Fault Diagnosis in Industrial Processes by Maximizing Pairwise Kullback–Leibler Divergence

Qiugang Lu, *Member, IEEE*, Benben Jiang, *Member, IEEE*, and Eranda Harinath

***Abstract*— Fault diagnosis gains increasing attention for its ability to enhance process safety and efficiency. This brief proposes a maximized ratio divergence analysis (MRDA) approach for fault diagnosis, which maximizes the pairwise ratio Kullback–Leibler (KL) divergence between each pair of classes during the dimensionality reduction step. In addition, an iterative algorithm based on deflation techniques is put forward for learning the loading vectors of MRDA. The proposed MRDA-based approach allows for improved power of fault diagnosis because of the following advantages over classical monitoring methods. First, MRDA maximizes the pairwise ratio divergence between each pair of classes, which directly leads to enhanced classification performance in the low-dimensional space. Moreover, MRDA is less likely to be dominated by "outlier" classes, since its objective is an average of ratio divergence, thereby facilitating the proposed method to be beneficial to the classification of imbalanced faulty classes. The effectiveness of the MRDA-based approach for fault diagnosis is verified by the Tennessee Eastman process (TEP).**

***Index Terms*— Data-driven method, fault diagnosis, Kullback–Leibler (KL) divergence, process monitoring, Tennessee Eastman Process (TEP).**

I. Introduction

FAULT diagnosis, which determines the type and cause of faults, can be rather challenging for modern industrial processes featured by a large number of process variables and complicated correlations among variables due to process dynamics and controllers [1]. With the availability of enormous process data collected from computer control systems, data-driven fault diagnosis has shown its exceptional value in promoting informed decision-making and enhancing the efficient and safe operation of industrial processes [2]–[4].

The performance of data-driven fault diagnosis can be improved by using dimension reduction methods. The most widely applied linear methods for fault diagnosis include principal component analysis (PCA) [5], [6], partial least square (PLS) analysis [7], and Fisher discriminant analysis (FDA) [8], [9]. A multivariate statistics approach that integrates PCA with discriminant analysis was developed for diagnosing anomalies [10], [11]. Alternately, a PLS method was proposed in conjunction with the discriminant algorithm for fault diagnosis [12]. The dimension reduction method based on the FDA technique has been extensively studied for diagnosing abnormal events [8], [9], [13], [14]. The objective of FDA is to extract a group of projection vectors that maximize the scatter between different classes while minimizing the scatter within each class. However, FDA only possesses optimality for the classification problem with equal covariance matrices for different classes [14], [15]. Classical nonlinear process-monitoring techniques include manifold learning-based and deep learning-based methods. Manifold learning methods discover and map data from the original high-dimensional space to the local low-dimensional manifolds [30]. Deep learning-based approaches employ deep neural networks to extract complex and nonlinear structures in the data for fault diagnosis, in which various autoencoder-based methods receive extensive attention for capturing nonlinear and structured manifold embedding [31]. Excellent reviews on data-driven fault diagnosis methods are available in [1], [4], [16], [17], and [25]–[27].

Recently, methods based on Kullback–Leibler (KL) divergence, an effective way to measure the difference between the probability density functions [18], have attracted the interest of many researchers and practitioners [28]. A large value of KL-divergence indicates that the distributions of two faulty data sets are well separated, thereby indicating that the faults are easy to diagnose. A value close to zero indicates that the two probability density functions are similar, i.e., the discrimination of one fault from another is difficult. KL divergence is used to measure the dissimilarity between the probability distributions of the faulty data collected from induction motor systems in [19], where enhanced diagnosis performance was observed. In addition, Zeng *et al.* [20] developed monitoring statistics based on KL divergence to diagnose large-scale processes, where KL divergence-based statistics were shown to be more sensitive than the conventional multivariate statistics. More recently, the authors developed a fault diagnosis method based on the KL-divergence for diagnosing multiple and unknown faults [3]. For all these contributions, KL divergence is mainly used as a measure of the difference between the underlying distributions of faulty data with that of the prescribed faulty classes. However, no dimensionality reduction, which plays a critical role in process monitoring, was involved in implementing KL divergence to fault

