

Structured Joint Sparse Orthogonal Nonnegative Matrix Factorization for Fault Detection

Xi Zhang^{ID}, Xianchao Xiu^{ID}, Member, IEEE, and Chao Zhang^{ID}

Abstract—As modern industrial processes become complicated, and some faults are difficult to be detected due to noises and nonlinearity of data, data-driven fault detection (FD) has been extensively used to detect abnormal events in functional units. To obtain better FD performance of nonnegative matrix factorization (NMF), this article first proposes an FD method using the structured joint sparse orthogonal NMF (SJSONMF). The core idea is to incorporate the graph regularization, sparsity, and orthogonality constraints into the classical NMF, which enjoys stronger discriminative ability, removes redundancy of different basis vectors, and improves fault interpretability. More importantly, an optimization algorithm based on the proximal alternating nonnegative least squares (PANLS) is developed, which can guarantee and speed up the convergence. Finally, the effectiveness of the proposed method is demonstrated by the experiments on the benchmark Tennessee Eastman Process (TEP) and two practical bearing datasets. Particularly, compared with the classical NMF, the T^2 statistic has a gain of 33.13% for the fault IDV(16) on the TEP. The results show that the proposed model and algorithms are promising for FD.

Index Terms—Fault detection (FD), non-Gaussian process, nonnegative matrix factorization (NMF), optimization algorithm, orthogonality constraint.

I. INTRODUCTION

FAULT detection (FD) is an essential step for industrial processes because it can ensure safe production and thus avoid economic losses to a certain extent [1]. Compared with the model-based FD, data-driven FD only relies on the measured variables, thereby making it more applicable for plant-wide processes [2]. As a data-driven method, multivariate analysis (MVA) has been widely used in FD, including principal component analysis (PCA) [3], [4], independent component analysis (ICA) [5], [6], partial least squares (PLS) [7], [8], Fisher discriminant analysis (FDA) [9], [10], canonical variate analysis (CVA) [11], [12], canonical correlation analysis (CCA) [13], [14], and nonnegative matrix factorization (NMF) [15], [16], where the structured joint sparse strategy appears in the applications of PCA [3], [4],

Manuscript received 17 October 2022; revised 2 January 2023; accepted 16 January 2023. Date of publication 3 February 2023; date of current version 15 February 2023. This work was supported in part by the National Science Foundation of Beijing, China, under Grant 1202021 and in part by the National Natural Science Foundation of China under Grant 12171027, Grant 52102472, and Grant 12001019. The Associate Editor coordinating the review process was Dr. Bruno Albuquerque De Castro. (*Corresponding author: Chao Zhang*)

Xi Zhang and Chao Zhang are with the School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China (e-mail: chzhang2@bjtu.edu.cn).

Xianchao Xiu is with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200072, China.

Digital Object Identifier 10.1109/TIM.2023.3241990

CCA [13], [14], and NMF [15] to FD. It should be pointed out that the PCA requires the process to obey the Gaussian distribution, while the ICA needs the process to follow the non-Gaussian distribution. By comparison, NMF has no other requirements except for nonnegativity, and thus, it can efficiently deal with the non-Gaussian and Gaussian processes. Now, NMF has attracted extensive attention in industry and academia.

NMF is a powerful dimensionality reduction technique designed to seek a low-dimensional subspace representation for original data information [17]. By combining regularization terms or constraints with NMF, it can improve the performance of matrix decomposition, better identify local features, and provide a more sparse representation.

For example, graph regularized NMF (GNMF) enjoys stronger discriminative ability than NMF, since the added graph Laplacian explicitly considers the local invariance and the internal geometric features of process data [18], [19], [20]. Sparse NMF (SNMF) helps in saving a great deal of storage space and enhancing the extraction of local features to better represent the latent features of the original data [21], [22], [23]. A soft orthogonal-regularized NMF (ONMF) for sparse modeling of spectral imaging data helps avoid chemical components' overlap [24]. These variants have been well-verified to be promising for machine learning and pattern recognition.

In addition, a sparse orthogonality-regularized joint NMF (SOJNMF) method was proposed to integratively analyze multidimensional omics data by introducing the sparsity and orthogonality regularization terms of the coefficient matrix. This method can not only identify multidimensional molecular regulatory modules but also reduce the overlap rate of features among the multidimensional modules while ensuring the sparsity of the coefficient matrix after decomposition [25].

However, the application to FD has not been extensively investigated yet. Li et al. [26] first extended NMF to the field of FD for the non-Gaussian processes. Numerical experiments on the benchmark Tennessee Eastman Process (TEP) demonstrated that compared with PCA and ICA, this NMF-based method obtains better FD performance and has wider applicability. After that, several NMF-based FD methods have been developed. Li et al. [27] constructed a generalized nonnegative matrix projection (GNMP) model by incorporating positive constrained projections with NMF, which can remove non-negative restrictions from the original data. Zhai et al. [28] presented a kernel variant of NMF, called KNMF, to project the low-dimensional original data into a higher space to deal with the nonlinear processes. Wang et al. [29] fully exploited

the local information within blocks and global information between blocks, and then suggested an adaptive partition NMF (APNMF) to show the performance improvement of the NMF-based FD. Ren et al. [16] considered a deep NMF (DNMF) framework by introducing deep autoencoders, which automatically realizes nonlinear mapping for input process data. Very recently, Xiu et al. [15] presented a structured joint SNMF (SJSNMF) model based on the graph Laplacian and joint sparsity, which not only retains the internal geometric structure but also determines the rowwise sparsity of the latent variables.

The successful applications of the NMF-based FD methods have been reported, and the following two modifications can further improve the NMF-based FD modeling performance. On one hand, it is well-known that the orthogonality plays a crucial role in PCA because it not only can remove redundant information of data but also can yield a more localized part-based representation. On the other hand, the algorithm for solving the NMF-based model with the nonconvex feasible set, SJSNMF, has not discussed the convergence properties. In numerical optimization, whether the algorithm converges will affect the stability of numerical results, and thus, a good algorithm should have a convergence guarantee. Therefore, it is a need to construct an NMF-based FD model with the nonconvex feasible set which has orthogonality constraint and convergence.

For FD, some faults are difficult to be detected due to noises and nonlinearity of data, which lead to low accuracy. Introducing orthogonality constraint is promising to remove redundant information. Inspired by the above analysis, we propose an effective data-driven NMF-based FD model, called the structured joint sparse orthogonal NMF (SJSONMF). The main contributions of this work are highlighted as follows.

- 1) It is the first work to integrate orthogonality constraint for the basis matrix and row sparsity constraint for the coefficient matrix into an NMF-based FD model.
- 2) It proposes a data-driven FD model that reduces the correlation between the basis vectors by the orthogonality constraint on the basis matrix and discards unimportant information in the rows of coefficient matrix by the row sparsity constraint on the coefficient matrix.
- 3) It provides a proximal alternating nonnegative least squares (PANLS) algorithm to solve the proposed SJSONMF model with the nonconvex feasible set, which guarantees and speeds up the convergence. Two effective algorithms for solving subproblems in the PANLS algorithm are also given.
- 4) It conducts the numerical experiments on the benchmark TEP dataset for complex chemical production process and two practical bearing datasets. There are actually many differences between the two types of faults. But the proposed model and algorithms all show good performances, which demonstrates the effectiveness and wide applicability of the proposed model and the algorithms.

Notation: All the matrices are denoted in bold uppercase letters, and all the vectors are denoted in bold lowercase letters.

$\mathbb{R}^{m \times n}$ denotes the set of all $m \times n$ real matrices, $\mathbb{R}_+^{m \times n}$ is the subset of $\mathbb{R}^{m \times n}$, which contains all $m \times n$ real matrices whose entries are all nonnegative. For a matrix \mathbf{Z} , the (i, j) entry, the i th row, and the j th column are denoted by \mathbf{Z}_{ij} , $\mathbf{Z}_{i\cdot}$ and $\mathbf{Z}_{\cdot j}$, respectively. Moreover, $\|\mathbf{Z}\|_{2,0}$ stands for the number of nonzero rows of \mathbf{Z} , $\text{tr}(\mathbf{Z})$ is the trace of \mathbf{Z} , and $\text{Diag}(\mathbf{Z})$ is a diagonal matrix with the main diagonal being the diagonal of \mathbf{Z} . In addition, for a set Ω , $\mathcal{P}_\Omega(\mathbf{Z})$ denotes the projection of \mathbf{Z} onto Ω .

II. PRELIMINARY AND MODEL STRUCTURE

A. NMF Basics

Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be a nonnegative process data matrix with m process variables and n samples. To seek nonnegative matrices \mathbf{W} and \mathbf{H} to approximate \mathbf{X} , i.e., $\mathbf{X} \approx \mathbf{WH}$, NMF can be described as minimizing $(1/2)\|\mathbf{X} - \mathbf{WH}\|_F^2$, where $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times n}$ are regarded as the basis matrix and coefficient matrix, respectively, and the selection of parameter r satisfies $r \ll \min\{m, n\}$ [30].

Based on the above methods, SJSNMF was proposed by incorporating the graph regularization and the joint sparse constraint to discard useless information and preserve the geometric structure [15]. Compared with the classical NMF, SJSNMF is able to capture the relationship between process variables by incorporating the information of a data into the Laplacian matrix and learn global structures of process data by adding the joint sparsity constraint.

B. Model Structure

One of the main shortcomings of the existing NMF-based FD models is that the redundancy between different bases has not been considered. Introducing orthogonality not only can yield potentially useful data representations as a superposition of disjoint parts but also can obtain a better clustering performance [31]. In this regard, the proposed model can be constructed as follows:

$$\begin{aligned} & \min_{\mathbf{W}, \mathbf{H}} \frac{1}{2} \|\mathbf{X} - \mathbf{WH}\|_F^2 + \lambda \text{tr}(\mathbf{HLH}^T) \\ & \text{s.t. } \mathbf{W} \geq \mathbf{0}, \quad \mathbf{H} \geq \mathbf{0}, \quad \mathbf{W}^T \mathbf{W} = \mathbf{I}, \quad \|\mathbf{H}\|_{2,0} \leq s \end{aligned} \quad (1)$$

where λ is the regularization parameter, \mathbf{L} denotes the graph Laplacian matrix, \mathbf{I} denotes the identity matrix, $\mathbf{W}^T \mathbf{W} = \mathbf{I}$ is the orthogonality constraint, and s is the sparsity that controls the number of nonzero rows. Note that the number of fault variables can be adjusted by tuning s . Without $\mathbf{W}^T \mathbf{W} = \mathbf{I}$ in (1), SJSONMF reduces to SJSNMF. If further deleting $\|\mathbf{H}\|_{2,0} \leq s$ in (1), SJSNMF reduces to GNMF.

As mentioned in the Introduction of [32], the combination of nonnegativity constraint $\mathbf{W} \geq \mathbf{0}$ and orthogonality constraint $\mathbf{W}^T \mathbf{W} = \mathbf{I}$ is equivalent to that each row of \mathbf{W} has at most one positive element and each column of \mathbf{W} takes the unit norm. This nice property theoretically guarantees the sparsity of \mathbf{W} and the nonoverlapping among basis vectors, i.e., column vectors of \mathbf{W} .

Although Xiu et al. [15] apply a similar NMF-based model, the main differences are given as follows.

- 1) It only considers the importance of incorporating graph Laplacian and joint sparsity in process control. As mentioned above, the newly added orthogonality constraint together with the original nonnegativity constraint in NMF can produce an orthogonal basis matrix with nice property, and consequently provide potentially useful data representations as a superposition of disjoint parts. Thus, the proposed SJSONMF model improves the detection performance, which will be verified in Sections V and VI.
- 2) It only develops an optimization algorithm based on the alternating minimization but lacks convergence analysis. On the contrary, a convergent PANLS algorithm will be established in Section III.

III. OPTIMIZATION ALGORITHM

Since problem (1) involves two nonconvex constraints, i.e., $\mathbf{W}^T \mathbf{W} = \mathbf{I}$, $\|\mathbf{H}\|_{2,0} \leq s$, and the objective function is not separable in terms of variables \mathbf{W} , \mathbf{H} , there exist no fast solvers to handle it directly. In what follows, an algorithm based on the PANLS framework motivated by Zhang et al. [33] is developed along with the convergence analysis.

