Wireless Communications and Mobile Computing

Scalable and Robust Optimization Techniques for Multidimensional Signal Processing

Lead Guest Editor: Xianpeng Wang
Guest Editors: Qi Liu, Bin Zhang, and Wei Liu

Scalable and Robust Optimization Techniques for Multidimensional Signal Processing

Wireless Communications and Mobile Computing

Scalable and Robust Optimization Techniques for Multidimensional Signal Processing

Lead Guest Editor: Xianpeng Wang
Guest Editors: Qi Liu, Bin Zhang, and Wei Liu

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7674421, 22 pages
https://doi.org/10.1155/2022/7674421

Research Article

Fault Diagnosis Algorithm of Gearbox Based on NAPSO-VMD Self-Adaptive Noise Reduction and Dual-Sensor Feature Fusion

Yangshuo Liu,[1] **Jianshe Kang**,[1] **Chiming Guo**,[1] **Yunjie Bai**,[2] **and Shuo Meng**[1]

[1]*Army Engineering University of PLA, Shijiazhuang 050003, China*
[2]*66029 Unit of the Chinese People's Liberation Army, Xilinguolemeng 011200, China*

Correspondence should be addressed to Chiming Guo; guochiming@nudt.edu.cn

Received 22 May 2022; Revised 19 June 2022; Accepted 27 June 2022; Published 24 August 2022

Academic Editor: Bin Zhang

Condition monitoring plays a very important role in equipment fault diagnosis technology. However, existing monitoring methods often collect equipment fault signals from a single dimension, resulting in a major lack of fault information. To improve the problem, we built a gearbox preset fault test bench and constructed a dual-sensor acquisition system to realize the multiple dimensions of vibration signal acquisition in the horizontal and vertical directions of the gearbox. At the same time, given the poor adaptability of most current signal preprocessing methods, the improved nonlinear adaptive inertial weight particle swarm optimization algorithm (NAPSO) and variational modal decomposition (VMD) are combined to optimize the key parameters in VMD with the maximum correlation kurtosis convolution (MCKD) as the fitness function. Further, after extracting fault features from the intrinsic mode functions (IMFS) decomposed by VMD, the single-layer sparse autoencoder network (SAE) and the double-layer stacked sparse autoencoder network (SSAE) with different structures are used to realize an effective fusion of multidimensional information and deep feature extraction. Finally, the hybrid fault diagnosis of gearboxes is realized by using the random forest algorithm (RF) as the classifier. The experimental results show that the accuracy of the method proposed in this paper can reach 96.0%, and the accuracy can be improved by 3.0% and 4.0%, respectively, when compared with a single horizontal or vertical sensor signal input.

1. Introduction

The health management of equipment plays an important role and significance in maintaining the reliability of the equipment. Among them, fault diagnosis technology, as a key technology for monitoring and judging equipment status, has long received extensive attention from industrial managers and researchers [1–3]. At present, the research on fault diagnosis mode based on the combination of equipment fault signal data-driven and deep learning has reached a relatively mature stage. Wang et al. [4] used the synchronous extraction transformation to convert the gearbox vibration signal into a time-frequency map of uniform size and finally learned the optimal classification strategy automatically under the framework of deep reinforcement learning. Y. Wang and S. Wang [5] used variable-parameter time-frequency analysis to expand the number of bearing abnormal samples to address the imbalance between normal and abnormal samples and then rebalanced the dataset to train a convolutional neural network (CNN) to classify bearing health. Deng et al. [6] first performed a continuous wavelet transform (CWT) on the vibration signal of the rolling bearing to obtain a time-frequency map and then input the time-frequency map as a feature map to the BP neural network and CNN to compare the feature extraction and classification effects of different networks. Bai et al. [7] used a multiscale crop fusion data enhancement method to enhance the rolling bearing fault signal before converting it into a short-time Fourier time-frequency image; following these steps, a multichannel deep network (DNN) was then used to fuse multisensor-derived image data for feature extraction and failure mode classification.

Although the above fault diagnosis algorithms based on deep learning have achieved good classification results, they only consider the single-dimensional signal of the equipment. When a device fails, it will generate fault excitations

in multiple directions. Only collecting signals from a single dimension of the device will undoubtedly result in the lack of important device failure information. In this regard, this paper builds a gearbox preset fault test bench and builds a dual-sensor acquisition system to realize the vibration signal acquisition in the horizontal and vertical directions of the gearbox, to ensure that there are sufficient multidimensional data for monitoring and judging the equipment operation status. Further, although deep learning methods can autonomously learn deep features from the raw signals of the device, they cannot effectively integrate the information of multisource signals. In addition, the signal preprocessing analysis technology can describe the fault state of the equipment from the perspective of the fault mechanism, which has its irreplaceable advantages. In this regard, Shao et al. [8] proposed a stacked wavelet autoencoder structure with a Morlet wavelet function for multisensory data fusion and a flexible weighted assignment of fusion strategies, which realized the effective extraction and fusion of multichannel sensor fault features. Aiming at the problem that the fault detection results of the vibration signal of a single sensor may be unreliable and unstable, Liu et al. [9] proposed an intelligent multisensor data fusion method based on the ant colony optimization algorithm and correlation vector machine (RVM) for gearbox faults detection. The method removes a large amount of redundant information in the features and improves the accuracy of fault diagnosis; Cheng et al. [10] combined wavelet correlation scale entropy (WCFSE) and Dempster-Shafer (DS) evidence theory algorithm and proposed a new method for gear fault diagnosis under strong noise conditions based on multisensor information fusion. The introduction of multisensor fusion technology eliminates the uncertainty and defects of single sensor identification and improves the accuracy of fault diagnosis. On the other hand, shallow classifiers often show better accuracy and stability than deep networks in discriminating fault mechanism features. Han et al. [11] compared the discriminative accuracy of random forest (RF), extreme learning machine (ELM), probabilistic neural network (PNN), and support vector machine (SVM) on different time-frequency domain features. Experimental results show that RF outperforms comparative classifiers in terms of recognition accuracy, stability, and robustness to features, especially when the training set is small; Yang et al. [12] proposed a fault diagnosis scheme based on a hierarchically structured ELM. Representation learning based on multilayer ELM covers functions such as data preprocessing, feature extraction, and dimensionality reduction; each ELM layer is processed independently for its corresponding role, thereby realizing data representation feature learning and fault classification. Compared with deep learning (DL) based on backpropagation, iterative fine-tuning of parameters is omitted and the efficiency of diagnosis is improved; Cerrada et al. [13] established a robust multilevel fault diagnosis system for gearboxes. The optimal set of conditional parameters in time, frequency, and time-frequency domains is extracted from vibration signals, and data classification is performed in a supervised environment using genetic algorithms and RF-based classifiers. High-efficiency fault diagnosis under variable working conditions of the gearbox is realized. It can be seen from the above research that the method of combining signal preprocessing and the shallow classifier can accurately and stably realize the fault diagnosis of equipment.

However, there are still some problems that can be improved. On the one hand, various types of fault signals of equipment will produce nonlinear fluctuations with changes in the external environment such as working conditions and operating time. However, in most cases, different types of signals have not been adaptively analyzed and preprocessed, and the differences between the data have not been further highlighted. In this regard, Wang et al. [14] first used ensemble empirical mode decomposition (EEMD) to decompose the signal and then used the kurtosis spectral entropy as the objective function to search for the filter length of multipoint optimal minimum entropy deconvolution adjusted (MOMEDA) in a grid format, which is used for searching complex fault pulse signal in strong noise environment; a comparative study of adaptive algorithms derived from empirical mode decomposition (EMD), empirical wavelet transform (EWT), variational mode decomposition (VMD), and Vold-Kalman filter order tracking was conducted by Liu et al. [15], and the improvement and application of these methods in dynamic analysis and fault diagnosis of the mechanical transmission system are elaborated; Ni et al. [16] proposed a fault information-guided VMD (FIVMD) method for extracting the weak bearing repetitive transient. Two nested statistical models based on the fault cyclic information, incorporated with the statistical threshold at a specific significance level, are used to approximately determine the mode number of VMD; then, the ratio of fault characteristic amplitude (RFCA) is defined and utilized to identify the optimal bandwidth control parameter of VMD, which enables the decomposed BLIMF to be sensitive to the bearing fault signature and less affected by the abnormal impulses and vibrations from other components. Cheng et al. [17] proposed an adaptive weighted symplectic geometric distribution (AWSGD) method, which can adaptively adjust the algorithm parameters according to periodic kurtosis (CK) and periodic impact intensity (PII) to effectively denoise gear signals. The above self-adaptive algorithm can perform self-adaptive preprocessing on the equipment fault signal according to different sensitive indicators, which can effectively improve the accuracy and effect of the diagnosis algorithm. However, based on the standard of adaptive preprocessing, the adjustment of the parameters of the above algorithm is still carried out using empirical values or grid traversal search, which is less efficient. In this regard, a method combining the population optimization algorithm with the adaptive optimization objective is proposed to efficiently process the effective fault information in the signal. In addition, the selection of signal adaptive optimization criteria is also very important. Feng et al. [18] used the vibration cyclostationarity index (CS) to track the wear evolution of the equipment and studied the relationship between the tribological characteristics of the two wear phenomena, fatigue pitting and abrasive wear, and gearmesh-modulated second-order cyclostationary (CS2) properties of the vibration signal. As a result, the mechanism recognition of the gearbox wear is implemented. Fan et al.

[19] proposed a sparse representation-based transient feature extraction technique for gearbox fault diagnosis. The method can effectively identify the transient pulse time and period based on wavelet, to extract transient features. The effectiveness of the proposed method is verified by simulation signals and actual gearbox vibration signals. Since the fault pulses in the vibration signal of rotating machinery often appear in a periodic form, how to extract and separate the fault pulse characteristics is a problem that needs to be focused on. Traditional time-domain features such as kurtosis can more sensitively detect pulse components in signals, but they cannot distinguish noise components from periodic components; although the minimum entropy deconvolution method (MED) can make the signal after filtering through the inverse filter, it prefers to deconvolve a single pulse or group of pulses rather than the expected periodic pulses that recur during the fault period [20]. To this end, Zhang et al. [21] proposed a new method called maximum correlation kurtosis deconvolution (MCKD). This method optimizes the correlation kurtosis of the signal by finding the optimal filter, which can more effectively extract the periodic pulse component of the signal and restrain the noise influence of the signal compared with the MED method. Therefore, this paper combines the good noise reduction performance of variational modal decomposition for mechanical equipment vibration signals and the optimization performance of a typical swarm optimization algorithm—particle swarm optimization (PSO)—and takes the MCKD value of the signal as the fitness function, so as to adaptively optimize the key parameters in VMD. The decomposed modal component (IMF) matrix retains the periodic fault pulse components in the equipment vibration signal. Further in terms of feature extraction, the time-frequency domain signal analysis method can accurately describe and cover the fault information of the signal from the perspective of the fault mechanism [22, 23]; therefore, this paper extracts the MCKD value, 14 common time-domain features, and wavelet packet decompose energy scale features from the IMF matrix decomposed by VMD to construct the feature matrix of the signal [24]. At the same time, to improve the global optimization ability of the PSO algorithm, the inertia weight in the PSO algorithm is nonlinearly improved (NAPSO).

On the other hand, there is redundant information between multisensor signals, and the accuracy and effect of equipment fault diagnosis can be improved through effective information fusion. In this regard, Xie et al. [25] used principal component analysis (PCA) to fuse and convert multisensor signal features into RGB images, and then, the image samples are input into a convolutional neural network (CNN) with residuals for further extraction of deep features; Li et al. [26] proposed an adaptive channel weighted neural network to study the importance of different sensor signals in the feature fusion method while maximizing the mining of the deep fault feature information of each sensor and finally realized the condition monitoring of the gearbox transmission system and the helicopter transmission system; Cao and Yunusa-Kaltungo [27] proposed a gearbox fault classification framework for the automatic fusion of multisensor data, generating features through coherent composite spectroscopy (CCS) and using PCA for data dimensionality reduction. The final diagnosis results were obtained from artificial neural network training feature samples. Although the above information fusion method has played a role in reducing the dimension of the data and extracting features, it still cannot achieve the effect of deep fusion and high-dimensional fault feature extraction for the effective information of the sensor; in addition, combined with the actual situation of this paper, the signal decomposed by adaptive VMD will have different feature dimensions due to the different decomposition layers. Therefore, aligning the feature dimensions of the fault data of different channels is also a key step in the diagnosis process; and while aligning the features, maintaining the original components and information of the signal is also an aspect that needs to be considered. Sparse autoencoder (SAE) is one of the important structures in deep networks. It adds sparsity constraints to traditional autoencoders and has powerful data reconstruction and feature reexpression capabilities, so it is widely used in various fields of production and life (pattern recognition, target detection, and natural language processing) [28–30]. Therefore, this paper uses a single-layer SAE to align data of different dimensions; at the same time, for the problem that the signals collected by multiple sensors may contain redundant information, this paper builds a two-layer stacked sparse autoencoder. In this way, the features of the dual sensors are fused to achieve the purpose of data compression and feature extraction. Finally, the random forest algorithm (RF), as an emerging and highly flexible machine learning algorithm, integrates the decision value of each decision tree into the final classification result through the idea of ensemble learning [31, 32]. Therefore, the fused feature samples are input into the RF classifier to realize the hybrid fault diagnosis under the variable working conditions of the gearbox. Experiments show that the fault diagnosis algorithm proposed in this paper can obtain a good classification effect for the mixed faults of the gearbox under variable working conditions, and the classification accuracy is improved compared with the single sensor signal input.

In summary, a hybrid fault diagnosis method for gearboxes based on NAPSO-VMD adaptive noise reduction of vibration signals and multitype sparse autoencoder dual-channel sensor feature alignment and fusion is proposed in this article. Experiments and diagnostic algorithms are verified on the gearbox preset fault test bench, and the experimental results demonstrate the effectiveness of the proposed dual-channel sensor data fusion diagnostic algorithm. The work in the rest of this paper is described as follows. Section 2 introduces the theory and the process of the fault diagnosis algorithm. Section 3 provides the experiment and data preparation process for the mixed failure of the gearbox under variable operating conditions. Section 4 analyzes and discusses the experimental results. Section 5 discusses the effectiveness of the diagnosis algorithm. Finally, the conclusions of the paper will be summarized in Section 6.

2. Methodology

2.1. Adaptive Noise Reduction of Signals. When using VMD decomposition and noise reduction, the noise reduction effect of the signal is more sensitive to the decomposition

layers K and the penalty factor alpha [33]; in addition, a unified decomposition parameter cannot be used for the vibration data of different types of gearbox faults. In this paper, the improved adaptive weight particle swarm optimization algorithm is used to optimize the parameters of the signal-to-VMD with the maximum correlation kurtosis deconvolution value of the signal as the optimization objective, so that each type of fault data can obtain appropriate decomposition parameters and adaptive noise reduction.

2.1.1. Maximum Correlation Kurtosis Deconvolution (MCKD). The maximum correlation kurtosis deconvolution is a new method that is extremely sensitive to the expected periodic pulses that recur during the fault cycle, and its purpose is to deconvolve the periodic fault pulses from the vibration signal of the equipment. The following is the principle and algorithm flow of MCKD [34], assuming that the input signal and the system are in the form of convolution, which is

$$x(n) = h(n) * y(n) + e(n). \tag{1}$$

In the case of ignoring noise, the purpose is to find an optimal filter $f = [f_1 f_2 \cdots f_L]^T$ to solve $y(n)$, that is,

$$y(n) = f(n) * x(n) = \sum_{k=1}^{L} f_k x_{n-k+1}. \tag{2}$$

Among them, $x(n)$ and $y(n)$ are the input signal and output signal, respectively, and f, L, and N are the filter coefficient, filter length, and sampling times, respectively. Then, the concept of correlation kurtosis (CK) is introduced, which is defined as

$$\mathrm{CK}_M(T_s) = \frac{\sum_{n=1}^{N}\left(\prod_{m=0}^{M} y_{n-mT_s}\right)^2}{\left(\sum_{n=1}^{N} {y_n}^2\right)^{M+1}}. \tag{3}$$

Among them, M is the shift number, and T_s is the sampling point corresponding to the iterative update period, which can be expressed as

$$T_s = f_s \cdot T, \tag{4}$$

where f_s is the sampling frequency and T is the epoch of fault feature detection. The entire MCKD algorithm adopts the objective function method; that is, a finite impulse response filter is selected iteratively to maximize the correlation kurtosis value of the filtered signal, and its expression is

$$\max_f \mathrm{CK}_M(T_s) = \max_f \frac{\sum_{n=1}^{N}\left(\prod_{m=0}^{M} y_{n-mT_s}\right)^2}{\left(\sum_{n=1}^{N} {y_n}^2\right)^{M+1}}. \tag{5}$$

Among them, the general value range of M is 1 to 7, and if it is greater than 7, the precision will be reduced. Then, take the derivative of the objective function:

$$\frac{d}{df_k}\mathrm{CK}_M(T_s) = 0\ k = 1, 2, \cdots, L. \tag{6}$$

Finally, the expression form of the matrix is obtained by derivation as

$$f = \frac{\left\|y^2\right\|}{(M+1)\|\beta\|^2}\left(X_0 {X_0}^T\right)^{-1}\sum_{m=0}^{M} X_{mT}\alpha_m, \tag{7}$$

where $X_0 {X_0}^{\mathrm{T}}$ is the Toeplitz autocorrelation matrix of the original signal $x(n)$, assuming that $\left(X_0 {X_0}^T\right)^{-1}$ exists; the superscript T denotes the transpose operation; in addition,

$$X_r = \begin{bmatrix} x_{1-r} & x_{1-r} & x_{1-r} & \cdots & x_{1-r} \\ 0 & x_{1-r} & x_{1-r} & \cdots & x_{1-r} \\ 0 & 0 & x_{1-r} & \cdots & x_{1-r} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & x_{1-r} \end{bmatrix}_{LbyN}, \quad r = mT_s, m = 1, 2, \cdots, M, \tag{8}$$

$$\alpha_m = \begin{bmatrix} {y^{-1}}_{1-mT_s}\left({y_1}^2 {y^2}_{1-T_s} \cdots {y^2}_{1-MT_s}\right) \\ {y^{-1}}_{2-mT_s}\left({y_2}^2 {y^2}_{2-T_s} \cdots {y^2}_{2-MT_s}\right) \\ \cdots \\ {y^{-1}}_{N-mT_s}\left({y_N}^2 {y^2}_{N-T_s} \cdots {y^2}_{N-MT_s}\right) \end{bmatrix}, \quad \beta = \begin{bmatrix} y_1 y_{1-T_s} \cdots y_{1-MT_s} \\ y_2 y_{2-T_s} \cdots y_{2-MT_s} \\ \cdots \\ y_N y_{N-T_s} \cdots y_{N-MT_s} \end{bmatrix}. \tag{9}$$

2.1.2. Adaptive Variational Mode Decomposition. Particle swarm optimization (PSO) is a swarm evolutionary computing technology derived from the study of predation behavior of bird flocks; its basic idea is to design a massless particle with only two properties—speed and position—to simulate the birds in the flock. The optimal solution is found through cooperation and information sharing among particle individuals [35]. For an optimization problem with n dimensions $x_1, x_2 \cdots, x_n$, x_i and v_i are used to represent the position and velocity of the i -th particle, respectively, and the expressions are as follows:

$$\begin{aligned} x_i &= \left({x_i}^1, {x_i}^2 \cdots, {x_i}^n\right), \\ v_i &= \left({v_i}^1, {v_i}^2 \cdots, {v_i}^n\right). \end{aligned} \tag{10}$$

The particles in the population find the optimal parameter value corresponding to the objective function by iteratively updating the position and velocity. At the time

Table 1: Function information.

