Sparsity-assisted Fault Feature Enhancement: Algorithm-aware versus Model-aware

Zhibin Zhao, Shibin Wang, *Member, IEEE,* Weixin Xu, Shuming Wu, *Student Member, IEEE,* David Wong, and Xuefeng Chen, *Senior Member, IEEE*

***Abstract*—Vibration signal analysis has become one of the important methods for machinery fault diagnosis. Extraction of weak fault features from vibration signals with heavy background noise remains a challenging problem. In this paper, we first introduce the idea of algorithm-aware sparsity-assisted methods for fault feature enhancement, which extends model-aware sparsity-assisted fault diagnosis and allows more flexible and convenient algorithm design. In the framework of algorithm-aware methods, we define the generalized structured shrinkage operators and construct the generalized structured shrinkage algorithm (GSSA) to overcome the disadvantages of l_1-norm regularization based fault feature enhancement methods. We then perform a series of simulation studies and two experimental cases to verify the effectiveness of the proposed method. Additionally, comparisons with model-aware methods, including basis pursuit denoising and windowed-group-lasso, and fast kurtogram further verify the advantages of GSSA for weak fault feature enhancement.**

***Index Terms*—Algorithm-aware method, fault diagnosis, generalized structured shrinkage operators, social sparsity**

I. Introduction

CONDITION-based maintenance (CBM) of rotating machinery has become an essential part of systems in a wide range of mechanical systems, such as wind turbines, helicopters, and aero-engines. Some key components of rotating machinery, including bearings and gearboxes, often operate in poor conditions and are likely to generate mechanical faults to make system failure. Therefore, it is important to detect these faults of key components quickly and accurately. In CBM, vibration signal analysis is a widely used and validated etchnique for fault diagnosis. Traditional vibration signal analysis methods, such as time domain statistics and classical spectrum analysis techniques, often fail to diagnose weak faults as fault features are often submerged by strong background noise and harmonic interference. Other signal processing algorithms such as spectral kurtosis (SK) [1], cyclostationary descriptors [2], minimum entropy deconvolution [2], time-frequency analysis [3], empirical mode decomposition [4], and stochastic resonance [5] have previously been proposed to address this issue.

Over the past two decades, wavelet denoising has been proven to be a powerful tool for signal processing. The main steps of wavelet denoising consist of transforming the signal into the wavelet domain, shrinking the wavelet coefficients [6], and calculating the inverse wavelet transform. However, because these strategies treat each coefficient independently, they usually fail to model the similarity between coefficients. Therefore, Cai et al. [7] incorporated information on neighboring coefficients into wavelet denoising and proposed the neighboring coefficients denoising (NCD) for image processing and further, they proposed a data-driven block thresholding approach to incorporate more coefficients. Sendur et al. [8] proposed bivariate shrinkage functions for wavelet denoising. For fault diagnosis, Chen et al. [9] and He et al. [10] used NCD to improve the performance of fault feature extraction. Sun et al. [11] and Chen et al. [12] used the data-driven block threshold for condition monitoring. Hussein et al. [13] considered the disadvantages of conventional thresholding functions and proposed the histogram-based threshold estimation method for signal denoising. In addition, Yu et al. [14] used the sparse coding shrinkage to improve the performance of weak fault feature extraction. Although researchers have tried to add more structural information into thresholding strategies there has been limited improvement in wavelet denoising methods.

Recently, sparsity-assisted signal processing methods are becoming increasingly popular in machinery fault diagnosis [15]–[18]. The core idea of sparsity-assisted fault diagnosis mainly consists of using sparse priors of fault features under some dictionaries (such as wavelet transform) to establish a sparse model and solving the established model by an optimization algorithm. We call these model-aware methods as they are mainly concerned with modeling the sparse prior by designing the corresponding regularization and establishing an explicit mathematical model. Among these methods, l_1-norm regularization methods, such as basis pursuit denoising (BPD) [19], are becoming increasingly popular and have been successfully used in machinery condition monitoring. Yang et al. [20] proposed a novel sparse time-frequency representation method based on l_1-norm regularization (BPD) and applied it to incipient fault diagnosis of wind turbine drive train. However, BPD has three serious problems for machinery fault diagnosis. First, BPD often underestimates the energy

Z. Zhao, S. Wang, W. Xu, Shu. W and X. Chen are with the State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China. E-mail: (zhibinzhao1993@gmail.com; wangshibin2008@gmail.com; xuwwwux@163.com; wushuming0309@gmail.com; chenxf@mail.xjtu.edu.cn)

D. Wong is with the Centre for Health Informatics, Univeristy of Manchester, Manchester, United Kingdom. E-mail: (david.wong@manchester.ac.uk)

This work was supported in part by the National Key Basic Research Program of China under Grant 2015CB057400, by the Natural Science Foundation of China under Grand 51605366 and 51705398, by the China Postdoctoral Science Foundation under Grant 2016M590937 and 2017T100740, and by the Fundamental Research Funds for the Central Universities.

S. Wang is the corresponding author.

of fault feature, leading to incorrect diagnosis. To enhance the energy of fault features, Zhang et al. [21], Huang et al. [22], Wang et al. [23], and Zhao et al. [24] proposed new penalties based on model-aware methods for rotating machinery fault diagnosis. Second, BPD assumes that each coefficient is independent and identically distributed (IID), but wavelet coefficients of fault features often appear in groups, especially under redundant transforms. To incorporate the relationship between coefficients, He et al. [25] and Sun et al. [26] used the structured information (They still established an explicit model to sovle the problem) to detect faults in rotating machines, including group sparsity and persistent sparsity. Finally, BPD does not consider the multiscale property of wavelet transform at all. However, the fault signal often shows the nonstationary property and the multiscale property is very useful for nonstationary signal processing. For the sake of using the multiscale property, He et al. [27] used the periodic group-sparse model in each wavelet transform layer and blended in the multiscale period sequences. In conclusion, previous published papers did not consider all the disadvantages of l_1-norm regularization methods together. In addition, an explicit model is critical to model-aware methods which means that we need to find penalties corresponding to sparse priors to form an explicit optimization problem. However, designing the penalties promoting properties simultaneously to overcome the disadvantages of l_1-norm regularization methods is difficult. Inversely, designing the thresholding functions from the aspect of wavelet denoising is simpler and has shown good performance in the previous studies.

In this paper, inspired by thresholding strategies from wavelet denoising, we introduce the idea of algorithm-aware methods for weak feature enhancement, which extends the traditional modeling framework (model-aware sparsity-assisted methods). To be more specific, in this paper, we propose generalized structured shrinkage operators and replace the operator in the proximal gradient descent algorithm with these proposed operators to form the iterative optimization algorithm called generalized structured shrinkage algorithm (GSSA) for weak fault feature enhancement. Finally, we perform a series of numerical simulations and experimental cases to further compare the performance of model-aware and algorithm-aware methods.

The main contributions of this paper are summarized in two categories:

1) We provide a new idea called the algorithm-aware method which breakthroughes the model-aware sparsity-assisted fault diagnosis and allows to design algorithms more flexible and convenient, and the algorithm-aware method uses existing algorithmic flows of sparsity-assisted methods through focusing on designing thresholding functions to develop new methods.
2) To introduce the structured information, we bring in the structured shrinkage operator. Using the idea of the algorithm-aware method, we further improve the structured shrinkage operator through considering unbiased and multiscale properties and construct the general expression of the structured shrinkage operator to copy with shortcomings of BPD (representing the model-aware method).

The remaining parts of this paper are organized as follows. In Section II, we review the sparsity-assisited fault diagnosis based on l_1-norm regularization. Section III explains the main idea of algorithm-aware methods, proposes generalized structured shrinkage operators, and forms GSSA for weak fault feature enhancement. In Section IV, we analyze the parameter selection in depth and provide the performance verification of the proposed algorithm. In Section V, we further verify the effectiveness and robustness of the proposed algorithm through experimental cases. Section VI states the conclusion of this paper.

