Fast Sparsity-Assisted Signal Decomposition with Non-Convex Enhancement for Bearing Fault Diagnosis

Zhibin Zhao, Shibin Wang, *Member, IEEE,* David Wong, *Member, IEEE,* Wendong Wang, Ruqiang Yan, *Senior Member, IEEE* and Xuefeng Chen, *Senior Member, IEEE*

***Abstract*—Sparsity-assisted signal decomposition based on morphological component analysis (MCA) for bearing fault diagnosis has been studied in depth. However, existing algorithms often use different combinations of representation dictionaries and priors, leading to difficult dictionary choice and high computational complexity. This paper aims to develop a fast sparsity-assisted algorithm to decompose a vibration signal into discrete frequency and impulse components for bearing fault diagnosis. We introduce the morphological discrimination of discrete frequency and impulse components in time and frequency domains respectively for the first time. To use this morphological discrimination, we establish a fast sparsity-assisted signal decomposition (SASD) based on MCA with non-convex enhancement. We further prove the necessary and sufficient condition to guarantee the convexity and use the majorization minimization (MM) algorithm to derive a fast solver. The proposed algorithm not only has low computational complexity, but also avoids choosing multiple dictionaries as well as underestimation of impulse features. Furthermore, an adaptive parameter selection algorithm to set parameters of our algorithm is designed for real applications. The effectiveness of fast SASD and its adaptive variant is verified by both simulation studies and bearing diagnosis cases. The source codes will be released at: https://github.com/ZhaoZhibin/Fast_SASD**

***Index Terms*—morphological component analysis, sparsity-assisted signal decomposition, convex optimization, enhanced fault diagnosis.**

I. Introduction

CONDITION monitoring and fault diagnosis are becoming increasingly popular for modern machinery, such as high-speed trains, helicopters, and aero-engines [1], [2]. Vibration signal analysis is one of the most important tools to realize effective condition monitoring [3]. Modern machinery is often complex and operates in harsh environment, causing that fault features are often submerged by strong background noise, especially at the early stage of faults. Such that traditional indicators and spectrum analysis methods perform poorly.

To address this problem, many advanced signal processing methods have been proposed in last twenty years. Fast spectrum kurtosis (fast SK), as a pioneer of resonance band selection methods, was originally proposed by Antoni [4], and improved variants have been widely studied [5]. Meanwhile, Antoni et al. [6] established the theory and application of cyclostationary approaches for rotating machinery diagnosis. Later, they also discussed a particular class of non-stationary signals called cyclo-non-stationary [7]. Wavelet transform [8] and time-frequency analysis [9] have also been well studied in the field of fault diagnosis. However, most of the above-mentioned methods often limit denoising performance due to the fact that they lack an iterative noise reduction process. It is worth mentioning that recently artificial intelligence methods, especially deep learning models, have attracted increasing popularity to automatically extract fault information for final fault diagnosis. Li et al. [10] proposed a novel deep distance metric learning method to learn fault features with small intra-class and large inter-calss variations. Zhao et al. [11] provided a benchmark accuracy of different deep learning models on seven open-source datasets, and released the whole code library for reproduction. However, these methods usually require plenty of labeled samples with different working conditions, which is beyond the scope of this paper.

Benefitting from excellent denoising performance, sparsity-assisted methods have shown to be effective for machinery fault diagnosis [12], [13]. Nevertheless, their applications are limited by choice of a suitable dictionary and high computational complexity. In response, He et al. [14] proposed a Periodicity-induced Overlapping Group Shrinkage (POGS) to model impulses in the time domain, which avoids choosing dictionaries and possesses low computational complexity. Ding et al. [15] extended POGS to internal encoder data. Wang et al. [16] extended the periodic information into the low rank representation. Zhao et al. [17] revealed that impulses actually have Periodic Sparsity Within and Across Groups (PSWAG) which was extended to the wavelet domain in [**?**], [18]. In addition, Hao et al. [19] proposed a step-by-step compound faults diagnosis method based on MM and constraint sparse component analysis for rotating machinery diagnosis. These methods only assume the noise interference, ignoring discrete frequency components such as pure harmonic (e.g. rotating

This work was supported in part by the Natural Science Foundation of China under Grand 91860125 and in part by the China Postdoctoral Science Foundation under Grand 2021M692557 and 2021TQ0263. *Corresponding Author: S. Wang*.

Z. Zhao, S. Wang R. Yan and X. Chen are with the State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. (Email: zhaozhibin@xjtu.edu.cn; wangshibin2008@gmail.com;yanruqiang@xjtu.edu.cn; chenxf@mail.xjtu.edu.cn).

D. Wong are with Centre for Health Informatics, University of Manchester, United Kingdom. (Email: david.wong@manchester.ac.uk)

W. Wang is with National Graphene Institute, University of Manchester, United Kingdom. (Email: wendong.wang@postgrad.manchester.ac.uk)

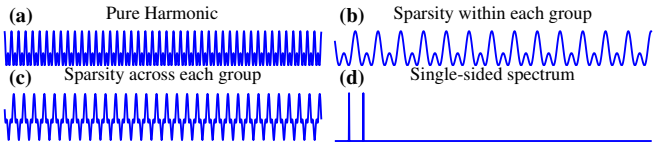


Fig. 1. Morphological discrimination of pure harmonic between time domain (not sparse within and across groups) and frequency domain (sparse in the single-sided spectrum).

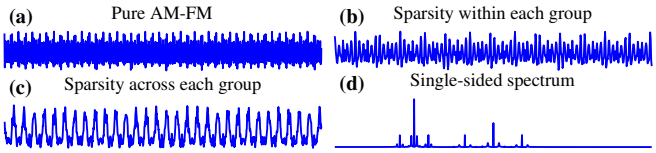


Fig. 2. Morphological discrimination of pure AM-FM between time domain (not sparse within and across groups) and frequency domain (sparse in the single-sided spectrum).

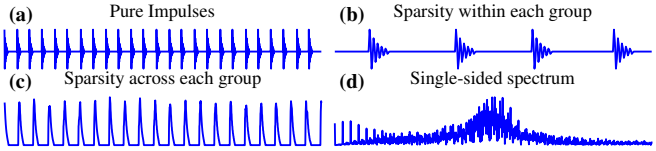


Fig. 3. Morphological discrimination of pure impulses between time domain (sparse within and across groups) and frequency domain (not sparse in the single-sided spectrum).

frequencies) as well as pure amplitude modulation and frequency modulation (AM-FM).

To remedy this drawback and maintain the excellent denoising performance, signal decomposition methods based on sparse representation and morphological component analysis (MCA) have been derived. MCA was proposed in image processing [20], [21] and was used in the field of fault diagnosis by Cai et al. [22]. Later, MCA-based methods continued gaining attention for bearing and gear fault diagnosis. Du et al. [23] used a union of redundant dictionaries to identify multiple components for wind turbine gearbox diagnosis. Shi et al. [24] proposed an iterative oscillatory behavior based signal decomposition for bearing fault diagnosis. Li et al. [25] considered the parameter selection of tunable Q-factor wavelet transform (TQWT) for MCA-based methods. Qin et al. [26] used a new family of impulse wavelets and Fourier bases based on MCA (IWF-MCA) for signal decomposition. To promote sparsity and maintain the amplitude, non-convex regularizers were also used in [27]–[31] for signal decomposition. However, there are two disadvantages of these MCA-based methods: choosing suitable representation dictionaries for multiple components is often very hard; the computational complexity is even higher than denoising methods.

In this paper, to avoid choosing multiple dictionaries and to reduce the computational complexity, we propose an algorithm called fast sparsity-assisted signal decomposition (fast SASD) based on the morphological discrimination between time and frequency domains. We first reveal that pure harmonic and AM-FM components (also called discrete frequency components) are very sparse in the frequency spectrum while they do not have PSWAG in the time domain shown in Fig. 1 and Fig. 2. Inversely, pure periodic impulses (also called impulse components) have PSWAG in the time domain while they are not sparse in the frequency spectrum shown in Fig. 3. Then, according to this morphological discrimination between PSWAG in the time domain and sparsity in the frequency domain, we construct a fast SASD model with non-convex enhancement for maintaining the amplitudes. After that, we prove the necessary and sufficient condition of the convexity of fast SASD and its extended variant. It is worth mentioning that we directly model the discrete frequency components in the frequency domain and model periodic impulses in the time domain, leading to determinate representation dictionaries (unit dictionary and Fourier dictionary). That is, we do not need to choose suitable representation dictionaries via tuning related parameters. MM is used to decouple multiple components and deduce a fast solver with low computational complexity. Because it is important to choose suitable penalty parameters in the decompositon performance, we establish a straight-forward strategy based on deterministic rules, the Golden ratio, and Trichotomy to set these parameters adaptively. Finally, a series of simulation studies and experiment cases are performed to verify the effectiveness and speed of the proposed algorithm.

In summary, the main contributions of this paper are listed as follows:

1) To the best of our knowledge, we introduce the morphological discrimination between PSWAG in the time domain and sparsity in the frequency domain for the first time and make use of this morphological discrimination to construct a fast SASD model which avoids choosing dictionaries and has low computational complexity.
2) We use non-convex regularizers to maintain the amplitude of impulse components, investigate the convexity of our constructed model, and prove the necessary and sufficient condition to guarantee the convexity. We further deduce a fast solver based on the MM algorithm.
3) A fast and adaptive algorithm to set the trade-off parameter of our method is also designed to allow the proposed approach more applicable. Furthermore, the effectiveness of fast SASD and its adaptive variant is verified by both simulation studies and bearing diagnosis cases.