Function	Expression	Parameter value range	Theoretical optimal solution (minimum value)
Goldstein and Price	$f_1(x) = \left\{ \left[1 + (x_1 + x_2 + 1)^2 \begin{pmatrix} 19 - 14x_1 + 3x_1^2 \\ -14x_2 + 6x_1x_2 + 3x_2^2 \end{pmatrix} \right] \times \left[30 + (2x_1 - 3x_2)^2 \times \begin{pmatrix} 18 - 32x_1 + 12x_1^2 \\ +48x_2 - 36x_1x_2 + 27x_2^2 \end{pmatrix} \right] \right\}$	[-2, 2]	3
Branin	$f_2(x) = \left\{ \left(x_2 - \left(5/4\pi^2 \right) x_1^2 + (5/\pi)x_1 - 6 \right)^2 + 10(1 - 1/8\pi) \cos(x_1) + 10 \right\}$	[-5, 15]	0.397887
Schaffer F6	$f_3(x) = \sin^2 \sqrt{x_1^2 + x_2^2} - 0.5 / \left[1 + 0.001 \left(x_1^2 + x_2^2 \right) \right]^2 - 0.5$	[-100, 100]	-1

Table 2: Average fitness of different inertia weights.

Inertia weight	Index	Test function f_1	f_2	f_3
Linear weight	*BE*	6.49010^{-7}	3.773×10^{-7}	1.072×10^{-3}
	WE	27.00	3.215	3.725×10^{-2}
	AE	0.271	4.249×10^{-1}	1.127×10^{-2}
Log weight	*BE*	1.510×10^{-6}	5.275×10^{-7}	3.860×10^{-3}
	WE	0.301	2.308	4.595×10^{-2}
	AE	4.866×10^{-2}	2.311×10^{-2}	1.622×10^{-2}
Square weight	*BE*	1.151×10^{-6}	3.686×10^{-7}	3.301×10^{-5}
	WE	3.957×10^{-2}	6.171×10^{-4}	9.716×10^{-3}
	AE	2.342×10^{-3}	1.880×10^{-4}	7.199×10^{-3}
Exponential weight	*BE*	1.337×10^{-8}	2.295×10^{-7}	0
	WE	0.144	5.217×10^{-4}	0
	AE	1.759×10^{-3}	1.009×10^{-4}	0

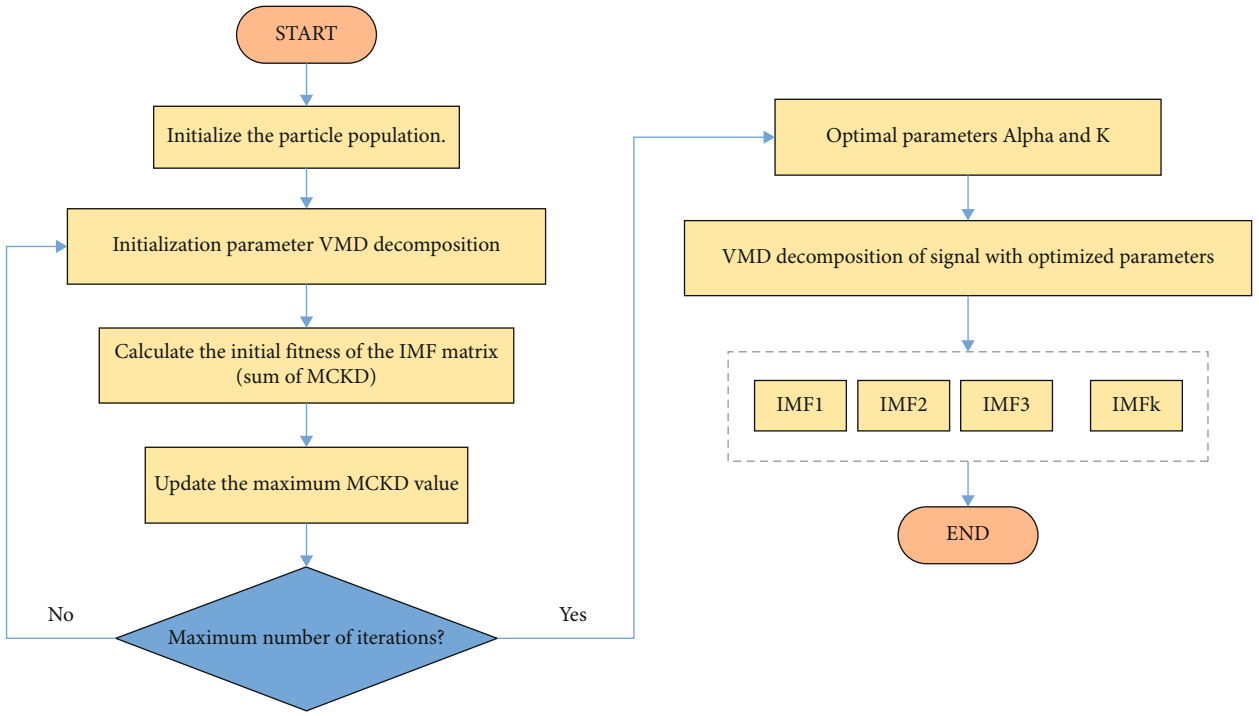

Figure 1: Parameter alpha and K value optimization process.

$t + \Delta t$, the particle velocity and position update formula are as follows:

$$v_i(t + \Delta t) = \omega v_i(t) + c_1 r_1(\overline{x_i} - x_i) + c_2 r_2(\overline{x_g} - x_i), \quad (11)$$

$$x_i(t + \Delta t) = x_i(t) + v_i(t + \Delta t)\Delta t, \quad \Delta t = 1. \quad (12)$$

t in the formula represents the current number of iterations, c_1, c_2 is a positive acceleration factor, usually taken as 2; r_1, r_2 is a random number between [0,1]; $\overline{x_i}, \overline{x_g}$ represents the historical optimal position of the i-th particle and the entire population, respectively; ω is the inertia weight, which changes linearly with the increase of the number of iterations to update the speed of the particle [36]. The basic expression of ω is

$$\omega = \omega_{\max} - (\omega_{\max} - \omega_{\min}) \times a = \omega_{\max} - (\omega_{\max} - \omega_{\min}) \times \frac{t}{T}. \quad (13)$$

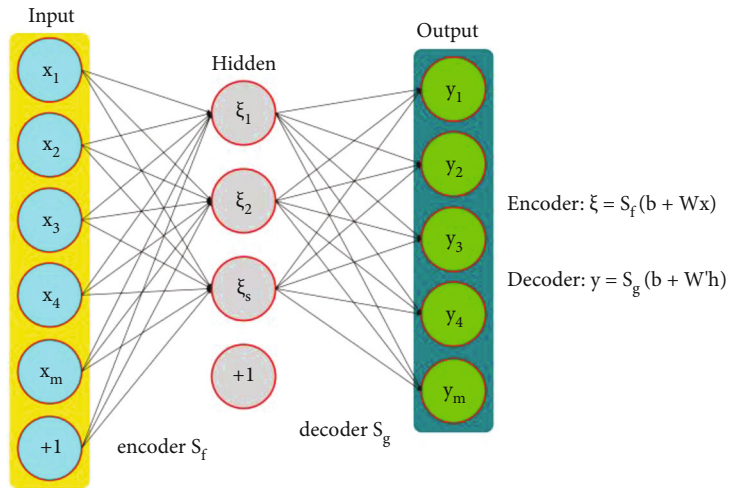

Figure 2: Structure of SAE.

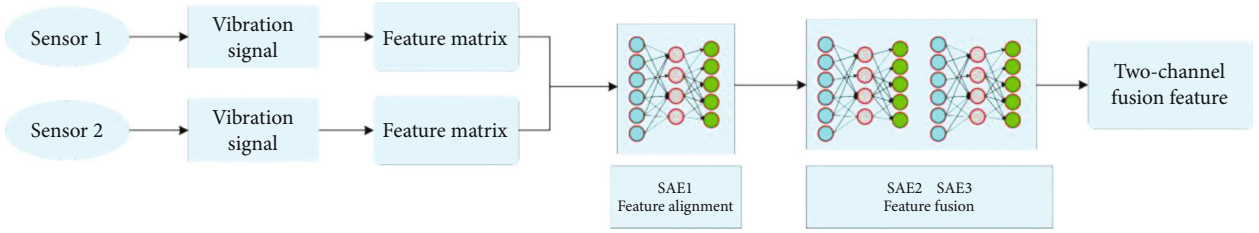

Figure 3: Feature fusion process.

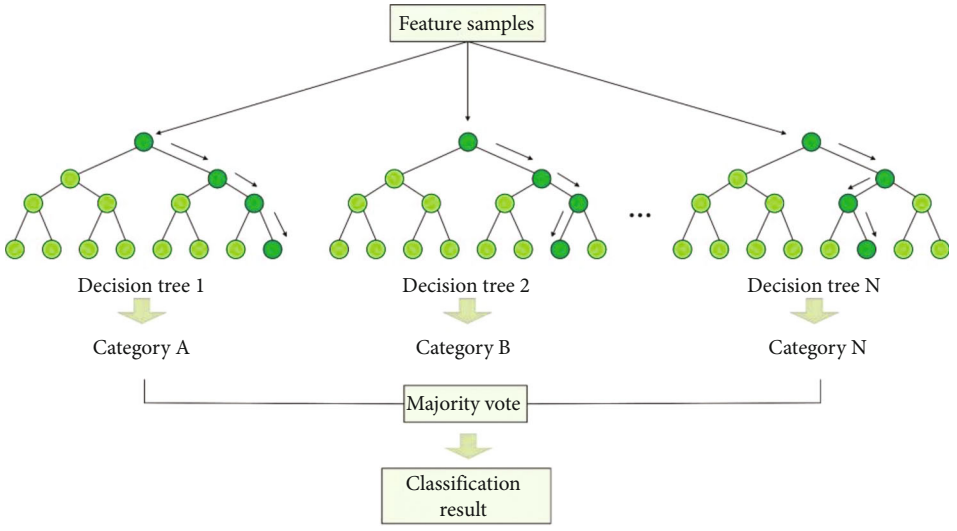

Figure 4: Schematic diagram of random forest algorithm.

In the formula, a is the weight decay method, and linear decay is the most commonly used method of the standard particle swarm algorithm. ω_{max}, ω_{min} are the maximum and minimum values of the inertia weight, usually taken as 0.9 and 0.4, and T is the maximum number of iterations. In recent years, with the deepening and diversification of research problems, the two biggest problems faced by the PSO algorithm are the speed of convergence and falling into a local minimum. Given this situation, some scholars have improved PSO from the aspects of evolution state evaluation, elite learning strategy, and system adaptive parameters. Among them, the method based on system adaptive parameters increases the diversity of particle swarms and accelerates the convergence speed by giving the inertial weight ω adaptiveness or controlling the acceleration factor [37]; in this paper, by changing the original linear inertia weight ω to exponential decay weight ω_1, it is

ensured that inertia weight has a larger value in the exploration stage, which is beneficial to the global search of the algorithm; in the development stage, a smaller value is beneficial to the local search of the algorithm and reduces fluctuations. At the same time, to compare with the weight values of other attenuation methods, this paper also lists two nonlinear weights besides linear weights: logarithmic attenuation weight ω_2 and square attenuation weight ω_3. Equation (14) lists the expressions of the four inertia weights.

$$\begin{cases} \omega = \omega_{\max} - (\omega_{\max} - \omega_{\min}) \times \dfrac{t}{T}, \\ \omega_1 = \omega_{\max} - (\omega_{\max} - \omega_{\min}) \times \dfrac{e^{t/T} - 1}{e^t - 1}, \\ \omega_2 = \omega_{\max} - (\omega_{\max} - \omega_{\min}) \times \dfrac{\ln \sqrt{T+1-t}}{\ln \sqrt{T+1}}, \\ \omega_3 = \omega_{\max} - (\omega_{\max} - \omega_{\min}) \times \dfrac{t^2}{T^2}. \end{cases} \tag{14}$$

To obtain more accurate results, this paper compares the optimization performance of different inertia weights through simulation experiments. The performance test function of the intelligent optimization algorithm used is shown in Table 1 [38, 39]. These test functions are complex and nonlinear, including multiple peaks and valleys, and there are a large number of local extreme points, which can be used to detect the global search performance, local optimization performance, and stability of the algorithm.

Taking the function value as the fitness of the PSO algorithm, the number of particles is set to 100, the maximum number of iterations is 500, and each test function is optimized 100 times. The performance of each optimization algorithm is measured by indicators such as the minimum error $BE = \min(F_i) - F_{\text{to}}$, the maximum error $\text{WE} = \max(F_i) - F_{\text{to}}$, and the average error $AE = \text{mean}(F_i) - F_{\text{to}}$. The smaller the minimum error, the better the local optimization ability of the optimization algorithm. The larger the maximum error, the worse the antipremature ability of the optimization algorithm (ability to resist premature convergence). The smaller the average error, the more stable the optimization performance of the optimization algorithm. In short, the smaller the value of these three indicators, the better the performance of the optimization algorithm, and the average fitness is shown in Table 2.

It can be seen from the overall test results that the three indicators of the nonlinear weight are smaller than those of the linear weight. Then comparing the test results of different nonlinear weights, it can be found that the square weight is not as effective as the other two types of weights in other test functions except that the maximum error WE corresponding to the test function f_1 is smaller; in addition, the indicators of exponential weight are further reduced as a whole compared with the logarithmic weight. Therefore, among the four types of inertia weights, the exponential weight has the best comprehensive optimization performance. In this paper, the exponential weight is used to replace the linear weight of the PSO algorithm to obtain a nonlinear adaptive inertia weight particle swarm optimization (NAPSO). Further, the MCKD value of the signal is used as the fitness function to optimize the parameters in the VMD. The specific steps are shown in Figure 1.

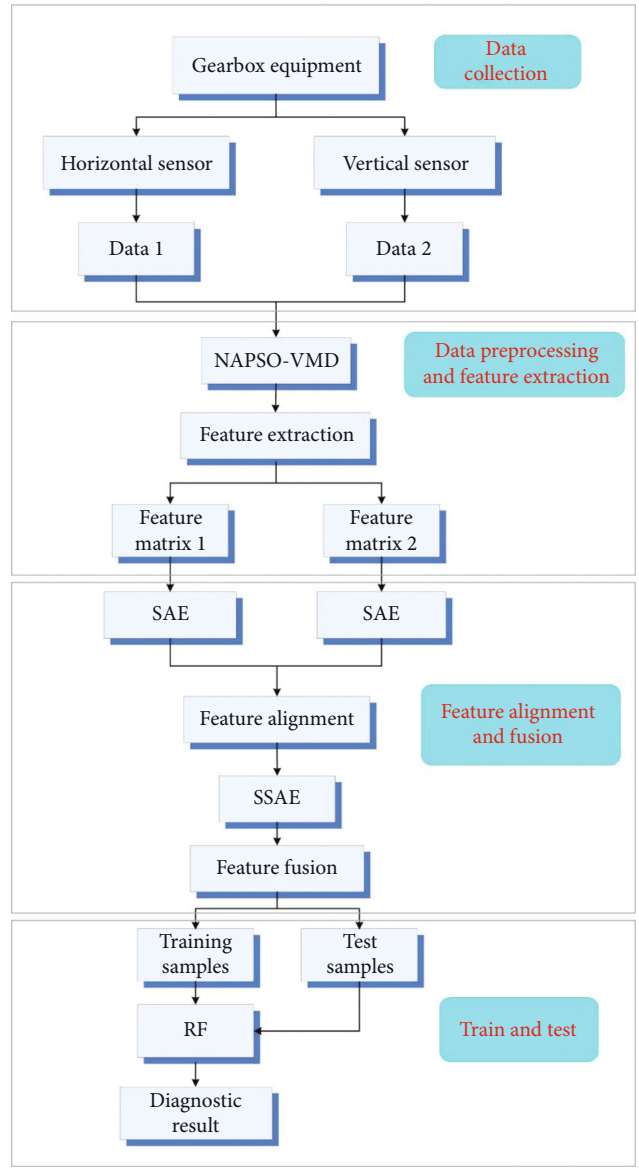

Figure 5: Diagnostic strategy flowchart.

2.2. Applications of Multitype Sparse Autoencoders. In this section, a single-layer SAE and a two-layer stacked sparse autoencoder (SSAE) are used to align and fuse the data of dual-channel sensors, using the powerful data depth reconstruction and feature extraction capabilities of the sparse autoencoder (SAE).

2.2.1. Autoencoder (AE). Autoencoder (AE), as one of the most typical unsupervised learning deep networks, its main role is to reconstruct the input data and learn its features and make the output data as equal to the input as possible. Its structure mainly includes input, hidden, and output

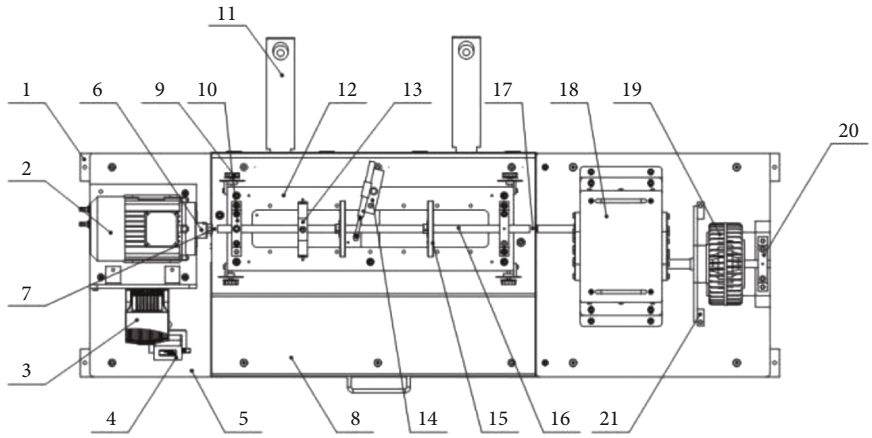

Figure 6: Structure diagram of the experimental platform.

layers (Figure 2); and its network learning includes encoding and decoding two processes. The data is encoded and reconstructed in the hidden layer to obtain a deep representation of the features [40–42].

