2020 IEEE 9th Data Driven Control and Learning Systems Conference
November 20-22, 2020, Liuzhou, China

A Highly Efficient Joint Sparsity Constrained Robust Principal Component Analysis for Fault Diagnosis

Xianchao Xiu [1], Ying Yang[1], Lingchen Kong[2], Wanquan Liu[3]

1. Department of Mechanics and Engineering Science, Peking University, Beijing 100871, P. R. China
E-mail: xcxiu@bjtu.edu.cn, yy@pku.edu.cn

2. Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, P. R. China
E-mail: lchkong@bjtu.edu.cn

3. Department of Computing, Curtin University, Perth WA 6102, Australia
E-mail: W.Liu@curtin.edu.au

Abstract: Principal component analysis (PCA) is one of the most commonly used techniques in high-dimensional data analysis, and it has been widely applied for fault diagnosis. However, the classical PCA has two drawbacks: sensitivity to outliers and non-interpretation of principal components. In this paper, a novel joint sparsity constrained robust principal component analysis (JSCRPCA) is proposed, in which the robust term makes it stable to outliers, the joint sparsity constraint controls row-wise sparsity, and the hypergraph Laplacian considers the structure information. In algorithm, an efficient alternating direction method of multipliers is designed to solve the proposed JSCRPCA. It is proved theoretically that the generated sequence converges to a local minimizer. Based on the T^2 and SPE statistics, an offline modelling and online monitoring procedure is presented. Numerical experiments on the Tennessee-Eastman process illustrate that JSCRPCA is able to improve the detection performance significantly in terms of fault detection rate and false alarm rate.

Key Words: Principal component analysis (PCA), Fault diagnosis (FD), Joint sparsity, Robust, Alternating direction method of multipliers.

1 Introduction

In modern industrial processes, fault diagnosis (FD) is becoming increasingly important to ensure safety and reliability. Generally speaking, FD techniques can be divided into two categories: model-based [5] and data-driven [6] ones. However, with the rapid development of data acquisition and sensor technologies, data-driven FD is becoming more and more dominant. A large number of multivariate analysis (MVA) have been proposed to improve the performance, such as principal component analysis (PCA) [12, 27], partial least squares (PLS) [4, 22], Fisher discriminant analysis (FDA) [24, 25], slow feature analysis (SFA) [2, 9], independent component analysis (ICA) [17, 22], canonical correlation analysis (CCA) [3, 16], and so on. It would not be exaggerating to say that PCA is one of the most effective and efficient techniques in FD.

The mathematical model of classical PCA can be presented as the form of

$$\begin{aligned} \min_{\mathbf{A}} \ & \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{A}\mathbf{A}^T\|_F^2 \\ \text{s.t.} \ & \mathbf{A}^T\mathbf{A}=\mathbf{I}_k, \end{aligned}$$

where $\mathbf{X}\in\mathbb{R}^{n\times p}$ is the given FD data matrix (n is the number of samples, p is the number of variables), $\mathbf{A}\in\mathbb{R}^{p\times k}$ is the loading matrix, and $\mathbf{I}_k$ is the k-dimensional identity matrix. It is worth noting that PCA can be interpreted as the best k low-rank approximation to $\mathbf{X}$. Therefore, the original data space contains a principal component (PC) subspace preserving most of the data information and a residual subspace keeping the rest data information. Accordingly, the T^2 and SPE statistics can be applied in the PC subspace or in the residual subspace. When a fault happens, the T^2 or SPE statistics will exceed the corresponding predefined control limits. The variable whose contribution index to T^2 and SPE statistics accounts for the major proportion will be isolated.

In high-dimensional data analysis, PCA cannot extract meaningful variables exactly, which often makes the interpretation ambiguous [28]. In order to overcome this shortcoming, Xie et al. [15] and Yan et al. [19] transformed the FD problem into sparse PCA (SPCA):

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B}} \ & \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^T\|_F^2+\lambda_1\|\mathbf{B}\|_1 \\ \text{s.t.} \ & \mathbf{A}^T\mathbf{A}=\mathbf{I}_k, \end{aligned}$$

where $\|\mathbf{B}\|_1$ is the sum of absolute values of all entries, and $\lambda_1>0$ is a penalty parameter for balancing the sparsity. By imposing the sparse regularizer on the PCA objective function, sparse loading matrix $\mathbf{B}$ can be obtained. As a remedy, the extracted PCs become linear combinations of a small number of process variables. Compared with PCA, SPCA enjoys higher computational efficiency and good isolation results. However, the loading matrix always has row-wise sparsity. That is, if variables are associated with all zero rows, they will be removed in the projected space. This inspired Liu et al. [11] to consider the joint sparse PCA (JSPCA), which is given by

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B}} \ & \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^T\|_F^2+\lambda_1\|\mathbf{B}\|_{2,1} \\ \text{s.t.} \ & \mathbf{A}^T\mathbf{A}=\mathbf{I}_k, \end{aligned}$$

where $\|\mathbf{B}\|_{2,1}$ denotes the sum of ℓ_2 norm of all rows. In fact, JSPCA acts like SPCA at the group level: depending on λ_1, an entire group of variables may be deleted from the model [10]. This makes JSPCA reduce the influence of outliers and reject useless feathers.

