this article, a maximized divergence analysis (MDA) method is proposed to detect faults in industrial systems. The objective of MDA is to directly maximizes the Kullback-Leibler (KL) divergence corresponding to the distributions of normal operating data and faulty data during the procedure of dimensionality reduction. An algorithm using eigenvalue-decomposition technique is put forward to efficiently solve the optimization problem of maximizing KL-divergence. Two-dimensional synthetic data and Tennessee Eastman process are used to demonstrate the effectiveness of the proposed MDA-based detection approach.

INDEX TERMS Dimensionality reduction technique, fault detection, fault diagnosis, process monitoring, Kullback-Leibler divergence, Tennessee Eastman process.

I. INTRODUCTION

Fault detection, which serves as the first procedure in process monitoring scheme, focuses on determining whether a fault has occurred [1], [2]. Detecting faults accurately and efficiently is crucial for the safety and reliability of practical industries, since a fault or change in a complicated industrial process may quickly evolve into a disastrous accident [3], [4]. An example is that the oil spill of Deepwater Horizon in 2010 caused economic loss over $90 billion [5].

Over last decades, data-driven process monitoring techniques have received growing research interest in academia and industry [6]–[8]. Various data-based approaches have been adopted in fault detection. Principal component analysis (PCA) is one of the most widely used fault detection approaches for industrial processes, and have been applied in the chemicals industry [9], [10], semiconductor manufacturing [11], and pharmaceutical manufacturing [12]. PCA aims to produce lower dimensional representations of the original data by maximizing the variance during the procedure of dimensionality reduction [2], [9]. However, the obtained projection vectors may not provide the optimal directions for separating faulty and normal classes. Although the variance is not guaranteed to be the best feature for determining directions in a lower dimensional space that is best for the detection of faults, the variance is related to the single-variable Shewart control chart. Both methods use the variance to define thresholds for fault detection, with a fault being identified as being an outlier compared to the variation observed during normal operating conditions. PCA can be interpreted as being the direct generalization of the single-variable Shewart control chart to multiple variables in which normal operations reside with a lower dimensional space.

Partial least squares (PLS) [13] is another dimensionality reduction technique commonly used for detecting faults in industrial processes. The objective of PLS is to generate lower dimensional data representations by maximizing the covariance between the original input and output data [14]. Therefore, the separability between faulty and healthy cases may not be maximized in the latent space. Unlike PCA, PLS uses an iterative algorithm such as nonlinear iterative partial least squares [14] to construct the lower dimensional representation.

The associate editor coordinating the review of this manuscript and approving it for publication was Baoping Cai.

Other dimensionality reduction techniques for detecting faults employ state-space models. These models are usually constructed using subspace identification methods such as canonical variate analysis (CVA) [15]–[17], numerical algorithm subspace-based state-space system identification (N4SID) [18], and multivariable output-error state-space (MOESP) [2]. CVA is one of the most used subspace identification techniques in industrial applications. The objective of CVA is to produce lower dimensional representations that maximize the correlation between the 'past' information of process input and output data and the 'future' information of output data [19], [20]. This method takes serial correlations into account during the dimensionality reduction procedure.

However, the objectives of PCA-, PLS-, and CVA-based dimensionality reduction techniques are not specific for fault detection in which the criterion is the maximization of the separability (i.e., difference) corresponding to the distributions of normal operating data and faulty data. A large value of separability leads to a good ability to detect faults [21]. Additionally, at least some quantity of faulty data is available from historical databases in industry, which is not optimally used in PCA-, PLS-, or CVA-based methods. The utilization of faulty data can improve the detection performance for the particular faults encountered when the data were collected. To the end, alternative methods that specifically maximize the difference between normal operating data and faulty data have promise for fault detection. Fisher discriminant analysis (FDA), which maximizes the separability between different classes while minimizing the scatter within each class, is one such method. However, FDA can only produce $m-1$ dimensional loading vectors for m-class classification problems [2], [22]. For fault detection (binary classification), FDA can only extract one loading vector, although multiple dimensions can be useful in some applications.

The Kullback-Leibler (KL) divergence, also known as discrimination information, is an efficient method to quantify the dissimilarity between two distributions of datasets [23]–[25]. A large value of KL-divergence corresponds to the probability density functions of normal operating data and faulty data being well isolated, which indicates it's easy for detecting faults. A value close to zero associates with the two distributions of datasets are similar, which means it is difficult to discriminate faulty data from normal operating data. KL divergence-based fault detection (and its direct relation with the generalized likelihood ratio) has been reported in the literature and shown exceptional performance. For instance, the incipient fault detection has been extensively based on KL divergence [30], [31] and its combination with the PCA, applied to CRH5 [31]–[33].

Previously, an effective fault detection method was developed based on the KL-divergence in which the value of KL-divergence was calculated between normal operating data and a uniform distribution data for single variables in training steps [26]. However, the information on faulty data was not utilized and no dimensionality reduction procedure on maximizing divergence was used [26]. In this regard, this article presents a novel KL-divergence-based method with the following contributions:

- We propose a fault detection method, termed as Maximized Divergence Analysis (MDA), that can directly maximize KL-divergence corresponding to the distributions of normal operating data and faulty data during the procedure of dimensionality reduction.
- An eigenvalue decomposition-based algorithm is put forward to efficiently solve the optimization problem of maximizing the KL-divergence.
- Two detection statistics are derived using the obtained and discarded loading scores in the MDA model.
- The proposed methods are verified based on a numerical example and the Tennessee Eastman process.

The rest of this article is organized as follows. Section II briefly revisits the Kullback-Leibler divergence. The MDA-based fault detection method is put forward in Section III. Section IV demonstrates the effectiveness of the developed MDA approach in two-dimensional synthetic datasets and the Tennessee Eastman process, followed by concluding remarks in Section V.

