A Sparsity-Aware Fault Diagnosis Framework Focusing on Accurate Isolation

Xianchao Xiu, *Member, IEEE*, Zhonghua Miao, and Wanquan Liu, *Senior Member, IEEE*

***Abstract*—In this article, we propose an efficient fault diagnosis framework to achieve accurate fault isolation. The core is to introduce the $\ell_{2,0}$-norm sparsity constrained optimization to reduce the variable redundancy and determine the variable number, which is different from the existing sparse variants. In order to illustrate the idea, this article takes principal component analysis (PCA) as an essential step. First, a sparsity-aware PCA is constructed by taking advantage of the $\ell_{2,0}$-norm constrained optimization. Afterward, a two-stage monitoring strategy is designed, including fault detection and fault isolation. Once the fault is detected, the sparsity level is then shrunk to achieve accurate fault isolation. Moreover, an alternating direction method of multipliers-based optimization algorithm is developed with detailed implementation. Finally, the detection improvement and accurate isolation performance are validated by two simulated examples, the Tennessee Eastman benchmark process, and a practical cylinder-piston process.**

***Index Terms*—$\ell_{2,0}$-norm, fault diagnosis (FD), optimization algorithm, two-stage principal component analysis (TSPCA).**

NOMENCLATURE

N	NUMBER of samples.
p	Number of variables.
r	Number of principal components.
s	Number of sparsity level.
k	Number of iterations.
$\mathbb{R}^p$	Set of p-dimensional vectors.
$\mathbb{R}^{N\times p}$	Set of $N \times p$ matrices.
$\mathbf{L}$	Graph Laplace matrix.
$\mathbf{I}$	Identity matrix.
$\mathbf{A}$	Loading matrix.
$\mathbf{B}$	Weight matrix.
$\mathbf{X}$	Matrix of process data.
$\mathbf{x}_i$	ith row of matrix $\mathbf{X}$.
$\mathbf{X}^\top$	Transpose of matrix $\mathbf{X}$.
$\mathbf{X}^{-1}$	Inverse of matrix $\mathbf{X}$.
$\mathrm{Tr}(\mathbf{X})$	Trace of matrix $\mathbf{X}$.
ℓ_1	Sum of absolute values.
$\ell_{2,1}$	Sum of least squares of rows.
$\ell_{2,0}$	Number of nonzero rows.
$\mathrm{prob}(e)$	Probability of stochastic event e.
$\mathcal{N}(\mu, \Sigma)$	Normal distribution with mean μ and covariance Σ.

Manuscript received 18 November 2021; revised 1 May 2022; accepted 31 May 2022. Date of publication 3 June 2022; date of current version 13 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 12001019. Paper no. TII-21-5103. *(Corresponding author: Wanquan Liu.)*

Xianchao Xiu and Zhonghua Miao are with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China (e-mail: xcxiu@shu.edu.cn; zhhmiao@shu.edu.cn).

Wanquan Liu is with the School of Intelligent Systems Engineering, Sun Yat-Sen University, Guangzhou 510275, China (e-mail: liuwq63@mail.sysu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2022.3180070.

Digital Object Identifier 10.1109/TII.2022.3180070

I. INTRODUCTION

Fault diagnosis (FD) has received extensive attention to satisfy product quality, efficiency maintenance, and safety levels in industrial engineering. During the past few decades, model-based FD approaches have been dominant because physical or chemical laws are usually available [1]. However, as modern industrial processes turn out to be more complex, it becomes extremely difficult to construct precise mathematical models [2]. With the development of sensor technology and computing capability, data-driven FD approaches have gained great success from both academia and industry. Unlike model-based FD approaches, data-driven FD approaches solely rely on process data, which makes them more applicable for complex industrial processes without physical or chemical priors. Popular multivariate statistical analysis-based FD approaches include principal component analysis (PCA), partial least squares, and canonical correlation analysis (see [3]–[8]). Generally speaking, FD contains fault detection and fault isolation. In the procedure of fault detection, T^2 statistic and squared prediction error (SPE) statistic are first defined, and then the corresponding control limits are determined. If the detection statistic violates the control limit, a fault is, thus, detected, otherwise, no fault occurs. Afterward, fault isolation will be performed according to the contribution-based plot. In comparison to fault detection, fault isolation is often more difficult and important. Incorrect isolation will seriously affect industrial processes, and even bring huge economic losses. Therefore, it comes the topic: how to achieve good detection as well as fulfill accurate isolation.

Now let us take PCA as an essential step. As a fundamental FD approach, PCA has illustrated its high efficiency due to the simple formulation and easy realization. Specifically, PCA seeks the maximum variance using orthogonal transformations to extract only a few principal components (PCs) to represent the original process data. However, interpreting the extracted PCs is still a challenging task [9]. The reason behind is that not

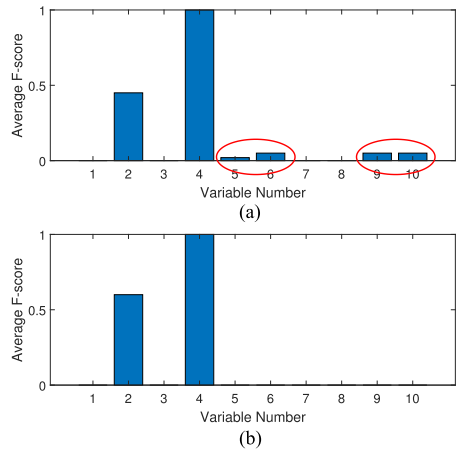

Fig. 1. Illustration of isolation performance obtained by (a) existing $\ell_{2,1}$-norm regularized PCA; (b) proposed approach. The process has a total of ten variables, and the average F-score is defined in Section IV-B.

all the PCs are significant for describing the fault, which may lead to a misunderstanding of industrial processes. To improve the interpretation of PCA-based FD, many researchers have proposed numerous strategies. Motivated by Lasso [10] and compressed sensing [11], Liu *et al.* [12] constructed a sparse variant of PCA called SPCA by imposing the $\ell_{2,1}$-norm joint sparse regularizer onto the PCA objective function. The obtained weight matrix becomes sparse, and thus, some fault-free process variables can be successfully discarded from the model. This in turn makes fault detection robust for outliers and improves the performance of fault isolation. In fact, the results are consistent with feature selection in computer vision [13]. Although SPCA enjoys a good performance, the variable structure relationship of process variables has not been fully explored. Followed by the inspiration of graph learning [14], Liu *et al.* [15] introduced a graph Laplace regularizer onto the SPCA objective and proposed the structured joint sparse PCA (SJSPCA), has the ability of capturing the cause-effect relationship among process variables. Furthermore, Zhai *et al.* [16] considered a Laplace sparse PCA (LSPCA) model by replacing the $\ell_{2,1}$-norm regularizer with the elastic net regularizer [17], thereby improving the fault variable interpretation.

Although the $\ell_{2,1}$-norm regularized PCA helps us to facilitate the extraction of intrinsic fault variables, it is quite difficult to draw conclusions about the exact relationship between regularized parameters and the number of fault process variables [18]. Additionally, there often generate small residuals [see Fig. 1(a)]. Naturally, it will be much better if the significant variables can be accurately extracted and the residuals can be discarded [19]. Thanks to the development of sparse optimization learning, the $\ell_{2,0}$-norm sparsity constrained optimization has been proposed and demonstrated the efficiency for image processing [20] and fault detection [21]. Compared with the previous $\ell_{2,1}$-norm regularizer, the advantages areas follows: 1) easier to control variables [22], and 2) faster to solve problems [23]. Unfortunately, its application to fault isolation has never been investigated yet. Considering the necessity of accurate fault isolation, a natural question is whether the $\ell_{2,0}$-norm sparsity constrained optimization can be used for PCA.

Inspired by the aforementioned observations, we develop an efficient FD framework by integrating the $\ell_{2,0}$-norm sparsity constrained optimization with PCA to achieve a good detection and accurate isolation. In the first stage, a sparse PCA variant is constructed for fault detection. In the second stage, through shrinking the sparsity level, fault isolation is performed. After integrating the $\ell_{2,0}$-norm optimization, the proposed approach not only can remove these small residuals in Fig. 1(a) but also can isolate accurate fault variables. For example, if two variables need to be isolated, just set the sparsity level $s = 2$ [see Fig. 1(b)].