This work is supported in part by the National Natural Science Foundation of China (61633001,U1713223) and the 111 Project of China (B16002).

978-1-7281-5922-5/20/$31.00 ©2020 IEEE

From the point view of optimization perspective, it is essential to combine the learning procedure with the prior structure information of data. Manifold learning can be applied to maintain local geometric and topological structures. Among them, graph learning is simple and easy to implement [21]. By combing the joint sparsity with graph Laplacian regularization, Liu et al. [12] studied a structured joint sparse PCA (SJSPCA) to learn a sparse representation that explicitly takes the local structures of data into account, i.e.,

$$\begin{aligned}&\min_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^T\|_F^2+\lambda_1\|\mathbf{B}\|_{2,1}+\lambda_2\text{tr}(\mathbf{B}^T\mathbf{L}\mathbf{B})\\&\text{s.t. }\ \mathbf{A}^T\mathbf{A}=\mathbf{I}_k,\end{aligned}$$

where $\mathbf{L}$ is the graph Laplacian matrix learned from data. It is verified that the graph Laplacian matrix not only captures the cause-effect relationship between process variables, but also discovers conditional independent process variable sets between operation units [16].

Another interesting work is called robust PCA (RPCA), which is proposed by Yan et al. [18]. They first removed outliers by decomposing the monitoring data $\mathbf{X}$ into low-rank plus sparse matrices:

$$\min_{\mathbf{C},\mathbf{D}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}-\mathbf{C}\|_F^2+\lambda_1\|\mathbf{C}\|_1+\lambda_2\|\mathbf{D}\|_*,$$

where $\mathbf{C}$ is a sparse matrix to represent outliers, $\mathbf{D}$ is a low-rank matrix to approximate $\mathbf{X}$, and $\|\mathbf{D}\|_*$ is the sum of all singular values of $\mathbf{D}$. Then PCA is performed on the cleaned $\mathbf{D}$ to extract PCs. The application in the Tennessee-Eastman (TE) process showed that, RPCA can obtain similar performance with PCA when the training data is clean, and RPCA still has its validity when the training data is polluted by outliers. Some popular PCA-based FD models are reviewed in Table 1. It is easy to see that SJSPCA is not robust to outliers while RPCA cannot learn structure information. The natural question is whether we can combine SJSPCA and RPCA, and then derive a variant of PCA that is able to capture "robust" structures and preserve local geometric structures.

In this paper, joint sparsity constrained robust principal component analysis (JSCRPCA) is considered as follows

$$\begin{aligned}&\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{B}\mathbf{A}^T-\mathbf{C}\|_F^2+\lambda_1\|\mathbf{C}\|_1\\&\qquad+\lambda_2\text{tr}(\mathbf{B}\mathbf{L}^h\mathbf{B}^T)\\&\text{s.t. }\ \mathbf{A}^T\mathbf{A}=\mathbf{I}_k,\|\mathbf{B}\|_{2,0}=s,\end{aligned}\tag{1}$$

where $\mathbf{C}$ is introduced to make it robust against outliers, $\mathbf{L}^h$ is a hypergraph Laplacian matrix, in which an edge can connect more than two vertices, and $\|\mathbf{B}\|_{2,0}$ is defined as the number of nonzeros of ℓ_2 norm of all rows. Note that, different from $\ell_{2,1}$ norm regularization, $\ell_{2,0}$ quasi-norm constrained optimization enjoys two benefits. One is that s is easily adjusted to control the number of variables, rather than selecting λ_1 in [11, 12]. The other one is that $\ell_{2,0}$ is a hard threshold such that "smearing effect" can be vanished. For example, if only one significant variable needs to be extracted, just set $s=1$.

Compared to previous work, the main innovations of this paper are summarized in the following four aspects:

Table 1: PCA-based FD models.

Model	Robust	(Joint) Sparse	Graph Laplacian
PCA			
RPCA	✓		
SPCA		✓	
JSPCA		✓	
SJSPCA		✓	✓
Our	✓	✓	✓

1) To improve the robustness to outliers or noises, a novel robust PCA framework is proposed, which is different from the existing two-stage procedure in [18].
2) Instead of considering $\ell_{2,1}$ norm regularization, $\ell_{2,0}$ constraint is embedded into JSCRPCA such that faulty variables can be accurately isolated.
3) A hypergraph Laplacian matrix, not a normal one in [12], is applied to explore the high order relations between process variables and operation units.
4) An efficient optimization algorithm is developed to solve the proposed JSCRPCA. In addition, the effectiveness on the TE process is tested.

