Locality Preserving Discriminative Canonical Variate Analysis for Fault Diagnosis

Qiugang Lu[a,b], Benben Jiang[b,c], R. Bhushan Gopaluni[a], Philip D. Loewen[d], and Richard D. Braatz[b,1]

[a] Dept. of Chemical and Biological Engineering, The University of British Columbia, Vancouver, BC, V6T 1Z3, Canada

[b] Dept. of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[c] Dept. of Automation, Beijing University of Chemical Technology, Beijing 100029, China

[d] Dept. of Mathematics, The University of British Columbia, Vancouver, BC, V6T 1Z3, Canada

Abstract

This paper proposes a locality preserving discriminative canonical variate analysis (LP-DCVA) scheme for fault diagnosis. The LP-DCVA method provides a set of optimal projection vectors that simultaneously maximizes the within-class mutual canonical correlations, minimizes the between-class mutual canonical correlations, and preserves the local structures present in the data. This method inherits the strength of canonical variate analysis (CVA) in handling high-dimensional data with serial correlations and the advantages of Fisher discriminant analysis (FDA) in pattern classification. Moreover, the incorporation of locality preserving projection (LPP) in this method makes it suitable for dealing with nonlinearities in the form of local manifolds in the data. The solution to the proposed approach is formulated as a generalized eigenvalue problem. The effectiveness of the proposed approach for fault classification is verified by the Tennessee Eastman process. Simulation results show that the LP-DCVA method outperforms the FDA, dynamic FDA (DFDA), CVA-FDA, and localized DFDA (L-DFDA) approaches in fault diagnosis.

[1] Corresponding author: R. D. Braatz. Telephone: +1-617-253-3112; fax: +1-617-258-0546; email: braatz@mit.edu.

Keywords: Fault diagnosis; canonical variate analysis; Fisher discriminant analysis; locality preserving projection; Tennessee Eastman process

1. Introduction

Data-driven process monitoring has shown high value in promoting informed decision-making and enhancing efficient and safe operations of industrial processes (e.g., for reviews and to gain a thorough perspective on the history of the field, see reviews [3] [4] [5] [6] [7] [8] [35] [36] and citations therein). The objective of most industrial process monitoring systems is the *detection* of faults, which are defined as abnormal process operations. Examples of data-driven fault detection methods include principal component analysis and partial least squares, which are multivariate statistical methods that are widely applied in industry, and state-space identification methods that have been widely studied in the academic literature, e.g., [1] [2]. Another objective of interest in process monitoring described in the above reviews is fault *diagnosis* – determining the type and root cause of faults – which can be challenging for modern industrial processes containing a large number of process variables and complicated correlations among variables due to process dynamics and controllers.

Among various methods for fault diagnosis, FDA has received extensive attention due to its efficiency and simplicity in fault classification [9]. Given labeled data sets from several faults, FDA provides projection vectors to map the original data into a lower-dimensional space in which the between-class scatter matrix is maximized while minimizing the within-class scatter matrix. FDA is particularly effective for data that are free of serial correlations [7]. Nevertheless, most industrial processes are slow in dynamics and equipped with fast-sampling sensors. To handle the serial correlations, dynamic FDA (DFDA) has been put forward to augment the observation with its lagged values to capture the dynamic information [3]. Incorporating time lags into auto-correlated data can attenuate the overlapping between different classes of augmented data, leading to improved fault classification [10]. However, similar to dynamic partial least-squares (PLS) and dynamic PCA [11], the performance of DFDA is limited by its implicit assumption of a restrictive noise structure [1].

On the other hand, the last decade has witnessed growing attention on CVA methods [3] [12]. In contrast to PCA and PLS, CVA constructs a more accurate and parsimonious state-space model that allows a general noise structure. CVA relies on maximizing the correlations between combinations of past and future data vectors, which can be transformed into a singular value decomposition (SVD) problem [13] [14]. CVA is mainly employed to estimate the canonical states of the process, which are further utilized to develop a state-space model from the process data. As CVA does not take account of the label information associated with data sets, the application of CVA to fault classification remains rare and is usually combined with FDA [1]. In addition, the potential loss of discriminative information in the CVA model requires extra attention since the CVA criterion may not be compatible with that of FDA [15]. However, the superiority of CVA in modeling dynamic relations in the data supplies a valuable resource to enhance the performance of current techniques for discriminant analysis with large-scale dynamic data.

CVA has a close link with canonical correlation analysis (CCA) [16]. The usage of CCA for discriminant analysis has been reported in the computer vision area. A technique known as discriminant CCA (DCCA) [17] incorporates the class label information into CCA to extract more discriminative features. In DCCA, for data sets with two views, a set of optimal projection vectors are obtained that maximize the canonical correlations between two views of within-class data and minimize those of between-class data, in an analogy to the idea of FDA. Other variants of DCCA have been presented in [18] [19]. It is shown that DCCA yields a better discriminant performance than CCA and PLS for feature recognition [20]. However, to the best of the authors' knowledge, including class label information into CVA as a discriminative CVA (DCVA) method to address the fault diagnosis problem has not been reported in the literature. Note that a critical distinction between DCVA and DCCA is that the data for DCVA usually involve serial (predictive) correlations due to the utilization of past and future data vectors, in addition to the spatial correlations, whereas DCCA only considers the spatial correlations between variables. Besides, DCVA differs from CVA-FDA [1] in that the goal of DCVA is not estimating the canonical states for a state-space model, but rather directly exploring the discriminant features by examining the relations between data sets from different classes.

All aforementioned methods only use the global structure information. To better mine the information hidden in the data, locality preserving methods have been proposed to handle nonlinearities in the form of local structures such as multi-modality [21]. Locality preserving projection (LPP) [22] paves the way for the research on local structure exploration in data analysis. LPP is a linear dimensionality reduction method that preserves local manifold structures of the original data in the lower-dimensional space after projection. Essentially, LPP decomposes nonlinear dimensionality reduction into a set of linear local dimensionality reductions. The combination of LPP and CCA has been explored in [23] [24]. In the realm of fault diagnosis, locality preserving methods have been merged with discriminant analysis methods such as FDA and kernel FDA to boost the fault classification performance [25] [26] [27]. In this article, we present a locality preserving discriminant CVA method, known as LP-DCVA, for fault diagnosis. This method extends the discriminant CCA idea in computer vision and image recognition to the field of fault classification. Specifically, we combine the strengths of CVA and FDA into DCVA to better handle the dynamic data with highly serial correlations. Besides, we present a way to integrate the objectives of DCVA and LPP together to explore local structures in the data to further improve the performance of fault classification.

The rest of this article is organized as follows. Section 2 briefly revisits CVA, FDA, and LPP. The proposed DCVA and LP-DCVA approaches are presented in Section 3. The effectiveness of the proposed approaches is demonstrated in the Tennessee Eastman process in Section 4, followed by conclusions in Section 5.

2. Review of CVA, FDA, and LPP

2.1. CVA

CVA is a well-known multivariate dimensionality reduction method that maximizes the correlation between two set of variables. CVA was first proposed by Hotelling [28] and then employed as a system identification approach to develop ARMA [29] or state-space models [12]. Suppose that the input data $\boldsymbol{u}(t) \in R^{n_u}$ and output data $\boldsymbol{y}(t) \in R^{n_y}$ are generated according to a linear state-space model

$$\boldsymbol{x}(t+1) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{v}(t), \tag{1}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) + \boldsymbol{E}\boldsymbol{v}(t) + \boldsymbol{w}(t), \tag{2}$$

where $\boldsymbol{x}(t) \in R^d$ is the state vector; $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$, and $\boldsymbol{E}$ are system matrices with compatible dimensions; and $\boldsymbol{v}(t)$ and $\boldsymbol{w}(t)$ are respectively the sequences of state and measurement noises with zero mean and constant covariances. A feature associated with the CVA approach is the separation of collected input-output data into past and future information vectors. The state is estimated by maximizing the predictive correlations between the past and future data with the CVA algorithm. Specifically, for a time instant t within the interval $1 \leq t \leq n$, where n is the number of samples, the past information vector $\boldsymbol{p}(t)$ consists of a window of past input and output data up to time $t-1$, i.e.,

$$\boldsymbol{p}(t) = [\boldsymbol{y}^{\mathrm{T}}(t-1), \dots, \boldsymbol{y}^{\mathrm{T}}(t-h), \boldsymbol{u}^{\mathrm{T}}(t-1), \dots, \boldsymbol{u}^{\mathrm{T}}(t-h)]^{\mathrm{T}}, \tag{3}$$

and $\boldsymbol{f}(t)$ contains a window of current and future outputs with the form

$$\boldsymbol{f}(t) = [\boldsymbol{y}^{\mathrm{T}}(t), \boldsymbol{y}^{\mathrm{T}}(t+2), \dots, \boldsymbol{y}^{\mathrm{T}}(t+l-1)]^{\mathrm{T}}. \tag{4}$$

where h and l represent the lags for the past and future vectors.

