IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 30, 2019, accepted December 30, 2019, date of publication January 22, 2020, date of current version January 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968726

A Bi-Level Nested Sparse Optimization for Adaptive Mechanical Fault Feature Detection

HAN ZHANG[1,2], XUEFENG CHEN[3], (Member, IEEE), XIAOLI ZHANG[1,2], (Member, IEEE), AND XINRONG ZHANG[1,2]

[1] Key Laboratory of Road Construction Technology and Equipment of Ministry of Education, Chang'an University, Xi'an 710064, China
[2] School of Construction Machinery, Chang'an University, Xi'an 710064, China
[3] State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Corresponding author: Han Zhang (zhanghan@chd.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51805040, in part by the Natural Science Foundation of Shanxi Province under Grant 2019JQ-673, in part by the Shaanxi Science and Technology Innovation Project under Grant 2016KTZDGY-02-03, in part by the Fundamental Research Funds for the Central Universities under Grant 300102258301, Grant 300102259402, and Grant 300102258305, and in part by the Shaanxi International Science and Technology Cooperation Project under Grant 2019KW-015.

ABSTRACT Denoising is a permanent topic and there are various denoisers proposed in the fault diagnosis of industrial systems. However, it is still ambiguous to evaluate their performance quantitatively in terms of mean square error (MSE) and further achieve their maximum gains, because it is always infeasible to obtain the MSE metric without real feature signals in the engineering practices. Therefore, leveraging Stein Unbiased Risk Estimator (SURE) theory, a bi-level nested sparse optimization framework (BiNSOF) is proposed to jointly optimize a parameterized sparse denoiser as well as its regularization parameter, further obtaining the near-optimal fault features with a minimum MSE. The inner level of BiNSOF utilizes a ℓ_1 regularized sparse denoiser to describe the intrinsic sparse structure of feature information, which can be effectively addressed by popular primal-dual splitting schemes. The core of the outer optimization level is a SURE-based unbiased estimator for MSE, and the minimum MSE search problem is transformed into a quadratic optimization problem which could be fast solved by classic golden section search schemes. The proposed BiNOSP can perfectly approximate the oracle MSE without any real feature information, and further provides a reliable way to obtain the optimal hyper-parameter sets for the maximum performance gains of the sparse denoiser. The computational complexity of the advocated approach is also investigated. Moreover, its feasibility and performances are profoundly evaluated by a set of comprehensive numerical studies. Lastly, two bearing fault detection cases confirm the applicability and superiority of the proposed framework.

INDEX TERMS Sparse optimization, Stein unbiased risk estimator (SURE), fault feature detection, primal-dual splitting, bi-level nested optimization, adaptive parameter selection.

I. INTRODUCTION

In modern industrial systems, fault diagnosis and isolation (FDI) has received an intensive amount of research interest during the last decades [1]–[3]. It is due to its ability to reduce maintenance costs and prevent the harmful and sometimes devastating consequences of faults and failures [4]. This topic has become far more attractive and critical for complex and expensive systems with less tolerance for performance degradation, productivity disease, and safety hazards [5]. The purpose of FDI is to identify which component is being subjected to malfunction or deviation from its normal working status and thus the benefits are improved operability and safety [6].

A primary issue in the FDI procedure is to detect feature information from noisy measurements and many advanced signal processing methods have been developed [7], such as spectrum analysis [8], time-frequency analysis [9], wavelet transform (WT) [10], spectrum kurtosis(SK) [11], adaptive

The associate editor coordinating the review of this manuscript and approving it for publication was Shaoyong Zheng.

mode decomposition [12], cyclostationary descriptors [13], and deep learning [14], [15], etc. In recent years, sparsity representation based fault diagnosis (SRFD) techniques have been one of the hottest topics in the signal processing society and aroused extensive interests. The core idea of SRFD is to explore the sparsity prior of fault feature signal in an elaborately designed transformation space, and then reconstruct the target feature through optimization strategy [16]. Research on SRFD mainly focuses on the design of the sparse representation dictionary and the exploit of feature prior information to construct a proper sparse regularized optimization model. As to the dictionary design, Cui constructed an impulse dictionary based on the exponentially decaying response of rolling bearing faults and then used matching pursuit methods to extract fault feature signal [17]. Qin designed an improved Morlet wavelet to sparsely represent weak transient features and meanwhile adopted an iterative thresholding algorithm to extract focused sparse components [18], [19]. Meanwhile, learned dictionary established from collected signals also attracts many researcher's interests because of its flexibility and adaptivity [20], [21]. On the other hand, various priori knowledges are exploited to design sparse regularized models. Enforcing different threshold levels on sparse coefficients according to their kurtosis values of envelope spectrum, Zhang *et al.* proposed a weighted sparse model [22] for bearing fault diagnosis. Sun *et al.* utilized mixed-norm priors on time-frequency coefficients to construct a structured sparsity time-frequency analysis model and meanwhile verified its effectiveness through extracting feature signals from strong noisy measurements [23]. Du explored the low-rank property of feature signal and proposed a $\ell_{2,1}$ norm based collaborative sparse model for robust feature identification [24]. Ding proposed a shock-response convolutional sparse coding model for the diagnosis of a wheelset bearing in a high-speed train [25]. Qing Li *et al.* developed a sparse low-rank model to extract two types of impulsive fault components [26]. Meanwhile, replacing the popular convex ℓ_1 norm regularization, non-convex sparsity-promoting regularization more recently has been extensively studied [27]. Zhao *et al.* proposed an adaptive enhanced sparse period-group lasso model which can promote the sparsity within and across groups of the impulsive bearing fault feature [28]. Wang proposed a non-convex sparse regularization method based on the generalized minimax convex penalty [29]. Zhang proposed a collaborative sparse classification method based on the low-rank property of two-dimensional fault feature signal and applied to the gear-hub crack fault detection [30].

In a nutshell, nearly all sparse feature detection techniques need to solve a regularized sparse optimization model (denoted as the sparse estimator in the following part) with a set of adjustable hyper-parameters. Meanwhile, their performances significantly depend on the hyper-parameter settings. However, how to confirm optimal parameter is a challenging problem as these parameters are always different for diverse fault feature detection tasks. Moreover, there lacks a reliable criterion to select optimal hyper-parameters. Currently, in most literatures of fault diagnosis society, hyper-parameter setting schemes can be categorized into two main strategies. One strategy is to select optimal hyper-parameters empirically by trial and error approaches. However, it requires much human labor to obtain only suboptimal results and it is often sensitive to noise levels. The other strategy is to perform an exhaust search among a set of pre-specific parameter grids based on minimum Mean Squared Error (MSE) $= \|x - \hat{x}\|_2^2$ or maximum SNR $= 20\log\frac{\|x\|_2}{\|x-\hat{x}\|_2}$ (x are real features and $\hat{x}$ are estimated features). Generally speaking, the latter strategy is very effective and ubiquitous in numeric simulation analysis, but it is unfavorable and even infeasible in practical engineering applications due to inaccessible to real features x for MSE. Therefore, it is a very necessary and attractive task to establish an adaptive strategy to select optimal hyper-parameter configuration effectively, which guarantees the optimal performance of sparse estimators.

To address theses problems above, a bi-level nested sparse optimization framework (BiNSOF) is proposed in this paper to jointly optimize a parameterized sparse denoiser as well as its regularization parameter. The proposed BiNSOF framework could be viewed as a bi-level nested optimization model. The inner optimization level utilizes the ℓ_1 regularized sparse estimator to entail the sparse structure of feature signals in the transformed domain. The core of the outer optimization level is that, without any knowledge of the noise-free feature signal x, an unbiased estimator for MSE could be established through Stein Unbiased Risk Estimator (SURE) theory [31]–[33], which effectively transforms minimum MSE estimation problem into a quadratic optimization problem. The proposed BiNSOF could adaptively tune its parameters to mitigate the limitation of regularization parameter selection in industrial applications. Meanwhile, extensive numeric analysis demonstrates that the proposed approach provides a nearly optimal way for adaptive feature detection problems. Lastly, the superiority of BiNSOF is further demonstrated through applying it to the CWRU benchmark bearing fault data and an aero-engine bearing test data.

The rest of the paper is organized as follows. Section II rigorously describes the nested optimization problems of BiNSOF and their relationship, and further designs a general feature detection framework for mechanical systems. Meanwhile, the complexity of the proposed technique is investigated. Section III investigates the performance of the proposed BiNSOF framework through extensive numerical experiments. Section IV is dedicated to applying BINSOF to two bearing fault data sets. Conclusions and future works are reported in Section V.

II. PROPOSED ALGORITHM

A. NOTATION AND PROBLEM FORMULATION

Fault features $x \in \mathbb{R}^{m\times 1}$ are often contaminated by relatively strong noises $w \in \mathbb{R}^{m\times 1}$ which may arise due to sensor imperfection, poor running environment or communication errors. Therefore, the measurements $y \in \mathbb{R}^{m\times 1}$ could be described by a linear data model

$$y = x + w \tag{1}$$

where $w \sim \mathcal{N}(\mu, \sigma^2)$ is an additive white Gaussian noises of mean μ and variance σ^2. Typically, the task to detect feature information x from y can be formulated as an operator $f_\theta(y) : \mathbb{R}^m \longrightarrow \mathbb{R}^m$ that maps observation data y to estimated features:

$$\hat{x} = f_\theta(y) \tag{2}$$

where $\theta \in \Theta$ represents the continuous hyper-parameter set of $f_\theta(y)$. Meanwhile, fault feature information often has an intrinsic sparse structure, strictly speaking, for a signal x, if we have a favorable over-complete dictionary $\Psi \in \mathbb{R}^{m\times n}$, most energy of x can be only concentrated on a few nonzero elements of coefficient $\alpha = \Psi^T x$ or α has many zero values. Thus a regularization term ($\ell_p, 0 \le p \le 1$) is often employed to describe the sparse structure of fault features and then the feature estimator can be formulated as a regularized optimization problem [34]:

$$f_\theta(y) = \arg\min_x \frac{1}{2}\|y - x\|_2^2 + \theta \left\|\Psi^T x\right\|_p \tag{3}$$

where θ is the regularization parameter. A large value of θ tends to over-smooth structures while a small value leads to noisy recovery. To make the problem (3) more tractable, moreover, the model parameter p is set as 1 since ℓ_1 norm yields the convex objective cost and we can employ many off-the-shelf optimization algorithms to solve it. The success of this model mainly depends on the dictionary Ψ and the regularization parameter θ. In this paper, we focus on the later problem and suppose the dictionary Ψ is appropriate to capture the sparse structure of x. However, the adaptive configuration of regularization parameter is generally not a trivial task.

A natural way is to design a performance criterion to evaluate the reconstruction quality and further confirm optimal regularization parameters. A popular criterion is the MSE which measures the difference between the reconstructed feature signal $f_\theta(y)$ and the original noise-free one x:

$$MSE(\hat{x}) = \mathbb{E}_w\left\{\|f_\theta(y) - x\|_2^2\right\} \tag{4}$$

where $\mathbb{E}_w\{\cdot\}$ stands for the mathematical expectation operator. However, due to the nonsmooth property of sparse regularizer ($\ell_p, 0 \le p \le 1$), it is impossible to directly obtain a closed solution for $f_\theta(y)$, and an iterative solver must be introduced. Therefore, an optimal parameter configuration can be obtained by the following bi-level nested optimization:

$$f_\theta(y) = \arg\min_x \frac{1}{2}\|y - x\|_2^2 + \theta\|\Psi^T x\|_1 \tag{5}$$

$$\theta^* = \arg\min_{f,\theta} \mathbb{E}_w\left\{\|f_\theta(y) - x\|_2^2\right\} \tag{6}$$

Obviously, this problem is highly ill-posedness as the MSE depends on the noise-free signal x which is generally unavailable or unknown in real engineering applications. Therefore, a practical way is to find an unbiased estimator to approximate the true MSE. Fortunately, a theoretical result developed by C. Stein makes this possible in Gaussian scenario [35] and provides a powerful scheme to obtain an unbiased estimator of true MSE criterion. Therefore, the sparse denoiser $f_\theta(y)$ with the regularization parameter θ could be optimized based on the SURE theory. Without ever requiring knowledge of the noise-free feature signals x, most importantly, this approach could achieve favorable accuracy to calculate MSE and meanwhile tack MSE evolution reliably. Moreover, its unbiasedness can be established rigorously, which makes it non-empirical.