The rest of this paper is organized as follows: In Section II, the basic concepts of MCA and enhanced sparse period-group lasso are introduced. Section III constructs the SASD model, gives the convexity condition of the model, and deduces the algorithmic solver. After that, simulation studies and parameters setting are investigated in Section IV followed by the experimental verification. Section VI concludes this paper.

II. Preliminaries

A. Signal decomposition based on MCA

Suppose that the vibration signal $\mathbf{y} \in \mathbb{R}^n$ can be modeled as:

$$\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2 + \mathbf{n}, \tag{1}$$

where $\mathbf{y}_1 \in \mathbb{R}^n$ represents discrete frequency components, $\mathbf{y}_2 \in \mathbb{R}^n$ represents periodic impulse components, and $\mathbf{n} \in \mathbb{R}^n$ is Gaussian noise. MCA aims to decompose $\mathbf{y}$ into two components accompanying with noise removal.

For the signal model (1), MCA assumes that components $\mathbf{y}_1$ and $\mathbf{y}_2$ have morphological discrimination under two specific dictionaries $\mathbf{D}_1 \in \mathbb{R}^{n\times m_1}$ and $\mathbf{D}_2 \in \mathbb{R}^{n\times m_2}$. Then, $\mathbf{y}_1$ and $\mathbf{y}_2$ can be recovered via calculating the following optimization problem:

$$\begin{aligned}\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2\} = &\arg\min_{\mathbf{x}_1,\mathbf{x}_2} \frac{1}{2}\|\mathbf{y} - \mathbf{D}_1\mathbf{x}_1 - \mathbf{D}_2\mathbf{x}_2\|_2^2 \\ &+ \lambda_1 P_1(\mathbf{x}_1) + \lambda_2 P_2(\mathbf{x}_2),\end{aligned} \tag{2}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are representative coefficients corresponding to two dictionaries $\mathbf{D}_1$ and $\mathbf{D}_2$, $P_1(\cdot)$ and $P_2(\cdot)$ are two regularizers of the representative coefficients, and $\lambda_1 > 0$ and $\lambda_2 > 0$ are two penalty parameters. After solving the optimization problem (2), two specific components can be estimated via $\hat{\mathbf{y}}_1 = \mathbf{D}_1\hat{\mathbf{x}}_1$ and $\hat{\mathbf{y}}_2 = \mathbf{D}_2\hat{\mathbf{x}}_2$.

The most common MCA model used in vibration signal decomposition is the sparsity-assisted method which assumes that $\mathbf{y}_1$ has a sparse representation with $\mathbf{D}_1$ while having a dense representaion with $\mathbf{D}_2$, and vice versa for $\mathbf{y}_2$. The regularizers are all set as l_1-norm, e.g. $P_1(\mathbf{x}_1) = \|\mathbf{x}_1\|_1$ and $P_2(\mathbf{x}_2) = \|\mathbf{x}_2\|_1$ where $\|\mathbf{x}\|_1 = \sum_i |x_i|$. Also, the most used dictionary is the TQWT dictionary due to its flexibility and efficiency in designing a suitable oscillatory waveform via tuning the Q-factor and the redundant factor r.

B. Enhanced sparse period-group lasso

The regularizer of enhanced sparse period-group lasso proposed in [17], [32] which possesses the PSWAG property is defined as follows:

$$P(\mathbf{x}) = \sum_{i\in\Omega} \phi\left(\|\mathbf{b}\odot\mathbf{x}_{\{i,n_2\}}\|_2; a\right) + \rho M n_1 \|\mathbf{x}\|_{1,\mathbf{W}}, \tag{3}$$

where Ω is a set of groups $\mathbf{x}_{\{i,n_2\}} = [x(i), ..., x(i+n_2-1)]^T \in \mathbb{R}^{n_2}$, $\mathbf{b}$ is a binary periodic sequence with the length of an estimated impulse n_1 and the number of impulses M ($n_1 = 4$ and $M = 4$ have been verified as an effective combination [17], and we simply follow this setting), $\odot$ represents the element-wise multiplication, $\rho = 9.235\times10^{-4}$ [17] is a trade-off parameter between sparisty within and across groups, $\phi(\cdot; a)$ is a penalty function defined in [33] ($a > 0$ is a parameter controlling the non-convex degree and we use the minimax convex penalty (MCP) [34] due to good properties in the unbiased estimation), and $\|\mathbf{x}\|_{1,\mathbf{W}}$ is defined as:

$$\|\mathbf{x}\|_{1,\mathbf{W}} = \|\mathbf{W}\mathbf{x}\|_1, \tag{4}$$

where $\mathbf{W} = \mathrm{diag}(\mathbf{w})$ is the diagonal matrix with $w_i = \frac{1}{|x_i|+\epsilon}$ ($\epsilon > 0$ is a small constant value).

III. Fast Sparsity-assisted Signal Decomposition

A. Model construction

Discrete frequency components are sparse in the frequency spectrum while they do not have PSWAG in the time domain shown in Fig. 1 and Fig. 2. Inversely, periodic impulse components have PSWAG in the time domain while they are not sparse in the frequency spectrum shown in Fig. 3. As a consequence, discrete frequency and periodic impulse components have the morphological discrimination between PSWAG in the time domain and sparsity in the frequency domain. Based on MCA, we let $\mathbf{D}_1 = \mathbf{D}$ be the Fourier dictionary and $\mathbf{D}_2$ be a unit matrix $\mathbf{I}$ and construct a fast SASD model as follows:

$$\begin{aligned}J(\mathbf{x}_1,\mathbf{x}_2) &= \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}_1 - \mathbf{x}_2\|_2^2 + \lambda_1 P_1(\mathbf{x}_1) + \lambda_2 P_2(\mathbf{x}_2), \\ P_1(\mathbf{x}_1) &= \|\mathbf{x}_1\|_1 + \frac{\eta}{2}\|\mathbf{x}_1\|_2^2, \\ P_2(\mathbf{x}_2) &= \sum_{i\in\Omega} \phi\left(\|\mathbf{b}\odot\mathbf{x}_{2\{i,n_2\}}\|_2; a\right) + \rho M n_1 \|\mathbf{x}_2\|_{1,\mathbf{W}},\end{aligned} \tag{5}$$

where $\eta \geq 0$ is a trade-off parameter and is set as $\eta = 1$ ($P_1(\mathbf{x}_1)$ is an elastic net penalty [35]), and a non-convex penalty, $P_2(\mathbf{x}_2)$, models PSWAG in the time domain to maintain the amplitude of periodic impulse components.

The main difference between our proposed fast SASD model (5) and traditional MCA-based methods is the setting of representation dictionaries. For traditional MCA-based methods, wavelet and time-frequency dictionaries are often used to model the sparsity, which suffers from choosing suitable dictionaries and high computational complexity. Conversely, fast SASD directly models PSWAG in the time domain and sparsity in the frequency domain, which avoids choosing multiple dictionaries and possesses low computational complexity.

B. Theoretical analysis of convexity condition

To establish the convex condition of the optimization problem (5), we first split the model (5) into two independent parts:

$$J(\mathbf{x}_1,\mathbf{x}_2) = J_1(\mathbf{x}_1,\mathbf{x}_2) + J_2(\mathbf{x}_2). \tag{6}$$

J_1 and J_2 are defined as follows:

$$\begin{aligned}J_1(\mathbf{x}_1,\mathbf{x}_2) &= \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}_1 - \mathbf{x}_2\|_2^2 + \lambda_1 P_1(\mathbf{x}_1) - \frac{\gamma}{2}\|\mathbf{x}_2\|_2^2, \\ J_2(\mathbf{x}_2) &= \frac{\gamma}{2}\|\mathbf{x}_2\|_2^2 + \lambda_2 P_2(\mathbf{x}_2),\end{aligned} \tag{7}$$

where γ is intermediate and positive variable connecting two parts. It follows that J is convex, if J_1 and J_2 are both convex.

We can use **Proposition 1** in [17] to establish the convex condition of J_2 directly. Therefore, we just need to prove the convex condition of J_1.

Proposition 1. *(Given in [17]) Assume ϕ is one of the penalty functions satisfying the properties in [33]. Then $J_2(\mathbf{x}_2)$ is convex if $a \leq \frac{\gamma}{M n_1 \lambda_2}$.*

Theorem 1. *$J_1(\mathbf{x}_1,\mathbf{x}_2)$ is convex if and only if $\gamma \leq \frac{\lambda_1\eta}{\lambda_1\eta+1}$.*

Proof of Theorem 1. We first introduce $P_1(\mathbf{x}_1)$ and rewrite J_1 as:

$$\begin{aligned}J_1(\mathbf{x}_1,\mathbf{x}_2) &= \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \\ &\quad + \lambda_1(\|\mathbf{x}_1\|_1 + \frac{\eta}{2}\|\mathbf{x}_1\|_2^2) - \frac{\gamma}{2}\|\mathbf{x}_2\|_2^2 \\ &= \frac{1}{2}\|\mathbf{D}\mathbf{x}_1 + \mathbf{x}_2\|_2^2 + \frac{\lambda_1\eta}{2}\|\mathbf{x}_1\|_2^2 - \frac{\gamma}{2}\|\mathbf{x}_2\|_2^2 \\ &\quad + \frac{1}{2}\|\mathbf{y}\|_2^2 - \mathbf{y}^T(\mathbf{D}\mathbf{x}_1 + \mathbf{x}_2) + \lambda_1\|\mathbf{x}_1\|_1.\end{aligned} \tag{8}$$

In (8), $\frac{1}{2}\|\mathbf{y}\|_2^2 - \mathbf{y}^T(\mathbf{D}\mathbf{x}_1 + \mathbf{x}_2) + \lambda_1\|\mathbf{x}_1\|_1$ is convex in $(\mathbf{x}_1, \mathbf{x}_2)$. Hence, J_1 is convex if and only if the former quadratic term, denoted as L, is convex.