Assume that the state order is k. For the CVA algorithm, a projection matrix $\boldsymbol{J}_k$ is computed to linearly map the past $\boldsymbol{p}(t)$ into the "memory" vector $\boldsymbol{m}(t)$ with the form

$$\boldsymbol{m}(t) = \boldsymbol{J}_k \boldsymbol{p}(t). \tag{5}$$

The $\boldsymbol{m}(t)$ is referred to as the memory vector instead of the state vector since in practice it may not necessarily contain all the information in the past and thus is regarded as an approximation of the state. With the memory vector, a state-space model is obtained by establishing the optimal prediction of the future based on the current memory. In other words, the goal of the CVA algorithm is seeking the optimal project matrix $\boldsymbol{J}_k$ to minimize the averaged prediction error [16]

$$\mathbb{E}\left\{\left[\boldsymbol{f}(t) - \hat{\boldsymbol{f}}(t)\right]^{\mathrm{T}} \Lambda^{\dagger} \left[\boldsymbol{f}(t) - \hat{\boldsymbol{f}}(t)\right]\right\}, \tag{6}$$

where $\hat{\boldsymbol{f}}(t)$ is the linear optimal forecast of $\boldsymbol{f}(t)$ based on the current memory, i.e., $\hat{\boldsymbol{f}}(t) = \boldsymbol{\Sigma}_{fm}\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{m}(t)$, where $\boldsymbol{\Sigma}_{fm}$ is the covariance between $\boldsymbol{f}(t)$ and $\boldsymbol{m}(t)$ and $\boldsymbol{\Sigma}_{mm}$ is defined similarly. The positive semidefinite weighting matrix Λ reflects the relative importance among output variables. With the CVA

algorithm, the optimal projection can be obtained by solving the singular value decomposition (SVD) problem

$$\boldsymbol{\Sigma}_{pp}^{-1/2}\boldsymbol{\Sigma}_{pf}\boldsymbol{\Sigma}_{ff}^{-1/2} = \boldsymbol{U}\Sigma\boldsymbol{V}^{T}, \tag{7}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are respectively the left and right singular vectors, Σ contains the singular values along its diagonal, and the projection matrix $\boldsymbol{J}_k$ (solution to (6)) is calculated as

$$\boldsymbol{J}_k = \boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{\Sigma}_{pp}^{-1/2}, \tag{8}$$

where $\boldsymbol{U}_k$ stands for the first k columns of the orthonormal matrix $\boldsymbol{U}$.

2.2. FDA

Process data collected under different faults are categorized into classes in which each class of data represents a particular fault. FDA is a classical pattern classification method that maximizes the separation among classes of data from different faults. This goal is achieved by finding linear transformation vectors to maximize the scatter between classes while minimizing the scatter within classes. Given n samples of m-dimensional observations from c classes stacked into a data matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, the element $\boldsymbol{x}_i^{(j)} \in \mathbb{R}^m$, $i = 1, \dots, n_j$, $j = 1, \dots, c$, of $\boldsymbol{X}$ refers to the i-th sample from class j, where n_j is the number of observations for the jth class. The total scatter matrix $\boldsymbol{S}_t$ is defined as

$$\boldsymbol{S}_t = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(\boldsymbol{x}_i^{(j)} - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i^{(j)} - \overline{\boldsymbol{x}}\right)^{T}, \tag{9}$$

where $\overline{\boldsymbol{x}}$ is the total mean of $\boldsymbol{X}$. The within-class scatter matrix is expressed as

$$\boldsymbol{S}_w = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(\boldsymbol{x}_i^{(j)} - \overline{\boldsymbol{x}}_j\right)\left(\boldsymbol{x}_i^{(j)} - \overline{\boldsymbol{x}}_j\right)^{T}, \tag{10}$$

where $\overline{\boldsymbol{x}}_j$ is the mean vector of class j. Similarly, the between-class scatter matrix is formulated as

$$\boldsymbol{S}_b = \sum_{j=1}^{c} n_j\left(\overline{\boldsymbol{x}}_j - \overline{\boldsymbol{x}}\right)\left(\overline{\boldsymbol{x}}_j - \overline{\boldsymbol{x}}\right)^{T}. \tag{11}$$

Note that the total scatter matrix is the sum of the within- and between-class scatter matrices, $\boldsymbol{S}_t = \boldsymbol{S}_w + \boldsymbol{S}_b$.

The objective of FDA is to supply a set of projection vectors, $\boldsymbol{W}$, to maximize the criterion

$$\max_{\boldsymbol{W}\neq 0} \frac{\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}}{\boldsymbol{W}^T \boldsymbol{S}_t \boldsymbol{W}}. \tag{12}$$

It is shown that this optimization is equivalent to a generalized eigenvalue problem,

$$\boldsymbol{S}_b \boldsymbol{w}_k = \lambda_k \boldsymbol{S}_w \boldsymbol{w}_k, \tag{13}$$

where $\boldsymbol{w}_k$ is the kth column of $\boldsymbol{W}$, and a larger eigenvalue λ_k indicates better separability among all classes by projecting the data onto $\boldsymbol{w}_k$. Note that the rank of $\boldsymbol{S}_b$ is less than c, thus there are at most $c-1$ nonzero eigenvalues and only the eigenvectors corresponding to nonzero eigenvalues are useful for separating these classes of data.

With the obtained projection vectors, the data in the $(c-1)$-dimensional space is represented as

$$\boldsymbol{z}_i = \boldsymbol{W}_a^T \boldsymbol{x}_i, \tag{14}$$

where $\boldsymbol{x}_i$ is the ith observation of $\boldsymbol{X}$, and $\boldsymbol{W}_a$ represents the first a columns of $\boldsymbol{W}$. To address the serial correlation in the dynamic data, DFDA has been proposed and widely used in fault diagnosis. The idea of DFDA is to append the data at time t with its past values and then apply FDA to this augmented data matrix. Defining the selected lags of past data as h, the augmented data matrix is

$$\boldsymbol{X}(h) = \begin{bmatrix} \boldsymbol{x}_t & \cdots & \boldsymbol{x}_{t+h-n} \\ \vdots & \ddots & \vdots\vdots \\ \boldsymbol{x}_{t-h} & \cdots & \boldsymbol{x}_{t-n} \end{bmatrix}. \tag{15}$$

The augmented vector provides richer information than a single observation and is effective to uncover the dynamic patterns in the process data. Thus, the DFDA can in general lead to better classification performance than traditional FDA when extensive serial correlations are present.

2.3. LPP

The LPP method is particularly useful for discovering local manifold structures in the original sample space and preserves such structures in the lower-dimensional space. Therefore, LPP can assist in decomposing the global problem into small local linear sub-problems. Define the data samples in the original space as $\boldsymbol{X} = [\boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \ldots \ \boldsymbol{x}_n]$, where n is the number of samples. We use $\boldsymbol{w}_x$ as the projection vector that preserves the manifold in the data set. The data after projection are denoted as $\boldsymbol{z} = [z_1 \ z_2 \ \ldots \ z_n]$, where $z_i = \boldsymbol{w}_x^{\mathrm{T}} \boldsymbol{x}_i$, $i = 1, \ldots, n$. The objective of LPP is to *minimize* the criterion

$$L = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(z_i - z_j\right)^2 S_{ij}^{x}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{w}_x^{\mathrm{T}}\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right) S_{ij}^{x}\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^{\mathrm{T}} \boldsymbol{w}_x, \tag{16}$$

where S_{ij}^{x} is the element of weighting matrix $\boldsymbol{S_x}$ in the ith row and jth column. A widely employed weighting function is the heat kernel, defined by [24]:

$$S_{ij}^{x} = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma}\right), & \text{if } \boldsymbol{x}_i \in \mathcal{N}_k\left(\boldsymbol{x}_j\right) \text{ or } \boldsymbol{x}_j \in \mathcal{N}_k(\boldsymbol{x}_i), \\ 0, & \text{otherwise}, \end{cases} \tag{17}$$

where $\mathcal{N}_k(\boldsymbol{x}_j)$ stands for the k-nearest neighbors of $\boldsymbol{x}_j$. Consider the case that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are within the k-nearest neighbors of either of them such that $S_{ij}^{x} \neq 0$. In such scenario, if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close to each other, then S_{ij}^{x} will be relatively large and the "distance" between z_i and z_j will be heavily penalized. As a result, the obtained projection vectors $\boldsymbol{w}_x$ are those that keep z_i and z_j close. On the other hand, if $\boldsymbol{x}_i$ is not within the k-nearest neighbors of $\boldsymbol{x}_j$ (or vice versa), then $S_{ij}^{x} = 0$ and the criterion (16) does not preserve any structure between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. With this idea, LPP is able to extract and keep the local structures among points in the data.

The objective function of LPP in (16) can be equivalently formulated as

$$L = \boldsymbol{w}_x^{\mathrm{T}} \boldsymbol{X} \boldsymbol{S}_{xx} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{w}_x, \tag{18}$$

where $\boldsymbol{S}_{xx} = \boldsymbol{D}_{xx} - \boldsymbol{S_x}$ with $\boldsymbol{D}_{xx}$ being a diagonal matrix, known as the Laplacian matrix, with each term representing the sum of the corresponding column (or row since $\boldsymbol{S_x}$ is symmetric) [23]. LPP is used in this paper to discover the local structures and enhance the discriminative features for data from different faults.