II. Sparsity-assisited fault diagnosis based on l_1-norm regularization

In this section, we briefly review the main idea of sparsity-assisited fault diagnosis based on l_1-norm regularization.

We assume that the measured vibration signal is $\mathbf{y} \in \mathbb{R}^N$ with a signal of interests, $\mathbf{s} \in \mathbb{R}^N$, corrupted by heavy background noise $\mathbf{n} \in \mathbb{R}^N$. Thus, the observed vibration signal can be modeled as:

$$\mathbf{y} = \mathbf{s} + \mathbf{n} = \mathbf{D}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{N\times M}$ denotes the matrix of the dictionary, and the representation coefficients of $\mathbf{s}$ in $\mathbf{D}$ are represented by $\mathbf{x} \in \mathbb{R}^M$.

A general relaxation and unconstrained formulation to estimate $\mathbf{x}$ from $\mathbf{y}$ can be written in the Lagrangian form:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda P(\mathbf{x}) \tag{2}$$

where $P(\cdot)$ denotes the penalty (regularization) which needs to be carefully designed according to sparse priors and $\lambda > 0$ is a trade-off regularization parameter. If the penalty $P(\cdot)$ is convex, standard algorithms such as the toolbox CVX [28] and the proximal gradient descent (PGD) [29] can be used to solve this problem. The most popular convex penalty for sparsity-assisited fault diagnosis is $\|\mathbf{x}\|_1$ where $\|\mathbf{x}\|_1 = \sum_n |x_n|$ is the l_1-norm of $\mathbf{x} \in \mathbb{R}^M$, and the optimization problem (2) is reduced to BPD [19]:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1. \tag{3}$$

As the algorithm proposed in this paper is based on PGD, we briefly describe PGD here. Recently research [30] has shown that even if the penalty $P(\mathbf{x})$ is not convex, PGD still displays good convergence properties. We first define the proximity operator for a proper and lower semicontinuous function $P(\cdot)$ as

$$\text{prox}_{P,\beta}(z) = \arg\min_{x} \frac{1}{2\beta}(z-x)^2 + P(x) \tag{4}$$

where $\beta > 0$ is the penalty parameter and z and x are scalar values. Proximity operators $\text{prox}_{P,\beta}(z)$ for $P(x)$ do not always have a closed-form solution and they are sometimes very difficult to calculate. $\text{prox}_{P,\beta}(\cdot)$ is element-by-element for any vector input and are separable such that $\text{prox}_{P,\beta}(\mathbf{z}) = [\text{prox}_{P,\beta}(z_1), \ldots, \text{prox}_{P,\beta}(z_M)]^{\mathrm{T}}$. One well-known operator is the soft-thresholding operator defined as:

$$\text{prox}_{\|\cdot\|_1,\beta}(z) = \text{sign}(z)\max\left(|z| - \beta, 0\right). \tag{5}$$

This operator is used in the Iterative Shrinkage-Thresholding Algorithm (ISTA) for the optimization problem (3) [31] and is a special case of PGD.

Algorithm 1 describes how PGD may be applied to solve the optimization problem in (2). μ is the updating step, $\|\mathbf{D}\|_2$ represents the square root of the maximum eigenvalue of the matrix $\mathbf{D}^{\mathrm{T}}\mathbf{D}$, and $Iter$ is the number of iterations.

Algorithm 1 : PGD

1: **Initialization:** $\mathbf{x}^{(0)}$, $\mu < \|\mathbf{D}\|_2^{-2}$, $Iter$
2: **Procedure:**
3: **for** $i = 0$ to $Iter$ **do**
4: $\quad \mathbf{z}^{(i)} = \mathbf{x}^{(i)} - \mu\mathbf{D}^{\mathrm{T}}(\mathbf{D}\mathbf{x}^{(i)} - \mathbf{y})$
5: $\quad \mathbf{x}^{(i+1)} = \mathrm{prox}_{P,\lambda\mu}(\mathbf{z}^{(i)})$ (for example, (5) for ISTA)
6: **end for**
7: **return** $\mathbf{x}^{(i+1)}$

The optimization problem described in equation (2) requires $P(\mathbf{x})$ to be explicitly defined (e.g. l_1-norm, group sparse [32] and periodic group sparse [25]), to generate the sparsity-assisted model. It is known as a model-aware method as one needs to establish the model before deducing a corresponding algorithm. It is worth mentioning that compressed sensing is not directly related to the model-aware method and compressed sensing is a more general concept which is not limited to the model-aware method.

Since the paper is not concerned with the choice of $\mathbf{D}$, we simply use a tunable $Q-$Factor wavelet transform (TQWT) [33], with parameters adopted to match the oscillatory waveform of fault features. Thus, the Q-factor, the redundant factor r and the decomposition levels J are set to be $2 \leq Q \leq 5$, $2 \leq r \leq 5$ and $J = 10$ respectively. Further details on parameter selection methods can refer to [34].

TQWT is not a "uniform" tight frame, which means that rows of $\mathbf{D}$ do not have the same energy. Therefore, we need to treat the regularization parameter λ differently at different levels, and we set the parameter λ as $\lambda_j = c\|\varphi_j\|_2, (j = 1, 2, \ldots, J+1)$ where φ_j with $j = 1, 2, \ldots, J$ is the wavelet function at the level j, φ_{J+1} is the scaling function at the level $J+1$, and c is a constant value.

III. Algorithm-aware Method

A. Core idea

As model-aware methods generate the corresponding algorithm through a certain optimization model and a specific penalty, therefore, it may not copy with complex priors, such as the multi-scale property [35].

The core idea of our algorithm-aware method is to use existing algorithmic flows of sparsity-assisted methods and to replace the thresholding functions in existing algorithmic flows to develop new methods. Fig. 1 shows the idea in mode detail. (k_1, k_2, and k_3 are the central coefficients in the neighborhoods $\mathcal{N}(k_1)$, $\mathcal{N}(k_2)$, and $\mathcal{N}(k_3)$ respectively, where $\mathcal{N}(\cdot)$ represents one type of the window, such as the Gaussian window) We summarize PGD in the form of a flowchart where the proximal projection plays the same role as the thresholding strategy of the wavelet denoising.

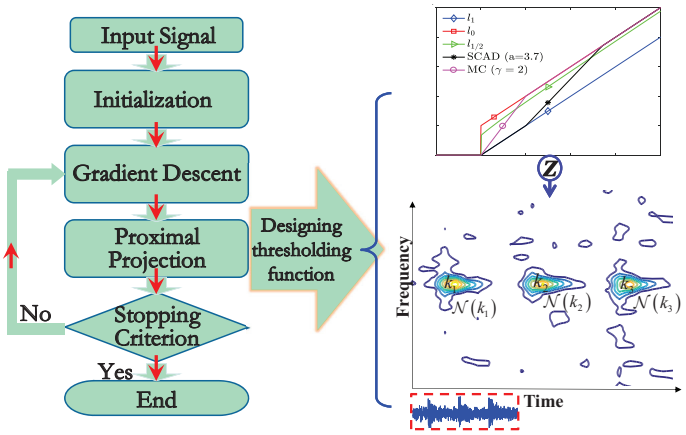

Fig. 1. Explanation of the algorithm-aware method and social sparsity structure of the time-frequency domain in the right bottom.

Here, we replace the original proximity operator used in PGD, based on thresholding strategies used for wavelet denoising. Because we directly design the algorithm from PGD, it is very difficult to find a model corresponding to the proposed algorithm like (3). This is the reason why this new idea is called the algorithm-aware method. Since we avoid to design the specific regularization term and the optimization model, the idea of the algorithm-aware method allows to design algorithms more flexible and convenient.