II. KULLBACK-LEIBLER DIVERGENCE REVISITED

The Kullback-Leibler (KL)-divergence, which quantifies the dissimilarity for two probability density functions, is defined as [23]

$$D(p_{nf}\,||\,p_f) = \int p_{nf}(\boldsymbol{x}) \ln \frac{p_{nf}(\boldsymbol{x})}{p_f(\boldsymbol{x})} d\boldsymbol{x}, \tag{1}$$

where p_{nf} and p_f denote the distributions of normal operating data and faulty data, respectively.

From the equation (1), the KL-divergence can be explained as the expected value of $\ln p_{nf}/p_f$ based on the probability density function of p_{nf} [23]. The KL-divergence is non-negative, i.e., $D(p_{nf}\,||\,p_f) \geq 0$, with equality if and only if $p_{nf} = p_f$.

If $p_{nf} \sim N(\boldsymbol{\mu}_{nf}, \boldsymbol{\Sigma}_{nf})$ and $p_f \sim N(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$ with $\boldsymbol{\Sigma}_{nf}$ and $\boldsymbol{\Sigma}_f$ positive definite, we can explicitly compute the KL-divergence as [23]

$$\begin{aligned} D(p_{nf}\,||\,p_f) &= \int p_{nf}(\boldsymbol{x}) \ln \frac{p_{nf}(\boldsymbol{x})}{p_f(\boldsymbol{x})} d\boldsymbol{x} \\ &= \frac{1}{2} \ln \frac{|\boldsymbol{\Sigma}_f|}{|\boldsymbol{\Sigma}_{nf}|} + \frac{1}{2} \mathrm{tr}\left\{ \boldsymbol{\Sigma}_{nf} \left(\boldsymbol{\Sigma}_f^{-1} - \boldsymbol{\Sigma}_{nf}^{-1} \right) \right\} \\ &\quad + \frac{1}{2} \mathrm{tr}\left\{ \boldsymbol{\Sigma}_f^{-1} \left(\boldsymbol{\mu}_{nf} - \boldsymbol{\mu}_f \right) \left(\boldsymbol{\mu}_{nf} - \boldsymbol{\mu}_f \right)^{\mathrm{T}} \right\} \end{aligned} \tag{2}$$

where $(\boldsymbol{\mu}_{nf}, \boldsymbol{\Sigma}_{nf})$ and $(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$ are the means and covariance matrices of the normal operating data and faulty data, respectively.

III. THE KL-DIVERGENCE MAXIMIZED ANALYSIS APPROACH FOR FAULT DETECTION

A. MAXIMIZED DIVERGENCE ANALYSIS

A large value of KL-divergence associates with the fault is easy to detect, whereas a value close to zero corresponds to the fault being hard to detect. Therefore,

maximizing the KL-divergence is beneficial for detecting faults of industrial processes. In other words, a method that can generate large values of divergence between normal operating data and faulty data during dimensionality reduction steps, possess good performance of fault detection.

To the end, we propose a fault detection approach termed as *Maximized Divergence Analysis* (MDA), which has the objective of maximizing KL-divergence during the dimensionality reduction step:

$$\max_{\boldsymbol{w}} D(p_{nf}\,||\,p_f) = \int p_{nf}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) \ln \frac{p_{nf}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})}{p_f(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})} d\boldsymbol{x}, \tag{3}$$

where $\boldsymbol{x} \in R^m$ is the vector consisting of m variables, $\boldsymbol{w} \in R^{m\times k}$ denotes the loading matrix, and k is the reduction order with $k \leq m$.

If p_{nf} and p_f follow multivariable Gaussian distributions, the objective of (3) can also be explicitly expressed as (2) by replacing $\boldsymbol{\Sigma}_f$ with $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_f\boldsymbol{w}$, $\boldsymbol{\Sigma}_{nf}$ with $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_{nf}\boldsymbol{w}$, $\boldsymbol{\mu}_{nf}$ with $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_{nf}$, and $\boldsymbol{\mu}_f$ with $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_f$.

The solution of the optimization (3) for the context when p_{nf} and p_f follow multivariable Gaussian distributions is investigated below. Since $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$ is positive definite, there exists a nonsingular transformation $\boldsymbol{Q} \in R^{k\times k}$ such that $(\boldsymbol{wQ})^{\mathrm{T}}\boldsymbol{wQ} = \boldsymbol{I}_k$, and because the invariance property of KL-divergence for nonsingular transformations [23], i.e., $D_{\boldsymbol{w}} = D_{\boldsymbol{wQ}}$, it suffices to consider only those matrices $\boldsymbol{w}$ belonging to the set $\Omega := \left\{\boldsymbol{w}|\,\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = \boldsymbol{I}_k\right\}$. Moreover, given any $\boldsymbol{w} \in \Omega$, it is possible to construct a matrix $\boldsymbol{v} \in R^{m\times(m-k)}$ satisfying $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{I}_{m-k}$, and $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v} = \boldsymbol{0}$, and such that the matrix

$$\boldsymbol{P} := \begin{bmatrix} \boldsymbol{w} & \boldsymbol{v} \end{bmatrix}, \tag{4}$$

satisfies $\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P} = \boldsymbol{I}_m$ (i.e., $\boldsymbol{P}$ is an orthonormal matrix). It follows that for any $\boldsymbol{w} \in \Omega$, it can be obtained that

$$\boldsymbol{w} = \boldsymbol{P}\begin{bmatrix} \boldsymbol{I}_k \\ \boldsymbol{0} \end{bmatrix}, \tag{5}$$

and the solution to the loading scores problem amounts to optimally rotating the original coordinates of the spectral measurement space (i.e., $\boldsymbol{x} \rightarrow \boldsymbol{P}^{\mathrm{T}}\boldsymbol{x}$) and then selecting the first k components of the resulting vectors.