For simplicity, denote the objective function to be

$$f(\mathbf{W}, \mathbf{H}) = \frac{1}{2} \|\mathbf{X} - \mathbf{WH}\|_F^2 + \lambda \text{tr}(\mathbf{HLH}^T) \quad (2)$$

and denote the feasible sets of \mathbf{W} and \mathbf{H} to be

$$\mathbb{O} := \{\mathbf{W} \in \mathbb{R}^{m \times r} : \mathbf{W} \geq 0, \mathbf{W}^T \mathbf{W} = \mathbf{I}\} \quad (3)$$

and $\mathbb{S}_+ := \mathbb{S} \cap \mathbb{R}_+^{r \times n}$ with

$$\mathbb{S} = \{\mathbf{H} \in \mathbb{R}^{r \times n} : \|\mathbf{H}\|_{2,0} \leq s\} \quad (4)$$

respectively. The PANLS algorithm for solving (1) is presented in Algorithm 1, where τ_1 and τ_2 are the proximal parameters corresponding to the proximal terms. It is necessary to mention that adding these proximal terms can guarantee and speed up the convergence.

Algorithm 1 PANLS Algorithm for Solving (1)

- 1: Given an initial point $(\mathbf{W}^0, \mathbf{H}^0)$, proximal parameters $\tau_1, \tau_2 \geq 0$.
- 2: **for** $k = 0, 1, \dots, k_{\max}$ **do**
- 3: Update \mathbf{W}^k by \mathbf{W}^{k+1} , which is an optimal solution of

$$\begin{aligned} \min_{\mathbf{W}} l(\mathbf{W}) &:= f(\mathbf{W}, \mathbf{H}^k) + \frac{\tau_1}{2} \|\mathbf{W} - \mathbf{W}^k\|_F^2 \\ \text{s.t. } \mathbf{W} &\geq \mathbf{0}, \quad \mathbf{W}^T \mathbf{W} = \mathbf{I}. \end{aligned} \quad (5)$$

- 4: Update \mathbf{H}^k by \mathbf{H}^{k+1} , which is an optimal solution of

$$\begin{aligned} \min_{\mathbf{H}} q(\mathbf{H}) &:= f(\mathbf{W}^{k+1}, \mathbf{H}) + \frac{\tau_2}{2} \|\mathbf{H} - \mathbf{H}^k\|_F^2 \\ \text{s.t. } \mathbf{H} &\geq \mathbf{0}, \quad \|\mathbf{H}\|_{2,0} \leq s. \end{aligned} \quad (6)$$

- 5: **end for**

- 6: Return $(\mathbf{W}^{k+1}, \mathbf{H}^{k+1})$.
-

A. Update \mathbf{W}

A large number of manifold learning methods have been proposed in recent years; see [34] for a review. However, there exist relatively few methods for the problems with both nonnegativity and orthogonality constraints, even if the objective function is continuously differentiable. Among them, the most impressive one is called the practical exact penalty (PEP) method, which is proposed by Jiang et al. [32]. Based on an equivalent optimization formulation, it uses the exact penalty approach by keeping partial constraints such that the projection is easy to compute.

From [32, eq. (1.8)], an equivalent optimization formulation of the optimization problem (5)

$$\min_{\mathbf{W} \in \mathcal{OB}_+^{m,r}} l(\mathbf{W}) \quad \text{s.t. } \|\mathbf{Wv}\|_2 = 1 \quad (7)$$

where the function $l(\cdot)$ is the objective function in problem (5) of Algorithm 1, $\mathcal{OB}_+^{m,r} = \mathcal{OB}^{m,r} \cap \mathbb{R}_+^{m \times r}$ with

$$\mathcal{OB}^{m,r} = \{\mathbf{W} \in \mathbb{R}^{m \times r} : \|\mathbf{W}_{\cdot j}\|_2 = 1, j = 1, \dots, r\} \quad (8)$$

and \mathbf{v} can be simply taken as $\mathbf{v} = \mathbf{e}/\sqrt{r}$, with $\mathbf{e} \in \mathbb{R}^r$ being the real vector whose elements are all 1. According to [32, eq. (1.9)], via keeping the constraint $\mathcal{OB}_+^{m,r}$ and penalizing the constraint $\|\mathbf{Wv}\|_2 = 1$, the exact penalty model of (7) is given as

$$\min_{\mathbf{W} \in \mathcal{OB}_+^{m,r}} l(\mathbf{W}) + \sigma (\|\mathbf{Wv}\|_2^2 - 1). \quad (9)$$

Futhermore, [32, Algorithm 2], subproblem (7) can be solved by Algorithm 2. Here, σ is the penalty parameter, γ is used to enlarge σ at each iteration, and ϵ_1 and ϵ_2 are the tolerance parameters. In addition

$$\begin{aligned} P_\sigma(\mathbf{W}) &= l(\mathbf{W}) + \sigma (\|\mathbf{Wv}\|_2^2 - 1) \\ \text{grad } P_\sigma(\mathbf{W}) &= \nabla P_\sigma(\mathbf{W}) - \mathbf{WDiag}(\mathbf{W}^T \nabla P_\sigma(\mathbf{W})) \end{aligned} \quad (10)$$

where $P_\sigma(\mathbf{W})$ is the exact penalty function, and $\text{grad } P_\sigma(\mathbf{W})$ is given in [32, eq. (2.5)], which is the Riemannian gradient of $P_\sigma(\cdot)$ at the point \mathbf{W} on the manifold $\mathcal{OB}^{m,r}$. Details for the convergence results can be found in [32].

Algorithm 2 PEP Method for Subproblem (5)

- 1: Given an initial point $\mathbf{W}^0 = \mathbf{W}^k$, parameters $\gamma > 1$, $\sigma, \epsilon_1, \epsilon_2 > 0$.
- 2: **for** $t = 0, 1, \dots, t_{w\max}$ **do**
- 3: Find \mathbf{W}^{t+1} by the projection gradient method such that

$$\|\min(\mathbf{W}^{t+1}, \text{grad } P_\sigma(\mathbf{W}^{t+1}))\|_F \leq \epsilon_1 \quad (11)$$

$$P_\sigma(\mathbf{W}^{t+1}) \leq P_\sigma(\mathbf{W}^t). \quad (12)$$

- 4: **if** $\|\mathbf{W}^{t+1}\|_2^2 - 1 \leq \epsilon_2$ **then**
 - 5: break
 - 6: **end if**
 - 7: Set $\sigma = \gamma \sigma$.
 - 8: **end for**
 - 9: $\mathbf{W}^{k+1} = \mathbf{W}^{t+1}$.
 - 10: Return \mathbf{W}^{k+1} .
-

B. Update \mathbf{H}

For the vector-based nonnegative optimization problem with the ℓ_0 -norm sparsity constraint and nonnegativity constraint, the improved iterative hard thresholding (IIHT) algorithm can be used [35]. To solve (6), the IIHT algorithm will be generalized to the matrix-based case with the $\ell_{2,0}$ -norm constraint and nonnegativity constraint outlined in Algorithm 3.

As shown in Appendix A-A, the objective function $q(\mathbf{H})$ in (6) of Algorithm 1 is strongly convex and also strongly smooth (see [35, Definition 2.2]). Thus, the IIHT algorithm can be generalized to the matrix-based case, which is outlined here in Algorithm 3. In Step 4, $\text{supp}(\mathbf{H}^{t+1}) := \{(i, j) : \mathbf{H}_{ij}^{t+1} \neq 0\}$, and $\nabla_{\text{supp}(\mathbf{H}^{t+1})} q(\mathbf{H}^{t+1}) \in \mathbb{R}^{r \times n}$ whose (i, j) entry equals the (i, j) entry of $\nabla q(\mathbf{H}^{t+1})$ if $(i, j) \in \text{supp}(\mathbf{H}^{t+1})$, and equals 0 otherwise. $\mathcal{P}_{\mathbb{S}_+}(\mathbf{Z}) = \mathcal{P}_{\mathbb{S}}(\mathcal{P}_{\mathbb{R}^{r \times n}}(\mathbf{Z}))$ can be obtained easily; see Appendix A-B for detail.

Algorithm 3 Matrix-Based IIHT Method for Subproblem (6)

- 1: Given an initial point $\mathbf{H}^0 = \mathbf{H}^k$, $\varepsilon > 0$, $\beta \in (0, 1)$, and $\rho_0 \in (0, 1/(1/2 + \lambda \|\mathbf{L}\|_F))$.
- 2: **for** $t = 0, 1, \dots, t_{\max}$ **do**
- 3: Compute $\mathbf{H}^{t+1} \in \mathcal{P}_{\mathbb{S}_+}(\mathbf{H}^t - \rho_t \nabla q(\mathbf{H}^t))$, where $\rho_t = \rho_0 \beta^{c_t}$ and c_t is the smallest nonnegative integer c such that
- 4:
$$q(\mathbf{H}^t(\rho_0 \beta^c)) \leq q(\mathbf{H}^t) - \frac{\rho_0 \beta^c}{2} \|\mathbf{H}^t(\rho_0 \beta^c) - \mathbf{H}^t\|_F^2 \quad (13)$$
- 5: and $\mathbf{H}^t(\rho) \in \mathcal{P}_{\mathbb{S}_+}(\mathbf{H}^t - \rho \nabla q(\mathbf{H}^t))$.
- 6: **if** $\|\nabla_{\text{supp}(\mathbf{H}^{t+1})} q(\mathbf{H}^{t+1})\|_F \leq \varepsilon$ **then**
- 7: break.
- 8: **end if**
- 9: **end for**
- 10: $\mathbf{H}^{k+1} = \mathbf{H}^{t+1}$.
- 11: Return \mathbf{H}^{k+1} .

C. Convergence Analysis

We provide rigorous convergence result of Algorithm 1 in this section.

By direct computation, the partial gradients of f with respect to \mathbf{W} and \mathbf{H} are

$$\begin{aligned} \nabla_{\mathbf{W}} f(\mathbf{W}, \mathbf{H}) &= -(\mathbf{X} - \mathbf{WH})\mathbf{H}^T \\ \nabla_{\mathbf{H}} f(\mathbf{W}, \mathbf{H}) &= -\mathbf{W}^T(\mathbf{X} - \mathbf{WH}) + 2\lambda\mathbf{HL}. \end{aligned} \quad (14)$$

Let us denote by $\mathcal{N}(\mathbf{W}, \mathbb{O})$ and $\mathcal{N}(\mathbf{H}, \mathbb{S}_+)$ the normal cones to \mathbb{O} at \mathbf{W} and to \mathbb{S}_+ at \mathbf{H} , respectively (see [36, Definition 6.3, p. 199]).

Definition 1: For problem (1), a feasible point (\mathbf{W}, \mathbf{H}) is called a stationary point if

$$\begin{aligned} -\nabla_{\mathbf{W}} f(\mathbf{W}, \mathbf{H}) &\in \mathcal{N}(\mathbf{W}, \mathbb{O}) \\ -\nabla_{\mathbf{H}} f(\mathbf{W}, \mathbf{H}) &\in \mathcal{N}(\mathbf{H}, \mathbb{S}_+). \end{aligned} \quad (15)$$

The convergence results of Algorithm 1 are shown in the following theorem, and its proof can be found in Appendix B.

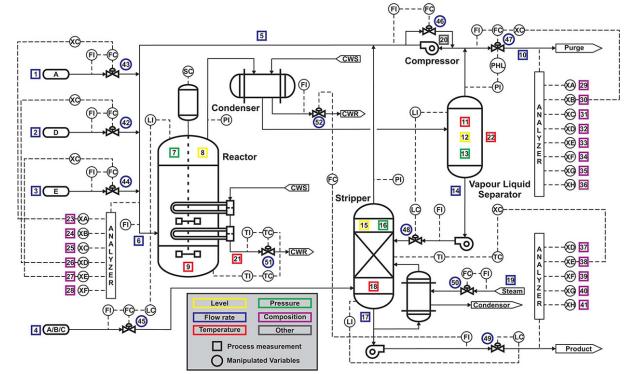


Fig. 1. Process layout of the TEP.

Theorem 1: Let $\{(\mathbf{W}^k, \mathbf{H}^k)\}$ be the sequence generated by Algorithm 1. Then the following statements hold.

- 1) $\{f(\mathbf{W}^k, \mathbf{H}^k)\}$ is nonincreasing. If $(\mathbf{W}^{k+1}, \mathbf{H}^{k+1}) \neq (\mathbf{W}^k, \mathbf{H}^k)$, then $f(\mathbf{W}^{k+1}, \mathbf{H}^{k+1}) < f(\mathbf{W}^k, \mathbf{H}^k)$.
- 2) There exists at least one accumulation point of $\{(\mathbf{W}^k, \mathbf{H}^k)\}$.
- 3) Any accumulation point $(\mathbf{W}^*, \mathbf{H}^*)$ of $\{(\mathbf{W}^k, \mathbf{H}^k)\}$ is a stationary point of (1).