B. Designing thresholding functions

The l_1-norm penalty and its soft-thresholding operator assume that each coefficient is independent and identically distributed, but wavelet coefficients of fault features often appear in groups [12]. To introduce the function of the block (or group sparse), Kowalski et al. [36] introduced the concept of "social sparsity" to shrink the coefficients by considering the weight of the coefficient's neighborhood. Its performance has been verified by different applications, such as audio inpainting [37] and audio denoising [32].

To explain the concept of social sparsity and define its induced shrinkage operator, we need to define a neighborhood. For an index k, we denote the weighted neighborhood as $\mathcal{N}(k) = \{k' \in \mathcal{I} : w_{k'} \geq 0\}$, and the weights $w_{k'}$ satisfy $\sum_{k' \in \mathcal{N}(k)} w_{k'}^2 = 1$. A visual interpretation of this is shown in the bottom-right of Fig. 1. Detailed choice of $\mathcal{N}(k)$ will be discussed in the parameter selection.

According to the defined neighborhood and the proximal operator of the l_1-norm regularization, the windowed-group-lasso (WGL) structured shrinkage operator, $\mathcal{S}(\cdot)$, is defined as [36]:

$$\mathcal{S}_{\mathrm{WGL},\lambda\beta}(z_k) = z_k \max\left(1 - \frac{\lambda\beta}{\sqrt{\sum_{k' \in \mathcal{N}(k)} w_{k'}|z_{k'}|^2}}, 0\right). \tag{6}$$

By taking weights of the neighborhood into consideration, a single large coefficient surrounded by small noise can be deleted, whereas a small coefficient containing feature information in the middle of large ones can be preserved. This

TABLE I
PROXIMITY OPERATORS

Penalty name	Proximity operator: $\text{prox}_{P,\beta}(z)$
l_0 [38]	$zI\left(\lvert z\rvert > \sqrt{2\beta}\right)$
l_1 [19]	$\text{sign}(z)\max\left(\lvert z\rvert - \beta, 0\right)$
MCP [40]	$\begin{cases} 0, & \lvert z\rvert < \beta \\ \frac{\text{sign}(z)(\lvert z\rvert-\beta)}{1-1/\gamma}, & \beta \le \lvert z\rvert < \gamma\beta \\ z, & \lvert z\rvert \ge \gamma\beta \end{cases}$

characteristic is good for enhancing the weak fault feature without also enhancing the interference.

Although the structured shrinkage operator (6) blends in the information of groups through the social sparsity, the induced shrinkage operator is based on soft-thresholding. It still suffers from underestimating the energy of feature information. In order to preserve the amplitude of feature information, we extend the structured shrinkage operator (6) to a generalized structured shrinkage operator through embedding the properties of other proximal operators. To express the operators concisely, we first define the structured threshold value as

$$T_k = \frac{\lambda\beta|z_k|}{\sqrt{\sum_{k'\in\mathcal{N}(k)} w_{k'}|z_{k'}|^2}} \tag{7}$$

where T_k is the structured threshold value of k-th coefficient. Therefore, the generalized structured shrinkage operators (thresholding functions) are defined as

$$\mathcal{S}_{P,T_k}(z_k) = \text{prox}_{P,T_k}(z_k) \tag{8}$$

where $\text{prox}(\cdot)$ is defined in Table I. It can be noted that only the l_1-norm penalty is convex. Among these penalties, l_0 is a natural selection for inducing sparsity, and its corresponding proximal operator is called the hard-thresholding operator introduced by Donoho [38]. The hard-thresholding operator leads to no bias on the large inputs, but its discontinuity makes it unstable during the optimization procedure [39].

The minimax concave penalty (MCP) [40] is widely used in statistical learning and feature selection, and it satisfies three principles proposed in [41]: unbiasedness, sparsity, and continuity. For MCP ($\gamma > 1$), when the parameter γ tends to 1, the induced operator tends to the hard-thresholding operator. Otherwise, when the parameter γ tends to positive infinity, the induced operator tends to the soft-thresholding operator. Therefore, in the remainder of this paper, we use the MCP and its induced proximal operator with the parameter $\gamma = 2$ recommended in [40] (Researchers can also use smoothly clipped absolute deviation (SCAD) [41] or $l_{1/2}$ [42], as shown in the right top of Fig. 1).

C. Generalized structured shrinkage algorithm

After designing the thresholding functions, we further consider the multiscale property of TQWT by redefining the thresholding functions and determine the selection of the regularization parameter c through K-sparsity [23] to construct GSSA.

In the wavelet domain (in this paper, TQWT), the total points of each level are different and sometimes appear radix 2. Thus, we need to embed this multiscale property into the generalized structured shrinkage operators. We first define the basic length l of the window $\mathcal{N}$ at the level $J+1$. If the length of the j-th level (denoted as M_j) is twice the length of the $j+1$-th level (denoted as M_{j+1}), the basic length l of the window $\mathcal{N}$ is multiplied by 2, whereas remaining unchanged. In order to distinguish the windows at different wavelet levels, we redefine the structured threshold value and the final generalized structured shrinkage operators as

$$\mathbf{T}_{j,\cdot} = \frac{\lambda_j\beta|\mathbf{z}_{j,\cdot}|}{\sqrt{\sum_{k'\in\mathcal{N}_j(\cdot)} w_{k'}|z_{j,k'}|^2}} \tag{9}$$

$$\mathcal{S}_{P,\mathbf{T}_{j,\cdot}}(\mathbf{z}_{j,\cdot}) = \text{prox}_{P,\mathbf{T}_{j,\cdot}}(\mathbf{z}_{j,\cdot}) \tag{10}$$

where j represents the $j-$th level in the wavelet domain, $\mathbf{T}_{j,\cdot}$ represents the vector $\left[T_{j,1},\ldots,T_{j,M_j}\right]^{\mathrm{T}}$, and $\mathbf{z}_{j,\cdot}$ is the wavelet coefficients at the $j-$th level. At different levels, the parameter λ_j is calculated by $c\|\varphi_j\|_2$.

In addition, we use the K-sparsity (K) strategy to determine the most important parameter c. The core idea of the K-sparsity strategy is that in each iteration, we only keep K largest coefficients and set other coefficients to be zero. This strategy is more robust to noise, and we suggest to keep $0.5\% - 5\%$ of total number of coefficients in the sparsity domain by our proposed method. The wavelet coefficients should be normalized before setting c as the K-th largest coefficient. Therefore, we define the parameter c as follows:

$$c = \max_{[\mathrm{K}]}\left(\frac{\mathbf{z}_{1,\cdot}}{\|\varphi_1\|_2},\ldots,\frac{\mathbf{z}_{J+1,\cdot}}{\|\varphi_{J+1}\|_2}\right) \tag{11}$$

where $\max_{[\mathrm{K}]}$ means to extract the K-th largest coefficient in the sequence.

After defining the structured threshold value and the generalized structured shrinkage operator at different levels, we can easily construct GSSA according to the idea of the algorithm-aware method with the multiscale window and the given generalized structured shrinkage operator. Algorithm 2 shows the details of the proposed algorithm, and the computational complexity is similar to PGD, because the calculation of the generalized structured shrinkage operators costs a little. Additionally, the convergence analysis of Algorithm 2 is very hard and is still an unsolved problem. However, through numerical simulations and experimental cases, Algorithm 2 shows good performance according to the results.