The remainder of this paper is organized as follows. Section 2 gives the optimization algorithm and discusses its convergence. In Section 3, an online FD strategy using JSCRPCA is presented. Section 4 reports numerical results to illustrate the advantages of JSCRPCA. Finally, Section 5 concludes this paper.

2 Optimization Algorithm

This section presents an alternating direction method of multipliers (ADMM) [1] for solving (1) discusses its local convergence guarantee.

Recall that JSCRPCA can be equivalently written as

$$\begin{aligned}&\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{D}\mathbf{A}^T-\mathbf{C}\|_F^2+\lambda_1\|\mathbf{C}\|_1\\&\qquad+\lambda_2\text{tr}(\mathbf{B}\mathbf{L}^h\mathbf{B}^T)\\&\text{s.t. }\ \mathbf{A}^T\mathbf{A}=\mathbf{I}_k,\|\mathbf{B}\|_{2,0}=s,\mathbf{B}=\mathbf{D},\end{aligned}\tag{2}$$

where $\mathbf{D}$ is an auxiliary variable to make the objective function separable. To describe the iterates of the ADMM, the augmented Lagrangian function is given by

$$\begin{aligned}&\mathcal{L}_\beta(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\mathbf{\Lambda})\\&=\frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{D}\mathbf{A}^T-\mathbf{C}\|_F^2+\lambda_1\|\mathbf{C}\|_1+\lambda_2\text{tr}(\mathbf{B}\mathbf{L}^h\mathbf{B}^T)\\&\quad-\langle\Lambda,\mathbf{B}-\mathbf{D}\rangle+\frac{\beta}{2}\|\mathbf{B}-\mathbf{D}\|_F^2,\end{aligned}$$

where $\mathbf{\Lambda}\in\mathbb{R}^{p\times k}$ is the Lagrangian multiplier and $\beta>0$ is a penalty parameter. The above expression always requires minimizations over $\mathcal{M}_1=\{\mathbf{A}\mid\mathbf{A}^T\mathbf{A}=\mathbf{I}_k\}$ and $\mathcal{M}_2=\{\mathbf{B}\mid\|\mathbf{B}\|_{2,0}=s\}$. The ADMM for solving (2) is thus described as follows:

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

where $\tau \in (0, \frac{1+\sqrt{5}}{2})$ is the dual step-size. In most cases, a larger τ results in faster convergence.

- For $\mathbf{A}$-subproblem, after trivial manipulation, it can be simplified to

$$\min_{\mathbf{A}\in\mathcal{M}_1} \left\{ \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{D}^k\mathbf{A}^T - \mathbf{C}^k\|_F^2 \right\}.$$

  This is called a reduced rank Procrustes rotation problem, and the solution is

$$\mathbf{A}^{k+1} = \mathbf{U}\mathbf{V}^T,$$

  where $\mathbf{U}$ and $\mathbf{V}$ are the associated singular vectors, i.e., $(\mathbf{X} - \mathbf{C}^k)^T\mathbf{X}\mathbf{D}^k = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.
- For $\mathbf{B}$-subproblem, it can be transformed as

$$\min_{\mathbf{B}\in\mathcal{M}_2} \left\{ \frac{\beta}{2}\|\mathbf{B} - \mathbf{D}^k - \mathbf{\Lambda}^k/\beta\|_F^2 + \lambda_2 \mathrm{tr}(\mathbf{B}\mathbf{L}^h\mathbf{B}^T) \right\},$$

  which is solved by

$$\mathbf{B}^{k+1} = \Pi_{\mathcal{M}_2}\left((\beta\mathbf{I} + 2\lambda_2\mathbf{L}^h)^{-1}(\beta\mathbf{D}^k + \mathbf{\Lambda}^k)\right),$$

  where $\Pi_{\mathcal{M}_2}$ denotes the projection onto manifold $\mathcal{M}_2$. For sparsity constrained optimization, it is efficiently computed by HTP [23] or Newton method [14] rowwisely. Notice that the coefficient matrix $\beta\mathbf{I} + 2\lambda_2\mathbf{L}^h$ is nonsingular because β, λ_2 are positive. Thus, the Moore-Penrose pseudo-inverse can be computed via the Cholesky decomposition or conjugate gradient method.
- For $\mathbf{C}$-subproblem, it can be rewritten as

$$\min_{\mathbf{C}} \left\{ \frac{1}{2}\|\mathbf{X} - \mathbf{X}\mathbf{D}^k(\mathbf{A}^{k+1})^T - \mathbf{C}\|_F^2 + \lambda_1\|\mathbf{C}\|_1 \right\}.$$

  The solution is given by

$$\mathbf{C}^{k+1} = \mathcal{S}_{\lambda_1}(\mathbf{X} - \mathbf{X}\mathbf{D}^k(\mathbf{A}^{k+1})^T),$$

  where the soft-shrinkage operator is defined as

$$\mathcal{S}_\lambda(x) = \mathrm{sign}(x) \circ \max\{0, |x| - \lambda\},$$

  with $\mathrm{sign}(\cdot)$ being the sign function and all operations are done componentwise. See [7] for more details.
- For $\mathbf{D}$-subproblem, the solution is

$$\mathbf{D}^{k+1} = (\beta\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1} \left(\mathbf{X}^T\mathbf{X}\mathbf{A}^{k+1} - \mathbf{X}^T\mathbf{C}^{k+1}\mathbf{A}^{k+1} + \beta\mathbf{B}^{k+1} - \mathbf{\Lambda}^k\right).$$

Overall, the ADMM for solving (2) (equivalently (1)) is established in Algorithm 1. To end this section, the convergence property is investigated, which results an efficient and convergent algorithm.