$$\begin{aligned} L &= \frac{1}{2}\|\mathbf{D}\mathbf{x}_1 + \mathbf{x}_2\|_2^2 + \frac{\lambda_1\eta}{2}\|\mathbf{x}_1\|_2^2 - \frac{\gamma}{2}\|\mathbf{x}_2\|_2^2 \\ &= \frac{1}{2}(\mathbf{x}_2;\ \mathbf{D}\mathbf{x}_1)^T \begin{pmatrix} \mathbf{I}-\gamma\mathbf{I} & \mathbf{I} \\ \mathbf{I} & \mathbf{I}+\lambda_1\eta\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}_2 \\ \mathbf{D}\mathbf{x}_1 \end{pmatrix}. \end{aligned} \tag{9}$$

If $\mathbf{H} = \begin{pmatrix} \mathbf{I}-\gamma\mathbf{I} & \mathbf{I} \\ \mathbf{I} & \mathbf{I}+\lambda_1\eta\mathbf{I} \end{pmatrix}$ is positive semidefinite, e.g. $\mathbf{H} \succeq 0$, then J_1 is convex. According to [36], there exists a permutation matrix $\mathbf{Q} \in \mathbb{R}^{2n\times 2n}$, such that

$$\mathbf{Q}\mathbf{H}\mathbf{Q}^T = \mathbf{H}' = \text{diag}(\mathbf{H}_1', \ldots, \mathbf{H}_n'), \tag{10}$$

where $\mathbf{H}_i' \in \mathbb{R}^{2\times 2}$ is defined as:

$$\mathbf{H}_i' = \begin{pmatrix} 1-\gamma & 1 \\ 1 & 1+\lambda_1\eta \end{pmatrix}, \quad i \in 1, \ldots, n. \tag{11}$$

Similar to [37], since $\mathbf{H}_i'$ is a real symmetric matrix, we can use the eigenvalue decomposition to further simplify the expression:

$$\mathbf{H}_i' = \mathbf{U}_i^T\mathbf{L}_i\mathbf{U}_i, \quad i \in 1, \ldots, n, \tag{12}$$

where $\mathbf{L}_i = \text{diag}(l_i^{(1)}, l_i^{(2)})$ ($l_i^{(1)}$ and $l_i^{(2)}$ are two real eigenvalues of $\mathbf{H}_i'$) and $\mathbf{U}_i$ is an orthonormal matrix consisting of eigenvectors of $\mathbf{H}_i'$. If we introduce two matrices $\mathbf{U}$ and $\mathbf{L}$ as follows:

$$\begin{aligned} \mathbf{U} &= \text{diag}(\mathbf{U}_1, \ldots, \mathbf{U}_n) \in \mathbb{R}^{2n\times 2n}, \\ \mathbf{L} &= \text{diag}(l_1^{(1)}, l_1^{(2)}, \ldots, l_n^{(1)}, l_n^{(2)}) \in \mathbb{R}^{2n\times 2n}, \end{aligned} \tag{13}$$

then $\mathbf{H}$ can be represented as:

$$\mathbf{H} = (\mathbf{U}\mathbf{Q})^T\mathbf{L}\mathbf{U}\mathbf{Q} = \mathbf{V}^T\mathbf{L}\mathbf{V}, \tag{14}$$

where $\mathbf{V}$ is a full-rank matrix. Therefore, $\mathbf{H} \succeq 0$ if and only if the matrix $\mathbf{L}$ is positive semidefinite, e.g. all the diagonal elements are non-negative. Due to the fact that all the diagonal elements come from matrices $\mathbf{H}_i', i \in 1, \ldots, n$, $\mathbf{H} \succeq 0$ if and only if all the matrices $\mathbf{H}_i', i \in 1, \ldots, n$ are positive semidefinite. Following Sylvester's criterion, this statement is satisfied if and only if

$$\begin{cases} 1-\gamma & \geq 0 \\ 1+\lambda_1\eta & \geq 0 \\ (1-\gamma)(1+\lambda_1\eta) - 1 & \geq 0 \end{cases} \Rightarrow \gamma \leq \frac{\lambda_1\eta}{\lambda_1\eta + 1}. \tag{15}$$

□

Theorem 2. *Assume $\phi(\cdot; a)$ is one of the penalty functions satisfying the properties in [33]. Then $J(\mathbf{x}_1, \mathbf{x}_2)$ is convex if and only if $a \leq \tau\frac{\lambda_1\eta}{Mn_1\lambda_2(1+\lambda_1\eta)}$, where $\tau \in [0, 1]$.*

Proof of Theorem 2. If we introduce another variable $\tau \in [0, 1]$, we can simply rewrite the parameter $\gamma = \tau\frac{\lambda_1\eta}{1+\lambda_1\eta}$. Then, combining with inequalities in **Proposition 1** and **Theorem 1**, $J(\mathbf{x}_1, \mathbf{x}_2)$ is convex if and only if $a \leq \tau\frac{\lambda_1\eta}{Mn_1\lambda_2(1+\lambda_1\eta)}$, where $\tau \in [0, 1]$. □

In addition, to make **Theorem 2** more complete, we also give the **Corollary 2.1** of the convex condition, when the l_1-norm regularizer in $P_1(\mathbf{x}_1)$ is also replaced with the non-convex penalty function $\phi(\cdot; a')$.

Corollary 2.1. *Assume $\phi(\cdot; a)$ and $\phi(\cdot; a')$ belong to the penalty functions satisfying the properties in [33]. Then $J(\mathbf{x}_1, \mathbf{x}_2)$ is convex if and only if $\lambda_1 a' + Mn_1\lambda_2 a \leq \tau\frac{\lambda_1\eta}{1+\lambda_1\eta}$, where $\tau \in [0, 1]$.*

From **Corollary 2.1**, it can be observed that two non-convex penalty functions $\phi(\cdot; a)$ and $\phi(\cdot; a')$ form a trade-off condition. Due to the good statistical properties of the elastic net [35], we mainly discuss $J(\mathbf{x}_1, \mathbf{x}_2)$ with $P_1(\mathbf{x}_1) = \|\mathbf{x}_1\|_1 + \frac{\eta}{2}\|\mathbf{x}_1\|_2^2$ in the following part. In addition, the trade-off parameter η is simply set as 1, and the non-convex degree a of $P_1(\mathbf{x}_2)$ is set as $a = 0.9\frac{\lambda_1\eta}{Mn_1\lambda_2(1+\lambda_1\eta)}$.

C. Model solving

Based on **Theorem 1**, we can preserve the convexity of the optimization problem (5), that is we can use convex optimization algorithms to achieve the optimal solution. Since the original optimization problem (5) is coupled, we use the MM algorithm to minimize the upper bound of (5) which means that the original optimization problem can be decoupled into two simple problems.

We first rewrite the data fidelity in (5) as:

$$\frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}_1 - \mathbf{x}_2\|_2^2 = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2, \tag{16}$$

where $\mathbf{A} = [\mathbf{D}, \mathbf{I}]$ and $\mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T]^T$. According to the descent lemma of the quadratic function, the majorizer of the data fidelity can be set as:

$$\begin{aligned} G_1(\mathbf{x}, \mathbf{u}) &= \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{u}\|_2^2 + (\mathbf{A}\mathbf{u} - \mathbf{y})^T\mathbf{A}(\mathbf{x} - \mathbf{u}) \\ &\quad + \frac{1}{2\mu}\|\mathbf{x} - \mathbf{u}\|_2^2 \\ &= \frac{1}{2\mu}\|\mathbf{x}\|_2^2 - \frac{1}{\mu}\left(\mathbf{u} - \mu\mathbf{A}^T(\mathbf{A}\mathbf{u} - \mathbf{y})\right)^T\mathbf{x} + C \\ &= \frac{1}{2\mu}\|\mathbf{x}\|_2^2 - \frac{1}{\mu}\mathbf{p}^T\mathbf{x} + C, \end{aligned} \tag{17}$$

where $\mathbf{u} = [\mathbf{u}_1^T, \mathbf{u}_2^T]^T$ is the iterative variable in MM, C is a value independent of $\mathbf{x}$, and $\mu \leq \frac{1}{L_\mathbf{A}} = 0.5$ is the step size of MM ($L_\mathbf{A}$ is the maximum eigenvalue of the matrix $\mathbf{A}^T\mathbf{A}$).