Manuscript received November 21, 2018; revised April 22, 2019 and July 29, 2019; accepted September 30, 2019. Date of publication November 22, 2019; date of current version February 9, 2021. Manuscript received in final form October 27, 2019. This work was supported by the National Natural Science Foundation of China under Grant 61603024. Recommended by Associate Editor A. Serrani. *(Corresponding author: Benben Jiang.)*

Q. Lu is with the General Motors of Canada Company, Markham, ON L3R 4H8, Canada (e-mail: qiuganglu@gmail.com).

B. Jiang is with the Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: bbjiang@mit.edu).

E. Harinath is with the Global Manufacturing Science and Technology, Sanofi Genzyme, Framingham, MA 01701 USA (e-mail: erandah@gmail.com).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCST.2019.2950403

diagnosis. For modern industrial processes featured by high dimensions, dimensionality reduction in general acts as a step of feature extraction to enhance the performance of fault diagnosis. Its advantages mainly arise when the dimension of the observation space is large while the number of samples is small, which is common in modern industrial processes. Under this circumstance, the statistical parameters of the process variables such as the mean and the covariance matrix are subject to large inaccuracies that often lead to adverse effects for statistical inference including fault classification. Moreover, the existence of correlation or even collinearity between variables is very common and this also motivates the dimensionality reduction to extract representative features prior to fault detection or diagnosis [1]. For fault diagnosis, FDA is widely used but it yields degraded performance when distributions of fault classes are imbalanced, e.g., an outlier class exists that is far from the rest of the classes. Moreover, FDA is only applicable when all fault classes share the same covariance matrix. To address these issues, in this brief, a fault diagnosis approach named maximized ratio divergence analysis (MRDA) is proposed, in which the objective is to maximize the pairwise ratio KL-divergence between each pair of classes during the dimensionality reduction procedure. The main contributions of this work are summarized as follows.

1) We propose an MRDA approach for dimensionality reduction for fault diagnosis, and this approach uses ratio KL-divergence to address imbalanced classification in which some fault classes are far away from the rest.
2) We provide an algorithm using the deflation technique to solve the optimization problem of maximizing the average ratio KL-divergence between classes.
3) We further present an extension of our method to the dynamic MRDA approach to handle serial correlations among data samples. It is shown from simulation examples that dynamic MRDA can significantly improve the fault diagnosis performance.

The rest of this brief is organized as follows. The KL divergence is briefly described in Section II. The proposed MRDA-based approach for fault diagnosis is developed in Section III. The effectiveness of the proposed method is demonstrated in the Tennessee Eastman process (TEP) in Section IV, which is followed by conclusions in Section V.

II. Symmetric Kullback–Leibler Divergence

The KL-divergence, which serves as the basis of the proposed fault diagnosis approach, is briefly reviewed in this section. More details of KL-divergence can be found in [18].

The symmetric KL-divergence is a measure of the difference between two probability density functions, and the divergence between the classes i and j is defined as [18]

$$D(p_i,\ p_j) = \int p_i(\boldsymbol{x}) \ln \frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})} + p_j(\boldsymbol{x}) \ln \frac{p_j(\boldsymbol{x})}{p_i(\boldsymbol{x})} d\boldsymbol{x} \quad (1)$$

where p_i and p_j denote the distributions of data of classes i and j, respectively.

The KL-divergence (1) is nonnegative, i.e.,$D(p_i,\ p_j) \geq 0$, with equality if and only if $p_i = p_j$. A large value of the KL-divergence is associated with the distributions p_i and p_j being well discriminated so that the diagnosis of faults is easy.

If p_i and p_j follow multivariable Gaussian distributions, i.e., $p_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $p_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ with $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_j$ being positive-definite, the KL-divergence can be computed explicitly as [18]

$$\begin{aligned} D(p_i,\ p_j) &= \int p_i(\boldsymbol{x}) \ln \frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})} + p_j(\boldsymbol{x}) \ln \frac{p_j(\boldsymbol{x})}{p_i(\boldsymbol{x})} d\boldsymbol{x} \\ &= \frac{1}{2}\mathrm{tr}\left\{\left(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j\right)\left(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_i^{-1}\right)\right\} \\ &\quad + \frac{1}{2}\mathrm{tr}\left\{\left(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1}\right)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}\right\} \end{aligned} \quad (2)$$

where $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ denote the means and covariance matrices of the classes i and j data, respectively, and $\boldsymbol{A}^{-1}$, $\boldsymbol{A}^{\mathrm{T}}$, and $\mathrm{tr}\{\boldsymbol{A}\}$ denote the inverse, transpose, and trace of matrix $\boldsymbol{A}$.