Compared to previous work, the contributions of this article can be summarized as follows.

1) It introduces a new $\ell_{2,0}$-norm constrained PCA model for good detection and accurate isolation.
2) It provides a two-stage detection and isolation strategy by shrinking the sparsity level.
3) It develops an efficient iterative optimization algorithm to solve the proposed model.
4) It verifies the superiority or comparability on the simulated and benchmark processes.

II. Preliminary and Problem Formulation

A. Basics of SPCA

Let $\mathbf{X} \in \mathbb{R}^{N\times p}$ be the process data with N samples and p variables. The mathematical model of PCA [24] can be

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}^\top\|^2 \quad \text{s.t.} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{p\times r}$ $(r \le p)$ is the loading matrix. To reduce the negative effect of outliers and improve the interpretation of PCA-based FD approaches, SPCA [12] is proposed as follows:

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|^2 + \mu\|\mathbf{B}\|_{2,1} \quad \text{s.t.} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I} \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{p\times r}$ is the weight matrix, $\|\mathbf{B}\|_{2,1}$ is the defined $\ell_{2,1}$-norm, and $\mu \ge 0$ is the parameter to balance the sparsity. By enforcing $\|\mathbf{B}\|_{2,1}$ and choosing different μ, SPCA can filter out some unimportant or useless variables, which improves the ability of fault interpretation. To preserve the graph structure prior, SJSPCA [15] considers the following minimization:

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|^2 + \mu\|\mathbf{B}\|_{2,1} + \lambda\mathrm{Tr}(\mathbf{B}^\top\mathbf{L}\mathbf{B}) \quad \text{s.t.} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I} \tag{3}$$

where $\mathbf{L} \in \mathbb{R}^{p\times p}$ is the graph Laplace matrix and $\mathrm{Tr}(\mathbf{B}^\top\mathbf{L}\mathbf{B})$ is the added term for structure preserving, because if samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are close in the data space, then their projections $\mathbf{b}_i$ and $\mathbf{b}_j$ are also close to each other in the feature space [25]. Later, LSPCA [16] is designed to facilitate a sparser solution by linking the elastic net regularizer with PCA as the form of

$$\min_{\mathbf{A},\mathbf{B}} \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|^2 + \mu\Theta(\mathbf{B}) + \lambda\mathrm{Tr}(\mathbf{B}^\top\mathbf{L}\mathbf{B}) \quad \text{s.t.} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I} \tag{4}$$

where $\Theta(\mathbf{B}) = \|\mathbf{B}\|_{2,1} + \kappa\|\mathbf{B}\|_1$ is the elastic net regularizer, which is a weighted sum of joint sparsity and Lasso, and $\kappa \geq 0$ is the weighted parameter.

B. Problem Formulation

Although the abovementioned PCA-based FD approaches may bring positive impact on fault detection, accurate fault isolation still cannot be achieved. Therefore, in this article, we construct a new PCA-based optimization model as

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B}} \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|^2 + \lambda \mathrm{Tr}(\mathbf{B}^\top\mathbf{L}\mathbf{B}) \\ &\text{s.t.} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I},\ \|\mathbf{B}\|_{2,0} \leq s. \end{aligned} \tag{5}$$

Here, s is the sparsity level, which can be tuned in different scenarios. $\|\mathbf{B}\|_{2,0}$ is the $\ell_{2,0}$-norm, which acts as $\|\mathbf{B}\|_{2,1}$ in models (2) and (3) and $\Theta(\mathbf{B})$ in model (4). Obviously, it can be regarded a variant of SJSPCA and LSPCA.

Indeed, it is the first work to introduce the $\ell_{2,0}$-norm sparsity constrained optimization into a PCA framework and attempt to realize accurate isolation. For instance, if only the most significant variable needs to be exploited, choose $s = 1$. By shrinking the sparsity level s, the proposed approach in (5) can retain at most s PCs and set the others to be zero, while these approaches in models (2)–(4) may generate residuals.

III. Optimization Algorithm

Although there exist some optimization algorithms for solving sparse problems, such as semidefinite programming and interior point methods, they often take substantial time for high-dimensional data [26]. Recently, alternating direction method of multipliers (ADMM), as a popular first-order optimization algorithm, has achieved widespread success in machine learning and industrial engineering. We refer the interested readers to [27], [28] and references therein. Moreover, the ADMM has shown its high efficiency in SPCA-based FD approaches [12], [15], [16].

However, due to the constraints $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$ and $\|\mathbf{B}\|_{2,0} \leq s$, it cannot be solved by directly applying the ADMM. Therefore, an iterative optimization algorithm based on the ADMM and the hard thresholding [29] will be developed to seek an approximate solution. First, an auxiliary variable $\mathbf{C} \in \mathbb{R}^{p\times r}$ is introduced to reformulate problem (5) as

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|^2 + \lambda \mathrm{Tr}(\mathbf{B}^\top\mathbf{L}\mathbf{B}) \\ &\text{s.t.} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I},\ \|\mathbf{C}\|_{2,0} \leq s,\ \mathbf{B} = \mathbf{C}. \end{aligned} \tag{6}$$

Note that, when the constraint $\mathbf{B} = \mathbf{C}$ is forced to be satisfied in the optimization, problem (6) will be simplified to problem (5). Therefore, this section only needs to consider how to solve the problem (6) in detail.

The augmented Lagrangian function associated with the abovementioned problem (6) can be defined as

$$\begin{aligned} \mathcal{L}_\beta(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}) = &\|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{A}^\top\|^2 + \lambda \mathrm{Tr}(\mathbf{B}^\top\mathbf{L}\mathbf{B}) \\ &- \langle \mathbf{D}, \mathbf{B} - \mathbf{C}\rangle + \frac{\beta}{2}\|\mathbf{B} - \mathbf{C}\|^2 \end{aligned} \tag{7}$$

where $\mathbf{D} \in \mathbb{R}^{p\times r}$ is the introduced Lagrangian dual multiplier, $\langle\cdot,\cdot\rangle$ is the inner product, and $\beta > 0$ is the penalty parameter to measure the distance between $\mathbf{B}$ and $\mathbf{C}$. According to the ADMM scheme, an approximate solution can be obtained by minimizing one variable with others fixed in a Gauss–Seidel manner as follows:

$$\begin{cases} \mathbf{A}^{k+1} = \underset{\mathbf{A}\in\mathcal{M}_1}{\mathrm{argmin}} \mathcal{L}_\beta(\mathbf{A}, \mathbf{B}^{\mathbf{k}}, \mathbf{C}^{\mathbf{k}}, \mathbf{D}^{\mathbf{k}}) \\ \mathbf{B}^{k+1} = \underset{\mathbf{B}}{\mathrm{argmin}}\ \mathcal{L}_\beta(\mathbf{A}^{k+1}, \mathbf{B}, \mathbf{C}^k, \mathbf{D}^k) \\ \mathbf{C}^{k+1} = \underset{\mathbf{C}\in\mathcal{M}_2}{\mathrm{argmin}}\ \mathcal{L}_\beta(\mathbf{A}^{k+1}, \mathbf{B}^{k+1}, \mathbf{C}, \mathbf{D}^k) \\ \mathbf{D}^{k+1} = \mathbf{D}^k - \beta(\mathbf{B}^{k+1} - \mathbf{C}^{k+1}) \end{cases} \tag{8}$$

where $\mathcal{M}_1 = \{\mathbf{A} \mid \mathbf{A}^\top\mathbf{A} = \mathbf{I}\}$ and $\mathcal{M}_2 = \{\mathbf{C} \mid \|\mathbf{C}\|_{2,0} \leq s\}$. Next, the resulting subproblems will be analyzed one by one.