2.2.2. Sparse Autoencoder (SAE). SAE is the same as AE in structure, including output layer, hidden layer, and output layer. The improvement is that SAE increases the sparsity limit based on AE. By constraining part of the network activation state in the hidden layer, the neuron nodes related to the input data are made active. On this basis, KL dispersion is introduced to measure the similarity between the average activation output of a hidden layer node and the sparse ρ we set. Therefore, SAE can learn more effective feature expressions and improve the efficiency of feature extraction. The training is achieved by minimizing the error as a loss function. Determine whether the error converges by calculating the output of SAE; the SAE networks are trained until the error converges [43, 44].

2.2.3. SAE-Based Feature Dimension Alignment. Based on the good data reconstruction ability of SAE, this paper uses a single-layer SAE as a feature dimension alignment tool. Firstly, the MCKD value feature, 14 commonly used time-domain features, and 8 energy scale features after 3-layer wavelet packet decomposition are extracted from the dual-channel samples after adaptive noise reduction, a total of 23 category features. Among them, time-domain features include dimensional features (mean, root mean square, root square amplitude, absolute mean, skewness, kurtosis, variance, maximum value, minimum value, and peak-to-peak value) and dimensionless features (shape factor, crest factor, margin factor, and impulse factor) [45, 46]; the "db3" wavelet basis is used in the wavelet packet decomposition [47]. In this way, after further feature extraction, a $K \times 23$-dimensional feature vector can finally be obtained.

However, since different decomposition level K values may be obtained, this paper uses the above SAE to align features of different data in each channel, so that the feature dimensions of data in the same channel are kept consistent.

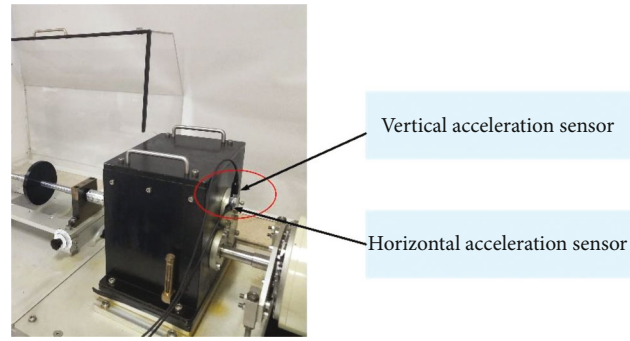

Figure 7: Two-channel sensor installation diagram.

2.2.4. SSAE-Based Deep Feature Fusion. In this section, two layers of SAE are stacked to form an SSAE network, which is used as a deep feature fusion network (DFFN). In the network, the output features of the hidden layer of the first SAE are input to the second SAE, and the deep feature extraction and fusion of multichannel data are realized through multiple unsupervised training of SSAE. For example, in Section 2.2.3, the data features of the horizontal channel sensor are unified to $K_1 \times 23$, and the data features of the vertical channel sensor are unified to $K_2 \times 23$; then, the features of the two channels are concatenated into a $(K_1 + K_2) \times 23$ vector and input to the DFFN network for feature fusion. The concatenated features are first input to the first layer of SAE in the DFFN network, and the output features of the hidden layer are used as the input of the second layer of SAE, and the final output of the second layer of SAE hidden layer is the dual-channel fusion feature. The process of feature fusion is shown in Figure 3.

2.3. Random Forest Algorithm (RF). Random forest is a machine learning algorithm based on the idea of ensemble learning. The basic unit of a random forest is a decision tree, and a random forest is formed by integrating multiple decision trees. For classification problems, first, use each decision tree in the random forest to classify the sample, each decision tree will randomly select some features of the sample, and make classification decisions based on them;

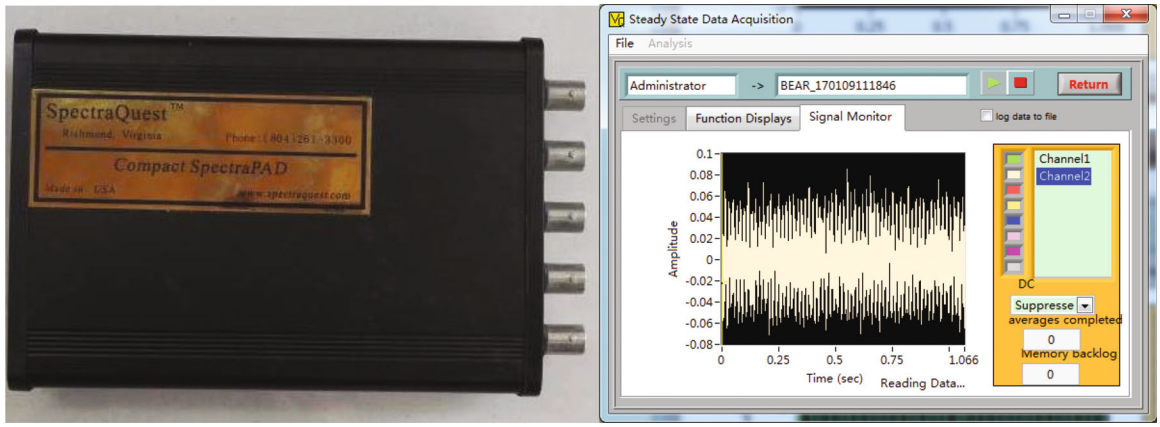

Figure 8: Data acquisition board and acquisition software.

these classification decisions are finally integrated to form the classification result. Figure 4 visually shows how random forests work [13, 48].

The random forest has two main elements: "random" and "forest." "Forest" is easy to understand, that is, multiple decision trees constitute a forest, which reflects the integration idea of a random forest. And "random" refers to the random selection of samples and the random selection of features to be selected. In recent years, random forests have been widely used in various fields because random forests have the following advantages [49]:

(1) It can run effectively on large datasets

(2) It can process input samples with high-dimensional features without dimensionality reduction

(3) Since each tree selects some samples and some features, overfitting can be avoided to a certain extent, and the performance is stable

In Section 2.2, the two-channel sensor fault samples have large dimensions after feature extraction and deep fusion, so random forest is selected as the classifier of the fault diagnosis method. The specific implementation process of random forest is as follows [50, 51]:

(1) Build a decision tree. Assume that the number of samples to be classified is N, and the number of features of the samples is U; when the decision tree analyzes and decides the samples, it needs to randomly extract u sample features, where u is much smaller than U, and then determines the decision result of a node on the decision tree according to the attributes of the selected features. After that, all samples are randomly sampled with a replacement for N times (that is, bootstrap sampling), thus forming a training sample set. The samples that are not drawn from the sample population are used as the test set to evaluate the error of the decision tree. So far, for all nodes in the decision tree, repeat the aforementioned analysis and decision-making process; that is, each node randomly selects u features. According to these u features, calculate the best splitting method; in addition, each decision tree will grow completely without being pruned

(2) Select important features. Among the randomly selected features in a tree, the importance of each feature is different, so it is necessary to select more important features. Arbitrarily take one of u features of a tree and randomly change the feature value, compare the error rate of the test set before and after the change, and the test set is the remaining samples after sampling. Then, calculate the error rate of the features of a tree, then calculate the importance of

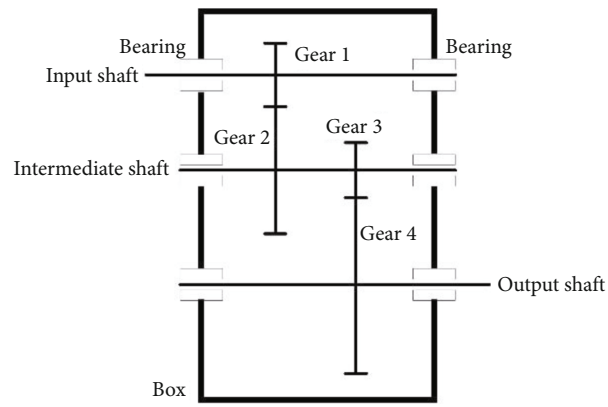

Figure 9: Internal structure diagram of experimental gearbox.

Table 3: The parameters of gears of the experimental gearbox.

Gear shaft	Gear number	Modulus	Number of teeth	Hub diameter (mm)
Input shaft	Gear 1	1.5	41	45
Intermediate shaft	Gear 2	1.5	79	60
Intermediate shaft	Gear 3	1.5	36	45
Output shaft	Gear 4	1.5	90	60

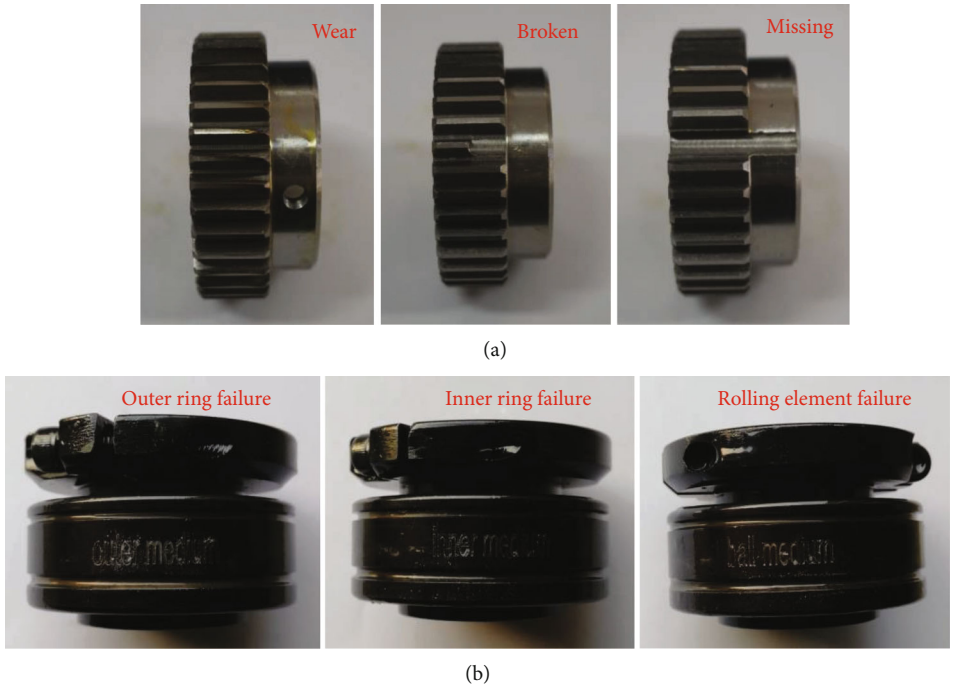

Figure 10: Preset failure experiments of (a) gears and (b) bearing.

the features in all trees in their respective trees, and finally calculate the mean of the importance of this feature value in multiple trees. This gives the importance of all features in the forest. Sort all the features according to their importance, remove some features with low importance in the forest, and get a new feature set

(3) Follow the above steps for multiple iterations, gradually remove features with relatively low importance, and generate a new forest each time until the number of remaining features is u. After each iteration of the forest, the out-of-bag error rate is also calculated to evaluate the performance of the forest, and the forest with the smallest out-of-bag error rate is selected as the final random forest model. The out-of-bag error rate is mainly calculated in the following ways: first, a sample is used as a decision tree of the test set to classify the sample, and the final result is obtained through the voting classification of multiple decision trees. Then, perform the first step on all samples, and finally, calculate the ratio of the number of misclassified samples to the total number of samples as the out-of-bag error rate

2.4. Fault Diagnosis Algorithm Flow. The dual-channel fusion feature samples obtained in Section 2.2 are input into the random forest classifier for training, to realize the fault diagnosis of the mixed fault of the gearbox. So far, the complete diagnosis process of this paper is shown in Figure 5.

(1) Install two accelerometers in different directions on the gearbox test bench to perform multichannel acquisition of vibration signals

(2) Perform adaptive VMD on the dual-channel fault vibration signal of the gearbox

(3) Fault features are extracted from the denoised signal; then use single-layer sparse autoencoder and double-layer stacked sparse autoencoder to align and fuse the features, respectively

(4) Use the RF classifier to train and test the feature samples to realize the fault diagnosis of the gearbox

3. Data Preparation

To research the hybrid fault diagnosis method of the gearbox, the preset hybrid fault experiment is carried out by relying on the simulation fault test bench of the general components of the mechanical transmission system. The simulation failure test bench for the general components of the mechanical transmission system is composed of gearboxes, bearings, motors, magnetic powder brakes, and sensors. The specific structure of the test bench is shown in Figure 6. The components corresponding to the numbers in the figure are as follows: 1, test bench foundation lifting ear; 2, motor; 3, motor frequency conversion controller; 4, speed sensor display table; 5, test bench foundation; 6, speed sensor; 7, coupling; 8, safety cover; 9, centering adjustment code disc (2 axial and 4 horizontal); 10, split rolling (sliding) bearing seat (2); 11, safety cover support; 12, rotor system base; 13, intermediate positioning device (for sliding bearings); 14, mechanical friction device; 15, rotors with evenly distributed screw holes (2); 16, rotor shaft (20 mm); 17, rotor shaft and gearbox connection coupling; 18, parallel gearbox (two-stage); 19, magnetic powder brake; and 20, gearbox support bearing seat (including rolling bearings).

Table 4: The parameters of bearings of the experimental gearbox.

Bearing pitch diameter D (mm)	Rolling body diameter d (mm)	Number of rolling elements Z/piece	Contact angle α (°)
38.5	7.9	9	0

Table 5: Details of 12 test working conditions.

Condition number	Rotating speed	Load current
1	1200 rpm (constant)	0.1 A
2	1500 rpm (constant)	0.1 A
3	1800 rpm (constant)	0.1 A
4	600 rpm~1200 rpm (even speed change)	0.1 A
5	1200 rpm (constant)	0.2 A
6	1500 rpm (constant)	0.2 A
7	1800 rpm (constant)	0.2 A
8	600 rpm~1200 rpm (even speed change)	0.2 A
9	1200 rpm (constant)	0.3 A
10	1500 rpm (constant)	0.3 A
11	1800 rpm (constant)	0.3 A
12	600 rpm~1200 rpm (even speed change)	0.3 A

The power of the test bench is provided by the motor, the magnetic powder brake is used to provide variable load, and the gearbox used in the experiment is a two-stage parallel shaft gearbox To realize the dual-channel vibration data acquisition, two acceleration vibration sensors are installed in the vertical and horizontal directions of the bearing housing of the preset fault bearing in the gearbox, respectively (Figure 7); the data acquisition board and data acquisition software interface are shown in Figure 8.

In this paper, a preset fault experiment is carried out on the gears and bearings in the gearbox. The gearbox used in the experiment is a two-stage parallel shaft gearbox. The internal structure diagram is shown in Figure 9, and the specific parameters of the gear are shown in Table 3. The preset faulty gear is the pinion of the intermediate shaft (gear 3 in Table 3). The preset faults of gears are set as gear wear, gear broken teeth, and gear missing teeth, as shown in Figure 10(a). Among them, the gear wear fault is that the upper and lower tooth surfaces of the gear teeth are ground equidistantly inward by 0.2 mm; the gear broken tooth fault is the removal of half of one of the gear teeth; the missing tooth fault is that one of the gear teeth is completely removed. The rolling bearing model of the gearbox used in the experiment is ER-16K, and its specific parameters are shown in Table 4. The bearing with the preset fault in the experiment is the bearing of the intermediate shaft close to the gear 3. The preset faults of the bearing mainly include outer ring fault, inner ring fault, and rolling element fault. The faulty bearing is shown in Figure 10(b). The outer ring fault is to machine a 0.5 mm deep groove at the center of the outer ring; the inner ring fault is to machine a 0.5 mm deep groove at the center of the inner ring; the rolling element fault is to machine a groove on one of the rolling elements 0.5 mm deep groove.

Table 6: Data category and operating condition information for fault diagnosis.

Fault type	Condition number
F1	4
F2	9
F3	5
F4	5
F5	4
F6	6
F7	10
F8	8
F9	2
F10	6

Table 7: Optimization results of parameters.