3. The Proposed Locality Preserving Discriminative Canonical Variate Analysis for Fault Diagnosis

3.1. Discriminative canonical variate analysis (DCVA) method

CVA is an efficient way to construct state-space models to capture the dynamic relationships among process variables. However, CVA does not take into account the class information associated with the data, and thus is not able to explore the discriminative patterns in the data for fault classification. In fact, applying CVA to the data from several classes may discard valuable information that characterizes the

distinctions between different classes and consequently make the data from different faults less distinguishable after processing [15]. In this section, we present a variant of the traditional CVA method, named discriminative CVA (DCVA), which incorporates the ideas of FDA with CVA and accounts for the label information associated with the data samples.

Consider collected input and output data from p classes. Similar to CVA, at time instant t, $\boldsymbol{p}_t^{(k)}$ represents the past vector from class k, $k = 1, \dots, c$. Denote n_k as the number of samples of past information vector for class k, and $n = \sum_{k=1}^{c} n_k$. Note that

$$\boldsymbol{p}_t^{(k)} = [\boldsymbol{y}_{t-1}^{(k)\mathrm{T}}, \dots, \boldsymbol{y}_{t-h}^{(k)\mathrm{T}}, \boldsymbol{u}_{t-1}^{(k)\mathrm{T}}, \dots, \boldsymbol{u}_{t-h}^{(k)\mathrm{T}}]^{\mathrm{T}}, \tag{19}$$

where h is the selected lags of past input and output. In an analogous way, at time t, the future information vector $\boldsymbol{f}_t^{(k)}$ for class k is defined as

$$\boldsymbol{f}_t^{(k)} = [\boldsymbol{y}_t^{(k)\mathrm{T}}, \boldsymbol{y}_{t+1}^{(k)\mathrm{T}}, \dots, \boldsymbol{y}_{t+l}^{(k)\mathrm{T}}]^{\mathrm{T}}, \tag{20}$$

where l is the selected lags of future output. The past information matrix $\boldsymbol{P}$ and future information matrix $\boldsymbol{F}$ are respectively defined as

$$\boldsymbol{P} = \left[\boldsymbol{p}_1^{(1)}, \boldsymbol{p}_2^{(1)}, \dots, \boldsymbol{p}_{n_1}^{(1)}, \boldsymbol{p}_1^{(2)}, \dots, \boldsymbol{p}_{n_2}^{(2)}, \dots, \boldsymbol{p}_{n_c}^{(c)}\right],$$

$$\boldsymbol{F} = \left[\boldsymbol{f}_1^{(1)}, \boldsymbol{f}_2^{(1)}, \dots, \boldsymbol{f}_{n_1}^{(1)}, \boldsymbol{f}_1^{(2)}, \dots, \boldsymbol{f}_{n_2}^{(2)}, \dots, \boldsymbol{f}_{n_c}^{(c)}\right].$$

Notice that traditional CVA maximizes the predictive relationship between pairwise $\boldsymbol{p}_t^{(k)}$ and $\boldsymbol{f}_t^{(k)}$, i.e., there exists a temporal one-to-one correspondence between past and future vectors at each time instant. This correspondence is essential for developing state estimates and process models. However, for DCVA, instead of seeking such relationships (since the objective of DCVA is not estimating the states), the interest is in discovering discriminative traits among classes. More formally, the goal of DCVA is maximizing the mutual correlations of past and future vectors within the class while minimizing the mutual correlations of those in different classes. The mutual correlation refers to the correlation between any past and future vectors without considering the temporal correspondence. It is apparent that using the

mutual correlations can thoroughly reveal the information in the data and thus facilitate the discovery of discriminative patterns for fault diagnosis.

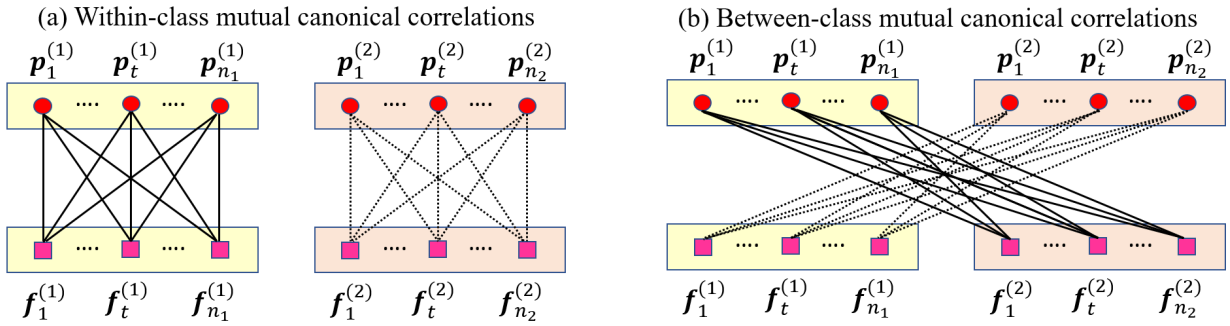


Figure 1. Illustration of the within-class and between-class mutual canonical correlations.

Without loss of generality, both future and past information data are assumed to have been mean-centered and auto-scaled. The DCVA aims at finding projection vectors $\boldsymbol{w}_p$ and $\boldsymbol{w}_f$ for two views $\boldsymbol{P}$ and $\boldsymbol{F}$ so as to maximize the discriminative canonical correlations, i.e., maximizing within-class mutual canonical correlations and simultaneously minimizing between-class mutual canonical correlations. The idea of DCVA is illustrated in Fig. 1. The expressions for within-class and between-class canonical *cross-covariance* matrices $\boldsymbol{C}_w$ and $\boldsymbol{C}_b$ are respectively defined as

$$\boldsymbol{C}_w = \sum_{k=1}^{c} \sum_{t=1}^{n_k} \sum_{s=1}^{n_k} \boldsymbol{p}_t^{(k)} \boldsymbol{f}_s^{(k)\mathrm{T}},$$

$$\boldsymbol{C}_b = \sum_{k=1}^{c} \sum_{p=1,p\neq k}^{c} \sum_{t=1}^{n_k} \sum_{s=1}^{n_p} \boldsymbol{p}_t^{(k)} \boldsymbol{f}_s^{(p)\mathrm{T}}.$$

It follows that $\boldsymbol{C}_w$ and $\boldsymbol{C}_b$ can be simplified as

$$\boldsymbol{C}_w = \sum_{k=1}^{c} \left(\boldsymbol{P}\boldsymbol{E}_{n_k}\right)\left(\boldsymbol{F}\boldsymbol{E}_{n_k}\right)^{\mathrm{T}} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{F}^{\mathrm{T}}, \tag{21}$$

$$\boldsymbol{C}_b = (\boldsymbol{P}\boldsymbol{1}_n)(\boldsymbol{F}\boldsymbol{1}_n)^{\mathrm{T}} - \boldsymbol{P}\boldsymbol{A}\boldsymbol{F}^{\mathrm{T}} = -\boldsymbol{P}\boldsymbol{A}\boldsymbol{F}^{\mathrm{T}}, \tag{22}$$

where $\boldsymbol{1}_n$ is a vector of ones with dimension n, $\boldsymbol{A} = diag\{\boldsymbol{E}_{n_1}, \ldots, \boldsymbol{E}_{n_c}\}$, and $\boldsymbol{E}_{n_k} = \boldsymbol{1}_n \boldsymbol{1}_n^{\mathrm{T}}$, $k = 1, \ldots, c$. The first term in $\boldsymbol{C}_b$ vanishes since both $\boldsymbol{P}$ and $\boldsymbol{F}$ have been centered. The objective function of DCVA is expressed as maximizing

$$\frac{\boldsymbol{w}_p^T \boldsymbol{C}_w \boldsymbol{w}_f - \eta \boldsymbol{w}_p^T \boldsymbol{C}_b \boldsymbol{w}_f}{\sqrt{\boldsymbol{w}_p^T \boldsymbol{P}\boldsymbol{P}^T \boldsymbol{w}_p}\sqrt{\boldsymbol{w}_f^T \boldsymbol{F}\boldsymbol{F}^T \boldsymbol{w}_f}} = \frac{(1+\eta)\boldsymbol{w}_p^T \boldsymbol{P}\boldsymbol{A}\boldsymbol{F}^{\mathrm{T}} \boldsymbol{w}_f}{\sqrt{\boldsymbol{w}_p^T \boldsymbol{P}\boldsymbol{P}^T \boldsymbol{w}_p}\sqrt{\boldsymbol{w}_f^T \boldsymbol{F}\boldsymbol{F}^T \boldsymbol{w}_f}}, \tag{23}$$

where η is a tuning parameter. From (23), it can be seen that the optimal projection vectors are independent of the tuning parameter η. Moreover, the denominator of mutual canonical correlations in (23) is the auto-covariance of latent variables, which is not able to reveal the local structures in the data. To further enhance the performance of DCVA, in the next subsection, we incorporate the idea of LPP in the formulation of within-class and between-class canonical correlations.