III. Maximized Ratio Divergence Analysis Method for Fault Diagnosis

A. Dimension Reduction Technique Based on MRDA

A large value of $D(p_i,\ p_j)$ implies that the faults p_i and p_j are easy to be diagnosed, while a value of $D(p_i,\ p_j)$ close to zero indicates that the faults p_i and p_j are difficult to be classified. Therefore, maximizing $D(p_i,\ p_j)$ during the dimension reduction procedure would be beneficial to diagnosing faults. In other words, the approach that produces a large KL-divergence [i.e., a large value of $D(p_i,\ p_j)$] during the dimensionality reduction step possesses good performance of fault diagnosis. The objective of maximizing KL-divergence can be described as

$$\begin{aligned} \max_{\boldsymbol{w}} D(p_i^w,\ p_j^w) = \int & p_i^w(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) \ln \frac{p_i^w(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})}{p_j^w(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})} \\ & + p_j^w(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}) \ln \frac{p_j^w(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})}{p_i^w(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})} d\boldsymbol{x} \end{aligned} \quad (3)$$

where p_i^w and p_j^w denote the probability density functions of classes i and j in the reduction space, respectively, each column of $\boldsymbol{w} \in R^{m\times k}$ represents a loading vector, and k denotes the order of dimension reduction.

In fault diagnosis, a common scenario is to handle data with multiple classes. To this end, objective (3) can be extended by employing the pairwise technique as

$$\max_{\boldsymbol{w}} D_w := \frac{2}{c(c-1)} \sum_{1\leq i<j\leq c} D\left(p_i^w,\ p_j^w\right) \quad (4)$$

where c denotes the number of faults. In order to avoid the issue in which several faulty classes inappropriately dominate the dimensionality reduction procedure, a ratio divergence instead of the original KL-divergence is used, which is defined as

$$\begin{aligned} \max_{\boldsymbol{w}} \bar{D}_w &:= \frac{2}{c(c-1)} \sum_{1\leq i<j\leq c} \bar{D}(p_i,\ p_j) \\ &= \frac{2}{c(c-1)} \sum_{1\leq i<j\leq c} \frac{D\left(p_i^w,\ p_j^w\right)}{D(p_i,\ p_j)} \end{aligned} \quad (5)$$

where $\bar{D}(p_i,\ p_j) := (D(p_i^w,\ p_j^w)/D(p_i,\ p_j))$ denotes the ratio KL-divergence between class i and class j. Since $0 \le D(p_i^w,\ p_j^w) \le D(p_i,\ p_j)$ [18], it can be obtained that ratio divergences $\bar{D}(p_i,\ p_j)$ possess

$$0 \le \bar{D}(p_i,\ p_j) \le 1 \tag{6}$$

and it can be further concluded that the average of ratio divergence $\bar{D}_w$ also satisfies $0 \le \bar{D}_w \le 1$.

If the data of class i follow a multivariable Gaussian distribution, i.e., $p_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, in the reduction space, the projected variable p_i^w follows $p_i^w \sim N(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu}_i, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{w})$. Then, objective (5) can be explicitly expressed as

$$\max_{\boldsymbol{w}} \bar{D}_w = \frac{1}{c(c-1)} \sum_{1\le i\le c} \sum_{j\ne i} \frac{\mathrm{tr}\{(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{w})^{-1}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{M}_{ij}\boldsymbol{w}\} - k}{D(p_i,\ p_j)} \tag{7}$$

where $\boldsymbol{M}_{ij} = \boldsymbol{\Sigma}_j + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}$ and $D(p_i,\ p_j)$ is explicitly expressed as (2).

It is worth mentioning that the form of average of ratio divergence used in the objective functions (5) or (7) facilitates the proposed approach to be less likely influenced by "outlier" data sets during the dimension reduction step, since each ratio divergence $\bar{D}(p_i,\ p_j)$ has a property of $0 \le \bar{D}(p_i,\ p_j) \le 1$. This enables the MRDA approach to be particularly suitable for the fault diagnosis tasks with imbalanced data, where some classes are relatively far away from the remaining ones. In contrast, standard FDA methods perform poorly on these data sets, because FDA can be easily dominant by outlier classes during the procedure of dimensionality reduction, thereby making it lose classification ability for the majority of classes [14], [21].