Type	Channel	Alpha value	K value	MCKD
F1	Horizontal	9016	6	2.88e-04
	Vertical	9082	5	5.62e-04
F2	Horizontal	9892	4	3.66e-05
	Vertical	9003	4	2.56e-05
F3	Horizontal	7629	3	3.41e-07
	Vertical	7682	4	2.57e-04
F4	Horizontal	9973	4	4.04e-04
	Vertical	9928	4	5.71e-05
F5	Horizontal	7839	4	3.86e-04
	Vertical	9374	3	3.44e-06
F6	Horizontal	7914	4	6.09e-05
	Vertical	7762	4	6.07e-04
F7	Horizontal	8538	4	6.33e-05
	Vertical	9714	5	2.81e-04
F8	Horizontal	8167	4	3.94e-05
	Vertical	8831	5	3.14e-04
F9	Horizontal	9416	5	3.05e-04
	Vertical	8801	5	1.99e-05
F10	Horizontal	8983	6	3.00e-06
	Vertical	9786	4	1.56e-06

Assembling the failure parts described above into the gearbox constitutes different mixed failure modes. A total of 9 mixed fault states and normal state data were collected in the experiment, of which 9 mixed fault states include broken gear teeth and bearing rolling element fault (hereinafter referred to as F1 for simplification, other fault forms are

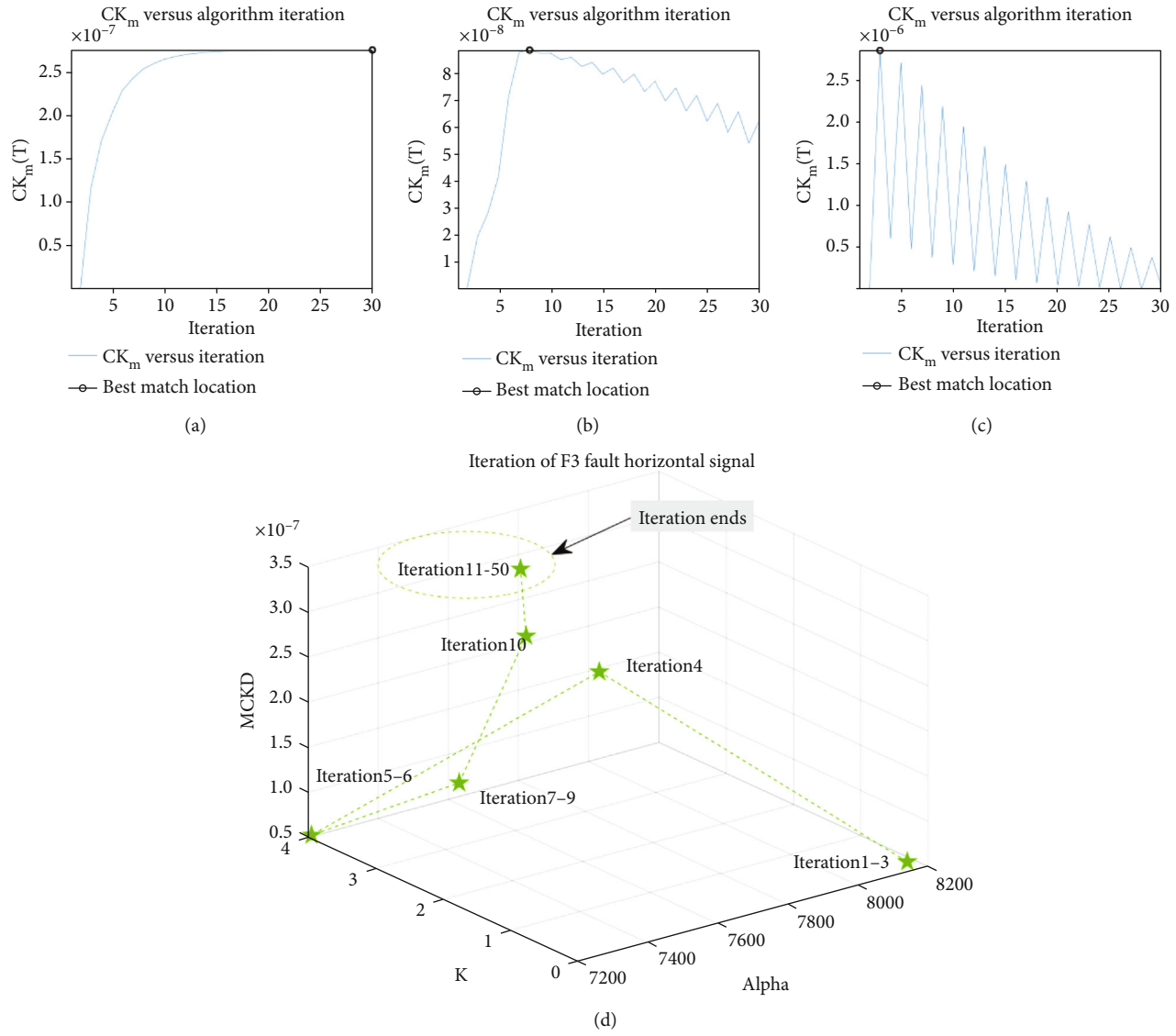

Figure 11: MCKD calculation of F3 fault horizontal channel signal (a–c) and iterative diagram of NAPSO algorithm (d).

simplified in this way), gear broken teeth and bearing inner ring fault (F2), gear broken teeth and bearing outer ring fault (F3), gear wear and bearing rolling element fault (F4), gear wear and bearing inner ring fault (F5), gear wear and bearing outer ring fault (F6), gear missing teeth and bearing rolling element fault (F7), gear missing teeth and bearing inner ring fault (F8), and gear missing teeth and bearing outer ring fault (F9), and the data in the normal state is expressed as F10. To more comprehensively diagnose different faults under the mixed working conditions of shaft teeth, mixed fault experiments under different working conditions were carried out. The specific working conditions are set as shown in Table 5, in which the speed of 600 rpm~1200 rpm (uniform speed change) means that the speed is increased uniformly from 600 rpm to 1200 rpm within 10 seconds, and changing the current simulates different loads.

In this paper, the data under a certain working condition is randomly selected from the 10 types of experimental data to realize the fault diagnosis under the mixed working condition of shaft teeth. Table 6 lists the working condition information of the selected data. The sampling frequency of the vibration signal is 20.48 kHz, the sampling length of each group of data is 49.6 s, 5000 points are taken as a sample, and each type of data of the two channels is taken 60 samples.

After the data samples are collected, adaptive VMD noise reduction is performed on them, the number of particles is set to 100, and the number of iterations is set to 50. Table 7 lists the average value of 10 calculations. It can be seen from the MCKD calculation of the two channels of F3 type data and the iterative process of the NAPSO algorithm shown in Figures 11 and 12 that after updating the iterative filter many times, each IMF component of the signal has obtained the MCKD value. This means that after the adaptive VMD decomposition, the signal retains a sufficient number of periodic fault pulses, which is conducive to feature extraction and fault classification. In terms of NAPSO iteration, on the one hand, with each fitness update, the

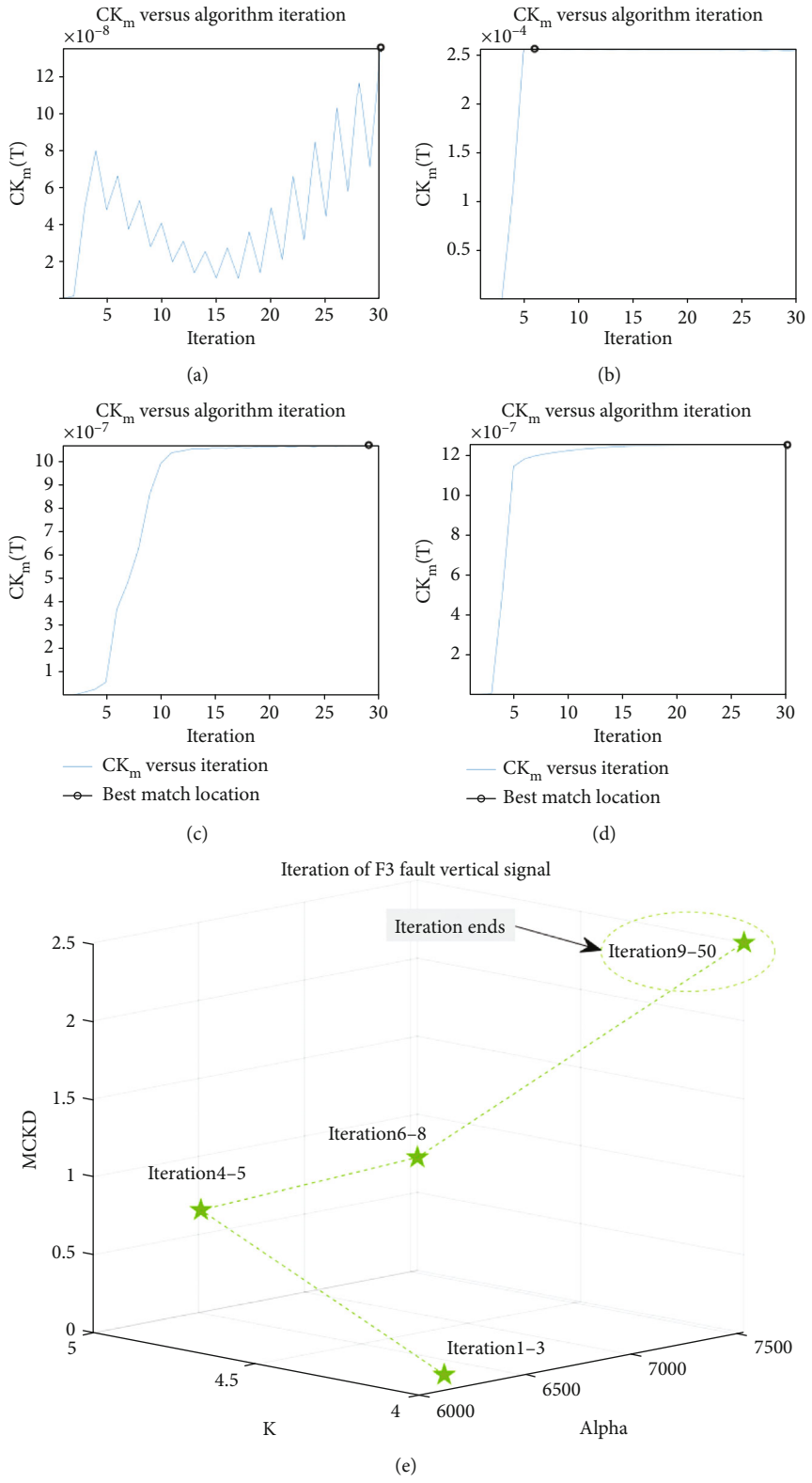

Figure 12: MCKD calculation of F3 fault vertical channel signal (a–d) and iterative diagram of NAPSO algorithm (e).

MCKD of the signal is iteratively computed; on the other hand, the horizontal signal and vertical signal reach the maximum MCKD value at the 11th and 9th iterations, respectively, and in the subsequent iterations, the MCKD value remains around the maximum value (iterative results overlap).

After extracting features and unifying dimensions, the feature dimension of the horizontal channel is unified to

Table 8: Parameter settings of sparse autoencoders with different functions.

Parameter	SAE1	SAE2	SAE3
Hidden layer node number	69/92	100	30
L2 regularization weight decay coefficient	0.002	0.001	0.001
Sparse penalty weight	0.003	0.002	0.002
Sparsity parameter	0.0001	0.0001	0.0001
Decoder transfer function	Purelin	Purelin	Purelin
The maximum number of iterations	100	100	100

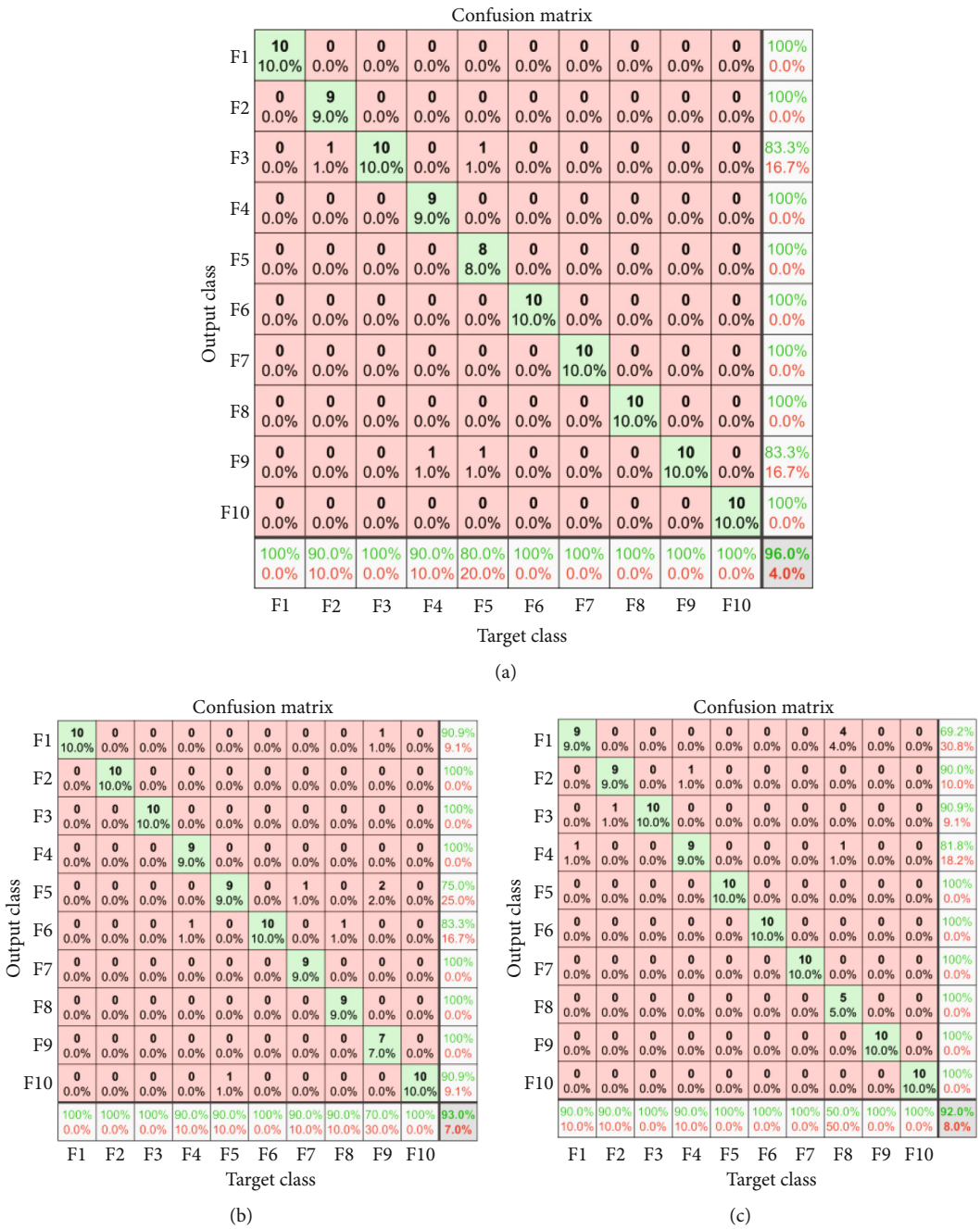

Figure 13: Confusion matrix of fault diagnosis results for different channels: (a) fusion features, (b) horizontal data features, and (c) vertical data features.

3×23 dimensions, and the feature dimension of the vertical channel is unified to 4×23 dimensions; that is, the hidden layer node number of the single-layer sparse autoencoder is set to 69 and 92, respectively, and they can be connected in series to form 7×23-dimensional feature samples; then, the series feature samples are normalized by Z-score and input into the DFFN network for feature fusion, and hidden layer node number of the two-layer sparse autoencoder in the DFFN network is 100 and 30, respectively. The final fault data can get a total of 600 30-dimensional feature samples of 10 types. Among them, the settings of the parameter of the three-layer sparse autoencoder are shown in Table 8 [52].

4. Fault Diagnosis Using Random Forest Classifiers

First, the samples are divided into two parts according to the ratio of 5 : 1, the training set, and the test set. The number of decision trees of the RF classifier is set to 1000, and the number of features randomly selected for each decision tree is set to 3; after training the training set with the RF classifier, the test set was input into the model, and the diagnostic accuracy was 96.00% (Figure 13(a)). At the same time, the feature matrices of the horizontal and vertical channels are input into RF for training and testing, and the diagnostic accuracy of the horizontal data is 93.00% (Figure 13(b)), and the diagnostic accuracy of the vertical data is 92.00% (Figure 13(c)). The diagnostic accuracy was 3% and 4% lower than that of fusion features, respectively. The out-of-bag error rates of the three types of features in RF are shown in Figure 14. It can be seen from the figure that the out-of-bag error rate of the fusion feature is significantly lower than that of the other two-directional sensor data features. The above experimental results demonstrate the effectiveness of the dual-channel feature fusion method in this paper.

5. Model Analysis

5.1. Effectiveness Analysis of Dual-Channel Fusion Features. To observe the fusion feature matrix extracted in this paper more intuitively, the data visualization method t-distributed stochastic neighborhood embedding (t-SNE) is used to visualize the fusion feature matrix and the feature matrix of the sensor data in the vertical and horizontal directions, and the results are shown in Figure 15. It can be seen from the figure that the fused feature scattergram can completely distinguish different fault states, while the data feature scattergram collected from the sensors in the vertical and horizontal directions cannot clearly distinguish some fault states. Therefore, the t-SNE analysis results intuitively show that the fused features perform better than the unfused features in distinguishing different mixed fault states. It also further illustrates the effectiveness of the feature extraction method in this paper.

5.2. Parameter Sensitivity Analysis of Sparse Autoencoder. The data reconstruction ability of SAE is closely related to the L2 regularization weight decay coefficient, sparse penalty weight, sparsity parameter, and other parameter values.

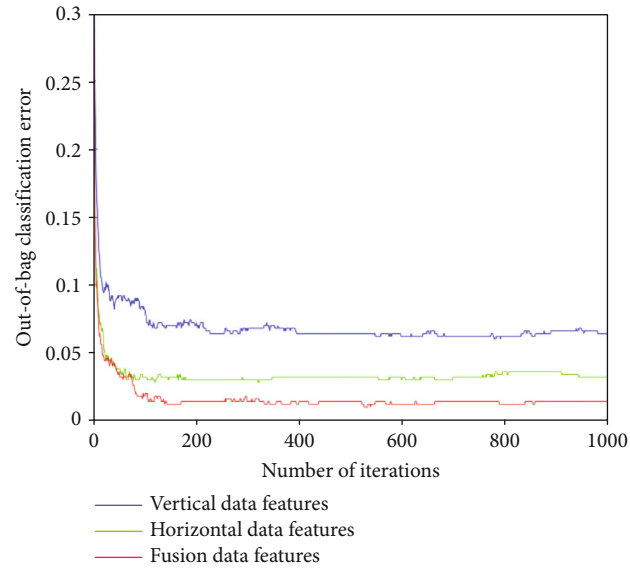

Figure 14: RF out-of-bag classification error for different channel data.

Therefore, to analyze the robustness of the diagnosis algorithm to the above parameters, this paper adjusts the above parameters of the sparse autoencoder in the DFNN network and calculates the final fault accuracy as shown in Tables 9–11 (when adjusting each parameter, the rest of the parameters remain unchanged).

It can be seen from Table 9 that when the L2 regularization weight decay coefficient is 3e-01 and 3e-02, the diagnostic accuracy is very low; when it is 1e-03 or even smaller, the diagnostic accuracy is basically consistent with the original accuracy, both of which are about 96.00%, indicating that the value of the L2 regularization weight decay coefficient should not be too large. Observing Tables 10 and 11, it is found that the diagnostic results corresponding to sparse penalty weight and sparsity parameters of different orders of magnitude are not much different, indicating that the diagnostic method in this paper is not sensitive to the changes in these two parameters. Based on the above analysis, the method in this paper has better robustness to sparse penalty weight and sparsity parameters, while the L2 decay coefficient should not be set too large, otherwise, the diagnosis effect will be affected.