From (2), the divergence $D(p_{nf}\,||\,p_f)$ in the space of optimal rotation by $\boldsymbol{P}$ can be expressed as

$$\begin{aligned} &D(p_{nf}\,||\,p_f) \\ &= \frac{1}{2}\ln\left|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\Sigma}_{nf}^{-1}\boldsymbol{\Sigma}_f\boldsymbol{P}\right| + \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\Sigma}_{nf}\boldsymbol{P}\right\} - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}\right\} \\ &\quad + \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\Sigma}_f^{-1}\left(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f\right)\left(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f\right)^{\mathrm{T}}\boldsymbol{P}\right\}. \end{aligned} \tag{6}$$

Proposition 1: Let $\boldsymbol{\mu}_{nf}$, $\boldsymbol{\mu}_f$, $\boldsymbol{\Sigma}_{nf}$, and $\boldsymbol{\Sigma}_f$ be defined above, and $\boldsymbol{\Sigma}_{nf}$ and $\boldsymbol{\Sigma}_f$ are positive definite. Let $\boldsymbol{P}_i \in R^{m\times 1}$ $(i = 1, 2, \ldots, m)$ be defined as the ith column (loading vector) of the optimal rotation matrix $\boldsymbol{P}$ in (4) or (6). Then the loading vector $\boldsymbol{P}_i$ can be obtained by solving the below equation:

$$\left[\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\Sigma}_{nf} + \boldsymbol{\Sigma}_f^{-1}\left(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f\right)\left(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f\right)^{\mathrm{T}}\right]\boldsymbol{P}_i = \lambda_i\boldsymbol{P}_i, \tag{7}$$

where the eigenvalues λ_i are the divergence degree by projecting the data onto $\boldsymbol{P}_i$.

Proof: From (6), the objective of the problem for optimally rotating the original coordinates of the spectral measurement space can be rewritten as

$$\begin{aligned} \max_{\boldsymbol{P}_i}\ & D(p_{nf}\,||\,p_f) = \frac{1}{2}\ln\left(\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{nf}^{-1}\boldsymbol{\Sigma}_f\boldsymbol{P}_i\right) + \frac{1}{2}\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\Sigma}_{nf}\boldsymbol{P}_i \\ &- \frac{1}{2}\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{P}_i + \frac{1}{2}\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{\Sigma}_f^{-1}(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f)(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f)^{\mathrm{T}}\boldsymbol{P}_i, \\ \text{Subject to } & \boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{P}_i = 1, \quad \text{and } \boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{P}_j = 0 \text{ for } 1 \leq j < i. \end{aligned} \tag{8}$$

The technique of Lagrange multipliers is employed and an auxiliary function is introduced as

$$\begin{aligned} &L(\boldsymbol{P}_i, \lambda_i, \mu_j) \\ &:= \frac{1}{2}\ln\left(\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{nf}^{-1}\boldsymbol{\Sigma}_f\boldsymbol{P}_i\right) + \frac{1}{2}\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\Sigma}_{nf}\boldsymbol{P}_i - \frac{1}{2}\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{P}_i \\ &\quad + \frac{1}{2}\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{\Sigma}_f^{-1}(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f)(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f)^{\mathrm{T}}\boldsymbol{P}_i \\ &\quad - \frac{\lambda_i}{2}(\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{P}_i - 1) - \sum_{j=1}^{i-1}\mu_j\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{P}_j. \end{aligned} \tag{9}$$

Taking the derivative of (9) with respect to $\boldsymbol{P}_i$, i.e.,

$$\frac{\partial}{\partial\boldsymbol{P}_i}L(\boldsymbol{P}_i, \lambda_i, \mu_j) = \boldsymbol{0}, \tag{10}$$

results in

$$\begin{aligned} &\boldsymbol{\Sigma}_{nf}^{-1}\boldsymbol{\Sigma}_f\boldsymbol{P}_i\left(\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{nf}^{-1}\boldsymbol{\Sigma}_f\boldsymbol{P}_i\right)^{-1} + \boldsymbol{\Sigma}_f^{-1}\boldsymbol{\Sigma}_{nf}\boldsymbol{P}_i - \boldsymbol{P}_i \\ &\quad + \boldsymbol{\Sigma}_f^{-1}\left(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f\right)\left(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f\right)^{\mathrm{T}}\boldsymbol{P}_i - \lambda_i\boldsymbol{P}_i \\ &\quad - \sum_{j=1}^{i-1}\mu_j\boldsymbol{P}_j = \boldsymbol{0}. \end{aligned} \tag{11}$$

Multiplying the left side of (11) by $\boldsymbol{P}_i^{\mathrm{T}}$, it follows that

$$\begin{aligned} &\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\Sigma}_{nf}\boldsymbol{P}_i + \boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{\Sigma}_f^{-1}\left(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f\right)\left(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f\right)^{\mathrm{T}} \\ &\quad \times\boldsymbol{P}_i - \lambda_i\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{P}_i = 0, \end{aligned} \tag{12}$$

which is satisfied if the loading vector $\boldsymbol{P}_i$ meets the condition:

$$\left[\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\Sigma}_{nf} + \boldsymbol{\Sigma}_f^{-1}\left(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f\right)\left(\boldsymbol{\mu}_{nf}-\boldsymbol{\mu}_f\right)^{\mathrm{T}}\right]\boldsymbol{P}_i - \lambda_i\boldsymbol{P}_i = \boldsymbol{0}. \tag{13}$$

In addition, taking the derivative of (9) with respect to λ_i and μ_j (i.e., $\frac{\partial}{\partial\lambda_i}L(\boldsymbol{P}_i, \lambda_i, \mu_j) = 0$ and $\frac{\partial}{\partial\mu_j}L(\boldsymbol{P}_i, \lambda_i, \mu_j) = 0$) give $\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{P}_i = 1$ and $\boldsymbol{P}_i^{\mathrm{T}}\boldsymbol{P}_j = 0$, respectively; and together with (13), it is equivalent to solve the eigenvalue problem of (7). The eigenvalue λ_i implies the divergence degree by projecting the data onto $\boldsymbol{P}_i$. □

Thus, the orthonormal loadings $\boldsymbol{w} \in R^{m\times k}$ in (3) that maximizes the value of KL-divergence can be selected as the first k components (i.e., equation (5)) of the optimal rotation matrix $\boldsymbol{P}$ obtained according to Proposition 1, and its associated loadings can be calculated as $\boldsymbol{d} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$. The reduction order k for MDA can be determined by cross-validation or the Akaike information criterion [1].