3.2. Locality preserving DCVA (LP-DCVA) method for fault diagnosis

Given that the past and future information data $\boldsymbol{P}$ and $\boldsymbol{F}$ are from p classes, for each class, the objective of LPP is stated as minimizing

$$L_p^{(k)} = \boldsymbol{w}_p^{\mathrm{T}} \boldsymbol{P}^{(k)} \boldsymbol{S}_{pp}^{(k)} \boldsymbol{P}^{(k)\mathrm{T}} \boldsymbol{w}_p, \;\; L_f^{(k)} = \boldsymbol{w}_f^{\mathrm{T}} \boldsymbol{F}^{(k)} \boldsymbol{S}_{ff}^{(k)} \boldsymbol{F}^{(k)\mathrm{T}} \boldsymbol{w}_f, \qquad k = 1, \dots c,$$

where $\boldsymbol{S}_{pp}^{(k)}$ is the Laplacian matrix for the kth class $\boldsymbol{P}^{(k)}$ and $\boldsymbol{P}^{(k)} = \left[\boldsymbol{p}_1^{(k)}, \boldsymbol{p}_2^{(k)}, \dots, \boldsymbol{p}_{n_k}^{(k)}\right]$. The term $L_f^{(k)}$ is defined analogously. Combining the objective functions of LPP for c classes of past and future data, the within-class locality preserving matrices are

$$\boldsymbol{S}_{pp} = \boldsymbol{P} diag\left\{\boldsymbol{S}_{pp}^{(1)}, \dots, \boldsymbol{S}_{pp}^{(c)}\right\} \boldsymbol{P}^{\mathrm{T}}, \;\; \boldsymbol{S}_{ff} = \boldsymbol{F} diag\left\{\boldsymbol{S}_{ff}^{(1)}, \dots, \boldsymbol{S}_{ff}^{(c)}\right\} \boldsymbol{F}^{\mathrm{T}}. \tag{24}$$

where $\boldsymbol{P} = \left[\boldsymbol{P}^{(1)}, \boldsymbol{P}^{(2)}, \dots, \boldsymbol{P}^{(c)}\right]$ and $\boldsymbol{F} = \left[\boldsymbol{F}^{(1)}, \boldsymbol{F}^{(2)}, \dots, \boldsymbol{F}^{(c)}\right]$. In the LP-DCVA method, the goal of locality preserving projection is integrated with that of DCVA as

$$\max_{\boldsymbol{w}_p, \boldsymbol{w}_f} \frac{\boldsymbol{w}_p^T \boldsymbol{P} \boldsymbol{A} \boldsymbol{F}^{\mathrm{T}} \boldsymbol{w}_f}{\sqrt{\boldsymbol{w}_p^T \boldsymbol{S}_{pp} \boldsymbol{w}_p \cdot \boldsymbol{w}_f^T \boldsymbol{S}_{ff} \boldsymbol{w}_f}}. \tag{25}$$

This optimization simultaneously maximizes the within-class mutual canonical correlations, preserves the local manifold in the original data after projection, and minimizes the between-class mutual canonical correlations. Following the standard procedures of CVA, (24) can be equivalently written as

$$\max_{\boldsymbol{w}_p, \boldsymbol{w}_f} \boldsymbol{w}_p^T \boldsymbol{P} \boldsymbol{A} \boldsymbol{F}^{\mathrm{T}} \boldsymbol{w}_f$$
$$s.t. \; \boldsymbol{w}_p^T \boldsymbol{S}_{pp} \boldsymbol{w}_p = 1, \; \boldsymbol{w}_f^T \boldsymbol{S}_{ff} \boldsymbol{w}_f = 1.$$

This problem can be readily solved by the generalized eigenvalue problem,

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{P} \boldsymbol{A} \boldsymbol{F}^{\mathrm{T}} \\ \boldsymbol{F} \boldsymbol{A} \boldsymbol{P}^{\mathrm{T}} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_p \\ \boldsymbol{w}_f \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{S}_{pp} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{ff} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_p \\ \boldsymbol{w}_f \end{bmatrix}. \tag{26}$$

Similar to FDA, the eigenvectors corresponding to the first a (where $1 \le a \le c-1$) largest eigenvalues are reserved as the projection vectors onto which the separation of data between classes is maximized. Define the set of a projection vectors as $\boldsymbol{W}_p = [\boldsymbol{w}_p^1, \dots, \boldsymbol{w}_p^a]$, $\boldsymbol{W}_f = [\boldsymbol{w}_f^1, \dots, \boldsymbol{w}_f^a]$, respectively, for the past and future information data $\boldsymbol{P}$ and $\boldsymbol{F}$. The transformed data for an example $[\boldsymbol{p}^{\mathrm{T}} \ \ \boldsymbol{f}^{\mathrm{T}}]^{\mathrm{T}}$ in the a-dimensional space is represented as $\boldsymbol{z} = \left[\boldsymbol{z}_p^{\mathrm{T}} \ \ \boldsymbol{z}_f^{\mathrm{T}}\right]^{\mathrm{T}}$ with

$$\boldsymbol{z}_p = \boldsymbol{W}_p^{\mathrm{T}}\boldsymbol{p}, \quad \boldsymbol{z}_f = \boldsymbol{W}_f^{\mathrm{T}}\boldsymbol{f}. \tag{27}$$

The discriminant function [30]:

$$g_j(\boldsymbol{x}) = -\frac{1}{2}\left(\boldsymbol{x} - \overline{\boldsymbol{x}}_j\right)^{\mathrm{T}} \boldsymbol{W}_a \left(\frac{1}{n_j - 1}\boldsymbol{W}_a^T \boldsymbol{S}_j \boldsymbol{W}_a\right)^T \boldsymbol{W}_a^T\left(\boldsymbol{x} - \overline{\boldsymbol{x}}_j\right) - \frac{1}{2} ln\left[det\left(\frac{1}{n_j - 1}\boldsymbol{W}_a^T \boldsymbol{S}_j \boldsymbol{W}_a\right)\right], \tag{28}$$

can be used to determine the classification of an example in the a-dimensional space, where $\boldsymbol{W}_a = [\boldsymbol{W}_p \ \boldsymbol{W}_f]$, $\boldsymbol{x} = [\boldsymbol{p}^{\mathrm{T}} \ \boldsymbol{f}^{\mathrm{T}}]^{\mathrm{T}}$ and $\overline{\boldsymbol{x}}_j$ is the mean value of class j. An observation $\boldsymbol{x}$ is classified into class j if $g_j(\boldsymbol{x}) > g_i(\boldsymbol{x}), \forall i \ne j$. The algorithm of LP-DCVA is shown in Algorithm 1, where N represents the number of samples of process variables.

Algorithm 1: Locality preserving discriminant canonical variate analysis

Input:	Process input and output data $[\boldsymbol{u}_1 \ \boldsymbol{u}_2 \ \dots \ \boldsymbol{u}_N]$, $[\boldsymbol{y}_1 \ \boldsymbol{y}_2 \ \dots . \boldsymbol{y}_N]$
1:	Given lags h, l, tuning parameters σ, a, κ, form past data $\boldsymbol{P}$ and future data $\boldsymbol{F}$
2:	Compute the weighting matrices $\boldsymbol{S}_p^{(k)}$ and $\boldsymbol{S}_f^{(k)}$, $k = 1, \dots, c$
3:	Compute the Laplacian matrices $\boldsymbol{S}_{pp}^{(k)}$ and $\boldsymbol{S}_{ff}^{(k)}$, $k = 1, \dots, c$
4:	Construct $\boldsymbol{A}$ according to (21), $\boldsymbol{S}_{pp}$ and $\boldsymbol{S}_{ff}$ according to (24)
5:	Solve the eigenvalue problem (26)
Output:	$\boldsymbol{W}_p \leftarrow [\boldsymbol{w}_p^1, \dots, \boldsymbol{w}_p^a]$, $\boldsymbol{W}_f \leftarrow [\boldsymbol{w}_f^1, \dots, \boldsymbol{w}_f^a]$

The LP-DCVA algorithm involves a set of tuning parameters that can impact the classification performance. A summary of these tuning parameters and their suggested values are listed in Table 1.

4. Application to the Tennessee Eastman Process

The Tennessee Eastman Process (TEP) is a well-known platform to validate and compare various fault detection and diagnosis techniques. For other validation synthetic examples than TEP, the readers can refer to [31] [32] and the references therein. This section applies the proposed LP-DCVA method for

fault diagnosis to simulated data from the TEP simulator. The diagram of TEP is shown in Fig. 2. The TEP has five major components, namely a two-phase reactor, a condenser, a compressor, a vapor/liquid separator, and a stripper. Since the TEP is open-loop unstable, a controller must be in the loop to generate simulation data. More information regarding the TEP and control strategy is provided in [3] and in the references therein. The TEP has 52 process variables, consisting of 41 process measurements and 11 manipulated variables. There are 21 pre-programed faults in the TEP simulator and a list of these faults is given in Table 2.