IV. SIMULATION STUDY

In this section, we perform a series of numerical simulations to discuss the parameter selection and verify the effectiveness of the proposed method. We first describe the model of the simulated signal according to [23] as follows:

$$\mathbf{y} = \mathbf{s} + \mathbf{n} = \sum\nolimits_k a_k\mathbf{h}\left(t - kT - \tau_k - \tau_0\right) + \mathbf{n} \tag{12}$$

where the fault characteristic period $T = 0.01$ with the random slip τ_k following the uniform distribution $U(-0.001, 0.001)$,

Algorithm 2 : GSSA

Initialization: $\mathbf{x}^{(0)}$, step size: $\mu < \|\mathbf{D}\|_2^{-2}$, number of iterations: $Iter$, TQWT parameters: Q, r, and J, K-sparsity parameter: K, length of windows: l
Procedure:
for $i = 0$ to $Iter$ **do**
  $\mathbf{z}^{(i)} = \mathbf{x}^{(i)} - \mu \mathbf{D}^{\mathrm{T}}(\mathbf{D}\mathbf{x}^{(i)} - \mathbf{y})$
  $c = \max_{[\mathrm{K}]}(\frac{\mathbf{z}_{1,\cdot}^{(i)}}{\|\varphi_1\|_2}, \ldots, \frac{\mathbf{z}_{J+1,\cdot}^{(i)}}{\|\varphi_{J+1}\|_2})$
  for $j = J+1$ to 1 **do**
    $\lambda_j = c\|\varphi_j\|_2$
    $\mathbf{T}_{j,\cdot} = \frac{\lambda_j \mu |\mathbf{z}_{j,\cdot}|}{\sqrt{\sum_{k' \in \mathcal{N}_j(\cdot)} w_{k'} |z_{j,k'}|^2}}$
    $\mathbf{x}_j^{(i+1)} = \mathcal{S}_{P,\mathbf{T}_{j,\cdot}}(\mathbf{z}_{j,\cdot}^{(i)})$
  end for
end for
return $\mathbf{x}^{(i+1)}$

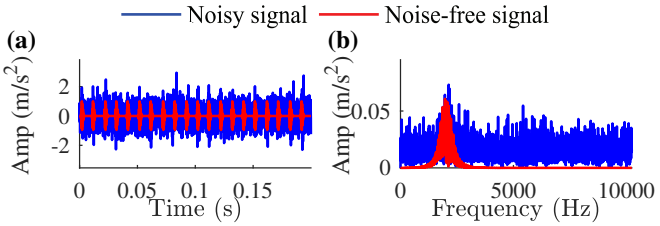

Fig. 2. (a) The simulation signal with the noise intensity equal to 0.6, and (b) its frequency spectrum.

TABLE II
EXPLANATION OF PARAMETERS

Parameters	Explanations
Q	the $Q-$factor of TQWT
r	the redundant factor
J	the decomposition levels
l	the basic length of the window
$\mathcal{N}$	the type of the window

τ_0 is the initial phase equal to 0.002, and the bilateral impulses $\mathbf{h}(t)$ is defined as

$$\mathbf{h}(t) = \begin{cases} e^{\frac{-\zeta_L}{\sqrt{1-\zeta_L^2}}(2\pi f_1 t)^2} \cos(2\pi f_1 t), & t < 0 \\ e^{\frac{-\zeta_R}{\sqrt{1-\zeta_R^2}}(2\pi f_1 t)^2} \cos(2\pi f_1 t), & t \geq 0 \end{cases} \quad (13)$$

with $\zeta_L = 0.02$, $\zeta_R = 0.005$, and $f_1 = 2000$ Hz. The noise $\mathbf{n}$ is generated by Gaussian distribution with zero mean and σ^2 variance. The sampling frequency and the length of the signal are set 20480 Hz and 4096 respectively. In addition, as shown in Fig. 2, the periodic impulses are submerged by the heavy background noise with the noise intensity equal to 0.6. In the discussion below, we use the index RMSE to evaluate the algorithms, and RMSE is defined as RMSE $= \sqrt{\frac{1}{N}||\mathbf{s} - \hat{\mathbf{s}}||_2^2}$, where $\hat{\mathbf{s}}$ is the extracted feature by the algorithm.

A. Parameter selection

We do not study the parameter selection of TQWT in this paper, interested researchers can further refer to [34]. Thus, the parameters required to be identified include the basic length l and the type of the window $\mathcal{N}$ shown in Table II.

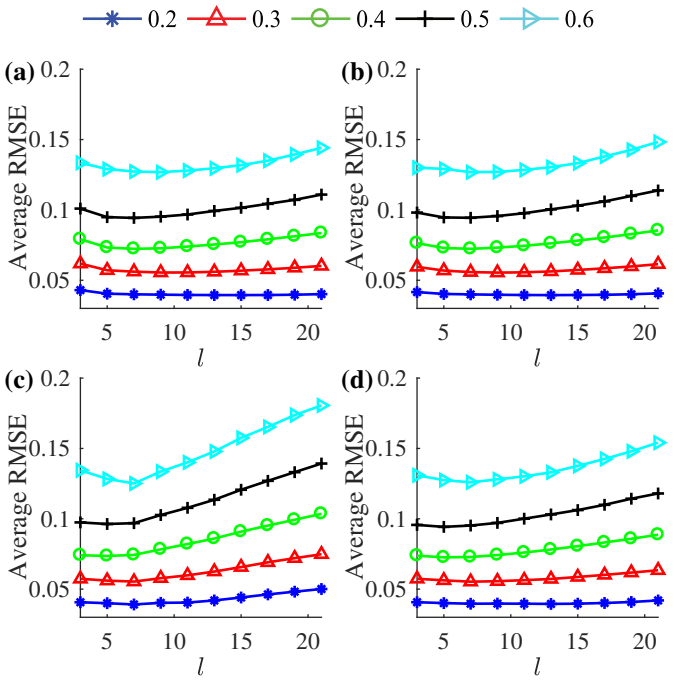

Fig. 3. Relationships between average RMSE and the basic length l under different windows: (a) Gaussian window, (b) Hamming window, (c) Triangular window, and (d) Rectangular window.

In order to discuss the selection of the basic length l, we use GSSA with the K-sparsity strategy to analyze the simulated signal. The TQWT parameters are set as $Q = 2$, $r = 5$, and $J = 10$. We test on four different windows (Gaussian window, Hamming window, Triangular window, and Rectangular window) and five different noise levels (σ from 0.2 to 0.6 with the increment equal to 0.1) to study the relationships between the basic length l and the average RMSE. In addition, we vary the K-sparsity parameter K (from 10 to 500 with the increment equal to 5) to achieve the minimal RMSE. To avoid the randomness, we perform 100 realizations under different random seeds to achieve the average RMSE. The results are shown in Fig. 3, and it is observed that when the noise level is relatively low (i.e., $\sigma = 0.2$), the average RMSE is almost the same under the different basic lengths l. However, as the noise increases, the average RMSE decreases first and then increases, especially in the strong noise interference, and this phenomenon helps us determine the range of the basic length l (i.e., $l = 5$, 7, or 9). In term of the calculation complexity (the larger the basic length, the greater the computational time), we recommend $l = 5$. Besides, according to the comparisons among four windows, we can conclude that the algorithm is almost robust to the type of the window. Therefore, in the analysis below, we simply use the Gaussian window.

In addition, we also perform 100 realizations for each method to compare the averaging computional time, and the results are listed in Table III. It can be observed that the computional times of GSSA, BPD and WGL are almost same since they are all based on the same optimization algorithm, and NCD is much faster than other three methods since it does not need any iteration and is just traditional wavelet denoising.

