Efficient and Fast Joint Sparse Constrained Canonical Correlation Analysis for Fault Detection

Xianchao Xiu, *Member, IEEE*, Lili Pan, Ying Yang, *Senior Member, IEEE*, and Wanquan Liu, *Senior Member, IEEE*

***Abstract*— The canonical correlation analysis (CCA) has attracted wide attention in fault detection (FD). To improve the detection performance, we propose a new joint sparse constrained CCA (JSCCCA) model that integrates the $\ell_{2,0}$-norm joint sparse constraints into classical CCA. The key idea is that JSCCCA can fully exploit the joint sparse structure to determine the number of extracted variables. We then develop an efficient alternating minimization algorithm using the improved iterative hard thresholding and manifold constrained gradient descent method. More importantly, we establish the convergence guarantee with detailed analysis. Finally, we provide extensive numerical studies on the simulated dataset, the benchmark Tennessee Eastman process, and a practical cylinder-piston process. In some cases, the computing time is reduced by 600 times, and the FD rate is increased by 12.62% compared with classical CCA. The results suggest that the proposed approach is efficient and fast.**

***Index Terms*— $\ell_{2,0}$-norm joint sparse, canonical correlation analysis (CCA), fault detection (FD), optimization algorithm.**

I. Introduction

FAULT detection (FD) is essential for modern industrial processes, such as microelectronics manufacturing, power systems, and agricultural production. Generally speaking, FD approaches can be categorized into model-based approaches [1] and data-driven approaches [2]. However, model-based FD approaches need precise system models, which makes practical applications difficult. Recent years have witnessed the rapid development of data-driven FD approaches, such as principal component analysis (PCA) [3], the Fisher discriminant analysis (FDA) [4], partial least squares (PLS) [5], nonnegative matrix factorization (NMF) [6], and canonical correlation analysis (CCA) [7], [8]. Unlike other data-driven FD approaches, CCA makes full use of the inputs and outputs to exploit the cause–effect relationship between variables, thus providing excellent detection performance. Now, CCA-based data-driven FD has made great success in research and industry.

However, for complicated processes, not all canonical variables have identical information to characterize the fault. In this sense, achieving a consistent common representation is not a trivial task if the canonical variables are not incorporated with discriminative information [9]. This observation has inspired many researchers to propose different CCA variants, such as graph CCA [10], [11], kernel CCA [12], deep CCA [13], [14], [15], [16], and sparse CCA (SCCA) [17], [18], [19], [20]. Among them, SCCA is frequently used since its formulation and implementation are simple. The key idea of SCCA is to integrate the ℓ_1-norm (sum of absolute values) regularization term with the CCA objective and then improve the variable interpretability [17]. From a variable selection point of view, each row of canonical matrices is closely related to a specific variable, and a zero row means that the variable is less important than others [21], [22], [23]. However, the above SCCA cannot capture the rowwise joint sparse structure. To overcome this issue, joint SCCA (JSCCA) was proposed by enforcing the $\ell_{2,1}$-norm (sum of ℓ_2-norm of each row) regularization term. The performance of JSCCA has been demonstrated in signal processing [24], [25], image classification [26], and FD [27]. In fact, JSCCA acts like SCCA at the row level to remove these unimportant or useless variables, thereby preserving the intrinsic variables. Although JSCCA brings promising detection improvement, it is difficult to determine the number of extracted variables.

In the last few years, ℓ_0-norm (number of nonzero elements) sparse constrained optimization has been illustrated to outperform the above ℓ_1-norm or $\ell_{2,1}$-norm regularizers [28], [29], [30], [31], [32]. Although these problems related to ℓ_0-norm are often NP-hard [33], some greedy methods can be applied to search for approximate solutions, including iterative hard thresholding [28], improved iterative hard thresholding (IIHT) [30], gradient hard thresholding pursuit [29], and the Newton hard thresholding pursuit [32]. Considering the joint sparse structure of canonical matrices, it is necessary to extend the sparse constrained optimization to the joint sparse constrained optimization version, i.e., $\ell_{2,0}$-norm (number of nonzero rows). The performance has been demonstrated in [34], [35], [36], and [37]. All these

Manuscript received 18 May 2021; revised 28 December 2021, 22 April 2022, and 27 June 2022; accepted 22 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 12001019, Grant 62173003, and Grant 12271309; and in part by the National Key Research and Development Program of China under Grant 2021YFB3301204. *(Corresponding author: Ying Yang.)*

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

Lili Pan is with the Department of Mathematics, Shandong University of Technology, Zibo 255049, China (e-mail: panlili1979@163.com).

Ying Yang is with the Department of Mechanics and Engineering Science, Peking University, Beijing 100871, China (e-mail: yy@pku.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/TNNLS.2022.3201881.

Digital Object Identifier 10.1109/TNNLS.2022.3201881

associated works show that, compared with the existing $\ell_{2,1}$-norm regularized variants, the $\ell_{2,0}$-norm constrained optimization has the following two benefits: 1) easier to determine variables and 2) available faster optimization algorithms. Although the $\ell_{2,0}$-norm constrained optimization has been combined with deep CCA [38] and generalized CCA [19], they lack detailed convergence guarantee.

Motivated by the discussion mentioned above, a joint sparse constrained CCA (JSCCCA)-based FD approach by constraining the $\ell_{2,0}$-norm optimization, denoted as JSCCCA, is proposed, which has not been systematically considered before. To deal with the proposed JSCCCA, an efficient optimization algorithm is designed, along with a detailed convergence analysis. Furthermore, the advantages of the proposed JSCCCA over some existing state-of-the-art approaches, including CCA, SCCA, and JSCCA, are verified by sufficient numerical examples.

Compared with the previous work, the main contributions of this article are summarized in the following three parts.

1) We propose an efficient CCA formulation by exploiting the joint sparse structure of variables.
2) We develop a fast iterative optimization algorithm using alternating minimization techniques.
3) We prove that the generated sequence can converge to a stationary point after finite iterations.

The rest of this article is structured as follows. Section II presents related notations and preliminaries. The optimization algorithm and its detailed theoretical analysis are provided in Sections III and IV, respectively. Section V validates the efficiency of different datasets. Finally, Section VI concludes this article with some remarks.

II. Notations and Preliminaries

This section first introduces some notations used throughout this article and then gives related preliminaries.

A. Notations

In this article, all matrices and vectors are represented by boldface letters, and all scalars are represented by lowercase letters. Let $\mathbb{R}^{N\times p}$ characterize the set of all $N \times p$ matrices. For a matrix $\mathbf{Z} \in \mathbb{R}^{N\times p}$, let $\mathbf{z}_i$ indicate its ith row and z_{ij} indicate its ijth element. The Frobenius norm of $\mathbf{Z}$ is $\|\mathbf{Z}\|_F^2 = \sum_{i=1}^{N}\sum_{j=1}^{p} z_{ij}^2$. The $\ell_{2,0}$-norm of $\mathbf{Z}$ is $\|\mathbf{Z}\|_{2,0} = |\{i : \|\mathbf{z}_i\|_2 \neq 0\}|$, where $|\cdot|$ denotes the cardinality. The support set of $\mathbf{Z}$ is $\text{supp}(\mathbf{Z}) = \{i : \|\mathbf{z}_i\|_2 \neq 0\}$. The tangent cone to set $\{\mathbf{Z} \mid \|\mathbf{Z}\|_{2,0} \leq s_1\}$ at $\mathbf{Z}$ is $\{\mathbf{U} \mid \mathbf{u}_i = 0,\ i \notin \text{supp}(\mathbf{Z})\}$. Moreover, for a closed and nonempty set Ω, the projection of $\mathbf{Z}$ onto Ω is $\Pi_\Omega(\mathbf{Z}) = \arg\min\{\|\mathbf{Z} - \mathbf{U}\|_F,\ \mathbf{U} \in \Omega\}$. In addition, for a manifold $\mathcal{M}$, the tangent space at $\mathbf{Z}$ is $\mathcal{T}_\mathcal{M}(\mathbf{Z})$. Furthermore, the restriction of $\mathbf{V} \in \mathcal{T}_\mathcal{M}(\mathbf{Z})$ onto $\mathbf{Z} \in \mathcal{M}$ is $\mathcal{R}_\mathbf{Z}(\mathbf{V})$.

B. Preliminaries

In sparse optimization, the restricted isometry property (RIP) is frequently used to guarantee the convergence [39]. Therefore, we first generalize this definition to the proposed JSCCCA.

Definition 1: For a matrix $\mathbf{Z}$, the scalable RIP (SRIP) with the upper isometry constant C_{2s} and the lower isometry constant c_{2s} is satisfied if, for any $\mathbf{u}$ with $\|\mathbf{u}\|_0 \leq 2s$, it holds

$$c_{2s}\|\mathbf{u}\|^2 \leq \|\mathbf{Z}\mathbf{u}\|^2 \leq C_{2s}\|\mathbf{u}\|^2.$$

In addition, if $\mathbf{Z}$ has SRIP with constants C_{2s} and c_{2s}, then, for any matrix $\mathbf{A}$ satisfying the joint sparse constraint $\|\mathbf{A}\|_{2,0} \leq 2s$, it derives that

$$c_{2s}\|\mathbf{A}\|_F^2 \leq \|\mathbf{Z}\mathbf{A}\|_F^2 \leq C_{2s}\|\mathbf{A}\|_F^2.$$

Next, we should address the definition of retraction in order to characterize the convergence analysis in manifold learning [40]. From the conceptual point of view, a retraction is indeed a projection from the tangent space to the manifold with a local rigidity condition that maintains the gradient at the given point. More discussions about the retraction can be found in [41] and the references therein.