Remark 1: The proposed method essentially treats the fault detection problem as a binary classification problem (normal vs faulty), any type of faulty data from the historian can be useful for training our MDA algorithm. In practice, we often have some available faulty data from the historian. Our proposed method will utilize this valuable resource to improve the fault detection performance. Such faulty data are not well explored by the PCA, PLS, or CVA methods, which are unsupervised requiring only normal data.

Remark 2: The Proposition 1 assumes that the density functions p_{nf} and p_f are Gaussian, which may not be valid in practice if the underlying process is nonlinear. However, the proposed MDA framework is still applicable for such cases. For non-Gaussian but known p_{nf} and p_f, the formulation of (3) still has a closed-form expression where nonlinear optimization techniques may be needed to solve it. In the worst case where the expressions of p_{nf} and p_f are unknown, one can use density function estimation techniques, such as the kernel density estimation, to obtain estimated probability functions p_{nf} and p_f based on faulty and normal datasets. Thus, the proposed MDA framework can be easily adapted to different scenarios with known or unknown non-Gaussian p_{nf} and p_f.

B. MDA-BASED STATISTICS FOR FAULT DETECTION

Similar to the detection statistics used in the PCA- and CVA-based monitoring method, we follow the similar manner to construct the detection statistics (i.e., T^2 and Q) for the proposed MDA-based method. Specifically, the statistics T^2 and Q, which are respectively constructed on the retained loading scores and discarded loading scores in the MDA model MDA, are defined as

$$T^2 = \boldsymbol{d}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{d}}^{-1}\boldsymbol{d}, \tag{14}$$

and

$$Q = \boldsymbol{r}^{\mathrm{T}}\boldsymbol{r}, \tag{15}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{d}}$ is the covariance matrix of $\boldsymbol{d}$, and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})\boldsymbol{x}$.

When the detection statistics $T^2$and Q are below the thresholds, namely, $T^2 \le T_\alpha^2$ and $Q \le Q_\alpha$, where T_α^2 and Q_α are the upper control limits for the metrics of $T^2$and Q with a significance level α, the process operations are regarded as normal. Otherwise, a fault occurs in the process operations. The upper control limits T_α^2 and $Q_\alpha$is calculated using the empirical approach based on the calibration samples under NOC [2]. For instance, a 99% confidence upper control limit can be derived as the T^2 value below which 99% of the calibration data are located.

Emerging data-driven fault detection methods can be classified into unsupervised (dimension reduction) approaches, e.g., PCA, PLS, and CVA, and supervised approaches, e.g., random forest, decision tree, and SVM [34]. The unsupervised techniques often optimize some objectives that are not directly related to maximizing the separability between normal and faulty classes, in contrast to our approach. The supervised approaches generally rely on statistical machine learning without dimension reduction. Thus, in the presence of high dimensions as often met in practice, such methods may require more data for training. Deep learning approaches, both supervised and unsupervised (e.g., autoencoder and deep neural networks), have shown prominent advantages for enhancing the performance. However, they often require a large amount of training data, which are expensive to gather in practice. Compared with existing approaches, our method based on dimension reduction does not require much training data. At the same time, it is supervised where any faulty data from the historian can be utilized to formulate the fault class. In particular, it does not require the labelling of the specific fault type for each faulty dataset. Thereby, it can fully uncover the values of past faulty data for promoting the fault detection.

IV. CASE STUDIES FOR SYNTHETIC DATASETS AND THE TENNESSEE EASTMAN PROCESS

A. TWO-DIMENSIONAL SYNTHETIC DATASETS

In this subsection, a comparison of the one-dimensional (1D) projections of PCA, FDA, and MDA for two-dimensional (2D) synthetic datasets is conducted. The 2D datasets are produced by

$$\text{Normal operating data:} \quad \begin{cases} y_1 \sim N\,(1, 0.25) \\ y_2 = y_1 + v, \end{cases} \tag{16}$$

and

$$\text{Faulty data:} \quad \begin{cases} y_1 \sim N\,(1, 1) \\ y_2 = -y_1 + 2 + w, \end{cases} \tag{17}$$

where v and w are white noise sequences that follow $v \sim N\,(1, 0.05)$ and $w \sim N\,(1, 0.2)$, respectively. Both the normal and faulty data had 100 samples.

In the 1D projections obtained by PCA, FDA, and MDA in Figure 1 indicate that MDA provided the best data separation for the two classes, followed by FDA, and then PCA. A visual inspection of Figure 1 indicates that the majority of the faulty data points lie outside of any reasonably defined T^2threshold for MDA, whereas all of the faulty data lie within the T^2 threshold for PCA and about half of the faulty data lie within the T^2threshold for FDA. The contrasting performance are mainly because that (*i*) the information on faulty data is not optimally used in PCA; and (*ii*) FDA is not so effective for discriminating the classes that share the same mean [27]. In contrast, MDA can perform well for such classes.

Remark 3: As shown by this example, the objectives of PCA and FDA (or PLS) in dimension reduction are not

TABLE 1. Faults used in case study 1.