TABLE III
COMPUTATIONAL TIMES OF DIFFERENT METHODS

Method	Averaging computional time (s)
GSSA	0.4720
BPD	0.3203
NCD	0.0674
WGL	0.4489

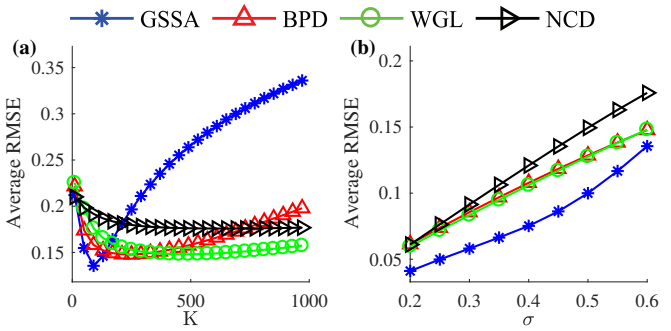

Fig. 4. (a) The relationship between average RMSE values and the K-sparsity values when $\sigma = 0.6$, and (b) average RMSE values of different methods under different noise levels.

B. Performance verification 1

In the performance verification, for comparison targets, we analyze the simulation signals through BPD [19], [20], WGL [36] (These two methods represent the model-aware methods), and neighboring coefficients denoising (NCD) representing the traditional wavelet denosing [10], [13]. Except for the k-sparsity parameter, other parameters are set as $Q = 2$, $r = 5$, and $J = 10$. On the one hand, we draw the relationships between the average RMSE and the K-sparsity parameter of four different methods under the noise intensity equal to 0.6. As shown in Fig. 4 (a), the best K is 90 for GSSA, 240 for BPD, 480 for WGL, and 630 for NCD. Moreover, the average RMSE of GSSA is obviously smaller than other methods. From the comparison results, we can conclude that the proposed algorithm can use much fewer coefficients to recover the more accurate features. On the other hand, we also evaluate the performance of four methods under different noise levels and vary the K-sparsity parameter K from 10 to 1000 with the increment equal to 10 (for NCD, we vary k from 10 to 11000 due to the fact that it does not promote the sparsity). We also perform 100 realizations for all four methods under different random seeds. Fig. 4 (b) compares the average RMSE of four different methods, and it is observed that GSSA achieves the best performance. Due to the disadvantage of not being able to promote the sparsity, NCD gets the worst performance. In addition, because BPD and WGL are both based on the l_1-norm regularization, their performance is almost the same, especially, when the noise intensity is relatively high.

For the purpose of further verifying the performance of the proposed algorithm, visualization results with the noise intensity $\sigma = 0.6$ are shown in Fig. 5. The parameter setting is the same as the performance verification above. From the visualization shown in Fig. 5, we can draw two conclusions: GSSA can maintain the amplitude of each impulse in the better way; From comparisons with other methods, GSSA achieves better denoising results and contains scarcely any pseudo-impulse in the extracted components.

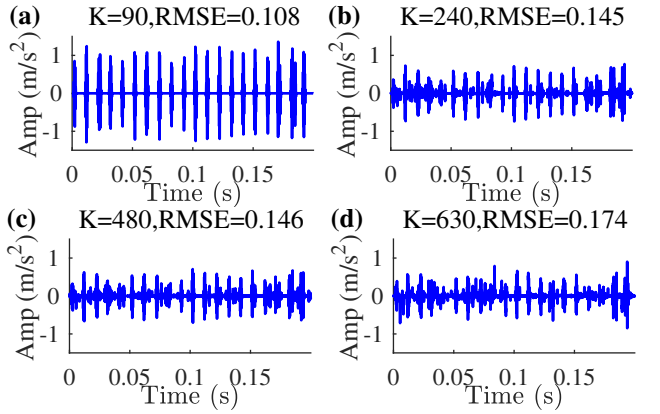

Fig. 5. The extracted results of different methods: (a) the result extracted by GSSA, (b) the result extracted by BPD, (c) the result extracted by WGL, and (d) the result extracted by NCD.

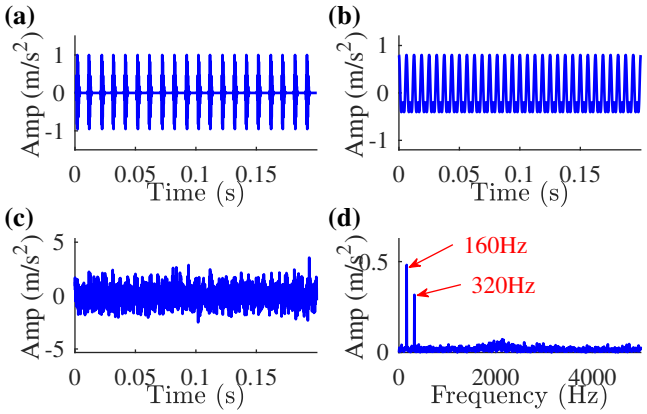

Fig. 6. The simulation signal corrupted by noise interference $\sigma = 0.6$ and discrete frequency interference: (a)the pure impulse signal, (b) discrete frequency interference, (c) the mixed signal, and (d) the frequency spectrum of the mixed signal.

C. Performance verification 2

In order to further testify the robustness of the proposed method, another numerical simulation which not only contains heavy background noise, but also suffers from discrete frequency interference is performed. We first define the discrete frequency interference $\mathbf{d}(t)$ as:

$$\mathbf{d}(t) = A_1 \cos(2\pi f_2 t) + A_2 \cos(4\pi f_2 t) \tag{14}$$

where $A_1 = 0.5$ and $A_2 = 0.3$ are the amplitudes of the interference, and $f_2 = 160$ Hz is the basic frequency of the interference. The simulation signal is shown in Fig. 6, and it is observed from the frequency spectrum (shown in Fig. 6 (d)) that the resonance frequency band is submerged by the discrete frequency components totally. The simulation signal is still handled by GSSA, BPD, WGL and NCD. Meanwhile, the parameter setting is the same as the previous discussion except that the K-sparsity parameter K is varied from 10 to 1000 (with the increment equal to 10) to find the best parameter for each method. The extracted results and their

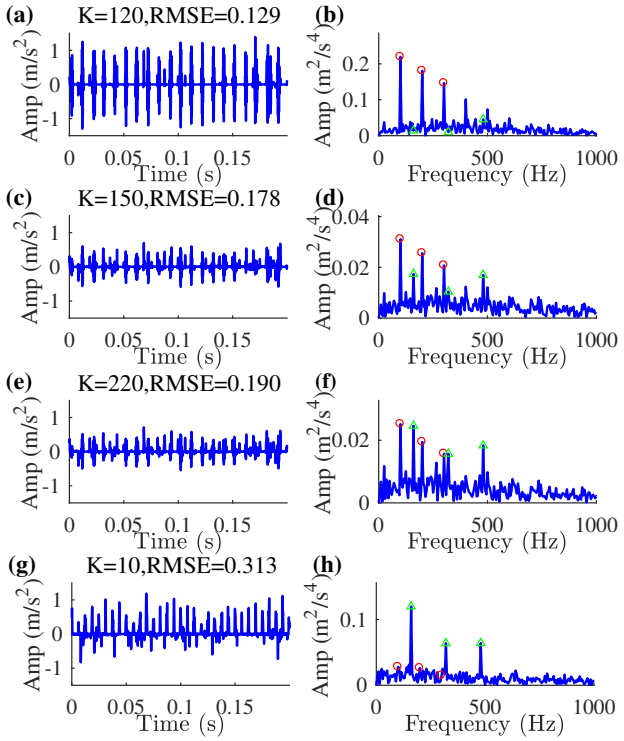

Fig. 7. The extracted results of different methods: (a) the fault feature extracted by GSSA, (b) SES of GSSA, (c) the fault feature extracted by BPD, (d) SES of BPD, (e) the fault feature extracted by WGL, (f) SES of WGL, (g) the fault feature extracted by NCD, and (h) SES of NCD.

square envelope spectra (SES) are displayed in Fig. 7 (The green triangle represents the harmonic interference, and the red circle represents the fault characteristic frequency). In conclusion, GSSA achieves the best performance either RMSE (according to the ability of maintaining the amplitude) or harmonic interference suppression since its SES is very clean and fault characteristic frequencies are extremely obvious. Comparisons with the proposed method, BDP and WGL are all affected by the discrete frequency interference exhibited in Fig. 7 (c)-(f). Meanwhile, there are lots of pseudo impulses in time domains which affect the judgment of the period, and the energy of fault characteristic frequencies is only a tenth of GSSA. Besides, as shown in Fig. 7 (g) and (h), NCD fails to extract the periodic impulses completely.