From [13], the first-order optimality condition of problem (1) at the local minimizer $(\overline{\mathbf{A}}, \overline{\mathbf{B}}, \overline{\mathbf{C}})$ is given by

$$\begin{cases} 0 \in \Pi_{\mathcal{M}_1}(-\mathbf{X}\overline{\mathbf{B}}(\mathbf{X} - \mathbf{X}\overline{\mathbf{B}\mathbf{A}}^T - \overline{\mathbf{C}})); \\ 0 \in \Pi_{\mathcal{M}_2}(-\mathbf{X}^T\overline{\mathbf{A}}(\mathbf{X} - \mathbf{X}\overline{\mathbf{B}\mathbf{A}}^T - \overline{\mathbf{C}}) + 2\lambda_2\mathbf{L}^h\overline{\mathbf{B}}); \\ 0 \in -(\mathbf{X} - \mathbf{X}\overline{\mathbf{B}\mathbf{A}}^T - \overline{\mathbf{C}}) + \lambda_2\partial\|\overline{\mathbf{C}}\|_1. \end{cases}$$

Hence, $(\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*)$ is called a stationary point of (1) if it satisfies the above relation in place of $(\overline{\mathbf{A}}, \overline{\mathbf{B}}, \overline{\mathbf{C}})$.

Algorithm 1 ADMM for solving JSCRPCA model (2)

Input: Data set $\mathbf{X}$, parameters $\lambda_1, \lambda_2 > 0$, and $\beta > 0$.
Output: Representation matrix $\mathbf{B}$.
Initialize: Compute $\mathbf{L}^h$, $\mathbf{C}^0 = \mathbf{D}^0 = 0$, $\mathbf{\Lambda}^0 = 0$, $\varepsilon_1 = 10^{-3}$, $\varepsilon_2 = 10^{-2}$.
While not converged **do**

1: update $\mathbf{A}^{k+1}$ according to

$$\left(\mathbf{X} - \mathbf{C}^k\right)^T\mathbf{X}\mathbf{D}^k = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$
$$\mathbf{A}^{k+1} = \mathbf{U}\mathbf{V}^T;$$

2: update $\mathbf{B}^{k+1}$ according to

$$\mathbf{B}^{k+1} = \Pi_{\mathcal{M}_2}\left((\beta\mathbf{I} + 2\lambda_2\mathbf{L}^h)^{-1}(\beta\mathbf{D}^k + \mathbf{\Lambda}^k)\right);$$

3: update $\mathbf{C}^{k+1}$ according to

$$\mathbf{C}^{k+1} = \mathcal{S}_{\lambda_1}\left(\mathbf{X} - \mathbf{X}\mathbf{D}^k(\mathbf{A}^{k+1})^T\right);$$

4: update $\mathbf{D}^{k+1}$ according to

$$\mathbf{D}^{k+1} = (\mathbf{X}^T\mathbf{X} + \beta\mathbf{I})^{-1} \left(\mathbf{X}^T\mathbf{X}\mathbf{A}^{k+1} - \mathbf{X}^T\mathbf{C}^k\mathbf{A}^{k+1} + \beta\mathbf{B}^{k+1} - \mathbf{\Lambda}_1^k\right);$$

5: update $\mathbf{\Lambda}^{k+1}$ via

$$\mathbf{\Lambda}^{k+1} = \mathbf{\Lambda}^k - \tau\beta\left(\mathbf{B}^{k+1} - \mathbf{D}^{k+1}\right);$$

6: check convergence: if

$$\frac{\left\|\mathbf{X} - \mathbf{X}\mathbf{D}^{k+1}(\mathbf{A}^{k+1})^T - \mathbf{C}^{k+1}\right\|_F}{\max\{\|\mathbf{X}\|_F, 1\}} < \varepsilon_1$$

and

$$\max\left\{ \begin{array}{l} \|\mathbf{A}^{k+1} - \mathbf{A}^k\|_F, \ \|\mathbf{B}^{k+1} - \mathbf{B}^k\|_F \\ \|\mathbf{C}^{k+1} - \mathbf{C}^k\|_F, \ \|\mathbf{D}^{k+1} - \mathbf{D}^k\|_F \end{array} \right\} < \varepsilon_2,$$

then stop;

End while

Theorem 2.1 *Assume that* $0 < \tau < \frac{1+\sqrt{5}}{2}$ *and the sequence* $\{(\mathbf{A}^k, \mathbf{B}^k, \mathbf{C}^k, \mathbf{D}^k, \mathbf{\Lambda}^k)\}_{k=1}^{\infty}$ *is generated by Algorithm 1. Then it converges to a stationary point of* (2) *(also* (1)*).*

The proof follows a similar line of arguments as in [20], thus it is omitted here for the sake of continuity.