5.3. Performance Comparison Analysis. To further prove the effectiveness of the fault diagnosis algorithm in this paper, this section first adopts different adaptive preprocessing methods to analyze the signal, including only VMD decomposition is performed on the signal; only empirical mode decomposition (EMD) is performed on the signal [53]; the particle swarm optimization algorithm is combined with VMD to decompose the signal, and the adaptive decomposition standard is the signal power spectrum entropy widely used in current research. It is noted as PSO-VMD-DE [54]. The particle swarm optimization algorithm is still combined with VMD, but the adaptive decomposition standard is consistent with the method in this paper, which is the MCKD

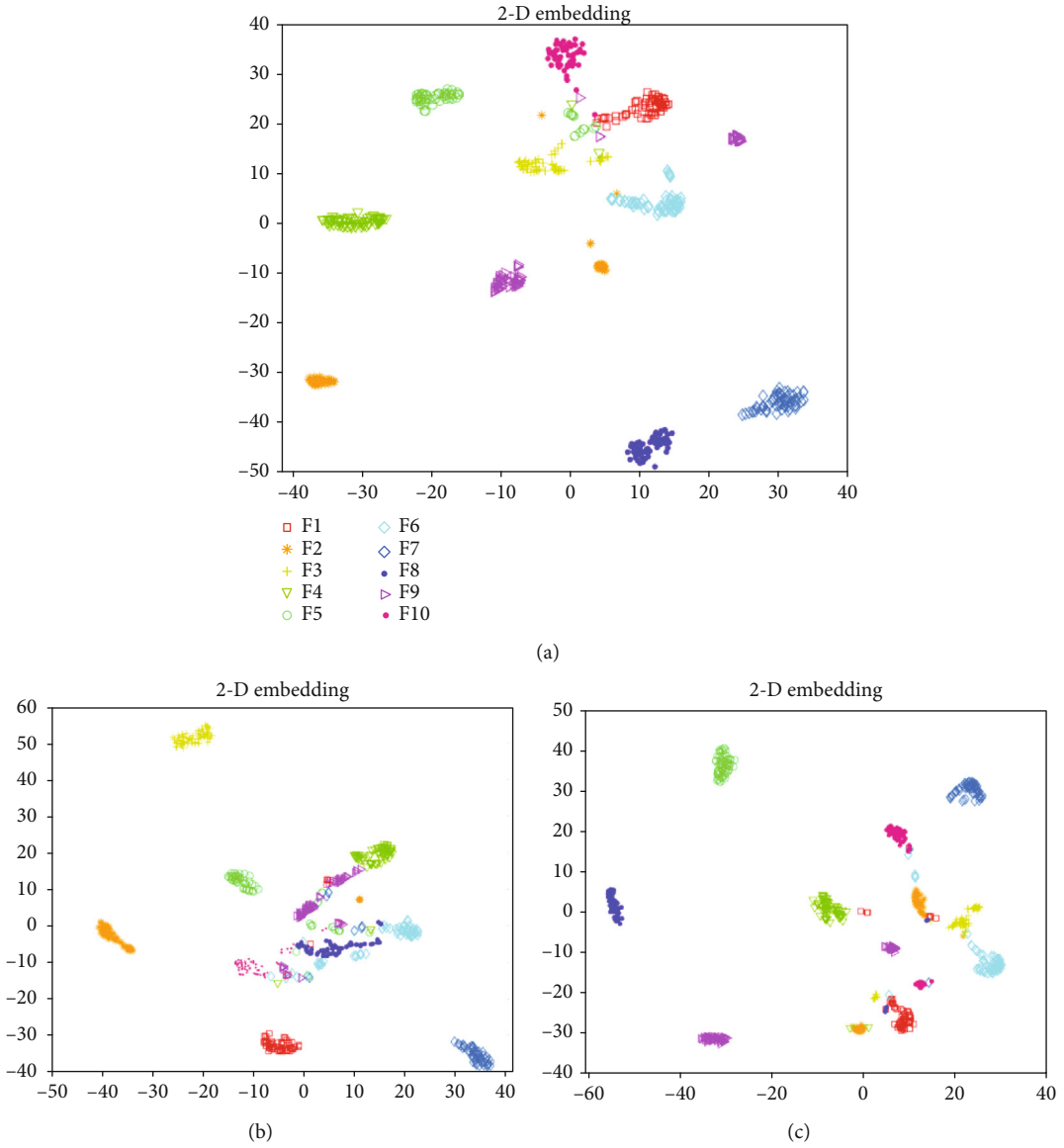

Figure 15: Feature visualization processing results: (a) fusion features, (b) horizontal data features, and (c) vertical data features.

value of the signal, which is recorded as PSO-VMD-MCKD; this method is denoted as NAPSO-VMD-MCKD. In addition, when the modal decomposition of the signal is performed directly, the parameter values such as the decomposition layers are taken as empirical values [55, 56]. The feature extraction and fusion steps are the same as those in Section 2.2, the classifier still uses RF, and the parameter settings are the same as the previous ones. The mean values of the multiple experimental results are shown in Table 12.

It can be seen from the above results that the VMD method has a better diagnosis effect than the EMD method because VMD effectively solves the modal aliasing problem in signal decomposition to a certain extent. However, since the signal is not decomposed adaptively, the diagnostic effect of the VMD method is lower than that of the PSO-VMD-DE and PSO-VMD-MCKD methods; the diagnostic effect of the PSO-VMD-MCKD method is higher than that of the PSO-VMD-DE method, which shows that in the fault diagnosis of the gearbox, compared with the power spectrum entropy, it is better to perform adaptive decomposition based on the MCKD value of the signal. It retains the fault information of the signal, thereby improving the accuracy of fault diagnosis. Finally, because the NAPSO algorithm has a stronger optimization ability than PSO, the NAPSO-VMD-MCKD method obtains more adaptive decomposition parameter values than the PSO-VMD-MCKD method, and thus, the diagnosis effect is better. The above experimental results demonstrate the effectiveness of the NAPSO-VMD-MCKD signal adaptive decomposition method in this paper. And on the whole, no matter which adaptive processing method is used, the accuracy of dual-channel data is higher than that of single-channel data. The validity of the establishment of the dual-channel sensor information acquisition system in this paper has been proved again.

On the other hand, this section adopts several typical feature fusion methods to fuse the channel feature data

Table 9: Fault diagnosis accuracy of methods with different L2 regularization weight decay coefficient.

L2 regularization weight decay coefficient	3e-01	3e-02	1e-03	1e-05	3e-06
Diagnostic accuracy (%)	55.23	62.64	96.00	95.33	96.00

Table 10: Fault diagnosis accuracy of methods with different sparse penalty weight.

Sparse penalty weight	2e-01	2e-02	2e-03	1e-05	2e-06
Diagnostic accuracy (%)	95.33	96.00	96.00	95.33	95.33

Table 11: Fault diagnosis accuracy of methods with different sparsity parameters.

Sparsity parameter	1e-01	1e-02	1e-04	1e-05	1e-06
Diagnostic accuracy (%)	96.00	96.00	96.00	95.33	96.00

Table 12: Fault diagnosis accuracy of the different adaptive preprocessing methods.

Method	VMD	EMD	PSO-VMD-DE	PSO-VMD-MCKD	NAPSO-VMD-MCKD
Horizontal data accuracy (%)	78.23	73.20	88.00	90.03	93.00
Vertical data accuracy (%)	74.00	73.00	89.30	90.20	92.00
Dual-channel data accuracy (%)	89.35	85.30	91.32	94.23	96.00

Table 13: Fault diagnosis accuracy of different feature fusion methods.

Method	KPCA	FA	LDA	MDS	MT-SAE
Horizontal data accuracy (%)	88.86	82.02	75.00	85.92	93.00
Vertical data accuracy (%)	89.02	83.00	76.23	86.50	92.00
Dual-channel data accuracy (%)	93.57	87.33	81.00	91.33	96.00

before alignment in Section 3 to verify the effectiveness of feature alignment and feature fusion methods based on multitype SAE. The methods mainly include kernel principal component analysis (KPCA), factor analysis (FA), linear discriminant analysis (LDA), and multidimensional scaling (MDS) [57–60]. The kernel function used in KPCA is the Gaussian kernel function. The method in this paper is denoted as multitype sparse autoencoder (MT-SAE). To maintain the principle of invariance, the classifier still uses RF, and the parameter settings are the same as the previous ones. The specific diagnosis results are shown in Table 13. It can be seen from the table that the KPCA method is higher than FA, LDA, and MDS. This is because the nonlinear mapping component is introduced in KPCA, which is more suitable for dimension reduction and fusion of nonstationary signal features. The accuracy of the MT-SAE method is still higher than that of KPCA because the data feature alignment and feature fusion based on MT-SAE can reexpress the data features while minimizing the data reconstruction error. The feature information of the data is nonlinearly reconstructed, and the deep-level features of the data will be further extracted, fully retaining the individual characteristics of the data, and the differences between different types of data will be effectively highlighted. Therefore, the effectiveness and superiority of the MT-SAE data feature fusion method have been proved.

Finally, different fault diagnosis methods are used to diagnose the fault of the gearbox to verify the superiority of the diagnosis algorithm proposed in this paper. These methods include inputting the raw vibration signal into a one-dimensional convolutional network (1D-CNN) and inputting the feature samples used in this paper into the SVM and Softmax classifiers, respectively [61–63]. In the training process of SVM, its two key parameters, the penalty parameter C and the kernel function parameter g, have been optimized by the NAPSO algorithm, and the objective function is the accuracy of fault diagnosis. The experimental results are shown in Table 14. It can be concluded that SVM cannot solve the multiclassification problem well, and the diagnosis result of dual-channel data only reaches 84.67%. In addition, although the diagnostic accuracy of 1D-CNN and Softmax classifiers can reach more than 90%, the accuracy of the two is still lower than that of RF classifiers because they have not been trained with sufficient data.

Table 14: Fault diagnosis accuracy of different classifiers.

Method	1D-CNN	SVM	Softmax	RF
Horizontal data accuracy (%)	85.36	79.32	89.00	93.00
Vertical data accuracy (%)	86.01	80.54	90.11	92.00
Dual-channel data accuracy (%)	91.22	84.67	93.21	96.00

The above experimental results demonstrate the effectiveness of the proposed algorithm for the gearbox hybrid fault classification problem.

In summary, the algorithm proposed in this paper can not only adaptively decompose and denoise various types of fault signals of the gearbox but also effectively and deeply fuse the data features of the dual-channel sensor. Finally, the hybrid fault diagnosis under variable working conditions of the gearbox is realized with high precision.

6. Conclusion

The condition monitoring is of great significance to the healthy and smooth operation of the equipment. A novel fault diagnosis algorithm of gearbox based on NAPSO-VMD self-adaptive noise reduction and dual-sensor feature fusion is proposed to solve the problem that adaptive noise reduction of the vibration signal of the equipment and the single-channel information cannot completely cover the signal fault information. Through the experimental verification of the gearbox, the following conclusions are obtained.

(1) By combining the nonlinear adaptive weight particle swarm algorithm and the variational mode decomposition method, the deconvolution of the maximum correlation spectrum kurtosis of the signal is used as the fitness function to perform adaptive VMD decomposition and noise reduction of the signal. The decomposed IMF matrix retains the periodic fault pulse components in the signal, which is very beneficial to the extraction and discrimination of fault features. In addition, after the inertia weight in particle swarm optimization is improved to exponential nonlinear weight, the optimization ability of the algorithm has been greatly improved

(2) In this paper, MCKD and common time-frequency domain features are extracted for the dual-channel signal, and the feature matrix can describe the fault feature performance from the original point of view of the signal. And when the time-frequency features of the signal are reconstructed and fused by multilayer sparse autoencoders and then output, the fault features are deeply expressed, which is very beneficial to the diagnosis of fault data

(3) In the process of method verification, on the one hand, this paper compares and analyzes the fault diagnosis results of fusion features and single sensor signal features, which proves that fusion features can cover and describe the fault features of signals more comprehensively and have higher diagnostic accuracy; on the other hand, the influence of the sensitivity parameters of SAE in the DFFN network on the diagnosis results is analyzed, and it is concluded that the robustness of the diagnostic algorithm in this paper to sparse penalty weight and sparsity parameter is better, but the diagnostic accuracy will be reduced if L2 regularization weight decay coefficient is set too large

(4) Sparse autoencoders have very powerful data feature reexpression capabilities. In this paper, two different sparse autoencoders, single-layer SAE and double-stacked SAE, are used to achieve the alignment and fusion of two-channel data features, respectively. Verified by RF classification experiments and compared with other common feature fusion methods, it is proved that the feature fusion method of DFFN in this paper is higher than other methods, and the effectiveness of the RF algorithm in classifying high-dimensional and multiclass samples has also been proved. To sum up, the fault diagnosis algorithm proposed in this paper can well realize the mixed fault diagnosis of gearbox equipment under variable working conditions

Effective fault diagnosis is of great significance to the smooth operation of equipment. In future work, further research will be done on the effective standard of adaptive preprocessing standard of the signal, the improvement of the optimization ability of the group optimization algorithm, the efficient fusion of multichannel signal features, and the improvement of the learning performance of the classifier.

Data Availability

The experimental data of this study can be obtained by contacting the corresponding author by email.

Conflicts of Interest

The authors of this article declare that there are no known competing financial interests or personal relationships that could influence the work of this article.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant No. 71871220.

References

[1] L. Wen, X. Li, and L. Gao, "A transfer convolutional neural network for fault diagnosis based on ResNet-50," *Neural Computing and Applications*, vol. 32, no. 10, pp. 6111–6124, 2020.

[2] Z. Liu, K. Ding, H. Lin, G. He, C. Du, and Z. Chen, "A novel impact feature extraction method based on EMD and sparse decomposition for gear local fault diagnosis," *Machines*, vol. 10, no. 4, article 242, 2022.

[3] J. Wu, Z. Zhao, C. Sun, R. Yan, and X. Chen, "Few-shot transfer learning for intelligent fault diagnosis of machine," *Measurement: Journal of the International Measurement Confederation*, vol. 166, article 108202, 2020.

[4] H. Wang, J. Xu, C. Sun, R. Yan, and X. Chen, "Intelligent fault diagnosis for planetary gearbox using time-frequency representation and deep reinforcement learning," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 2, pp. 985–998, 2022.

[5] Y. Wang and S. Wang, "Soft sensor for VFA concentration in anaerobic digestion process for treating kitchen waste based on SSAE-KELM," *IEEE Access*, vol. 9, pp. 36466–36474, 2021.

[6] H. Deng, W. X. Zhang, and Z. F. Liang, "Application of BP neural network and convolutional neural network (CNN) in bearing fault diagnosis," *IOP Conference Series: Materials Science and Engineering*, vol. 1043, no. 4, 2021.

[7] R. Bai, Q. Xu, Z. Meng, L. Cao, K. Xing, and F. Fan, "Rolling bearing fault diagnosis based on multi-channel convolution neural network and multi-scale clipping fusion data augmentation," *Measurement: Journal of the International Measurement Confederation*, vol. 184, article 109885, 2021.

[8] H. Shao, J. Lin, L. Zhang, D. Galar, and U. Kumar, "A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance," *Information Fusion*, vol. 74, pp. 65–76, 2021.

[9] Z. Liu, W. Guo, Z. Tang, and Y. Chen, "Multi-sensor data fusion using a relevance vector machine based on an ant colony for gearbox fault detection," *Sensors (Switzerland)*, vol. 15, no. 9, pp. 21857–21875, 2015.

[10] G. Cheng, X. H. Chen, X. L. Shan, H. G. Liu, and C. F. Zhou, "A new method of gear fault diagnosis in strong noise based on multi-sensor information fusion," *Journal of Vibration and Control*, vol. 22, no. 6, pp. 1504–1515, 2016.

[11] T. Han, D. Jiang, Q. Zhao, L. Wang, and K. Yin, "Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery," *Transactions of the Institute of Measurement and Control*, vol. 40, no. 8, pp. 2681–2693, 2018.

[12] Z. X. Yang, X. B. Wang, and J. H. Zhong, "Representational learning for fault diagnosis of wind turbine equipment: a multi-layered extreme learning machines approach," *Energies*, vol. 9, no. 6, article 379, 2016.

[13] M. Cerrada, G. Zurita, D. Cabrera, R. V. Sánchez, M. Artés, and C. Li, "Fault diagnosis in spur gears based on genetic algorithm and random forest," *Mechanical Systems and Signal Processing*, vol. 70, pp. 87–103, 2016.

[14] Z. Wang, W. Du, J. Wang et al., "Research and application of improved adaptive MOMEDA fault diagnosis method," *Measurement*, vol. 140, pp. 63–75, 2019.

[15] T. Liu, Z. Luo, J. Huang, and S. Yan, "A comparative study of four kinds of adaptive decomposition algorithms and their applications," *Sensors*, vol. 18, no. 7, article 2120, 2018.

[16] Q. Ni, J. C. Ji, K. Feng, and B. Halkon, "A fault information-guided variational mode decomposition (FIVMD) method for rolling element bearings diagnosis," *Mechanical Systems and Signal Processing*, vol. 164, article 108216, 2022.

[17] J. Cheng, Y. Yang, N. Hu, Z. Cheng, and J. Cheng, "A noise reduction method based on adaptive weighted symplectic geometry decomposition and its application in early gear fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 149, article 107351, 2021.

[18] K. Feng, W. A. Smith, P. Borghesani, R. B. Randall, and Z. Peng, "Use of cyclostationary properties of vibration signals to identify gear wear mechanisms and track wear evolution," *Mechanical Systems and Signal Processing*, vol. 150, article 107258, 2021.

[19] W. Fan, G. Cai, Z. K. Zhu, C. Shen, W. Huang, and L. Shang, "Sparse representation of transients in wavelet basis and its application in gearbox fault feature extraction," *Mechanical Systems and Signal Processing*, vol. 56, pp. 230–245, 2015.

[20] Y. Miao, M. Zhao, J. Lin, and Y. Lei, "Application of an improved maximum correlated kurtosis deconvolution method for fault diagnosis of rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 92, pp. 173–195, 2017.

[21] X. Zhang, J. Zhao, X. Ni, F. Sun, and H. Ge, "Fault diagnosis for gearbox based on EMD-MOMEDA," *International Journal of Systems Assurance Engineering and Management*, vol. 10, no. 4, pp. 836–847, 2019.

[22] S. Zhou, M. Xiao, P. Bartos, M. Filip, and G. Geng, "Remaining useful life prediction and fault diagnosis of rolling bearings based on short-time Fourier transform and convolutional neural network," *Shock and Vibration*, vol. 2020, Article ID 8857307, 14 pages, 2020.

[23] Y. Yao, H. Wang, S. Li et al., "End-to-end convolutional neural network model for gear fault diagnosis based on sound signals," *Applied Sciences (Switzerland)*, vol. 8, no. 9, article 1584, 2018.

[24] Q. Hu, A. Qin, Q. Zhang, J. He, and G. Sun, "Fault diagnosis based on weighted extreme learning machine with wavelet packet decomposition and KPCA," *IEEE Sensors Journal*, vol. 18, no. 20, pp. 8472–8483, 2018.

[25] T. Xie, X. Huang, and S. -K. Choi, "Intelligent mechanical fault diagnosis using multisensor fusion and convolution neural network," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3213–3223, 2022.

[26] T. Li, Z. Zhao, C. Sun, R. Yan, and X. Chen, "Adaptive channel weighted CNN with multisensor fusion for condition monitoring of helicopter transmission system," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8364–8373, 2020.

[27] R. Cao and A. Yunusa-Kaltungo, "An automated data fusion-based gear faults classification framework in rotating machines," *Sensors*, vol. 21, no. 9, article 2957, 2021.

[28] P. Yang, C. Wen, H. Geng, and P. Liu, "Intelligent fault diagnosis method for blade damage of quad-rotor UAV based on stacked pruning sparse denoising autoencoder and convolutional neural network," *Machines*, vol. 9, no. 12, article 360, 2021.