ID	Fault description	Type
Fault 3	D Feed Temperature (Stream 2)	Step change
Fault 9	D Feed Temperature (Stream 2)	Random variation
Fault 11	Reactor cooling water inlet temperature	Random variation
Fault 15	Condenser cooling water Valve	Sticking
Fault 19	Unknown	

directly related to the separability between faulty and normal classes, as reflected by the distribution of data points of the two classes in the latent space. In contrast, the proposed MDA approach can better distinguish the classes in the latent space. For our method, by maximizing the KL-divergence, the resultant loading vectors can project the faulty and normal data into the low-dimensional space with the largest separability.

B. TENNESSEE EASTMAN PROCESS

Here MDA is compared to PCA- and FDA-based methods for the Tennessee Eastman process (TEP), which is a benchmark widely used for the comparison of various fault detection methods [2]. The TEP was created based on a simulation of a real industrial process, a detailed description of which can be found in [2].

Twenty-one fault datasets were produced using the preprogrammed faults (Faults 1–21). In addition, a normal operating dataset was produced under NOC (with no faults). For each fault, three data sets (training, validation, and testing data) are produced. The training and validation data are for model building and the testing data are for model testing. The training and validation datasets contain 400 observations. The testing dataset for each fault contains 960 observations, where each data set starts with normal operations and after the 160th observation, faults are triggered. Each data sample consists of all manipulated and measurement variables except the agitation speed of the reactor's stirrer, i.e., 52 process variables. The KL-divergence for the MDA-based method was calculated by approximating data as Gaussian distributions.

1) CASE STUDY 1: FAULTS 3, 9, 11, 15, AND 19

In this case study, the performance of MDA-, PCA-, and FDA-based methods for detecting faults is examined for Faults 3, 9, 11, 15, and 19 generated by the TEP (see Table 1). These five faults were selected as being among the most difficult to detect [2], [28]. To make a fair comparison of different detection methods, we compare the fault detection rates (FDR) of detection algorithms while maintaining the false alarm rates in the same level of $\alpha = 1\%$ in this work. The reduction order for PCA is selected as k=11 according to [29], and for FDA is selected as k=1 since FDA can only extract one loading vector for binary classification. The reduction order for MDA determined by cross-validation is depicted in Table 2.

The FDR produced by the three detection approaches are provided in Table 3, where the overall FDR for PCA and FDA for the five faults was only 16.2% and 22.3%, respectively.

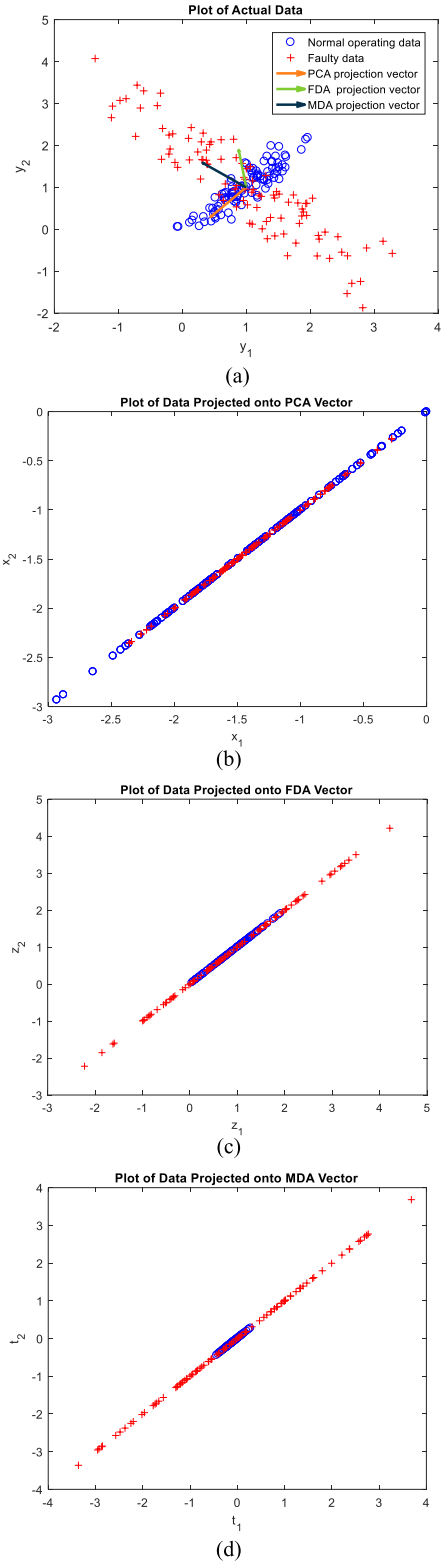

FIGURE 1. A comparison of the 1D projections of PCA, FDA, and MDA for the 2D datasets ('o' denotes normal operating data, and '+' denotes faulty data): (a) actual data, (b) data projected onto the PCA vector, (c) data projected onto the FDA vector, and (d) data projected onto the MDA vector.

MDA outperformed PCA and FDA in detecting all five faults, and had an overall FDR a factor of 2.7 and 2.0 higher than

TABLE 2. The reduction order of MDA model for case study 1 in TEP.

	Fault 3	Fault 9	Fault 11	Fault 15	Fault 19
Reduction order k	25	26	17	26	22

TABLE 3. The FDR for PCA, FDA, and MDA for Faults 3, 9, 11, 15, and 19 (For each method, a fault was indicated if either the T^2 or Q statistics violated threshold).

FDR (%)	PCA	FDA	MDA
Fault 3	1.7	3.2	6.5
Fault 9	0.8	2.4	5.4
Fault 11	55.5	66.5	76.3
Fault 15	15.0	27.5	39.9
Fault 19	7.8	11.6	92.0
Overall FDR	16.2	22.3	44.1

PCA and FDA, respectively. The FDR for MDA is a factor of 3.8, 6.8, 1.4, 2.7, and 11.8 higher for Faults 3, 9, 11, 15, and 19 in comparison with PCA. Compared to FDA, the fault detection rate for MDA is a factor of 2.0, 2.3, 1.1, 1.5, and 7.9 higher for Faults 3, 9, 11, 15, and 19, respectively.