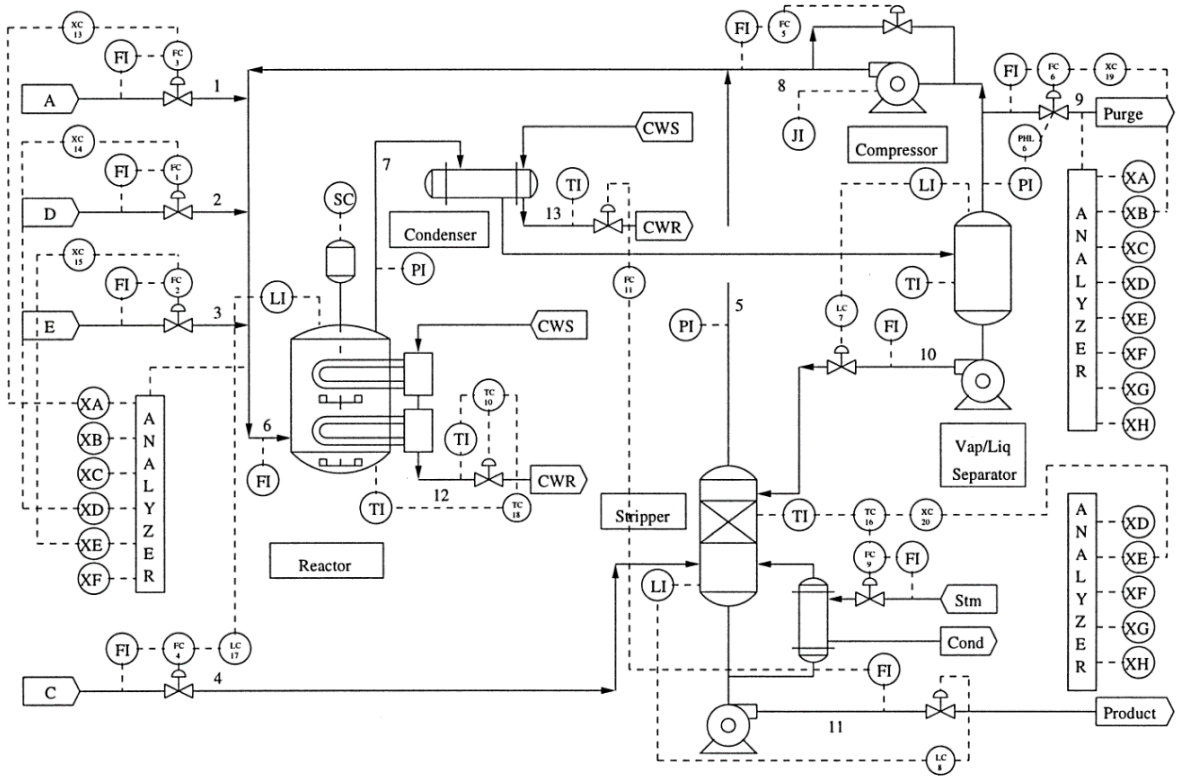


Figure 2. Flow chart for the Tennessee Eastman Process *[3]*.

Table 1. A summary of tuning parameters for the LP-DCVA algorithm

Tuning parameters	Note
Lags h and l in (3) and (4)	Determined by cross validation
The parameter σ in the heat kernel (17)	Suggested value $\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/(n^2-n)$ [23]
The # of nearest neighbors κ in (17)	Determined by cross-validation
The # of projection vectors a	Suggested value $(c-1)$, where c is the # of classes

For each fault, there are three types of data: training data, validation data, and test data. Each training dataset contains 480 observations and is used to build statistical models for fault diagnosis. Each

validation dataset contains 480 observations and is used to cross-verify the performance of the trained models and determine the values of the tuning parameters. The testing dataset contains 800 observations to test the performance of the fault diagnosis techniques. The sampling interval is 3 minutes. In this section, two examples are provided to compare the fault classification performance of FDA, DFDA, CVA-FDA, L-DFDA [26], and LP-DCVA.

Table 2. The process faults involved in the simulation [20].

Variables	Description	Type
Case study 1:		
IDV(3)	D Feed Temperature (Stream 2)	Step
IDV(4)	Reactor Cooling Water Inlet Temperature	Step
IDV(11)	Reactor Cooling Water Inlet Temperature	Random variation
Case study 2:		
IDV(2)	B Composition, A/C Ratio Constant (Stream 4)	Step
IDV(5)	Condenser Cooling Water Inlet Temperature	Step
IDV(8)	A, B, C Feed Composition (Stream 4)	Random variation
IDV(12)	Condenser Cooling Water Inlet Temperature	Random variation
IDV(13)	Reaction Kinetics	Slow drift
IDV(14)	Reactor Cooling Water Valve	Sticking

4.1 Case study 1: Faults 3, 4 and 11

Faults 3, 4, and 11 have significant overlap since both Faults 4 and 11 are associated with reactor cooling water inlet temperature. For the training data from the three faults, FDA, DFDA, CVA-FDA, L-DFDA, and LP-DCVA are applied to establish the fault diagnosis models. The validation data are used to specify the best tuning parameters. For simplicity, we set the lags h and l to be equal. The optimal values of lags for DFDA in this case study are shown to be $h = l = 9$ from cross-validation. The lags for CVA-FDA, L-DFDA and LP-DCVA are chosen to be the same as for DFDA. The optimal number $\kappa = 6$ of nearest neighbors for LP-DCVA was determined by cross-validation. The heat kernel parameter for LP-DCVA and the reserved number of projection vectors for these methods are chosen according to Table 1. The kernel parameter $\sigma = 335$ for L-DFDA was chosen from cross-validation.

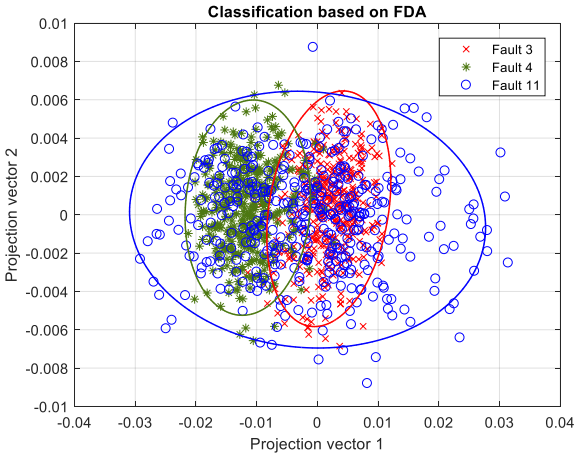
With the selected tuning parameters, Fig. 3a-e demonstrate the scores on the first two projected vectors based on FDA, DFDA, CVA-FDA, L-DFDA, and LP-DCVA, respectively, for the validation data. The ellipse encompassing each data set indicates the 95% confidence threshold. For FDA, a large portion

of overlapping between Fault 4 (or Fault 3) with Fault 11 is observed in the score space. This observation is mainly because FDA does not take account of the serial correlations among samples, thus failing to extract this information from the data. Fig. 3b illustrates that the separation is improved after accounting for the dynamic relationship in the data with DFDA, but there still exists a large degree of overlap among these data sets. Fig. 3c demonstrates that CVA-FDA method can well distinguish Fault 3 and Fault 4, but a significant amount of overlap still exists between those faults and Fault 11. Fig. 3d shows that with L-DFDA the intersections decline furthermore but the improvement is not large. Fig. 3e shows that, with LP-DCVA, the separation between these clusters becomes more distinct.

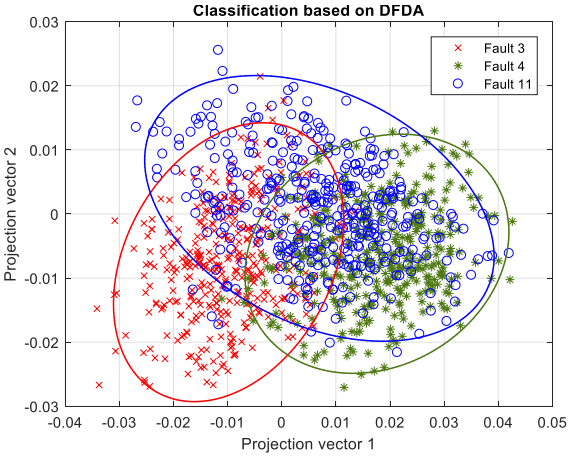
The test data for three faults are further employed to validate the performance of these methods. The comparison results are shown in Fig. 4 and Table 3. As seen in Fig. 4, Fault 4 is easier to identify than the other two faults. Specifically, for the FDA method, Faults 3 and 11 are incorrectly classified most of the time. DFDA, CVA-FDA, and L-DFDA can effectively increment the classification performance for Faults 3 and 11 compared with FDA. The LP-DCVA method gives the best classification performance, which is consistent with its full exploration of local structures of the data and simultaneously consideration of global discriminant information.

Table 3 shows the misclassification rates for three faults with above methods. FDA can recognize Fault 4 reasonably well with only 11.25% misclassification rate. However, FDA has high misclassification rates for Faults 3 and 11. DFDA reduces the misclassification rates for Faults 3 and 11 but slightly increases the rate for Fault 4. CVA-FDA significantly decreases the misclassification rate for Fault 4 but with a degraded performance in recognizing Fault 3. A possible explanation is that, for this two-stage method, some critical information in distinguishing Fault 3 is lost when building the CVA model. L-DFDA further decreases the misclassification rate for Fault 11 compared with the former three methods but the performance for classifying Fault 3 has a small deterioration. In contrast, LP-DCVA reduces the misclassification rates for all three faults at the same time compared with the other methods. Note that DFDA, CVA-FDA, and L-DFDA are almost on the same level (between 25% and 28%) in the performance of misclassification rate, which is due to the inherent difficulty in separating these three

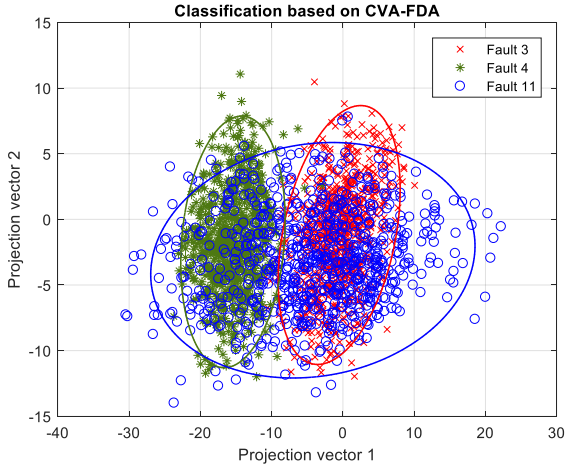
faults. However, LP-DCVA drastically improves the performance by almost 20% relative to L-DFDA. This example clearly shows the advantage of using LP-DCVA for fault diagnosis.