Definition 2: For a manifold $\mathcal{M}$, a retraction from the tangent space $\mathcal{T}_\mathcal{M}$ onto $\mathcal{M}$ satisfies: 1) $\mathcal{R}_\mathbf{Z}(0) = \mathbf{Z}$ and 2) $\lim_{\mathbf{V}\to 0}(\|\mathcal{R}_\mathbf{Z}(\mathbf{V}) - \mathbf{Z} - \mathbf{V}\|_F/\|\mathbf{V}\|_F) = 0$ for all $\mathbf{Z} \in \mathcal{M}$ and $\mathbf{V} \in \mathcal{T}_\mathcal{M}(\mathbf{Z})$. Here, 0 is the zero element of $\mathcal{T}_\mathcal{M}(\mathbf{Z})$.

According to Boumal *et al.* [42], the retraction has the following property, which is often used to describe the boundedness of variables.

Lemma 3: For a manifold $\mathcal{M}$, there exist two constants $M_1 > 0$ and $M_2 > 0$ such that: 1) $\|\mathcal{R}_\mathbf{Z}(\mathbf{V}) - \mathbf{Z}\|_F \leq M_1\|\mathbf{V}\|_F$ and 2) $\|\mathcal{R}_\mathbf{Z}(\mathbf{V}) - (\mathbf{Z} + \mathbf{V})\|_F \leq M_2\|\mathbf{V}\|_F^2$ for all $\mathbf{Z} \in \mathcal{M}$ and $\mathbf{V} \in \mathcal{T}_\mathcal{M}(\mathbf{Z})$.

Finally, we would like to point out that, since the proposed JSCCCA in this article involves both joint sparse and manifold constraints, which brings significant challenges to the convergence analysis, the above definitions and lemma play a crucial role in the following discussion.

III. Joint Sparse Constrained CCA

This section first introduces the proposed JSCCCA model. To deal with it, an alternating minimization algorithm (AMA) using the IIHT and the manifold constrained gradient descent (MCGD) method is presented to approximate the optimal solutions.

A. Model Construction

Assume that $\mathbf{X} \in \mathbb{R}^{N\times p}$ and $\mathbf{Y} \in \mathbb{R}^{N\times q}$ are the given inputs and outputs, where N denotes the number of samples and p and q denote the numbers of variables. In mathematics, the classical CCA model can be described by

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B}} \quad & -\frac{1}{N}\text{Tr}(\mathbf{A}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{B}) \\ \text{s.t.} \quad & \mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{A} = \mathbf{I}, \quad \mathbf{B}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{B} = \mathbf{I} \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{p\times m}$ and $\mathbf{B} \in \mathbb{R}^{q\times m}$ are the canonical matrices, and $\mathbf{I} \in \mathbb{R}^{m\times m}$ is the identity matrix ($m \leq p, q$). The target of CCA is to seek a pair of canonical matrices $\mathbf{A}$ and $\mathbf{B}$ to make $\mathbf{X}\mathbf{A}$ and $\mathbf{Y}\mathbf{B}$ have the maximum correlation.

In order to exploit the joint sparse information of $\mathbf{X}$ and $\mathbf{Y}$, we construct a JSCCCA model, which is given by

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B}} \quad & -\frac{1}{N}\text{Tr}(\mathbf{A}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{B}) \\ \text{s.t.} \quad & \mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{A} = \mathbf{I}, \quad \mathbf{B}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{B} = \mathbf{I} \\ & \|\mathbf{A}\|_{2,0} \leq s_1, \quad \|\mathbf{B}\|_{2,0} \leq s_2 \end{aligned} \tag{2}$$

where $s_1, s_2 > 0$ are the joint sparse parameters, and $\|\cdot\|_{2,0}$ is the $\ell_{2,0}$-norm defined in Section II.

Different from JSCCA studied in [24], [25], [26], and [27], the proposed JSCCCA in (2) considers a different formulation, which shares more advantages, mainly shown in the following two aspects.

Algorithm 1 AMA for Solving (4)

Input: Data $\mathbf{X}, \mathbf{Y}$, parameter $\beta > 0$, joint sparse $s_1, s_2 > 0$, and $\varepsilon = 10^{-5}$.

Initialize: $(\mathbf{A}^0, \mathbf{B}^0, \mathbf{C}^0, \mathbf{D}^0)$.

While not converged **do**

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

$$\mathbf{A}^{k+1} \in \arg\min_{\mathbf{A}\in\mathcal{S}_1} F(\mathbf{A}, \mathbf{B}^k, \mathbf{C}^k, \mathbf{D}^k).$$

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

$$\mathbf{B}^{k+1} \in \arg\min_{\mathbf{B}\in\mathcal{S}_2} F(\mathbf{A}^{k+1}, \mathbf{B}, \mathbf{C}^k, \mathbf{D}^k).$$

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

$$\mathbf{C}^{k+1} \in \arg\min_{\mathbf{C}\in\mathcal{M}_1} F(\mathbf{A}^{k+1}, \mathbf{B}^{k+1}, \mathbf{C}, \mathbf{D}^k).$$

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

$$\mathbf{D}^{k+1} \in \arg\min_{\mathbf{D}\in\mathcal{M}_2} F(\mathbf{A}^{k+1}, \mathbf{B}^{k+1}, \mathbf{C}^{k+1}, \mathbf{D}).$$

5: Check convergence: if

$$|F(\mathbf{A}^{k+1}, \mathbf{B}^{k+1}, \mathbf{C}^{k+1}, \mathbf{D}^{k+1}) - F(\mathbf{A}^k, \mathbf{B}^k, \mathbf{C}^k, \mathbf{D}^k)| \le \varepsilon$$

then stop.

End while

Output: $(\mathbf{A}^{k+1}, \mathbf{B}^{k+1}, \mathbf{C}^{k+1}, \mathbf{D}^{k+1})$.

1) JSCCCA uses the $\ell_{2,0}$-norm to measure rowwise sparse, while JSCCA applies the $\ell_{2,1}$-norm by approximating the $\ell_{2,0}$-norm. In fact, the $\ell_{2,0}$-norm is the essential description, but finding solutions to the proposed JSCCCA is more complicated due to its discontinuity and nonconvexity. Therefore, we develop an efficient optimization algorithm to solve it in Section III-B.
2) JSCCCA constructs the $\ell_{2,0}$-norm constrained optimization problem, while JSCCCA studies the $\ell_{2,1}$-norm regularized version. In comparison, the convergence analysis of the proposed JSCCCA is more difficult. We prove that the objective is strictly decreasing, and the generated sequence converges to a stationary point in Section IV.

B. Optimization Algorithm

From an optimization perspective, the proposed JSCCCA in (2) is not trivial for directly solving variables $\mathbf{A}$ and $\mathbf{B}$. Inspired by Boyd *et al.* [43], we first introduce matrices $\mathbf{C}$ and $\mathbf{D}$ to make the constraints separable, which gives the equivalent form

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}} \quad & -\frac{1}{N}\mathrm{Tr}(\mathbf{C}^\top\mathbf{D}) \\ \text{s.t.} \quad & \mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{A} = \mathbf{I}, \quad \mathbf{B}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{B} = \mathbf{I} \\ & \|\mathbf{A}\|_{2,0} \le s_1, \quad \|\mathbf{B}\|_{2,0} \le s_2 \\ & \mathbf{X}\mathbf{A} = \mathbf{C}, \quad \mathbf{Y}\mathbf{B} = \mathbf{D}. \end{aligned} \tag{3}$$

The regularized version of (3) can be formulated as

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}} \quad & -\frac{1}{N}\mathrm{Tr}(\mathbf{C}^\top\mathbf{D}) + \frac{\beta}{2}\|\mathbf{X}\mathbf{A} - \mathbf{C}\|_F^2 + \frac{\beta}{2}\|\mathbf{Y}\mathbf{B} - \mathbf{D}\|_F^2 \\ \text{s.t.} \quad & \|\mathbf{A}\|_{2,0} \le s_1, \quad \|\mathbf{B}\|_{2,0} \le s_2 \\ & \mathbf{C}^\top\mathbf{C} = \mathbf{I}, \quad \mathbf{D}^\top\mathbf{D} = \mathbf{I} \end{aligned} \tag{4}$$

Algorithm 2 IIHT for Solving (5)

Input: Data $\mathbf{X}$, $\mathbf{C}$, parameters $\eta \in (0, 1)$, $\sigma > 0$, joint sparse s_1, and $\varepsilon_1 = 10^{-3}$.

Initialize: $\mathbf{A}^0 = \mathbf{0}$, $\alpha_0 \in (0, 1/C_{2s_1})$.

While not converged **do**

1: Compute

$$\mathbf{A}^{k+1} = \Pi_{\mathcal{S}_1}\big(\mathbf{A}^k - \alpha_k \nabla G(\mathbf{A}^k)\big),$$

where $\alpha_k = \alpha_0\eta^{q_k}$ and q_k is the smallest nonnegative integer q such that

$$G\big(\mathbf{A}^k(\alpha_0\eta^q)\big) \le G(\mathbf{A}^k) - \frac{\sigma}{2}\|\mathbf{A}^k(\alpha_0\eta^q) - \mathbf{A}^k\|_F^2$$

with $\mathbf{A}^k(\alpha) = \Pi_{\mathcal{S}_1}(\mathbf{A}^k - \alpha\nabla G(\mathbf{A}^k))$.

2: Check convergence: if

$$\|\mathbf{A}^{k+1} - \mathbf{A}^k\|_F / \|\mathbf{A}^k\|_F \le \varepsilon_1,$$

then stop.

End while

Output: $\mathbf{A}^{k+1}$.

where $\beta > 0$ is the penalty parameter. When β is large enough, (4) is equivalent to (3).