The sufficient conditions of MM for this data fidelity are satisfied as:

$$G_1(\mathbf{x}, \mathbf{u}) \geq \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2, \quad G_1(\mathbf{u}, \mathbf{u}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{u}\|_2^2. \tag{18}$$

For regularizers, $P_1(\mathbf{x}_1)$ is already uncoupled and we just need to construct the majorizer of $P_2(\mathbf{x}_2)$. According to [17], the majorizer of $P_2(\mathbf{x}_2)$ can be constructed as:

$$\begin{aligned} G_2(\mathbf{x}_2, \mathbf{u}_2) &= \sum_{i\in\Omega}\frac{\phi'\left(\|\mathbf{b}\odot\mathbf{u}_{2\{i,n_2\}}\|_2\right)}{\|\mathbf{b}\odot\mathbf{u}_{2\{i,n_2\}}\|_2}\|\mathbf{b}\odot\mathbf{x}_{2\{i,n_2\}}\|_2^2 \\ &\quad + \rho Mn_1\|\mathbf{x}_2\|_{1,\mathbf{W}} + C \\ &= \sum_i\frac{1}{2}r(i;\mathbf{u}_2)x_2^2(i) + \rho Mn_1\|\mathbf{W}\mathbf{x}_2\|_1 + C \\ &= \frac{1}{2}\mathbf{x}^T\mathbf{R}\mathbf{x} + \rho Mn_1\|\mathbf{W}\mathbf{x}_2\|_1 + C, \end{aligned} \tag{19}$$

where $\mathbf{R} = \text{diag}\left(\mathbf{r}(\cdot;\mathbf{u}_2)\right)$ and $\mathbf{r}(\cdot;\mathbf{u}_2)$ is defined as:

$$\mathbf{r}(\cdot;\mathbf{u}_2) = \sum_{j=0}^{n_2-1} b(j)\frac{\phi'\left(\|\mathbf{b}\odot\mathbf{u}_{2\{\cdot-j,n_2\}}\|_2\right)}{\|\mathbf{b}\odot\mathbf{u}_{2\{\cdot-j,n_2\}}\|_2}. \tag{20}$$

The sufficient conditions of MM for this regularizer are also satisfied as:

$$G_2(\mathbf{x}_2,\mathbf{u}_2) \geq P_2(\mathbf{x}_2), \quad G_2(\mathbf{u}_2,\mathbf{u}_2) = P_2(\mathbf{u}_2). \tag{21}$$

To sum up, the overall majorizer of the problem (5) can be constructed as:

$$G(\mathbf{x},\mathbf{u}) = G_1(\mathbf{x},\mathbf{u}) + \lambda_1 P_1(\mathbf{x}_1) + \lambda_2 G_2(\mathbf{x}_2,\mathbf{u}_2). \tag{22}$$

Therefore, the optimal solution of the optimization problem (5) can be solved by iterating the following step:

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} G\left(\mathbf{x},\mathbf{x}^k\right). \tag{23}$$

Since the iterative step (23) is separable between $\mathbf{x}_1$ and $\mathbf{x}_2$, (23) is equivalent to the following two steps:

$$\begin{aligned}\mathbf{x}_1^{k+1} &= \arg\min_{\mathbf{x}_1}\frac{1+\mu\eta\lambda_1}{2}\|\mathbf{x}_1\|_2^2 - \mathbf{p}_1\mathbf{x}_1 + \mu\lambda_1\|\mathbf{x}_1\|_1,\\ \mathbf{x}_2^{k+1} &= \arg\min_{\mathbf{x}_2}\frac{1}{2}\mathbf{x}_2^T(\mathbf{I}+\mu\lambda_2\mathbf{R})\mathbf{x}_2 - \mathbf{p}_2\mathbf{x}_2\\ &\quad + \mu\rho\lambda_2 M n_1\|\mathbf{W}\mathbf{x}_2\|_1,\end{aligned} \tag{24}$$

where $\mathbf{p}_1 = \mathbf{x}_1^k - \mu\mathbf{D}^T(\mathbf{A}\mathbf{x}^k - \mathbf{y})$ and $\mathbf{p}_2 = \mathbf{x}_2^k - \mu(\mathbf{A}\mathbf{x}^k - \mathbf{y})$.

Two optimization problems (24) both can be separated into series of single variable optimization problems which have the closed-form solutions, and hence we can directly calculate the solutions as:

$$\begin{aligned}\mathbf{x}_1^{k+1} &= \frac{1}{1+\mu\eta\lambda_1}\text{soft}(\mathbf{p}_1,\mu\lambda_1)\\ \mathbf{x}_2^{k+1} &= \mathbf{\Sigma}^{-1}\text{soft}(\mathbf{p}_2,\mu\rho\lambda_2 M n_1\mathbf{w})\end{aligned} \tag{25}$$

where $\mathbf{\Sigma} = \text{diag}\left(1+\mu\lambda_2\mathbf{r}(\cdot;\mathbf{x}_2^k)\right)$ and the soft-thresholding operator $\text{soft}(\cdot,\cdot)$ for one scalar is defined as:

$$\text{soft}(x,t) = \text{sign}(x)\max(|x|-t,0), \tag{26}$$

where $\text{sign}(\cdot)$ is a sign function.

In conclusion, the algorithmic process of the proposed fast SASD is formed in **Algorithm 1**.

D. Computational complexity

To proof the rapidity of fast SASD, we perform a line-by-line analysis of computational complexity. It should be mentioned that operations $\mathbf{D}$ and $\mathbf{D}^T$ can be realized effectively by iFFT and FFT with computational complexity $O(n\log n)$. Therefore, for line 6 and line 7 in **Algorithm 1**, computational complexity is to perform FFT and iFFT, which approximates $O(n\log n)$. For line 9, the computational cost mainly contains two convolutional operations, which approximate $O(n_2 n)$. The other lines mainly consist of simple additions and multiplications, and computational complexity approximates $O(n)$. To sum up, the total computational complexity of the proposed fast SASD (I is the iteration number) is:

$$\begin{aligned}CC &= O(I(2n\log n + n\log n + n + 2n_2 n + n + n))\\ &= O(In(3\log n + 3 + 2n_2))\\ &\approx O(In(\log n + n_2)).\end{aligned} \tag{27}$$

Algorithm 1 Fast SASD

1: **Input:** The vibration signal $\mathbf{y}\in\mathbb{R}^n$, $\lambda_1,\lambda_2>0,\rho = 9.235\times10^{-4}$, $\mu\leq\frac{1}{L_{\mathbf{A}}}=0.5$, $M=4$, $n_1=4$, $\eta=1$, and Iteration number $iter$.
2: **Initialization:** $\mathbf{x}_1^0=\mathbf{D}^T\mathbf{y}, \mathbf{x}_2^0=\mathbf{y}$
3: **Procedure:**
4: $a = 0.9\frac{\lambda_1\eta}{Mn_1\lambda_2(1+\lambda_1\eta)}$
5: **for** each $k\in[0,iter]$ **do**
6: $\quad\mathbf{p}_1 = \mathbf{x}_1^k - \mu\mathbf{D}^T(\mathbf{A}\mathbf{x}^k-\mathbf{y})$
7: $\quad\mathbf{p}_2 = \mathbf{x}_2^k - \mu(\mathbf{A}\mathbf{x}^k-\mathbf{y})$
8: $\quad\mathbf{w} = 1/\left(\left|\mathbf{x}_2^k\right|+\varepsilon\right)$
9: $\quad\mathbf{r}(\cdot;\mathbf{x}_2^k) = \sum_{j=0}^{n_2-1} b(j)\frac{\phi'\left(\|\mathbf{b}\odot\mathbf{x}_{2\{i-j,n_2\}}^k\|_2\right)}{\|\mathbf{b}\odot\mathbf{x}_{2\{i-j,n_2\}}^k\|_2}$
10: $\quad\mathbf{\Sigma} = \text{diag}\left(1+\mu\lambda_2\mathbf{r}(\cdot;\mathbf{x}_2^k)\right)$
11: $\quad\mathbf{x}_1^{k+1} = \frac{1}{1+\mu\eta\lambda_1}\text{soft}(\mathbf{p}_1,\mu\lambda_1)$
12: $\quad\mathbf{x}_2^{k+1} = \mathbf{\Sigma}^{-1}\text{soft}(\mathbf{p}_2,\mu\rho\lambda_2 Mn_1\mathbf{w})$
13: **end for**
14: **Output:** $\hat{\mathbf{x}}_1 = \mathbf{x}_1^{k+1}$ and $\hat{\mathbf{x}}_2 = \mathbf{x}_2^{k+1}$

IV. Simulation study and parameter selection

To verify the performance and set parameters of the proposed fast SASD, we perform a series of simulation studies.

A. Simulation construction

As defined in (1), the vibration signal $\mathbf{y}$ consists of discrete frequency components $\mathbf{y}_1$, periodic impulse components $\mathbf{y}_2$, and Gaussian noise $\mathbf{n}$, which are defined as follows:

- *Discrete frequency components:*

$$\begin{aligned}\mathbf{y}_1 &= (0.5+0.25\cos(2\pi\times180\mathbf{t}))\times\cos(2\pi\times1000\mathbf{t}\\ &+ 0.5\cos(2\pi\times35\mathbf{t})) + (0.25+0.25\cos(2\pi\times360\mathbf{t}))\\ &\times\cos(2\pi\times2000\mathbf{t}+0.5\cos(2\pi\times70\mathbf{t}))\\ &+ 0.5\cos(2\pi\times180\mathbf{t}) + 0.5\cos(2\pi\times360\mathbf{t}),\end{aligned} \tag{28}$$

  where carrier frequencies (CF) are 1000 Hz and 2000 Hz, amplitude modulation frequencies (AMF) are 180 Hz and 360 Hz, frequency modulation frequencies (FMF) are 35 Hz and 70 Hz, and pure harmonic frequencies (PHF) are 180 Hz and 360 Hz.

- *Periodic impulse components:*

$$\mathbf{y}_2 = \sum\nolimits_k \mathbf{Imp}\left(\mathbf{t}-kT-\tau_k\right), \tag{29}$$

  where $T = 0.01$ s represents the period of impulse components (we simulate the outer race fault whose characteristic frequency is denoted as BPFO). τ_k represents the slip effect, and $\mathbf{Imp}(\mathbf{t})$ is a single impulse defined as:

$$\mathbf{Imp}(\mathbf{t}) = e^{(-1000\mathbf{t})}\sin(4000\pi\times\mathbf{t}+20), \tag{30}$$

  where the initial phase is 20, the resonant frequency is 2000 Hz, and the decay factor is 1000.

- *Gaussian noise:*

$$\mathbf{n} = \mathcal{N}(0,\sigma^2), \tag{31}$$

  where σ is the standard deviation of Gaussian noise.