V. Experimental Verification

A. Case 1: Experiment description

First, we analyze the measured vibration signal collected by the NSF I/UCR Center for Intelligent Maintenance Systems [43]. The test rig which installed four bearing on the shaft was shown in Fig. 8. Besides, all the bearings were forced lubricated and were loaded with 6000 lb radial load by the sping mechanism. Eight accelerometers (PCB 353B33 High Sensitivity Quartz ICP) were installed on four bearings and each bearing contained two accelerometers on the horizontal X and vertical Y shown in Fig. 8. The sampling frequency and the rotating speed are equal to 20.48 kHz and 2000 rpm, respectively. The acquisition system collected one second data every 10 minutes by a NI DAQ Card 6062E. At the end of

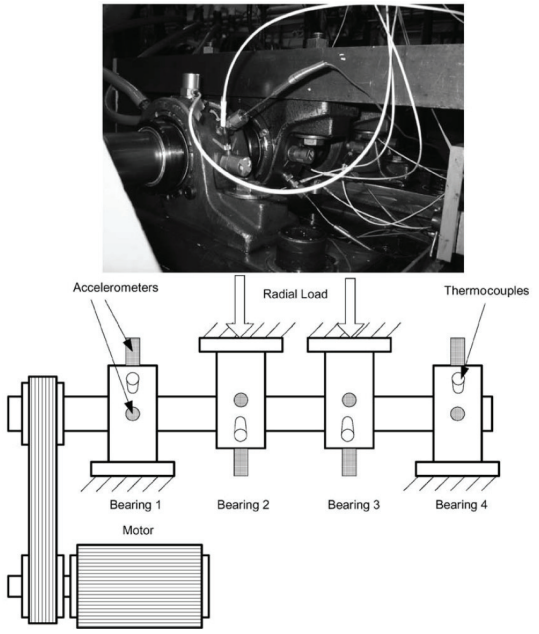

Fig. 8. Test rig and its structural sketch with sensor arrangement [43].

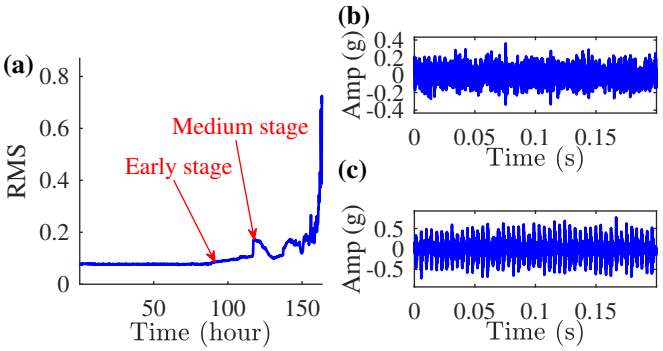

Fig. 9. (a) RMS values of the measured vibration signals in the run-to-failure data, (b) the measured signal in the early stage of the fault, and (c) the measured signal in the medium stage of the fault.

the test, a fault occurred at the outer race of the bearing, and we analyze the measured signal from the accelerometer 1 which was just mounted on the fault bearing and the horizontal X. According to Ref. [43], the fundamental outer race fault frequency of the test bearing (BPFO) was approximately equal to 236.4 Hz.

The root-mean-square (RMS) values of the run-to-failure data and two segments of the measured signal in the early and medium stage are shown in Fig. 9. We can observe that in the early stage of the fault (in file No.550 corresponding to 91.7 hours after the experiment started), the periodic impulses are submerged by the strong background noise. In the medium stage of the fault (in file No.703 corresponding to 117.3 hours after the experiment started), the periodic impulses are obvious and the amplitude of the signal is larger than that in the early stage of the fault. In order to verify the performance of GSSA, we analyze the measured signal in the early stage of the fault to extract fault features.

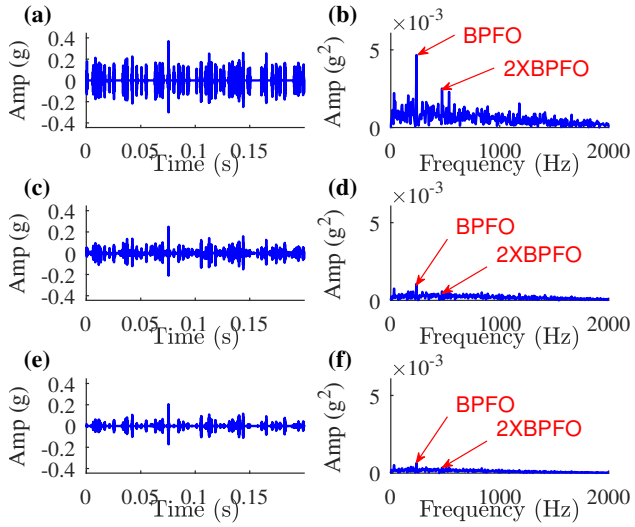

Fig. 10. The extracted results of different methods: (a) fault features extracted by GSSA, (b) SES of GSSA, (c) fault features extracted by BPD, (d) SES of BPD, (e) fault features extracted by WGL, and (f) SES of WGL.

B. Case 1: Results

In this case, the parameters of TQWT are the same as the simulation condition ($Q = 2$, $r = 5$, and $J = 10$) and Gaussian window with the length $l = 5$ is used as the neighboring window. Meanwhile, the K-sparsity parameters (K) of GSSA, BPD and WGL are set as 2% of the total number of wavelet coefficients. The extracted results of GSSA, BPD and WGL are shown in Fig. 10. It is observed that fault features extracted by GSSA are more obvious than those extracted by BPD and WGL in the time domain. Moreover, from the comparison of SES, the BPFO and 2xBPFO are successfully extracted by GSSA and their amplitudes are much larger than those extracted by BPD and WGL. Therefore, the results of GSSA are more accurate than the results of BPD and WGL which underestimate the energy of fault features.

C. Case 2: Experiment description

In this subsection, we verify the performance of the proposed method through another complete life test of the aeroengine bearing. As shown in Fig. 11, the test rig which is controlled by the Industrial Personal Computer (IPC) can simulate the load spectrum, rotation spectrum, and temperature spectrum of the aeroengine bearing. Meanwhile, the test rig consists of the main body of the testing machine, cooling and lubrication system, electrical control system, IPC, and data acquisition instrument. The main body of the test rig and its structural sketch are shown in Fig. 12(a-b). The shaft is driven by a high-speed motor and hosts one test bearing (bearing 3) and two support bearing (bearing 1 and bearing 2). One run-to-failure experiment was performed with the radial load and the axial load equal to 11 kN and 2 kN added to the test bearing by the load system. Two accelerometers (Lance LC0401) were mounted on the sleeve and the vibration signal was collected every 5 minutes using the econ data collector. The sampling frequency equal to 20.48 kHz. It is worth mentioning that the analysis frequency should be larger than the resonance

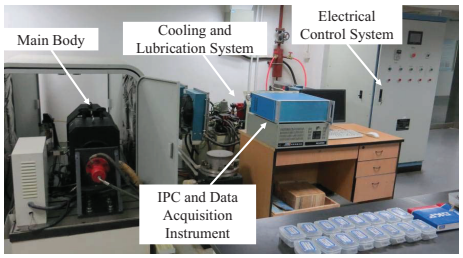

Fig. 11. The test rig of the aeroengine bearing.