B. Optimization Algorithm for Learning Loading Vectors

This section investigates solving the optimization problem (7). A numerical algorithm based on the quasi-Newton technique is used for learning a single loading vector and then is extended to learn multiple loading vectors (i.e., loading matrix) by employing the well-known deflation technique [1].

From (7), the objective function for learning a single loading vector $\boldsymbol{w} \in R^{m\times 1}$ is

$$\max_{\boldsymbol{w}} \bar{D}_w = \frac{1}{c(c-1)} \sum_{1\le i\le c} \sum_{j\ne i} \frac{\mathrm{tr}\{(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{w})^{-1}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{M}_{ij}\boldsymbol{w}\} - 1}{D(p_i,\ p_j)}. \tag{8}$$

The gradient of $\bar{D}_w$ with respect to $\boldsymbol{w}$ is computed as

$$\begin{aligned}\frac{\partial \bar{D}_w}{\partial \boldsymbol{w}} = {} & \frac{1}{c(c-1)} \sum_{1\le i\le c} \sum_{j\ne i} \frac{2}{D(p_i,\ p_j)} \\ & \times (\boldsymbol{M}_{ij}\boldsymbol{w}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{w})^{-1} - \boldsymbol{\Sigma}_i\boldsymbol{w}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{w})^{-1} \\ & \times (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{M}_{ij}\boldsymbol{w})(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{w})^{-1}).\end{aligned} \tag{9}$$

Thus, the loading vector $\boldsymbol{w} \in R^{m\times 1}$ can be obtained by the quasi-Newton algorithms, and its corresponding loading score $\boldsymbol{d}$ can be computed as $\boldsymbol{d} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$.

Learning the loading matrix $\boldsymbol{w} \in R^{m\times k}$ by using the deflation technique is investigated below. Let the data of class i with n_i samples be stacked into a matrix as $\boldsymbol{X}_i \in R^{n_i\times m} (i = 1, 2, \ldots, c)$. Similar to PCA and PLS [1], the data matrices for $\boldsymbol{X}_i$ can be represented as the sum of a series of rank one matrices, namely

$$\boldsymbol{X}_i = \sum_{r=1}^{k} \boldsymbol{d}_{i,r}\boldsymbol{w}_r^{\mathrm{T}} + \boldsymbol{E}_i, \tag{10}$$

where $\boldsymbol{w}_r$ $(r = 1, 2, \ldots, k)$ denotes the r^{th} loading vector and $\boldsymbol{d}_{i,r}$ is the corresponding r^{th} loading scores for class i data (i.e., $\boldsymbol{d}_{i,r} = \boldsymbol{X}_i\boldsymbol{w}_r$); $\boldsymbol{E}_i$ denotes the residual matrices.

Thus, the data matrices $\boldsymbol{X}_i$ remained for calculating the $(r+1)^{\mathrm{th}}$ loading vector $\boldsymbol{w}_{r+1}$ can be expressed as

$$\boldsymbol{X}_{i,r} = \boldsymbol{X}_{i,r-1} - \boldsymbol{d}_{i,r}\boldsymbol{w}_r^{\mathrm{T}}. \tag{11}$$

Then, the means and covariance matrices for $\boldsymbol{X}_{i,r}$, i.e., $(\boldsymbol{\mu}_{i,r}, \boldsymbol{\Sigma}_{i,r})$ can be obtained. Therefore, the objective function for learning the $(r+1)^{\mathrm{th}}$ loading vector $\boldsymbol{w}_{r+1}$ becomes

$$\max_{\boldsymbol{w}} \bar{D}_w = \frac{1}{c(c-1)} \sum_{1\le i\le c} \sum_{j\ne i} \frac{\mathrm{tr}\{(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_{i,r}\boldsymbol{w})^{-1}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{M}_{ij,r}\boldsymbol{w}\} - 1}{D(p_i,\ p_j)} \tag{12}$$

where $\boldsymbol{M}_{ij,r} := \boldsymbol{\Sigma}_{j,r} + (\boldsymbol{\mu}_{i,r} - \boldsymbol{\mu}_{j,r})(\boldsymbol{\mu}_{i,r} - \boldsymbol{\mu}_{j,r})^{\mathrm{T}}$. Similarly, $\boldsymbol{w}_{r+1}$ can be computed by using the aforementioned numerical optimization algorithm for solving (8). This procedure is continued until all k loading vectors $\boldsymbol{w} \in R^{m\times k}$ are obtained. The corresponding loading score $\boldsymbol{d}$ can be obtained as $\boldsymbol{d} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$.