1) *Update* $\mathbf{A}$*:* When the variables $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are fixed, the variable $\mathbf{A}^{k+1}$ can be updated by solving

$$\begin{aligned} &\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{X}\mathbf{B}^k\mathbf{A}^\top\|^2 \\ &\text{s.t.} \quad \mathbf{A}^\top\mathbf{A} = \mathbf{I}. \end{aligned} \tag{9}$$

The objective is convex and quadratic in terms of $\mathbf{A}$, and can be expanded into

$$\mathrm{Tr}(\mathbf{X}^\top\mathbf{X}) - 2\mathrm{Tr}(\mathbf{X}^\top\mathbf{X}\mathbf{B}^k\mathbf{A}^\top) + \mathrm{Tr}((\mathbf{B}^k)^\top\mathbf{X}^\top\mathbf{X}\mathbf{B}^k). \tag{10}$$

It is found that the first item and the last item are both not related with $\mathbf{A}$, then problem (10) only needs to maximize $\mathrm{Tr}(\mathbf{X}^\top\mathbf{X}\mathbf{B}^k\mathbf{A}^\top)$. Denote $\mathbf{X}^\top\mathbf{X}\mathbf{B}^k = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$, then the middle item can be rewritten as

$$\mathrm{Tr}(\mathbf{X}^\top\mathbf{X}\mathbf{B}^k\mathbf{A}^\top) = \mathrm{Tr}(\mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{B}^k) = \mathrm{Tr}(\hat{\mathbf{A}}^\top\mathbf{U}\mathbf{\Lambda}) \tag{11}$$

where $\hat{\mathbf{A}} = \mathbf{A}\mathbf{V}$. Together with the fact that $\mathbf{V}$ is orthonormal, it derives that $\hat{\mathbf{A}}^\top\hat{\mathbf{A}} = \mathbf{I}$. Hence, the maximum can be achieved when the diagonal of $\hat{\mathbf{A}}^\top\mathbf{U}$ is positive and maximized [24]. Following the Cauchy–Schwartz inequality, it implies that the optimal solution is obtained when $\hat{\mathbf{A}} = \mathbf{U}$. Therefore, problem (9) can be solved by a closed-form solution as

$$\mathbf{A}^{k+1} = \mathbf{U}\mathbf{V}^\top. \tag{12}$$

2) *Update* $\mathbf{B}$*:* When the variables $\mathbf{A}$, $\mathbf{C}$, and $\mathbf{D}$ are fixed, the variable $\mathbf{B}^{k+1}$ can be updated by minimizing

$$\begin{aligned} \min_{\mathbf{B}} &\|\mathbf{X} - \mathbf{X}\mathbf{B}(\mathbf{A}^{k+1})^\top\|^2 + \lambda \mathrm{Tr}(\mathbf{B}^\top\mathbf{L}\mathbf{B}) \\ &+ \frac{\beta}{2}\|\mathbf{B} - \mathbf{C}^k - \mathbf{D}^k/\beta\|^2. \end{aligned} \tag{13}$$

According to the first-order optimality condition, the solution can be given by

$$\begin{aligned} \mathbf{B}^{k+1} = &(2\mathbf{X}^\top\mathbf{X} + 2\lambda\mathbf{L} + \beta\mathbf{I})^{-1} \\ &\times (2\mathbf{X}^\top\mathbf{X}\mathbf{A}^{k+1} + \beta\mathbf{C}^k + \mathbf{D}^k). \end{aligned} \tag{14}$$

Notice that the coefficient matrix is positive definite because both λ and β are positive. In the process of optimization, the inverse only requires to be calculated once.

Algorithm 1: Optimization Algorithm.

Input: Process data $\mathbf{X}$, parameters λ, β, sparsity level s, and compute Laplace matrix $\mathbf{L}$.
Output: $(\mathbf{A}^{k+1}, \mathbf{B}^{k+1})$.
While not converged **do**
1: Update $\mathbf{A}^{k+1}$ by (12);
2: Update $\mathbf{B}^{k+1}$ by (14);
3: Update $\mathbf{C}^{k+1}$ by (16);
4: Update $\mathbf{D}^{k+1}$ by (17);
5: Check convergence.
End while

3) *Update* $\mathbf{C}$*:* When the variables $\mathbf{A}, \mathbf{B}$, and $\mathbf{D}$ are fixed, the variable $\mathbf{C}^{k+1}$ can be updated by optimizing

$$\min_{\mathbf{C}} \|\mathbf{B}^{k+1} - \mathbf{C} - \mathbf{D}^k/\beta\|^2 \quad \text{s.t.} \quad \|\mathbf{C}\|_{2,0} \leq s. \tag{15}$$

According to the hard thresholding [29], a projected gradient descent method can be applied to solve the abovementioned problem (15), which is described as

$$\mathbf{C}^{k+1} = \Pi_{\mathcal{M}_2}(\mathbf{B}^{k+1} - \mathbf{D}^k/\beta) \tag{16}$$

where $\Pi_{\mathcal{M}_2}(\cdot)$ is a projection onto $\mathcal{M}_2$, which means that if $\|\mathbf{B}^{k+1} - \mathbf{D}^k/\beta\|_{2,0} \leq s$, then $\mathbf{C}^{k+1} = \mathbf{B}^{k+1} - \mathbf{D}^k/\beta$, otherwise truncate $\mathbf{B}^{k+1} - \mathbf{D}^k/\beta$ with top s (in row-level) entries preserved.

4) Update $\mathbf{D}$*:* When the variables $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ are fixed, the dual variable $\mathbf{D}^{k+1}$ can be updated by computing

$$\mathbf{D}^{k+1} = \mathbf{D}^k - \beta(\mathbf{B}^{k+1} - \mathbf{C}^{k+1}). \tag{17}$$

To sum up, all the steps for solving the proposed model (5) are provided in Algorithm 1, whose convergence criterion is determined by checking the relative differences. On the one hand, since each subproblem can be solved by fast solvers, the speed of the algorithm is guaranteed. On the other hand, from the perspective of optimization theory [30], it can be concluded that the sequence generated by Algorithm 1 converges to a local minimum of problem (6) [equivalent to (5)]. This shows that the proposed algorithm is both fast and convergent.

IV. Monitoring Framework

This section develops an efficient FD framework based on the proposed approach, including the following two stages.

1) Perform fault detection using residual generators.
2) Do fault isolation by shrinking the sparsity level.

A. Stage I: Fault Detection

According to [2] and [31], the proposed approach can separate the process data into a PC subspace and a residual subspace, where the PC subspace preserves most of the process information and the residual subspace keeps the remaining process information. Once the proposed model (5) is solved, the weight matrix $\mathbf{B}$ is obtained. Based on the abovementioned arguments, it can be rewritten as $[\mathbf{B}_r, \mathbf{B}_d]$, where $\mathbf{B}_r, \mathbf{B}_d$ denote the vectors of the retained PCs and the discarded PCs, respectively.

For the purpose of fault detection, two classical statistical monitoring metrics [32], i.e., T^2 test statistic and SPE test statistic, are defined as

$$\begin{aligned} \mathrm{T}^2 &= \mathbf{x}^\top \mathbf{B}_r \mathbf{\Lambda}^{-1} \mathbf{B}_r^\top \mathbf{x} \\ \mathrm{SPE} &= \mathbf{x}^\top (\mathbf{I} - \mathbf{B}_r \mathbf{B}_r^\top) \mathbf{x} \end{aligned} \tag{18}$$

where $\mathbf{x} \in \mathbb{R}^p$ is the test sample, and $\mathbf{\Lambda}$ is the variance matrix defined as $\mathbf{\Lambda} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2)$. According to [33], the control limit for the abovementioned T^2 test statistic is related to the F distribution and can be estimated by

$$J_{\mathrm{th},\mathrm{T}^2} = \frac{r(n^2-1)}{n(n-1)} F_\alpha(r, n-r) \tag{19}$$

where the F distribution depends on the degrees of freedom $r, n-r$ and the significance level α. The control limit for the SPE test statistic [34] can be determined by

$$J_{\mathrm{th,SPE}} = \theta_1 \left(\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right)^{1/h_0} \tag{20}$$

where c_α is the normal deviate corresponding to the upper $1-\alpha$ percentile and

$$\theta_i = \sum_{j=r+1}^{p} (\sigma_j^2)^i \ (i = 1, 2, 3), \ h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2}. \tag{21}$$

Once the test statistics and the corresponding control limits are determined, the fault detection logic can be checked as

$$\begin{cases} \mathrm{T}^2 > J_{\mathrm{th},\mathrm{T}^2} \text{ and } \mathrm{SPE} > J_{\mathrm{th,SPE}} \ \Rightarrow \ \text{faulty} \\ \mathrm{T}^2 \leq J_{\mathrm{th},\mathrm{T}^2} \text{ or } \mathrm{SPE} \leq J_{\mathrm{th,SPE}} \ \Rightarrow \ \text{fault-free} \end{cases}. \tag{22}$$

The detection performance is commonly measured by the fault detection rate (FDR) and the false alarm rate (FAR). See [2] for detailed information.