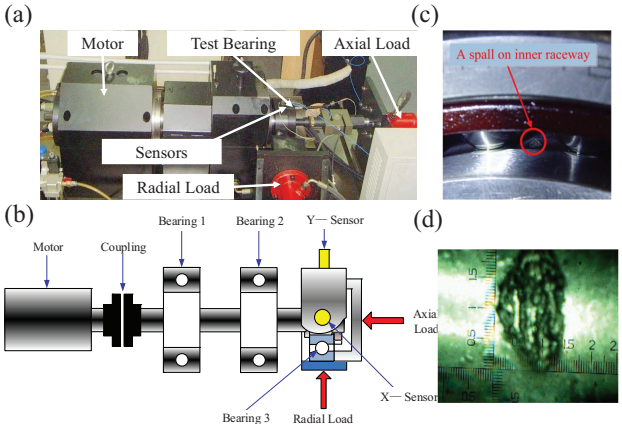

Fig. 12. (a) The zoom-in view of the main body of the experimental rig, (b) the structural sketch of the main body, (c) A spall on inner raceway, and (d) the area of the spall.

frequency plus the fault characteristic frequency for detecting the bearing faults. Therefore, the frequency response of the accelerometer and the data acquisition system need to meet this requirement.

After the total running time of the test bearing was 156.3 hours, a spall on the inner raceway was found in the test bearing, as shown in Fig. 12(c-d). Besides, the area of the spall was about 3 mm^2 measured by the monocular video microscope system. The type of the test bearing was an H7018C angular contact ball bearing, and its pitch diameter (D), ball diameter (d), number of balls (n), and contact angle (ϕ) are 117 mm, 11.12 mm, 15°, and 27 respectively.

$$\text{BPFI} = \frac{nf_r}{2}\left(1 + \frac{d}{D}\cos\phi\right) \tag{15}$$

where f_r is the rotating frequency.

According to the rotating Frequency (RF) 100 Hz and the geometrical parameters of the test bearing, the characteristic frequency of the fault located on the inner race (ball pass frequency of inner race, denoted by BPFI) was about 1474 Hz calulated by (15). We use one segment of the vibration signal (The length of the signal is 32768) collected during the run-to-failure experiment to verify the performance of the proposed method, as shown in Fig. 13.

D. Case 2: Results

We firstly use GSSA, BPD and WGL to analyze the collected vibration signal, and parameters of TQWT are $Q = 2$, $r = 2$, and $J = 10$. The K-sparsity parameters of GSSA, BPD and WGL are 4% of the total number of wavelet coefficients.

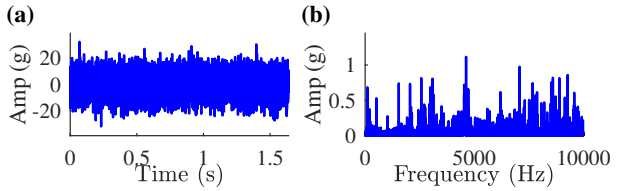

Fig. 13. (a) The original vibration signal, and (b) its spectrum.

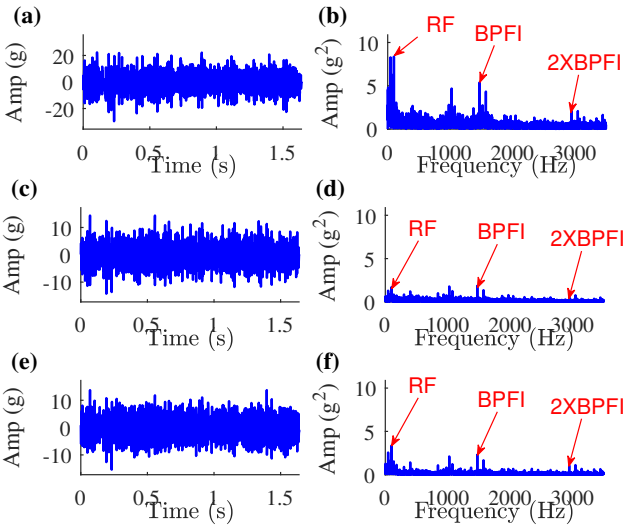

Fig. 14. The extracted results of different methods: (a) fault features extracted by GSSA, (b) SES of GSSA, (c) fault features extracted by BPD, (d) SES of BPD, (e) fault features extracted by WGL, and (f) SES of WGL.

The results of GSSA, BPD and WGL are shown in Fig. 14. Fig. 14(a) indicate that GSSA can preserve the amplitude of bearing fault features and reduce the noise interference more effectively than BPD and WGL. Besides, as shown in Fig. 14(b), BPFI and its high-order frequency (2xBPFI) extracted by GSSA are obvious in the SES. Meanwhile, the amplitudes of BPFI and 2xBPFI extracted by BPD and WGL are much smaller than those extracted by GSSA. It is worth mentioning that because the fault of the test bearing occurs on the inner raceway of the failure bearing, amplitudes of the periodic impulses may be modulated by the rotating speed, and thus RF is also dominant in the SES. In conclusion, the proposed GSSA method can extract the fault feature of the bearing effectively.

Similar to the above case, SK is also applied to extract the fault feature of the bearing. The kurtogram, the filtered signal, and its SES are shown in Fig. 15 (the optimal filter with the center frequency is equal to 8333.33 Hz and the bandwidth is equal to 3333.33 Hz). It is observed that we can also find the RF, BPFI, and 2xBPFI in the SES. However, fault characteristic frequencies extracted by GSSA are more dominant than these extracted by SK and the interference frequencies in the SK result are more complex than these in GSSA results. Thus, the proposed GSSA method is better than SK.

VI. Conclusion

In this paper, we interpret the concept of the algorithm-aware method which allows to design algorithms more flexible, in contrast to the model-aware method which needs to construct the specific penalty to model the sparse prior (if the prior is complex, it is really difficult to construct the expression). In addition, on the basis of the idea of algorithm-aware methods, we generalize structured shrinkage operators and form GSSA through rewriting PGD. The main advantages of the proposed method consist of introducing the group structure (also called weighted neighborhood structure), embedding the properties (unbiasedness, sparsity, continuity) of other proximal operators, and considering the multiscale property of the wavelet transform. Moreover, its satisfying performance is further verified by simulation studies, experimental cases, and comparisons with model-aware methods and fast kurtogram. More theoretical works and application developments including verifying the convergence of the proposed method, designing more complex denoisers, rewriting broader sparsity-assisted optimization frameworks, and studying the mathematical relationship between the defined operators and the corresponding penalties need to be carried out in the further research.

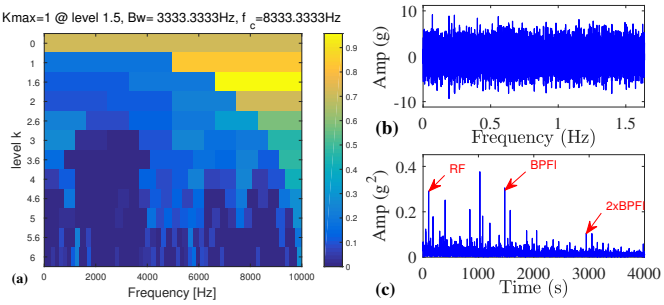

Fig. 15. The extracted fault feature by SK: (a) kurtogram of the measured signal, (b) the filtered signal, and (c) SES of SK.

References

[1] Jérôme Antoni, "The spectral kurtosis: a useful tool for characterising non-stationary signals," *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 282–307, 2006.

[2] D Abboud, M Elbadaoui, WA Smith, and RB Randall, "Advanced bearing diagnostics: A comparative study of two powerful approaches," *Mechanical Systems and Signal Processing*, vol. 114, pp. 604–627, 2019.