Observations are then classified in the k-dimensional space of MRDA using the discriminant function [1], [8]

$$\begin{aligned} g_i(\boldsymbol{x}) = {} & -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{w}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{w})^{-1}\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{\mu}_i) \\ & - \frac{1}{2}\ln[\det(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}_i\boldsymbol{w})]. \end{aligned} \tag{13}$$

An observation $\boldsymbol{x}$ is categorized to class i if

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \quad \forall j \ne i. \tag{14}$$

Remark 1: The MRDA method can be extended to handle serial correlations in the data by augmenting the observation vectors $\boldsymbol{x}$ in (3) with time lags, that is

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_t^{\mathrm{T}} & \boldsymbol{x}_{t-1}^{\mathrm{T}} & \cdots & \boldsymbol{x}_{t-l}^{\mathrm{T}} \\ \boldsymbol{x}_{t-1}^{\mathrm{T}} & \boldsymbol{x}_{t-2}^{\mathrm{T}} & \cdots & \boldsymbol{x}_{t-l-1}^{\mathrm{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{t+l-n}^{\mathrm{T}} & \boldsymbol{x}_{t+l-n-1}^{\mathrm{T}} & \cdots & \boldsymbol{x}_{t-n}^{\mathrm{T}} \end{bmatrix}. \tag{15}$$

Remark 2: Compared with traditional methods such as FDA for fault diagnosis, our MRDA method involves a slightly heavier computation when solving the iterative optimizations. However, this does not become a critical issue in that the computation mainly lies in the offline training stage. Moreover, the extension of our approach to non-Gaussian variables may require density-estimation techniques such as kernel density estimation, which requires a careful theoretical treatment, which is one future direction to improve the performance of our current method.

Remark 3: In practice, imbalanced fault classes may be present in which some fault classes have far more samples than the others. Common approaches to resolve such issues is to oversample the minority classes or undersample the majority classes. Another solution is to add different weights to the KL-divergence of fault classes with imbalanced samples. This idea is analogous to the one in [29], which modifies the KL-divergence as a distance metric for imbalanced classes.

IV. Application to Tennessee Eastman Process

The TEP is a well-known benchmark process for comparing various fault diagnosis algorithms by simulating a realistic industrial process with high fidelity [1], [22], [23]. In this section, TEP is adopted to evaluate the MRDA-based approach for diagnosing faults in comparison with various other methods. Two data sets (i.e., training and testing data) were generated with a sampling interval of 3 min for each fault. The training data are for developing diagnostic models and the testing data is for testing the proficiency of the models. Each training and testing data set contains 480 and 800 samples, respectively, in which each data observation consists of 52 process variables including all the manipulated and measurement variables except for the reactor's stirrer agitation speed. Note that process faults are defined as the operation events that deviate from normal operating conditions [1], [9]. All the faults considered in the examples below represent certain irregular scenarios that upset the process and drive it to the out-of-control status, indicated by the drifted values of the corresponding variables against nominal values.

A. Case Study 1: Faults 3, 4, and 11

In this case study, Faults 3, 4, and 11 produced by the TEP (see Table I) are used to examine the performance of various fault classification methods discussed in this brief with the imbalanced distribution of data. The methods under investigation are FDA, method from [3] (which does not involve dimensionality reduction), classification without dimensionality reduction [i.e., only using the discriminant function (13) with $\boldsymbol{w}$ as an identity matrix], MRDA without ratio, MRDA, and dynamic MRDA with lag $l = 3$. Faults 4 and 11 are associated with the inlet temperature of the reactor cooling water but different in the type of faults [1]. In contrast to these two faults, Fault 3 involves a different process variable—D feed temperature [8].