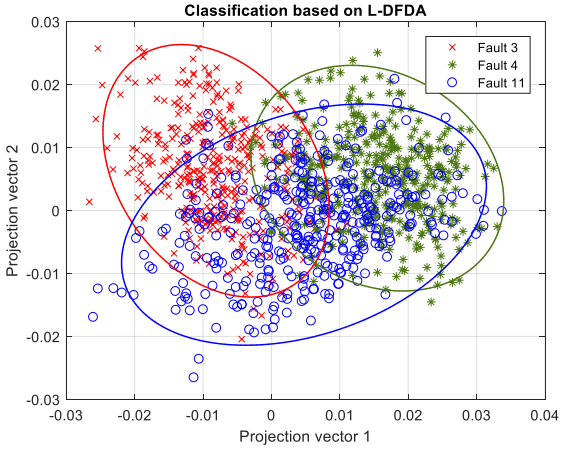
(a)



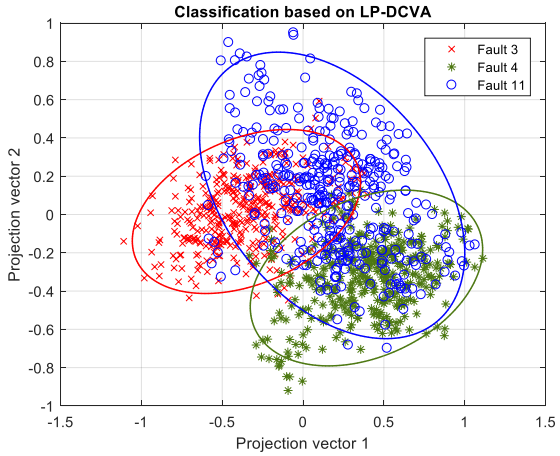
(b)



(c)



(d)

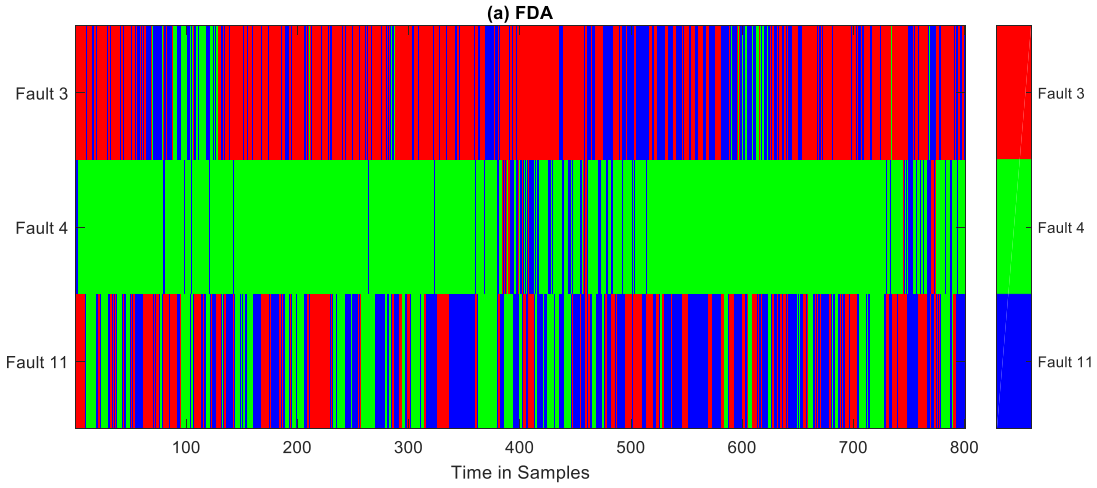


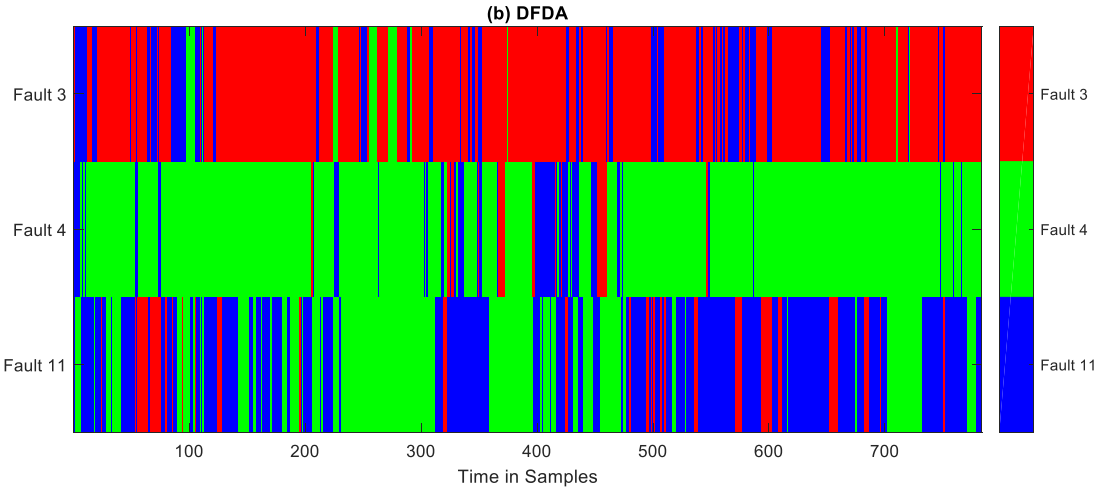
(e)

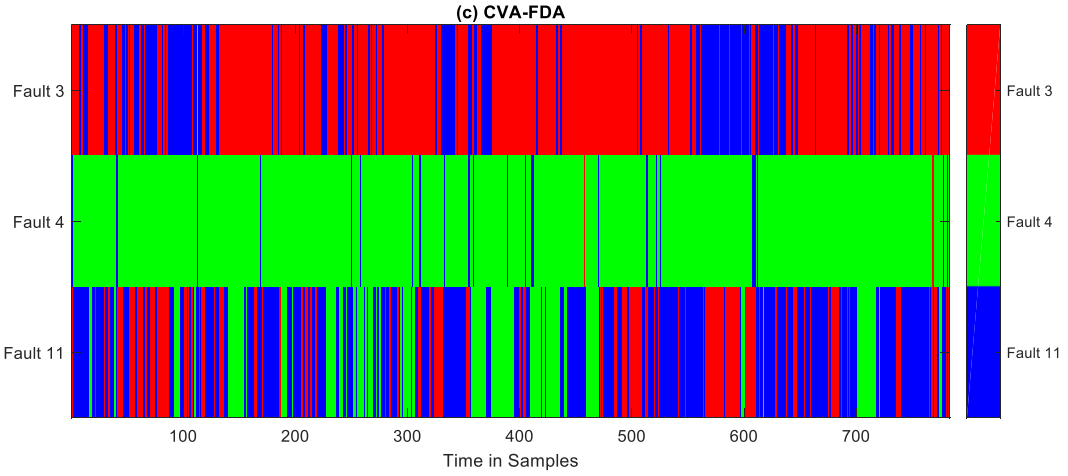
Figure 3. Classification results with three methods on the validation data.