However, there remain two problems need to be addressed. Firstly, it is a necessary procedure to develop efficient algorithms to jointly optimize Eq.(5) and Eq.(6). On the other hand, the divergence of the operator $f_\theta(y)$ with respect to y is one key ingredient of SURE calculation, but the explicit evaluation of the divergence is out of reach because the sparse estimator $f_\theta(y)$ is an iterative optimization procedure. Therefore, another challenge is to propose an feasible scheme to evaluate divergence operation without heavy computational cost. All detailed information about the two problems is shown in the subsequent sections.

B. STEIN'S UNBIASED RISK ESTIMATE - SURE

In this section, we discuss an unbiased estimator for the MSE, which will later be used to optimize the sparse denoiser as well as its regularization parameter. The Stein Unbiased Risk Estimator (SURE) was proposed in [35]. One of the best known algorithms that uses SURE is Donoho's SureShrink denoising algorithm [36]. The core idea of the SURE principle is to seek an unbiased estimator for the MSE that is only a function of the observation y and the feature estimator $f_\theta(y)$. Denote such an MSE estimator as $\eta(f_\theta(y))$. Then, if we have accessed to such a function $\eta(f_\theta(y))$ that estimates the MSE, while being also dependent on a set of parameters θ that control the feature detection accuracy. Naturally we could choose one optimal value θ^* to minimize that function, which guarantees the estimator $f_\theta(y)$ to achieve the near-optimal detection accuracy in terms of MSE.

To cover most of nonsmooth estimators with sparse regularization penalties, we introduce the following assumptions.

Assumption 1: The established feature estimator $f_\theta(y)$ could be simplified as a m-dimensional vector mapping and it has following characteristics:

A.1 The estimator $f_\theta(y)$ is always a single-valued mapping though $f_\theta(y)$ is possibly multivalued.

A.2 The estimator $f_\theta(y)$ is weakly differentiable with respect to y.

A.3 The estimator $f_\theta(y)$ is bounded by some fast increasing functions, typically such that:

$$\lim_{|z|\to\infty} f_\theta(x+z)e^{-\frac{z^2}{2\sigma^2}} = 0.$$

A.4 The estimator $f_\theta(y)$ is uniformly Lipschitz continuous with Lipschitz constant $L_1 > 0$.

A.5 The estimator $f_\theta(y)$ is such that $f_\theta(0) = 0$ for any θ.

Moreover, we also require that the divergence of $f_\theta(y)$ with respect to the observation y is given by

$$\mathbf{div}\,\{f_\theta(y)\} = \sum_{i=1}^{m} \frac{\partial f_{\theta_i}(y)}{\partial y_i} \tag{7}$$

where $f_{\theta_i}(y)$ and y_i denotes the i-th component of the vectors $f_\theta(y)$ and y, respectively.

Definition 1: Given y as in (1), $w \sim \mathcal{N}(0, \sigma^2)$ is an additive white Gaussian noises of mean 0 and variance σ^2. SURE corresponding to $f_\theta(y)$ is defined as:

$$\eta(f_\theta(y)) = \|y - f_\theta(y)\|_2^2 - m\sigma^2 + 2\sigma^2 \mathbf{div}\,\{f_\theta(y)\} \tag{8}$$

Then, the following description, according to Stein theory [35], states that η is indeed unbiased

Theorem 1: If $f_\theta(y)$ satisfies the assumption 1, then the random variable $\eta(f_\theta(y))$ is an unbiased estimator of the expected MSE, i.e.

$$\mathbb{E}_w\,\{\eta(f_\theta(y))\} = \mathbb{E}_w\left\{\|x - f_\theta(y)\|_2^2\right\} \tag{9}$$

Remark 1: Without any assumptions on the noise-free features x, it is possible to reliably approximate the oracle MSE by the unbiased estimator (8) which is a function of y only. This has an important consequence: contrary to what is frequently done in the literature, feature signal x is not modeled as a random process in our estimator, i.e., we do not even require x belong to a specific class of signals.

Remark 2: The constant energy term in the unbiased estimator (8), $m\sigma^2$, is irrelevant to the estimator process $f_\theta(y)$, therefore, there is no need to estimate it since it will disappear in the minimization problem (6). Consequently, we will consider these terms which is the only part of the MSE estimator that depends on the choice of the unbiased estimator $f_\theta(y)$. Naturally, the optimization problem (6) could be formulated as follows,

$$\begin{aligned}\theta^* &= \arg\min_{f_\theta(y)} \mathbb{E}_w\left\{\|f_\theta(y) - x\|_2^2\right\}\\ &= \arg\min_{f_\theta(y)} \mathbb{E}_w\left\{\|(y - f_\theta(y))\|_2^2 + 2\sigma^2 \mathbf{div}\,\{f_\theta(y)\}\right\}\end{aligned} \tag{10}$$

Remark 3: The unbiased estimator and optimization problem (10) require the knowledge of the variance σ^2, and numerous works have entirely dedicated to σ estimation problem [37]–[39]. One popular strategy is to firstly address the noisy signals with a unit-norm highpass filter, and then estimate the noise variance $\hat{\sigma}$ from the filtered components. This method has been popularized in wavelet-based denoising algorithms developed by Donoho [37], and the median of the absolute deviation (MAD) of the highest frequency subband w_1 has become as a benchmark method for $\hat{\sigma}$:

$$\hat{\sigma} = 1.4826 * Median\,\{|w_1 - Median(w_1)|\} \tag{11}$$

For a wide range of noise levels, the wavelet-domain MAD estimator usually gives an effective estimation of noise variance in most engineering applications. Thus, this approach has been adopted for noise variance estimation throughout this paper.

The reliability of the unbiased estimator is another important topic and now we evaluate its reliability by computing the expected squared error between $\eta(f_\theta(y))$ and the actual MSE. The reliability is guaranteed by the following theorem.

Theorem 2: Under the same hypotheses as Definition 1, the expected squared error between the estimator $\eta(f_\theta(y))$ and the actual MSE is given by [32]:

$$\mathbb{E}_w\left\{\left(\frac{\eta(f_\theta(y)) - \|x - f_\theta(y)\|_2^2}{m\sigma^2}\right)^2\right\} = O(\frac{1}{m}) \tag{12}$$

Remark 4: Generally, the estimated MSE, $\eta(f_\theta(y))$, is inversely proportional to the sample number m. Moreover, the number of samples in practical applications is usually large, and thus the estimator has a small variance, typically $\propto 1/m$. This designed estimator is therefore close to its expectation, which indeed effectively describes the true MSE.

Now, we can estimate the MSE purely based on the input data y, the divergence of $f_\theta(y)$, and the noise statistics, meanwhile don't require knowledges whatsoever of the noise-free signal x. However, the evaluation of $\mathbf{div}\,\{f_\theta(y)\}$ is difficult or even infeasible when there is no explicit form for the estimator, as is the case for the sparse optimization model (5). Therefore, we will demonstrates an iterative procedure to compute $\mathbf{div}\,\{f_\theta(y)\}$ in the following parts.

C. ℓ_1PARSE OPTIMIZATION ALGORITHM

In recent years, how to design numeric solvers for sparse model (5) has attracted lots of researchers and many useful algorithms have been proposed. An important progress in the last decades is the primal-dual method, which offers significantly computational advantages [40], [41]. Now we illustrate detailed optimization procedures for the problem (5) via the primal-dual strategy.

Based on Legendre-Fenchel transforms, the nonsmooth term $\left\|\Psi^T x\right\|_1$ could be rewritten as

$$\left\|\Psi^T x\right\|_1 = \sup_{\lambda\in\mathbb{R}^n} \left\langle \Psi^T x, \lambda\right\rangle - R^*(\lambda) \tag{13}$$

where $R^*(\lambda)$ is the convex conjugate function of $\|\lambda\|_1$ and defined as:

$$R^*(\lambda) = \sup_{r\in\mathbb{R}^n} \langle\lambda, r\rangle - \max_{\|s\|_\infty\le 1} \langle r, s\rangle \tag{14}$$

Substituting (13) into (3), we obtain one equivalent saddle point problem:

$$\min_x \max_\lambda \frac{1}{2\theta}\|y - x\|_2^2 + \left\langle \Psi^T x, \lambda\right\rangle - R^*(\lambda) \tag{15}$$

Then, two subproblems are formulated through alternately optimizing a single block of variables and keeping the rest of variables fixed.

$$\arg\max_\lambda \left\langle \Psi^T x, \lambda\right\rangle - R^*(\lambda) - \frac{1}{2\tau}\left\|\lambda - \lambda^k\right\|_2^2 \tag{16}$$

$$\arg\min_x \frac{1}{2\theta}\|y - x\|_2^2 + \left\langle \Psi^T x, \lambda\right\rangle + \frac{1}{2\xi}\left\|x - x^k\right\|_2^2 \tag{17}$$

where the auxiliary terms $\|\lambda-\lambda^k\|_2^2$ and $\|x-x^k\|_2^2$ could guarantee the optimal point of every subproblem is not away from previous points, which often provides an important condition for algorithmic convergence. In the following, we will provide the implementation details to obtain efficient solutions to each separated sub-problem. For simplicity, the iteration subscript k is omitted without confusion.

The λ-subproblem (16) is just the proximal operator of $R^*(\cdot)$ and thus we can obtain a closed-form solution as follows

$$U^{k+1} = \left(\lambda^k + \tau\Psi^T x\right) \tag{18}$$

$$\lambda^{k+1} = U^{k+1} - \tau Prox_{R/\tau}\left(U^{k+1}/\tau\right) \tag{19}$$

where the $Prox_{\tau R}(U)$ is defined as

$$Prox_{\tau R}(U) = \begin{cases} U_i+\tau & \text{if } U_i \le -\tau, \\ 0 & \text{if } -\tau < U_i < \tau, \\ U_i - \tau & \text{otherwise.} \end{cases} \tag{20}$$

The x-subproblem (17) can be reformulated as a least-square problem and a closed form solution thus can be straightforwardly obtained

$$V^{k+1} = \left(x^k - \xi\Psi\lambda^{k+1}\right) \tag{21}$$

$$x^{k+1} = (\theta+\xi)^{-1}\left(\theta V^{k+1} + \xi y\right) \tag{22}$$

Moreover, in order to improve the numerical efficiency, an extrapolation step is performed for x,

$$\tilde{x}^{k+1} = x^{k+1} + \zeta\left(x^{k+1} - x^k\right) \tag{23}$$

Finally, performing the above iterations leads to an optimal point (x^*, λ^*). Based on the Fermat's rule, it can be proved that, if $\left(\hat{x}, \lambda^*\right)$ is the solution of problem (15), then $\hat{x}$ is a solution to the primal problem (3) and λ^* is a solution to the dual one. Moreover, a detailed description of the proposed algorithm for sparse estimator $\hat{x} = f_\theta(y)$ is illustrated in algorithm 1.

```
Algorithm 1 Sparse Denoiser $f_\theta(y)$ Solver (SDS)
Require: Observation signal $y$, dictionary $\Psi$, parameters $\theta$.
Ensure: set $k = 0, \lambda^0 = \Psi^T y, x^0 = y, \tilde{x}^0 = y, L = \|\Psi\|$,
    $\tau > 0, \xi > 0, \tau\xi L^2 < 1$ and $\zeta \in [0,1], K = 50$.
1: for $k \le K$ do
2:     $U^{k+1} \leftarrow \lambda^k + \tau\Psi^T\tilde{x}^k$.
3:     $\lambda^{k+1} \leftarrow U^{k+1} - \tau Prox_{R/\tau}\left(U^{k+1}/\tau\right)$.
4:     $V^{k+1} \leftarrow x^k - \xi\Psi\lambda^{k+1}$.
5:     $x^{k+1} \leftarrow (\theta+\xi)^{-1}(\theta V^{k+1} + \xi y)$.
6:     $\tilde{x}^{k+1} \leftarrow x^{k+1} + \zeta\left(x^{k+1} - x^k\right)$
7:     $k \leftarrow k + 1$.
8: end for
Output: Feature estimate $\hat{x}$.
```

Since all the subproblems of the proposed algorithm have closed-form solutions, its convergence is guaranteed by the primal-dual splitting theory given in [40].