3 Fault Diagnosis Using JSCRPCA

This section gives an offline modelling and online monitoring procedure based on the proposed JSCRPCA. As is known to us that normalization is important to eliminate the effects of engineering units and measurement ranges. Let $\mathbf{X}_{obs}$ be the observation matrix, then perform

$$x_{ij} = \frac{x_{obs,ij} - \overline{x}_{obs,j}}{\sigma_{obs,j}},$$

where $x_{obs,ij}$ is the ijth entry of the observation matrix $\mathbf{X}_{obs}$, $\overline{x}_{obs,j}$ is the mean value of the jth process variable, $\sigma_{obs,j}$ is the corresponding standard deviation, and x_{ij} is the ijth entry of the normalized matrix $\mathbf{X}$.

By solving optimization problem (1), the loading matrix $\mathbf{B}$ can be obtained. For convenience of expression, $\mathbf{B}$ is di-

Algorithm 2 Fault diagnosis using JSCRPCA

Offline modelling:

1: Normalize the training data $\mathbf{X}$;
2: Specify the retained number k;
3: Solve JSCRPCA using Algorithm 1;
4: Compute control limits for T^2 and SPE statistics by (4);

Online monitoring:

1: Normalize the test data $\mathbf{x}$;
2: Calculate the T^2 and SPE values based on (3);
3: Compare them with control limits to detect a fault via (5);
4: Compute fault score for each variable using (6), and thus the fault is isolated.

vided into $\mathbf{B} = [\mathbf{B}_r, \mathbf{B}_d]$, where $\mathbf{B}_r$ denotes the loading vectors of the retained PCs and $\mathbf{B}_d$ denotes the loading vectors of the discarded PCs. To detect abnormalities, the following T^2 and SPE statistics [12] are usually used:

$$T^2 = \mathbf{x}_i^T \mathbf{B}_r \mathbf{\Lambda}^{-1} \mathbf{B}_r^T \mathbf{x}_i,$$
$$SPE = \mathbf{x}_i^T (\mathbf{I} - \mathbf{B}_r \mathbf{B}_r^T) \mathbf{x}_i, \quad (3)$$

where $\mathbf{x}_i \in \mathbb{R}^n$ $(i = 1, \ldots, p)$ is the data vector associated with the ith variable, and $\mathbf{\Lambda}$ is defined as $\mathbf{\Lambda} = \text{diag}(\text{var}(PC_1), \text{var}(PC_2), \ldots, \text{var}(PC_r))$.

According to [6], the control limits for T^2 and SPE statistics are chosen as

$$J_{th,T^2} = \frac{r(n^2-1)}{n(n-1)} F_\alpha(r, n-r),$$
$$J_{th,SPE} = \theta_1 \left(\frac{C_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right)^{1/h_0} \quad (4)$$

with

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

where $F_\alpha(r, n-r)$ is the F-distribution with r and $n-r$ degrees of freedom under the significance level α. For the purpose of FD, the following logic is adopted

$$T^2 \le J_{th,SPE} \text{ and } SPE \le J_{th,SPE}$$
$$\Rightarrow \text{fault-free, otherwise faulty.} \quad (5)$$

Once a fault is detected, the contribution plot is then applied to determine which variable leads to the fault with a high probability. To measure the degree for each variable, a fault score can be formulated by

$$\delta_i = \sum_{j=1}^{k} |b_{ij}|/k, \ i = 1, \ldots, p. \quad (6)$$

For each faulty data matrix $\mathbf{X}$, the resulting fault score vector can be normalized by dividing the maximum of δ_i. A high score indicates the variable is more likely to be faulty.

Finally, the FD procedure is summarized in Table 2. By integrating the constraint $\|\mathbf{B}\|_{2,0} = s$, only s variables are extracted. Compared with $\ell_{2,1}$ regularization, it is highly flexible and efficient in modern industrial processes. Note that, after conducting JSCRPCA, the loading vectors only contain a few nonzero row-wise elements, thus it is convenient to find the root cause of the incipient fault. That's why our proposed JSCRPCA is called "highly efficient".