B. Stage II: Fault Isolation

For the existing SPCA-based variants, the sparsity regularization parameter μ in models (2)–(4) should be adjusted carefully in order to extract different fault variables. To be honest, it is not a trivial task to distinguish the exact relationship between the parameter μ and the accurate sparsity level, which may make fault isolation a bit tricky. However, for the proposed approach, fault isolation becomes relatively easy to implement, because the existence of $\|\mathbf{B}\|_{2,0} \leq s$ enforces the number of extracted variables to reach s. This means that the extracted variables can be accurately isolated by tuning the sparsity level s. Therefore, in real-world industrial processes, it can be recommended to determine s according to the actual demand. In other words, if one needs to isolate at most s significant variables, just set the sparsity level to be s.

In addition, to isolate the fault variables, the following F-score function can be defined as

$$\delta_i = \|\mathbf{b}_i\|_1, \ i = 1, \ldots, p \tag{23}$$

Algorithm 2: Monitoring Framework.

Stage I: Fault detection

1: Collect and normalize the process data $\mathbf{X}$;
2: Compute the graph Laplace matrix $\mathbf{L}$;
3: Solve the proposed approach via Algorithm 1, and obtain the weight matrix $\mathbf{B}$;
4: Estimate the control limits J_{th,T^2}, $J_{th,\mathrm{SPE}}$ according to (19)–(21);
5: Collect and normalize the test sample $\mathbf{x}$;
6: Generate the T^2 test statistic and SPE test statistic according to (18);
7: Compare these test statistics with the corresponding control limits J_{th,T^2}, $J_{th,\mathrm{SPE}}$;
8: Check the decision logic (22);

Stage II: Fault isolation

1: Calculate the average F-score values according to (23);
2: Determine and analyze the fault variables;
3: Next sample.

where $\mathbf{b}_i$ is the ith row of matrix $\mathbf{B}$. Compared with the existing contribution-plot methods [35], the abovementioned F-score is just derived from the weight matrix. Of course, for the convenience of visualization, the F-score values are normalized, called average F-score here. In this sense, a high F-score value reflects that the extracted variable is likely to be faulty with high probability.

In the end, the detailed two-stage monitoring framework of the proposed approach can be summarized in Algorithm 2.

V. Simulation Examples

A. Data Generation

In this section, a total of ten process variables are simulated, where the first four variables are only associated with the first hidden variable h_1, the middle four variables are only associated with the second hidden variable h_2, and the last two variables are linear combinations of the first hidden variable h_1 and the second hidden variable h_2. To be specific, the process data are generated according to the following relations:

$$\begin{aligned}\mathbf{x} &= \mathbf{S}\mathbf{h} + \mathbf{e}\\ \mathbf{S} &= \begin{bmatrix}1&1&1&1&0&0&0&0&-0.6&-0.6\\0&0&0&0&1&1&1&1&0.8&0.8\end{bmatrix}^\top\\ \mathbf{h} &= \begin{bmatrix}h_1\\h_2\end{bmatrix} \sim \mathcal{N}\left\{\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.98&0\\0&1\end{bmatrix}\right\}\end{aligned} \tag{24}$$

where $\mathbf{h}$ obeys the multivariate normal distribution, and $\mathbf{e}$ is the introduced Gaussian noise with mean 0 and variance $\mathbf{\Sigma} = \mathrm{diag}\left([0.09\,0.09\,0.09\,0.09\,0.16\,0.16\,0.16\,0.16\,0.25\,0.25]\right)$.

In order to validate PCA-based FD approaches, 500 samples are first generated for offline training, and 300 samples are then generated for online testing. Finally, two types of faults are introduced into the testing process data, which are described as follows.

TABLE I
Quantitative Comparisons for Type I and Type II

Approaches	Type I		Type II	
	$\mathrm{FDR}_{\mathrm{SPE}}$	$\mathrm{FAR}_{\mathrm{SPE}}$	$\mathrm{FDR}_{\mathrm{T}^2}$	$\mathrm{FAR}_{\mathrm{T}^2}$
PCA	86.67%	0.67%	21.00%	1.00%
SPCA	90.67%	0.67%	24.50%	0.00%
SJSPCA	93.33%	0.67%	27.50%	0.00%
LSPCA	95.67%	0.67%	29.00%	0.00%
TSPCA	**98.67%**	**0.67%**	**32.50%**	**0.00%**

1) Type I: Two sensor biases are introduced to variables x_2 and x_4 at the 151th sample, i.e.,

$$\mathbf{x}_f = \mathbf{S}\mathbf{h} + \mathbf{e} + [0\,1.2\,0\,1.8\,0\,0\,0\,0\,0\,0]^\top. \tag{25}$$

2) Type II: A step change is introduced to hidden variable h_1 at the 101th sample, i.e.,

$$\mathbf{x}_f = \mathbf{S}(\mathbf{h} + [2\,0]^\top) + \mathbf{e}. \tag{26}$$

B. Parameter Settings

For the sake of simplicity, the proposed approach is referred to as TSPCA for short. In this article, it will be compared with some state-of-the-art approaches, including PCA [2], SPCA [12], SJSPCA [15], and LSPCA [16]. For all the mentioned PCA-based approaches, the first two PCs can extract at least 85% variance, thus, the retained number can be determined to be 2. In particular, for the $\ell_{2,1}$-norm regularized approaches, i.e., SPCA, SJSPCA, and LSPCA, the sparsity parameter μ is chosen using five-fold cross-validation and fixed in the optimization procedure. For the proposed TSPCA, the sparsity level parameter s should be chosen by a two-stage strategy. More discussion can be found in Section V-D.

In addition, the graph Laplace matrix $\mathbf{L}$ is constructed based on [15]. For process data $\mathbf{X}$, it can be described as

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \tag{27}$$

where $\mathbf{D}$ is the diagonal matrix, whose entries are composed by the sum of each row of matrix $\mathbf{W}$, and $\mathbf{W}$ is the weight matrix, whose ijth entry is chosen to be $e^{-\|\mathbf{x}_i-\mathbf{x}_j\|^2/\sigma}$ if $\mathbf{x}_i \in \mathrm{KNN}(\mathbf{x}_j)$ or $\mathbf{x}_j \in \mathrm{KNN}(\mathbf{x}_i)$, otherwise 0.