None of the methods obtained a high FDR for Faults 3 and 9. Given that MDA maximizes the distance between the distributions for the normal and fault data, the main cause for the bad performance in detection of Faults 3 and 9 reported in [29] is from the property of faults (i.e., the faults are inherently hard to be detect/separate). However, our method still shows significant improvement compared with PCA and FDA (e.g., for Fault 3, our method gives 6.5% fault detection rate, against 3.2% for FDA and 1.7% for PCA). Such improvement is dramatic given the inherent difficulty of detecting these two faults. For Faults 11, 15, and 19, the FDRs for the MDA-based approach increase considerably compared with the PCA-, and FDA-based methods. For Fault 19, MDA had about an order of magnitude higher successful detection rate than PCA and FDA. The extent to which the methods are sensitive to Fault 19 is shown in Figure 2. The MDA-based approach much more persistently indicates a fault in comparison with the other methods. The differing results indicate that the poor fault detection performance for Faults 11, 15, and 19 was from limitations of the PCA- and FDA-based methods rather than because the normal and fault data are inseparable. PCA and FDA do not determine projections that optimally separate the data to detect the faults, and significantly improved fault detection is obtained by using optimal separation, i.e., MDA.

2) CASE STUDY 2: ALL 21 FAULTS

All 21 faults from the TEP simulator are used to further verify the effectiveness of the proposed MDA-based approach in comparison with the PCA- and FDA-based approaches for fault detection (see Figure 3). The MDA-based method outperformed both the PCA- and CVA-based methods for each of the 21 faults. It confirms that the superior fault detection obtained by the objective of directly maximizing the divergence between normal operating data and faulty data during dimensionality reduction. For example, the FDR for

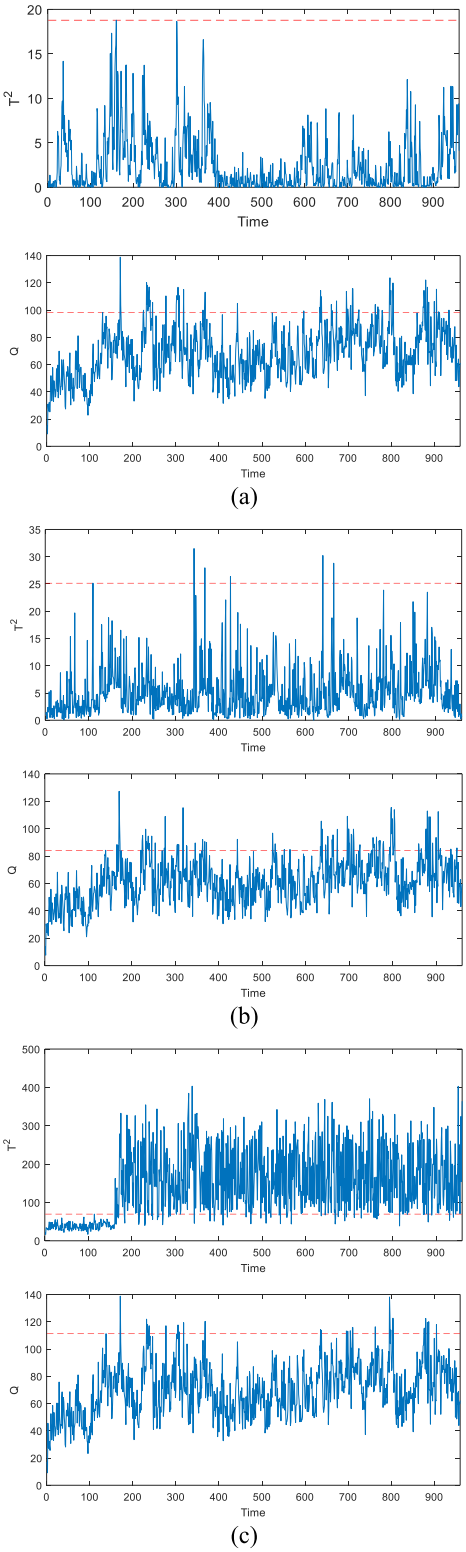

FIGURE 2. Detection results for Fault 19 for (a) PCA-based, (b) FDA-based, and (c) MDA-based approaches.

the MDA detection method for Fault 10 is 73.2% in contrast to 57.1% to the PCA-based method, which is more than a factor of 1.3 improved detection. The FDR for the MDA detection approach is factor of 1.3 better than that for the FDA detection method (73.2% and 57.9%).

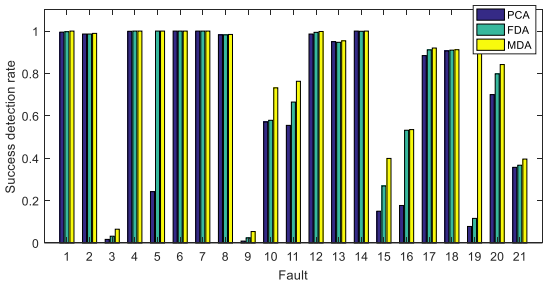

FIGURE 3. Fault detection rates of the PCA-, FDA-, and MDA detection approaches for the 21 faults.

TABLE 4. Summary of detection results for PCA, MDA, and FDA for all 21 faults.

Fault detection rate (%)	PCA	FDA	MDA
Overall fault detection rate	64.5	70.9	79.1

Table 4 displays the overall FDRs for all faults using the PCA-, FDA-, and MDA-based approaches. The FDR of the MDA method is 14.6% and 8.2% higher than the PCA and FDA approaches respectively, which further verify the better performance of the proposed MDA-based detection method.

In addition, Table 5 displays the detection delays for the PCA-, FDA-, and MDA-based methods, in which the detection delay is recorded when the first three consecutive values for the statistics (T^2or Q) have violated the threshold. For nearly all of the faults, the MDA-based method had the smallest detection delay than the PCA- and FDA-based methods. By examining Figure 3 and Table 5, it can be concluded that the fault detection approach with the highest FDRs usually have the shortest detection delays.