As shown in Fig. 4, the sampling frequency and length of the simulated signal are 16000 Hz and 8000 respectively, and

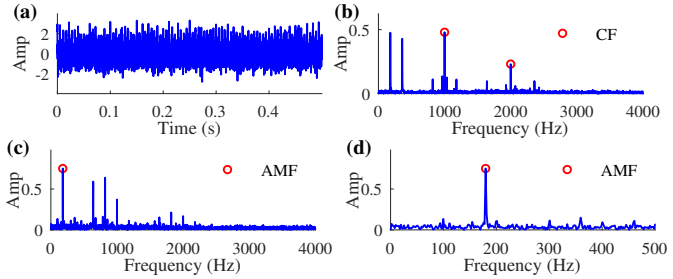


Fig. 4. The simulated vibration signal $\mathbf{y}$ and its spectra: (a) time waveform, (b) frequency spectrum, (c) square envelope spectrum, and (d) enlargement of SES.

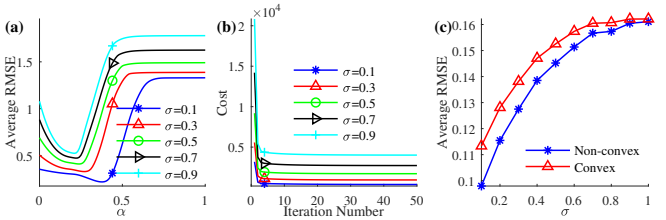


Fig. 5. (a) Average RMSE versus the trade-off paramter α, (b) the cost values versus the iteration number, and (c) comparison of non-convex and convex penalties.

the standard deviation of $\mathbf{n}$ is 0.5. We can observe that BPFO is completely submerged by the discrete frequency components in the square envelope spectrum (SES). Furthermore, because the resonant band couples with carrier frequencies, filter-based methods, such as SK, will fail to locate the impulse components.

All simulation and experiment studies were carried out under macOS Catalina 10.15.2 and MATLAB 2017b running on a personal computer with an Intel Core i7 6-CPU at 2.6 GHz and 16 GB RAM.

B. Parameter setting strategy

Parameters of fast SASD consist of λ_1, λ_2, and $iter$. We first consider the setting of two penalty parameters λ_1 and λ_2, separately. Here, we consider three scenarios. First, the vibration signal contains discrete frequency components and Gaussian noise. Due to the orthonormality of the FFT dictionary $\mathbf{D}$, we can use the universal estimation proposed by Donoho and Johnstone [38] to set λ_1 as:

$$\lambda_1 = \sigma\sqrt{2\ln n}, \tag{32}$$

where n is the length of the input signal. Second, the vibration signal contains periodic impulse components and Gaussian noise. Under this condition, we can directly use the fitting function in [17] which is defined as:

$$\lambda_2 = 0.272\sigma + 0.044. \tag{33}$$

Finally, we consider the combination of all three different parts, and redefine the penalty parameters λ_1 and λ_2 as:

$$\begin{aligned} \lambda_1 &= \alpha\sigma\sqrt{2\ln n}, \\ \lambda_2 &= (1-\alpha)(0.272\sigma + 0.044), \end{aligned} \tag{34}$$

where $0 \leq \alpha \leq 1$ is a trade-off parameter controlling the weight of discrete frequency and periodic impulse components.

To investigate the influence of α, we performed a series of numerical simulations with different σ to test the sum over the root mean square error (RMSE) of discrete frequency and periodic impulse components. In these simulation studies, α varies from 0 to 1 with the step 0.01 and σ varies from 0.1 to 0.9 with the step 0.2. To avoid randomness, 50 random tests are performed and the average RMSE is used as the metric. The results, in which the best α under different σ lies in a small interval, are shown in Fig. 5 (a). It means that the best α mainly depends on the weight of discrete frequency and periodic impulse components, which corresponds to the construction of the parameter selection strategy.

For real application, we further design a fast and adaptive algorithm described in **Algorithm 2** to search the optimal α via the Golden ratio and Trichotomy. It aims to maximize the value of fault characteristic energy ratio (FCER) defined as follows:

$$\text{FCER} = \frac{\sum_{i=1}^{floor(\frac{f_s}{2f_c})-1} A_{i\times f_c}}{\sum_{j=1}^{n} A_{j\times\Delta f}}, \tag{35}$$

where f_s represents the sampling frequency, $floor(\cdot)$ is to round down the value, and $A_{i\times f_c}$ and $A_{j\times\Delta f}$ represent amplitudes of the $i-$th order fault characteristic frequency f_c and the $j-$th spaced frequency Δf in the square envelope spectrum. In addition, fast SASD embedding **Algorithm 2** with the adaptive parameter setting strategy is denoted as fast adaptive SASD.

To further prove the superiority of the proposed method, we use another quantitative indicator called envelope kurtosis (EK) [4] to evaluate the impact in the time domain, and the defination can be formulated as follows:

$$\text{EK} = \frac{\mathbb{E}(\mathbf{h} - \text{Mean}(\mathbf{h}))^4}{\text{Var}(\mathbf{h})^2}, \tag{36}$$

where $\mathbb{E}(\cdot)$ denotes the calculation of mathematical expectation, $\mathbf{h}$ is the Hilbert envelope of extracted signals, $\text{Mean}(\cdot)$ denotes the calculation of mean, and $\text{Var}(\cdot)$ denotes the calculation of variance. To sum up, FCER and EK can be used to describe the fault characteristic energy in the square envelope spectrum and in the time domain, respectively.

For the iteration number $iter$, we perform similar simulation experiments with different σ to test the cost. In this simulation study, σ varies from 0.1 to 0.9 with the step 0.2. Fifty random tests are performed and the average cost is used as the metric. As shown in Fig. 5 (b), our algorithm has fast convergence, and thus we simply set $iter$ as 100 for more robust results.

C. Comparisons

To verify the performance adequately, we compare our algorithm with three methods including our proposed method with a convex penalty, the MCA-based method (IWF-MCA) [26], and the filtered-based method (fast SK) [4] with two quantitative indicators defined in (35) and (36).

Impulse extraction performance: The results of fast SASD, IWF-MCA, and fast SK are shown in Fig. 6, Fig. 7, and Fig. 8,

Algorithm 2 Adaptive and fast search algorithm of α

Input: Lower L and upper U bounds of α, $\epsilon = 0.01$.
Symbol: $\mathbf{x}_\alpha$ means impulse features obtained by fast SASD with α.
Procedure:
$\alpha_1 = L + 0.382(U - L)$, $\alpha_2 = L + 0.618(U - L)$
$v_1 = \text{FCER}(\mathbf{x}_{\alpha_1})$, $v_2 = \text{FCER}(\mathbf{x}_{\alpha_2})$
while $|U - L| > \epsilon$ **do**
  if $v_1 > v_2$ **then**
    $U = \alpha_2$, $\alpha_2 = \alpha_1$, $v_2 = v_1$
    $\alpha_1 = L + 0.382(U - L)$, $v_1 = \text{FCER}(\mathbf{x}_{\alpha_1})$
  else
    $L = \alpha_1$, $\alpha_1 = \alpha_2$, $v_1 = v_2$
    $\alpha_2 = L + 0.618(U - L)$, $v_2 = \text{FCER}(\mathbf{x}_{\alpha_2})$
  end if
end while
Output: $\alpha = (U + L)/2$

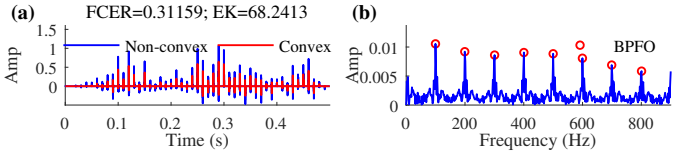


Fig. 6. The extracted impulses of fast SASD and its spectrum: (a) time waveform, and (b) enlargement of SES.

respectively. First of all, our proposed method can achieve highest FCER and EK among different methods. Because the resonance band and discrete frequency components overlap, the filtered-based methods, such as fast SK, cannot separate these two parts. As shown in Fig. 8 (b), the fault characteristic frequencies are still submerged by discrete frequency components. Although IWF-MCA can also extract BPFO and its high order frequencies, the extracted impulses in Fig. 7 (a) still contain lots of interference, which leads to unknown frequencies and amplitude attenuation of fault characteristic frequencies shown in Fig. 7 (b). Inversely, The impulses extracted by fast SASD show good periodicity in Fig. 6 (a) and Fig. 6 (b) also shows clean BPFO and its high order frequencies. In addition, as shown in Fig. 6 (a), fast SASD with a non-convex penalty can maintain the amplitude more effective than that with a convex penalty, which indicates the enhancement ability of our proposed method. To further verify the effectiveness of non-convex enhancement, we perform numerical studies with different noise intensities σ injected into the clean simulation signal, and σ varies from 0.1 to 1.0 with the step 0.1. Fifty random tests are conducted and average RMSE is also used to evaluate the performance of non-convex and convex penalties. The results are shown in Fig. 5 (c), and we can observe that the performance of the non-convex penalty is always better than that of the convex penalty with the noise intensities ranging from 0.1 to 1.

Computational complexity: To avoid the randomness, we run each method 10 times, shown in Fig. 4. The average times of four methods are listed in Table I. From these results, we can see that fast SASD and fast adaptive SASD are much quicker than IWF-MCA and fast SASD is slightly slower

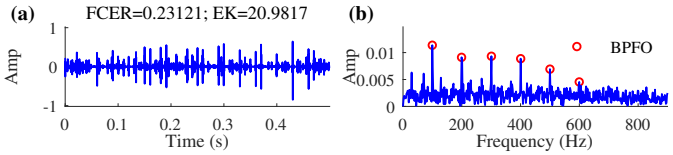


Fig. 7. The extracted impulses of IWF-MCA and its spectrum: (a) time waveform, and (b) enlargement of SES.