D. OPTIMAL FEATURE DETECTION BASED ON SPARSE DENOISER

After the explicit formulation of sparse estimator $f_\theta(y)$ has been constructed in the subsection II-C, the optimal point of optimization problem (10) could be obtained through the golden-section [42] or quasi-Newton optimization techniques [43]. However, a major practical difficulty when computing the objective values (10) lies in the numerical method of $\mathbf{div}\,\{f_\theta(y)\}$ defined by (7). Due to that there is no closed-form expressions for the sparse estimator (3), it is not possible to evaluate $\mathbf{div}\,\{f_\theta(y)\}$ directly. Therefore, we adopt an iterative way developed from chain rules to compute the weak directional derivative of $f_\theta(y)$. Based on step 5 of algorithm 1 and the linearity of Jacobian matrix $\partial f_\theta(y)/\partial y$, we can obtain

$$\frac{\partial f_\theta(y)}{\partial y} = \mathcal{D}_x^{(k+1)} = (\theta+\xi)^{-1}(\theta\mathcal{D}_V^{(k+1)} + \xi) \tag{24}$$

For other steps in the algorithm 1, we have

$$\mathcal{D}_U^{(k+1)} = \mathcal{D}_\lambda^{(k)} + \tau\Psi^T\mathcal{D}_{\tilde{x}}^{(k)} \tag{25}$$

$$\mathcal{D}_\lambda^{(k+1)} = \mathcal{D}_U^{(k+1)} - \tau\frac{\partial Prox_{R/\tau}\,(U/\tau)}{\partial U}\mathcal{D}_U^{(k+1)} \tag{26}$$

$$\mathcal{D}_V^{(k+1)} = \mathcal{D}_x^{(k)} - \xi\Psi\mathcal{D}_\lambda^{(k+1)} \tag{27}$$

$$\mathcal{D}_{\tilde{x}}^{(k+1)} = \mathcal{D}_x^{(k+1)} + \zeta\left(\mathcal{D}_x^{(k+1)} - \mathcal{D}_x^{(k)}\right) \tag{28}$$

It can be shown that soft-thresholding function (20) is weakly differentiable and thus its weak Jacobian $\partial Prox_{\tau R}\,(U)\,/\partial U$ is diagonal, with diagonal elements, for $1 \le i \le n$,

$$\frac{\partial Prox_{\tau R}\,(U)}{\partial U} = \begin{cases} -1 & \text{if } U_i \le -\tau, \\ 0 & \text{if } -\tau < U_i < \tau, \\ 1 & \text{otherwise.} \end{cases} \tag{29}$$

Now that a feasible and effective way to obtain the $\mathbf{div}\,\{f_\theta(y)\}$ has been established, and therefore the SURE-based MSE can be estimated by (8). Algorithm 2 summarizes the procedure of the sparse solver with SURE-based MSE estimation under one fixed parameter $\hat{\theta}$.

Since an analytical solution for problem (10) is difficult to obtain, another iterative optimization technique is exploited. The objective function with respective to θ in essence is piecewise-affine [32] and furthermore there is only one minimum point in the interval $\theta \in [0, |\Psi^T y|_\infty]$. Thus, the golden section search method is adopted to find the optimal solution $f_{\theta^*}(y)$. Substituting the iterative schemes of algorithm 2 into the second-stage optimization (10), then, a Bi-level nested sparse optimization framework is obtained and shown in algorithm 3. The mapping *SureSDS* shown in algorithm 2 denotes the sparse solver with SURE-based MSE estimation, and the mapping *GSS* is the update criterion in golden section search method. With GSS's iteration increases, the interval $[\theta_L^{i+1}, \theta_H^{i+1}]$ is gradually reduced to one optimal parameter θ^*. The update steps in the line 6 $\sim$ 12 of Algorithm 3 is performed cyclically until the interval length is less than one pre-specified tolerance level $\epsilon = 10^{-3}$, one optimal

Algorithm 2 The Sparse Solver With SURE-Based MSE Estimation (SureSDS)

```
Require: Observation signal $y$, sparse representation dictionary $\Psi$ and parameter $\hat{\theta}$.
Ensure: let $k = 0, \lambda^0 = \Psi^T y, x^0 = y, \tilde{x}^0 = y, L = \|\Psi\|, \tau > 0, \xi > 0, \tau\xi L^2 < 1, \zeta \in [0,1], K = 50; \mathcal{D}_U, \mathcal{D}_\lambda, \mathcal{D}_x, \mathcal{D}_{\tilde{x}} \leftarrow \mathbf{0}$;
 1: for $k \le K$ do
 2:     $U^{k+1} \leftarrow \lambda^k + \tau \Psi^T \tilde{x}^k$.
 3:     $\mathcal{D}_U^{(k+1)} \leftarrow \mathcal{D}_\lambda^{(k)} + \tau \Psi^T \mathcal{D}_{\tilde{x}}^{(k)}$.
 4:     $\lambda^{k+1} \leftarrow U^{k+1} - \tau Prox_{R/\tau}\left(U^{k+1}/\tau\right)$.
 5:     $\mathcal{D}_\lambda^{(k+1)} \leftarrow \mathcal{D}_U^{(k+1)} - \tau \frac{\partial Prox_{R/\tau}(U/\tau)}{\partial U} \mathcal{D}_U^{(k+1)}$.
 6:     $V^{k+1} \leftarrow x^k - \xi \Psi \lambda^{k+1}$.
 7:     $\mathcal{D}_V^{(k+1)} \leftarrow \mathcal{D}_x^{(k)} - \xi \Psi \mathcal{D}_\lambda^{(k+1)}$.
 8:     $x^{k+1} \leftarrow (\theta + \xi)^{-1}(\theta V^{(k+1)} + \xi y)$
 9:     $\mathcal{D}_x^{(k+1)} \leftarrow (\theta + \xi)^{-1}(\theta \mathcal{D}_V^{(k+1)} + \xi)$
10:     $\tilde{x}^{k+1} \leftarrow x^{k+1} + \zeta\left(x^{k+1} - x^k\right)$.
11:     $\mathcal{D}_{\tilde{x}}^{(k+1)} \leftarrow \mathcal{D}_x^{(k+1)} + \zeta\left(\mathcal{D}_x^{(k+1)} - \mathcal{D}_x^{(k)}\right)$.
12:     $k \leftarrow k + 1$.
13: end for
14: $\hat{x} \leftarrow x^K$.
15: div $\left\{f_{\hat{\theta}}(y)\right\} \leftarrow \mathcal{D}_x^K$
16: $\sigma \leftarrow$ MAD estimator by (11).
17: $\eta(f_{\hat{\theta}}(y)) \leftarrow \left\|\left(y - \hat{x}\right)\right\|_2^2 + 2\sigma^2 \mathbf{div}\left\{f_{\hat{\theta}}(y)\right\}$.
Output: Feature information $\hat{x}$,
        Estimated MSE $\eta(f_{\hat{\theta}}(y))$,
```

Algorithm 3 Adaptive Feature Detection Based on the Bi-Level Nested Sparse Optimization Framework (BiNSOF)

```
Require: Observation signal $y$, dictionary $\Psi$.
Ensure: set $\theta_L^0$ by the formula (30) and $\theta_U^0 = |\Psi^T y|_\infty$ and $\epsilon = 10^{-3}, I^0 = \theta_U^0 - \theta_L^0, \theta_a^0 = \theta_U^0 - 0.618 I^0, \theta_b^0 = \theta_L^0 + 0.618 I^0$
 1: while $I^i = \theta_U^i - \theta_L^i \ge \epsilon$ do
 2:     $\left(\theta_L^{i+1}, \theta_U^{i+1}, \theta_a^{i+1}, \theta_b^{i+1}\right) \leftarrow GSS(\theta_L^i, \theta_U^i, SureSDS)$.
 3: end while
 4: $\left(\sim, \eta_a^{i+1}\right) \leftarrow SureSDS\left(\theta_a^{i+1}\right)$.
 5: $\left(\sim, \eta_b^{i+1}\right) \leftarrow SureSDS\left(\theta_b^{i+1}\right)$.
 6: if $\eta_a^{i+1} > \eta_b^{i+1}$ then
 7:     $\theta^* = \frac{1}{2}(\theta_a^{i+1} + \theta_U^{i+1})$.
 8: else if $\eta_a^i < \eta_b^i$ then
 9:     $\theta^* = \frac{1}{2}(\theta_L^{i+1} + \theta_b^{i+1})$.
10: else
11:     $\theta^* = \frac{1}{2}(\theta_a^{i+1} + \theta_b^{i+1})$.
12: end if
13: $(x^*, \eta^*) \leftarrow SureSDS\,(\theta^*)$.
Output: Optimal feature information $x^*$,
        Optimal parameter $\theta^*$.
        Minmimum MSE $\eta^*$.
```

parameter is then obtained. More detailed description about golden section search could be found in the chapter 4 of [44].

Another important issue is how to choose efficient initialization values and reliable stopping criterion. If $\theta \to 0$, the recovered features x^* are often meaningless due to that interferences and noises are inevitable, and thus we empirically select a lower bound for the feasible space of θ as

$$\theta_L = \frac{m\sigma^2}{4\left\|\Psi^T y\right\|_1} \tag{30}$$

Moreover, the upper bound θ_U of parameter θ is set as $|\Psi^T y|_\infty$ since the solutions $f_\theta(y)$ are always zeros when θ is over θ_U. In addition, a medium accuracy solution of iterative algorithm 1 is often attained after only a few iteration times and thus the iteration number is fixed as $K = 50$. Lastly, the proposed algorithm BiNSOF is stopped if the interval $[\theta_L, \theta_U]$ reduces to a sufficiently precise value which is often set as the order of 10^{-3}.

E. COMPUTATIONAL COMPLEXITY

The main computational cost (CC) of the proposed algorithm BiNSOF is mainly composed of three parts, i.e., sparse estimator solver, MSE estimation and the golden section search. The sparse estimator solver only requires a small number of inner products, vector-scalar multiplications and vector additions, and every iteration needs $O(n)$ or $O(m)$ floating-point operations plus a modest number of multiplications by Ψ and Ψ^T. In order to effectively reduce the overall computational complexity, we only consider the analytic formulation of dictionary Ψ and its explicit dictionary will be a future topic, therefore, performing Ψ or Ψ^T product only needs $O(n)$ or $O(n \log n)$ computational cost, where the reasonability of cost approximation is originated from that redundant dictionary Ψ $(m \le n)$ often provides a more sparse representation expansion. In the case of the soft-thresholding operation, the computational cost is approximate $O(n)$. Therefore, every iteration cost of the Sparse Denoiser Solver from line 2 to line 6 of algorithm 1 is $O(n \log n) + O(m) + O(n)$. As to the estimation procedure of MSE in algorithm 2, fortunately, it doesn't introduce more complex operations besides some linear partial derivations. Therefore, the main cost of sparse denoiser along with MSE estimation from line 2 to line 11 of algorithm 2 is $O(n \log n) + O(m) + O(n)$. Lastly, the cost of golden section search in line 2 of algorithm 3 is $O(1)$. Therefore, the global complexity of the algorithm is

$$\begin{aligned} CC &= O\left((2J+1)K(n \log n + m + n) + JO(1) + O(m)\right) \\ &\approx O\left((2J+1)K(n \log n + m + n)\right) \\ &\approx O\left((2J+1)Kn \log n\right) \end{aligned} \tag{31}$$

where J is the iteration times from line 4 to line 2 of algorithm 3.

In summary, for a wide choice of fast dictionaries with $O(n)$ or $O(n \log n)$ computational cost, the BiNSOF algorithm has $O\left(JKn \log n\right)$ computational complexity.

III. PERFORMANCE ANALYSIS

In this section, extensive computational experiments are implemented to investigate the performance behavior of the proposed algorithm. Impulsive components are always viewed as one type of critical signatures when there is an anomaly in the mechanical systems [11], [45], therefore, fault feature x and the measurements y are designed as follows:

$$\begin{aligned} y &= x + w \\ h(t) &= \exp(-1500t)\sin(2\pi \times 3000t) \\ x(t) &= 2\cos(20\pi t)\sum_k 10 \times h\left(t - \frac{k - 0.1 \times RC}{201}\right) \end{aligned} \tag{32}$$

where 201 Hz is the period of the impulse signals and w is the additive zero-mean white noises with variance σ. RC is a random number sampled from the uniform probability distribution on $[-1, 1]$. The simulation signals are sampled at a frequency of 8192 Hz and have $m = 32768$ points. Moreover, eight different noise variance σ varying from 0.1 to 0.8 in 8 steps, corresponding to SNR from 10.89 dB to -7.18 dB, are employed to study the performance robustness of the BiNSOF under various noise levels.