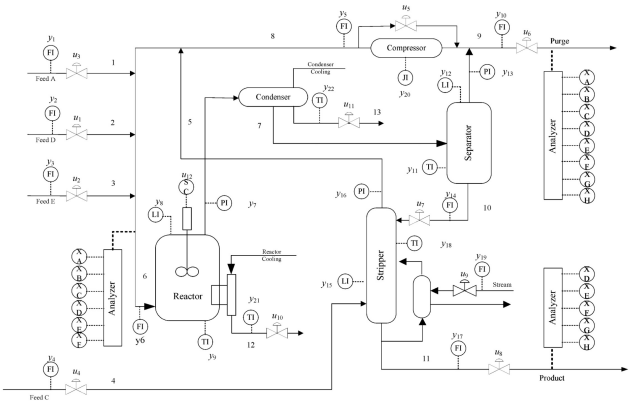

Fig. 1: TE process and monitoring variables.

Table 2: The TE process faults.

Fault No.	Description	Type
1	A/C feed ratio	step change
2	component B	step change
3	feed D temperature	step change
4	RCW inlet temperature	step change
5	CCW inlet temperature	step change
6	feed A loss	step change
7	C header pressure loss	step change
8	feed A−C components	random variation
9	feed D temperature	random variation
10	feed C temperature	random variation
11	RCW inlet temperature	random variation
12	CCW inlet temperature	random variation
13	reaction kinetics	slow drift
14	RCW valve	sticking
15	CCW valve	sticking
16	unknown fault	unknown
17	unknown fault	unknown
18	unknown fault	unknown
19	unknown fault	unknown
20	unknown fault	unknown
21	unknown fault	constant

4 Application to TE Process

This section demonstrates the superiority of our proposed JSCRPCA over the classical PCA on the benchmark Tennessee Eastman (TE) process.

4.1 Data Description

The TE process has five major units, i.e. reactor, condenser, separator, stripper, and compressor. Besides, there exist eight components, including four reactants, two products, a major byproduct, and an inert component. Fig. 1 provides a overview and more detailed information can be found in [8].

The TE process consists of 1 normal data set and 21 faulty data sets. The sampling interval for all data sets is 3 minutes and the total time is 48 hours, so there are 960 samples for each data set. A fault is introduced to the process at the 161st sampling time point. The normal data set is usually employed for offline modelling, while fault data sets are utilized for online monitoring. Table 2 gives detailed descriptions of the 21 faults. Here, RCW represents reactor cooling water, and CCW stands for condenser cooling water.

Table 3: Selected variables in the TE process.

No.	Description	No.	Description
1	feed A	18	stripper temperature
2	feed D	19	stripper steam flow
3	feed E	20	compressor work
4	total feed	21	RCW outlet temperature
5	recycle flow	22	product separator temperature
6	reactor feed rate	23	feed D flow valve
7	reactor pressure	24	feed E flow valve
8	reactor level	25	feed A flow valve
9	reactor temperature	26	total feed flow valve
10	purge rate	27	compressor recycle valve
11	RCW outlet temperature	28	purge valve
12	product separator level	29	separator pot liquid flow valve
13	product separator pressure	30	stripper liquid product flow valve
14	product separator underflow	31	stripper steam valve
15	stripper level	32	RCW flow
16	stripper pressure	33	CCW flow
17	stripper underflow		

4.2 Implementation Details

In this study, 33 variables are considered as shown in Table 3. For purpose of comparison, the parameters are empirically set $\lambda_1 = 200$ and $\lambda_2 = 50$. In addition, the hypergraph Laplacian matrix $\mathbf{L}^h$ is generated as follows. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a hypergraph, where $\mathcal{V}$ is the vertex set of n samples, and $\mathcal{E}$ is the hyperedge set of p variables. For a subset $\mathbf{e} \in \mathcal{E}$, the weight matrix $\mathbf{W}_{\mathcal{E}}$ consists of diagonal entries $W(\mathbf{e})$. For a vertex $v \in \mathcal{V}$, the degree is defined as $d(v) = \sum_{\mathbf{e}\in\mathcal{E}} W(\mathbf{e})h(v, \mathbf{e})$ with

$$h(v, \mathbf{e}) = \begin{cases} 1 & \text{if } v \in \mathbf{e}; \\ 0 & \text{otherwise.} \end{cases}$$

Denote $\mathbf{D}_{\mathcal{V}}$ as the diagonal matrix whose diagonal entries correspond to the degree of each vertex. For a hyperedge $\mathbf{e}$, the degree is denoted by $d(\mathbf{e}) = \sum_{v\in\mathcal{V}} h(v, \mathbf{e})$ and the diagonal degree matrix $\mathbf{D}_{\mathcal{E}}$ is thus obtained. Based on [26], the hypergraph Laplacian matrix is presented as

$$\mathbf{L}^h = \mathbf{D}_{\mathcal{V}} - \mathbf{H}\mathbf{W}_{\mathcal{E}}\mathbf{D}_{\mathcal{E}}^{-1}\mathbf{H}^T.$$

Clearly, it is more general than normal graph Laplacian which is used in [12, 16].