Now, it can be updated by minimizing the objective function with other variables fixed in a Gauss-Seidel manner, which is called AMA. For notation simplicity, let the objective function of (4) be $F(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ and the constraints of (4) be

$$\begin{aligned}\mathcal{S}_1 &= \{\mathbf{A} \mid \|\mathbf{A}\|_{2,0} \le s_1\}, \quad \mathcal{S}_2 = \{\mathbf{B} \mid \|\mathbf{B}\|_{2,0} \le s_2\} \\ \mathcal{M}_1 &= \{\mathbf{C} \mid \mathbf{C}^\top\mathbf{C} = \mathbf{I}\}, \quad \mathcal{M}_2 = \{\mathbf{D} \mid \mathbf{D}^\top\mathbf{D} = \mathbf{I}\}.\end{aligned}$$

Therefore, the iterative scheme for solving (4) can be presented in Algorithm 1.

In the following, we will explain how to efficiently compute $\mathbf{A}$-subproblem and $\mathbf{C}$-subproblem, and these algorithms work similarly for the $\mathbf{B}$-subproblem and the $\mathbf{D}$-subproblem.

1) When $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ have been computed, the $\mathbf{A}$-subproblem can be simplified to

$$\min_{\mathbf{A}\in\mathcal{S}_1} G(\mathbf{A}) = \frac{1}{2}\|\mathbf{X}\mathbf{A} - \mathbf{C}\|_F^2. \tag{5}$$

Estimating the variable $\mathbf{A} \in \mathcal{S}_1$ is a not easy task since (5) is NP-hard in general. According to Pan *et al.* [30], an efficient IIHT can be used. It first calculates the gradient descent at $\mathbf{A}^k$ with a step size of $\alpha_k > 0$ and then adopts a truncation projection to obtain its approximate solution. Algorithm 2 provides a detailed framework for solving the $\mathbf{A}$-subproblem, where C_{2s_1} is the upper isometry constant of $\mathbf{X}$ and $\Pi_{\mathcal{S}_1}(\cdot)$ sets all but the s_1 largest rows to zero in the sense of ℓ_2-norm.

2) When $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}$ have been computed, the $\mathbf{C}$-subproblem can be solved by

$$\min_{\mathbf{C}\in\mathcal{M}_1} H(\mathbf{C}) = -\frac{1}{N}\mathrm{Tr}(\mathbf{C}^\top\mathbf{D}) + \frac{\beta}{2}\|\mathbf{X}\mathbf{A} - \mathbf{C}\|_F^2. \tag{7}$$

However, the problem is nonconvex because there exists a manifold constraint $\mathbf{C} \in \mathcal{M}_1$. This makes this subproblem hard to compute. Although singular value decomposition (SVD) can be applied [44], it brings tremendous computation, which limits the scalability

Algorithm 3 MCGD for Solving (7)

Input: Data $\mathbf{X}$, $\mathbf{A}$, $\mathbf{D}$, parameters $\eta \in (0,1)$, $\beta > 0$, $\delta > 0$, and $\varepsilon_2 = 10^{-3}$.

Initialize: $\mathbf{C}^0 = \mathbf{0}$, γ_0.

While not converged **do**

1: Compute

$$\mathbf{V}^k \in \arg\min_{\mathbf{V}\in\mathcal{T}_{\mathcal{M}_1}(\mathbf{C}^k)} -\frac{1}{N}\mathrm{Tr}((\mathbf{C}^k+\mathbf{V})^T\mathbf{D}) + \frac{\beta}{2}\|\mathbf{X}\mathbf{A}-\mathbf{C}^k-\mathbf{V}\|_F^2. \tag{6}$$

2: Compute

$$\mathbf{C}^{k+1} = \mathcal{R}_{\mathbf{C}^k}\left(\gamma_k \mathbf{V}^k\right),$$

where $\gamma_k = \gamma_1 \eta^{q_k}$ and q_k is the smallest nonnegative integer q such that

$$H\left(\mathbf{C}^{k+1}\right) \le H(\mathbf{C}^k) - \delta\gamma_1\eta^q\|\mathbf{V}^k\|_F^2.$$

3: Check convergence: if

$$\|\mathbf{C}^{k+1}-\mathbf{C}^k\|_F / \|\mathbf{C}^k\|_F \le \varepsilon_2,$$

then stop.

End while

Output: $\mathbf{C}^{k+1}$.

to large-scale applications [40], [45]. According to Definition 2, a retraction on the manifold $\mathcal{M}_1$ can be set as

$$\mathcal{R}_{\mathbf{X}}(\mathbf{V}) = (\mathbf{X}+\mathbf{V})(\mathbf{I}+\mathbf{V}^\top\mathbf{V})^{-1/2}.$$

Afterward, the solution for (7) can be obtained using the manifold constrained gradient descent (MCGD) method [25], [46]. Refer to Algorithm 3 for the general iterative optimization scheme.

Therefore, the proposed JSCCCA in (4) [also (2)] can be efficiently solved by Algorithms 1–3, where the **A**-subproblem and the **B**-subproblem can be computed according to Algorithm 2, and the **C**-subproblem and the **D**-subproblem can be computed according to Algorithm 3.

IV. Convergence Analysis

This section first defines the optimality condition and then establishes the detailed convergence guarantee theoretically.

A. Optimality Condition

Let $\mathbf{W} = (\mathbf{A}^\top, \mathbf{B}^\top, \mathbf{C}^\top, \mathbf{D}^\top)^\top$, and thus, the objective is $F(\mathbf{W})$. Note that $F(\mathbf{W})$ is continuously differentiable, and the gradient can be given by

$$\nabla F(\mathbf{W}) = \begin{bmatrix} \frac{\partial}{\partial \mathbf{A}}F(\mathbf{W}) \\ \frac{\partial}{\partial \mathbf{B}}F(\mathbf{W}) \\ \frac{\partial}{\partial \mathbf{C}}F(\mathbf{W}) \\ \frac{\partial}{\partial \mathbf{D}}F(\mathbf{W}) \end{bmatrix} = \begin{bmatrix} \beta\mathbf{X}^\top(\mathbf{X}\mathbf{A}-\mathbf{C}) \\ \beta\mathbf{Y}^\top(\mathbf{Y}\mathbf{B}-\mathbf{D}) \\ -\mathbf{D}-\beta(\mathbf{X}\mathbf{A}-\mathbf{C}) \\ -\mathbf{C}-\beta(\mathbf{Y}\mathbf{B}-\mathbf{D}) \end{bmatrix}.$$

Let $\mathbf{\Omega} = \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{M}_1 \times \mathcal{M}_2$; then, the tangent space of $\mathbf{\Omega}$ at $\mathbf{W}$ can be expressed as

$$\mathcal{T}_{\mathbf{\Omega}}(\mathbf{W}) = \mathcal{T}_{\mathcal{S}_1}(\mathbf{A}) \times \mathcal{T}_{\mathcal{S}_2}(\mathbf{B}) \times \mathcal{T}_{\mathcal{M}_1}(\mathbf{C}) \times \mathcal{T}_{\mathcal{M}_2}(\mathbf{D})$$

in which $\mathcal{T}_{\mathcal{S}_1}(\mathbf{A})$ and $\mathcal{T}_{\mathcal{S}_2}(\mathbf{B})$ are the tangent cones for joint sparse sets $\mathcal{S}_1$ and $\mathcal{S}_2$, and $\mathcal{T}_{\mathcal{M}_1}(\mathbf{C})$ and $\mathcal{T}_{\mathcal{M}_2}(\mathbf{D})$ are the tangent spaces of manifold sets $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively.

According to Zhang *et al.* [47], the first-order optimality condition associated with (4) can be defined as follows.

Definition 4: We call $\mathbf{W} \in \mathbf{\Omega}$ a stationary point of (4) if it satisfies

$$\mathbf{0} = \mathrm{grad}F(\mathbf{W})$$

where $\mathrm{grad}F(\mathbf{W}) = \Pi_{\mathcal{T}_{\mathbf{\Omega}}(\mathbf{W})}\left(\nabla F(\mathbf{W})\right)$ is the Riemannian gradient of F at $\mathbf{W}$.

From the expression of $\nabla F(\mathbf{W})$ and the tangent space of $\mathbf{\Omega}$, the above first-order condition is equivalent to

$$\begin{cases} (\mathbf{X}^\top(\mathbf{X}\mathbf{A}-\mathbf{C}))_i = 0, & i \in \mathrm{supp}(\mathbf{A}) \\ (\mathbf{Y}^\top(\mathbf{Y}\mathbf{B}-\mathbf{D}))_i = 0, & i \in \mathrm{supp}(\mathbf{B}) \\ \Pi_{\mathcal{T}_{\mathcal{M}_1}(\mathbf{C})}(-\mathbf{D}-\beta(\mathbf{X}\mathbf{A}-\mathbf{C})) = \mathbf{0} \\ \Pi_{\mathcal{T}_{\mathcal{M}_2}(\mathbf{D})}(-\mathbf{C}-\beta(\mathbf{Y}\mathbf{B}-\mathbf{D})) = \mathbf{0}. \end{cases}$$

Before proceeding, another definition of ϵ-stationary point should be given.

Definition 5: We call $(\mathbf{A}^k, \mathbf{B}^k, \mathbf{C}^k, \mathbf{D}^k)$ an ϵ-stationary point of (4) if it satisfies

$$\left\|\left(\frac{\partial}{\partial \mathbf{A}}F(\mathbf{W})\right)_{\mathrm{supp}(\mathbf{A}^k)}\right\|_F^2 + \left\|\mathbf{V}_{\mathbf{C}}^k\right\|_F^2 + \left\|\left(\frac{\partial}{\partial \mathbf{B}}F(\mathbf{W})\right)_{\mathrm{supp}(\mathbf{B}^k)}\right\|_F^2 + \left\|\mathbf{V}_{\mathbf{D}}^k\right\|_F^2 \le \epsilon$$

where $\mathbf{V}_{\mathbf{C}}^k$ and $\mathbf{V}_{\mathbf{D}}^k$ are the $\mathbf{V}$ in Algorithm 3 for solving the **C**-subproblem and the **D**-subproblem.

It is worth noting that the ϵ-stationary point is an approximation of the stationary point in Definition 4, which turns out to be a stationary point of (4) if $\epsilon = 0$.