C. Detection and Isolation Performance

Type I simulates a ramp change, which will slightly affect the T^2 statistic, but greatly affect the SPE statistic. Therefore, only the quantitative comparisons related to the SPE statistic are listed in Table I. Moreover, the results of the proposed TSPCA are indicated in bold. By comparison with the existing PCA, SPCA, SJSPCA, and LSPCA, the proposed TSPCA obtains a higher $\mathrm{FDR}_{\mathrm{SPE}}$ value with an increase of at least 3%. Moreover, the corresponding detection performance is presented in Fig. 2. It can be found that all five approaches can successfully detect this fault at the 151st sample, which indicates that the PCA-based FD approaches can achieve the satisfactory detection performance for this fault. Nevertheless, TSPCA can detect most of the fault samples that violate the control limits after the fault is

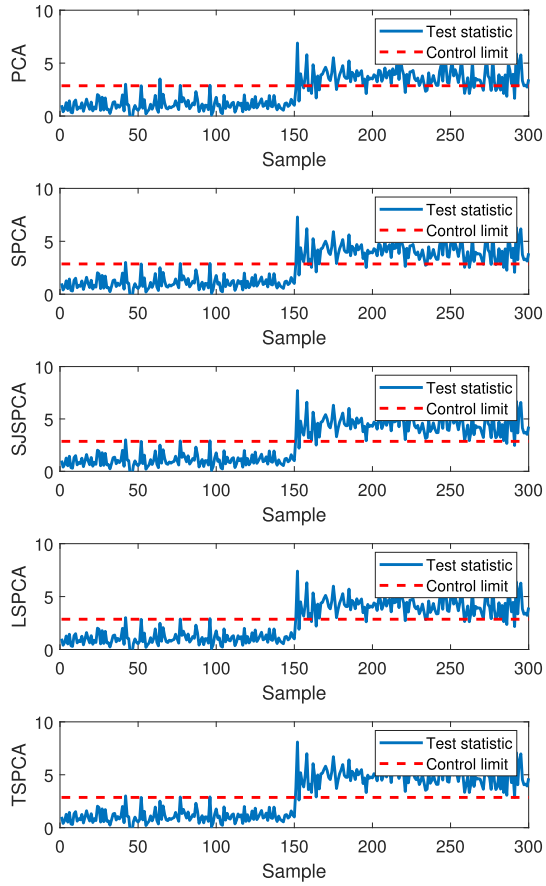

Fig. 2. Detection performance of the SPE statistic for Type I.

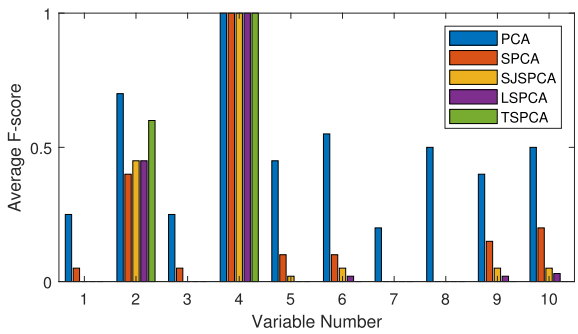

Fig. 3. Isolation results for Type I.

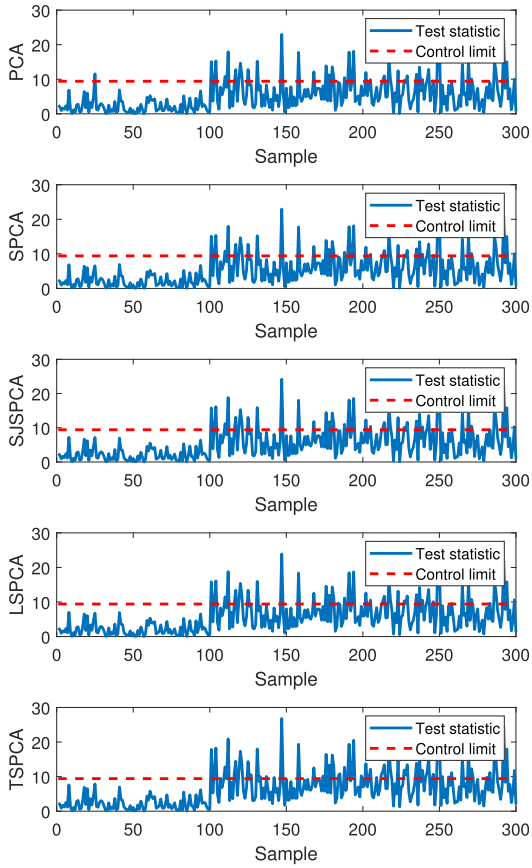

Fig. 4. Detection performance of the T^2 statistic for Type II.

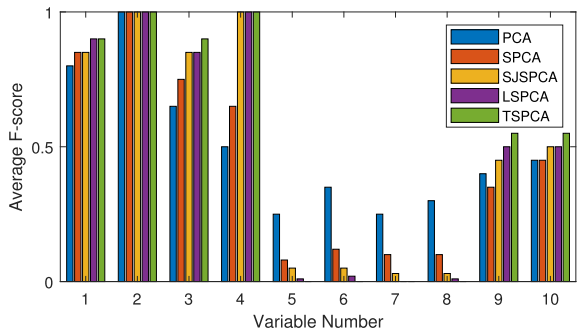

Fig. 5. Isolation results for Type II.

introduced, which verifies that the proposed TSPCA is effective for this fault.

To compare the isolation results, the average F-score values for Type I are shown in Fig. 3. Although SJSPCA and LSPCA can achieve good isolation results, there often generate small residuals, such as x_5, x_6, x_9, and x_{10}. After embedding the $\ell_{2,0}$-norm constrained optimization and setting the sparsity level $s = 2$, only x_2 and x_4 are retained, leading to an easier interpretation of the fault. In fact, this is consistent with the fact that two sensor biases of 1.2 and 1.8 are introduced to variables x_2 and x_4.

Type II simulates a step change, which will slightly affect the SPE statistic, but greatly affect the T^2 statistic. As a result, the quantitative comparisons corresponding to the T^2 statistic are also listed in Table I. A similar conclusion can be drawn that the proposed TSPCA has a better detection capability in terms of FDR_{T^2}. Furthermore, the detection performance is presented in Fig. 4. Although a fault is introduced at the 101st sample, all of them fail to detect it immediately. After a few samples, the fault is detected but very unstable. Obviously, the proposed TSPCA can detect more violated samples, especially for samples between 200 and 250.

Fig. 5 shows the average F-score values for Type II. Since the fault is introduced to the hidden variable h_1, thus, it will affect x_1, x_2, x_3, x_4, x_9, and x_{10}. It can be observed that the isolation results of TSPCA are more interpretable than other PCA-based approaches. This, together with the abovementioned detection performance, verifies that the proposed TSPCA is able to achieve good detection and accurate isolation.

TABLE II
QUANTITATIVE COMPARISONS IN TERMS OF FDR_{T^2} (FAR_{T^2}) AND FDR_{SPE} (FAR_{SPE})