4.3 Monitoring Results

The fault detection rate (FDR) and false alarm rate (FAR) are adopted to measure the performance of FD [6]. The monitoring results of PCA and JSCRPCA are reported in Table 4. It is concluded that, for all the selected faults, the proposed JSCRPCA always achieves higher FDR and lower FAR. In particular, for T^2 statistics of fault 4, the gains of FDR values are around 18.50%. This good FD performance is attributed to the incorporation of the robust term, $\ell_{2,0}$ joint sparsity and graph Laplacian regularization term. While the robust term can filter out gross noises, the $\ell_{2,0}$ constraint can retain the fault variables and exclude normal ones, the graph Laplacian can preserve the structured correlation relationship between process variables.

Let's take fault 10 for example. It is a random variation in the feed C temperature in stream 4. The variation in the feed C temperature causes a change in the conditions of the stripper and then the condenser. See Fig. 2 for monitoring results of PCA and JSCRPCA. The fault is successfully detected at around 161th sample, but JSCRPCA gives much more samples violated the control limits, which shows that the monitoring performance of JSCRPCA is superior to that of PCA.

Table 4: Detection results in terms of FDR and FAR(%).

No.	PCA(FDR)		JSCRPCA(FDR)		PCA(FAR)		JSCRPCA(FAR)	
	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE
1	99.13	99.88	99.38	100	0.00	0.63	0.00	0.00
2	98.38	95.75	98.38	99.75	1.88	0.63	0.00	0.63
3	0.88	2.63	3.75	6.75	0.00	1.25	0.00	0.00
4	20.88	100	39.38	100	0.63	1.25	0.63	0.00
5	24.13	20.88	35.75	26.50	0.63	1.88	0.00	0.00
6	99.13	100	99.38	100	0.00	1.25	0.00	0.63
7	100	100	100	100	0.00	1.25	0.00	0.00
8	96.88	83.63	97.50	96.88	0.00	0.63	0.00	0.63
9	1.75	1.75	3.75	2.75	1.88	1.25	0.63	0.00
10	29.63	25.75	36.13	34.58	0.00	0.63	0.00	0.00
11	40.63	74.88	48.50	90.25	0.63	2.50	0.63	0.63
12	98.38	89.50	99.25	92.38	0.00	1.25	0.00	0.63
13	93.63	95.25	93.63	99.75	0.63	0.00	0.63	0.00
14	99.25	100	99.88	100	0.00	1.25	0.00	0.63
15	1.38	3.00	3.88	7.25	0.00	1.25	0.00	0.00
16	13.50	27.38	18.38	37.50	2.50	1.88	0.63	1.25
17	76.25	95.38	88.50	96.75	1.25	2.50	0.63	1.88
18	89.25	90.13	90.63	93.50	0.00	2.50	0.00	1.25
19	14.13	18.50	16.75	28.25	0.00	0.63	0.00	0.00
20	31.75	49.75	48.38	69.63	0.00	1.25	0.00	0.63
21	39.25	47.25	47.63	53.75	0.00	2.50	0.00	1.25

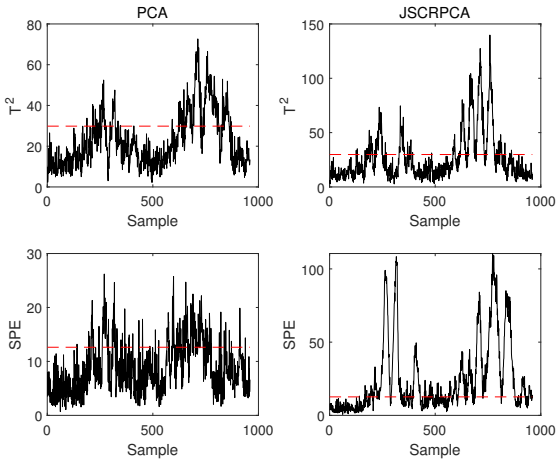

Fig. 2: Monitoring results for fault 10 with respect to T^2 and SPE. The red dotted lines denote the control limits.

The fault scores are displayed in Fig. 3. All of them illustrate that v_{18} (stripper temperature) is the largest contributor. Therefore, v_{18} is viewed as the faulty variable responsible, which is closely related to the real faulty variable (C feed temperature) of fault 10. By choosing different s, JSCRPCA is able to diagnose the root cause of a fault, see the right column of Fig. 3. This makes the interpretation of faults easier to understand.