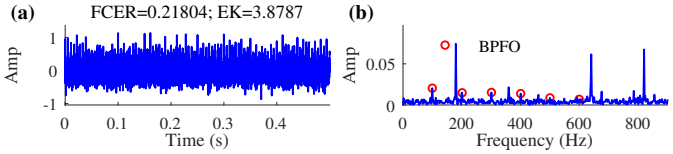


Fig. 8. The extracted impulses of Fast SK and its spectrum: (a) time waveform, and (b) enlargement of SES.

TABLE I
COMPARISON OF COMPUTATION TIMES

Method	Fast SASD	Fast adaptive SASD	Fast SK	IWF-MCA
Time (s)	0.1036	1.3120	0.0483	77.5799

than fast SK due to the fact that our proposed method is an iterative optimization approach and fast SK is just a filter-based approach. Meanwhile, fast adaptive SASD finding the optimal α adaptively only adds a little computational time compared with fast SASD, which indicates that **Algorithm 2** is very efficient for searching a suitable parameter in real application. It also worth mentioning that we can use other tricks to accelerate fast adaptive SASD, such as warm start and a smaller interval.

V. EXPERIMENTAL VERIFICATION

In this section, to verify the effectiveness and adaptation of our proposed algorithm, we apply it to different bearing fault diagnosis cases and compare with the MCA-based method (IWF-MCA) [26], and the filtered-based method (fast SK) [4]. What we expect is that the proposed method can separate periodic impulses as well as discrete frequency components effectively and enhance the fault information.

A. Case1

1) Experimental description: We performed a motor bearing fault experiment on the mechanical failure simulator of Spectra Quest, Inc. (SQI), which consists of the testing motor, the speed controller, the shaft, disks, the belt drive system, and the gearbox system, shown in Fig. 9 (a). During the experiment, two PCB accelerometers were mounted on horizontal and vertical directions of the testing motor shown in Fig. 9 (b). The CoCo80 data acquisition system was used to collect the vibration signal with the sampling frequency and the rotating frequency (RF) equivalent to 6400 Hz and 23.88 Hz, respectively. According to geometric parameters and RF, fault characteristic frequencies of the inner race (BPFI), the outer race (BPFO), rolling elements (BSF) and the cage (FTF) of the fault bearing in the testing motor are 117.8 Hz, 73.2Hz,

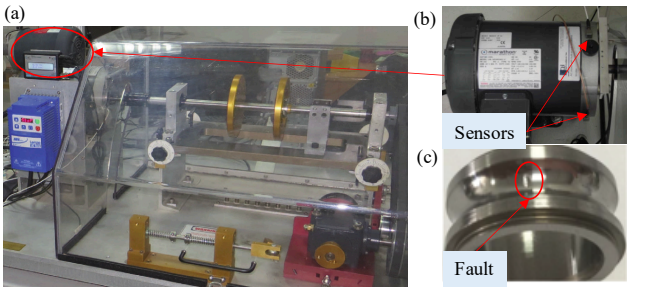


Fig. 9. (a) mechanical failure simulator, (b) the testing motor, and (c) pitting fault on the inner race.

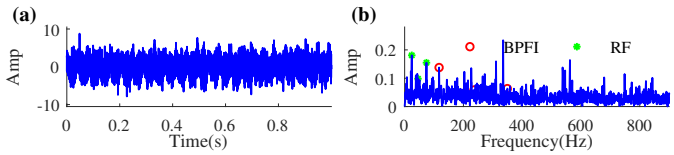


Fig. 10. The vibration signal and its spectrum: (a) time waveform, and (b) enlargement of SES.

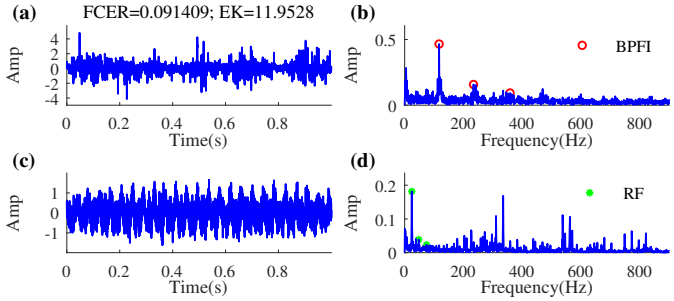


Fig. 11. Results extracted by fast adaptive SASD and its spectra: (a) time waveform of impulse components, (b) enlargement of SES of impulse components, (c) time waveform of discrete frequency components, and (d) enlargement of SES of discrete frequency components.

48.5 Hz, and 9.2Hz, respectively. After dismantling the testing motor, we found a pitting fault on the inner race described in Fig. 9 (c).

Fig. 10 shows the vibration signal on the vertical direction with 6400 points. From Fig. 10 (a), we can observe that the vibration signal is obviously coupled with discrete frequency components, and meanwhile, BPFI and its high order frequencies are submerged by those interference components shown in Fig. 10 (b).

2) Results: We first apply fast adaptive SASD to decompose the vibration signal into periodic impulse and discrete frequency components. The results, in which fast adaptive SASD can decompose two components successfully, are shown in Fig. 11. From Fig. 11 (a) and (b), impulse components are tremendously highlighted in the time domain, and BPFI in SES is much clearer than that in Fig. 10 (b).

To further verify the performance of our proposed algorithm, we also apply IWF-MCA and fast SK to analyse the vibration signal, and performance indicators defined in (35) and (36) are introduced to accurately compare their performance against each other. It is worth mentioning that we search the optimal trade-off parameter for IWF-MCA in the interval [0, 1] with the step 0.01, and the final optimal parameter is 0.01. First of all, our proposed method can achieve highest FCER

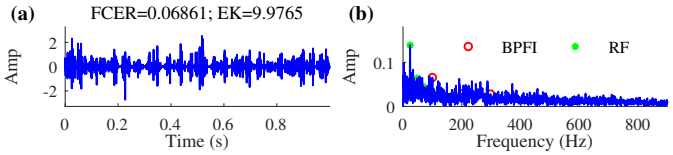


Fig. 12. Results extracted by IWF-MCA and its spectrum: (a) time waveform, and (b) enlargement of SES.

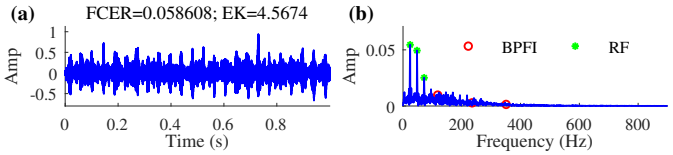


Fig. 13. Results extracted by fast SK and its spectrum: (a) time waveform, and (b) enlargement of SES.

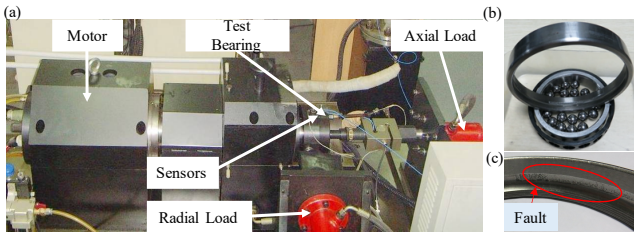


Fig. 14. (a) the main body of the test rig, (b) the fault on the outer race, and (c) the area of the fault.

and EK among different methods. The results of IWF-MCA and fast SK, in which both of them fail to extract periodic impulse components, are shown in Fig. 12 and Fig. 13. In addition, from Fig. 12 (b) and Fig. 13 (b), only RF and its high order frequencies are highlighted.

B. Case2

1) Experimental description: We performed a high-precision bearing fault experiment on the aero-engine bearing fault test rig, which was controlled by industrial personal computer (IPC) to simulate the load, temperature, and rotation spectra. Fig. 14 (a) shows the main body of the test rig, which is driven by the high-speed motor. The axial and radial loads were applied to the test bearing via the lubrication system. Two accelerometers were mounted on the horizontal and vertical directions of the bearing seat. The sampling frequency and the rotating speed are 20 kHz and 2000 r/min. Thus the calculated BPFO of the fault bearing is 275.9 Hz. As shown in Fig. 14 (b) and (c), the outer race of the test bearing had a local spalling.

As shown in Fig. 15 (a), we can observe that there are no obvious periodic impulses in the time domain. There exists complex low frequency interference and RF predominates in the enlargement of SES shown in Fig. 15 (b). Meanwhile, the twice order BPFO is submerged by the background noise, which makes diagnosis difficult and unreliable. It is necessary to apply advanced signal processing methods to remove the background noise and low frequency interference.

2) Results: We also first apply the fast adaptive SASD to extract impulse components, and the results are shown in Fig. 16. We can observe that periodic impulses are highlighted obviously in the time domain shown in Fig. 16 (a). BPFO and

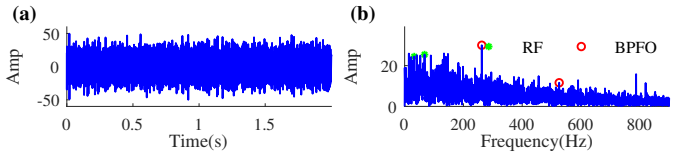


Fig. 15. The vibration signal and its spectrum: (a) time waveform, and (b) enlargement of SES.

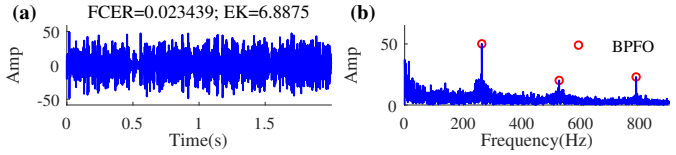


Fig. 16. Results extracted by fast adaptive SASD and its spectrum: (a) time waveform, and (b) enlargement of SES.