The proximal terms in (5) and (6) of the PANLS algorithm in Algorithm 1 play important roles to get that $(\mathbf{W}^*, \mathbf{H}^*)$ is a stationary point of (1) in the convergence analysis. The reason that the traditional ANLS algorithm [37], [38], [39], [40] is not applicable to (1) with rigorous convergence result due to the nonconvexity of the feasible set; see explanation in detail in Remark 1 after the proof of Theorem 1 in Appendix B.

Although the technique of proximal terms has been used in [33], the model in [33] is much easier to handle with, because it does not consider the orthogonality constraint and the sparsity constraint addressed in this article. Moreover, compared with the convergence result given in [33, Proposition 2.1], the convergence result 2) is novel, which guarantees the existence of an accumulation point.

IV. DETECTION STRATEGY FOR SJSONMF

As discussed in Section III, problem (1) can be computed by Algorithm 1, where the subproblems are solved by Algorithms 2 and 3, respectively. Given input data \mathbf{X} , when problem (1) has been solved, we can get the basis matrix \mathbf{W} and the coefficient matrix \mathbf{H} . For the coefficient matrix \mathbf{H} , it can be regarded as the low-rank approximation of the original data \mathbf{X} and reflects the status of industrial processes corresponding to \mathbf{X} .

The computed basis matrix \mathbf{W} corresponding to given data \mathbf{X} will be further used when new data sample comes. Given a new sample $\mathbf{X}_{\text{new}} \in \mathbb{R}^{m \times n}$, according to [15], the reconstruction $\widehat{\mathbf{H}}$ of \mathbf{H} is expressed as

$$\widehat{\mathbf{H}} = (\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T \mathbf{X}_{\text{new}} = \mathbf{W}^T \mathbf{X}_{\text{new}}. \quad (16)$$

Then, the monitoring model based on NMF is formulated as

$$\widehat{\mathbf{X}} = \mathbf{W}\widehat{\mathbf{H}}. \quad (17)$$

In addition, two monitoring metrics, including T^2 statistic and squared prediction error (SPE) statistic, are used to monitor

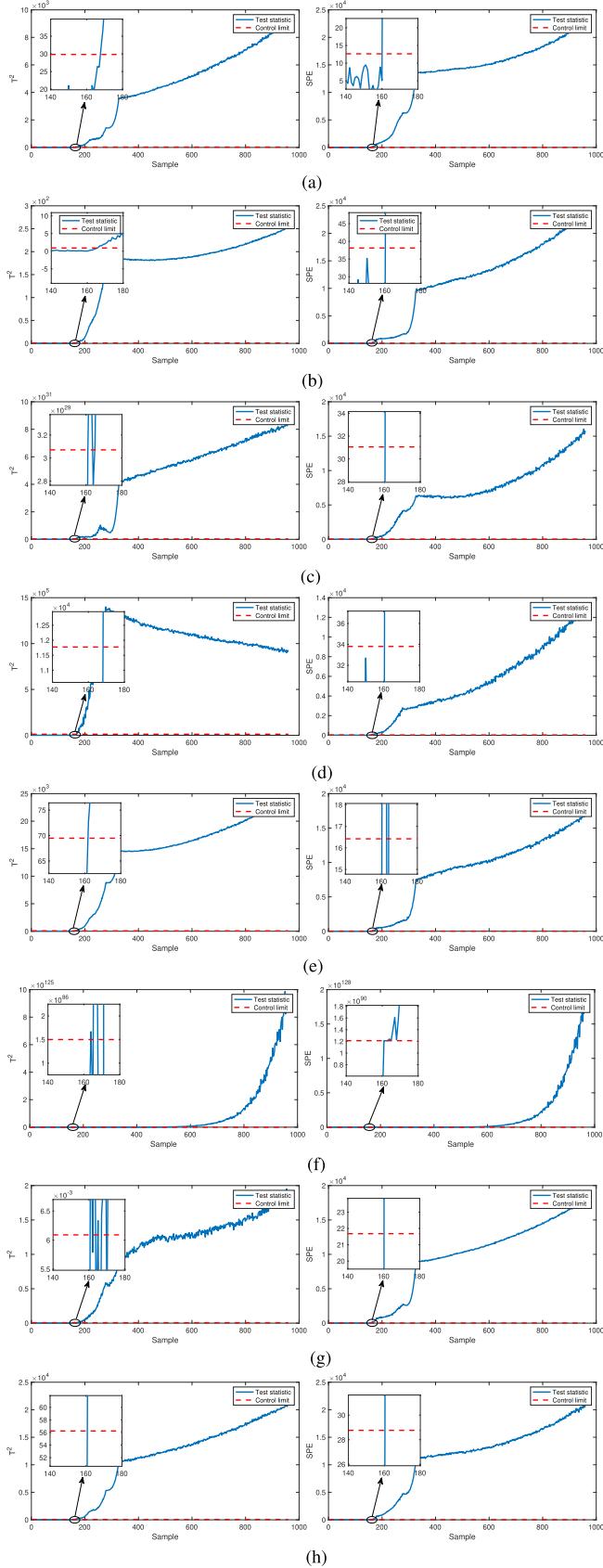


Fig. 2. Detection performance of fault IDV(6) for the TEP. (a) PCA. (b) NMF. (c) GNMF. (d) SNMF. (e) ONMF. (f) KNMF. (g) SJSNMF. (h) SJSONMF.

of mean integrated square error, that is,

$$h_{\text{opt}} = 1.06\sigma n^{-\frac{1}{5}} \quad (20)$$

TABLE III
ORTHOGONALITY AND SPARSITY OF THE BASIS MATRIX

Model	PCA	NMF	GNMF	SNMF	ONMF	KNMF	SJSNMF	SJSONMF
Ortho	-2e-16	0.36	0.3028	0.85	0.0459	1.51	3.41	0
Spar	0	0.8021	0	0.8735	0.9323	/	0.5687	0.9412

where σ is the standard deviation of x [41]. For the given confidence limit α , the control limit J_{th} is calculated as

$$P(J < J_{\text{th}}) = \int_{-\infty}^{\text{th}} p(J)dJ = \alpha. \quad (21)$$

For the corresponding control limits J_{th,T^2} and $J_{\text{th},\text{SPE}}$, the detection logic is given by

$$\begin{cases} T^2 < J_{\text{th},T^2} \text{ and } \text{SPE} < J_{\text{th},\text{SPE}} \Rightarrow \text{fault-free} \\ T^2 \geq J_{\text{th},T^2} \text{ or } \text{SPE} \geq J_{\text{th},\text{SPE}} \Rightarrow \text{faulty}. \end{cases} \quad (22)$$

Overall, the proposed detection approach is summarized as below, which contains offline training and online detecting.

Algorithm 4 SJSONMF-Based Detection Approach

Offline training

- 1: Collect training data \mathbf{X} .
- 2: Compute \mathbf{W} and \mathbf{H} by Algorithm 1 for solving (1) with training data \mathbf{X} .
- 3: Determine the training statistics J_{th,T^2} and $J_{\text{th},\text{SPE}}$ using (21).

Online detecting

- 4: Collect online new samples \mathbf{X}_{new} as the test data.
- 5: Compute $\widehat{\mathbf{H}}$ and $\widehat{\mathbf{X}}$ according to (16) and (17) using \mathbf{W} obtained in offline training.
- 6: Calculate the test statistics T^2 and SPE using (18).
- 7: Detect faults according to the decision logic using (22).

Since training can be finished in offline training, online detecting really becomes very fast. Therefore, the given detection approach guarantees the efficiency of detection. It is worth mentioning that only the basis matrix \mathbf{W} obtained in the offline phase is used repeatedly in the online detecting phase. Hence, it is appropriate to impose the orthogonality constraint on \mathbf{W} for removing redundancy of different basis vectors, not on the coefficient matrix \mathbf{H} as the SOJNMF model done in [25]. Moreover, the convergence of Algorithm 1 guarantees a more reliable output of computed \mathbf{W} in offline training, than the algorithm for solving the SJSNMF model with nonconvex feasible set that without the convergence result.

V. APPLICATION ON THE BENCHMARK TEP

This section demonstrates the superiority of SJSONMF over NMF [26], GNMF [18], SNMF [21], ONMF [24], KNMF [28], and SJSNMF [15] on the benchmark TEP. Meanwhile, PCA is also reported to further illustrate the effectiveness of the NMF-based methods. Note that some nonlinear NMF models such as DNMF are not considered due to their long computational time required [16]. All the experiments of this article are performed in MacBook Pro

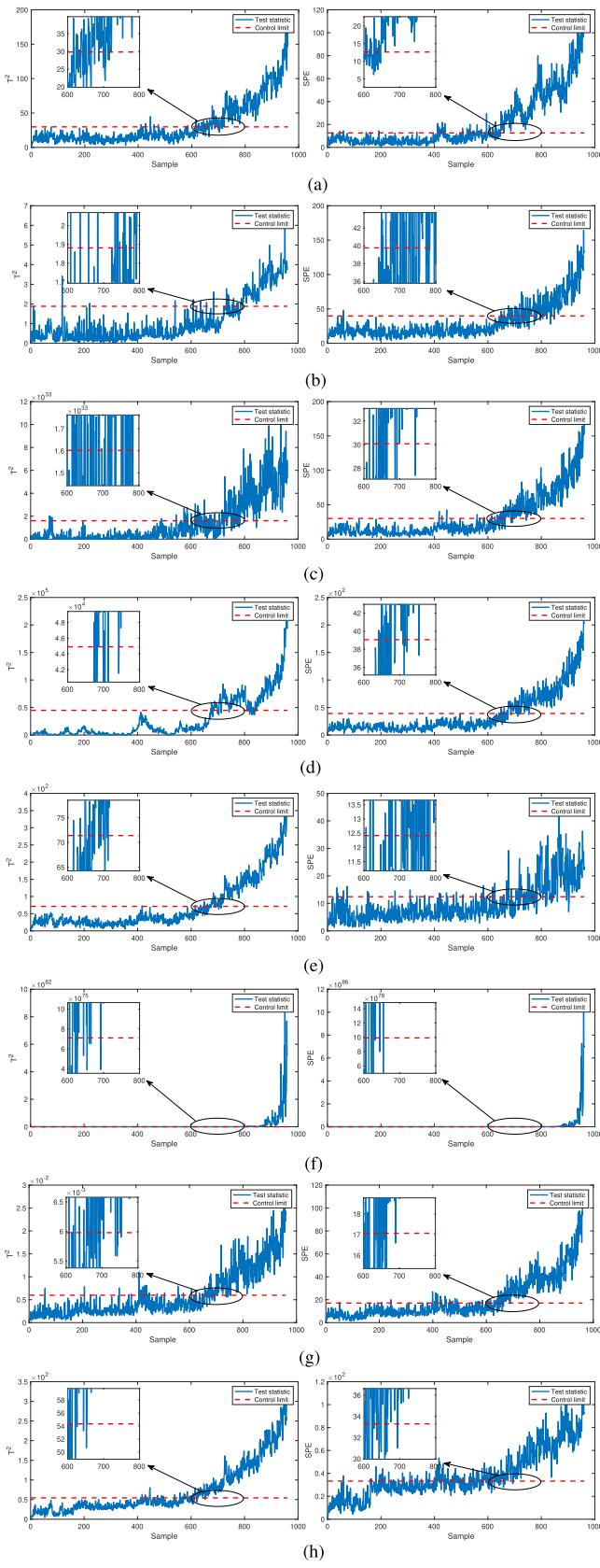


Fig. 3. Detection performance of fault IDV(21) for the TEP. (a) PCA. (b) NMF. (c) GNMF. (d) SNMF. (e) ONMF. (f) KNMF. (g) SJSNMF. (h) SJSNMF.

10.15.3 on an Intel Core 4 CPU at 1.4 GHz with 8 GB of RAM, using MATLAB 2020a.

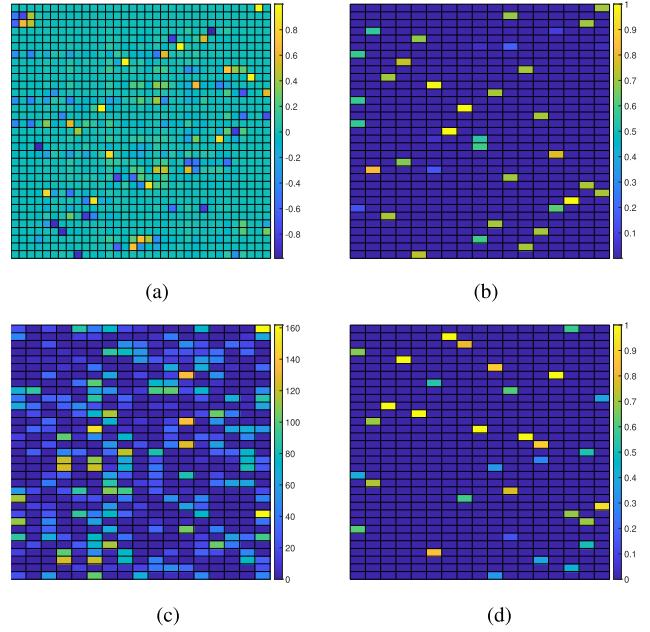


Fig. 4. Obtained orthogonal matrices by different methods. (a) PCA. (b) NMF. (c) SJSNMF. (d) SJSNMF.