Moreover, the dictionary Ψ plays an important role in the proposed method and thus how to select an impulse-similarity dictionary becomes a fundamental problem. Once the dictionary Ψ is specified, a sparse estimator is then designed and further obtained feature signals are unique. Among current analytical dictionaries, wavelet dictionary shows excellent performances in sparsely representing the impulsive components, since the morphology of wavelet atoms is similar to the impulsive waveform. It is thus an important task to construct or select one proper wavelet basis for focused fault features, consequently various wavelets have been investigated [46], such as dB family, lifting scheme, multi-wavelets. As a new branch of wavelet family, tunable Q-factor wavelet transform (TQWT) could generate various basis with different oscillating patterns through adjusting three flexible parameters [47], i.e., quality factor Q, redundant factor R and decomposition level J. Therefore, each combination (Q, R, J) is a feasible sparse estimator and then the BiNSOF's performance could be evaluated quantitatively. Every parameter in (Q, R, J) ranges through 5 equispaced points in $[1, 9]$, and only one parameter is considered at one time while other parameters are fixed as initialization values ($Q = 3$, $R = 3, J = 5$).

To profoundly evaluate the performance of the proposed method, some quantitative criterion are designed.

$$\delta_1 = \frac{\left\|\theta^{Exh} - \theta^{Ora}\right\|_2}{\theta^{Ora}} \tag{33}$$

$$\delta_2 = \frac{\left\|\theta^{*} - \theta^{Ora}\right\|_2}{\theta^{Ora}} \tag{34}$$

where θ^{Ora} is the optimal hyper-parameter that achieves the minimum MSE with x is known and serves as a benchmark on the best estimator performance. It is obtained through implementing algorithm 1 with an exhaustive search of θ in one interval from θ_L to $\theta_U$in 100 steps. θ^{Exh} is the optimal hyper-parameter that achieves the minimum $\eta(f_{\hat\theta}(y))$. It is obtained by implementing algorithm 2 with exhaustive search of $\hat\theta$ in the same interval. θ^* is the underlying parameter corresponding to the resulting feature information x^* through the proposed algorithm 3. Therefore, δ_1 evaluates the accuracy of the SURE-based MSE estimation. δ_2 evaluates the effectiveness of the proposed bi-level nested sparse optimization algorithm.

Moreover, the denoising performance of the proposed method is also evaluated by the improvement in SNR (ISNR) calculated as,

$$ISNR = 20\log_{10}\left(\frac{\|y - x\|_2}{\|x^* - x\|_2}\right) \tag{35}$$

Based on these criterions, we now make a quantitative analysis through a series of numerical experiments for various noise levels and parameter settings. The statistical results are tabulated in Tables 1-3 where the superior δ_2 with respect to δ_1 is indicated in bold-face font for all combination of parameter and noise variance, and also the minimum ISNR value among eight tests with only different noise level (i.e., the parameter Q, R, J are specific) is marked with boxes. Several observations are in order:

- Firstly, in most parameter configurations, θ^{Exh} is nearly equal to θ^{Ora}, which demonstrates that $\eta(f_{\hat\theta}(y))$ can accurately approximate the true MSE without knowledge of the real feature signals x.
- Noticeably, the optimal parameters obtained based on the true MSE and estimated MSE, θ^{Ora} and θ^* respectively, are either equal or different only in the second decimal place for nearly all tested cases. This indicates the reliability and robustness of the proposed method. Moreover, it is also interesting to see that δ_2 sometimes is even less than δ_1, highlighted in bold-face font, which means that the proposed BiNSOF can outperform the exhaustive search strategy. This phenomenon can be attributed to that the revolution of θ in $[\theta_L, \theta_H]$ can not as high as possible in the exhaustive search considering the computation complexity. Therefore, it can be asserted that BiNSOF provides an efficient and reliable way to adaptively tune the regularization parameter, and guarantees the recovered feature signals are near-optimal.
- Evidently, the average ISNR is 8 dB and this gain is very noticeable, especially at high noise levels. Moreover, the obtained ISNR from every test is the maximum level and cannot be increased significantly, since the BiNSOF method sufficiently takes advantage of both the sparse structure of feature information and meanwhile the near-optimal parameter configuration of the sparse denoiser.
- Furthermore, the ISNR of each experiment under the same noise level is nearly same with only a slight difference of 0.6 dB. These results highlight that the proposed BiNSOF is robust with TQWT dictionary parameters.
- Surprisingly, under specific dictionary parameter cases, the tendency of ISNR doesn't decrease continuously

TABLE 1. Comparison of feature detection accuracy under five different *Q*-factors and various noise levels.

Q[a]	σ	θ^{Ora}	θ^{Exh}	θ^*	$\delta_1(\%)$[b]	$\delta_2(\%)$[b]	ISNR
1	0.1	0.05278	0.05278	0.05398	0	2.267	13.93
	0.2	0.1214	0.1214	0.1302	0	7.27	10.19
	0.3	0.2102	0.2381	0.1997	13.26	**4.969**	8.889
	0.4	0.329	0.329	0.3301	0	0.3518	$\boxed{8.576}$
	0.5	0.4281	0.4281	0.4421	0	3.269	8.759
	0.6	0.5222	0.5222	0.5058	0	3.151	8.94
	0.7	0.6657	0.6315	0.6466	5.132	**2.873**	9.222
	0.8	0.7741	0.7741	0.7571	0	2.198	9.853
3	0.1	0.05495	0.05495	0.05135	0	6.556	13.39
	0.2	0.1267	0.1267	0.1196	0	5.605	9.428
	0.3	0.1937	0.1937	0.2098	0	8.273	8.029
	0.4	0.297	0.3303	0.3128	11.2	**5.307**	7.623
	0.5	0.3985	0.3985	0.3997	0	0.282	$\boxed{7.611}$
	0.6	0.5262	0.5563	0.5391	5.703	**2.441**	7.942
	0.7	0.6447	0.6447	0.6374	0	1.129	8.381
	0.8	0.8199	0.7527	0.7576	8.204	**7.608**	8.855
5	0.1	0.03522	0.03522	0.04467	0	26.83	12.96
	0.2	0.09997	0.09997	0.1004	0	0.4466	8.934
	0.3	0.1651	0.1651	0.1669	0	1.088	7.364
	0.4	0.2542	0.2542	0.2472	0	2.768	6.86
	0.5	0.321	0.3398	0.339	5.859	**5.595**	$\boxed{6.81}$
	0.6	0.4293	0.4515	0.4516	5.174	5.21	7.047
	0.7	0.547	0.571	0.5303	4.38	**3.054**	7.464
	0.8	0.7228	0.6462	0.6358	10.6	12.04	8.035
7	0.1	0.05022	0.05022	0.04634	0	7.716	12.79
	0.2	0.1061	0.1226	0.1233	15.57	16.26	8.694
	0.3	0.1955	0.1787	0.2002	8.639	**2.373**	7.178
	0.4	0.2737	0.2528	0.2535	7.618	**7.377**	$\boxed{6.626}$
	0.5	0.3837	0.4048	0.3942	5.508	**2.739**	6.634
	0.6	0.5005	0.4788	0.474	4.334	5.299	6.875
	0.7	0.6373	0.7046	0.7471	10.56	17.23	7.318
	0.8	0.9638	0.7966	0.7948	17.35	17.54	7.998
9	0.1	0.04199	0.04199	0.03862	0	8.026	12.37
	0.2	0.09661	0.09661	0.09001	0	6.835	8.229
	0.3	0.1688	0.1688	0.1524	0	9.718	6.651
	0.4	0.2754	0.2754	0.2709	0	1.643	6.183
	0.5	0.3782	0.4025	0.4016	6.408	**6.194**	$\boxed{6.162}$
	0.6	0.5171	0.5171	0.439	0	15.1	6.472
	0.7	0.8	0.6839	1.724	14.51	115.5[c]	7.07
	0.8	1.239	0.887	0.9021	28.4	**27.17**	7.975

[a] Other parameters of wavelet dictionary (R and J) are fixed as 3 and 5, respectively

[b] Relative error indexes are normalized as the percenter ratio to make the resulting difference distinct.

[c] This singular case is originated from that the lower bound θ_L is not very reasonable and thus initial θ_L^0 should be reduced in high noise level cases.

TABLE 2. Comparison of feature detection accuracy under five different *R*-factors and various noise levels.

R[a]	σ	θ^{Ora}	θ^{Exh}	θ^*	$\delta_1(\%)$[b]	$\delta_2(\%)$[b]	ISNR
1.01[c]	0.1	0.04113	0.04113	0.04995	0	21.46	12.8
	0.2	0.1622	0.1622	0.179	0	10.41	9.271
	0.3	0.2849	0.2849	0.2554	0	10.37	8.1431.01
	0.4	0.4123	0.4486	0.4799	8.784	16.37	$\boxed{7.942}$
	0.5	0.6128	0.5674	0.6406	7.415	**4.534**	8.046
	0.6	0.7617	0.8394	0.834	10.2	**9.497**	8.331
	0.7	0.968	0.968	1.018	0	5.116	8.828
	0.8	1.202	1.246	1.274	3.646	5.973	9.232
3	0.1	0.05548	0.05548	0.04197	0	24.35	13.45
	0.2	0.1318	0.1318	0.1205	0	8.534	9.45
	0.3	0.209	0.209	0.1885	0	9.796	8.054
	0.4	0.3035	0.3035	0.3073	0	1.254	$\boxed{7.616}$
	0.5	0.4185	0.4185	0.4243	0	1.385	7.623
	0.6	0.5377	0.5709	0.5173	6.178	**3.789**	7.934
	0.7	0.6531	0.6202	0.67	5.042	**2.58**	8.33
	0.8	0.819	0.9376	0.8015	14.48	**2.141**	8.847
5	0.1	0.05066	0.05066	0.04731	0	6.612	13.67
	0.2	0.1169	0.1169	0.09873	0	15.54	9.583
	0.3	0.1831	0.1831	0.1834	0	0.1648	8.101
	0.4	0.2619	0.2619	0.2583	0	1.39	7.594
	0.5	0.3604	0.3345	0.3352	7.187	**6.99**	$\boxed{7.519}$
	0.6	0.45	0.4823	0.4763	7.189	**5.853**	7.811
	0.7	0.5598	0.5942	0.5626	6.145	**0.4962**	8.163
	0.8	0.7102	0.7463	0.6538	5.095	7.934	8.633
7	0.1	0.04775	0.04775	0.04553	0	4.646	13.67
	0.2	0.1068	0.1068	0.1066	0	0.2328	9.453
	0.3	0.1787	0.1787	0.2044	0	14.38	7.842
	0.4	0.2679	0.2942	0.2873	9.841	**7.239**	7.33
	0.5	0.3386	0.3649	0.3806	7.752	12.39	$\boxed{7.229}$
	0.6	0.4555	0.4555	0.4567	0	0.2701	7.49
	0.7	0.5725	0.6011	0.5313	4.996	7.193	7.867
	0.8	0.7136	0.7819	0.781	9.57	**9.432**	8.355
9	0.1	0.04742	0.04742	0.04648	0	1.989	13.54
	0.2	0.1058	0.1058	0.09916	0	6.244	9.344
	0.3	0.1777	0.1777	0.1756	0	1.229	7.732
	0.4	0.2593	0.2846	0.274	9.737	**5.67**	7.177
	0.5	0.3551	0.3829	0.3799	7.84	**6.992**	$\boxed{7.056}$
	0.6	0.4474	0.4474	0.4232	0	5.417	7.274
	0.7	0.5859	0.5565	0.5403	5.023	7.79	7.7
	0.8	0.7264	0.6376	0.6957	12.23	**4.22**	8.227

[a] Other parameters of wavelet dictionary (Q and J) are fixed as 3 and 5, respectively

[b] Relative error indexes are normalized as the percenter ratio to make the resulting difference distinct.

[c] The lower band value of redundant factor R is set as 1.01 because $R = 1$ cannot be accessible.

as the noise variance σ increases, but achieves the minimum point (marked in the boxes) at one moderate noise level and then increases regardless of the noise energy. This phenomenon could be explained as follows. To keep more feature information in the extracted features x^*, noises are introduced inevitably and a good balance between the estimator bias and noise variance should be achieved. When the noise variance is small or medium, the obtained x^* includes a small number of noises and generates uniform bias for the whole feature signals x, the ISNR thus decreases. As the noise level increases, the details of x are submerged

TABLE 3. Comparison of feature detection accuracy under five different *J*-factors and various noise levels.