Remark 4: For the example on TEP, for Case Study 1, we used the data from each of the Faults 3, 9, 11, 15, 18 as the faulty data. For each fault, we trained the MDA algorithm and used it for detection. Similarly, for Case Study 2, we used each of the 21 classes of faulty data for training the MDA algorithm and then testing the fault detection performance. Overall, the size of the faulty data in our case studies is balanced with the normal data for each training. It is thus an interesting future topic to study the impact of overly small faulty data size on the performance of the proposed MDA algorithm.

Remark 5: Note that for our two case studies on the TEP, some faults such as Faults 3, 9, 13, and 15 are incipient faults with slight features [35]. These respective faults are: step change of D feed temperature, random variation of D feed temperature, slow drift in the reaction kinetics, and sticking fault of condenser cooling water valve. All these faults are small in amplitudes and are thus not easy to be detected [35]. However, the results from the two case studies above have shown the advantageous performance of our MDA algorithm against the PCA and FDA. Specifically, for Faults 3 and 9, the MDA gives 6.5% and 5.4% fault detection rate (Table 3), much higher that 1.7% and 0.8% for the PCA method, and 3.2% and 2.4% for the FDA method. For Faults 13 and 15, as shown in Figure 3, the fault detection rate of MDA is also higher than that from FDA and PCA. These results together show that the proposed approach is also effective in fault detection if we only have normal and incipient faulty data for training.

TABLE 5. Summary of detection delays for PCA, FDA, and MDA for all 21 faults.

Fault ID	PCA	FDA	MDA
1	7	5	3
2	14	14	14
3	713	88	88
4	3	3	3
5	5	3	3
6	3	3	3
7	3	3	3
8	17	15	14
9	765	154	138
10	23	17	17
11	13	13	8
12	9	4	5
13	41	47	40
14	3	4	3
15	580	94	94
16	291	18	22
17	27	27	27
18	87	60	57
19	74	13	13
20	81	81	77
21	490	490	478

V. CONCLUSION

This article presents a fault detection approach based on maximized divergence analysis (MDA). MDA generates lower dimensional representations of the original data that maximize the retained KL-divergence between normal operating data and faulty data during the step of dimensionality reduction. Two detection statistics, i.e., T^2 and Q, are put forward using the retained and discarded loading scores in the MDA model. The proposed MDA-based approach outperforms the PCA- and FDA-based approaches for 2D synthetic datasets and for the 21 faults in the Tennessee Eastman process, particularly for faults that are difficult to detect. To further enhance the detection performance on data with nonlinear dynamics, MDA can be extended by utilizing kernel techniques.

ACKNOWLEDGMENT

The authors acknowledge Prof. Richard D. Braatz at the Massachusetts Institute of Technology for helpful discussions.

REFERENCES

[1] S. Tang, S. Yuan, and Y. Zhu, "Convolutional neural network in intelligent fault diagnosis toward rotatory machinery," *IEEE Access*, vol. 8, pp. 86510–86519, 2020.
[2] L. H. Chiang, E. L. Russell, and R. D. Braatz, *Fault Detection and Diagnosis in Industrial Systems*. London, U.K.: Springer, 2001.
[3] H. Zheng, R. Wang, Y. Yang, J. Yin, Y. Li, and M. Xu, "Cross-domain fault diagnosis using knowledge transfer strategy: A review," *IEEE Access*, vol. 7, pp. 129260–129290, 2019.
[4] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly accurate machine fault diagnosis using deep transfer learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2446–2455, Apr. 2018.
[5] R. K. Perrons, "Assessing the damage caused by *Deepwater Horizon*: Not just another *Exxon Valdez*," *Mar. Pollut. Bull.*, vol. 71, nos. 1–2, pp. 20–22, Jun. 2013.