B. Theoretical Guarantee

Compared with JSCCA in [24], there exist two challenging issues for analyzing the convergence of the proposed JSCCCA. One is that it has four variables. The other one is that all the constraints are nonconvex. To deal with these difficulties, we first demonstrate that the successive differences of variables are bounded and then prove the convergence property.

Theorem 6: Suppose that $\{\mathbf{A}^k\}$ is a generated sequence by Algorithm 2, and $\mathbf{X}$ has an upper restricted isometry constant C_{2s_1}. Whenever $0 < \alpha_k \le (1/C_{2s_1} + \sigma)$, it holds that

$$G(\mathbf{A}^{k+1}) \le G(\mathbf{A}^k) - \frac{\sigma}{2}\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2. \tag{8}$$

Furthermore, when $k \to \infty$, it is easy to conclude the following.

1) $\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F \to 0$.
2) $\|(\nabla G(\mathbf{A}^k))_{\mathrm{supp}(\mathbf{A}^k)}\|_F \to 0$.

Proof: From $\mathbf{A}^{k+1} = \Pi_{\mathcal{S}_1}(\mathbf{A}^k - \alpha_k\nabla G(\mathbf{A}^k))$, it derives

$$\|\mathbf{A}^{k+1}-\mathbf{A}^k+\alpha_k\nabla G(\mathbf{A}^k)\|_F^2 \le \|\mathbf{A}^k-\mathbf{A}^k+\alpha_k\nabla G(\mathbf{A}^k)\|_F^2$$

which gives

$$\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2 \le -2\alpha_k\langle\nabla G(\mathbf{A}^k),\mathbf{A}^{k+1}-\mathbf{A}^k\rangle.$$

Then, for $0<\alpha_k\le(1/C_{2s_1}+\sigma)$, it follows that

$$\begin{aligned}G(\mathbf{A}^{k+1}) &\le G(\mathbf{A}^k)+\langle\nabla G(\mathbf{A}^k),\mathbf{A}^{k+1}-\mathbf{A}^k\rangle\\ &\quad+\frac{C_{2s_1}}{2}\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2\\ &\le G(\mathbf{A}^k)-\frac{1}{2\alpha_k}\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2\\ &\quad+\frac{C_{2s_1}}{2}\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2\\ &\le G(\mathbf{A}^k)-\frac{\sigma}{2}\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2.\end{aligned}$$

This shows that (8) is guaranteed. It then obtains

$$\sum_{k=0}^{\infty}\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2\le\frac{2}{\sigma}\sum_{k=0}^{\infty}(G(\mathbf{A}^k)-G(\mathbf{A}^{k+1}))<+\infty$$

which implies $\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2\to 0$.

Moreover, the projection $\mathbf{A}^{k+1}=\Pi_{\mathcal{S}_1}(\mathbf{A}^k-\alpha_k\nabla G(\mathbf{A}^k))$ sets all but s_1 largest absolute components of $\mathbf{A}^k-\alpha_k\nabla G(\mathbf{A}^k)$ to zero; hence,

$$\|(\nabla G(\mathbf{A}^k))_{\mathrm{supp}(\mathbf{A}^k)}\|_F\to 0$$

and the desired conclusion follows. ∎

Theorem 6 has proved that the sequence of the variable $\mathbf{A}$ is bounded, provided that $0<\alpha_k\le(1/C_{2s_1}+\sigma)$ is satisfied. This idea comes from [39] but with some differences because the $\ell_{2,0}$-norm constraint is considered in the proposed JSCCCA. Next, the boundedness of the variable $\mathbf{C}$ will be established.

Theorem 7: Suppose that $\{\mathbf{C}^k\}$ is a generated sequence by Algorithm 3. Then, there exist $\bar\gamma_1>0$ and $\bar\beta>0$ such that

$$H(\mathbf{C}^{k+1})-H(\mathbf{C}^k)\le-\bar\beta\|\mathbf{V}^k\|_F^2.$$

Proof: Let $\mathbf{C}_+^{k+1}=\mathbf{C}^k+\gamma_k\mathbf{V}^k$ and $\nabla H(\mathbf{C})=-(1/N)\mathbf{D}-\beta(\mathbf{XA}-\mathbf{C})$. Then, the following relationships hold:

$$\begin{aligned}&H(\mathbf{C}^{k+1})-H(\mathbf{C}^k)\\ &=\langle\nabla H(\mathbf{C}^k),\mathbf{C}^{k+1}-\mathbf{C}^k\rangle+\frac{\beta}{2}\|\mathbf{C}^{k+1}-\mathbf{C}^k\|_F^2\\ &=\left\langle\nabla H(\mathbf{C}^k),\mathbf{C}^{k+1}-\mathbf{C}_+^{k+1}\right\rangle+\left\langle\nabla H(\mathbf{C}^k),\mathbf{C}_+^{k+1}-\mathbf{C}^k\right\rangle\\ &\quad+\frac{\beta}{2}\|\mathbf{C}^{k+1}-\mathbf{C}^k\|_F^2\\ &\le\|\nabla H(\mathbf{C}^k)\|_F\left\|\mathbf{C}^{k+1}-\mathbf{C}_+^{k+1}\right\|_F\\ &\quad-\langle\mathbf{D}-\beta(\mathbf{C}^k-\mathbf{XA}),\gamma_k\mathbf{V}^k\rangle+\frac{\beta}{2}\left\|\mathbf{C}^{k+1}-\mathbf{C}^k\right\|_F^2.\end{aligned}$$

Since $\nabla H(\mathbf{C}^k)$ is continuous on the compact set $\mathcal{M}_1$, there exists a constant $\overline{M}>0$ such that $\|\nabla H(\mathbf{C})\|_F\le\overline{M}$ for all $\mathbf{C}\in\mathcal{M}_1$. The optimality condition of (6) yields that

$$0=\Pi_{\mathcal{T}_{\mathcal{M}_1}(\mathbf{C}^k)}(-\mathbf{D}+\beta(\mathbf{V}^k+\mathbf{C}^k-\mathbf{XA})).$$

It is equivalent to

$$0=\langle-\mathbf{D}+\beta(\mathbf{V}^k+\mathbf{C}^k-\mathbf{XA}),\mathbf{V}^k\rangle$$

which means that

$$-\langle\mathbf{D}-\beta(\mathbf{C}^k-\mathbf{XA}),\mathbf{V}^k\rangle=-\beta\langle\mathbf{V}^k,\mathbf{V}^k\rangle.$$

From Lemma 3, there exist $M_1>0$ and $M_2>0$ satisfying

$$\begin{aligned}\left\|\mathbf{C}^{k+1}-\mathbf{C}^k\right\|_F&\le M_1\left\|\gamma_k\mathbf{V}^k\right\|_F\\ \left\|\mathbf{C}^{k+1}-\mathbf{C}_+^{k+1}\right\|_F&\le M_2\left\|\gamma_k\mathbf{V}^k\right\|_F^2.\end{aligned}$$

This, together with (9), follows that

$$\begin{aligned}&H(\mathbf{C}^{k+1})-H(\mathbf{C}^k)\\ &\le-\left(\beta-\overline{M}M_2\gamma_k^2-\frac{\beta}{2}M_1^2\gamma_k^2\right)\|\mathbf{V}^k\|_F^2\\ &<-\bar\beta\|\mathbf{V}^k\|_F^2\end{aligned}$$

whenever $0<\gamma_k<\bar\gamma_1:=\gamma\left((\beta/\overline{M}M_2+(\beta/2)M_1^2)\right)^{1/2}$, and thus, the proof is completed. ∎

With the above two theorems, the following theorem will illustrate that the generated sequence is convergent under mild conditions. The proof makes full use of the retraction property, which was previously adopted in [25] to prove the convergence of JSCCA. Though it follows a similar line, JSCCCA is much more intricate because it involves the $\ell_{2,0}$-norm constraints, and the successive changes of variables are not easy to bound.

Theorem 8: Suppose that $\{(\mathbf{A}^k,\mathbf{B}^k,\mathbf{C}^k,\mathbf{D}^k)\}$ is a sequence generated by Algorithm 1. Moreover, $\mathbf{X}$ and $\mathbf{Y}$ satisfy SRIP with constants C_{2s_1}, c_{2s_1} and C_{2s_2}, c_{2s_2}, respectively. Then, the sequence converges to a stationary point of (4). Furthermore, Algorithm 1 returns an ϵ-stationary point in at most $\lfloor(F(\mathbf{A}^0,\mathbf{B}^0,\mathbf{C}^0,\mathbf{D}^0)-F^*)/((\bar\sigma_1+\bar\sigma_2+2\bar\beta)\epsilon)\rfloor+1$ iterations, where F^* denotes a lower bound of the optimal value of (4) with $\bar\sigma_1$, $\bar\sigma_2$, and $\bar\beta$ being constants.