[29] S. Ullah and D. H. Kim, "Benchmarking Jetson platform for 3D point-cloud and hyper-spectral image classification," in *2020 IEEE International conference on big data and smart computing (BigComp)*, pp. 477–482, Busan, Korea (South), February 2020.

[30] G. Zhang, S. Zhao, W. Li, Q. Du, Q. Ran, and R. Tao, "HTD-Net: a deep convolutional neural network for target detection in hyperspectral imagery," *Remote Sensing*, vol. 12, no. 9, article 1489, 2020.

[31] S. Liu, X. Qian, H. Wan, Z. Ye, S. Wu, and X. Ren, "NPC three-level inverter open-circuit fault diagnosis based on adaptive electrical period partition and random forest," *Journal of Sensors*, vol. 2020, Article ID 9206579, 18 pages, 2020.

[32] T. Han and D. Jiang, "Rolling bearing fault diagnostic method based on VMD-AR model and random forest classifier," *Shock and Vibration*, vol. 2016, Article ID 5132046, 11 pages, 2016.

[33] X. Zhang, Q. Miao, H. Zhang, and L. Wang, "A parameter-adaptive VMD method based on grasshopper optimization algorithm to analyze vibration signals from rotating machinery," *Mechanical Systems and Signal Processing*, vol. 108, pp. 58–72, 2018.

[34] Y. He, H. Wang, H. Xue, and T. Zhang, "Research on unknown fault diagnosis of rolling bearings based on parameter-adaptive maximum correlation kurtosis deconvolution," *Review of Scientific Instruments*, vol. 92, no. 5, article 055103, 2021.

[35] H. Chen, Y. Xiong, S. Li, Z. Song, Z. Hu, and F. Liu, "Multi-sensor data driven with PARAFAC-IPSO-PNN for identification of mechanical nonstationary multi-fault mode," *Machines*, vol. 10, no. 2, article 155, 2022.

[36] M. Ye, X. Yan, and M. Jia, "Rolling bearing fault diagnosis based on VMD-MPE and PSO-SVM," *Entropy*, vol. 23, no. 6, article 762, 2021.

[37] M. Zellagui, A. Lasmari, S. Settoul, C. Z. El-Bayeh, and R. Chenni, "Assessment integration of hybrid PV-DSTATCOM-BES-DG system in EDS under uncertainties using chaotic adaptive inertia weight PSO algorithms," in *2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, pp. 1–8, Bucharest, Romania, March 2021.

[38] D. Połap and M. Woźniak, "Red fox optimization algorithm," *Expert Systems with Applications*, vol. 166, article 114107, 2021.

[39] J. Xiong, D. Yu, S. Liu, L. Shu, X. Wang, and Z. Liu, "A review of plant phenotypic image recognition technology based on deep learning," *Electronics*, vol. 10, no. 1, pp. 1–19, 2021.

[40] Y. Mao, D. Xiao, J. Cheng et al., "Multigrades classification model of magnesite ore based on SAE and ELM," *Journal of Sensors*, vol. 2017, Article ID 9846181, 9 pages, 2017.

[41] X. Wang, Z. Chu, B. Han, J. Wang, G. Zhang, and X. Jiang, "A novel data augmentation method for intelligent fault diagnosis under speed fluctuation condition," *IEEE Access*, vol. 8, pp. 143383–143396, 2020.

[42] M. Sun, H. Wang, P. Liu, S. Huang, P. Wang, and J. Meng, "Stack autoencoder transfer learning algorithm for bearing fault diagnosis based on class separation and domain fusion," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 3, pp. 3047–3058, 2022.

[43] S. Pang, X. Yang, X. Zhang, and Y. Sun, "Fault diagnosis of rotating machinery components with deep ELM ensemble induced by real-valued output-based diversity metric," *Mechanical Systems and Signal Processing*, vol. 159, article 107821, 2021.

[44] Z. Dong, Y. Liu, J. Kang, and S. Zhang, "A novel generative method for machine fault diagnosis," *Journal of Sensors*, vol. 2022, Article ID 5420478, 11 pages, 2022.

[45] Y. Li, X. Chen, J. Yu, and X. Yang, "A fusion frequency feature extraction method for underwater acoustic signal based on variational mode decomposition, duffing chaotic oscillator and a kind of permutation entropy," *Electronics (Switzerland)*, vol. 8, no. 1, article 61, 2019.

[46] X. Yan and M. Jia, "A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing," *Neurocomputing*, vol. 313, pp. 47–64, 2018.

[47] X. Sui, K. Wan, and Y. Zhang, "Pattern recognition of SEMG based on wavelet packet transform and improved SVM," *Optik*, vol. 176, pp. 228–235, 2019.

[48] S. Ma, M. Chen, J. Wu, Y. Wang, B. Jia, and Y. Jiang, "Intelligent fault diagnosis of HVCB with feature space optimization-based random forest," *Sensors*, vol. 18, no. 4, article 1221, 2018.

[49] J. Tian, L. Liu, F. Zhang, Y. Ai, R. Wang, and C. Fei, "Multi-domain entropy-random forest method for the fusion diagnosis of inter-shaft bearing faults with acoustic emission signals," *Entropy*, vol. 22, no. 1, article 57, 2020.

[50] N. Liu, B. Liu, and C. Xi, "Fault diagnosis method of rolling bearing based on the multiple features of LMD and random forest," *IOP Conference Series: Materials Science and Engineering*, vol. 892, no. 1, 2020.

[51] R. H. Lin, Z. X. Pei, Z. Z. Ye, C. C. Guo, and B. D. Wu, "Hydrogen fuel cell diagnostics using random forest and enhanced feature selection," *International Journal of Hydrogen Energy*, vol. 45, no. 17, pp. 10523–10535, 2020.

[52] C. Sun, M. Ma, Z. Zhao, S. Tian, R. Yan, and X. Chen, "Deep transfer learning based on sparse autoencoder for remaining useful life prediction of tool in manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2416–2425, 2019.

[53] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," *Mechanical Systems and Signal Processing*, vol. 35, no. 1–2, pp. 108–126, 2013.

[54] F. Zhang, W. Sun, H. Wang, and T. Xu, "Fault diagnosis of a wind turbine gearbox based on improved variational mode algorithm and information entropy," *Entropy*, vol. 23, no. 7, article 794, 2021.

[55] S. Mohanty, K. Kumar Gupta, and K. Solomon Raju, "Comparative study between VMD and EMD in bearing fault diagnosis," in *2014 9th International Conference on Industrial and Information Systems (ICIIS)*, pp. 1–6, Gwalior, India, December 2014.

[56] X. Li, Z. Ma, D. Kang, and X. Li, "Fault diagnosis for rolling bearing based on VMD-FRFT," *Measurement*, vol. 155, article 107554, 2020.

[57] Y. Sun, W. Qin, Z. Zhuang, and H. Xu, "An adaptive fault detection and root-cause analysis scheme for complex industrial processes using moving window KPCA and information geometric causal inference," *Journal of Intelligent Manufacturing*, vol. 32, no. 7, pp. 2007–2021, 2021.

[58] R. Yan, F. Shen, C. Sun, and X. Chen, "Knowledge transfer for rotary machine fault diagnosis," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8374–8393, 2020.

[59] J. Liu, Y. Hu, B. Wu, and Y. Wang, "An improved fault diagnosis approach for FDM process with acoustic emission," *Journal of Manufacturing Processes*, vol. 35, pp. 570–579, 2018.

[60] Y. Qu, Y. Zhang, M. He, D. He, C. Jiao, and Z. Zhou, "Gear pitting fault diagnosis using disentangled features from

unsupervised deep learning," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 233, no. 5, pp. 719–730, 2019.

[61] L. Eren, T. Ince, and S. Kiranyaz, "A generic intelligent bearing fault diagnosis system using compact adaptive 1D CNN classifier," *Journal of Signal Processing Systems*, vol. 91, no. 2, pp. 179–189, 2019.

[62] L. M. R. Baccarini, V. V. R. Silva, B. R. de Menezes, and W. M. Caminhas, "SVM practical industrial application for mechanical faults diagnostic," *Expert Systems with Applications*, vol. 38, no. 6, pp. 6980–6984, 2011.

[63] S. Tao, T. Zhang, J. Yang, X. Wang, and W. Lu, "Bearing fault diagnosis method based on stacked autoencoder and softmax regression," in *2015 34th Chinese Control Conference (CCC)*, pp. 6331–6335, Hangzhou, China, July 2015.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7012261, 12 pages
https://doi.org/10.1155/2022/7012261

Research Article

Amplitude-Based Smoothed Sparse Phase Retrieval Algorithm

Zhuolei Xiao and Zhengyan Liu

School of Computer and Information Engineering, Fuyang Normal University, Fuyang 236037, China

Correspondence should be addressed to Zhuolei Xiao; sky_south@163.com

Received 29 March 2022; Accepted 28 May 2022; Published 9 June 2022

Academic Editor: Xianpeng Wang

Sparse phase retrieval mainly solves nonconvex and nonsmooth problems. Aiming at the nonsmooth problem in sparse phase retrieval, we propose a smoothing algorithm which is called sparse smoothed amplitude flow (SPSAF). The proposed SPSAF algorithm is an amplitude-based nonconvex sparse smoothing phase retrieval algorithm. First, the original phase retrieval loss function is smoothed without modifying the gradient in the gradient refinement stage, thereby reducing the computational complexity of the overall algorithm. Secondly, the support of the original signal is estimated by differential analysis of the gaps, and the initialization can be obtained through a carefully designed method based on this support. Finally, we get sparse estimates by gradient descent based on hard thresholding. Numerical experiments show that the proposed SPSAF algorithm has significant improvements in recovery performance, convergence speed, and sampling complexity. Further, the SPSAF algorithm is stable in noisy environments.

1. Introduction

In real-world applications, especially in the field of channel estimation and image processing, the primary signal is naturally sparse or accepts sparse representation after some known deterministic linear transformations [1–5]. When the measurements are undersampled, phase retrieval will introduce sparse assumption or direct sparse transformation for the signal [6–11]. For instance, the sparsely distributed stars observed in astronomy and the sparsely distributed atoms or distributions observed in crystallography can all be regarded as applications of sparse phase retrieval. This recovery method for solving sparse signals is called sparse phase retrieval or compressed phase retrieval, which has fundamental significance in signal recovery and reconstruction [12, 13].

Sparse prior is essential for solving sparse phase retrieval problems, and many methods for sparse phase retrieval have been developed. Based on the PhaseLift algorithm, the CPRL algorithm [14] is designed by utilizing the ℓ_1-norm regularization term, which improves the one-dimensional (1D) signal to $N \times N$ matrix, and finally obtains the sparse solution. Aiming at the application problem of the PhaseMax algorithm under sparse conditions, the SparsePhaseMax algorithm was proposed in [15, 16] by utilizing ℓ_1-norm regularization. Based on the alternating minimization algorithm, [17] proposed the Sparse AltMinPhase algorithm, which solves the sparse solution of the equation by alternating between the missing phase information and the candidate solution. [18] developed a probabilistic phase retrieval algorithm based on generalized approximate message passing. [19] proposed a greedy sparse phase retrieval algorithm (GESPAR) by combining the damped Gauss-Newton method with the two-element optimization method. It dynamically updated the support of the signal by using the two-element optimization method and solved the optimal value of the current support by using the damped Gauss-Newton method. [20] proposed a soft threshold TWF algorithm for sparse phase retrieval. Qiu and Palomar developed C-PRIME and SC-PRIME algorithms based on the majorization-minimization method [21]. They replaced the objective function approximately with the convex function. [22] proposed the sparse truncated amplitude flow algorithm (SPARTA), a sparse phase retrieval algorithm based on the TAF algorithm. The SPARTA algorithm improves the performance of sparse phase retrieval by introducing

1. **Input:** Data $\{(\mathbf{a}_i, \psi_i)\}_{i=1}^m$ and sparsity k; learning rate μ; the maximum number of iterations T; subset cardinality $|J| \coloneqq \lfloor 3m/13 \rfloor$; adaptive smoothing factor p; smoothing vector $\varepsilon_i \coloneqq \tau\psi_i$, and τ is a constant;
2. **Construct** the support index set $\tilde{S}$, which includes the index corresponding to the largest k in
$$\{\tilde{Z}_{i,j}\}_{i=1}^m \coloneqq \{1/m\textstyle\sum_{i=1}^m \psi_i^2 a_{i,j}^2\}_{i=1}^m;$$
3. **Calculate** the sparse initial estimate
$$\tilde{\mathbf{z}}_{\tilde{S}}^0 \coloneqq \arg \max_{\|\mathbf{z}\|=1} \mathbf{z}^T (1/|J| \textstyle\sum_{i\in J} r_i \mathbf{a}_{i,\tilde{S}} \mathbf{a}_{i,\tilde{S}}^T)\mathbf{z},$$
where the weighting factor $r_i = \psi_i^{1/2}$;
4. **Initialize**
$$\mathbf{z}_0 = \sqrt{\textstyle\sum_{i=1}^m \psi_i^2/m}\, \tilde{\mathbf{z}}_0$$
where $\tilde{\mathbf{z}}_0$ is obtained by zero-filling $\tilde{\mathbf{z}}_{\tilde{S}}^0$;
5. **for** $t = 0$ **to** $T - 1$ **do**
$$\mathbf{z}_{t+1} \coloneqq \mathscr{H}_k(\mathbf{z}_t - \mu \nabla \ell_{\varepsilon,p}(\mathbf{z}_t))$$
where $\mathscr{H}_k(\cdot)$ is the hard threshold operator, and the gradient
$$\nabla \ell_{\varepsilon,p}(\mathbf{z}) = 1/m \textstyle\sum_{i=1}^m \begin{array}{l}[(|\mathbf{a}_i^T\mathbf{z}|^p + \varepsilon_i^p)^{1/p} - (\psi_i^p + \varepsilon_i^p)^{1/p}] \\ (|\mathbf{a}_i^T\mathbf{z}|^p + \varepsilon_i^p)^{1/p-1} |\mathbf{a}_i^T\mathbf{z}|^{p-1} \mathbf{a}_i \operatorname{sgn}(\mathbf{a}_i^T\mathbf{z})\end{array}$$
6. **end for**
7. **Output**: $\mathbf{z}_T$.

Algorithm 1: The proposed SPSAF algorithm.

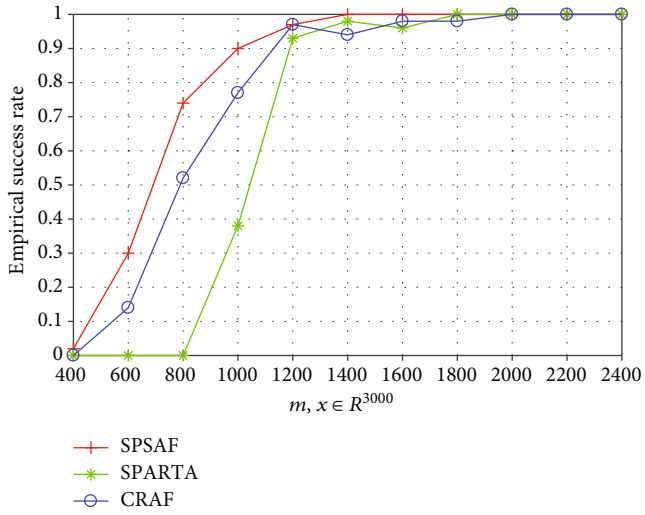

Figure 1: The success rate of signal recovery with sparsity $k = 30$ and estimated sparsity $\tilde{k} = 55$ in the real case.

truncation procedures. Zhang et al. proposed compressive reweighted amplitude flow (CRAF) based on the RAF algorithm in [23], making the gradient descent direction of sparse phase retrieval more effective and accurate by introducing a weighted program.

When phase retrieval adopts a noise-free Gaussian random measurement model, the CPRL algorithm can accurately recover any k-sparse signal from $\mathcal{O}(k^2 \log n)$ measurements, and the computational complexity is $\mathcal{O}(n^3)$ [24]. Sparse AltMinPhase and soft threshold TWF algorithm need at least $\mathcal{O}(k^2 \log n)$ measurements, and the computational complexity is $\mathcal{O}(k^2 n \log n)$. The SPARTA algorithm accurately recovers the signal from about $\mathcal{O}(k^2 \log n)$ random Gaussian measurements with a computational complexity of about $\mathcal{O}(k^2 n \log n)$, and the total running time is a positive correlation to the time required to read the data. The CRAF algorithm can accurately recover the signal from $\mathcal{O}(k^2 \log mn)$ random Gaussian measurements.

The performance of the amplitude-based phase retrieval method is better than that of the intensity-based phase retrieval algorithm in both numerical and experimental verification, which is the same in a sparse environment [25, 26]. The soft threshold TWF algorithm in [20] solves the smoothing problem in the sparse phase retrieval model based on the intensity $\{y_i = |\mathbf{a}_i^T\mathbf{z}|^2\}_{i=1}^m$. It adopts the adaptive hard threshold iterative algorithm based on compressed sensing. In the gradient optimization stage, each iteration only retains some maximum indexes.

On the other hand, the Sparse AltMinPhase algorithm in [17] estimates the support of the original signal and solves the k-sparse phase retrieval problem. The Sparse AltMinPhase algorithm solves the sparse phase retrieval problem by alternating minimization and resampling conditions and performs matrix inversion in each iteration. Numerical experiments show that resampling is a necessary condition for the Sparse AltMinPhase algorithm; that is, many measurements are needed to estimate the support accurately, and the importance of the recovery of the support for signal recovery and even the whole phase retrieval is self-evident.

Among the above methods, the illustrious amplitude-based sparse phase retrieval algorithms include the SPARTA algorithm [22] and CRAF algorithm [23]. The SPARTA algorithm applies the TAF algorithm in a sparse environment, divided into two stages: initialization and gradient refinement. Firstly, a reasonable rule is used to solve the support, and some simple power iterations are used to solve the initialization problem to obtain the sparse initialization. Through a series of truncated gradient iterations and the hard threshold of each iteration, all indexes except k

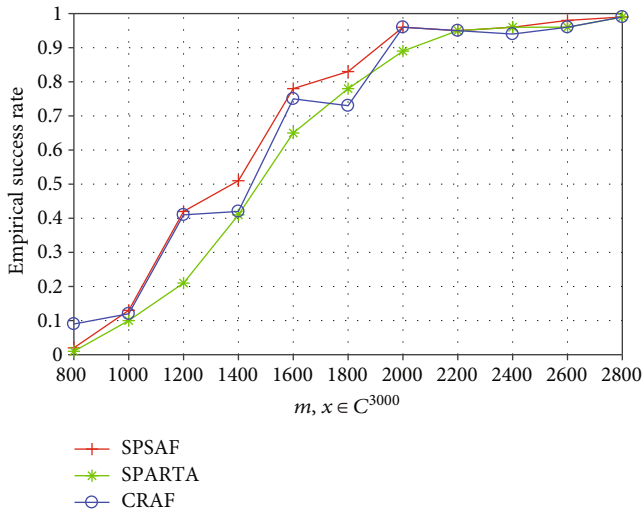

Figure 2: The success rate of signal recovery with sparsity $k = 30$ and estimated sparsity $\tilde{k} = 55$ in the complex case.