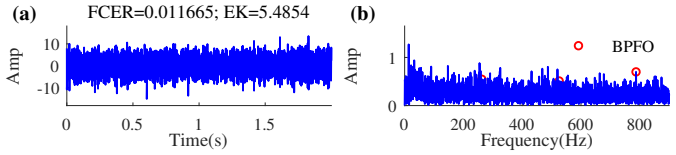


Fig. 17. Results extracted by IWF-MCA and its spectrum: (a) time waveform, and (b) enlargement of SES.

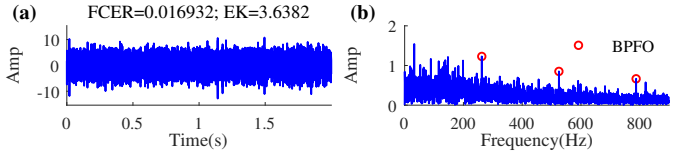


Fig. 18. Results extracted by fast SK and its spectrum: (a) time waveform, and (b) enlargement of SES.

its high order frequencies in SES shown in Fig. 16 (b) are much chearer than those in Fig. 15 (b). Meanwhile, the low frequency interference has a high degree of removed.

Similarly, we also apply IWF-MCA and fast SK to analyze the original signal to verify the performance of the proposed algorithm. We can observe that IWF-MCA fail to highlight the fault information and remove the background noise completely, as shown in Fig. 17. In addition, the quantitative indicators, FCER and EK, are smaller than that of the proposed method. Fast SK can also partially extract the fault information, but the low frequency interference in SES still predominates, and FCER as well as EK are also one half of the proposed method. It is also worth mentioning that the absolute amplitude of fault information extracted by IWF-MCA and fast SK in SES is much smaller than that extracted by the proposed method.

C. Case3

To evaluate the rate of false diagnosis and diagnose the rolling element fault, we further apply the proposed method to the open-source dataset from the Case Western Reserve University (CWRU) Bearing Data Center [39], which contains both normal and fault bearings.

1) Experimental description: The CWRU dataset was generated using a 2 hp Reliance Electric motor, and motor

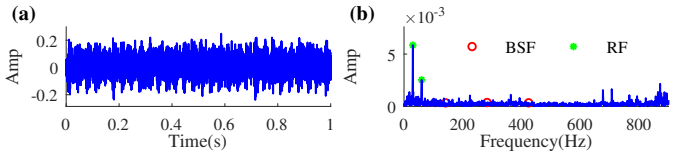


Fig. 19. The vibration signal and its spectrum of the normal bearing from CWRU: (a) time waveform, and (b) enlargement of SES.

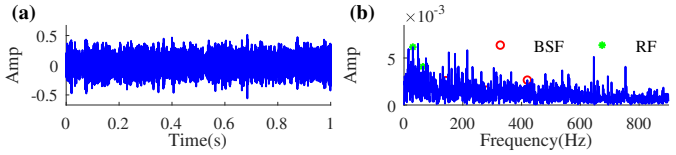


Fig. 20. The vibration signal and its spectrum of the fault bearing from CWRU: (a) time waveform, and (b) enlargement of SES.

bearings were seeded with faults using electro-discharge machining (EDM). Local faults ranging from 0.007 inches to 0.04 inches in diameter were implanted in different components of bearings, including the inner race, the outer race, and rolling elements, respectively. In this paper, according to the benchmark study of CWRU data [40], we used two data files numbered 97.mat (denoted as the normal bearing) and 118.mat (denoted as the rolling element fault with 0.007 inches in diameter at the drive end bearing). 118.mat cannot be diagnosed with any of applied methods in [40]. These two data files were collected for the motor load of 0 horsepower and the motor speed of 1797 r/min by the accelerometer, mounted on the drive end of the motor housing. The sampling frequency is 12000 Hz, and the BSF is 141.2 Hz. More detailed information can be found in the dataset website [39].

For signals from the normal bearing, as shown in Fig. 19, we can observe that there only exist RF and its high order frequencies in the enlargement of SES. Meanwhile, Fig. 20 reveals that the SES structure of the fault bearing is more complex than that of the normal bearing. However, the BSF and its high order frequencies are still submerged by other interference, that is why none of applied methods in [40] could diagnose this rolling element fault.

2) Results: To evaluate the rate of false diagnosis, we first apply the proposed fast adaptive SASD to the signal measured from the normal bearing, and the extracted results are shown in Fig. 21. In Fig. 21 (a) and (b), there is no fault characteristic frequency in the enlargement of SES, which shows that the proposed method has a relatively low rate of false diagnosis. Besides, from Fig. 21 (c) and (d), the proposed method can successfully separate the discrete frequency components.

Similarly, we further apply the proposed method to analyse the signal measured from the bearing with rolling element fault. As shown in Fig. 22 (a) and (b), we can observe that the impulses from the time domain are obviously enhanced, and BSF as well as its high order frequencies are much clearer than that in Fig. 20 (b). Therefore, we can easily judge the existence of rolling element fault.

D. Computational Time

To further explain the low computational complexity of the proposed method, we list the computational time of four

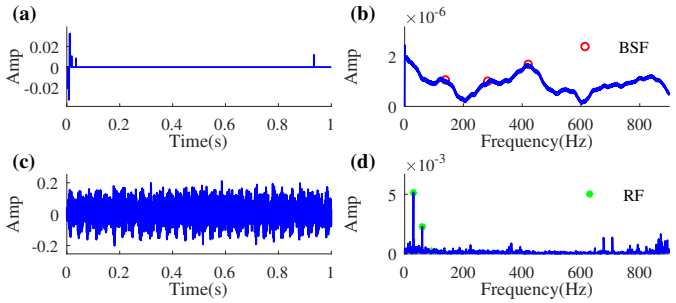


Fig. 21. Results extracted by fast adaptive SASD and its spectra for the normal bearing: (a) time waveform of impulse components, (b) enlargement of SES of impulse components, (c) time waveform of discrete frequency components, and (d) enlargement of SES of discrete frequency components.

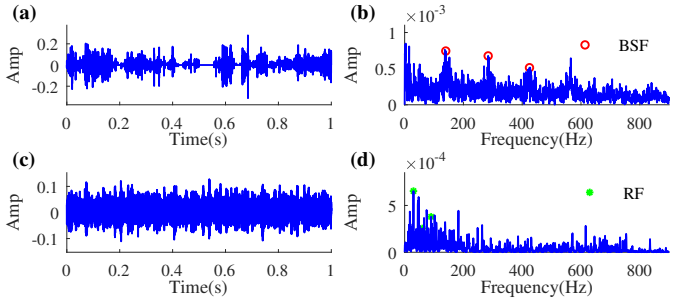


Fig. 22. Results extracted by fast adaptive SASD and its spectra for the fault bearing: (a) time waveform of impulse components, (b) enlargement of SES of impulse components, (c) time waveform of discrete frequency components, and (d) enlargement of SES of discrete frequency components.

TABLE II
COMPARISON OF COMPUTATION TIMES

Method	Fast SASD	Fast adaptive SASD	Fast SK	IWF-MCA
Case1 (s)	0.081	1.087	0.031	83.869
Case2 (s)	0.485	6.231	0.081	457.358

methods in Table II, and the listed time is the average of ten random tests. First of all, the proposed method is much faster than the traditional MCA-based method (IWF-MCA), which indicates the superiority of our method. In addition, fast SK is the fastest method among four methods due to the fact that it does not need the process of iterative optimization, resulting in worse denoising performance.

VI. CONCLUSIONS

In this paper, we propose a fast sparsity-assisted signal decomposition method with non-convex enhancement for bearing fault diagnosis. This method possesses three advantages: without choosing multiple dictionaries for signal representation; low computational complexity; using non-convex enhancement while preserving the overall convexity of the model. Additionaly, we prove the necessary and sufficient condition to guarantee the convexity of our proposed model and derive a fast solver to solve the constructed model. Complete numerical simulations and experimental studies are performed to verify the effectiveness and practicability of our method. Further applications need to be exploited using the proposed method.

REFERENCES

[1] O. Abdeljaber, S. Sassi, O. Avci, S. Kiranyaz, A. A. Ibrahim, and M. Gabbouj, "Fault detection and severity identification of ball bearings by online condition monitoring," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 10, pp. 8136–8147, 2018.

[2] J. Li, R. Huang, G. He, Y. Liao, Z. Wang, and W. Li, "A two-stage transfer adversarial network for intelligent fault diagnosis of rotating machinery with multiple new faults," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 3, pp. 1591–1601, 2021.

[3] B. Ghalamchi, Z. Jia, and M. W. Mueller, "Real-time vibration-based propeller fault diagnosis for multicopters," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 1, pp. 395–405, 2019.

[4] J. Antoni, "Fast computation of the kurtogram for the detection of transient faults," *Mechanical Systems and Signal Processing*, vol. 21, no. 1, pp. 108–124, 2007.

[5] J. Antoni, "The infogram: Entropic evidence of the signature of repetitive transients," *Mechanical Systems and Signal Processing*, vol. 74, pp. 73–94, 2016.

[6] J. Antoni, G. Xin, and N. Hamzaoui, "Fast computation of the spectral correlation," *Mechanical Systems and Signal Processing*, vol. 92, pp. 248–277, 2017.

[7] D. Abboud, S. Baudin, J. Antoni, D. Rémond, M. Eltabach, and O. Sauvage, "The spectral analysis of cyclo-non-stationary signals," *Mechanical Systems and Signal Processing*, vol. 75, pp. 280–300, 2016.