[3] Weiguo Huang, Guanqi Gao, Ning Li, Xingxing Jiang, and Zhongkui Zhu, "Time-frequency squeezing and generalized demodulation combined for variable speed bearing fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, 2018.

[4] Yaguo Lei, Jing Lin, Zhengjia He, and Ming J Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 35, no. 1, pp. 108–126, 2013.

[5] Jun Wang, Qingbo He, and Fanrang Kong, "Adaptive multiscale noise tuning stochastic resonance for health diagnosis of rolling element bearings," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 2, pp. 564–577, 2014.

[6] David L Donoho, "De-noising by soft-thresholding," *IEEE transactions on information theory*, vol. 41, no. 3, pp. 613–627, 1995.

[7] T Tony Cai and Bernard W Silverman, "Incorporating information on neighbouring coefficients into wavelet estimation," *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 127–148, 2001.

[8] Levent Sendur and Ivan W Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Transactions on signal processing*, vol. 50, no. 11, pp. 2744–2756, 2002.

[9] Binqiang Chen, Zhousuo Zhang, Chuang Sun, Bing Li, Yanyang Zi, and Zhengjia He, "Fault feature extraction of gearbox by using overcomplete rational dilation discrete wavelet transform on signals measured from

vibration sensors," *Mechanical Systems and Signal Processing*, vol. 33, pp. 275–298, 2012.

[10] WangPeng He, YanYang Zi, BinQiang Chen, Shuai Wang, and ZhengJia He, "Tunable q-factor wavelet transform denoising with neighboring coefficients and its application to rotating machinery fault diagnosis," *Science China Technological Sciences*, vol. 56, no. 8, pp. 1956–1965, 2013.

[11] Hailiang Sun, Yanyang Zi, and Zhengjia He, "Wind turbine fault detection using multiwavelet denoising with the data-driven block threshold," *Applied Acoustics*, vol. 77, pp. 122–129, 2014.

[12] Jinglong Chen, Zhiguo Wan, Jun Pan, Yanyang Zi, Yu Wang, Binqiang Chen, Hailiang Sun, Jing Yuan, and Zhengjia He, "Customized maximal-overlap multiwavelet denoising with data-driven group threshold for condition monitoring of rolling mill drivetrain," *Mechanical Systems and Signal Processing*, vol. 68, pp. 44–67, 2016.

[13] Ramy Hussein, Khaled Bashir Shaban, and Ayman H El-Hag, "Wavelet transform with histogram-based threshold estimation for online partial discharge signal denoising," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 12, pp. 3601–3614, 2015.

[14] Jianbo Yu and Haiqiang Liu, "Sparse coding shrinkage in intrinsic time-scale decomposition for weak fault feature extraction of bearings," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 7, pp. 1579–1592, 2018.

[15] Lin Wang, Gaigai Cai, Jun Wang, Xingxing Jiang, and Zhongkui 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.

[16] Ning Li, Weiguo Huang, Wenjun Guo, Guanqi Gao, and Zhongkui Zhu, "Multiple enhanced sparse decomposition for gearbox compound fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, 2019.

[17] Lingli Cui, Jing Wang, and Seungchul Lee, "Matching pursuit of an adaptive impulse dictionary for bearing fault diagnosis," *Journal of Sound and Vibration*, vol. 333, no. 10, pp. 2840–2862, 2014.

[18] Debarshi Sen, Amirali Aghazadeh, Ali Mousavi, Satish Nagarajaiah, and Richard Baraniuk, "Sparsity-based approaches for damage detection in plates," *Mechanical Systems and Signal Processing*, vol. 117, pp. 333–346, 2019.

[19] Scott Shaobing Chen, David L Donoho, and Michael A Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[20] Boyuan Yang, Ruonan Liu, and Xuefeng Chen, "Sparse time-frequency representation for incipient fault diagnosis of wind turbine drive train," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 11, pp. 2616–2627, 2018.

[21] Han Zhang, Xuefeng Chen, Zhaohui Du, and Ruqiang Yan, "Kurtosis based weighted sparse model with convex optimization technique for bearing fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 80, pp. 349–376, 2016.

[22] Weiguo Huang, Shijun Li, Xiangyu Fu, Cheng Zhang, Juanjuan Shi, and Zhongkui Zhu, "Transient extraction based on minimax concave regularized sparse representation for gear fault diagnosis," *Measurement*, p. 107273, 2019.

[23] Shibin Wang, Ivan Selesnick, Gaigai Cai, Yining Feng, Xin Sui, and Xuefeng Chen, "Nonconvex sparse regularization and convex optimization for bearing fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 9, pp. 7332–7342, 2018.

[24] Zhibin Zhao, Shuming Wu, Baijie Qiao, Shibin Wang, and Xuefeng Chen, "Enhanced sparse period-group lasso for bearing fault diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 3, pp. 2143–2153, 2019.

[28] Michael Grant, Stephen Boyd, and Yinyu Ye, "Cvx: Matlab software for disciplined convex programming," 2008.

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

[26] Ruobin Sun, Zhibo Yang, Xuefeng Chen, Shaohua Tian, and Yong Xie, "Gear fault diagnosis based on the structured sparsity time-frequency analysis," *Mechanical Systems and Signal Processing*, vol. 102, pp. 346–363, 2018.

[27] Wangpeng He, Binqiang Chen, and Yanyang Zi, "Enhancement of fault vibration signature analysis for rotary machines using an improved wavelet-based periodic group-sparse signal estimation technique," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 232, no. 6, pp. 941–951, 2018.

[29] Olivier Fercoq and Peter Richtárik, "Accelerated, parallel, and proximal coordinate descent," *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 1997–2023, 2015.

[30] Huan Li and Zhouchen Lin, "Accelerated proximal gradient methods for nonconvex programming," in *Advances in neural information processing systems*, 2015, pp. 379–387.

[31] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[32] Kai Siedenburg and Monika Dörfler, "Persistent time-frequency shrinkage for audio denoising," *Journal of the Audio Engineering Society*, vol. 61, no. 1/2, pp. 29–38, 2013.

[33] Ivan W Selesnick, "Wavelet transform with tunable q-factor," *IEEE transactions on signal processing*, vol. 59, no. 8, pp. 3560–3575, 2011.

[34] Yongbo Li, Xihui Liang, Minqiang Xu, and Wenhu 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.

[35] Boaz Ophir, Michael Lustig, and Michael Elad, "Multi-scale dictionary learning using wavelets," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1014–1024, 2011.

[36] Matthieu Kowalski, Kai Siedenburg, and Monika Dörfler, "Social sparsity! neighborhood systems enrich structured shrinkage operators," *IEEE transactions on signal processing*, vol. 61, no. 10, pp. 2498–2511, 2013.

[37] Florian Lieb and Hans-Georg Stark, "Audio inpainting: Evaluation of time-frequency representations and structured sparsity approaches," *Signal Processing*, vol. 153, pp. 291–299, 2018.

[38] David L Donoho and Jain M Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[39] Joseph Woodworth and Rick Chartrand, "Compressed sensing recovery via nonconvex shrinkage penalties," *Inverse Problems*, vol. 32, no. 7, pp. 075004, 2016.

[40] Cun-Hui Zhang et al., "Nearly unbiased variable selection under minimax concave penalty," *The Annals of statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[41] Jianqing Fan and Runze Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[42] Zongben Xu, Xiangyu Chang, Fengmin Xu, and Hai Zhang, "$l_{1/2}$ regularization: A thresholding representation theory and a fast solver," *IEEE Transactions on neural networks and learning systems*, vol. 23, no. 7, pp. 1013–1027, 2012.

[43] Hai Qiu, Jay Lee, Jing Lin, and Gang Yu, "Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics," *Journal of sound and vibration*, vol. 289, no. 4-5, pp. 1066–1090, 2006.