Faults	PCA		SPCA		SJSPCA		LSPCA		TSPCA	
	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE
IDV(1)	99.13(0.00)	99.88(0.63)	99.25(0.00)	99.88(0.63)	99.25(0.00)	99.88(0.63)	99.25(0.00)	99.88(0.63)	**99.25(0.00)**	**100(0.63)**
IDV(2)	98.38(1.25)	95.75(0.63)	98.38(0.63)	97.75(0.63)	98.38(0.63)	99.00(0.63)	98.38(0.63)	99.25(0.63)	**98.38(0.63)**	**99.25(0.63)**
IDV(3)	0.88(0.00)	2.63(1.25)	1.75(0.00)	3.50(0.63)	2.25(0.00)	4.25(0.63)	3.75(0.00)	4.25(0.00)	**4.13(0.00)**	**5.38(0.00)**
IDV(4)	20.88(0.63)	100(1.25)	28.63(0.63)	100(0.63)	34.50(0.63)	100(0.63)	36.13(0.63)	100(0.00)	**38.75(0.63)**	**100(0.00)**
IDV(5)	24.13(0.63)	20.88(1.88)	28.75(0.00)	24.25(0.63)	30.25(0.00)	24.25(0.00)	30.25(0.00)	24.50(0.00)	**34.50(0.00)**	**24.50(0.00)**
IDV(6)	99.13(0.63)	100(1.25)	99.13(0.63)	100(1.25)	99.38(0.00)	100(1.25)	99.50(0.63)	100(1.25)	**99.50(0.00)**	**100(1.25)**
IDV(7)	100(0.00)	100(1.25)	100(0.00)	100(0.00)	100(0.00)	100(0.63)	100(0.00)	100(0.63)	**100(0.00)**	**100(0.63)**
IDV(8)	96.88(0.00)	83.63(0.63)	97.13(0.00)	91.50(0.63)	97.25(0.00)	96.75(0.63)	97.75(0.00)	96.75(0.63)	**98.50(0.00)**	**96.75(0.63)**
IDV(9)	1.75(1.88)	1.75(1.88)	2.25(1.25)	2.50(1.25)	1.75(0.63)	2.50(0.63)	2.25(0.63)	2.50(0.63)	**3.88(0.63)**	**2.75(0.00)**
IDV(10)	29.63(0.00)	25.13(0.63)	30.75(0.00)	28.00(0.00)	34.13(0.00)	30.38(0.00)	35.63(0.00)	30.75(0.00)	**37.50(0.00)**	**35.63(0.00)**
IDV(11)	40.63(0.63)	74.88(2.50)	45.13(0.63)	79.75(2.25)	48.38(0.63)	84.88(1.88)	49.50(0.63)	86.13(1.25)	**50.13(0.63)**	**88.50(1.25)**
IDV(12)	98.38(0.00)	89.50(1.25)	98.38(0.00)	89.75(0.63)	98.50(0.00)	90.75(0.63)	99.00(0.00)	90.75(0.63)	**99.00(0.00)**	**91.25(0.63)**
IDV(13)	93.63(0.63)	95.25(0.00)	93.63(0.63)	95.88(0.00)	93.63(0.63)	97.50(0.00)	93.63(0.63)	97.50(0.00)	**93.63(0.63)**	**98.63(0.00)**
IDV(14)	99.25(0.00)	100(1.25)	99.75(0.00)	100(1.25)	99.75(0.00)	100(0.63)	99.75(0.00)	100(0.63)	**99.75(0.00)**	**100(0.63)**
IDV(15)	1.38(0.00)	3.00(1.25)	1.63(0.00)	3.88(0.63)	2.25(0.00)	3.88(1.25)	2.63(0.00)	4.25(0.00)	**3.50(0.00)**	**6.88(0.00)**
IDV(16)	13.50(3.75)	27.38(1.88)	14.25(3.25)	33.63(1.88)	14.63(1.88)	39.50(1.88)	15.75(0.63)	39.50(1.88)	**15.88(0.63)**	**40.13(1.88)**
IDV(17)	76.25(1.25)	95.38(2.50)	77.13(1.25)	95.75(1.88)	83.50(1.25)	96.25(1.88)	85.38(1.25)	96.25(1.88)	**87.13(1.25)**	**96.25(1.25)**
IDV(18)	89.25(0.00)	90.13(2.50)	89.38(0.00)	91.25(1.25)	89.38(0.00)	92.50(1.25)	90.25(0.00)	92.50(1.63)	**90.63(0.00)**	**92.50(1.63)**
IDV(19)	14.13(0.00)	18.50(0.63)	16.50(0.00)	22.13(0.00)	16.25(0.00)	22.38(0.63)	17.25(0.00)	23.75(0.63)	**17.25(0.00)**	**24.88(0.00)**
IDV(20)	31.75(0.00)	49.75(1.25)	36.25(0.00)	55.63(1.25)	39.38(0.00)	57.63(0.63)	42.63(0.00)	57.63(0.63)	**48.50(0.00)**	**57.63(0.63)**
IDV(21)	39.25(0.00)	47.25(3.13)	39.50(0.00)	48.13(2.50)	44.63(0.00)	49.25(2.50)	44.63(0.00)	50.13(1.88)	**44.88(0.00)**	**52.50(1.88)**

Note that % is omitted for simplicity.

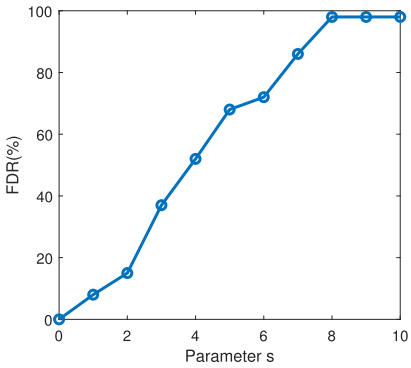

Fig. 6. Illustration of parameter selecting.

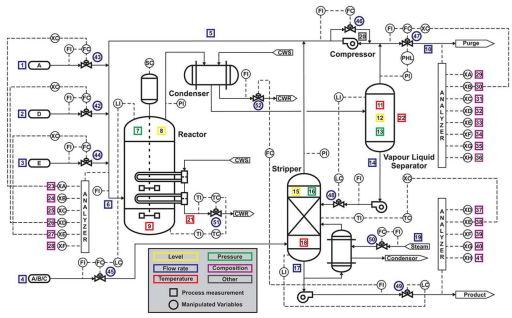

Fig. 7. Detailed flowchart of the TE process.

D. Discussion

For the proposed TSPCA, the sparsity level parameter s has great effects on the detection and isolation performance. If it is too large, the constraint will be meaningless, that is, all variables are extracted, which does not contribute to the fault interpretation. If it is too small, the constraint will be very strict, that is, only a few variables are useful, which makes fault detection performance poor. Therefore, in the detection stage, one can possibly initialize the algorithm with a small s, and then increase it by a constant ratio until the best detection performance is obtained. For example, the FDR values of Type I for different s are plotted in Fig. 6. It can be seen that if s is too small, FDR is less than 20%. As s increases, FDR will become larger. When $s = 8$, it reaches the maximum value and then remains unchanged. In this sense, $s = 8$ can be chosen for Type I.

However, in the isolation stage, the sparsity level parameter s should be shrunk by the desired fault variables. For example, if two fault variables need to be isolated, just set $s = 2$ (see Fig. 3); if six fault variables need to be isolated, just set $s = 6$ (see Fig. 5). From this point of view, the proposed TSPCA is more feasible to implement and easier to interpret what causes the fault than the existing regularized approaches.

VI. APPLICATION STUDIES

A. Tennessee Eastman (TE) Benchmark Process

1) Data Description: The TE benchmark process has been widely used for testing different FD approaches [36]; see Fig. 7 for a detailed flowchart. In the TE process, 21 fault datasets with 960 samples are collected and a fault is introduced at the 161st sample. In general, the fault-free dataset is used for offline modeling and the fault datasets are used for online detecting. Each dataset contains 41 process variables (22 process variables and 19 analysis variables) and 11 manipulated variables. In this case, 22 process variables and 11 manipulated variables are chosen as $\mathbf{X}$.

2) Detection and Isolation Performance: Table II provides the quantitative comparisons achieved by the compared approaches. Moreover, the results of the proposed TSPCA are labeled in bold. It can be easily found that all the FAR_{T^2} and FAR_{SPE} values appear to be low, which suggests that the

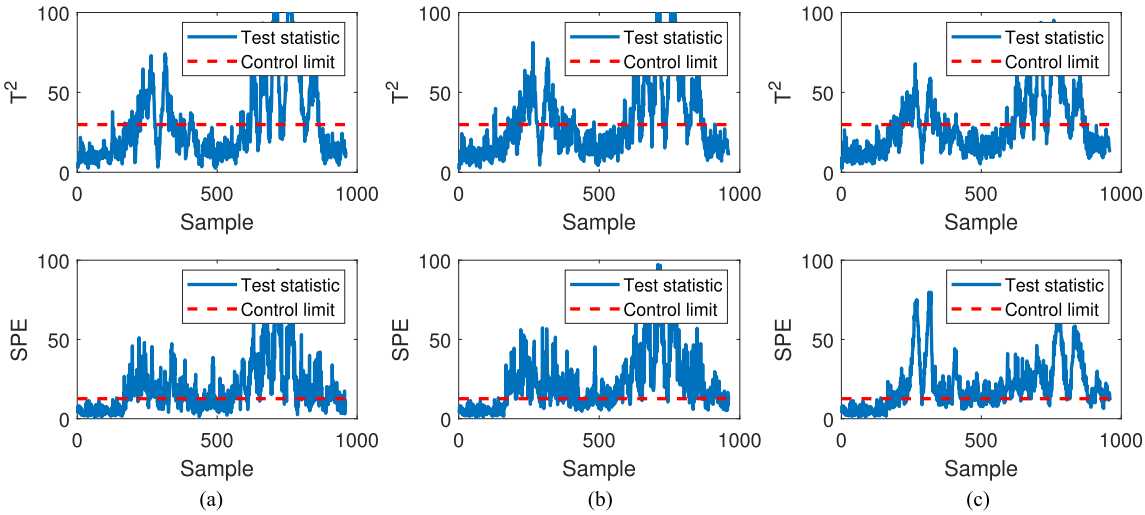

Fig. 8. Detection performance for IDV(10) in the TE process. (a) SJSPCA. (b) LSPCA. (c) TSPCA.