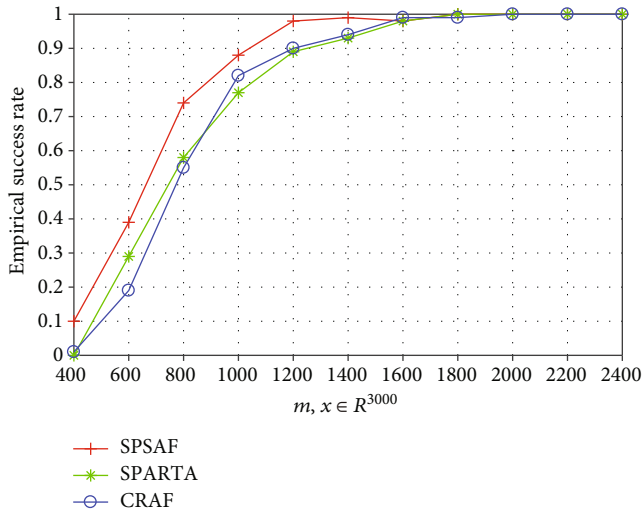

Figure 3: The success rate of signal recovery with prior sparsity $k = 30$ in the real case.

maximum values are set to zero to realize the continuous update of the initial value, where k refers to the signal sparsity. The CRAF algorithm further considers the structured sparsity pattern based on the SPARTA and RAF algorithms. It proposes the amplitude-based (block) sparse phase retrieval problem. The CRAF algorithm developed a new sparse spectrum initialization method in the initialization phase, wisely assigning negative or positive weights to each sample. By this method, the mean value of the initialization matrix obtained by the CRAF algorithm has improved performance. Then, the CRAF algorithm uses the reweighted gradient in the gradient refinement stage to gradually improve the initialization of the hard threshold iteration.

Based on the SAFPR algorithm and CRAF algorithm, we propose a new sparse phase retrieval algorithm, called sparse smoothed amplitude flow (SPSAF). The SPSAF algorithm is an amplitude-based nonconvex sparse smoothing phase

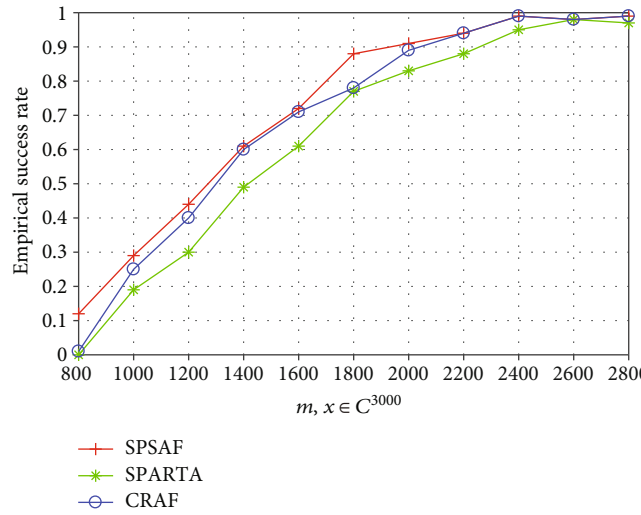

Figure 4: The success rate of signal recovery with prior sparsity $k = 30$ in the complex case.

retrieval algorithm with two stages: initialization and gradient refinement. In the initialization stage, we estimate the support by a reasonable rule. Then, the initial estimation is obtained by a carefully designed initialization method based on the support. The gradient descent method updates the initialization estimate based on the hard threshold in the gradient refinement stage. Numerical experiments show that the SPSAF algorithm is robust to additional noise in the finite support. Compared with the existing typical algorithms, the recovery performance and speed of the SPSAF algorithm are significantly improved.

The rest of this paper is organized as follows. In Section 2, the sparse smoothing phase retrieval problem is formulated. Section 3 introduces the SPSAF algorithm in initialization and gradient refinement two stages, respectively. Numerical experiments are provided in Section 4. Finally, Section 5 summarizes this paper.

2. Sparse Phase Retrieval Problem

2.1. Amplitude-Based Sparse Phase Retrieval. The amplitude-based sparse phase retrieval problem is a kind of problem to reconstruct the sparse signal from the phaseless measurements [19, 20, 22]. Mathematically, we can describe the sparse phase retrieval problem as a set of phaseless quadratic equations, namely,

$$\psi_i = \left|\mathbf{a}_i^T \mathbf{x}\right|, \quad 1 \le i \le m \text{ subject to } \|\mathbf{x}\|_0 \le k, \tag{1}$$

where k represents the level of sparsity. $\|\cdot\|_0$ express a zero-norm operator, that is, the number of nonzero elements. The signal $\mathbf{x}$ is k-sparse, consisting of up to k nonzero elements and $n - k$ zero elements. The goal of amplitude-based sparse phase retrieval is to reconstruct the sparse signal $\mathbf{x} \in \mathbb{R}^n$ or $\mathbb{C}^n$ based on the given measurements ψ_i.

For theoretical analysis, suppose sparsity k is prior known. In the real case, the number of measurements required to the k-sparse signal $\mathbf{x}$ recovery is at least $m \ge$

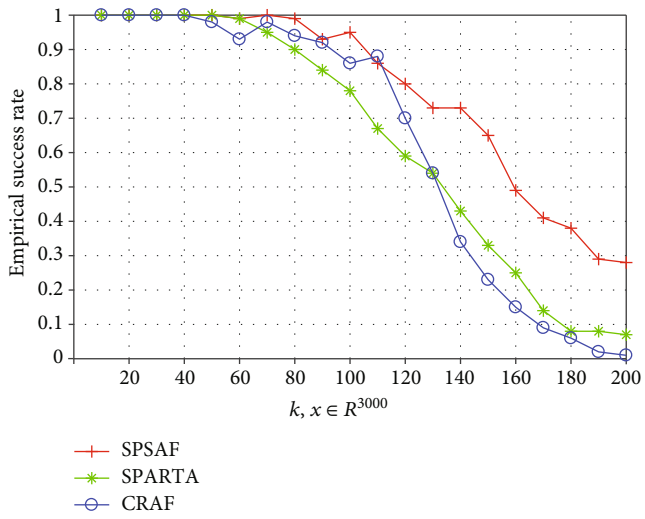

Figure 5: Comparison of signal recovery success rate under different sparsity k selection.

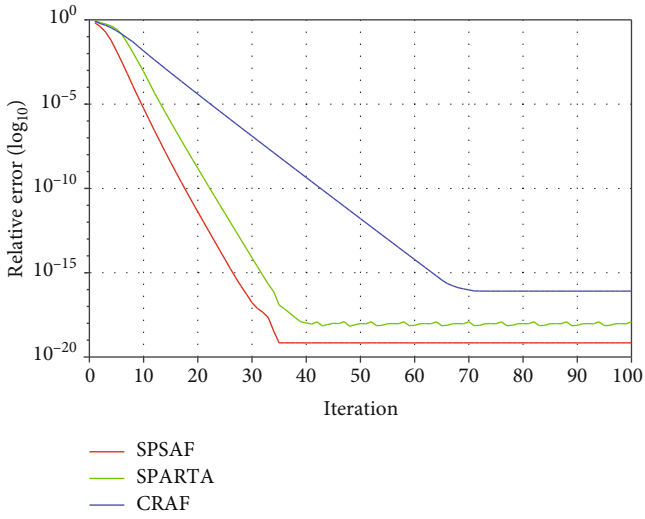

Figure 6: Comparison of convergence rate in the real case.

min $\{2k, 2n-1\}$. [27] pointed out that k-sparse signals can be reconstructed by selecting $m \geq \min\{2k, 2n-1\}$ measurements on general positions in real number space $\mathbb{R}^n$. Similarly, in the complex cases, the number of measurements required for k-sparse phase retrieval is at least $m \geq 4k-2$ [28]. Due to the lack of more phase information in the noise environment, stable sparse or compressed phase retrieval needs as many measurements as the relevant compressed sensing problem. Therefore, as with compressed sensing, stable sparse phase retrieval requires at least $\mathcal{O}(k \log(n/k))$ measurements [29]. [30] proves that sparse phase retrieval requires $\mathcal{O}(k \log(n/k))$ measurements to recover sparse signals stably in the real case.

We adopt a real-valued Gaussian model to analyze sparse phase retrieval in this paper. The model assumes sparse signal $\mathbf{x} \in \mathbb{R}^n$ and i.i.d. standard Gaussian sensing vector $\mathbf{a}_i \sim \mathcal{N}(0, I_n), i = 1, \cdots, m$. Nevertheless, the proposed algorithm is also applicable to the complex-valued Gaussian

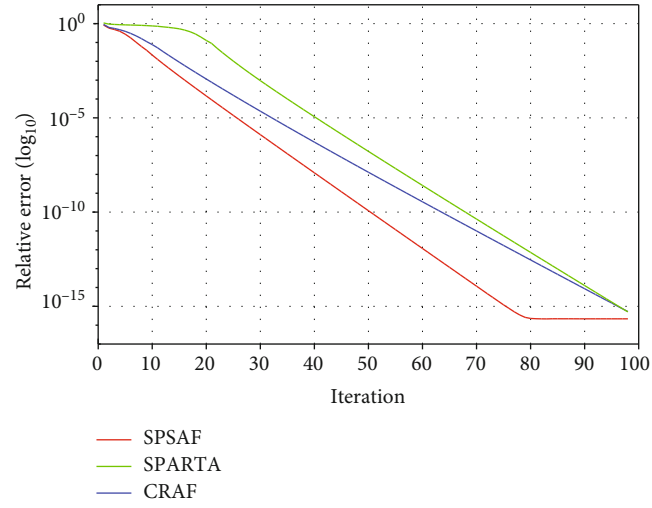

Figure 7: Comparison of convergence rate in the complex case.

phase retrieval model, namely, signal $\mathbf{x} \in \mathbb{C}^n$ and i.i.d. standard Gaussian sensing vector $\mathbf{a}_i \sim \mathcal{CN}(0, I_n) = \mathcal{N}(0, I_n/2) + j\mathcal{N}(0, I_n/2), i = 1, \cdots, m$. Given the data $\{(\mathbf{a}_i, \psi_i)\}_{i=1}^m$ and assuming that the equation has a unique k-sparse solution, our goal is to develop a simple and efficient algorithm that can recover any k-sparse n-dimensional signal $\mathbf{x}$ from as few amplitude measurements as possible in (1).

Using the least square criterion, the problem of reconstructing k-sparse solution from phaseless quadratic equation (1) can be naturally transformed into the problem of minimizing the amplitude-based empirical loss function:

$$\begin{aligned} \min_{\mathbf{z}} \ell(\mathbf{z}) &:= \frac{1}{2m} \sum_{i=1}^{m} \left(\psi_i - \left|\mathbf{a}_i^T \mathbf{z}\right|\right)^2 \\ \text{s.t.} \|\mathbf{z}\|_0 &\leq k \\ S(\mathbf{z}) &\subseteq [n] := \{1, 2, \cdots, n\} \\ \mathbf{z} &\in \mathbb{R}^n, \end{aligned} \tag{2}$$

where $S(\mathbf{z})$ represents the support of $\mathbf{z}$ and k is the signal sparsity. The objective function $\ell(\mathbf{z})$ in (2) is nonconvex and nonsmooth, which is subject to the combined constraint of zero norm $\|\mathbf{z}\|_0 \leq k$. Therefore, the sparse phase retrieval optimization problem is NP-hard, which is difficult to solve in the calculation [31–33]. The methods to solve these problems include the following:

(1) In the initialization stage, the support of the signal needs to be accurately estimated, and the accurate support information can correctly restore the original sparse signal. The selection of initialization methods is also critical. At present, there are many initialization methods, including spectral initialization, orthogonal promotion initialization, and reweighted maximum correlation initialization, which can be applied to nonconvex sparse phase retrieval models

Table 1: Comparison of convergence speed and time cost.

Algorithms	Real case		Complex case	
	Iterations	Time (s)	Iterations	Time (s)
SPSAF	26	0.1035	75	1.3525
CRAF	64	0.2013	94	1.6509
SPARTA	27	0.1352	96	1.8528

(2) In the gradient refinement stage, inspired by the iterative hard threshold algorithm of compressed sensing [34, 35], in the descent process, the adaptive hard threshold process that only retains some maximum indexes in each iteration is proved to be effective

2.2. Smoothing Scheme of Sparse Phase Retrieval. Due to the existence of modulus in the sparse phase retrieval objective function (2), there will be dramatic changes and discontinuous changes in the gradient function at point $\mathbf{z}$ when $\mathbf{a}_i^T\mathbf{z} = \mathbf{0}$, which makes the normal gradient descent method ineffective. Therefore, the sparse optimization problem in (2) is nonconvex and nonsmooth. Usually, the way to solve this kind of nonsmooth function is to bypass this nonsmooth function, such as the SPARTA or CRAF algorithm. [22] points out that if $|\mathbf{a}_i^T\mathbf{z}|/\psi_i$ is less than a certain threshold, the SPARTA algorithm will remove these "bad" gradient components by the truncation program. The CRAF algorithm [23] is based on the same principle, but it is not a simple removal of these "bad" gradient components, but according to the size of $|\mathbf{a}_i^T\mathbf{z}|/\psi_i$ to reweight gradient components, the effect is better than the SPARTA algorithm. However, the above two methods will introduce additional calculation for the gradient, resulting in changes in the search direction and thus affecting the update of the entire gradient direction.

In recent years, there have been many research on the nonconvex phase retrieval algorithm, among which the optimization of the nonsmooth term in the empirical loss function has made great progress. [36] proposed a mixed optimization method to modify the empirical loss function and solve the nonsmooth problem through the classical proximal method. Pinilla et al. [37] proposed to smooth the phase retrieval problem. They introduced a special smoothing function to replace the nonsmooth term in the original loss empirical function and solved the related problems using the projection conjugate gradient method. [38] adopted the same smoothing strategy. By introducing the smoothing approximation function, it replaced the nonsmooth term and used the same operation to put forward a new smoothing loss empirical function. [39] introduced two smoothing functions for the nonconvex phase retrieval loss function and proposed two smoothing algorithms. [40] further improves the smoothing operation for the loss function and proposes a faster stochastic smoothing phase retrieval algorithm.

By introducing the smoothing function, we approximate and replace the absolute value function $|\mathbf{a}_i^T\mathbf{x}|$ in (2) to obtain an approximate sparse smoothing phase retrieval loss empirical function. Therefore, no additional operation on the gradient is needed in the iterative update process, which reduces the computational complexity. There are many substitution methods for the absolute value function. We mainly use the following method to smooth the phase retrieval loss function (2).

We first define the concept of smoothing function to be used.

Definition 1. Define the absolute value function $f(\mathbf{u}) = |\mathbf{u}|$. If $g_\varepsilon(\mathbf{u})$ is smooth for any real constant $\varepsilon \geq 0$ *in* $\mathbb{C}^n$ and for any fixed $\mathbf{u} \in \mathbb{C}^n$

$$\lim_{\varepsilon \longrightarrow 0} g_\varepsilon(\mathbf{u}) = f(\mathbf{u}), \tag{3}$$

then $g_\varepsilon(\mathbf{u})$: $\mathbb{C}^n \times \mathbb{R} \longrightarrow \mathbb{R}$ is the smoothing function of $f(\mathbf{u})$.

By Definition 1, the smooth substitution function of the absolute value function $f(u) = |u|$ is defined as

$$g_{\varepsilon,p}(u) := \left(|u|^p + \varepsilon^p\right)^{1/p}, \tag{4}$$

where the relaxation factor $\varepsilon \geq 0$ is a real constant and the smoothing parameter $p > 0$.

As a smooth approximation function of the absolute value function $f(u)$, $g_{\varepsilon,p}(u)$ must have the following properties:

(1) $g_{\varepsilon,p}(u)$ is a Lipschitz continuous function

(2) $g_{\varepsilon,p}(u)$ is even, namely, $g_{\varepsilon,p}(u) = g_{\varepsilon,p}(-u)$

(3) $g_{\varepsilon,p}(u)$ uniformly converges to $f(u)$ in $\mathbb{R}$

Proof.

(1) By Definition 1, $g_{\varepsilon,p}(u)$ is smooth for any real constant $\varepsilon \geq 0$ in $\mathbb{C}^n$, so we can have

$$\nabla g_{\varepsilon,p}(u) = \frac{|u|^{p-1}}{\left(|u|^p + \varepsilon^p\right)^{1-1/p}} = \left(\frac{|u|}{\left(|u|^p + \varepsilon^p\right)^{1/p}}\right)^{p-1}. \tag{5}$$

Since the real constants $\varepsilon \geq 0$, $p > 0$, therefore,

$$\left(|u|^p + \varepsilon^p\right)^{1/p} \geq \left(|u|^p\right)^{1/p} = |u|. \tag{6}$$

According to (5) and (6), it can be obtained that

$$\left|\nabla g_{\varepsilon,p}(u)\right| = \left|\frac{g_{\varepsilon,p}(u_1) - g_{\varepsilon,p}(u_2)}{u_1 - u_2}\right| \leq 1. \tag{7}$$

Therefore, $g_{\varepsilon,p}(u)$ is a Lipschitz continuous function with bounded first-order function, and its Lipschitz constant is 1.

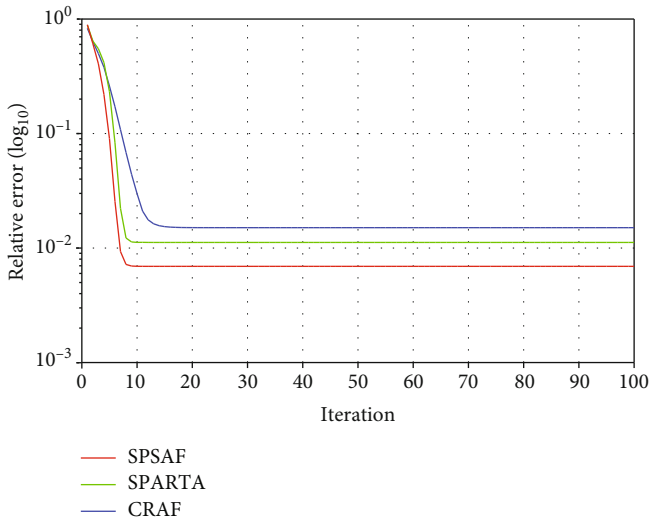

Figure 8: Relative error of additional noise in the real case.