[6] Q. Jiang, X. Yan, and B. Huang, "Review and perspectives of data-driven distributed monitoring for industrial plant-wide processes," *Ind. Eng. Chem. Res.*, vol. 58, no. 29, pp. 12899–12912, Jul. 2019.
[7] Y. Li, X. Wang, Z. Liu, X. Liang, and S. Si, "The entropy algorithm and its variants in the fault diagnosis of rotating machinery: A review," *IEEE Access*, vol. 6, pp. 66723–66741, 2018.
[8] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 136–144, Jan. 2017.
[9] Y. Dong and S. J. Qin, "A novel dynamic PCA algorithm for dynamic data modeling and process monitoring," *J. Process Control*, vol. 67, pp. 1–11, Jul. 2018.
[10] S. Gajjar, M. Kulahci, and A. Palazoglu, "Real-time fault detection and diagnosis using sparse principal component analysis," *J. Process Control*, vol. 67, pp. 112–128, Jul. 2018.
[11] C. Zhang, X. Gao, Y. Li, and L. Feng, "Fault detection strategy based on weighted distance of k nearest neighbors for semiconductor manufacturing processes," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 1, pp. 75–81, Feb. 2019.
[12] F. Tahir, M. T. Islam, J. Mack, J. Robertson, and D. Lovett, "Process monitoring and fault detection on a hot-melt extrusion process using in-line Raman spectroscopy and a hybrid soft sensor," *Comput. Chem. Eng.*, vol. 125, pp. 400–414, Jun. 2019.
[13] Z. Chen, S. X. Ding, T. Peng, C. Yang, and W. Gui, "Fault detection for non-Gaussian processes using generalized canonical correlation analysis and randomized algorithms," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1559–1567, Feb. 2018.
[14] G. Li, S. J. Qin, and D. Zhou, "Geometric properties of partial least squares for process monitoring," *Automatica*, vol. 46, no. 1, pp. 204–210, Jan. 2010.
[15] B. Jiang and R. D. Braatz, "Fault detection of process correlation structure using canonical variate analysis-based correlation features," *J. Process Control*, vol. 58, pp. 131–138, Oct. 2017.
[16] B. Jiang, D. Huang, X. Zhu, F. Yang, and R. D. Braatz, "Canonical variate analysis-based contributions for fault identification," *J. Process Control*, vol. 26, pp. 17–25, Feb. 2015.
[17] K. E. S. Pilario and Y. Cao, "Canonical variate dissimilarity analysis for process incipient fault detection," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5308–5315, Dec. 2018.
[18] P. Van Overschee and B. De Moor, "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems," *Automatica*, vol. 30, no. 1, pp. 75–93, Jan. 1994.
[19] B. Jiang, X. Zhu, D. Huang, and R. D. Braatz, "Canonical variate analysis-based monitoring of process correlation structure using causal feature representation," *J. Process Control*, vol. 32, pp. 109–116, Aug. 2015.
[20] B. Jiang, X. Zhu, D. Huang, J. A. Paulson, and R. D. Braatz, "A combined canonical variate analysis and Fisher discriminant analysis (CVA–FDA) approach for fault diagnosis," *Comput. Chem. Eng.*, vol. 77, pp. 1–9, Jun. 2015.
[21] D. Eriksson, E. Frisk, and M. Krysander, "A method for quantitative fault diagnosability analysis of stochastic linear descriptor models," *Automatica*, vol. 49, no. 6, pp. 1591–1600, Mar. 2013.
[22] F. De la Torre and T. Kanade, "Multimodal oriented discriminant analysis," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, Bonn, Germany, Mar. 2005, pp. 177–184.
[23] S. Kullback, *Information Theory and Statistics*. New York, NY, USA: Wiley, 1959.
[24] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Estimation of KL divergence: Optimal minimax rate," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2648–2674, Apr. 2018.
[25] L. Feng, H. Wang, B. Jin, H. Li, M. Xue, and L. Wang, "Learning a distance metric by balancing KL-divergence for imbalanced datasets," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 12, pp. 2384–2395, Dec. 2019.
[26] L. H. Chiang and R. D. Braatz, "Process monitoring using causal map and multivariate statistics: Fault detection and identification," *Chemometrics Intell. Lab. Syst.*, vol. 65, no. 2, pp. 159–178, 2003.
[27] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA, USA: Academic, 1990.
[28] L. H. Chiang, B. Jiang, X. Zhu, D. Huang, and R. D. Braatz, "Diagnosis of multiple and unknown faults using the causal map and multivariate statistics," *J. Process Control*, vol. 28, pp. 27–39, Apr. 2015.
[29] E. L. Russell, L. H. Chiang, and R. D. Braatz, "Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 51, no. 1, pp. 81–93, May 2000.
[30] H. Chen, B. Jiang, and N. Lu, "An improved incipient fault detection method based on Kullback–Leibler divergence," *ISA Trans.*, vol. 79, pp. 127–136, Aug. 2018.
[31] H. Chen, B. Jiang, N. Lu, and W. Chen, "Real-time incipient fault detection for electrical traction systems of CRH2," *Neurocomputing*, vol. 306, pp. 119–129, Sep. 2018.
[32] J. Harmouche, C. Delpha, and D. Diallo, "Incipient fault detection and diagnosis based on Kullback–Leibler divergence using principal component analysis: Part I," *Signal Process.*, vol. 94, pp. 278–287, Jan. 2014.
[33] W. Bounoua, A. B. Benkara, A. Kouadri, and A. Bakdi, "Online monitoring scheme using principal component analysis through Kullback–Leibler divergence analysis technique for fault detection," *Trans. Inst. Meas. Control*, vol. 42, no. 6, pp. 1225–1238, Apr. 2020.
[34] S. A. A. Taqvi, H. Zabiri, L. D. Tufa, F. Uddin, S. A. Fatima, and A. S. Maulud, "A review on data-driven learning approaches for fault detection and diagnosis in chemical processes," *ChemBioEng Rev.*, vol. 8, no. 3, pp. 239–259, Jun. 2021.
[35] W. S. Ge, J. Wang, J. L. Zhou, H. Wu, and Q. B. Jin, "Incipient fault detection based on fault extraction and residual evaluation," *Ind. Eng. Chem. Res.*, vol. 54, no. 14, pp. 3664–3677, 2015.

BENBEN JIANG (Member, IEEE) received the B.Eng. degree in automation from Zhejiang University, China, in 2010, and the Ph.D. degree in control science and engineering from Tsinghua University, China, in 2015. He was a Postdoctoral Associate at the Massachusetts Institute of Technology (MIT), from 2016 to 2020. He is an Assistant Professor with the Department of Automation, Tsinghua University. His research interests include computational energy intelligence, machine learning with application to advanced battery systems, and fault diagnosis for manufacturing processes. He was a recipient of the 2015 Outstanding Doctoral Dissertation from the Department of Automation, Tsinghua University, for being a Promising Young Researcher, and the 2016 Prize in Informatics/Computer Science from the Dimitris Chorafas Foundation, Switzerland.

QIUGANG LU received the B.Eng. degree in control science and engineering from the Harbin Institute of Technology, China, in 2011, and the Ph.D. degree in chemical and biological engineering from the University of British Columbia, Canada, in 2018. He was a Prognostics Engineer at the General Motors of Canada, Toronto, from 2018 to 2019, and a Postdoctoral Research Associate at the University of Wisconsin–Madison, from 2019 to 2020. He is an Assistant Professor with the Department of Chemical Engineering, Texas Tech University. He has published over 30 articles in renowned journal and conferences. His research interests include advanced process control, data analytics, system identification, and process monitoring. He was a recipient of the Vanier Canada Graduate Scholarship, in 2015, the Best Presenter Award at PACWEST Conference, in 2015, and the Certificate of Service Award from the Journal of the Franklin Institute, in 2019.