Proof: From Theorems 6 and 7, the generated sequence $\{(\mathbf{A}^k,\mathbf{B}^k,\mathbf{C}^k,\mathbf{D}^k)\}$ satisfies

$$\begin{aligned}&F(\mathbf{A}^{k+1},\mathbf{B}^k,\mathbf{C}^k,\mathbf{D}^k)\\ &\quad\le F(\mathbf{A}^k,\mathbf{B}^k,\mathbf{C}^k,\mathbf{D}^k)-\frac{\sigma_1}{2}\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2\\ &F(\mathbf{A}^{k+1},\mathbf{B}^{k+1},\mathbf{C}^k,\mathbf{D}^k)\\ &\quad\le F(\mathbf{A}^{k+1},\mathbf{B}^k,\mathbf{C}^k,\mathbf{D}^k)-\frac{\sigma_2}{2}\|\mathbf{B}^{k+1}-\mathbf{B}^k\|_F^2\\ &F(\mathbf{A}^{k+1},\mathbf{B}^{k+1},\mathbf{C}^{k+1},\mathbf{D}^k)\\ &\quad\le F(\mathbf{A}^{k+1},\mathbf{B}^{k+1},\mathbf{C}^k,\mathbf{D}^k)-\bar\beta\left\|\mathbf{V}_C^k\right\|_F^2\\ &F(\mathbf{A}^{k+1},\mathbf{B}^{k+1},\mathbf{C}^{k+1},\mathbf{D}^{k+1})\\ &\quad\le F(\mathbf{A}^{k+1},\mathbf{B}^{k+1},\mathbf{C}^{k+1},\mathbf{D}^k)-\bar\beta\left\|\mathbf{V}_D^k\right\|_F^2.\end{aligned}$$

From the lower boundedness of $F(\mathbf{W})$, it follows that

$$\begin{aligned}\lim_{k\to\infty}&\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2+\|\mathbf{B}^{k+1}-\mathbf{B}^k\|_F^2\\ &+\bar\beta\left\|\mathbf{V}_C^k\right\|_F^2+\bar\beta\left\|\mathbf{V}_D^k\right\|_F^2=0.\end{aligned}$$

Since $\mathbf{X}$ and $\mathbf{Y}$ satisfy the SRIP and $\mathcal{M}_1$, and $\mathcal{M}_2$ are compact, there exists an accumulate point of the generated sequence $\{(\mathbf{A}^k,\mathbf{B}^k,\mathbf{C}^k,\mathbf{D}^k)\}$, denoted as $(\mathbf{A}^*,\mathbf{B}^*,\mathbf{C}^*,\mathbf{D}^*)$. This, together with Theorem 6, derives that

$$\begin{aligned}\|(\nabla G(\mathbf{A}^*))_{\mathrm{supp}(\mathbf{A}^*)}\|_F&=0\\ \|(\nabla G(\mathbf{B}^*))_{\mathrm{supp}(\mathbf{B}^*)}\|_F&=0.\end{aligned}$$

From Theorem 7, if $\|\mathbf{V}_C^*\|_F^2=0$ and $\|\mathbf{V}_D^*\|_F^2=0$, then

$$\begin{aligned}\Pi_{\mathcal{T}_{\mathcal{M}_1}(\mathbf{C}^*)}(-\mathbf{D}^*-\beta(\mathbf{XA}^*-\mathbf{C}^*))&=0\\ \Pi_{\mathcal{T}_{\mathcal{M}_2}(\mathbf{D}^*)}(-\mathbf{C}^*-\beta(\mathbf{YB}^*-\mathbf{D}^*))&=0.\end{aligned}$$

It is concluded immediately that $(\mathbf{A}^*,\mathbf{B}^*,\mathbf{C}^*,\mathbf{D}^*)$ is a stationary point of (4).

Suppose that the proposed algorithm cannot terminate after K iterations, i.e.,

$$\frac{\sigma_1}{2}\|\mathbf{A}^{k+1}-\mathbf{A}^k\|_F^2+\frac{\sigma_2}{2}\|\mathbf{B}^{k+1}-\mathbf{B}^k\|_F^2 + \bar{\beta}\left\|\mathbf{V}_C^k\right\|_F^2+\bar{\beta}\left\|\mathbf{V}_D^k\right\|_F^2>\epsilon$$

for all $k<K$. In addition,

$$\begin{aligned}&F(\mathbf{A}^0,\mathbf{B}^0,\mathbf{C}^0,\mathbf{D}^0)-F^*\\&\geq F(\mathbf{A}^0,\mathbf{B}^0,\mathbf{C}^0,\mathbf{D}^0)-F(\mathbf{A}^K,\mathbf{B}^K,\mathbf{C}^K,\mathbf{D}^K)\\&>(\bar{\sigma}_1+\bar{\sigma}_2+2\bar{\beta})K\epsilon.\end{aligned}$$

Therefore, the proposed algorithm can find an ϵ-stationary point after $\lfloor(F(\mathbf{A}^0,\mathbf{B}^0,\mathbf{C}^0,\mathbf{D}^0)-F^*)/((\bar{\sigma}_1+\bar{\sigma}_2+2\bar{\beta})\epsilon)\rfloor+1$ iterations. This completes the proof. ■

From Theorem 8, matrices $\mathbf{X}$ and $\mathbf{Y}$ should satisfy SRIP with constants C_{2s_1}, c_{2s_1} and C_{2s_2}, c_{2s_2}, which are easy to check in practice. Another issue is that all subproblems are nonconvex, and this results in that the proposed algorithm may converge to one of the local minimizers. Therefore, a good initialization will give satisfactory performance, which will be discussed in Section V.

V. Numerical Studies

This section conducts numerical experiments to demonstrate the superiority of the proposed JSCCCA over state-of-the-art approaches, including CCA, SCCA, and JSCCA. Although some nonlinear CCA-based approaches, such as [12], [14], and [16], are offered, these variants are not mentioned because this article only focuses on sparse linear approaches. Nevertheless, this idea can be easily extended to these variants.

First, Section V-A performs some simulation examples to illustrate that the proposed JSCCCA is fast. Then, Sections V-B and V-C conduct FD experiments on the benchmark Tennessee Eastman process (TEP) and a cylinder-piston process (CPP) to show that the proposed JSCCCA is efficient.

A. Simulation Examples

1) Dataset: According to Li *et al.* [39] and Cai *et al.* [49], datasets $\mathbf{X}$ and $\mathbf{Y}$ can be generated as follows:

$$\mathbf{X}=\mathbf{u}^\top(\mathbf{v}_1+\mathbf{e}_1),\quad \mathbf{Y}=\mathbf{u}^\top(\mathbf{v}_2+\mathbf{e}_2)$$

where $\mathbf{u}\in\mathbb{R}^{1\times 1000i}$ is a random vector that satisfies standard normal distributions, $\mathbf{v}_1,\mathbf{v}_2\in\mathbb{R}^{1\times 500j}$ are given by

$$\begin{aligned}\mathbf{v}_1&=[\underbrace{1,\ldots,1}_{50j}\ \underbrace{-1,\ldots,-1}_{50j}\ \underbrace{0,\ldots,0}_{400j}]\\\mathbf{v}_2&=[\underbrace{0,\ldots,0}_{400j}\ \underbrace{1,\ldots,1}_{50j}\ \underbrace{-1,\ldots,-1}_{50j}]\end{aligned}$$

and $\mathbf{e}_1,\mathbf{e}_2\in\mathbb{R}^{1\times 500j}$ are the noise vectors. If there exists no noise, the datasets $\mathbf{X}$ and $\mathbf{Y}$ are columnwise sparse, and then, the canonical matrices $\mathbf{A}$ and $\mathbf{B}$ are rowwise sparse. As a consequence, the generated datasets are suitable for validating the efficiency of the proposed approach.

For SCCA and JSCCA, the regularized parameters are determined using fivefold cross-validation. Moreover, the solution of CCA is used to initialize the proposed JSCCCA. For all the mentioned approaches, the maximum iteration number is set to 500, and the stopping criterion is defined in the algorithms. In addition, for all the testing cases, the sparse parameters s_1 and s_2 are chosen as 10.

TABLE I
Time (s) for the Simulation Dataset

Problem Scale	CCA	SCCA	JSCCA	JSCCCA
(1,000;500;500)	0.08	0.09	0.09	**0.03**
(5,000;500;500)	0.35	0.32	0.36	**0.09**
(10,000;500;500)	0.65	0.61	0.62	**0.12**
(50,000;500;500)	4.20	3.28	3.15	**0.46**
(100,000;500;500)	7.22	6.44	6.34	**0.89**
(1,000;2,500;2,500)	1.16	1.25	1.19	**0.13**
(5,000;2,500;2,500)	2.95	2.84	2.73	**0.20**
(10,000;2,500;2,500)	5.47	5.35	5.22	**0.61**
(50,000;2,500;2,500)	25.82	23.66	20.45	**1.89**
(100,000;2,500;2,500)	127.04	45.39	40.10	**3.60**
(1,000;5,000;5,000)	5.25	5.97	5.74	**0.09**
(5,000;5,000;5,000)	10.78	9.85	9.72	**0.38**
(10,000;5,000;5,000)	17.25	16.73	14.95	**0.78**
(50,000;5,000;5,000)	113.06	58.12	54.07	**3.01**
(100,000;5,000;5,000)	8331.17	5249.65	4,989.36	**8.03**

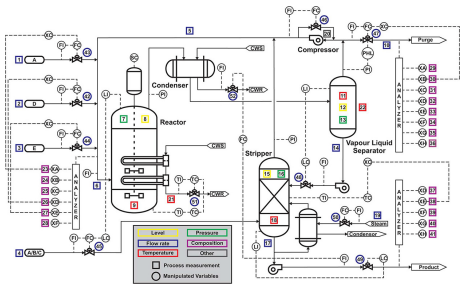


Fig. 1. Framework of the TEP.

2) Computational Results: Table I lists the running time in seconds for $i=1,5,10,50,100$ and $j=1,5,10$, where the first column presents the problem scale under different $(N;p;q)$. The best results are indicated in bold. It can be seen that the proposed JSCCCA needs less time compared to the existing CCA, SCCA, and JSCCA. In particular, when the problem scale is $(100,000;5,000;5,000)$, the improvement has reached more than 600 times because all subproblems in Algorithm 1 can be efficiently calculated by Algorithms 2 and 3. However, CCA, SCCA, and JSCCA involve the computation of SVDs, which are numerically expensive for large-scale datasets. This is why our proposed algorithm runs faster than the existing CCA-based approaches.

B. Benchmark Tennessee Eastman Process

1) Dataset: The TEP has been extensively used as a benchmark dataset for validating different FD techniques [49]. Fig. 1 draws a framework of the TEP. In the process, one fault-free dataset and 21 fault datasets are simulated. Table II provides the selected faults, where RCW denotes reactor cooling water and CCW denotes condenser cooling water. For each fault dataset, it collects 960 samples, and a fault is introduced into the process at the 161st sample. In general, the fault-free dataset is applied for off-line training, while fault datasets are used for online detection. As suggested in [27], 22 process variables are selected as $\mathbf{X}$, and 11 manipulated variables are selected as $\mathbf{Y}$.