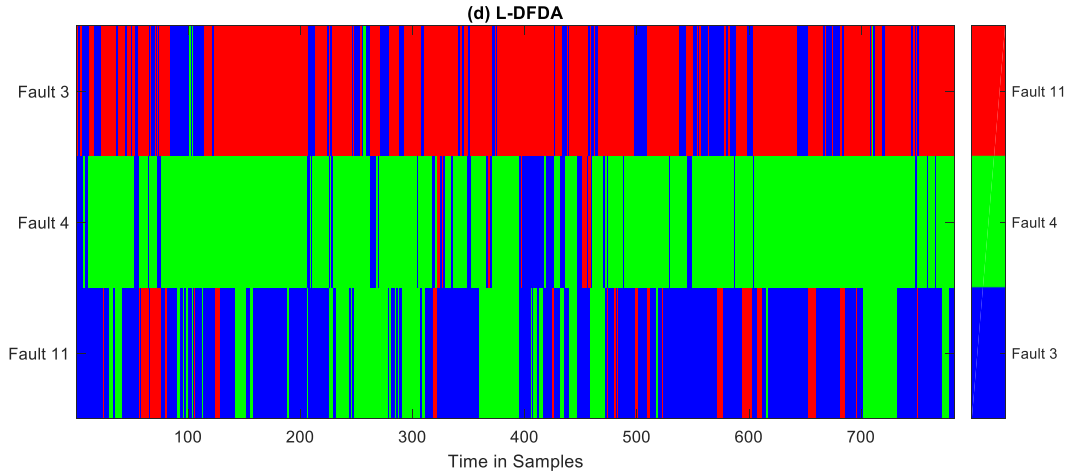
Table 3. Misclassification rates for Faults 3, 4, and 11

Method	Misclassification rates for testing data			
	Fault 3	Fault 4	Fault 11	Overall
FDA	0.3738	0.1125	0.5687	0.3517
DFDA	0.2286	0.1456	0.4687	0.2810
CVA-FDA	0.3103	0.0421	0.4674	0.2733
L-DFDA	0.2656	0.1507	0.3627	0.2597
LP-DCVA	0.2259	0.0945	0.3052	0.2085




(b) DFDA
Fault 3
Fault 4
Fault 11
Fault 3
Fault 4
Fault 11
100
200
300
400
500
600
700
Time in Samples


(c) CVA-FDA
Fault 3
Fault 4
Fault 11
Fault 3
Fault 4
Fault 11
100
200
300
400
500
600
700
Time in Samples


(d) L-DFDA
Fault 3
Fault 4
Fault 11
Fault 11
Fault 4
Fault 3
100
200
300
400
500
600
700
Time in Samples

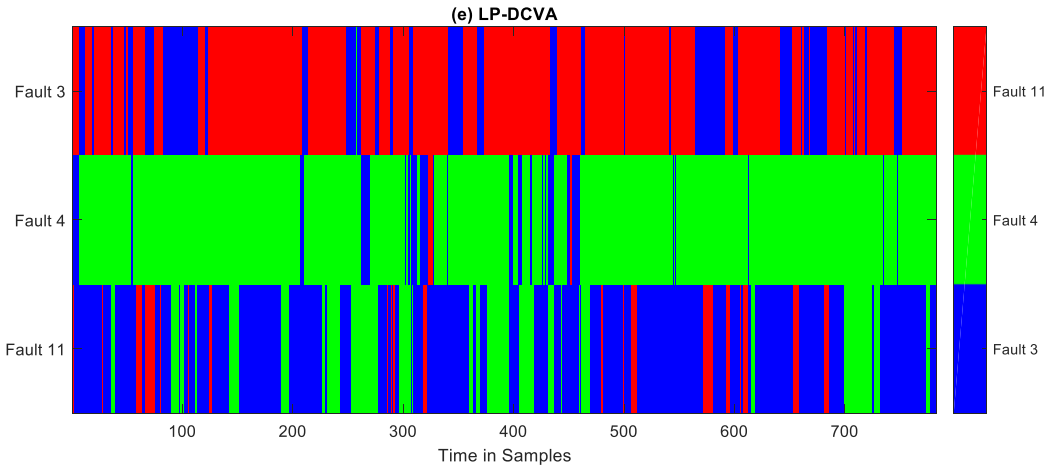


Figure 4. Classification results on the test data for Faults 3, 4, and 11.

4.2 Case study 2: Faults 2, 5, 8, 12, 13, and 14

This case study evaluates the fault diagnosis performance for Faults 2, 5, 8, 12, 13, and 14. Faults 2 and 8 are associated with the faults occurred in the feed composition in Stream 4. Faults 5, 12, and 14 are relevant to the cooling water for the condenser and reactor. The lags are determined from cross validation as $h = l = 3$ for DFDA, CVA-FDA, L-DFDA, and LP-DCVA. The number κ of nearest neighbors is chosen as 10. The heat kernel parameter for LP-DCVA is specified according to the rule-of-thumb in Table 1 and the kernel parameter for L-DFDA is selected as $\sigma = 100$.

Fig. 5 displays the fault classification results for these six faults with $a = 5$. It is observed that Faults 2 and 5 are correctly recognized most of the time by these methods. FDA yields a large number of false classifications for Faults 8, 12, and 13. DFDA slightly improves the performance by reducing the amount of incorrect categorizations for these three faults. The overall misclassification rate is still at a high level, observed from Fig. 5b. CVD-FDA further enhances the classification performance for Fault 8 and Fault 13 but the overall performance for these six faults is only slightly better than DFDA. L-DFDA improves the classification performance by considering the local structures in the data, as shown in Fig. 5d. On the other hand, with LP-DCVA, the misclassification rate for Fault 13 is dramatically decreased. The obtained misclassification rates for each fault from these methods are illustrated in Table 4. LP-DCVA

provides a comparable performance with FDA and DFDA for Faults 2, 5, and 14 that are easy to group. Moreover, LP-DCVA significantly improves the classification performance for Fault 13 by reducing nearly 20% misclassification rates compared with the other four methods. The overall misclassification rate from LP-DCVA is almost 10% lower than those from FDA, DFDA, and CVA-FDA.

Table 4. Misclassification rates for Faults 2, 5, 8, 12, 13, and 14

Fault	Misclassification rates for testing data				
	FDA	DFDA	CVA-FDA	L-DFDA	LP-DCVA
Fault 2	0.0238	0.0189	0.0240	0.0138	0.0377
Fault 5	0.0225	0.0176	0.0227	0.0189	0.0201
Fault 8	0.3350	0.3182	0.2951	0.1371	0.2000
Fault 12	0.2500	0.1698	0.2346	0.1484	0.1484
Fault 13	0.6687	0.5711	0.5284	0.4730	0.2503
Fault 14	0.0813	0.1082	0.0542	0.0214	0.0239
Overall	0.2302	0.2006	0.1931	0.1354	0.1134

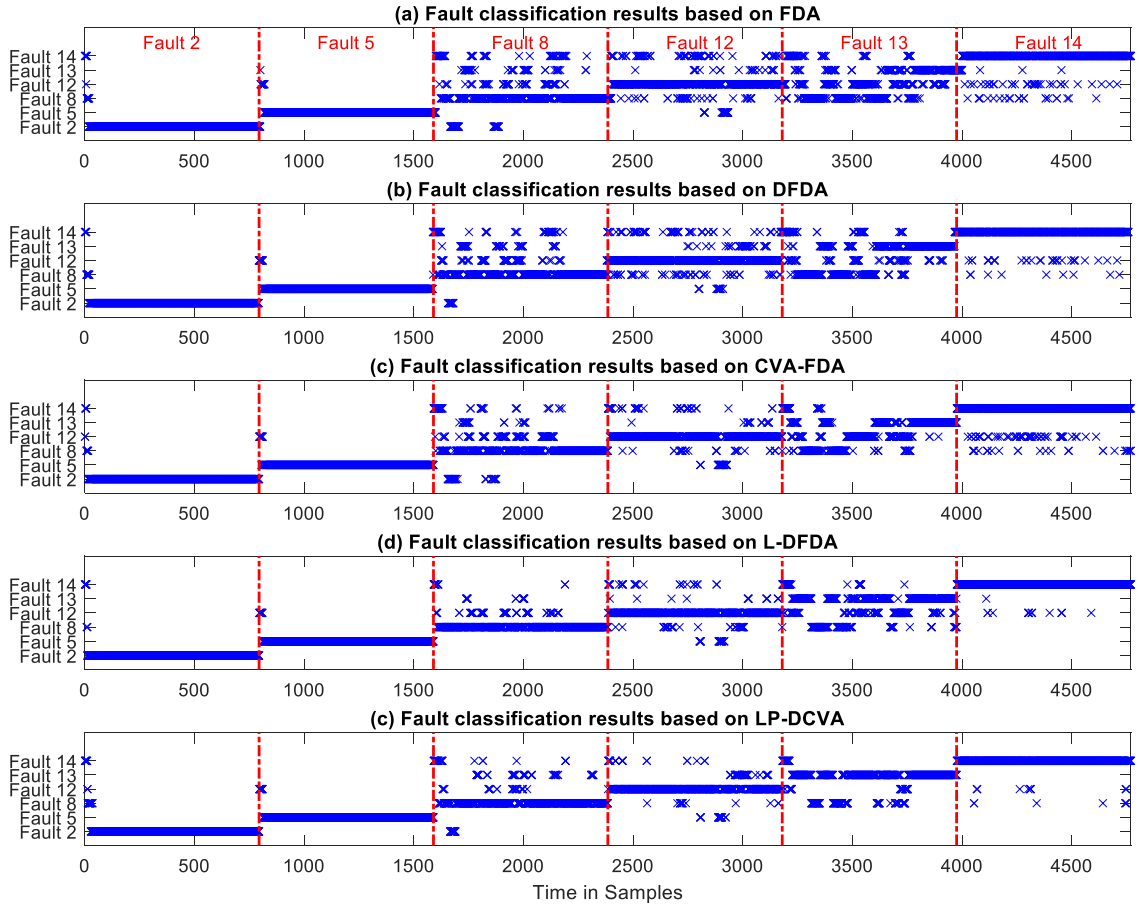


Figure 5. Classification results on the test data for Faults 2, 5, 8, 12, 13, and 14.