J[a]	σ	θ^{Ora}	θ^{Exh}	θ^{*}	$\delta_1(\%)$[b]	$\delta_2(\%)$[b]	ISNR
1	0.1	0.05786	0.05786	0.07207	0	24.56	13.41
	0.2	0.1603	0.1603	0.1698	0	5.965	9.557
	0.3	0.2644	0.2644	0.2684	0	1.525	8.139
	0.4	0.3815	0.4114	0.4017	7.835	**5.301**	$\boxed{7.686}$
	0.5	0.5344	0.5041	0.4829	5.685	9.644	7.839
	0.6	0.6527	0.6223	0.6326	4.656	**3.079**	8.065
	0.7	0.8101	0.8101	0.7813	0	3.548	8.503
	0.8	0.956	0.9893	0.9786	3.486	**2.359**	8.976
3	0.1	0.05763	0.05763	0.05938	0	3.042	13.77
	0.2	0.1225	0.1469	0.141	19.85	**15.07**	9.753
	0.3	0.2145	0.2145	0.2205	0	2.815	8.335
	0.4	0.2928	0.2928	0.3107	0	6.11	$\boxed{7.84}$
	0.5	0.4058	0.4058	0.4006	0	1.277	7.89
	0.6	0.5214	0.487	0.4865	6.591	6.68	8.132
	0.7	0.6545	0.6195	0.6278	5.34	**4.069**	8.557
	0.8	0.7811	0.7417	0.7247	5.04	7.224	9.07
5	0.1	0.05434	0.05434	0.04988	0	8.194	13.46
	0.2	0.1273	0.1016	0.1183	20.22	**7.08**	9.368
	0.3	0.2022	0.2022	0.197	0	2.554	7.979
	0.4	0.3048	0.3048	0.2954	0	3.106	$\boxed{7.567}$
	0.5	0.3961	0.4307	0.4106	8.733	**3.659**	7.599
	0.6	0.5341	0.5671	0.5094	6.177	**4.619**	7.947
	0.7	0.6679	0.6679	0.6675	0	0.05732	8.329
	0.8	0.8081	0.7346	0.7538	9.092	**6.718**	8.836
7	0.1	0.05628	0.0318	0.04187	43.51	**25.61**	13.22
	0.2	0.09754	0.09754	0.1075	0	10.19	9.395
	0.3	0.1749	0.1749	0.1682	0	3.82	8.093
	0.4	0.2677	0.2937	0.2918	9.743	**9.018**	$\boxed{7.82}$
	0.5	0.3838	0.3498	0.3505	8.853	**8.673**	7.906
	0.6	0.4758	0.4758	0.4942	0	3.869	8.133
	0.7	0.5882	0.5882	0.5976	0	1.591	8.629
	0.8	0.7369	0.7369	0.8012	0	8.723	9.17
9	0.1	0.0306	0.0306	0.03997	0	30.64	12.98
	0.2	0.09741	0.1224	0.1208	25.7	**24.06**	9.294
	0.3	0.1797	0.1797	0.1804	0	0.3876	8.061
	0.4	0.2666	0.2666	0.2666	0	0.001854	$\boxed{7.752}$
	0.5	0.3646	0.3646	0.3633	0	0.3566	7.836
	0.6	0.476	0.5106	0.498	7.262	**4.627**	8.184
	0.7	0.6223	0.7003	0.7011	12.55	12.67	8.516
	0.8	0.7407	0.7049	0.6763	4.824	8.689	9.005

[a] Other parameters of wavelet dictionary (Q and R) are fixed as 3 and 3, respectively

[b] Relative error indexes are normalized as the percenter ratio to make the resulting difference distinct.

completely and only its dominant parts could be preserved, the bias of dominant parts is thus reduced to achieve the minimum objective cost (10), and consequently, the ISNR index begins to increase. From this analysis, it can be found that the proposed BiNSOF algorithm could effectively search the optimal bias-variance trade-off and thus the feature detection procedure is adaptive.

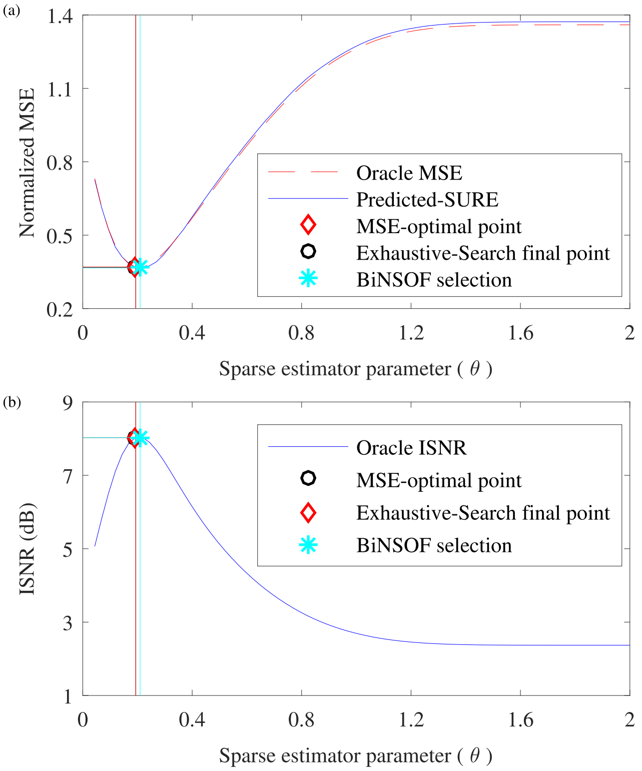

FIGURE 1. Plot of MSE and ISNR as functions of sparse denoiser parameter θ. (a) evolution of two types of MSE and three minimum points based on different schemes, (b) evolution of oracle ISNR and three maximum SNR points based on different schemes. These curves are adopted from the eleven row of Tables 1 corresponding to the $\sigma = 0.3$ condition, and the dictionary parameters are $Q = 3, R = 3, J = 5$. The MSE in (a) is normalized through dividing by the factor $m\sigma^2$. Moreover, the predicted-SURE curve denotes the 100 MSE sequences that originated from the exhaustive-search method. These plots demonstrate that Predicated-SURE closely captures the trend of Oracle-MSE and the minima of BiNSOF selection are close to that of the oracle points indicated by the solid vertical lines.

To further highlight the characteristic of the proposed feature detection method, the evolution of two criterions (MSE and ISNR) versus the sparse denoiser parameter θ under three different (Q, R, J) settings are illustrated in Fig. 1, Fig. 2, Fig. 3, respectively. It is observed that the SURE-based MSE estimator (8) follows the true MSE curve remarkably well in all the cases and is indeed a good estimator to approximate the oracle MSE. Moreover, the resulting ISNR of the proposed BiNSOF is sufficiently closed to the corresponding oracle ISNR, which illustrates that the BiNSOF has an excellent capability in maximally detecting feature information for fault diagnosis.

In summary, the proposed BiNSOF algorithm provides a near-optimal sparse estimator for MSE and thus it is indeed a good objective to maximize the ISNR. Moreover, the BiNSOF algorithm could adaptively strike a good balance between the estimator bias and noise variance and thus achieves the maximum gain of ISNR, which yields an adaptive feature detection technique. Lastly, the BiNSOF algorithm is insensitive to the TQWT dictionary parameters.

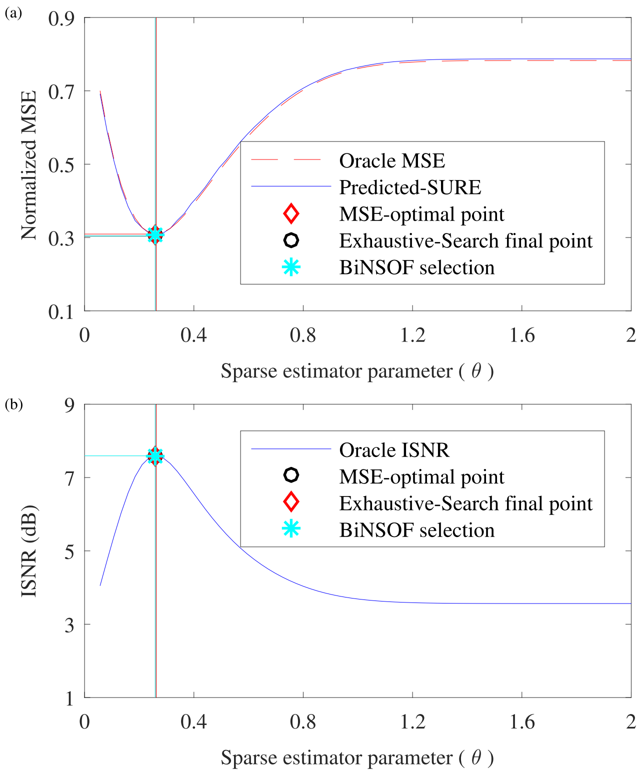

FIGURE 2. **Plot of MSE and ISNR as functions of sparse denoiser parameter θ. (a) evolution of two types of MSE and three minimum points based on different schemes, (b) evolution of oracle ISNR and three maximum SNR points based on different schemes. These curves are adopted from the twenty row of Tables 2 corresponding to the $\sigma = 0.4$ condition, and the dictionary parameters are $Q = 3, R = 5, J = 5$. The MSE in (a) is normalized through dividing by the factor $m\sigma^2$. Moreover, the predicted-SURE curve denotes the 100 MSE sequences that originated from the exhaustive-search method. These plots demonstrate that Predicated-SURE closely captures the trend of Oracle-MSE and the minima of BiNSOF selection are close to that of the oracle points indicated by the solid vertical lines.**

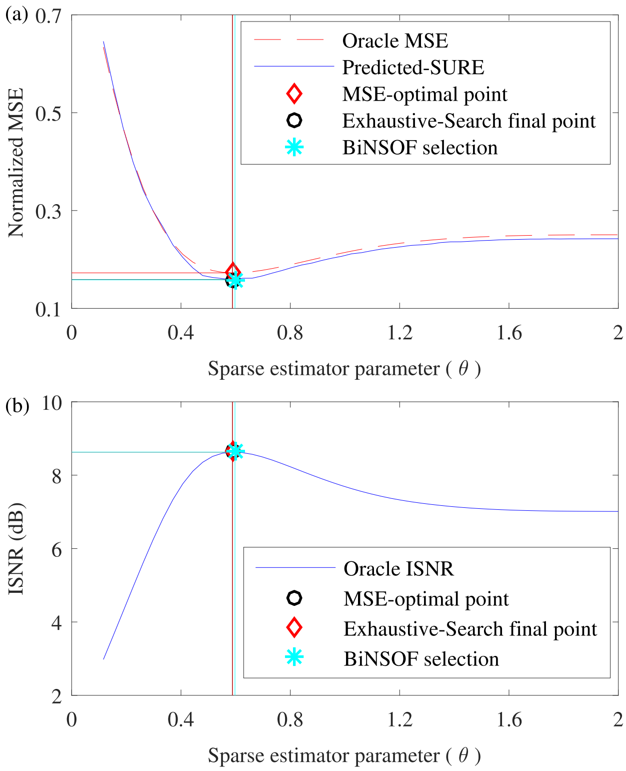

FIGURE 3. **Plot of MSE and ISNR as functions of sparse denoiser parameter θ. (a) evolution of two types of MSE and three minimum points based on different schemes, (b) evolution of oracle ISNR and three maximum SNR points based on different schemes. These curves are adopted from the thirty-one row of Tables 3 corresponding to the $\sigma = 0.7$ condition, and the dictionary parameters are $Q = 3, R = 3, J = 7$. The MSE in (a) is normalized through dividing by the factor $m\sigma^2$. Moreover, the predicted-SURE curve denotes the 100 MSE sequences that originated from the exhaustive-search method. These plots demonstrate that Predicated-SURE closely captures the trend of Oracle-MSE and the minima of BiNSOF selection are close to that of the oracle points indicated by the solid vertical lines.**

IV. EXPERIMENTAL VALIDATION

A. CASE 1

In this section, a bearing vibration data collected from the data center of Case Western Reserve University (CWRU) is used to evaluate the proposed method. This dataset is publicly available and widely used. Therefore, a thorough examination of the dataset was performed by Smith and Randall [48]. Every data record has been studied and categorized into three groups, labeled as "Y", "P" and "N" respectively. "Y" means that the data is diagnosable, "P" denotes that the feature is weak and the data is probable or potentially diagnosable and "N" is labeled to data which can not be diagnosable. The data sets in the "P" or "N" categories are suggested for validation of newly developed algorithms. Consequently, one data record labeled as "P" is taken. The experimental bearing is SKF deep groove ball bearing 6203-2RS JEM and its characteristic frequencies are presented in Table 4. The vibration signal from one bearing with a 0.36-mm-diameter fault on the inner race was collected with a sampling frequency of 12 kHz. The shaft rotational speed is 1730 rev/min. The shaft rotation frequency R and the ball pass frequency in the inner race (BPFI) are 28.83 Hz and 142.64 Hz, respectively.