The classification results for Faults 3, 4, and 11 using the six methods are listed in Table II. For the FDA method, Faults 3 and 11 are misclassified most of the time and Fault 4 has a relatively low misclassification rate. This may be because the objective of FDA does not consider the uneven separations between the classes and may be dominated by some outlier classes. The method from [3], although incorrectly diagnosing Fault 4 most of the time, can separate Faults 3 and 11 with high accuracy, leading to slightly improved performance relative to FDA. The method without dimensionality reduction, only relying on discriminant function (13), yields superior classification performance compared with the former two. One possible explanation is that the number of variable dimensions without stacking dynamic lags, i.e., $l = 1$, is small enough compared with the number of samples, and thus, the gained information in most directions in the observation space can outweigh the inaccuracies in the estimated statistical parameters with finite samples. However, it does not imply that dimensionality reduction is not beneficial to fault classification. In fact, as shown below, the dimensionality reduction can indeed elevate the classification performance with properly chosen loadings. For the MRDA without ratio, it can significantly improve the classification performance by reducing the misclassification rate to 6.1%, in contrast to 39.3% for FDA and 34.2% for the method from [3]. MRDA with ratio can further reinforce the classification performance. However, its increment in classification performance compared with MRDA without ratio is minor. The reason is that, for this case study, the pairwise KL-divergences between the three faults do not differ enough to the extent of causing excessive difference between these two methods in classifying these faults. In fact, the computed pairwise KL-divergence values [see (2)] for this case study are, respectively, $D_{3,4} = 127.7$, $D_{3,11} = 64.2$, and $D_{4,11} = 153.7$. In the next case study, the necessity of including a ratio into the objective function of MRDA will be more clearly revealed with a highly uneven separability among faults. To account for the serial correlations in data samples, we test the performance of dynamic MRDA with ratio, which can further boost the fault classification performance, as shown in Table II.

TABLE I
Description of Faults Selected for Case Studies

ID	Fault description	Type
Fault 1	A/C feed ratio, B composition constant (Stream 4)	Step change
Fault 3	D Feed Temperature	Step change
Fault 4	Reactor cooling water inlet temperature	Step change
Fault 11	Reactor cooling water inlet temperature	Random variation
Fault 13	Reactor kinetics	Slow drift
Fault 14	Reactor cooling water valve	Sticking

TABLE II
Classification Results From Different Methods for Case Study 1

Misclassification rate (%)	Fault3	Fault4	Fault11	Overall
FDA	46.0	13.1	58.9	39.3
Method from [3]	7.63	85.5	9.5	34.2
No dimensionality reduction	36.9	17.6	9.4	21.3
MRDA without ratio	3.0	2.0	13.4	6.1
MRDA	2.9	1.4	12.8	5.7
Dynamic MRDA (lag $l = 3$)	6.4	1.4	4.0	3.9

B. Case Study 2: Faults 1, 3, 4, 11, 13, and 14

In this case study, six faults (i.e., Faults 1, 3, 4, 11, 13, and 14, as specified in Table I) generated by the TEP simulator are further used to evaluate the performance of the proposed MRDA-based approach.

The first two loading vectors ($a = 2$) of each faulty data set extracted by four methods, FDA, MRDA without ratio, MRDA, and dynamic MRDA (lag $l = 3$), are displayed in Figs. S1(a)–(d) of the Supplementary Material, respectively. The bar plots in Fig. 1 represent the pairwise KL-divergences

TABLE III

KL-DIVERGENCE VALUES AMONG ALL PAIRS OF FAULTS 1, 3, 4, 11, 13, AND 14

KL-divergence	Fault 1	Fault 3	Fault 4	Fault 11	Fault 13	Fault 14
Fault 1	0	849.9	846.0	704.4	564.9	952.2
Fault 3	-	0	127.6	64.2	2804.9	320.6
Fault 4	-	-	0	153.7	1831.1	592.5
Fault 11	-	-	-	0	1525.6	123.6
Fault 13	-	-	-	-	0	1750.9
Fault 14	-	-	-	-	-	0

Note: The dash "-" represents symmetric values of corresponding pairs.