2) Detection Strategy: According to Chen *et al.* [50], the residual vector for the testing sample vectors $\mathbf{x}$ and $\mathbf{y}$ can be constructed by the following generator:

$$\mathbf{r}=\mathbf{A}^\top\mathbf{x}-\mathbf{\Lambda}\mathbf{B}^\top\mathbf{y}$$

TABLE II
SELECTED FAULTS IN THE TEP

Fault No.	Description	Type
IDV(01)	A/C feed ratio	step change
IDV(02)	component B	step change
IDV(03)	feed D temperature	step change
IDV(04)	RCW inlet temperature	step change
IDV(05)	CCW inlet temperature	step change
IDV(06)	feed A loss	step change
IDV(07)	C header pressure loss	step change
IDV(08)	feed A−C components	random variation
IDV(09)	feed D temperature	random variation
IDV(10)	feed C temperature	random variation
IDV(11)	RCW inlet temperature	random variation
IDV(12)	CCW inlet temperature	random variation
IDV(13)	reaction kinetics	slow drift
IDV(14)	RCW valve	sticking
IDV(15)	CCW valve	sticking
IDV(16)	unknown fault	unknown
IDV(17)	unknown fault	unknown
IDV(18)	unknown fault	unknown
IDV(19)	unknown fault	unknown
IDV(20)	unknown fault	unknown
IDV(21)	unknown fault	constant

where $\mathbf{\Lambda} = \text{diag}(\rho_1, \ldots, \rho_\kappa)$ with $1 \geq \rho_1 \geq \cdots \geq \rho_\kappa \geq 0$ being the κ largest canonical correlation coefficients, which are obtained from the classical CCA. If no prior knowledge of the fault is available, the Hotelling T^2 test statistic obtains the best detection performance. Thus, the following T^2 statistic can be adopted as the form of:

$$T^2 = \mathbf{r}^\top (\mathbf{I} - \mathbf{\Lambda}^2)^{-1} \mathbf{r}.$$

The corresponding control limit can be chosen by the standard $\mathcal{X}^2$ distribution, i.e.,

$$J_{\text{th},T^2} = \mathcal{X}^2_\alpha(m)$$

where m is the freedom and α is the significance level. It is obvious that the control limit is the same for all approaches. After obtaining the T^2 statistic and the corresponding control limit, the FD logic can be given by

$$\begin{cases} T^2 > J_{\text{th},T^2} \Rightarrow \text{ fault occurs} \\ T^2 \leq J_{\text{th},T^2} \Rightarrow \text{ fault-free.} \end{cases}$$

To evaluate the detection performance, two indicators, i.e., the fault detection rate (FDR) and the false alarm rate (FAR) [2], are defined as

$$\text{FDR} = \text{prob}(T^2 > J_{\text{th},T^2} \mid f \neq 0)$$
$$\text{FAR} = \text{prob}(T^2 > J_{\text{th},T^2} \mid f = 0).$$

It is obvious that FDR represents the percentage of fault samples that are correctly detected under fault conditions ($f \neq 0$), while FAR represents the percentage of samples that are incorrectly identified as faults under fault-free conditions ($f = 0$). Thus, a higher FDR value or a lower FAR value indicates better detection performance. See Fig. 2 for visual illustration, where a fault is introduced at the 161st sample.

3) Detection Results: The detection results of the proposed JSCCCA and other CCA approaches in terms of FDR and FAR are listed in Table III. The best results are indicated in bold. As for FAR, the values of all compared approaches are relatively low. This reflects that the CCA-based approaches are useful for detecting false alarms in fault-free conditions.

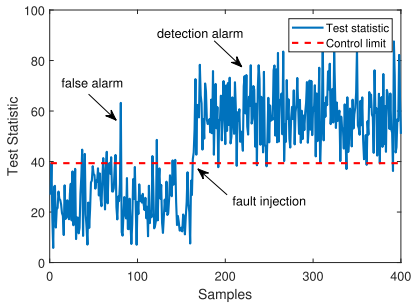


Fig. 2. Illustration of detection indicators.

Compared with the classical CCA, the FDR values of SCCA variants are larger. In particular, the gains of SCCA, JSCCA, and JSCCCA for fault IDV(10) are 3.88%, 5.30%, and 12.62%, respectively. This shows that sparse is beneficial for FD, which can reduce the noise and, thus, improve the representation of variables. The performance of JSCCA and JSCCCA concerning FDR is better than or as good as that of SCCA, which can be attributed to the fact that the joint sparse can preserve key variables and dropout the unimportant variables. Even for some difficult cases, such as fault IDV(03), the increases in JSCCA and JSCCCA are still positive in comparison to SCCA. For all faults in the TEP, the proposed JSCCCA obtains the best detection results. This considerable performance comes from the incorporation of the $\ell_{2,0}$-norm constrained optimization and CCA. Moreover, it is derived that the $\ell_{2,0}$-norm joint sparse constraint is more efficient than the $\ell_{2,1}$-norm regularization in the TEP.

To demonstrate the performance achieved by the proposed JSCCCA, three representative faults, i.e., IDV(02), IDV(10), and IDV(20), are chosen as examples. Fault IDV(02) is a step change, which is an easy fault for data-driven FD approaches. The T^2 test statistic and the corresponding control limit are presented in Fig. 3. It is found that all four approaches can detect IDV(02) at the 161st sample, and the FDR values are larger than 95%, which shows that the CCA-based approaches can obtain excellent detection results. However, compared with CCA, these sparse variants have fewer false alarms; see the subfigures in Fig. 3. Fault IDV(10) involves a random variation in the feed C temperature, which is a difficult task. The T^2 test statistic is given in Fig. 4. Compared with SCCA, both JSCCA and JSCCCA can achieve much more samples that exceed the control limit. It is convinced that the joint sparse is much more efficient than the sparse itself because the latent structures of variables are explored. Fault IDV(20) involves an unknown-type fault, and the detection results are presented in Fig. 5. Looking at the resulting values between 400 and 450 samples, JSCCCA is able to detect more fault samples than others, which illustrates that the proposed approach is efficient for detecting this fault.

4) Parameter Selection: Compared with the existing CCA-based FD approaches, parameters s_1 and s_2 should be selected carefully. If s_1 and s_2 are too large, all variables are extracted, which derives bad fault interpretation. If s_1 and s_2 are small, only a few variables are extracted, which makes detection performance poor. Therefore, in this article, one can possibly initialize the algorithm with small s_1 and s_2, and then increase them until obtaining the best performance.

For example, the FDR values of fault IDV(10) for different s_1 and fixed s_2 are shown in Fig. 6. It is obtained that, if s_1 is too small, the FDR value is less than 20%.

TABLE III
DETECTION RESULTS FOR THE TEP

Fault No.	CCA		SCCA		JSCCA		JSCCCA	
	FDR	FAR	FDR	FAR	FDR	FAR	FDR	FAR
IDV(01)	99.25%	**0.00%**	99.38%	**0.00%**	99.50%	**0.00%**	**99.62%**	**0.00%**
IDV(02)	98.62%	0.63%	99.25%	**0.00%**	99.25%	**0.00%**	**99.37%**	**0.00%**
IDV(03)	32.80%	2.50%	36.64%	1.88%	38.20%	**0.00%**	**43.98%**	**0.00%**
IDV(04)	**100%**	1.88%	**100%**	0.63%	**100%**	**0.00%**	**100%**	**0.00%**
IDV(05)	27.63%	1.88%	31.00%	0.63%	34.50%	**0.00%**	**38.36%**	**0.00%**
IDV(06)	99.75%	0.63%	99.90%	**0.00%**	**100%**	**0.00%**	**100%**	**0.00%**
IDV(07)	**100%**	1.88%	**100%**	0.63%	**100%**	0.63%	**100%**	**0.00%**
IDV(08)	93.25%	1.88%	95.00%	0.63%	95.24%	**0.00%**	**97.88%**	**0.00%**
IDV(09)	31.20%	3.13%	35.25%	2.50%	38.50%	0.63%	**39.87%**	**0.00%**
IDV(10)	27.50%	1.25%	34.82%	**0.63%**	36.24%	**0.63%**	**40.12%**	**0.63%**
IDV(11)	66.43%	0.63%	68.71%	**0.00%**	74.00%	**0.00%**	**77.89%**	**0.00%**
IDV(12)	90.87%	1.88%	93.87%	1.25%	94.50%	**0.63%**	**95.75%**	**0.63%**
IDV(13)	91.36%	0.63%	92.12%	0.63%	93.87%	**0.00%**	**95.61%**	**0.00%**
IDV(14)	86.00%	1.88%	87.25%	**0.63%**	88.12%	**0.63%**	**90.37%**	**0.63%**
IDV(15)	35.10%	2.50%	38.57%	1.25%	41.23%	0.63%	**43.57%**	**0.00%**
IDV(16)	15.78%	7.50%	18.23%	4.38%	22.75%	3.13%	**25.24%**	**1.25%**
IDV(17)	33.75%	3.13%	35.12%	3.13%	37.00%	3.13%	**39.75%**	**2.50%**
IDV(18)	87.88%	1.88%	90.18%	0.63%	92.54%	0.63%	**95.86%**	**0.00%**
IDV(19)	22.24%	**1.25%**	25.87%	**1.25%**	28.06%	**1.25%**	**29.78%**	**1.25%**
IDV(20)	47.50%	0.63%	52.75%	**0.00%**	54.37%	**0.00%**	**55.63%**	**0.00%**
IDV(21)	89.62%	1.25%	90.36%	1.25%	93.75%	**0.63%**	**96.85%**	**0.63%**

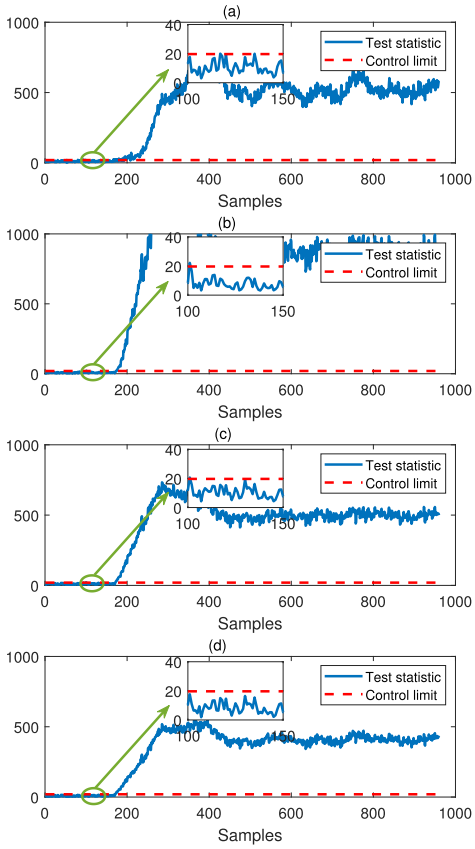


Fig. 3. Detection performance for IDV(02) in the TEP. (a) CCA. (b) SCCA. (c) JSCCA. (d) JSCCCA.