TABLE 4. **Basic characteristic frequencies of different rolling bearings components for 6203 − 2*RSJEM* type.**

Fault patterns	Relative frequency[a]
Ball pass frequency at outer ring (BPFO)	7.643
Ball pass frequency at inner ring (BPFI)	4.9469
Fundamental train frequency (FTF)	0.3817
Ball spin frequency (BSF)	3.9874

[a] Relative frequency indicates that the characteristic frequency is normalized through dividing by the rotational speed.

One segment of vibration signals and its corresponding spectrum are illustrated in Fig. 4. The time view of the signal shown in Fig. 4(a) does not reveal fault symptoms due to high-level noises and fault frequency bins are also indistinct in the frequency spectrum. Moreover, only rotation frequency R and its harmonics could be found in the envelope spectrum. Moreover, the characteristic frequency BPFI is

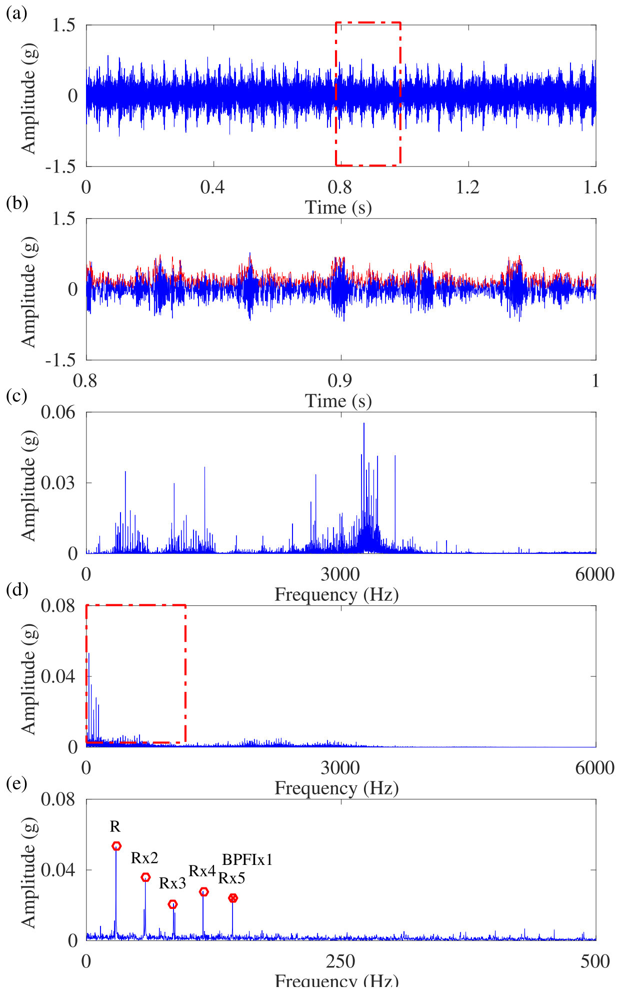

FIGURE 4. Vibration signals of the fault bearing and its corresponding spectrums: (a) original waveforms, (b) zoom-in view of waveforms with its corresponding envelope components, (c) frequency spectrum, (d) envelope spectrum of original signals, (e) zoom-in view of envelope spectrum indicated by the dotted rectangle. Moreover, R denotes the rotational speed.

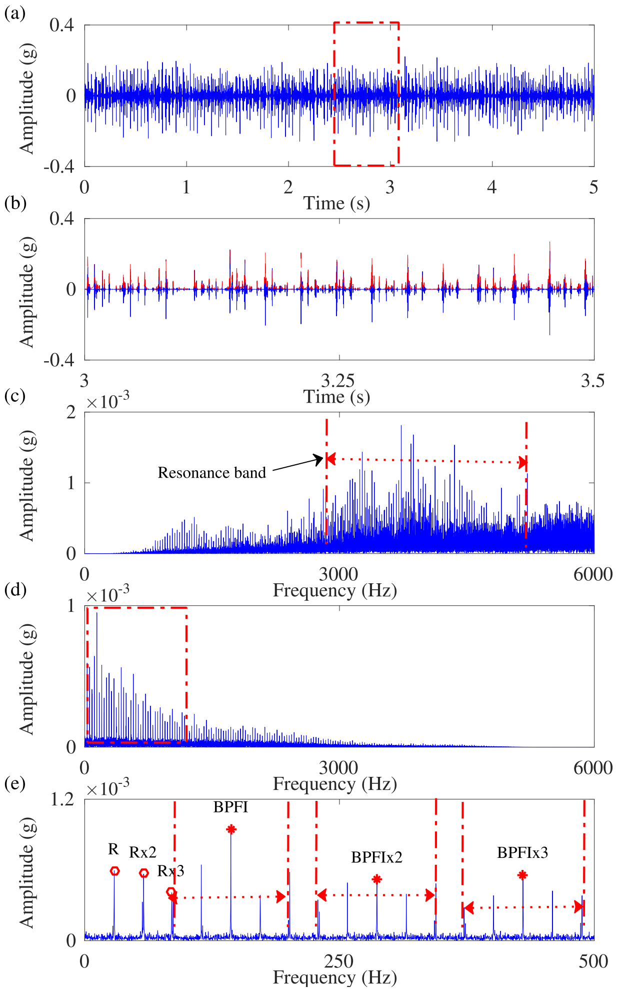

FIGURE 5. Feature signals extracted through the proposed BiNSOF and its corresponding spectrums: (a) waveforms of feature signals, (b) zoom-in view of waveforms and corresponding envelope components, (c) frequency spectrum, (d) envelope spectrum of feature signals, (e) zoom-in view of envelope spectrum indicated by the polygon. Resonance band indicates the critical region where feature information concentrates. BPFI × 1 is the inner race fault frequency of bearings. The interval period in the vicinity of BPFI and its multiplies (BPFI × 1 to 3) is the rotational speed R, which is displayed in (e).

nearly coincided with the 5-th order of rotation frequency. Thus, there is no significant feature information to perform diagnosis. Due to strong rotational frequency bins, the pre-whited strategy is firstly adopted and the proposed method BiNSOF is then applied to detect fault information. Due to that tunable Q-factor wavelet transform (TQWT) could generate various basis with different oscillating patterns through adjusting three flexible parameters [47], i.e., quality factor Q, redundant factor R and decomposition level J, sparse dictionary Ψ is set as TQWT with parameters $(Q, R, J) = (3, 3, 5)$. The resulting feature signals and its spectrum are depicted in Fig. 5. It can be seen from time waveforms that feature signals are mainly composed of quasi-periodic impulses and furthermore are modulated by low-frequency components. As seen from the spectrum of Fig. 5(c), the resonance band of feature signals is extracted and noises are alleviated. Compared with the results of Fig. 5(d-e) with Fig. 4(d-e), the envelope spectrum of extracted feature signals clearly indicates the characteristic frequency 142.64 Hz (i.e., BPFI × 1 $= 28.83 \times 4.9469$ Hz) and its multiples (BPFI × 2 and ×3) with the side-bands composed of equispaced modulated frequency 28.83 Hz. According to the fault mechanism [49], the detected feature patterns thus demonstrate that a localized fault exists on the inner race of the rolling bearings.

To gain more insight into the effectiveness of BiNSOF, the evolution of estimated MSE and the optimal points are depicted in Fig. 6. The MSE curve demonstrates the relationship between the estimator bias and noise variance and this is a popular phenomenon in the signal processing community. However, it is a very difficult problem to achieve an equilibrium point, and fortunately, the resulting point obtained through BiNSOF is remarkably near to desired values, which validates the effectiveness of the proposed feature detection algorithm.

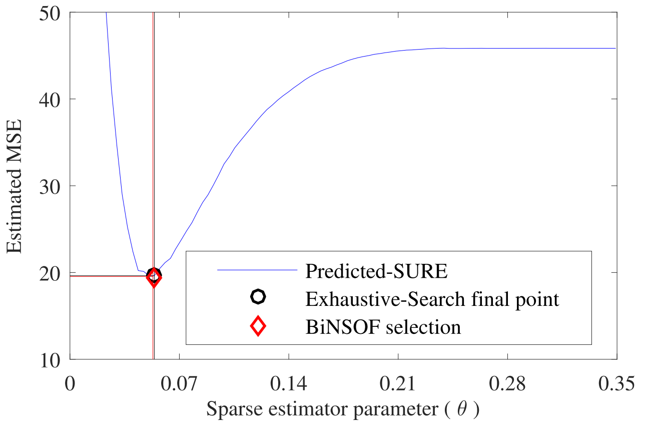

FIGURE 6. Evolution of MSE versus sparse estimator parameter θ. The estimated MSE is calculated based on the formula (8). Moreover, the predicted-SURE curve denotes the 100 MSE sequences that originated from the exhaustive-search method. Strikingly, the minima of BiNSOF selection is close to the empirical minimal MSE point indicated by the solid vertical line.

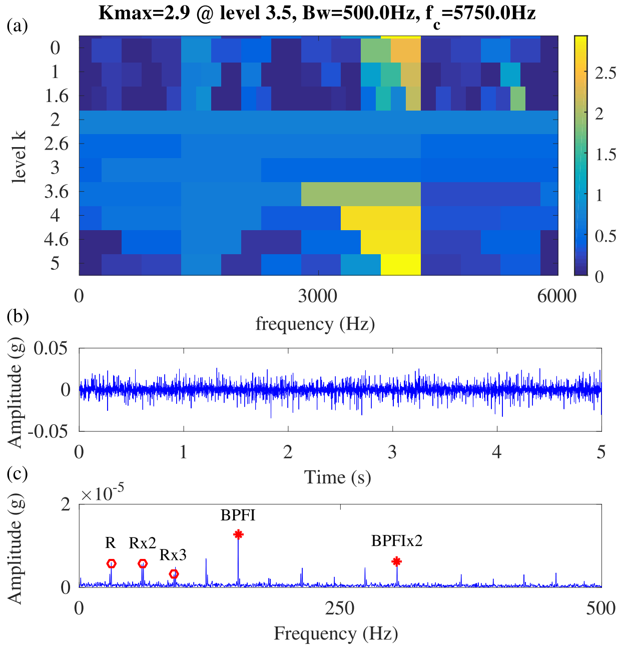

FIGURE 7. The fault information obtained by the SK technique: (a) Fast Kurtogram of the original signals, (b) filtered signals according to the optimal filter with the center frequency 1520 Hz and the bandwidth 160 Hz, (c) envelope spectrum of the filtered signals. R denotes the rotational speed and BPFI is the characteristic frequency of faults located in the inner race of rolling bearings.

Moreover, spectral kurtosis [46], one of state-of-the-art fault diagnosis techniques, is introduced to detect the feature information from the same vibration signals. The results are illustrated in Fig. 7. As can be seen, the fault information of rolling bearings is not significant and its multiples cannot be detected. Meanwhile, compared with the results shown in Fig. 5(e), the energy of BPFI depicted in Fig. 7(c) is weak and not easy to identify, which is originated from the fact that most of feature information is discarded to remove the noises. Therefore, the SK technique is inferior to extract fault signals compared with the BiNSOF method.

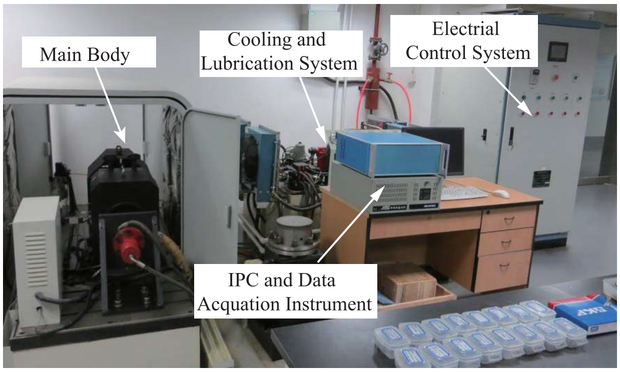
Main Body
Cooling and Lubrication System
Electrial Control System
IPC and Data Acquation Instrument

FIGURE 8. The aero-engine bearing test rig.