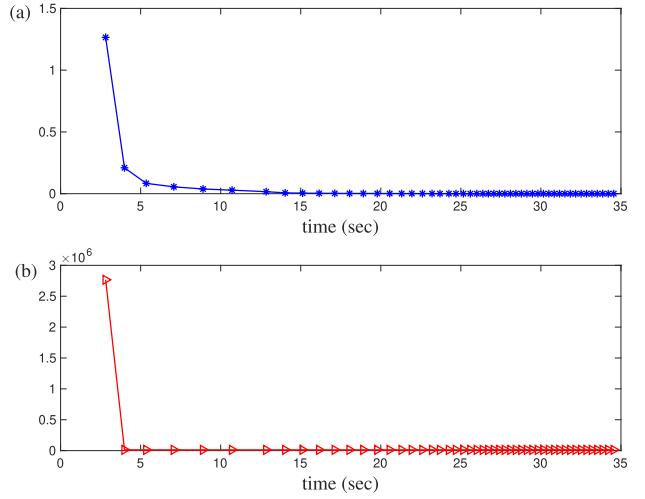


Fig. 5. Illustrations of convergence. (a) Relative error. (b) Objective error.

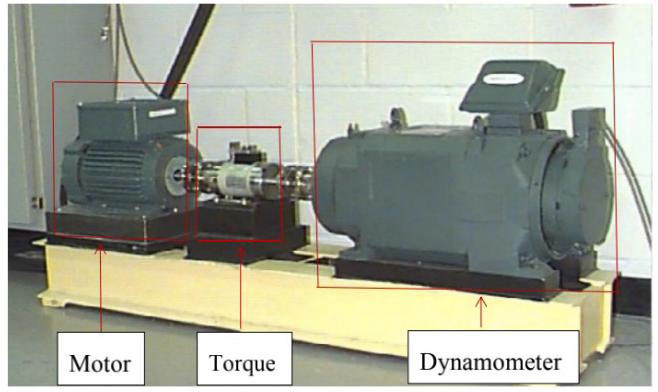


Fig. 6. Experimental platform of the CWRU bearing dataset.

A. Data Preparation

The TEP is a complex chemical production process, which has been extensively applied to test different FD methods; see Fig. 1 for the process layout and [42] for detailed information.

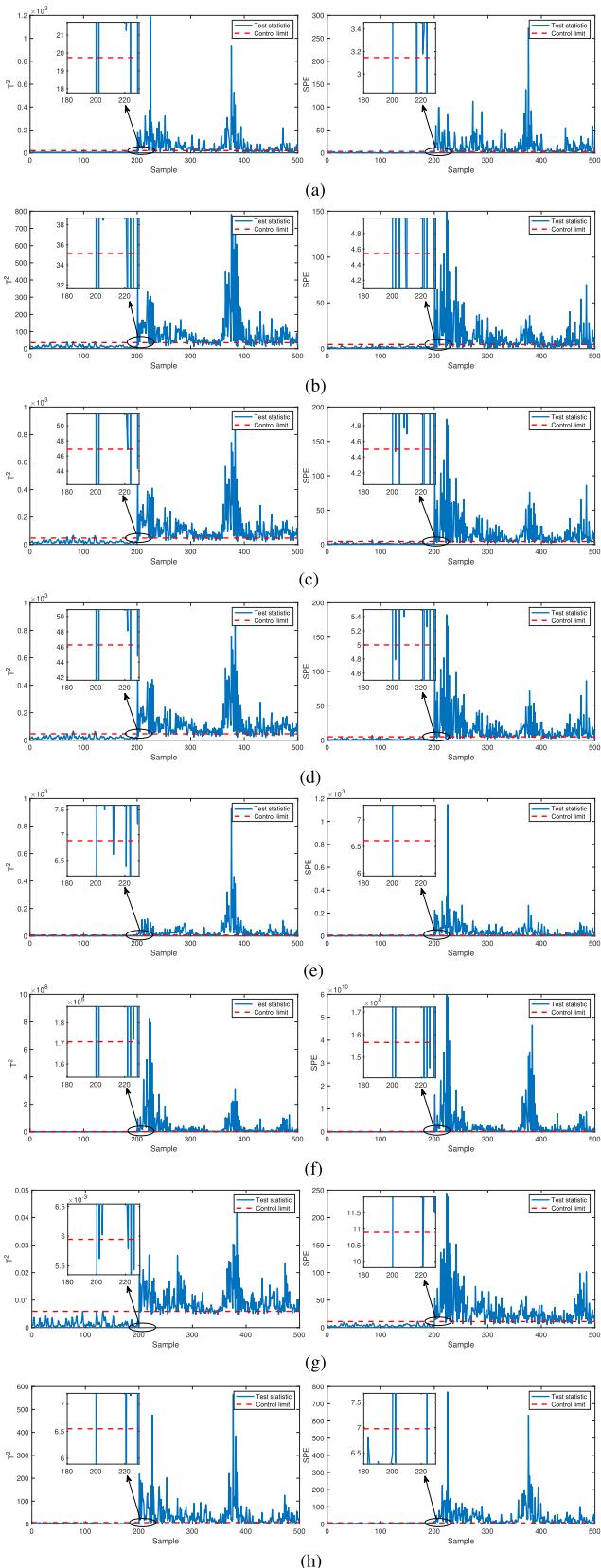


Fig. 7. Detection performance of fault IDV(3) for CWRU bearing. (a) PCA. (b) NMF. (c) GNMF. (d) SNMF. (e) ONMF. (f) KNMF. (g) SJSNMF. (h) SJSONMF.

In this simulation, the normal operating condition is used for offline training, and 21 faulty conditions are used for online

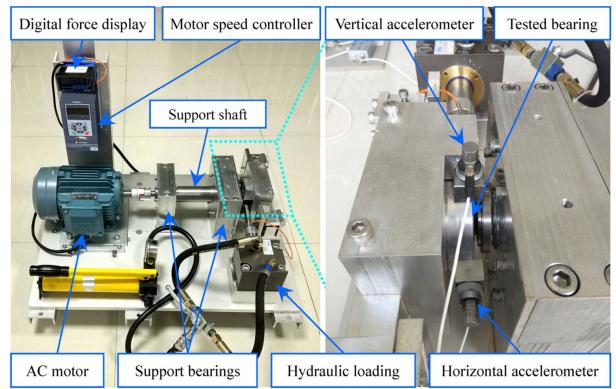


Fig. 8. Experimental platform of the XJTU-SY bearing dataset.

testing. In addition, 11 manipulated variables and 22 process variables are considered as suggested in [26].

For the proposed method, the confidence limit α is set as 0.99, the graph Laplacian matrix \mathbf{L} is obtained as in [3], and the parameters λ , τ_1 , and τ_2 are determined by the fivefold cross-validation technique for fault-free data within the candidate set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. The value of sparsity level s is chosen from a smaller value and is increased until attaining the best FD performance. Meanwhile, the maximum number of iterations for Algorithms 1–3 is 500, the limits ϵ_1 , ϵ_2 and ε for Algorithms 2 and 3 are all 10^{-4} , and the stopping criterion for Algorithm 1 is that the relative error

$$\max \left\{ \frac{\|\mathbf{W}^{k+1} - \mathbf{W}^k\|_F}{\|\mathbf{W}^k\|_F}, \frac{\|\mathbf{H}^{k+1} - \mathbf{H}^k\|_F}{\|\mathbf{H}^k\|_F} \right\} \leq 10^{-5}. \quad (23)$$

For all the other NMF-based FD methods, the parameters are also obtained by the fivefold cross-validation technique if they involve parameters, and the maximum numbers of iterations are set to be 500, and (23) is used as the stopping criterion, for fair comparison with SJSONMF proposed in this article. In addition, the FD rate (FDR) and the false alarm rate (FAR) are used as measurement metrics to evaluate the FD ability [43]. Taking T^2 statistic for example, FDR is defined as

$$\frac{\text{number of samples } (T^2 \geq J_{\text{th}, T^2} | f \neq 0)}{\text{total of samples } (f \neq 0)} \times 100\% \quad (24)$$

and FAR is defined as

$$\frac{\text{number of samples } (T^2 \geq J_{\text{th}, T^2} | f = 0)}{\text{total of samples } (f = 0)} \times 100\%. \quad (25)$$

It is evident that higher FDR value or lower FAR value demonstrates a better performance for FD.

B. Implementation Results

The detection results of PCA and NMF-based methods are shown in Tables I and II, respectively. Furthermore, the best performances are highlighted in bold.

From Tables I and II, we obtain that the FDR average values of the NMF-based methods are higher than the FDR average values of PCA. This illustrates the NMF-based methods are effective for FD. It is obvious to see that these NMF variants

outperform the classical NMF in terms of FDR and FAR. Especially, from FDR values for the fault IDV(16) in Table I, the T^2 statistic of NMF is 41.12%, and the T^2 statistic of SJSONMF is 74.25%. Thus, the T^2 statistic has a gain of 33.13% for the fault IDV(16) on the TEP. This shows that imposing proper constraints or regularization terms can improve the performance of FD. Furthermore, ONMF performs better for SPE statistic compared with GNMF and SNMF, which suggests that the orthogonality constraint has stronger detection potential. More importantly, the FDR average values of SJSNMF and SJSONMF are higher than those of KNMF, and the proposed SJSONMF model in this article obtains the best performance in most cases. It convinces us to believe that the combination of graph Laplacian, joint sparsity, and orthogonality into NMF is promising.

Then, two faults are chosen to further show the importance of SJSONMF in Figs. 2 and 3, where the red line and blue line represent the control limit and statistical value T^2 or SPE, respectively. Fig. 2 shows the detection performance of fault IDV(6). It is found that the values of FDR are over 99% for T^2 and SPE statistics for IDV(6), which reflects that the NMF-based methods complete the detection tasks successfully. Although this fault is relatively easy to detect, SJSONMF exhibits better detection ability; see the regional enlarged drawings in Fig. 2. Fault IDV(21) involves a constant position fault that the value of stream 4 was fixed at the steady-state position, and it is relatively difficult to detect. The detection performance is presented in Fig. 3. By observing the values between 600 and 800 samples for T^2 and SPE statistics, it is evident that SJSONMF detects more fault samples, which verifies the superiority over the other NMF-based methods.

Below, we aim to provide the reason why introducing the orthogonality constraint into the original SJSNMF model is helpful, from the computational results. Fig. 4 shows comparison of the obtained orthogonal basis matrix by PCA and the basis matrices by ONMF, SJSNMF, and SJSONMF for the TEP. Recall that the singular value decomposition (SVD) method is used to solve PCA. The mathematical expression of PCA is written as $(\mathbf{X}/(n-1)^{1/2}) = \mathbf{V}\Sigma^T\mathbf{U}^T$, where $\mathbf{V} \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{n \times m}$, $\mathbf{U} \in \mathbb{R}^{n \times n}$, and \mathbf{U} and \mathbf{V} are the orthogonal matrices. Here, \mathbf{V} and $(n-1)^{1/2}\Sigma^T\mathbf{U}^T$ can be viewed as \mathbf{W} and \mathbf{H} in the NMF-based methods, respectively.