5 Conclusions

In this paper, a novel joint sparsity constrained principal component analysis (JSCRPCA) is proposed by integrating ℓ_1 noisy term, $\ell_{2,0}$ joint sparsity constraint and hypergraph Laplacian regularization. To the best of our knowledge, it is the first time to model PCA in such a JSCRPCA framework. An efficient ADMM is established with convergence analysis. Numerical comparisons between our approach and PCA, on the TE process, are conducted to demonstrate its ef-

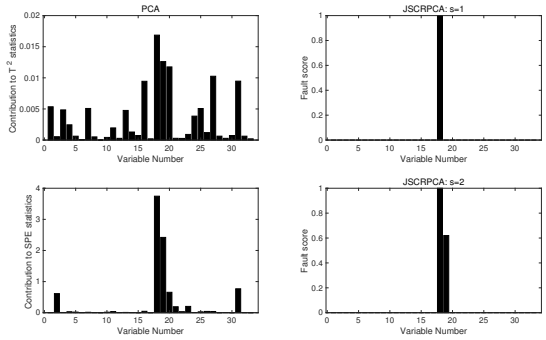

Fig. 3: Contribution plots for fault 10. For JSCRPCA, s is set as 1 or 2.

fectiveness. For these reasons, it is believed that JSCRPCA is a valuable approach for FD.

In the future work, several interesting questions and extensions remain open. First, extend the prior information to CCA. Second, develop deep PCA to learn nonlinear structure. Finally, apply it to more practical FD scenarios.

References

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

[2] Z. Chai and C. Zhao, "Enhanced random forest with concurrent analysis of static and dynamic nodes for industrial fault classification," *IEEE Transactions on Industrial Informatics*, 2019.

[3] Z. Chen, S. X. Ding, T. Peng, C. Yang, and W. Gui, "Fault detection for non-Gaussian processes using generalized canonical correlation analysis and randomized algorithms," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 2, pp. 1559–1567, 2017.

[4] 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.

[5] S. X. Ding, *Model-based fault diagnosis techniques: design schemes, algorithms, and tools*. Springer Science & Business Media, 2008.

[6] ——, *Data-driven design of fault diagnosis and fault-tolerant control systems*. Springer, 2014.

[7] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[8] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Computers & Chemical Engineering*, vol. 17, no. 3, pp. 245–255, 1993.

[9] Z. Li and X. Yan, "Complex dynamic process monitoring method based on slow feature analysis model of multi-subspace partitioning," *ISA Transactions*, vol. 95, pp. 68–81, 2019.

[10] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization," *arXiv:1205.2631*, 2012.

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

[12] Y. Liu, J. Zeng, L. Xie, S. Luo, and H. Su, "Structured joint sparse principal component analysis for fault detection and isolation," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 2721–2731, 2018.

[13] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009.

[14] R. Wang, N. Xiu, and C. Zhang, "Greedy projected gradient-newton method for sparse logistic regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 2, pp. 527–538, 2020.

[15] L. Xie, X. Lin, and J. Zeng, "Shrinking principal component analysis for enhanced process monitoring and fault isolation," *Industrial & Engineering Chemistry Research*, vol. 52, no. 49, pp. 17 475–17 486, 2013.

[16] X. Xiu, Y. Yang, L. Kong, and W. Liu, "Data-driven process monitoring using structured joint sparse canonical correlation analysis," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2020.

[17] Y. Xu and X. Deng, "Fault detection of multimode non-Gaussian dynamic process using dynamic Bayesian independent component analysis," *Neurocomputing*, vol. 200, pp. 70–79, 2016.

[18] Z. Yan, C.-Y. Chen, Y. Yao, and C.-C. Huang, "Robust multivariate statistical process monitoring via stable principal component pursuit," *Industrial & Engineering Chemistry Research*, vol. 55, no. 14, pp. 4011–4021, 2016.

[19] Z. Yan and Y. Yao, "Variable selection method for fault isolation using least absolute shrinkage and selection operator (LASSO)," *Chemometrics and Intelligent Laboratory Systems*, vol. 146, pp. 136–146, 2015.

[20] L. Yang, T. K. Pong, and X. Chen, "Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction," *SIAM Journal on Imaging Sciences*, vol. 10, no. 1, pp. 74–110, 2017.

[21] L. Yang, C. Li, J. Han, C. Chen, Q. Ye, B. Zhang, X. Cao, and W. Liu, "Image reconstruction via manifold constrained convolutional sparse coding for image sets," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1072–1081, 2017.

[22] 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," *Journal of Process Control*, vol. 22, no. 9, pp. 1567–1581, 2012.

[23] X.-T. Yuan, P. Li, and T. Zhang, "Gradient hard thresholding pursuit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6027–6069, 2017.

[24] J. Zheng, H. Wang, Z. Song, and Z. Ge, "Ensemble semi-supervised Fisher discriminant analysis model for fault classification in industrial processes," *ISA Transactions*, vol. 92, pp. 109–117, 2019.

[25] S. Zhong, Q. Wen, and Z. Ge, "Semi-supervised Fisher discriminant analysis model for fault classification in industrial processes," *Chemometrics and Intelligent Laboratory Systems*, vol. 138, pp. 203–211, 2014.

[26] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Advances in Neural Information Processing Systems*, 2007, pp. 1601–1608.

[27] J. Zhu, Z. Ge, and Z. Song, "Distributed parallel PCA for modeling and monitoring of large-scale plant-wide processes with big data," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1877–1885, 2017.

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