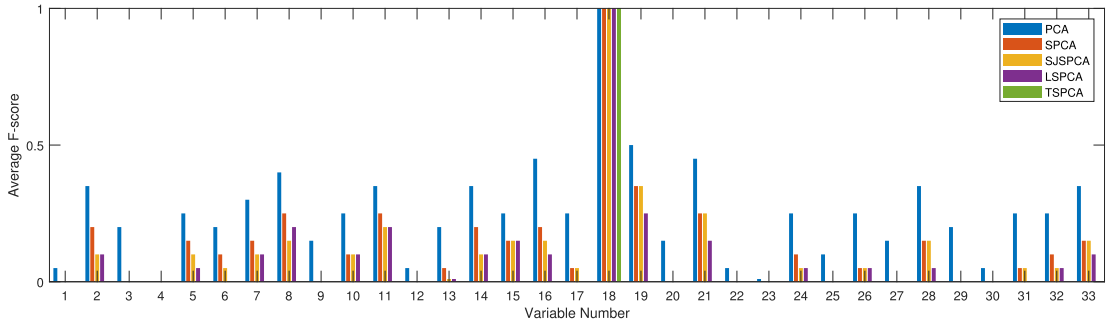

Fig. 9. Isolation results for IDV(10) in the TE process.

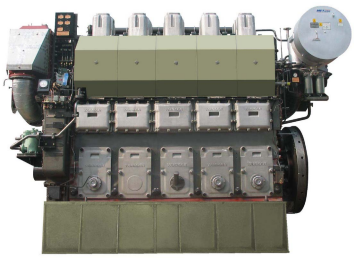

Fig. 10. Framework of the CP process.

PCA-based approaches are capable of detecting false alarms. For most faults in the TE process, the proposed TSPCA can obtain superior or comparable detection results than others. This considerable performance shows that the combination of the $\ell_{2,0}$-norm sparsity constrained optimization and the graph Laplace regularizer into a PCA framework has strong potential in fault detection.

To demonstrate the performance obtained by the proposed TSPCA, fault IDV(10) is chosen as an example. It occurs a random change in the feed C temperature of the stream 4, which causes a variation in the conditions of the stripper and the condenser. Fig. 8 plots the T^2 statistic and SPE statistic of fault IDV(10) to compare the performance of SJSPCA, LSPCA, and TSPCA. In comparison, the proposed TSPCA achieves very outstanding detection results, that is, numbers that violated the control limits between 600 and 800 samples.

In addition, the fault isolation results of fault IDV(10) are shown in Fig. 9. Compared with other PCA variants, the proposed TSPCA can isolate the root cause of this fault by selecting the sparsity level $s = 1$. In this article, the variable v_{18} gives the most significant contributor with the largest value. Therefore, v_{18} can be regarded as the possible faulty variable. As is described in the TE process, fault IDV(10) is closely correlated with v_{18}, that is, stripper temperature. This convinces that the proposed TSPCA makes it easier to interpret faults, and brings benefits to industrial processes.

B. Cylinder-Piston (CP) Process

1) Data Description: The CP is an important component of diesel engines, which bears a lot of pressure in the normal working process. The CP process dataset is collected in a practical two-stroke low-speed marine diesel engine [21]; see Fig. 10 for a schematic diagram.

This marine diesel engine contains five cylinders, and each has five process variables, including exhaust gas temperature, cooling oil inlet and outlet temperature difference, oil inlet pressure, JCW inlet and outlet temperature difference, JCW inlet

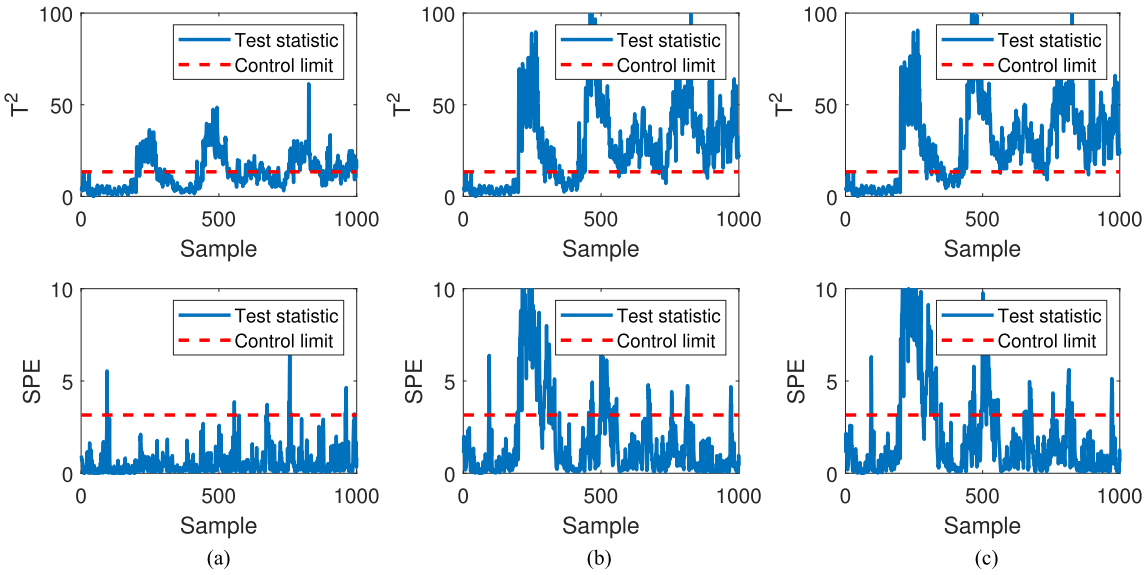

Fig. 11. Detection performance for the CP process. (a) SJSPCA. (b) LSPCA. (c) TSPCA.

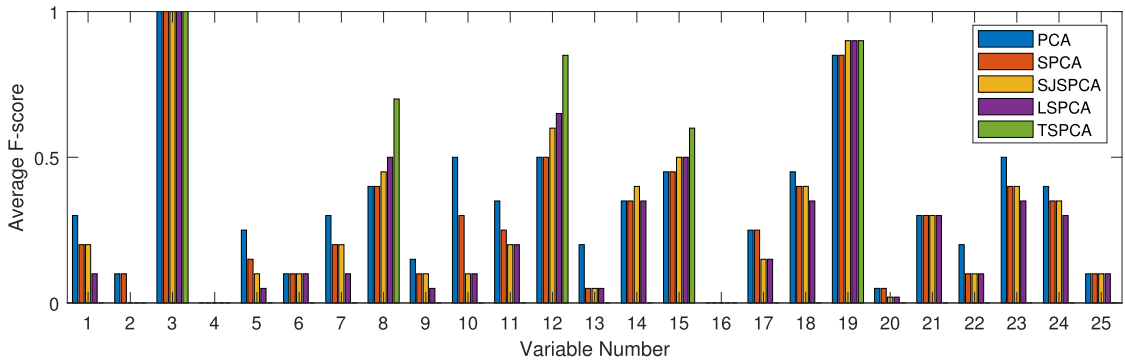

Fig. 12. Isolation results for the CP process.

pressure, where JCW stands for jacket cooling water. Therefore, a total of 25 variables are selected in this experiment. For offline modeling, 1000 process samples are collected. It is noted that a fault of five major bias and random small noise is introduced at the 201st sample to check the isolation performance.

2) Detection and Isolation Performance: Fig. 11 illustrates the detection performance for the practical CP application. It can be concluded that all the PCA-based approaches can detect this fault by monitoring the T^2 statistic at the 201st sample but the subsequent performance is different. Compared with SJSPCA and LSPCA, the proposed TSPCA is able to detect more fault samples under fault conditions.