Each entry in the basis matrix corresponds to a rectangular region in the image, and the values of entries are reflected by the colors of the rectangular. Fig. 4(b) and (d) demonstrates that the orthogonality regularization term/constraint together with the nonnegativity constraint provide the basis matrices that remove redundancy data information. Especially, Fig. 4(d) illustrates that Algorithm 2 ensures at most one positive element in each row of the basis matrix. It is found that by introducing the orthogonality constraint, the obtained matrix \mathbf{W} in Fig. 4(d) is more informative and discriminative. In contrast, Fig. 4(b) shows that there is a row that has more positive elements in the basis matrix provided by ONMF [24], which may be caused by not directly dealing with the orthogonality constraint in ONMF [24], but using a

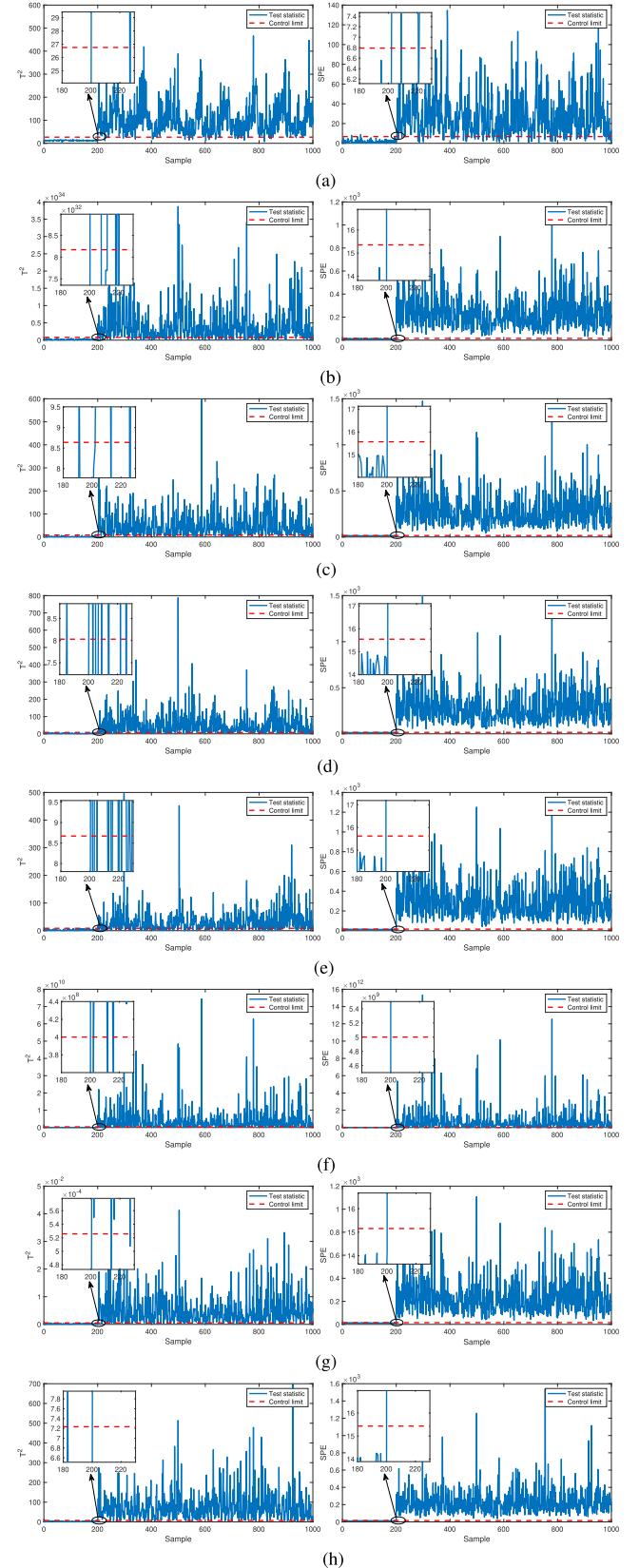


Fig. 9. Detection performance of Bearing 3_1 for XJTU-SY bearing. (a) PCA. (b) NMF. (c) GNMF. (d) SNMF. (e) ONMF. (f) KNMF. (g) SJSNMF. (h) SJSONMF.

regularized term in the objective to penalize the violation of orthogonality.

TABLE IV

FDR OF (a) PCA, (b) NMF, (c) GNMF, (d) SNMF, (e) ONMF, (f) KNMF, (g) SJSNMF, AND (h) SJSONMF FOR CWRU BEARING

Fault No.	(a) [PCA]		(b) [NMF]		(c) [GNMF]		(d) [SNMF]		(e) [ONMF]		(f) [KNMF]		(g) [SJSNMF]		(h) [SJSONMF]	
	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE
IDV(1)	91.33%	87.67%	94.00%	89.00%	95.88%	96.12%	94.67%	94.33%	89.00%	94.67%	94.00%	94.67%	90.33%	94.00%	96.33%	94.67%
IDV(2)	41.00%	48.00%	36.67%	28.33%	41.50%	48.33%	42.75%	54.00%	66.33%	45.67%	45.33%	55.33%	53.67%	61.67%	58.67%	63.67%
IDV(3)	74.67%	79.67%	77.00%	69.00%	80.88%	80.67%	82.00%	78.00%	77.67%	83.33%	83.00%	85.67%	81.67%	83.67%	83.67%	86.00%
Average	69.00%	71.78%	69.22%	62.11%	72.75%	75.04%	73.14%	75.44%	77.66%	74.56%	74.11%	78.56%	75.22%	79.78%	79.56%	81.45%

TABLE V

FAR OF (a) PCA, (b) NMF, (c) GNMF, (d) SNMF, (e) ONMF, (f) KNMF, (g) SJSNMF, AND (h) SJSONMF FOR CWRU BEARING

Fault No.	(a) [PCA]		(b) [NMF]		(c) [GNMF]		(d) [SNMF]		(e) [ONMF]		(f) [KNMF]		(g) [SJSNMF]		(h) [SJSONMF]	
	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE
IDV(1)	0.00%	0.50%	1.50%	1.00%	3.50%	2.50%	0.00%	4.00%	0.00%	0.00%	0.50%	0.50%	0.00%	0.00%	0.00%	0.00%
IDV(2)	0.00%	0.50%	2.50%	0.50%	4.00%	1.50%	0.00%	1.00%	0.00%	0.00%	1.50%	2.50%	0.50%	0.50%	0.00%	0.00%
IDV(3)	0.00%	0.50%	2.00%	0.50%	2.00%	1.50%	2.50%	0.50%	0.00%	0.00%	0.50%	0.50%	0.00%	0.00%	0.00%	0.00%
Average	0.00%	0.50%	2.00%	0.67%	3.17%	1.83%	0.83%	1.83%	0.00%	0.00%	0.83%	1.17%	0.33%	0.17%	0.00%	0.00%

TABLE VI

OPERATING CONDITIONS OF THE TESTED BEARING

Operating condition	Radial force (kN)	Rotating speed (rpm)	Bearing dataset
Condition 1	12	2100	Bearing 1_1 Bearing 1_2 Bearing 1_3 Bearing 1_4 Bearing 1_5 Bearing 2_1 Bearing 2_2 Bearing 2_3 Bearing 2_4 Bearing 2_5 Bearing 3_1 Bearing 3_2 Bearing 3_3 Bearing 3_4 Bearing 3_5
Condition 2	11	2250	
Condition 3	10	2400	

Furthermore, the orthogonality of the matrix \mathbf{W} can be defined as follows:

$$\text{Ortho}(\mathbf{W}) = \frac{\sum_{i=1}^r \sum_{j=1}^r (\mathbf{W}^T \mathbf{W})_{ij} - \text{tr}(\mathbf{W}^T \mathbf{W})}{\text{tr}(\mathbf{W}^T \mathbf{W})}. \quad (26)$$

The closer the matrix \mathbf{W} is to the orthogonal, the closer the value is to zero. The sparsity of \mathbf{W} can be measured as

$$\text{Spar}(\mathbf{W}) = \frac{\text{number of zero entries of } \mathbf{W}}{m \times r}. \quad (27)$$

Here, we take \mathbf{W}_{ij} as a zero entry as long as $|\mathbf{W}_{ij}| < 1.0e-16$. Table III gives the orthogonality and the sparsity of the basis matrix for the TEP, and it is obvious to find that the basis matrix obtained by SJSONMF has the best orthogonality and sparsity. Note that it is almost impossible to obtain the factors \mathbf{W} for KNMF since the explicit form of the nonlinear mapping is unknown [28]. Hence, we do not compute its sparsity here. However, $\mathbf{W}^T \mathbf{W}$ can be obtained by the kernel matrix. Thus, the orthogonality of \mathbf{W} is calculated in Table III.

It is clear that the combination of the newly added orthogonality constraint and the original nonnegativity constraint in the NMF-based model really provides a basis matrix that owns both the orthogonality and the sparsity, which greatly removes redundancy information among different basis vectors.

To verify the convergence of the proposed Algorithm 1, the relative error defined by (23) and the objective value versus elapsed time are plotted in Fig. 5. Obviously, both of them decrease rapidly after few steps, which shows that the proposed algorithms can efficiently solve the SJSONMF

model with nonconvex feasible set, which is rather difficult to be solved.

VI. APPLICATION ON THE BEARING DATASETS

In this section, two practical bearing datasets, CWRU¹ and XJTU-SY,² are used to verify the FD performance of the proposed method. The parameters and stopping criterion are determined in the same way as that in Section V.

The normal operation of rolling bearings has a great impact on the reliability, accuracy, and life of the equipment. The bearing condition monitoring technology can be used to determine whether they are in a state of health and further improve the equipment management level and maintenance efficiency [44]. Thus, FD methods based on the vibration for bearing are widely studied [44], [45], [46], [47], [48]. However, under the actual working conditions of the rotating machine, the collected bearing vibration signals are complex, the multisource information is coupled, and the characteristic components are weak. An FD method based on NMF, called semi-NMF, was proposed for bearing FD [49]. When the classical NMF method is used to process actual vibration signal, data redundancy will occur, and the features' extraction is not ideal. Hence, introducing orthogonality and sparsity can reduce the correlation between basis vectors and discard unimportant information.

A. Application on the CWRU Bearing Dataset

1) Data Preparation: Fig. 6 shows the platform which contains a motor, a torque, and a dynamometer. The data acquisition equipment (accelerometers) are installed at the drive end and fan end to acquire different working conditions including normal and faulty conditions. The CWRU bearing dataset contains three categories of fault data. The sampling frequency of the drive end fault data is 12 and 48 kHz, and the sampling frequency of the normal data and fan end fault data is 12 kHz. In this study, we consider the following three faults: IDV(1), the 12-kHz drive end bearing fault data, is composed of outer race (@3:00) fault with the motor load being 0, 1, 2, and 3 under the fault diameter being 0.021 in; IDV(2), the 48-kHz drive end bearing fault data, is composed of ball

¹<https://engineering.case.edu/bearingdatacenter/download-data-file>

²<https://biaowang.tech/xjtu-sy-bearing-datasets>

TABLE VII

FDR OF (a) PCA, (b) NMF, (c) GNMF, (d) SNMF, (e) ONMF, (f) KNMF, (g) SJSNMF, AND (h) SJSONMF FOR XJTU-SY BEARING

Fault No.	(a) [PCA]		(b) [NMF]		(c) [GNMF]		(d) [SNMF]		(e) [ONMF]		(f) [KNMF]		(g) [SJSNMF]		(h) [SJSONMF]	
	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE
Bearing 1_1	92.50%	88.00%	85.88%	100%	93.00%	100%	78.00%	100%	92.38%	100%	92.75%	100%	91.75%	100%	95.75%	100%
Bearing 1_2	98.25%	92.00%	94.25%	100%	93.88%	100%	97.62%	100%	92.25%	100%	94.75%	100%	99.00%	100%	99.25%	100%
Bearing 1_3	99.33%	92.00%	95.25%	100%	95.67%	100%	97.88%	100%	98.12%	100%	97.75%	100%	99.12%	100%	99.75%	100%
Bearing 1_4	39.67%	55.67%	28.75%	71.38%	33.12%	76.00%	45.62%	68.62%	22.12%	79.25%	22.75%	75.25%	23.12%	75.50%	44.25%	79.50%
Bearing 1_5	100%	99.75%	92.50%	100%	91.50%	100%	97.88%	100%	94.25%	100%	96.25%	100%	98.75%	100%	99.38%	100%
Bearing 2_1	100%	99.00%	89.62%	100%	93.38%	100%	94.88%	100%	99.62%	100%	97.25%	100%	97.62%	100%	97.25%	100%
Bearing 2_2	99.00%	92.50%	98.33%	100%	97.50%	100%	96.88%	100%	96.12%	100%	97.00%	100%	99.38%	100%	99.62%	100%
Bearing 2_3	99.75%	95.00%	95.25%	100%	91.88%	100%	96.88%	100%	91.88%	100%	96.62%	100%	97.62%	100%	99.50%	100%
Bearing 2_4	100%	99.75%	92.25%	100%	93.75%	100%	89.62%	100%	92.50%	100%	93.25%	100%	95.00%	100%	97.88%	100%
Bearing 2_5	99.88%	99.25%	88.88%	100%	89.00%	100%	87.75%	100%	88.25%	100%	87.75%	100%	93.38%	100%	96.25%	100%
Bearing 3_1	95.88%	88.75%	80.38%	100%	87.33%	100%	85.50%	100%	77.12%	100%	86.62%	100%	90.00%	100%	96.50%	100%
Bearing 3_2	90.88%	94.12%	74.25%	100%	77.12%	100%	71.50%	100%	75.25%	100%	83.75%	100%	87.38%	100%	94.25%	100%
Bearing 3_3	90.88%	94.12%	92.75%	100%	95.33%	100%	93.88%	100%	92.88%	100%	93.12%	100%	94.62%	100%	98.75%	100%
Bearing 3_4	95.88%	91.75%	91.50%	100%	93.38%	100%	95.25%	100%	93.50%	100%	97.12%	100%	98.50%	100%	98.25%	100%
Bearing 3_5	100%	99.88%	92.60%	100%	89.62%	100%	83.62%	100%	84.75%	100%	87.00%	100%	90.38%	100%	98.25%	100%
Average	93.62%	92.10%	86.16%	98.09%	87.70%	98.40%	87.50%	97.91%	86.07%	98.62%	88.25%	98.35%	90.37%	98.37%	94.33%	98.63%