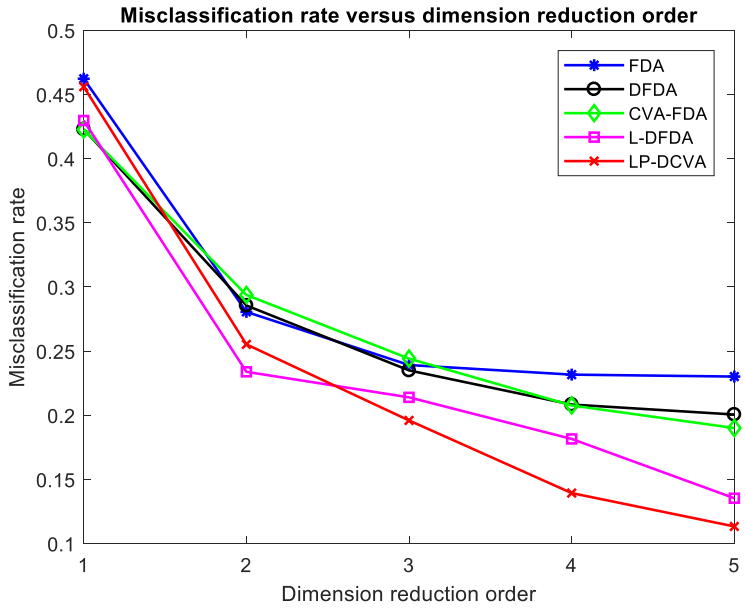


Figure 6. Misclassification rates for different orders of dimension reduction with different methods.

Fig. 6 displays the overall misclassification rates based on five methods under different numbers of projection vectors. These misclassification rates decrease monotonically as the order of dimension reduction increases. For low reduction order, the performance of these four methods does not show significant distinctions. It is observed that CVA-FDA method gives almost the same performance as DFDA and the reason may be, as explained in previous example, due to the loss of discriminative information during the dimensionality reduction in obtaining the CVA model. As the reduction order increases, the superior performance of L-FDFA and LP-DCVA becomes evident. This observation verifies the advantages of using local information in the data for separating different faults. Moreover, the superior performance of LP-DCVA than L-DFDA further motivates the use of LP-DCVA for fault classification.

5. Conclusions

This article presents a locality preserving discriminative CVA approach for fault diagnosis, which combines the merits of CVA in handling the serial and spatial correlations in high-dimensional data and the merits of FDA in maximizing the separations among different classes of data. Similar to CVA,

collected input and output data are split into past and future information vectors in the LP-DCVA approach. This method simultaneously maximizes the within-class mutual canonical correlations, minimizes the between-class mutual canonical correlations and keeps the local manifolds in the data. It is shown that the LP-DCVA method can be transformed into a generalized eigenvalue problem and thus closed-form solutions are obtained. An algorithm is presented to implement the proposed LP-DCVA method. In two simulation examples on the TEP, the LP-DCVA method provides superior performance for fault classifications than FDA, DFDA, CVA-FDA, and L-DFDA for fault classification.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by the Vanier Canada Graduate Scholarships (Vanier CGS). The second author is grateful for the financial support from the National Natural Science Foundation of China (61603024). The last author acknowledges the Edwin R. Gilliland Professorship.

References

[1] 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," *Computers & Chemical Engineering,* vol. 77, no. 9, pp. 1-9, 2015.

[2] R. J. Treasure, U. Kruger and J. E. Cooper, "Dynamic multivariate statistical process control using subspace identification," *Journal of Process Control,* vol. 14, no. 3, pp. 279-292, 2004.

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

[4] S. Joe Qin, "Survey on data-driven industrial process monitoring and diagnosis," *Annual Reviews in Control,* vol. 36, no. 2, pp. 220-234, 2012.

[5] S. Joe Qin, "Statistical process monitoring: basics and beyond," *Journal of Chemometrics,* vol. 17, no. 8-9, pp. 480-502, 2003.

[6] B. Wise and N. Gallagher, "The process chemometrics approach to process monitoring and fault detection," *Journal of Process Control,* vol. 6, no. 6, pp. 329-348, 1996.

[7] R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification. 2nd ed., New York: John Wiley & Sons,

Inc., 2001.

[8] P. Nomikos and J. MacGregor, "Monitoring of batch processes using multi-way principal component analysis," *AIChE Journal,* vol. 40, no. 8, pp. 1361-1375, 1994.

[9] X. B. He, W. Wang, Y. P. Yang and Y. H. Yang, "Variable-weighted Fisher discriminant analysis for process fault diagnosis," *Journal of Process Control,* vol. 19, no. 6, pp. 923-931, 2009.

[10] L. H. Chiang, M. E. Kotanchek and A. K. Kordon, "Fault diagnosis based on Fisher discriminant analysis and support vector machines," *Computers & Chemical Engineering,* vol. 28, no. 8, pp. 1389-1401, 2004.

[11] W. Ku, R. H. Storer and C. Georgakis, "Disturbance detection and isolation by dynamic principal component analysis," *Chemometrics and Intelligent Laboratory Systems,* vol. 30, no. 1, pp. 179-196, 1995.

[12] W. E. Larimore, "Canonical variate analysis in control and signal processing," in *Statistical Methods in Control & Signal Processing*, New York, Marcel Dekker, Inc., 1997, pp. 83-120.

[13] A. Simoglou, E. B. Martin and A. J. Morris, "Statistical performance monitoring of dynamic multivariate processes using state space modelling," *Computers & Chemical Engineering,* vol. 26, no. 6, pp. 909-920, 2002.

[14] A. Negiz and A. Çinar, "Statistical monitoring of multivariable dynamic processes with state-space models," *AIChE Journal,* vol. 43, no. 8, pp. 2002-2020, 1997.

[15] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—with application to face recognition," *Pattern Recognition,* vol. 34, no. 10, pp. 2067-2070, 2001.

[16] W. E. Larimore, "Statistical optimality and canonical variate analysis system identification," *Signal Processing,* vol. 52, no. 2, pp. 131-144, 1996.

[17] T. Sun, S. Chen, J. Yang and P. Shi, "A novel method of combined feature extraction for recognition," in *Proceedings of the Eighth IEEE International Conference on Data Mining*, Pisa, Italy, 2008.

[18] M. Kan, S. Shan, H. Zhang, S. Lao and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 38, no. 1, pp. 188-194, 2016.

[19] T.-K. Kim, J. Kittler and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, pp. 1005-1018, 2007.

[20] S. Sun, X. Xie and M. Yang, "Multiview uncorrelated discriminant analysis," *IEEE Transactions on Cybernetics,* vol. 46, no. 12, pp. 3272-3284, 2016.

[21] K. McClure, R. B. Gopaluni, T. Chmelyk, D. Marshman and S. L. Shah, "Nonlinear process monitoring using supervised locally linear embedding projection," *Industrial & Engineering Chemistry Research,*

vol. 53, no. 13, pp. 5205-5216, 2014.

[22] X. He and P. Niyogi, "Locality preserving projections," in *Proceedings of the Advances in Neural Information Processing Systems*, 2004.

[23] T. Sun and S. Chen, "Locality preserving CCA with applications to data visualization and pose estimation," *Image and Vision Computing,* vol. 25, no. 5, pp. 531-543, 2007.

[24] Y. Yuan, C. Ma and D. Pu, "A novel discriminant minimum class locality preserving canonical correlation analysis and its applications," *Journal of Industrial & Management Optimization,* vol. 12, no. 1, pp. 251-268, 2016.

[25] M. Van and H.-J. Kang, "Wavelet kernel local Fisher discriminant analysis with particle swarm optimization algorithm for bearing defect classification," *IEEE Transactions on Instrumentation and Measurement,* vol. 64, no. 12, pp. 3588-3600, 2015.

[26] J. Yu, "Localized Fisher discriminant analysis based complex chemical process monitoring," *AIChE Journal,* vol. 57, no. 7, pp. 1817-1828, 2011.

[27] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *Journal of Machine Learning Research,* vol. 8, no. 5, pp. 1027-1061, 2007.

[28] H. Hotelling, "Relations between two sets of variates," *Biometrika,* vol. 28, no. 3/4, pp. 321-377, 1936.

[29] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control,* vol. 19, no. 6, pp. 716-723, 1974.

[30] 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 and Intelligent Laboratory Systems,* vol. 50, no. 2, pp. 243-252, 2000.

[31] S. Joe Qin and Y. Zheng, "Quality-relevant and process-relevant fault monitoring with concurrent projection to latent structures," *AIChE,* vol. 59, no. 1, pp. 496-504, 2013.

[32] G. Li, B. Liu, S. Joe Qin and D. Zhou, "Quality relevant data-driven modeling and monitoring of multivariate dynamic processes: Dynamic T-PLS approach," *IEEE Transactions on Neural Networks,* vol. 22, no. 12, pp. 2262-2271, 2011.

[33] B. Jiang, D. Huang, X. Zhu, F. Yang and R. D. Braatz, "Canonical variate analysis-based contributions for fault identification," *Jounral of Process Control,* vol. 26, pp. 17-25, 2015.

[34] R. Dunia, S. Joe Qin, T. Edgar and T. McAvoy, "Identification of faulty sensors using principal component analysis," *AIChE Journal,* vol. 42, no. 10, pp. 2797-2812, 1996.

[35] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis Part III: Process history based methods," C*omputers and Chemical Engineering*, vol. 27, pp. 327-334, 2003.

[36] D. M. Himmelblau, *Fault Detection and Diagnosis in Chemical and Petrochemical Processes*, Elsevier Scientific Pub. Co., 1978.