To validate the effectiveness of the proposed TSPCA, the isolation results are given in Fig. 12. Although PCA, SPCA, SJSPCA, and LSPCA can isolate this fault, some fault-free variables are also isolated. For example, SPCA isolates more than 15 fault variables, but most of these variables are caused by the introduced noise. In contrast, the proposed TSPCA can discard these small residuals and isolate the major five fault variables. This indicates that the proposed TSPCA is also robust to noise. However, it should also be admitted that the proposed TSPCA cannot distinguish minor faults from noise, it may bring a limitation, that is, minor faults will be ignored.

Through the abovementioned simulation examples and application studies, it convinces us to believe that the proposed approach is promising for both fault detection and fault isolation with the better performance.

VII. Conclusion

In this article, we have proposed a new FD framework by introducing the $\ell_{2,0}$-norm sparsity constrained optimization. Unlike the existing $\ell_{2,1}$-norm regularizer, it has the capability to constrain the weight vectors row-wisely and detect the fault variables exactly. In algorithms, an iterative algorithm based on the ADMM scheme has been developed and discussed. In experiments, sufficient numerical cases have conducted to verify the superiority of the proposed approach. Now, it can be concluded that the $\ell_{2,0}$-norm sparsity constrained optimization enables the classical PCA approach to achieve good detection and fulfill accurate isolation. Naturally, this idea can also be applied to other data-driven FD approaches.

In the future, several interesting and important questions need to be further studied. First,this work is developed from engineering intuitions, fault tolerance, and error prone should be analyzed theoretically. Second, although the superiority has been illustrated by numerical cases, more efforts on practical platform verification are worth considering. Finally, it is possible to construct a robust PCA variant to deal with noisy process data.

REFERENCES

[1] S. X. Ding, *Model-Based Fault Diagnosis Techniques: Design Schemes, Algorithms and Tools*. Berlin, Germany: Springer-Verlag, 2008.

[2] S. X. Ding, *Data-Driven Design of Fault Diagnosis and Fault-Tolerant Control Systems*. London, U.K.: Springer-Verlag, 2014.

[3] S. Yin, S. X. Ding, A. Haghani, H. Hao, and P. Zhang, "A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process," *J. Process Control*, vol. 22, no. 9, pp. 1567–1581, 2012.

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

[5] 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.

[6] L. Li, S. X. Ding, H. Luo, K. Peng, and Y. Yang, "Performance-based fault-tolerant control approaches for industrial processes with multiplicative faults," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4759–4768, Jul. 2020.

[7] Z. Chen *et al.*, "A distributed canonical correlation analysis-based fault detection method for plant-wide process monitoring," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 2710–2720, May 2019.

[8] Q. Sun and Z. Ge, "A survey on deep learning for data-driven soft sensors," *IEEE Trans. Ind. Informat.*, vol. 17, no. 9, pp. 5853–5866, Sep. 2021.

[9] J. Wu, S. Sfarra, and Y. Yao, "Sparse principal component thermography for subsurface defect detection in composite products," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5594–5600, Dec. 2018.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Society: Ser. B. (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[11] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[12] Y. Liu, G. Zhang, and B. Xu, "Compressive sparse principal component analysis for process supervisory monitoring and fault detection," *J. Process Control*, vol. 50, pp. 1–10, 2017.

[13] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[14] D. West, *Introduction to Graph Theory*, vol. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[15] Y. Liu, J. Zeng, L. Xie, S. Luo, and H. Su, "Structured joint sparse principal component analysis for fault detection and isolation," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 2721–2731, May 2019.

[16] R. Zhai, J. Zeng, and Z. Ge, "Structured principal component analysis model with variable correlation constraint," *IEEE Trans. Control Syst. Technol.*, vol. 30, no. 5, pp. 558–569, Mar. 2022.

[17] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc.: Ser. B. (Statist. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.

[18] A. Beck and N. Hallak, "Optimization problems involving group sparsity terms," *Math. Program.*, vol. 178, no. 1, pp. 39–67, 2019.

[19] J. Li *et al.*, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, 2017.

[20] F. Nie, X. Dong, L. Tian, R. Wang, and X. Li, "Unsupervised feature selection with constrained $\ell_{2,0}$-norm and optimized graph," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1702–1713, Apr. 2022.

[21] X. Xiu, J. Fan, Y. Yang, and W. Liu, "Fault detection using structured joint sparse nonnegative matrix factorization," *IEEE Trans. Instrum. Meas.*, vol. 70, Mar. 2021, Art. no. 3513011.

[22] T. Pang, F. Nie, J. Han, and X. Li, "Efficient feature selection via $\ell_{2,0}$-norm constrained sparse regression," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 880–893, May 2019.

[23] S. Zhou, N. Xiu, and H. Qi, "Global and quadratic convergence of newton hard-thresholding pursuit," *J. Mach. Learn. Res.*, vol. 22, no. 12, pp. 1–45, 2021.

[24] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graphical Statist.*, vol. 15, no. 2, pp. 265–286, 2006.

[25] M. Yin, J. Gao, and Z. Lin, "Laplacian regularized low-rank representation and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 504–517, Mar. 2016.

[26] H. Zou and L. Xue, "A selective overview of sparse principal component analysis," *Proc. IEEE*, vol. 106, no. 8, pp. 1311–1320, Aug. 2018.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[28] D. Han, "A survey on some recent developments of alternating direction method of multipliers," *J. Operations Res. Soc. China*, vol. 10, no. 1, pp. 1–52, 2022.

[29] X. Yuan, P. Li, and T. Zhang, "Gradient hard thresholding pursuit," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6027–6069, 2017.

[30] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *J. Sci. Comput.*, vol. 78, no. 1, pp. 29–63, 2019.

[31] L. Chiang, E. Russell, and R. Braatz, *Fault Detection and Diagnosis in Industrial Systems*. Berlin, Germany: Springer, 2000.

[32] S. Qin, "Statistical process monitoring: Basics and beyond," *J. Chemometrics Soc.*, vol. 17, no. 8/9, pp. 480–502, 2003.

[33] N. Tracy, J. Young, and R. Mason, "Multivariate control charts for individual observations," *J. Qual. Technol.*, vol. 24, no. 2, pp. 88–95, 1992.

[34] J. Jackson and G. Mudholkar, "Control procedures for residuals associated with principal component analysis," *Technometrics*, vol. 21, no. 3, pp. 341–349, 1979.

[35] P. Van den Kerkhof, J. Vanlaer, G. Gins, and J. F. Van Impe, "Analysis of smearing-out in contribution plot based fault isolation for statistical process control," *Chem. Eng. Sci.*, vol. 104, pp. 285–293, 2013.

[36] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Comput. Chem. Eng.*, vol. 17, no. 3, pp. 245–255, 1993.

Xianchao Xiu (Member, IEEE) received the Ph.D. degree in operations research from Beijing Jiaotong University, Beijing, China, in 2019.

From June 2019 to May 2021, he was a Postdoctoral Researcher with Peking University, Beijing, China.

He is currently a Faculty Member with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China. His current research interests include large-scale sparse optimization, signal processing, deep learning, and data-driven fault detection.

Zhonghua Miao received the Ph.D. degree in mechatronic engineering from Shanghai Jiaotong University, Shanghai, China, in 2010.

He is currently a Full Professor and a Doctoral Advisor with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China. His research interests include intelligent robotics, measurement and control, fault diagnosis, and agricultural machinery equipment.

Wanquan Liu (Senior Member, IEEE) received the B.S. degree in applied mathematics from Qufu Normal University, Jining, China, in 1985, the M.S. degree in control theory and operation research from the Chinese Academy of Science, Beijing, China, in 1988, and the Ph.D. degree in electrical engineering from Shanghai Jiaotong University, Shanghai, China, in 1993.

He is currently a Full Professor with the School of Intelligent Systems Engineering, Sun Yat-Sen University, Guangzhou, China. His current research interests include large-scale pattern recognition, signal processing, machine learning, and control systems.

Dr. Liu once held the ARC Fellowship, U2000 Fellowship, and JSPS Fellowship and received the research funds from different resources more than 2.4 million dollars.