TABLE VIII

FAR OF (a) PCA, (b) NMF, (c) GNMF, (d) SNMF, (e) ONMF, (f) KNMF, (g) SJSNMF, AND (h) SJSONMF FOR XJTU-SY BEARING

Fault No.	(a) [PCA]		(b) [NMF]		(c) [GNMF]		(d) [SNMF]		(e) [ONMF]		(f) [KNMF]		(g) [SJSNMF]		(h) [SJSONMF]	
	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE	T^2	SPE
Bearing 1_1	0.00%	0.50%	0.50%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.00%	0.00%	0.00%	0.00%
Bearing 1_2	0.00%	0.50%	1.00%	0.00%	1.50%	0.00%	1.50%	0.00%	2.50%	0.00%	0.00%	0.00%	2.50%	0.00%	0.00%	0.00%
Bearing 1_3	0.00%	1.50%	1.00%	0.00%	1.00%	0.00%	1.50%	0.00%	0.50%	0.00%	0.00%	0.00%	3.00%	0.00%	0.50%	0.00%
Bearing 1_4	0.00%	0.00%	1.00%	0.00%	0.50%	0.00%	0.50%	0.00%	0.50%	0.00%	0.50%	0.00%	1.00%	0.00%	0.00%	0.00%
Bearing 1_5	0.00%	1.00%	0.50%	0.00%	1.50%	0.00%	1.50%	0.00%	2.00%	0.00%	1.00%	0.00%	1.00%	0.00%	0.50%	0.00%
Bearing 2_1	0.00%	2.00%	1.00%	0.00%	0.50%	0.00%	1.50%	0.00%	0.00%	0.00%	0.00%	0.00%	2.50%	0.00%	0.00%	0.00%
Bearing 2_2	0.00%	0.50%	0.00%	0.00%	1.50%	0.00%	1.50%	0.00%	0.50%	0.00%	3.00%	0.00%	1.50%	0.00%	1.00%	0.00%
Bearing 2_3	0.00%	1.00%	1.50%	0.00%	3.00%	0.00%	1.50%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Bearing 2_4	0.00%	1.00%	0.50%	0.00%	1.00%	0.00%	0.50%	0.00%	0.50%	0.00%	0.50%	0.00%	2.50%	0.00%	0.00%	0.00%
Bearing 2_5	0.00%	0.50%	2.50%	0.00%	0.00%	0.00%	0.50%	0.00%	1.50%	0.00%	0.00%	0.00%	1.00%	0.00%	0.00%	0.00%
Bearing 3_1	0.00%	1.00%	0.00%	0.00%	1.00%	0.00%	0.50%	0.00%	0.50%	0.00%	0.00%	0.00%	1.00%	0.00%	0.00%	0.00%
Bearing 3_2	0.00%	1.00%	1.00%	0.00%	2.50%	0.00%	1.50%	0.00%	0.00%	0.00%	1.50%	0.00%	0.50%	0.00%	1.50%	0.00%
Bearing 3_3	0.00%	1.00%	2.00%	0.00%	0.00%	0.00%	0.50%	0.00%	2.00%	0.00%	0.00%	0.00%	0.50%	0.00%	1.50%	0.00%
Bearing 3_4	0.00%	0.50%	1.50%	0.00%	1.00%	0.00%	1.00%	0.00%	1.00%	0.00%	1.00%	0.00%	1.00%	0.00%	1.00%	0.00%
Bearing 3_5	0.00%	0.50%	2.50%	0.00%	1.50%	0.00%	0.50%	0.00%	0.00%	0.00%	2.50%	0.00%	3.00%	0.00%	2.50%	0.00%
Average	0.00%	0.83%	1.10%	0.00%	1.10%	0.00%	0.97%	0.00%	0.77%	0.00%	0.67%	0.00%	1.47%	0.00%	0.57%	0.00%

fault with the motor load being 0, 1, 2, and 3 under the fault diameter being 0.014 in; IDV(3), the 12-kHz fan end bearing fault data, is composed of inner race fault with the motor load being 0, 1, 2, and 3 under the fault diameter being 0.007 in. In all, 1500 samples are selected from normal data as training samples, and 500 samples including 200 samples selected from normal data and 300 samples selected from fault data as testing samples.

2) *Implementation Results:* Tables IV and V give the values of FDR and FAR for all the methods, respectively, and Fig. 7 presents the detailed monitoring performance of fault IDV(3). Through Tables IV and V, it is obvious that the NMF-based methods perform better than PCA in most cases. Meanwhile, compared with PCA and other NMF-based models, SJSONMF has the best detection performance in most cases, and its FDR average values for the above faults are the highest. For the relatively difficult fault IDV(2), ONMF obtains the highest T^2 value, but the proposed SJSONMF derives the highest SPE value. It is worth emphasizing that the proposed SJSONMF has the best detection performance on average. The good performance of this practical bearing data also shows that the proposed SJSONMF model is promising.

B. Application on the XJTU-SY Bearing Dataset

1) *Data Preparation:* The XJTU-SY bearing dataset is the full life cycle dataset, which is collected from the normal operation state to complete failure [47]. The experimental platform is shown in Fig. 8. This testbed is designed to conduct the accelerated degradation tests of rolling element bearings under different operating conditions (i.e., different radial force and rotating speed). The radial force is generated by the hydraulic loading system and applied to the housing of tested bearings, and the rotating speed is set and kept by the speed controller of the alternating current (ac) induction motor. In this study, a total of 15 faults are considered in Table VI, which are tested under three different operating conditions. For each fault, the training samples include 1000 normal data, and the testing samples include 200 normal data and 800 fault data.

2) *Implementation Results:* The values of FDR and FAR for all the models are given in Tables VII and VIII, and the detailed monitoring performance for Bearing 3_1 is presented in Fig. 9. The values of FDR for all the models are more than 90% in most cases, and the SPE values for the NMF-based models are all 100% except for Bearing 1_4. The FAR values

for all the models are less than 3.00%, where the SPE values for the NMF-based models are all 0.00%. This illustrates that the NMF-based models have better performance than PCA in terms of the SPE statistic. Especially, the proposed SJSONMF model solved by the proposed algorithms has obtained the highest FDR values for the majority of faults and performs the best in average. Therefore, the effectiveness of SJSONMF can be further illustrated by the XJTU-SY bearing dataset.

At the end of numerical experiments, it is worth pointing out that there are actually many differences between the two cases in Sections V and VI for the numerical experiments, including the TEP dataset for complex chemical production process and two practical bearing datasets. But the model and the algorithms proposed in this article all show good performances. The reason behind this can be explained as follows. First, our model extends the popular NMF model and inherits its ability to automatically extract sparse and easily interpretable factors, as well as its wide applicability. Unlike the squared envelope analysis [44], [45], [46] that focuses on bearing faults, our model is flexible for different types of faults including that from the chemical production process as well. Second, through incorporating graph regularization, row sparsity, and orthogonality constraints into the NMF model, the proposed SJSONMF model enjoys stronger discriminative ability, removes redundancy of different basis vectors, and improves the fault interpretability. Third, the PANLS algorithm can guarantee and speed up the convergence. Hence, the PANLS algorithm provides a reliable output of computed basis matrix, which is also important for FD performance.

VII. CONCLUSION

In this article, a data-driven FD model using the SJSONMF has been proposed to discard redundant information and improve detection interpretation and reliability. In algorithms, an optimization algorithm with solid convergence results has been designed. In experiments, the superiority has been validated on the benchmark TEP and two practical bearing datasets in comparison to the existing NMF-based methods. The practical importance is to improve the detection performance and verify the necessity of introducing orthogonality. In fact, this technique can also be applied to other data-driven FD methods.

Naturally, some faults are relatively difficult to detect as observed in this article, and thus, the combination with new techniques such as deep neural networks should be further investigated.

APPENDIX A

A. $q(\cdot)$ in (6) Is Strongly Convex and Strongly Smooth

Proof: By direct computation, for any \mathbf{H}^1 and \mathbf{H}^2

$$\begin{aligned} D_q &= q(\mathbf{H}^1) - q(\mathbf{H}^2) - \langle \nabla q(\mathbf{H}^2), \mathbf{H}^1 - \mathbf{H}^2 \rangle \\ &= \frac{1}{2} \left\langle \mathbf{H}^1 - \mathbf{H}^2, (\mathbf{W}^{k+1T} \mathbf{W}^{k+1} + \tau_2 \mathbf{I}) (\mathbf{H}^1 - \mathbf{H}^2) \right\rangle \\ &\quad + \lambda \text{tr} \left((\mathbf{H}^1 - \mathbf{H}^2) \mathbf{L} (\mathbf{H}^1 - \mathbf{H}^2)^T \right). \end{aligned} \quad (\text{A1})$$

Recall that \mathbf{L} is the graph Laplacian matrix in (2), which is positive semidefinite. Note that $\mathbf{W}^{k+1T} \mathbf{W}^{k+1} = \mathbf{I}$. Let $\underline{l} =$

$((1 + \tau_2)/2)$, and $\bar{L} = ((1 + \tau_2)/2) + \lambda \|\mathbf{L}\|_F$. It is clear that

$$\underline{l} \|\mathbf{H}^1 - \mathbf{H}^2\|_F^2 \leq D_q \leq \bar{L} \|\mathbf{H}^1 - \mathbf{H}^2\|_F^2. \quad (\text{A2})$$

The left inequality implies that the function $q(\cdot)$ is strongly convex, and the right inequality indicates that the function $q(\cdot)$ is strongly smooth.

B. Calculation of $\mathcal{P}_{\mathbb{S}_+}(\cdot)$ in Section III-B

$\mathcal{P}_{\mathbb{S}_+}(\mathbf{Z})$ may be a set due to the nonconvexity of \mathbb{S}_+ . In the numerical experiments, one matrix $\bar{\mathcal{P}}_{\mathbb{S}_+}(\mathbf{Z}) \in \mathcal{P}_{\mathbb{S}_+}(\mathbf{Z})$ is computed as follows. First, compute $\mathbf{y} \in \mathbb{R}^r$ with $\mathbf{y}_i = \|(\mathcal{P}_{\mathbb{R}_+^{r \times n}}(\mathbf{Z}))_{i \cdot}\|, i \in \{1, \dots, r\}$. Then, the MATLAB command MAXK is used to find the index set Γ corresponding to s largest elements of \mathbf{y} . Finally, $(\bar{\mathcal{P}}_{\mathbb{S}_+}(\mathbf{Z}))_{i \cdot} = (\mathcal{P}_{\mathbb{R}_+^{r \times n}}(\mathbf{Z}))_{i \cdot}$ when $i \in \Gamma$, and other elements are zero.

APPENDIX B

PROOF OF THEOREM 1

Before giving the proof of Theorem 1, it is necessary to introduce the following notations which will be used in the proof.

Let $\mathcal{L}(\mathbf{W}, \mathbb{O})$ be the linearized cone of \mathbb{O} at \mathbf{W} (see 2(9) in [51]. Let $\mathcal{T}(\mathbf{W}, \mathbb{O})$ be the tangent cone of \mathbb{O} at \mathbf{W} (see [36, Definition 6.1, p. 197]). Given a cone \mathcal{K} , \mathcal{K}^o represents the polar cone of \mathcal{K} (see 6(14) of [36].