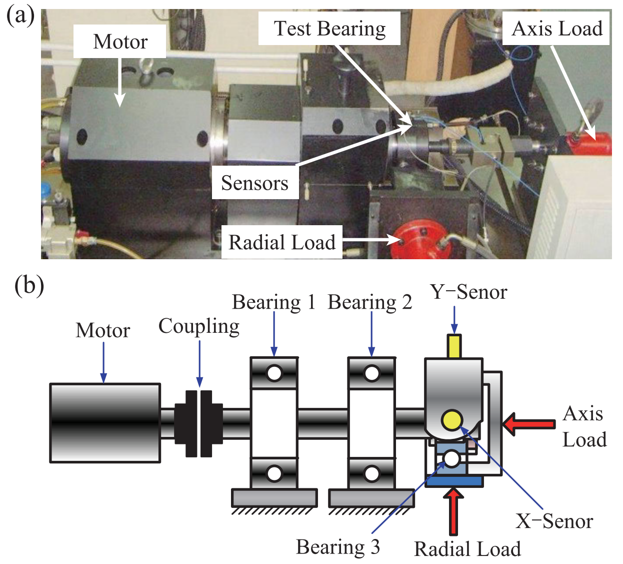

FIGURE 9. The internal structure of the aero-engine bearing test rig.

B. CASE 2

In this section, we use the data generated from the aero-engine bearing fault test to investigate the effectiveness of BiNSOF. The photo of the test rig is shown in Fig. 8. It consists of the main body of the test rig, the cooling and lubrication system, the industrial personal computer (IPC), and the data acquisition system. The temperature, load and the rotational speed of the test rig are controlled by the IPC. The internal structure of the test rig is also shown in Fig. 9. The spindle of the test rig is driven by the high-speed motor. The bearing 3 is the test bearing and the bearing 1 and the bearing 2 are two support bearings. The axial load and radial load are simultaneously applied to the test bearing by the lubrication system. During one accelerated life test, a local defect on the outer raceway is detected, which is shown in Fig. 10. The type of the test bearing is H7015. The patch diameter, the ball diameter, the angle of contact and the number of balls are

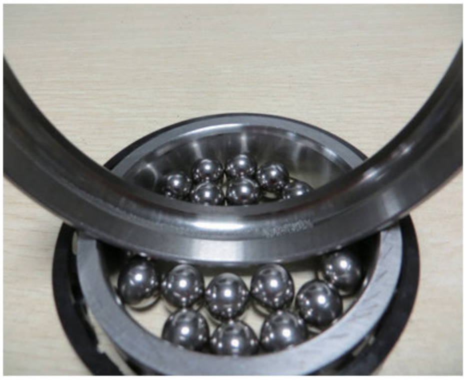

FIGURE 10. The local fault on the outer raceway.

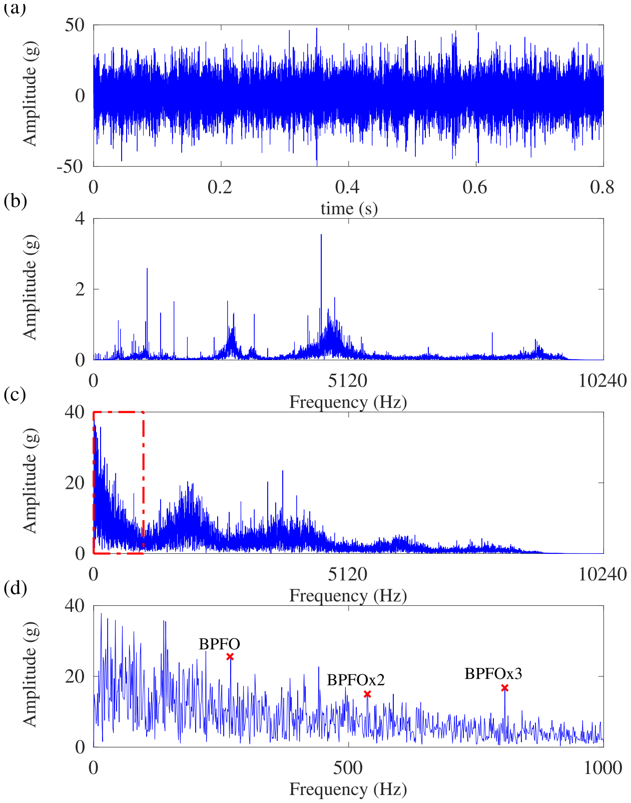

FIGURE 11. Vibration signals of the fault bearing and its corresponding spectrums: (a) original waveforms, (b)frequency spectrum, (c) envelope spectrum of original signals, (d) zoom-in view of envelope spectrum indicated by the dotted rectangles.

95mm, 12.65mm, 15° and 19 respectively. The characteristic frequency of the outer race under running speed of 2000 r/min is 275.94Hz. Vibration signals are sampled at 20.48Hz by a data acquisition system.

One segment of the vibration signal with a length of 0.8s is adopted for basic spectrum analysis, which is shown in Fig. 11. However, fault information cannot be found and our desired feature component BPFO is fully submerged into various noises and interferences. To highlight the feature information, a band-pass filter with center frequency 2700Hz and bandwidth 2000Hz is designed and its filtered results are shown in Fig. 12. However, feature signals still cannot be recognized. To further mitigate the discrete harmonic components, a pre-whited operation is then performed to the filtered signals and then the proposed BiNSOF is applied to extract fault information. The TQWT transform with parameters $(Q, R, J) = (2, 6, 19)$ is adopted as the sparse dictionary. The extracted feature signal and its spectrum are depicted in Fig. 13. Compared with the results of Fig. 12(c) and (d) with Fig. 13(c) and (d), the fault characteristic frequency BPFO and its higher orders are significantly highlighted. Moreover, discrete interference frequencies and strong noises are well eliminated.

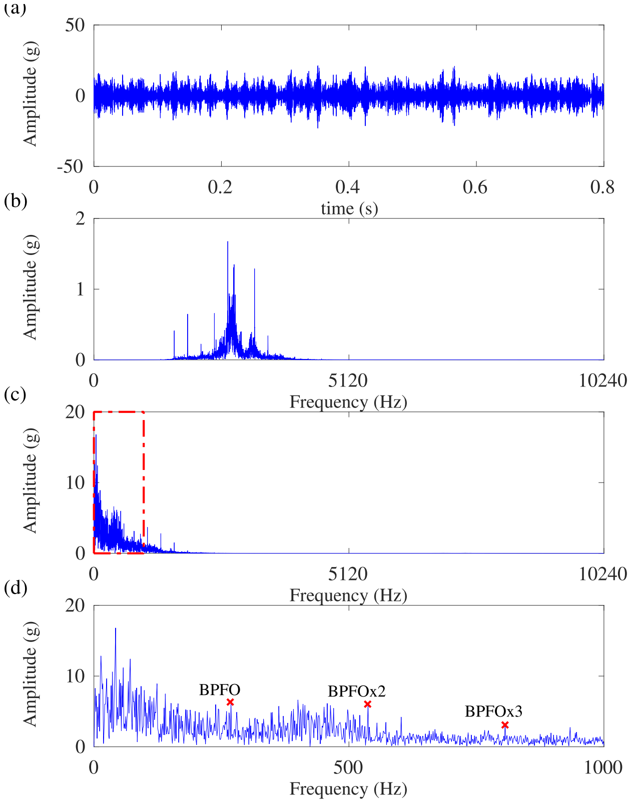

FIGURE 12. Filtered signals of the fault bearing and its corresponding spectrums: (a) time waveforms, (b) frequency spectrum, (c) envelope spectrum of original signals, (d) zoom-in view of envelope spectrum indicated by the dotted rectangle.

For comparison, SK is also used to analyze the same signal and its results are shown in Fig. 14. It can be seen that the center frequency and bandwidth of the optimal filter are 3840Hz and 2560Hz, respectively. Moreover, feature information cannot be recognized from the spectrum of the envelope signal. Therefore, the effectiveness of the SK algorithm is unsatisfactory compared with the proposed BiNSOF.

In addition, the above two cases are performed under Windows 10 and MATLAB 2016*b* running on a computer equipped with an Intel Core 7 CPU at 2.93 GHz and 16 GB of RAM. The running time of the proposed method is 66.73s

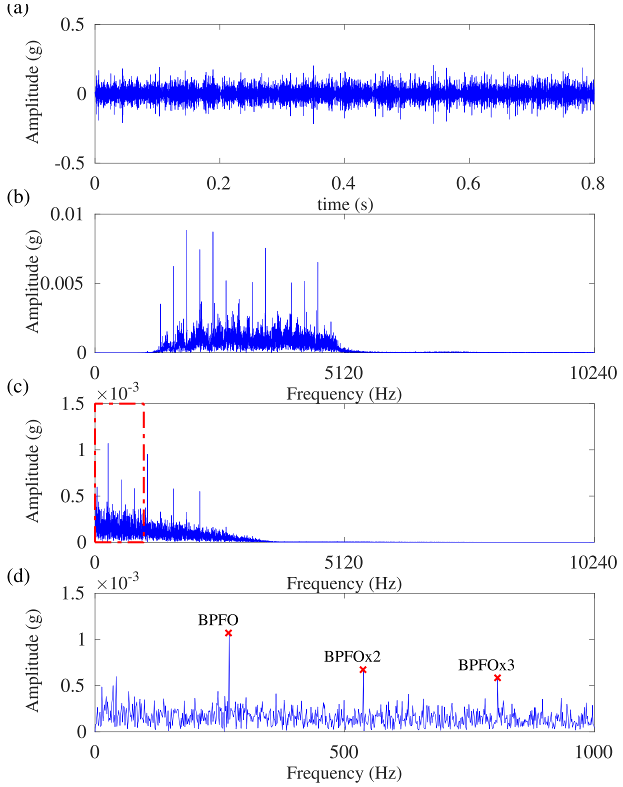

FIGURE 13. Feature signals extracted through the proposed BiNSOF: (a) waveforms of feature signals, (b) frequency spectrum, (c) envelope spectrum of feature signals, (d) zoom-in view of envelope spectrum. Resonance band indicates the critical region where feature information concentrates. BPFO × 1 is the outer race fault frequency of bearings.

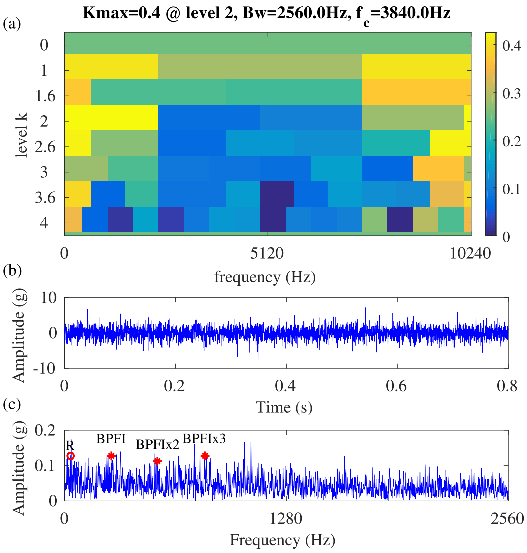

FIGURE 14. The fault information obtained by the SK technique: (a) Fast Kurtogram of the original signals, (b) filtered signals according to the optimal filter with the center frequency 3840 Hz and the bandwidth 2560 Hz, (c) envelope spectrum of the filtered signals.

with the signal length of 60000 and 27.95s with the signal length of 16283, which is acceptable for offline condition monitoring Systems.

V. CONCLUSION

In this paper, a SURE based bi-level nested sparse optimization framework (BiNSOF) is proposed for adaptive fault feature recognition from its noisy collections. The proposed BiNSOF jointly optimizes the ℓ_1 regularized sparse estimator as well as its hyper-parameter. Its highlights is to almost perfectly approximate the oracle MSE without any real feature information, and meanwhile provide a reliable way to adaptively confirm optimal hyper-parameter sets for the maximum performance gains of fault detection. The BiNSOF, in essence, adopts a bi-level nested optimization strategy with an inner primal-dual splitting scheme and an outer golden section search scheme. Moreover, an iterative procedure is proposed to gradually calculate the divergence of sparse estimator while this procedure incurs no significantly computational cost, which guarantees the effectiveness of MSE estimator. The BiNSOF's convergence is evident based on convex optimization theory and its computational complexity has the same order of popular first-order algorithms. Extensive numeric simulation shows that the BiNSOF algorithm could reliably track the trend of oracle MSE and rapidly achieve the minimum MSE point to extract near-optimal feature signals. Moreover, the BiNSOF's performance is robust with dictionary types and noise levels, which significantly boosts its generalization capability. Lastly, the effectiveness and applicability of the BiNSOF are demonstrated through applying it to two bearing fault data sets from CWRU benchmark bearing data and aero-engine bearing test data. All fault detection results confirm that the BiNSOF outperforms state-of-the-art fault detection techniques.