[8] P. Sangeetha B. and H. S., "Rational-dilation wavelet transform based torque estimation from acoustic signals for fault diagnosis in a three-phase induction motor," *IEEE Transactions on industrial informatics*, vol. 15, no. 6, pp. 3492–3501, 2018.

[9] S. Wang, X. Chen, C. Tong, and Z. Zhao, "Matching synchrosqueezing wavelet transform and application to aeroengine vibration monitoring," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 2, pp. 360–372, 2016.

[10] X. Li, W. Zhang, and Q. Ding, "A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning," *Neurocomputing*, vol. 310, pp. 77–95, 2018.

[11] Z. Zhao, T. Li, J. Wu, C. Sun, S. Wang, R. Yan, and X. Chen, "Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study," *ISA transactions*, vol. 107, pp. 224–255, 2020.

[12] Y. Qin, J. Zou, B. Tang, Y. Wang, and H. Chen, "Transient feature extraction by the improved orthogonal matching pursuit and k-svd algorithm with adaptive transient dictionary," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 215–227, 2019.

[13] Z. Zhao, S. Wang, W. Xu, S. Wu, D. Wong, and X. Chen, "Sparsity-assisted fault feature enhancement: Algorithm-aware versus model-aware," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 7004–7014, 2020.

[14] W. He, Y. Ding, Y. Zi, and I. W. Selesnick, "Sparsity-based algorithm for detecting faults in rotating machines," *Mechanical Systems and Signal Processing*, vol. 72, pp. 46–64, 2016.

[15] C. Ding, M. Zhao, J. Lin, J. Jiao, and K. Liang, "Sparsity-based algorithm for condition assessment of rotating machinery using internal encoder data," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 9, pp. 7982–7993, 2019.

[16] B. Wang, Y. Liao, C. Ding, and X. Zhang, "Periodical sparse low-rank matrix estimation algorithm for fault detection of rolling bearings," *ISA transactions*, vol. 101, pp. 366–378, 2020.

[17] Z. Zhao, S. Wu, B. Qiao, S. Wang, and X. Chen, "Enhanced sparse period-group lasso for bearing fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 3, pp. 2143–2153, 2018.

[18] W. Huang, N. Li, I. Selesnick, J. Shi, J. Wang, L. Mao, X. Jiang, and Z. Zhu, "Nonconvex group sparsity signal decomposition via convex optimization for bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 4863–4872, 2019.

[19] Y. Hao, L. Song, B. Ren, H. Wang, and L. Cui, "Step-by-step compound faults diagnosis method for equipment based on majorization-minimization and constraint sca," *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 6, pp. 2477–2487, 2019.

[20] J.-L. Starck, M. Elad, and D. L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," *IEEE transactions on image processing*, vol. 14, no. 10, pp. 1570–1582, 2005.

[21] J. Bobin, J.-L. Starck, J. M. Fadili, Y. Moudden, and D. L. Donoho, "Morphological component analysis: An adaptive thresholding strategy," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2675–2681, 2007.

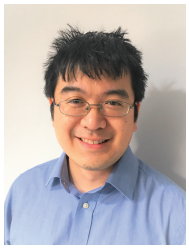
[22] G. Cai, X. Chen, and Z. He, "Sparsity-enabled signal decomposition using tunable q-factor wavelet transform for fault feature extraction of gearbox," *Mechanical Systems and Signal Processing*, vol. 41, no. 1-2, pp. 34–53, 2013.
[23] Z. Du, X. Chen, H. Zhang, and R. Yan, "Sparse feature identification based on union of redundant dictionary for wind turbine gearbox fault diagnosis," *IEEE transactions on industrial electronics*, vol. 62, no. 10, pp. 6594–6605, 2015.
[24] J. Shi and M. Liang, "Intelligent bearing fault signature extraction via iterative oscillatory behavior based signal decomposition (iobsd)," *Expert Systems with Applications*, vol. 45, pp. 40–55, 2016.
[25] Y. Li, X. Liang, M. Xu, and W. Huang, "Early fault feature extraction of rolling bearing based on icd and tunable q-factor wavelet transform," *Mechanical Systems and Signal Processing*, vol. 86, pp. 204–223, 2017.
[26] Y. Qin, "A new family of model-based impulsive wavelets and their sparse representation for rolling bearing fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 3, pp. 2716–2726, 2017.
[27] L. Wang, G. Cai, J. Wang, X. Jiang, and Z. Zhu, "Dual-enhanced sparse decomposition for wind turbine gearbox fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 2, pp. 450–461, 2018.
[28] N. Li, W. Huang, W. Guo, G. Gao, and Z. Zhu, "Multiple enhanced sparse decomposition for gearbox compound fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 3, pp. 770–781, 2019.
[29] S. Wang, I. W. Selesnick, G. Cai, B. Ding, and X. Chen, "Synthesis versus analysis priors via generalized minimax-concave penalty for sparsity-assisted machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 127, pp. 202–233, 2019.
[30] G. Cai, S. Wang, X. Chen, J. Ye, and I. W. Selesnick, "Reweighted generalized minimax-concave sparse regularization and application in machinery fault diagnosis," *ISA transactions*, vol. 105, pp. 320–334, 2020.
[31] W. Huang, Z. Song, C. Zhang, J. Wang, J. Shi, X. Jiang, and Z. Zhu, "Multi-source fidelity sparse representation via convex optimization for gearbox compound fault diagnosis," *Journal of Sound and Vibration*, vol. 496, p. 115879, 2021.
[32] Z. Zhao, S. Wang, C. Sun, R. Yan, and X. Chen, "Sparse multiperiod group lasso for bearing multifault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 2, pp. 419–431, 2019.
[33] P.-Y. Chen and I. W. Selesnick, "Group-sparse signal denoising: non-convex regularization, convex optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 13, pp. 3464–3478, 2014.
[34] C.-H. Zhang *et al.*, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of statistics*, vol. 38, no. 2, pp. 894–942, 2010.
[35] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
[36] N. L. Tsitsas, "On block matrices associated with discrete trigonometric transforms and their use in the theory of wave propagation," *Journal of Computational Mathematics*, pp. 864–878, 2010.
[37] M. Huska, A. Lanza, S. Morigi, and I. Selesnick, "A convex-nonconvex variational method for the additive decomposition of functions on surfaces," *Inverse Problems*, vol. 35, no. 12, p. 124008, 2019.
[38] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
[39] "Case Western Reserve University (CWRU) Bearing Data Center, [Online]," Available: https://csegroups.case.edu/bearingdatacenter/pages/download-data-file/, Accessed 2021, May.
[40] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the case western reserve university data: A benchmark study," *Mechanical Systems and Signal Processing*, vol. 64, pp. 100–131, 2015.

Zhibin Zhao received the B.S. and Ph.D. degrees in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2015 and 2020, respectively. He was also a Visiting Ph.D. Student with the University of Manchester, Manchester, U.K., from 2019 to 2020.

He is currently an assistant professor of mechanical engineering with Xi'an Jiaotong University. His current research is focused on sparse signal processing and machine learning algorithms for machinery health monitoring.

Shibin Wang (M'15) received the B.S. and M.S. degrees in electrical engineering from Soochow University, Suzhou, China, in 2008 and 2011, respectively, and the Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2015.

He then joined the School of Mechanical Engineering, Xi'an Jiaotong University, where he is a Lecturer. In 2017, he was a Visiting Scholar at the Tandon School of Engineering, New York University, NY, USA. His research interests include time-frequency analysis and sparsity-assisted signal processing for machine condition monitoring and fault diagnosis.



David Wong is a lecturer in AI for Healthcare at the University of Manchester. He graduated from the University of Oxford with Meng and DPhil degrees. His research interests are in machine learning and time series analysis, particularly applied to biomedical data.

Wendong Wang received the M.S. degree from Xi'an Jiaotong University, Xi'an, China, in 2018. He is currently working towards the Ph.D. degree in the National Graphene Institute, the University of Manchester, Manchester, UK. His research interests are in signal processing and experimental condensed matter physics.

Ruqiang Yan (M'07-SM'11) received the M.S. degree in precision instrument and machinery from the University of Science and Technology of China, Hefei, China, in 2002, and the Ph.D. degree in mechanical engineering from the University of Massachusetts Amherst, Amherst, MA, USA, in 2007.

(M'07-SM'11) received the M.S. degree in precision instrument and machinery from the University of Science and Technology of China, Hefei, China, in 2002, and the Ph.D. degree in mechanical engineering from the University of Massachusetts Amherst, Amherst, MA, USA, in 2007.

Dr. Yan is an Associate Editor of the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT. He received the New Century Excellent Talents in University Award from the Ministry of Education in China, in 2009. He is a member of the ASME.

Xuefeng Chen (M'12) is a full professor and dean of School of Mechanical Engineering in Xi'an Jiaotong University , P.R.China, where he received his Ph.D. Degree in 2004. He works as the executive director of the Fault Diagnosis Branch in China Mechanical Engineering Society. Besides, he is also a member of ASME and IEEE, and the chair of IEEE the Xian and Chengdu Joint Section Instrumentation and Measurement Society Chapter.

He has authored over 100 SCI publications in areas of composite structure, aero-engine, wind power equipment, etc. He won National Excellent Doctoral Thesis Award in 2007, First Technological Invention Award of Ministry of Education in 2008, Second National Technological Invention Award in 2009, First Provincial Teaching Achievement Award in 2013, First Technological Invention Award of Ministry of Education in 2015, and he was awarded as Science & Technology Award for Chinese Youth in 2013. Additionally, he hosted a National Key 973 Research Program of China as principal scientist in 2015.