Proof:

- 1) In view of \mathbf{W}_{k+1} being an optimal solution of (5) and \mathbf{H}_{k+1} being an optimal solution of (6) in Algorithm 1, it follows that

$$\begin{aligned} f(\mathbf{W}^{k+1}, \mathbf{H}^{k+1}) &\leq f(\mathbf{W}^{k+1}, \mathbf{H}^{k+1}) + \frac{\tau_2}{2} \|\mathbf{H}^{k+1} - \mathbf{H}^k\|_F^2 \\ &\leq f(\mathbf{W}^{k+1}, \mathbf{H}^k) \\ &\leq f(\mathbf{W}^{k+1}, \mathbf{H}^k) + \frac{\tau_1}{2} \|\mathbf{W}^{k+1} - \mathbf{W}^k\|_F^2 \\ &\leq f(\mathbf{W}^k, \mathbf{H}^k). \end{aligned} \quad (\text{A3})$$

Thus, $\{f(\mathbf{W}^k, \mathbf{H}^k)\}$ is nonincreasing. If $(\mathbf{W}^{k+1}, \mathbf{H}^{k+1}) \neq (\mathbf{W}^k, \mathbf{H}^k)$, then it is easy to see that $f(\mathbf{W}^{k+1}, \mathbf{H}^{k+1}) < f(\mathbf{W}^k, \mathbf{H}^k)$. Thus, the first one holds.

- 2) Note that the feasible region \mathbb{O} in (3) for \mathbf{W} is compact. Hence, $\{\mathbf{W}^k\} \subseteq \mathbb{O}$ is bounded, and \mathbf{W}^k is an orthogonal matrix for each k . The following arguments show that $\{\mathbf{H}^k\}$ is also bounded.

Assume on the contrary that $\{\mathbf{H}^k\}$ is unbounded. Then there exists an infinite subsequence $K \subseteq \{1, 2, \dots\}$ such that $\lim_{k \rightarrow \infty, k \in K} \|\mathbf{H}^k\|_F = +\infty$. In view of the boundedness of $\{\mathbf{W}^k\}$, there exists an infinite subsequence $K_1 \subseteq K$ such that

$$\lim_{k \rightarrow \infty, k \in K_1} \mathbf{W}^k = \mathbf{W}^*, \quad \text{where } \mathbf{W}^{*T} \mathbf{W}^* = \mathbf{I}. \quad (\text{A4})$$

Moreover, since $\{\mathbf{H}^k / \|\mathbf{H}^k\|_F\}$ is bounded, there exists an infinite subsequence $K_2 \subseteq K_1$ such that

$$\lim_{k \rightarrow \infty, k \in K_2} \frac{\mathbf{H}^k}{\|\mathbf{H}^k\|_F} = \bar{\mathbf{H}}, \quad \text{where } \|\bar{\mathbf{H}}\|_F = 1. \quad (\text{A5})$$

By the definition of $f(\mathbf{W}, \mathbf{H})$ in (2) and the fact that \mathbf{L} is positive semidefinite, it is easy to know that

$$\begin{aligned} f(\mathbf{W}^k, \mathbf{H}^k) &= \frac{1}{2} \|\mathbf{X} - \mathbf{W}^k \mathbf{H}^k\|_F^2 + \lambda \text{tr}(\mathbf{H}^k \mathbf{L} \mathbf{H}^{kT}) \\ &\geq \frac{1}{2} \|\mathbf{X} - \mathbf{W}^k \mathbf{H}^k\|_F^2. \end{aligned} \quad (\text{A6})$$

Hence

$$\begin{aligned} \lim_{k \rightarrow \infty, k \in K_2} \frac{f(\mathbf{W}^k, \mathbf{H}^k)}{\|\mathbf{H}^k\|_F^2} \\ \geq \lim_{k \rightarrow \infty, k \in K_2} \frac{1}{2} \left\| \frac{\mathbf{X}}{\|\mathbf{H}^k\|_F} - \mathbf{W}^k \frac{\mathbf{H}^k}{\|\mathbf{H}^k\|_F} \right\|_F^2 \\ = \frac{1}{2} \|\mathbf{W}^* \bar{\mathbf{H}}\|_F^2 = \frac{1}{2} \|\bar{\mathbf{H}}\|_F^2 = \frac{1}{2}. \end{aligned} \quad (\text{A7})$$

This implies that $\lim_{k \rightarrow \infty, k \in K_2} f(\mathbf{W}^k, \mathbf{H}^k) = +\infty$, which contradicts to 1) that $\{f(\mathbf{W}^k, \mathbf{H}^k)\}$ is bounded. Hence $\{(\mathbf{W}^k, \mathbf{H}^k)\}$ is bounded and there is at least one accumulation point. Thus, 2) holds.

- 3) Let $(\mathbf{W}^*, \mathbf{H}^*)$ be an arbitrary accumulation point of $\{(\mathbf{W}^k, \mathbf{H}^k)\}$. Then there exists an infinite subsequence $K \subseteq \{1, 2, \dots\}$ such that $\lim_{k \rightarrow \infty, k \in K} (\mathbf{W}^k, \mathbf{H}^k) = (\mathbf{W}^*, \mathbf{H}^*)$. Since $f(\mathbf{W}, \mathbf{H})$ is a continuous function, then

$$\lim_{k \rightarrow \infty, k \in K} f(\mathbf{W}^k, \mathbf{H}^k) = f(\mathbf{W}^*, \mathbf{H}^*). \quad (\text{A8})$$

From (A3), the nonincreasing property of $\{f(\mathbf{W}^k, \mathbf{H}^k)\}$ holds. Moreover, $\{f(\mathbf{W}^k, \mathbf{H}^k)\}$ is bounded from below because \mathbf{L} is a positive semidefinite matrix. Thus, $\{f(\mathbf{W}^k, \mathbf{H}^k)\}$ converges to $f(\mathbf{W}^*, \mathbf{H}^*)$. Taking $k \rightarrow \infty$ in (A3), the following results hold that:

$$\lim_{k \rightarrow \infty} \|\mathbf{H}^{k+1} - \mathbf{H}^k\|_F = 0 \quad (\text{A9})$$

and

$$\lim_{k \rightarrow \infty} \|\mathbf{W}^{k+1} - \mathbf{W}^k\|_F = 0. \quad (\text{A10})$$

Hence

$$\lim_{k \rightarrow \infty, k \in K} \mathbf{H}^{k+1} = \mathbf{H}^* \quad (\text{A11})$$

and

$$\lim_{k \rightarrow \infty, k \in K} \mathbf{W}^{k+1} = \mathbf{W}^*. \quad (\text{A12})$$

Since \mathbf{W}^{k+1} is a global minimizer of (5), it is obvious that \mathbf{W}^{k+1} is a local minimizer of (5). This, combined with [32, Th. 2.1], yields that

$$\begin{aligned} -\nabla l(\mathbf{W}^{k+1}) &= -\nabla w f(\mathbf{W}^{k+1}, \mathbf{H}^k) - \tau_1(\mathbf{W}^{k+1} - \mathbf{W}^k) \\ &\in \mathcal{L}^o(\mathbf{W}^{k+1}, \mathbb{O}). \end{aligned} \quad (\text{A13})$$

It has been shown in [32, Sec. II-A] that the feasible set \mathbb{O} satisfies the Guignard constraint qualifications (GCQs), i.e., $\mathcal{T}^o(\mathbf{W}^{k+1}, \mathbb{O}) = \mathcal{L}^o(\mathbf{W}^{k+1}, \mathbb{O})$. It is known that $\mathcal{N}(\mathbf{W}^{k+1}, \mathbb{O}) = \mathcal{T}^o(\mathbf{W}^{k+1}, \mathbb{O})$ (see [36, Proposition 6.5]). Thus,

$$-\nabla l(\mathbf{W}^{k+1}) \in \mathcal{N}(\mathbf{W}^{k+1}, \mathbb{O}). \quad (\text{A14})$$

Note that the objective function $q(\mathbf{H})$ in (6) is strongly convex and strongly smooth as shown in Appendix A-A. Meanwhile, \mathbf{H}^{k+1} is a local minimizer of (6) due to \mathbf{H}^{k+1} being a global minimizer of (6). Thus, integrated with the relationship (7) of Theorem 2.1 and Definition 2.1 in [35] yields that

$$0 \in \mathcal{P}_{\mathcal{T}(\mathbf{H}^{k+1}, \mathbb{S}_+)}(-\nabla q(\mathbf{H}^{k+1})). \quad (\text{A15})$$

By [51, Th. 3.4 (i)], (A15) holds if and only if

$$\begin{aligned} -\nabla q(\mathbf{H}^{k+1}) &= -\nabla_{\mathbf{H}} f(\mathbf{W}^{k+1}, \mathbf{H}^{k+1}) - \tau_2(\mathbf{H}^{k+1} - \mathbf{H}^k) \\ &\in \mathcal{N}(\mathbf{H}^{k+1}, \mathbb{S}_+). \end{aligned} \quad (\text{A16})$$

Taking $k \rightarrow \infty, k \in K$ in (A14) and (A16), and using (A11) and (A12), it follows that (15) in Definition 1 holds by replacing \mathbf{W} by \mathbf{W}^* and \mathbf{H} by \mathbf{H}^* , respectively. Hence, $(\mathbf{W}^*, \mathbf{H}^*)$ is a stationary point of (1). That is, statement 3) holds. \square

Remark 1: We explain in detail why the traditional ANLS algorithm cannot guarantee the convergence results in Theorem 1. The two proximal terms in the PANLS algorithm are necessary to show that (A11) and (A12) hold, that is, $(\mathbf{W}^{k+1}, \mathbf{H}^{k+1})$ converges to the same accumulation point $(\mathbf{W}^*, \mathbf{H}^*)$, provided that $(\mathbf{W}^k, \mathbf{H}^k)$ converges to $(\mathbf{W}^*, \mathbf{H}^*)$ as $k \rightarrow \infty, k \in K$. Thus, taking $k \rightarrow \infty, k \in K$ in (A14) yields

$$-\nabla_{\mathbf{W}} f(\mathbf{W}^*, \mathbf{H}^*) \in \mathcal{N}(\mathbf{W}^*, \mathbb{O}) \quad (\text{A17})$$

which is the requirement of \mathbf{W}^* for stationary point of (5).

The traditional ANLS algorithm has no proximal terms in the PANLS algorithm in Algorithm 1. If it is used, in contrast to (A14)

$$-\nabla_{\mathbf{W}} f(\mathbf{W}^{k+1}, \mathbf{H}^k) \in \mathcal{N}(\mathbf{W}^{k+1}, \mathbb{O}) \quad (\text{A18})$$

will be obtained. It cannot deduce (A17) without (A11) and (A12).

For the classical NMF or its variants that only have the nonnegativity constraints, without the orthogonality constraint on \mathbf{W} as well as the row sparsity constraint on \mathbf{H} as in the SJSONMF model, [39, Corollary 2] guarantees that any accumulation point of the sequence generated by the ANLS algorithm is a stationary point, as pointed out in the explanation before [40, Th. 2]. The orthogonality constraint on \mathbf{W} and the row sparsity constraint on \mathbf{H} make the feasible sets for subproblems of \mathbf{W} and \mathbf{H} not convex, while the convexities of feasible sets for subproblems are the basic assumptions throughout [39].

REFERENCES

- [1] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757–3767, Jun. 2015, doi: [10.1109/TIE.2015.2417501](https://doi.org/10.1109/TIE.2015.2417501).
- [2] S. J. Qin, "Survey on data-driven industrial process monitoring and diagnosis," *Annu. Rev. Control*, vol. 36, no. 2, pp. 220–234, Dec. 2012, doi: [10.1016/j.arcontrol.2012.09.004](https://doi.org/10.1016/j.arcontrol.2012.09.004).
- [3] 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, doi: [10.1109/TII.2018.2868364](https://doi.org/10.1109/TII.2018.2868364).