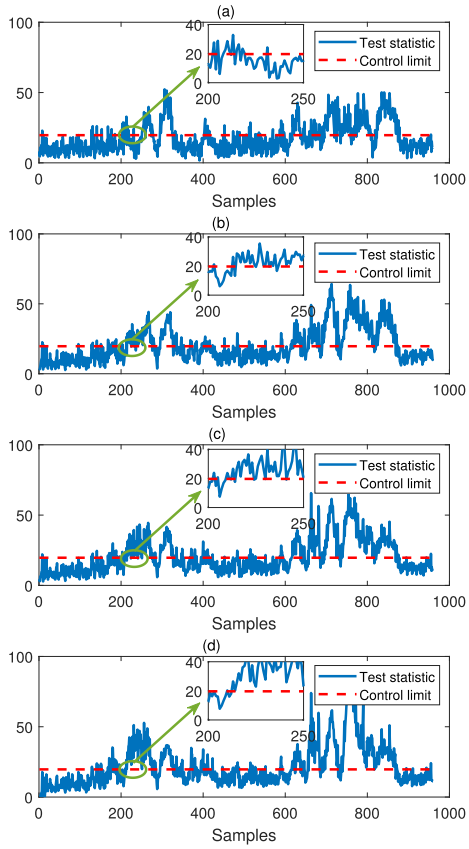


Fig. 4. Detection performance for IDV(10) in the TEP. (a) CCA. (b) SCCA. (c) JSCCA. (d) JSCCCA.

As s_1 increases, the FDR value will become more significant. When $s_1 = 9$, it reaches the largest FDR value and then remains unchanged. This reflects that $s_1 = 9$ obtains the best performance. In this case, after solving the proposed JSCCCA, the canonical matrix $\mathbf{A}$ is sparse, as demonstrated in Fig. 7. Here, Fig. 7(a) shows the first 11 rows, and Fig. 7(b) shows the last 11 rows. It can be seen that there exist only nine nonzero rows since $s_1 = 9$. That is to say, JSCCCA can determine the selected canonical variables as defined in the optimization model.

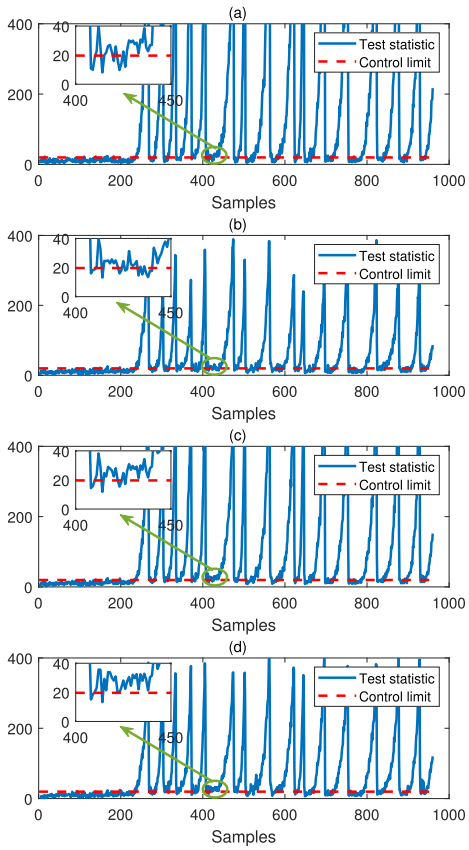


Fig. 5. Detection performance for IDV(20) in the TEP. (a) CCA. (b) SCCA. (c) JSCCA. (d) JSCCCA.

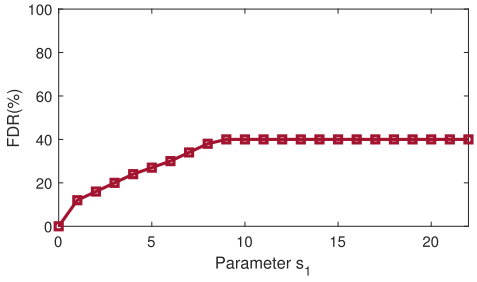


Fig. 6. Demonstration of parameter selecting.

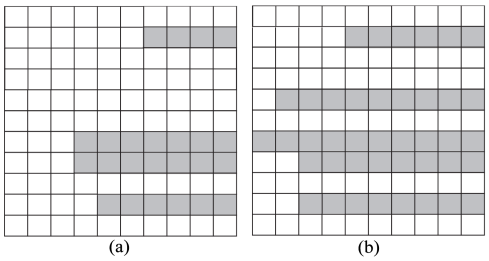


Fig. 7. Canonical matrices by JSCCCA with $s1 = 9$. (a) First 11 rows. (b) Last 11 rows.

5) Discussion: From the TEP benchmark studies, the proposed JSCCCA works better than CCA, SCCA, and JSCCA. In particular, compared with CCA, the FDR value for the fault IDV(10) is increased by 12.62%. Of course, it is admitted that the improvements for some cases are not significant. However, the increases in JSCCCA are still positive. As stated in [51], these faults are relatively tricky for data-driven FD approaches. Therefore, it needs to develop a different strategy to deal with these faults, which are hard to be detected with the existing CCA-based techniques.

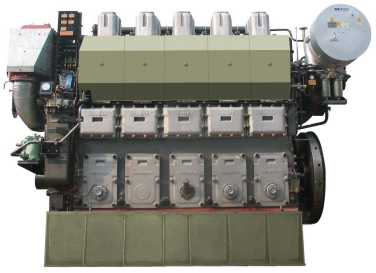


Fig. 8. Diagram of the CPP.

TABLE IV
SELECTED VARIABLES IN THE CPP

Variable No.	Description
01–05	exhaust gas temperature
06–10	cooling oil inlet and outlet temperature difference
11–15	oil inlet pressure
16–20	JCW inlet and outlet temperature difference
21–25	JCW inlet pressure

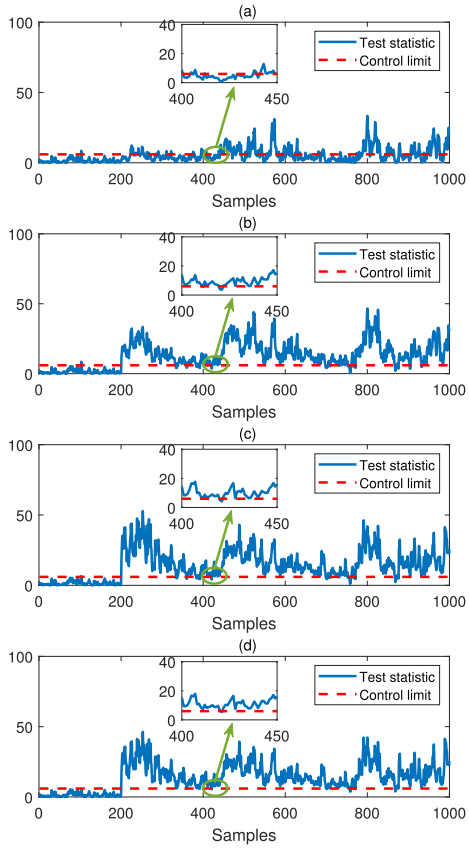


Fig. 9. Detection performance for the CPP. (a) CCA. (b) SCCA. (c) JSCCA. (d) JSCCCA.

C. Cylinder-Piston Application

1) Dataset: The cylinder piston is a vital component of diesel engines, which holds great pressure during normal operation. The CPP dataset is recorded in a practical

TABLE V
DETECTION RESULTS FOR THE CPP

	CCA	SCCA	JSCCA	JSCCCA
FDR	38.25%	93.25%	94.75%	**98.38%**
FAR	0.50%	0.50%	**0.00%**	**0.00%**

two-stroke low-speed marine diesel engine [38]. Fig. 8 shows a schematic of the CPP that contains five cylinders.

In the process, 1000 samples are chosen for off-line training, and a step change of 5 is introduced at the 201st sample for online detection. For each cylinder, five variables are selected, so there exist 25 variables in total; see Table IV for the detailed descriptions. Here, JCW denotes jacket cooling water. In this study, variable Nos. 01–10 are set to $\mathbf{X}$, and variable Nos. 11–25 are set to $\mathbf{Y}$.

2) Detection Results: The detection results are presented in Table V, and the related performance is shown in Fig. 9. It is concluded that all CCA-based approaches can successfully detect this fault in the 201st sample. In comparison, JSCCCA can obtain better detection performance than others because most of the fault samples under fault conditions are detected.