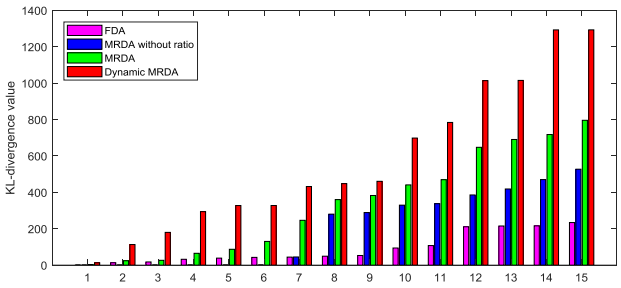

Fig. 1. KL-divergence values of each pair of classes in the lower dimensional space for Case Study 2 produced by the methods of FDA, MRDA without ratio, MRDA, and dynamic MRDA.

in the reduction space of each two different data sets in an ascending order for these methods. Table III summarizes the KL-divergence values for all pairs among these six faults in the original observation space. The KL-divergence has a highly imbalanced distribution among all pairs with the maximum as 2804.9 and the minimum as 64.2.

As shown in Figs. S1(a)–(c) of the Supplementary Material, the projected data generated by MRDA or dynamic MRDA have better visual performance than those by FDA or MRDA without ratio, in terms of separating classes from each other. Furthermore, the bar plots in Fig. 1 show that the KL-divergence values in the lower dimensional space generated by MRDA without ratio are larger than those by FDA for class pairs from 8 to 15. However, for the pairs from 2 to 7, the values from FDA are slightly larger than those from MRDA without ratio, which has almost zero KL-divergence values for these pairs. These small values from MRDA without ratio significantly worsen the classification performance of this method, as shown in Table III. On the other hand, for FDA, the KL-divergence values for pairs 2–7, although small, can assist FDA in separating relevant classes, as supported by the superior performance of FDA than MRDA without ratio in Table IV. This observation clearly highlights the importance of balanced distribution of KL-divergence among all pairs after dimensionality reduction. Therefore, by purely seeking maximization of KL-divergence during dimensionality reduction, as is the method of MRDA without ratio, may contrarily degrade the classification performance. In contrast, the MRDA method results in a much more balanced distribution of KL-divergence among all pairs after dimensionality reduction, as shown in green bars in Fig. 1. As a result, the average misclassification rate is reduced to 22.6%. Furthermore, the dynamic MRDA method generates the largest KL-divergence values for all pairs with the most balanced distribution of KL-divergence, and thus leads to the best misclassification performance (11.2%).

TABLE IV

MISCLASSIFICATION RATE FOR FDA, MRDA WITHOUT RATIO, MRDA, AND DYNAMIC MRDA FOR CASE STUDY 2

Faults	FDA	MRDA without ratio	MRDA	Dynamic MRDA
Fault 1	3.0	12.5	6.2	7.0
Fault 3	42.8	71.4	25.5	13.1
Fault 4	5.0	56.5	7.5	2.0
Fault 11	46.6	85.6	57.0	20.4
Fault 13	58.6	5.5	5.4	18.7
Fault 14	54.2	80.6	34.1	6.1
Overall	35.0	50.2	22.6	11.2

The classification results of these fours methods for Faults 1, 3, 4, 11, 13, and 14 are detailed in Table IV, where the dimensionality reduction order a for the four methods is selected as $a = 3$. As displayed in Table IV, FDA incorrectly diagnoses Faults 3, 11, 13, and 14. MRDA without ratio gives even worse results than FDA, for the reason given in the previous section. In contrast, MRDA with ratio presents a much better performance, with 22.6% misclassification rate, with a factor of 1.5 lower than that of FDA. The main reason is that MRDA employs an objective function in the form of averaged pairwise ratio divergence between each pair of classes, making it robust to outlier faulty data sets. For the dynamic MRDA, the classification performance is further accelerated by almost a factor of 2 than MRDA, by addressing the temporal correlations of process variables.

V. CONCLUSION

This brief presents a fault diagnosis approach based on MRDA. MRDA generates lower dimensional representations of the original data that maximizes the averaged pairwise ratio KL-divergence between each pair of faulty data sets during the dimensionality reduction procedure. In order to solve the problem of maximizing ratio divergence, an iterative algorithm is proposed to learn projection loadings. The advantage of using the MRDA-based fault diagnosis approach is demonstrated with the data collected from TEP, and the simulation results show that the proposed method outperforms FDA-based and other KL-divergence-based methods. It is also shown that using dynamic MRDA can yield even superior performance for fault classification. In the future work, MRDA methods for non-Gaussian process variables can be investigated by kernel density estimation techniques. MRDA-based fault diagnosis with imbalanced sample numbers of fault classes is another direction for our future work.

ACKNOWLEDGMENT

The authors would like to thank Prof. R. Braatz at the Massachusetts Institute of Technology for his guidance and help.