- [4] W. Jiang, T. Zhang, and H. Wang, "Joint sparse principal component analysis based robust sparse fault detection," in *Proc. IEEE 9th Data Driven Control Learn. Syst. Conf. (DDCLS)*, Nov. 2020, pp. 1234–1239, doi: [10.1109/DDCLS49620.2020.9275214](https://doi.org/10.1109/DDCLS49620.2020.9275214).
- [5] X. Kong, Z. Yang, J. Luo, H. Li, and X. Yang, "Extraction of reduced fault subspace based on KDICA and its application in fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 3505212, doi: [10.1109/TIM.2022.3150589](https://doi.org/10.1109/TIM.2022.3150589).
- [6] J. E. Garcia-Bracamonte, J. M. Ramirez-Cortes, J. de Jesus Rangel-Magdaleno, P. Gomez-Gil, H. Peregrina-Barreto, and V. Alarcon-Aquino, "An approach on MCSA-based fault detection using independent component analysis and neural networks," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 5, pp. 1353–1361, May 2019, doi: [10.1109/TIM.2019.2900143](https://doi.org/10.1109/TIM.2019.2900143).
- [7] G. Li, S. J. Qin, and D. Zhou, "Geometric properties of partial least squares for process monitoring," *Automatica*, vol. 46, no. 1, pp. 204–210, 2010, doi: [10.1016/j.automatica.2009.10.030](https://doi.org/10.1016/j.automatica.2009.10.030).
- [8] G. Lee, C. H. Han, and E. S. Yoon, "Multiple-fault diagnosis of the Tennessee Eastman process based on system decomposition and dynamic PLS," *Ind. Eng. Chem. Res.*, vol. 43, no. 25, pp. 8037–8048, 2004, doi: [10.1021/ie030084v](https://doi.org/10.1021/ie030084v).
- [9] J. Liu, C. Song, J. Zhao, and P. Ji, "Manifold-preserving sparse graph-based ensemble FDA for industrial label-noise fault classification," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 2621–2634, Jun. 2020, doi: [10.1109/TIM.2019.2930157](https://doi.org/10.1109/TIM.2019.2930157).
- [10] L. Yin, H. Wang, W. Fan, J. Zhou, and G. Yu, "Combining active learning and Fisher discriminant analysis for the semi-supervised process monitoring," *IFAC-PapersOnLine*, vol. 51, no. 21, pp. 147–151, 2018, doi: [10.1016/j.ifacol.2018.09.407](https://doi.org/10.1016/j.ifacol.2018.09.407).
- [11] Q. Lu, B. Jiang, R. B. Gopaluni, P. D. Loewen, and R. D. Braatz, "Sparse canonical variate analysis approach for process monitoring," *J. Process Control*, vol. 71, pp. 90–102, Nov. 2018, doi: [10.1016/j.jprocont.2018.09.009](https://doi.org/10.1016/j.jprocont.2018.09.009).
- [12] P.-E. P. Odiwei and Y. Cao, "Nonlinear dynamic process monitoring using canonical variate analysis and kernel density estimations," *IEEE Trans. Ind. Informat.*, vol. 6, no. 1, pp. 36–45, Feb. 2010, doi: [10.1109/TII.2009.2032654](https://doi.org/10.1109/TII.2009.2032654).
- [13] X. Xiu, L. Pan, Y. Yang, and W. Liu, "Efficient and fast joint sparse constrained canonical correlation analysis for fault detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 12, 2022, doi: [10.1109/TNNLS.2022.3201881](https://doi.org/10.1109/TNNLS.2022.3201881).
- [14] 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, doi: [10.1109/TCSII.2020.2988054](https://doi.org/10.1109/TCSII.2020.2988054).
- [15] X. Xiu, J. Fan, Y. Yang, and W. Liu, "Fault detection using structured joint sparse nonnegative matrix factorization," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 3505212, doi: [10.1109/TIM.2021.3067218](https://doi.org/10.1109/TIM.2021.3067218).
- [16] Z. Ren, W. Zhang, and Z. Zhang, "A deep nonnegative matrix factorization approach via autoencoder for nonlinear fault detection," *IEEE Trans. Instrum. Meas.*, vol. 16, no. 8, pp. 5042–5052, Aug. 2020, doi: [10.1109/TII.2019.2951011](https://doi.org/10.1109/TII.2019.2951011).
- [17] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013, doi: [10.1109/TKDE.2012.51](https://doi.org/10.1109/TKDE.2012.51).
- [18] J. Wang, I. Almasri, and X. Gao, "Adaptive graph regularized nonnegative matrix factorization via feature selection," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 963–966.
- [19] X. Zhang, D. Chen, and K. Wu, "Incremental nonnegative matrix factorization based on correlation and graph regularization for matrix completion," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 6, pp. 1259–1268, Mar. 2018, doi: [10.1007/s13042-018-0808-7](https://doi.org/10.1007/s13042-018-0808-7).
- [20] H. Jin, W. Yu, and S. Li, "Graph regularized nonnegative matrix tri-factorization for overlapping community detection," *Phys. A, Stat. Mech. Appl.*, vol. 515, pp. 376–387, Feb. 2019, doi: [10.1016/j.physa.2018.09.093](https://doi.org/10.1016/j.physa.2018.09.093).
- [21] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007, doi: [10.1109/TASL.2006.885253](https://doi.org/10.1109/TASL.2006.885253).
- [22] M. T. Belachew, "Efficient algorithm for sparse symmetric nonnegative matrix factorization," *Pattern Recognit. Lett.*, vol. 125, pp. 735–741, Jul. 2019, doi: [10.1016/j.patrec.2019.07.026](https://doi.org/10.1016/j.patrec.2019.07.026).
- [23] L. Dong, Y. Yuan, and X. Lu, "Spectral-spatial joint sparse NMF for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2391–2402, Mar. 2021, doi: [10.1109/TGRS.2020.3006109](https://doi.org/10.1109/TGRS.2020.3006109).
- [24] S. Motoki et al., "Sparse modeling of EELS and EDX spectral imaging data by nonnegative matrix factorization," *Ultramicroscopy*, vol. 170, pp. 43–59, Nov. 2016, doi: [10.1016/j.ultramic.2016.08.006](https://doi.org/10.1016/j.ultramic.2016.08.006).
- [25] Y. Wang, T. Guan, G. Zhou, H. Zhao, and J. Gao, "SOJNMF: Identifying multidimensional molecular regulatory modules by sparse orthogonality-regularized joint non-negative matrix factorization algorithm," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 6, pp. 3695–3703, Nov./Dec. 2022, doi: [10.1109/TCBB.2021.3114146](https://doi.org/10.1109/TCBB.2021.3114146).
- [26] X.-B. Li, Y.-P. Yang, and W.-D. Zhang, "Fault detection method for non-Gaussian processes based on non-negative matrix factorization," *Asia-Pacific J. Chem. Eng.*, vol. 8, no. 3, pp. 362–370, May 2013, doi: [10.1002/apj.1669](https://doi.org/10.1002/apj.1669).
- [27] X. Li, Y. Yang, and W. Zhang, "Statistical process monitoring via generalized non-negative matrix projection," *Chemometrics Intell. Lab. Syst.*, vol. 121, no. 7, pp. 15–25, 2013, doi: [10.1016/j.chemolab.2012.11.018](https://doi.org/10.1016/j.chemolab.2012.11.018).
- [28] L. Zhai, Y. Zhang, S. Guan, Y. Fu, and L. Feng, "Nonlinear process monitoring using kernel nonnegative matrix factorization," *Can. J. Chem. Eng.*, vol. 96, no. 2, pp. 554–563, Feb. 2018, doi: [10.1002/cje.22938](https://doi.org/10.1002/cje.22938).
- [29] Y. Wang, S.-M. Yuan, D. Ling, Y.-B. Zhao, X.-G. Gu, and B. Y. Li, "Fault monitoring based on adaptive partition non-negative matrix factorization for non-Gaussian processes," *IEEE Access*, vol. 7, pp. 32783–32795, 2019, doi: [10.1109/ACCESS.2019.2900499](https://doi.org/10.1109/ACCESS.2019.2900499).
- [30] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: [10.1038/44565](https://doi.org/10.1038/44565).
- [31] M. Asteris, D. Papailiopoulos, and A. G. Dimakis, "Orthogonal NMF through subspace exploration," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1–9.
- [32] B. Jiang, X. Meng, Z. Wen, and X. Chen, "An exact penalty approach for optimization with nonnegative orthogonality constraints," *Math. Program.*, early access, Feb. 2022, doi: [10.1007/s10107-022-01794-8](https://doi.org/10.1007/s10107-022-01794-8).
- [33] C. Zhang, L. Jing, and N. Xiu, "A new active set method for non-negative matrix factorization," *SIAM J. Sci. Comput.*, vol. 36, no. 6, pp. A2633–A2653, Jan. 2014, doi: [10.1137/130930212](https://doi.org/10.1137/130930212).
- [34] J. Hu, X. Liu, Z. Wen, and Y. Yuan, "A brief introduction to manifold optimization," *J. Oper. Res. Soc. China*, vol. 8, no. 2, pp. 199–248, 2020.
- [35] L. Pan, S. Zhou, N. Xiu, and H. Qi, "A convergent iterative hard thresholding for sparsity and nonnegativity constrained optimization," *Pacific J. Optim.*, vol. 13, no. 2, pp. 325–353, 2017.
- [36] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Berlin, Germany: Springer, 2009.
- [37] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 713–730, 2008, doi: [10.1137/07069239X](https://doi.org/10.1137/07069239X).
- [38] J. Kim and H. Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM J. Sci. Comput.*, vol. 21, no. 11, pp. 4508–4521, Nov. 2012, doi: [10.1137/110821172](https://doi.org/10.1137/110821172).
- [39] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss-Seidel method under convex constraints," *Oper. Res. Lett.*, vol. 26, no. 3, pp. 127–136, Apr. 2000, doi: [10.1016/S0167-6377\(99\)00074-7](https://doi.org/10.1016/S0167-6377(99)00074-7).
- [40] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, Oct. 2007, doi: [10.1162/neco.2007.19.10.2756](https://doi.org/10.1162/neco.2007.19.10.2756).
- [41] A. R. Mugdadi and I. A. Ahmad, "A bandwidth selection for kernel density estimation of functions of random variables," *Comput. Statist. Data Anal.*, vol. 47, no. 1, pp. 49–62, 2004, doi: [10.1016/j.csda.2003.10.013](https://doi.org/10.1016/j.csda.2003.10.013).
- [42] 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, doi: [10.1016/0098-1354\(93\)80018-I](https://doi.org/10.1016/0098-1354(93)80018-I).
- [43] S. X. Ding, *Data-Driven Design of Fault Diagnosis and Fault-Tolerant Control Systems*. London, U.K.: Springer, 2014.
- [44] B. An, Z. Zhao, S. Wang, S. Chen, and X. Chen, "Sparsity-assisted bearing fault diagnosis using multiscale period group lasso," *ISA Trans.*, vol. 98, pp. 338–348, Mar. 2020, doi: [10.1016/j.isatra.2019.08.042](https://doi.org/10.1016/j.isatra.2019.08.042).
- [45] Z. Diwu, H. Cao, L. Wang, and X. Chen, "Collaborative double sparse period-group lasso for bearing fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 3507110, doi: [10.1109/TIM.2020.3043940](https://doi.org/10.1109/TIM.2020.3043940).

- [46] X. Huang et al., "Memory residual regression autoencoder for bearing fault detection," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 3515512, doi: [10.1109/TIM.2021.3072131](https://doi.org/10.1109/TIM.2021.3072131).
- [47] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Trans. Rel.*, vol. 69, no. 1, pp. 401–412, Mar. 2020, doi: [10.1109/TR.2018.2882682](https://doi.org/10.1109/TR.2018.2882682).
- [48] J. Li, Y. Wang, Y. Zi, X. Sun, and Y. Yang, "A current signal-based adaptive semisupervised framework for bearing faults diagnosis in drivetrains," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 3508012, doi: [10.1109/TIM.2020.3046051](https://doi.org/10.1109/TIM.2020.3046051).
- [49] A. Rai and S. H. Upadhyay, "The application of semi-nonnegative matrix factorization for detection of incipient faults in bearings," *Proc. Inst. Mech. Eng., C, J. Mech. Eng. Sci.*, vol. 233, no. 13, pp. 4543–4555, Jul. 2019, doi: [10.1177/0954406219827332](https://doi.org/10.1177/0954406219827332).
- [50] R. Andreani, J. M. Martínez, A. Ramos, and P. J. S. Silva, "A cone-continuity constraint qualification and algorithmic consequences," *SIAM J. Optim.*, vol. 26, no. 1, pp. 96–110, Jan. 2016, doi: [10.1137/15m1008488](https://doi.org/10.1137/15m1008488).
- [51] L. Pan, N.-H. Xiu, and S.-L. Zhou, "On solutions of sparsity constrained optimization," *J. Oper. Res. Soc. China*, vol. 3, pp. 421–439, Dec. 2015, doi: [10.1007/s40305-015-0101-3](https://doi.org/10.1007/s40305-015-0101-3).



Xi Zhang received the B.S. degree in mathematics and applied mathematics from the University of Jinan, Jinan, China, in 2017, and the M.S. degree in mathematics from the Shandong University of Technology, Zibo, China, in 2020. She is currently pursuing the Ph.D. degree with the School of Mathematics and Statistics, Beijing Jiaotong University, Beijing, China.

Her current research interests include algorithms for sparse and Riemannian optimization and applications in fault detection.



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.



Chao Zhang received the B.S. degree in mathematics from the Normal School, Qingdao University, Qingdao, China, in 2001, the M.S. degree from Beijing Jiaotong University, Beijing, China, in 2004, and the Ph.D. degree from Hirosaki University, Hirosaki, Japan, in 2008.

She is currently a Professor with the School of Mathematics and Statistics, Beijing Jiaotong University. Her current research interests include algorithms for nonsmooth and nonconvex optimization and applications in data and image processing.