VI. CONCLUSION

In this article, we have proposed a JSCCCA model to improve the detection performance by incorporating the $\ell_{2,0}$-norm joint sparse with classical CCA. In algorithms, we have developed an efficient optimization framework based on the AMA with detailed convergence analysis. In applications, we have validated the efficiency of simulation examples, the benchmark TEP, and a practical CPP. This show that the proposed JSCCCA has more satisfactory detection capability than the existing CCA-based approaches.

In the future, several interesting directions need to be further investigated. First, although the proposed JSCCCA achieves satisfactory performance in numerical cases, more efforts on real-world verification should be studied. Second, the fault isolation and diagnosis can be considered to demonstrate the superiority of $\ell_{2,0}$-norm constrained approaches over existing $\ell_{2,1}$-norm regularized formulations. Third, developing a new detection strategy for non-Gaussian processes is necessary. Finally, the proposed approach is only used for FD and can also be applied to other multiview applications, including face recognition, image denoising, and social networks.

REFERENCES

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

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

[3] X. Deng, X. Tian, S. Chen, and C. J. Harris, "Nonlinear process fault diagnosis based on serial principal component analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 560–572, Mar. 2018.

[4] K. Zhong, M. Han, T. Qiu, and B. Han, "Fault diagnosis of complex processes using sparse kernel local Fisher discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1581–1591, May 2020.

[5] Y. Si, Y. Wang, and D. Zhou, "Key-performance-indicator-related process monitoring based on improved kernel partial least squares," *IEEE Trans. Ind. Electron.*, vol. 68, no. 3, pp. 2626–2636, Mar. 2021.

[6] Z. Ren, W. Zhang, and Z. Zhang, "A deep nonnegative matrix factorization approach via autoencoder for nonlinear fault detection," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5042–5052, Aug. 2020.

[7] Z. Chen, K. Zhang, S. X. Ding, Y. A. W. Shardt, and Z. Hu, "Improved canonical correlation analysis-based fault detection methods for industrial processes," *J. Process Control*, vol. 41, pp. 26–34, May 2016.

[8] Q. Jiang, S. Yan, H. Cheng, and X. Yan, "Local–global modeling and distributed computing framework for nonlinear plant-wide process monitoring with industrial big data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3355–3365, Aug. 2021.

[9] X. Yang, W. Liu, W. Liu, and D. Tao, "A survey on canonical correlation analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2349–2369, Jun. 2021.

[10] C. O. Sakar and O. Kursun, "Discriminative feature extraction by a neural implementation of canonical correlation analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 1, pp. 164–176, Jan. 2017.

[11] J. Chen, G. Wang, and G. B. Giannakis, "Graph multiview canonical correlation analysis," *IEEE Trans. Signal Process.*, vol. 67, no. 11, pp. 2826–2838, Jun. 2019.

[12] S. Mehrkanoon and J. A. K. Suykens, "Regularized semipaired kernel CCA for domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 3199–3213, Jul. 2018.

[13] Y. Yu, S. Tang, K. Aizawa, and A. Aizawa, "Category-based deep CCA for fine-grained venue discovery from multimodal data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1250–1258, Apr. 2018.

[14] P. Wu, S. Lou, X. Zhang, J. He, Y. Liu, and J. Gao, "Data-driven fault diagnosis using deep canonical variate analysis and Fisher discriminant analysis," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3324–3334, May 2021.

[15] H. S. Wong, L. Wang, R. Chan, and T. Zeng, "Deep tensor CCA for multi-view learning," *IEEE Trans. Big Data*, early access, May 11, 2021, doi: 10.1109/TBDATA.2021.3079234.

[16] Z. Chen, K. Liang, S. X. Ding, C. Yang, T. Peng, and X. Yuan, "A comparative study of deep neural network-aided canonical correlation analysis-based process monitoring and fault detection methods," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 22, 2021, doi: 10.1109/TNNLS.2021.3072491.

[17] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, Apr. 2009.

[18] V. Uurtio, S. Bhadra, and J. Rousu, "Large-scale sparse kernel canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6383–6391.

[19] X. Li, X. Xiu, W. Liu, and Z. Miao, "An efficient Newton-based method for sparse generalized canonical correlation analysis," *IEEE Signal Process. Lett.*, vol. 29, pp. 125–129, 2022.

[20] Z. Ni, X. Xiu, and Y. Yang, "Towards efficient state of charge estimation of lithium-ion batteries using canonical correlation analysis," *Energy*, vol. 254, Sep. 2022, Art. no. 124415.

[21] 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.*, vol. 23, 2010, pp. 1–9.

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

[23] 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.

[24] C. Gao, Z. Ma, and H. H. Zhou, "Sparse CCA: Adaptive estimation and computational barriers," *Ann. Statist.*, vol. 45, no. 5, pp. 2074–2101, Oct. 2017.

[25] S. Chen, S. Ma, L. Xue, and H. Zou, "An alternating manifold proximal gradient method for sparse principal component analysis and sparse canonical correlation analysis," *INFORMS J. Optim.*, vol. 2, no. 3, pp. 192–208, Jul. 2020.

[26] M. Xu, Z. Zhu, X. Zhang, Y. Zhao, and X. Li, "Canonical correlation analysis with $\ell_{2,1}$-norm for multiview data representation," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4772–4782, Nov. 2020.

[27] X. Xiu, Y. Yang, L. Kong, and W. Liu, "Data-driven process monitoring using structured joint sparse canonical correlation analysis," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 1, pp. 361–365, Jan. 2021.

[28] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, Nov. 2009.

[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] L. Pan, N. Xiu, and J. Fan, "Optimality conditions for sparse nonlinear programming," *Sci. China Math.*, vol. 60, no. 5, pp. 759–776, May 2017.

[31] R. Wang, N. Xiu, and C. Zhang, "Greedy projected gradient-Newton method for sparse logistic regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 2, pp. 527–538, Feb. 2020.

[32] S. Zhou, Z. Luo, N. Xiu, and G. Y. Li, "Computing one-bit compressive sensing via double-sparsity constrained optimization," *IEEE Trans. Signal Process.*, vol. 70, pp. 1593–1608, 2022.

[33] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[34] M. Hyder and K. Mahata, "Direction-of-arrival estimation using a mixed $\ell_{2,0}$ norm approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4646–4655, Sep. 2010.

[35] 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 2018.

[36] 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.

[37] H. Zhang, Z. Yuan, and N. Xiu, "Recursion Newton-like algorithm for $\ell_{2,0}$-ReLU deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 13, 2021, doi: 10.1109/TNNLS.2021.3131406.

[38] X. Xiu, Z. Miao, Y. Yang, and W. Liu, "Deep canonical correlation analysis using sparsity-constrained optimization for nonlinear process monitoring," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 6690–6699, Oct. 2022, doi: 10.1109/TII.2021.3121770.

[39] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[40] J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan, "A brief introduction to manifold optimization," *J. Operations Res. Soc. China*, vol. 8, no. 2, pp. 199–248, Jun. 2020.

[41] P. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ, USA: Princeton Univ. Press, 2007.

[42] N. Boumal, P.-A. Absil, and C. Cartis, "Global rates of convergence for nonconvex optimization on manifolds," *IMA J. Numer. Anal.*, vol. 39, no. 1, pp. 1–33, Feb. 2018.

[43] S. Boyd, N. Parikh, and E. Chu, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. New York, NY, USA: Now, 2011.

[44] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 265–286, 2004.

[45] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, no. 1, pp. 397–434, Dec. 2013.

[46] S. Chen, S. Ma, A. Man-Cho So, and T. Zhang, "Proximal gradient method for nonsmooth optimization over the stiefel manifold," *SIAM J. Optim.*, vol. 30, no. 1, pp. 210–239, Jan. 2020.

[47] W. H. Yang, L.-H. Zhang, and R. Song, "Optimality conditions for the nonlinear programming problems on Riemannian manifolds," *Optim. Online*, vol. 10, no. 2, pp. 415–434, 2012.

[48] J. Cai, W. Dan, and X. Zhang, "ℓ_0-based sparse canonical correlation analysis with application to cross-language document retrieval," *Neurocomputing*, vol. 329, pp. 32–45, Feb. 2019.

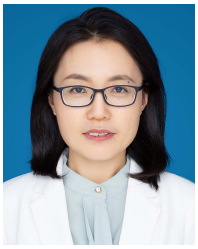
[49] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Comput. Chem. Eng.*, vol. 17, no. 3, pp. 245–255, Mar. 1993.

[50] Z. Chen, S. X. Ding, T. Peng, C. Yang, and W. Gui, "Fault detection for non-Gaussian processes using generalized canonical correlation analysis and randomized algorithms," *IEEE Trans. Ind. Electron.*, vol. 65, no. 2, pp. 1559–1567, Feb. 2018.

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

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 Post-Doctoral Researcher with Peking University, Beijing. 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.



Lili Pan received the Ph.D. degree in operations research from Beijing Jiaotong University, Beijing, China, in 2017.

She is currently an Associate Professor with the School of Mathematics and Statistics, Shandong University of Technology, Zibo, China. Her research interests include sparse optimization theory and algorithm, and sparse image recovery.

Ying Yang (Senior Member, IEEE) received the Ph.D. degree in control theory from Peking University, Beijing, China, in 2002.

From January 2003 to November 2004, she was a Post-Doctoral Researcher with Peking University. From 2005 to 2014, she was an Associate Professor with the Department of Mechanics and Engineering Science, College of Engineering, Peking University. Since 2014, she has been a Full Professor with the Department of Mechanics and Engineering Science, Peking University. Her current research interests include robust and optimal control, nonlinear systems control, numerical analysis, fault detection, and fault-tolerant systems.

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 Sciences, Beijing, China, in 1988, and the Ph.D. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1993.

He once held the ARC Fellowship, the U2000 Fellowship, and the JSPS Fellowship and attracted research funds from different resources for over 2.4 million dollars. 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.