REFERENCES

[1] L. H. Chiang, E. L. Russell, and R. D. Braatz, *Fault Detection and Diagnosis in Industrial Systems*. London, U.K.: Springer-Verlag, 2001.

[2] 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.
[3] 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.
[4] K. Serverson, P. Chaiwatanodom, and R. D. Braatz, "Perspectives on process monitoring of industrial systems," *Annu. Rev. Control*, vol. 42, pp. 190–200, Dec. 2016.
[5] P. Honeine, "Online kernel principal component analysis: A reduced-order model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1814–1826, Sep. 2012.
[6] A. Papaioannou and S. Zafeiriou, "Principal component analysis with complex kernel: The widely linear model," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 9, pp. 1719–1726, Sep. 2014.
[7] 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, 2010.
[8] 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.
[9] L. H. Chiang, M. E. Kotanchek, and A. K. Kordon, "Fault diagnosis based on Fisher discriminant analysis and support vector machines," *Comput. Chem. Eng.*, vol. 28, no. 8, pp. 1389–1401, Jul. 2004.
[10] A. Raich and A. Çinar, "Statistical process monitoring and disturbance diagnosis in multivariable continuous processes," *AIChE J.*, vol. 42, no. 4, pp. 995–1009, Apr. 1996.
[11] L. H. Chiang, E. L. Russell, and R. D. Braatz, "Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 50, no. 2, pp. 243–252, Mar. 2000.
[12] M. Barker and W. Rayens, "Partial least squares for discrimination," *J. Chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.
[13] Q. P. He, S. J. Qin, and J. Wang, "A new fault diagnosis method using fault directions in Fisher discriminant analysis," *AIChE J.*, vol. 51, no. 2, pp. 555–571, 2005.
[14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA, USA: Academic, 1990.
[15] F. De la Torre and T. Kanade, "Multimodal oriented discriminant analysis," in *Proc. 22nd Int. Conf. Mach. Learn.*, Bonn, Germany, Aug. 2005, pp. 177–184.
[16] S. J. Qin, "Survey on data-driven industrial process monitoring and diagnosis," *Annu. Rev. Control*, vol. 36, no. 2, pp. 220–234, Dec. 2012.
[17] S. Yin, S. X. Ding, X. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6418–6428, Nov. 2014.
[18] S. Kullback, *Information Theory and Statistics*. New York, NY, USA: Wiley, 1959.
[19] A. Giantomassi, F. Ferracuti, S. Iarlori, G. Ippoliti, and S. Longhi, "Electric motor fault detection and diagnosis by kernel density estimation and Kullback–Leibler divergence based on stator current measurements," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1770–1780, Nov. 2014.
[20] J. Zeng, U. Kruger, J. Geluk, X. Wang, and L. Xie, "Detecting abnormal situations using the Kullback–Leibler divergence," *Automatica*, vol. 50, no. 11, pp. 2777–2786, Nov. 2014.
[21] G. R. G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan, "A robust minimax approach to classification," *J. Mach. Learn. Res.*, vol. 3, pp. 555–582, Mar. 2003.
[22] P. R. Lyman and C. Georgakis, "Plant-wide control of the Tennessee Eastman problem," *Comput. Chem. Eng.*, vol. 19, no. 3, pp. 321–331, Mar. 1995.
[23] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Comput. Chem. Eng.*, vol. 17, no. 3, pp. 245–255, Mar. 1993.
[24] 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, 2000.
[25] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017.
[26] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757–3767, Jun. 2015.
[27] M. C. Thomas, W. Zhu, and J. A. Romagnoli, "Data mining and clustering in chemical process databases for monitoring and knowledge discovery," *J. Process Control*, vol. 67, pp. 160–175, Jul. 2018.
[28] L. Xie, J. Zeng, U. Kruger, X. Wang, and J. Geluk, "Fault detection in dynamic systems using the Kullback–Leibler divergence," *Control Eng. Pract.*, vol. 43, pp. 39–48, Oct. 2015.
[29] 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.*, to be published.
[30] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
[31] W. Yan, P. Guo, L. Gong, and Z. Li, "Nonlinear and robust statistical process monitoring based on variant autoencoders," *Chemometrics Intell. Lab. Syst.*, vol. 158, pp. 31–40, Nov. 2016.