By product, the inner sparse estimator of BiNSOF is not restricted to ℓ_1 regularized sparse estimator, various recently developed sparse estimators, such as weighted sparse estimator [22], enhanced sparse period-group lasso estimator [28] and generalized minimax convex penalty based sparse estimator [29] can also be embedded into BiNSOF to adaptively attain optimal hyper-parameters as well as achieve its near-optimal performances. Moreover, it is a possible direction for future research to extend the applicability of our current adaptive fault diagnosis technique to other industrial systems.

REFERENCES

[1] Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757–3767, Jun. 2015.

[2] Z. Gao, C. Cecati, and S. Ding, "A survey of fault diagnosis and fault-tolerant techniques—Part II: Fault diagnosis with knowledge-based and hybrid/active approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3768–3774, Jun. 2015.

[3] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 42, nos. 1–2, pp. 314–334, Jan. 2014.

[4] Y. Yuan, X. Liu, S. Ding, and B. Pan, "Fault sdetection and location system for diagnosis of multiple faults in aeroengines," *IEEE Access*, vol. 5, pp. 17671–17677, 2017.

[5] M. Riera-Guasp, J. Antonino-Daviu, and G.-A. Capolino, "Advances in electrical machine, power electronic, and drive condition monitoring and fault detection: State of the art," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1746–1759, Mar. 2015.

[6] S. Yin, S. Ding, X. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6418–6428, Nov. 2014.

[7] D. Wang, K.-L. Tsui, and Q. Miao, "Prognostics and health management: A review of vibration based bearing and gear health indicators," *IEEE Access*, vol. 6, pp. 665–676, 2018.

[8] L. Ciabattoni, F. Ferracuti, A. Freddi, and A. Monteriu, "Statistical spectral analysis for fault diagnosis of rotating machines," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4301–4310, May 2018.

[9] S. Wang, X. Chen, I. W. Selesnick, Y. Guo, C. Tong, and X. Zhang, "Matching synchrosqueezing transform: A useful tool for characterizing signals with fast varying instantaneous frequency and application to machine fault diagnosis," *Mech. Syst. Signal Process.*, vol. 100, pp. 242–288, Feb. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888327017303734

[10] N. Su, X. Li, and Q. Zhang, "Fault diagnosis of rotating machinery based on wavelet domain denoising and metric distance," *IEEE Access*, vol. 7, pp. 73262–73270, 2019.

[11] V. Leite, J. B. da Silva, G. C. Veloso, L. B. da Silva, G. Lambert-Torres, E. Bonaldi, and L. De L. de Oliveira, "Detection of localized bearing faults in induction machines by spectral kurtosis and envelope analysis of stator current," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1855–1865, Mar. 2015.

[12] X. Yu, F. Dong, E. Ding, S. Wu, and C. Fan, "Rolling bearing fault diagnosis using modified LFDA and EMD with sensitive feature selection," *IEEE Access*, vol. 6, pp. 3715–3730, 2018.

[13] D. Abboud, M. Elbadaoui, W. A. Smith, and R. B. Randall, "Advanced bearing diagnostics: A comparative study of two powerful approaches," *Mech. Syst. Signal Process.*, vol. 114, pp. 604–627, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888327018302619

[14] H. Qiao, T. Wang, P. Wang, L. Zhang, and M. Xu, "An adaptive weighted multiscale convolutional neural network for rotating machinery fault diagnosis under variable operating conditions," *IEEE Access*, vol. 7, pp. 118954–118964, 2019.

[15] Z. Meng, X. Guo, Z. Pan, D. Sun, and S. Liu, "Data segmentation and augmentation methods based on raw data using deep neural networks approach for rotating machinery fault diagnosis," *IEEE Access*, vol. 7, pp. 79510–79522, 2019.

[16] Z. Feng, Y. Zhou, M. J. Zuo, F. Chu, and X. Chen, "Atomic decomposition and sparse representation for complex signal analysis in machinery fault diagnosis: A review with examples," *Measurement*, vol. 103, pp. 106–132, Jun. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0263224117301252

[17] L. Cui, J. Wang, and S. Lee, "Matching pursuit of an adaptive impulse dictionary for bearing fault diagnosis," *J. Sound Vib.*, vol. 333, no. 10, pp. 2840–2862, May 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0022460X13010729

[18] Y. Qin, "A new family of model-based impulsive wavelets and their sparse representation for rolling bearing fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 65, no. 3, pp. 2716–2726, Mar. 2018.

[19] Y. Qin, J. Zou, and F. Cao, "Adaptively detecting the transient feature of faulty wind turbine planetary gearboxes by the improved kurtosis and iterative thresholding algorithm," *IEEE Access*, vol. 6, pp. 14602–14612, 2018.

[20] H. Wang, P. Wang, L. Song, B. Ren, and L. Cui, "A novel feature enhancement method based on improved constraint model of online dictionary learning," *IEEE Access*, vol. 7, pp. 17599–17607, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8630917/

[21] H. Zhang, X. Chen, Z. Du, and B. Yang, "Sparsity-aware tight frame learning with adaptive subspace recognition for multiple fault diagnosis," *Mech. Syst. Signal Process.*, vol. 94, pp. 499–524, Sep. 2017.

[22] H. Zhang, X. Chen, Z. Du, and R. Yan, "Kurtosis based weighted sparse model with convex optimization technique for bearing fault diagnosis," *Mech. Syst. Signal Process.*, vol. 80, pp. 349–376, Dec. 2016.

[23] R. Sun, Z. Yang, X. Chen, S. Tian, and Y. Xie, "Gear fault diagnosis based on the structured sparsity time-frequency analysis," *Mech. Syst. Signal Process.*, vol. 102, pp. 346–363, Mar. 2018.

[24] Z. Du, X. Chen, H. Zhang, R. Yan, and W. Yin, "Learning collaborative sparsity structure via nonconvex optimization for feature recognition," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4417–4430, Oct. 2018.

[25] J. Ding, "Fault detection of a wheelset bearing in a high-speed train using the shock-response convolutional sparse-coding technique," *Measurement*, vol. 117, pp. 108–124, Mar. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0263224117307819

[26] Q. Li and S. Y. Liang, "Multiple faults detection for rotating machinery based on bicomponent sparse low-rank matrix separation approach," *IEEE Access*, vol. 6, pp. 20242–20254, 2018.

[27] W. He, Y. Ding, Y. Zi, and I. W. Selesnick, "Sparsity-based algorithm for detecting faults in rotating machines," *Mech. Syst. Signal Process.*, vols. 72–73, pp. 46–64, May 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888327015005440

[28] Z. Zhao, S. Wu, B. Qiao, S. Wang, and X. Chen, "Enhanced sparse period-group lasso for bearing fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 66, no. 3, pp. 2143–2153, Mar. 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," *Mech. Syst. Signal Process.*, vol. 127, pp. 202–233, Jul. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888327019301451

[30] H. Zhang, X. Chen, W. Chen, and Z. Shen, "Collaborative sparse classification for aero-engine's gear hub crack diagnosis," *Mech. Syst. Signal Process.*, to be published, doi: 10.1109/TTHZ.2016.2544142.

[31] Y. Eldar, "Generalized SURE for exponential families: Applications to regularization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 471–481, Feb. 2009.

[32] S. Vaiter, C.-A. Deledalle, G. Peyré, C. Dossal, and J. Fadili, "Local behavior of sparse analysis regularization: Applications to risk estimation," *Appl. Comput. Harmon. Anal.*, vol. 35, no. 3, pp. 433–451, Nov. 2013.

[33] S. Vaiter, C. Deledalle, J. Fadili, G. Peyré, and C. Dossal, "The degrees of freedom of partly smooth regularizers," *Ann. Inst. Stat. Math.*, vol. 69, no. 4, pp. 791–832, Aug. 2017.

[34] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal.*, vol. 31, no. 1, pp. 59–73, Jul. 2011.

[35] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, no. 6, pp. 1135–1151, Nov. 1981.

[36] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statistical Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.

[37] D. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[38] M. Hashemi and S. Beheshti, "Adaptive noise variance estimation in bayesshrink," *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 12–15, Jan. 2010.

[39] S. Pyatykh, J. Hesser, and L. Zheng, "Image noise level estimation by principal component analysis," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 687–699, Feb. 2013.

[40] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, May 2011.

[41] D. Davis, "Convergence rate analysis of primal-dual splitting schemes," *SIAM J. Optim.*, vol. 25, no. 3, pp. 1912–1943, Jan. 2015.

[42] R. Giryes, M. Elad, and Y. Eldar, "The projected GSURE for automatic parameter tuning in iterative shrinkage methods," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 3, pp. 407–422, May 2011.

[43] C.-A. Deledalle, S. Vaiter, J. Fadili, and G. Peyre, "Stein unbiased gradient estimator of the risk (SUGAR) for multiple parameter selection," *SIAM J. Imag. Sci.*, vol. 7, no. 4, pp. 2448–2487, 2014.

[44] A. Antoniou and W. S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. New York, NY, USA: Springer, 2007, pp. 92–94.

[45] X. Chen, Z. Du, J. Li, X. Li, and H. Zhang, "Compressed sensing based on dictionary learning for extracting impulse components," *Signal Process.*, vol. 96, pp. 94–109, Mar. 2014.

[46] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: A review with applications," *Signal Process.*, vol. 96, pp. 1–15, Mar. 2014.

[47] I. W. Selesnick, "Wavelet transform with tunable Q-factor," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3560–3575, Aug. 2011.

[48] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the case western reserve University data: A benchmark study," *Mech. Syst. Signal Process.*, vols. 64–65, pp. 100–131, Dec. 2015.

[49] R. B. Randall, *Vibration-Based Condition Monitoring: Industrial, Aerospace and Automotive Applications*. Hoboken, NJ, USA: Wiley, 2011.

HAN ZHANG received the Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2017. She is currently a Lecturer with the Department of Mechanical and Electronic Engineering, School of Construction Machinery, Chang'an University, Xi'an. Her current research interests include sparse signal representation, convex optimization, and low-rank matrix factorization for mechanical fault diagnosis.

XUEFENG CHEN (Member, IEEE) received the Ph.D. degree from Xi'an Jiaotong University, China, in 2004. He is currently a Full Professor and the Dean of the School of Mechanical Engineering, Xi'an Jiaotong University. He has authored over 100 SCI publications in the areas of composite structure, aeroengine, wind power equipment, and so on. He is also a member of the ASME. He received the National Excellent Doctoral Thesis Award, in 2007, the First Technological Invention Award of Ministry of Education, in 2008, the Second National Technological Invention Award, in 2009, the First Provincial Teaching Achievement Award, in 2013, the First Technological Invention Award of Ministry of Education, in 2015, and the Science and Technology Award for Chinese Youth, in 2013. He hosted the National Key 973 Research Program of China as a Principal Scientist, in 2015. He is also the Chair of the IEEE Xian and Chengdu Joint Section Instrumentation and Measurement Society Chapter. He is also the Executive Director of the Fault Diagnosis Branch, China Mechanical Engineering Society.

XIAOLI ZHANG (Member, IEEE) received the Ph.D. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2011. She is currently an Associate Professor with the Department of Mechanical and Electronic Engineering, School of Construction Machinery, Chang'an University, Xi'an. Her research interests include machinery condition monitoring, intelligent fault diagnosis, and life prediction.

XINRONG ZHANG received the B.S. degree in mechanical engineering from the Jilin University of Technology, Changchun, China, in 1990, the M.S. degree in mechanical engineering from the Xi'an Highway Institute, Xi'an, China, in 1993, and the Ph.D. degree in mechanical engineering from Chang'an University, Xi'an, in 2000. He was a Postdoctoral Fellow in mechanical engineering with Tongji University, Shanghai, China, from 2001 to 2003. He is currently a Professor with the Construction Machinery School, Chang'an University. His research interest includes mechanical system dynamics and control.