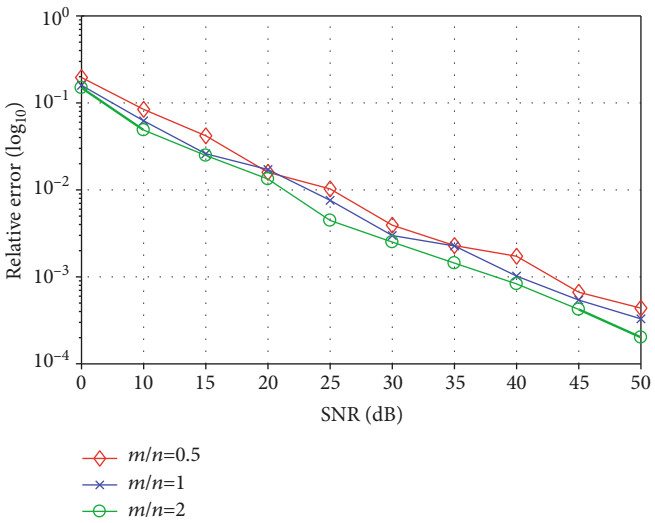

Figure 9: Relationship between relative error and SNR.

(2) By definition of the absolute value, it is obvious that $g_{\varepsilon,p}(u) = g_{\varepsilon,p}(-u) = \left(|u|^p + \varepsilon^p\right)^{1/p}$

(3) According to Definition 1 and (4), we can have

$$\left|g_{\varepsilon,p}(u) - f(u)\right| = \left|\left(|u|^p + \varepsilon^p\right)^{1/p} - |u|\right|. \tag{8}$$

According to the Minkowski inequality, it can be obtained that

$$\left|g_{\varepsilon,p}(u) - f(u)\right| = \left|\left(|u|^p + \varepsilon^p\right)^{1/p} - |u|\right| \le \left|\left(|u|^p\right)^{1/p} + \left(\varepsilon^p\right)^{1/p} - |u|\right| = \varepsilon. \tag{9}$$

Therefore, $g_{\varepsilon,p}(u)$ can converge uniformly to $f(u)$ and only depends on the value of ε. □

Let $\mathbf{u} = |\mathbf{a}_i^T \mathbf{z}|$, according to (2) and (4), the smoothing sparse phase retrieval loss function can be expressed as

$$\begin{aligned} &\min_{\mathbf{z}} \frac{1}{2m} \sum_{i=1}^{m} \left(\psi_i - \left|\mathbf{a}_i^T \mathbf{z}\right|_{\varepsilon,p}\right)^2 \\ &\text{s.t.} \|\mathbf{z}\|_0 \le k \\ &S(\mathbf{z}) \subseteq [n] \coloneqq \{1, 2, \cdots, n\} \\ &\mathbf{z} \in \mathbb{R}^n. \end{aligned} \tag{10}$$

To ensure that the introduction of $g_{\varepsilon,p}(|\mathbf{a}_i^T \mathbf{z}|)$ does not affect the global minimum of the original nonconvex loss function (2), the measurement ψ_i adopts the same smoothing strategy, i.e.,

$$(\psi_i)_{\varepsilon,p} \coloneqq \left(\psi_i^p + \varepsilon_i^p\right)^{1/p}. \tag{11}$$

Due to the fact that smoothing vector ε_i is a positive correlation to the amplitude ψ_i of signal $\mathbf{x}$, we approximate it as

$$\varepsilon_i := \tau\psi_i, \tag{12}$$

where the coefficient τ is a constant.

Combining (10) and (11), we can obtain the amplitude-based sparse smoothing phase retrieval least-square optimization problem:

$$\begin{aligned}\min_{\mathbf{z}} \ell_{\varepsilon,p}(\mathbf{z}) := \frac{1}{2m}\sum_{i=1}^{m}\left(\left(\left|\mathbf{a}_i^T\mathbf{z}\right|^p+\varepsilon_i^p\right)^{1/p}-\left(\psi_i^p+\varepsilon_i^p\right)^{1/p}\right)^2\\ \text{s.t.}\|\mathbf{z}\|_0 \le k\\ S(\mathbf{z}) \subseteq [n] := \{1,2,\cdots,n\}\\ \mathbf{z} \in \mathbb{R}^n.\end{aligned} \tag{13}$$

Function $\ell_{\varepsilon,p}(\mathbf{z})$ is an approximate smoothing function of the original loss function $\ell(\mathbf{z})$. Both of $\ell_{\varepsilon,p}(\mathbf{z})$ and $\ell(\mathbf{z})$ have the same global minimum. When the smoothing relaxation factor $\varepsilon_i \longrightarrow 0$ or the smoothing parameter $p \longrightarrow 1$, $\ell_{\varepsilon,p}(\mathbf{z})$ will be reduced to the original function $\ell(\mathbf{z})$.

3. The Proposed SPSAF Algorithm

To solve the smoothing sparse phase retrieval least-square optimization problem (13), we propose a new smoothing sparse phase retrieval algorithm, named SPSAF. It can be classified into two periods: initialization and gradient refinement, which are described in detail.

3.1. Sparse Weighted Spectrum Initialization. Based on the weighted initialization method proposed in [41], the SPSAF algorithm proposes an improved sparse weighted initialization estimation method. Compared with the orthogonal promotion initialization method used in the SPARTA algorithm, this method abandons the truncation procedure. It adopts the reweighted strategy so that the information of all samples is effectively applied. We divide the initialization algorithm of SPSAF into three parts, including the general initialization method, support recovery method, and sparse initialization method.

3.1.1. General Initialization Method. First of all, assume $\|\mathbf{x}\| = 1$ without loss of generality. It can be seen from [25, 41] that there is a certain correlation between $\mathbf{a}_i$ and $\mathbf{x}$. The larger the amplitude $\psi_i = |\mathbf{a}_i^T\mathbf{z}|$ of the inner product of $\mathbf{a}_i$ and $\mathbf{x}$ is, the higher the correlation between $\mathbf{a}_i$ and the unknown solution $\mathbf{x}$ is, so it carries more profitable direction information of signal $\mathbf{x}$. Similar to the weighted maximum correlation method in [41], the corresponding sequence of $\mathbf{a}_i$ can be obtained by sorting the amplitude ψ_i, and then, the unknown solution $\mathbf{x}$ can be estimated.

Let set $J := \{1, 2,\cdots,m\}$ denote the set of indexes that participate in the calculation initialization of the picking sensing vector $\mathbf{a}_i$. The vector $\mathbf{a}_i$ corresponding to the largest $|J|$ measurements $\{\psi_{[i]}\}_{1\le i\le|J|}$ have the largest correlation with the real solution $\mathbf{x}$, where $|J|$ is an integer of m. Therefore, we can estimate the real solution $\mathbf{x}$ by indexing $\mathbf{a}_i$ with $|J|$, that is,

$$\tilde{\mathbf{z}}_0 := \arg\max_{\|\mathbf{z}\|=1}\frac{1}{|J|}\sum_{i\in J}\left|\mathbf{a}_i^T\mathbf{z}\right|^2 = \arg\max_{\|\mathbf{z}\|=1}\mathbf{z}^T\left(\frac{1}{|J|}\sum_{i\in J}\mathbf{a}_i\mathbf{a}_i^T\right)\mathbf{z}. \tag{14}$$

Secondly, to further strengthen the ability of $\mathbf{a}_i$ to point to the real solution $\mathbf{x}$, the initialization method uses the reweighted strategy to add weights to different $\{\mathbf{a}_i\}_{i\in J}$. Sorting index set J according to the size of $\{\psi_{[i]}\}_{1\le i\le|J|}$. The larger J corresponds to the larger weight, and the smaller index corresponds to the smaller weight or even negative weight, namely,

$$\tilde{\mathbf{z}}_0 := \arg\max_{\|\mathbf{z}\|=1}\mathbf{z}^T\left(\frac{1}{|J|}\sum_{i\in J}r_i\mathbf{a}_i\mathbf{a}_i^T\right)\mathbf{z}, \tag{15}$$

where $r_i \in \mathbb{R}$ is the weight corresponding to different indexes. According to the size of ψ_i, we define the weight

$$r_i = \psi_i^{1/2}. \tag{16}$$

When $\|\mathbf{x}\| \neq 1$, utilizing the strong law of large numbers and rotation invariance of Gaussian sampling vector $\mathbf{a}_i$, we can get the initiation

$$\mathbf{z}_0 = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\psi_i^2}\tilde{\mathbf{z}}_0. \tag{17}$$

From Theorem 1 in [23], we can see that given a constant $\delta_0 \in (0, 1)$, there exist constants $c_0 > 0$ and C_0, and when $m \ge C_0 n$, the error $\text{dist}(\mathbf{z}_0, \mathbf{x})$ between the initial estimate $\mathbf{z}_0$ and the real solution $\mathbf{x}$ meets

$$\text{dist}(\mathbf{z}_0, \mathbf{x}) \le \delta_0\|\mathbf{x}\|_2, \tag{18}$$

with probability exceeding $1 - 10\exp(-c_0 m)$.

Finally, it is worth emphasizing that since the signal $\mathbf{x}$ is k-sparse and $k \ll n$, the initialization method in (15) becomes

$$\begin{aligned}\tilde{\mathbf{z}}_0 := \arg\max_{\|\mathbf{z}\|=1}\mathbf{z}^T\left(\frac{1}{|J|}\sum_{i\in J}r_i\mathbf{a}_i\mathbf{a}_i^T\right)\mathbf{z}\\ \text{s.t.}\|\mathbf{z}\|_0 \le k,\end{aligned} \tag{19}$$

by utilizing ℓ_0 regularization to represent sparse prior information.

3.1.2. Accurate Support Recovery. When the measurements are undersampled, the number of measurements is less than

the dimension of the signal; that is, $m < n$. At this time, the sparse assumption of the signal needs to be introduced.

Without loss of generality, we set the support $S \subseteq [n] \coloneqq \{1,\cdots,n\}$ of unknown k-sparse solution $\mathbf{x}$, and $|S| = k \ll n$. According to [22], we define random variables:

$$Z_{i,j} \coloneqq \psi_i^2 a_{i,j}^2, \quad j = 1, \cdots, n. \tag{20}$$

For the normalized Gaussian sensing vector $\mathbf{a}_i \sim \mathcal{N}(0, I_n)$, let p be a positive integer; we can get

$$\mathbb{E}\left(a_{i,j}^{2p}\right) = (2p-1)!!, \tag{21}$$

where !! denotes two-order multiplication. According to the rotation invariance of Gaussian distribution, it can be proved that for all $1 \le j \le n$,

$$\mathbb{E}\left(Z_{i,j}\right) = \mathbb{E}\left(\left(\mathbf{a}_i^T\mathbf{x}\right)^2 a_{i,j}^2\right) = \mathbb{E}\left(a_{i,j}^4 x_j^2 + \left(\mathbf{a}_{i,/j}^T\mathbf{x}_{/j}\right)^2 a_{i,j}^2\right) = 3x_j^2 + \left\|\mathbf{x}_{/\mathbf{j}}\right\|_2^2 = 2x_j^2 + \|\mathbf{x}\|_2^2, \tag{22}$$

where $\mathbf{a}_{i,/j}^T, \mathbf{x}_{/j} \in \mathbb{R}^{n-1}$ is $\mathbf{a}_i^T, \mathbf{x} \in \mathbb{R}^n$ after removing the jth element.

If index $j \in S$ which means the corresponding $\mathbf{x}_j$ is a nonzero term, that is, $\mathbf{x}_j \ne 0$, then (22) can be converted to

$$\mathbb{E}\left(Z_{i,j}\right) = 2x_j^2 + \|\mathbf{x}\|_2^2, \quad j \in S. \tag{23}$$

Conversely, if $j \notin S$ which means $\mathbf{x}_j = 0$, we can get

$$\left(Z_{i,j}\right) = \|\mathbf{x}\|_2^2, \quad j \notin S. \tag{24}$$

According to formulas (23) and (24), we can find that for $j \notin S$ and $j \notin S$, the expected value of $Z_{i,j}$ has at least $2x_j^2$ intervals, which we call the gap. As long as the gap $2x_j^2$ is large enough, the support S can be restored accurately in this way [22]. That is, the index set corresponding to k maximum $\mathbb{E}(Z_{i,j})$ can restore the support S of the original signal $\mathbf{x}$. However, since $\{\mathbb{E}(Z_{i,j})\}$ is actually unavailable, it is replaced by their independent implementation. At the same time, following the law of strong numbers, the sample mean should approach a whole $\tilde{Z}_{i,j}$.

To estimate the support S of the original signal $\mathbf{x}$, we first calculate the empirical estimate of the so-called sample mean

$$\tilde{Z}_{i,j} \coloneqq \frac{1}{m}\sum_{i=1}^{m} Z_{i,j} = \frac{1}{m}\sum_{i=1}^{m} \psi_i^2 a_{i,j}^2 \tag{25}$$

as the expected $\mathbb{E}(Z_{i,j})$. The larger $\tilde{Z}_{i,j}$, the greater probability of nonzero the corresponding element x_j is. Therefore, we need to collect indexes corresponding to k maximums in $\{\tilde{Z}_{i,j}\}$, which form an estimated support $\tilde{S}$ that can be expressed as

$$\tilde{S} \coloneqq \left\{ j \in [n] \middle| \text{indexes corresponding to } k\text{maximums in } \left\{\tilde{Z}_{i,j}\right\} \right\}. \tag{26}$$

[22] proves that to improve the probability of accurately recovering the support S, $\mathcal{O}(k^2 \log n)$ measurements are needed. At the same time, in order to ensure the separation of the index of the support, the smallest nonzero item

$$x_{\min} \coloneqq \min_{j \in S} \left|\mathbf{x}_j\right| \tag{27}$$

of the signal is approximately $1/\sqrt{k}\|\mathbf{x}\|_2$; that is, for some constants $C_0 > 0$, there are

$$x_{\min} = \frac{C_0}{\sqrt{k}}\|\mathbf{x}\|_2. \tag{28}$$

3.1.3. Sparse Initialization Method. After obtaining the support domain estimate $\tilde{S}$ of the original signal, we can estimate the initialization according to $\tilde{S}$. Specifically, for all $1 \le i \le m$, we rewrite the measurements ψ_i as

$$\psi_i = \left|\mathbf{a}_i^T\mathbf{x}\right| = \left|\mathbf{a}_{i,\tilde{S}}^T\mathbf{x}_{\tilde{S}}\right|, \tag{29}$$

where $\mathbf{a}_{i,\tilde{S}}$ includes the element of $\mathbf{a}_{i,j}$ whose index $j \in \tilde{S}$, and $\mathbf{x}_{\tilde{S}}$ is the element of $\mathbf{x}$ whose index $j \in \tilde{S}$.

Apply the general weighted initialization method in (19) to the reduced data $\{\mathbf{a}_{i,\tilde{S}}, \psi_i\}_{i=1}^m$, that is,

$$\begin{aligned} \tilde{\mathbf{z}}_{\tilde{S}}^0 \coloneqq \arg \max_{\|\mathbf{z}\|=1} & \mathbf{z}^T\left(\frac{1}{|J|}\sum_{i\in J} r_i \mathbf{a}_{i,\tilde{S}}\mathbf{a}_{i,\tilde{S}}^T\right)\mathbf{z} \\ \text{s.t.} & \|\mathbf{z}\|_0 \le k. \end{aligned} \tag{30}$$

By zero-filling the element corresponding to the index $j \notin \tilde{S}$ of $\tilde{\mathbf{z}}_{\tilde{S}}^0$, we can construct k-sparse n-dimensional initialization $\tilde{\mathbf{z}}_0$. When $\|\mathbf{x}\| \ne 1$, the final k-sparse initialization $\mathbf{z}_0$ can be obtained through the norm estimation of $\mathbf{x}$ in (17).

3.2. Thresholded Gradient Stage. After obtaining an accurate k-sparse initialization $\mathbf{z}_0$, we put it into the gradient iteration for continuous refinement. Our SPSAF algorithm introduces a smoothing strategy for the amplitude-based phase retrieval loss function, avoiding the nondifferentiable objective function. It significantly simplifies the convergence analysis of SPSAF, eliminating the need to perform other operations on the gradient like other algorithms.

The method we adopt is to iteratively refine through a series of gradient iterations based on the k-sparse hard threshold, that is,

$$\mathbf{z}_{t+1} \coloneqq \mathscr{H}_k\left(\mathbf{z}_t - \mu\nabla\ell_{\varepsilon,p}(\mathbf{z}_t)\right), \tag{31}$$

where t is the number of iterations and constant μ is the step

size. $\mathscr{H}_k(\mathbf{u})$ denotes the k-sparse hard threshold operator, which can transform all elements except k largest elements in $\mathbf{u}$ into zero, thus transforming n-dimensional vector $\mathbf{u}$ into k-sparse n-dimensional vector $\bar{\mathbf{u}}$, i.e.,

$$\bar{\mathbf{u}}_j = \begin{cases} \mathbf{u}_j, & \text{if } j \in S, \\ 0, & \text{if } j \notin S. \end{cases} \tag{32}$$

Combining (13), the Wirtinger gradient $\nabla \ell_{\varepsilon,p}(\mathbf{z})$ of the SPSAF algorithm can be expressed as

$$\nabla \ell_{\varepsilon,p}(\mathbf{z}) = \frac{1}{m}\sum_{i=1}^{m} \begin{matrix} \left[\left(\left|\mathbf{a}_i^T\mathbf{z}\right|^p + \varepsilon_i^p\right)^{1/p} - \left(\psi_i^p + \varepsilon_i^p\right)^{1/p}\right] \\ \left(\left|\mathbf{a}_i^T\mathbf{z}\right|^p + \varepsilon_i^p\right)^{1/p-1}\left|\mathbf{a}_i^T\mathbf{z}\right|^{p-1}\mathbf{a}_i \operatorname{sgn}\left(\mathbf{a}_i^T\mathbf{z}\right) \end{matrix}. \tag{33}$$

It is worth emphasizing that for the smoothed sparse phase retrieval model, the selection of different p will affect the recovery performance and calculated efficiency of the algorithm. Specifically, the larger the p value, the better the restore performance of the algorithm, but the convergence rate will slow down; the smaller the p value, the faster the algorithm converges, but the recovery performance decreases. We set the p value selection according to the ratio of measurements to sparsity m/k. When the ratio m/k is less than 1.9 in the real case or 2.8 in the complex case, $p = 2$ or 3 ; otherwise, $p = 4$ or 5. The specific SPSAF algorithm is described as follows.

4. Experimental Results

This section will introduce the relevant numerical results of the SPSAF algorithm. To reflect the superiority of the SPSAF algorithm, we compare it with the SPARTA algorithm [16] and CRAF algorithm [20], which are the latest methods of sparse phase retrieval. The parameters of all algorithms will use their recommended values. All simulation experiments were conducted with 100 independent Monte Carlo experiments. In each experiment, all algorithms' initialization power iteration numbers and gradient refinement iteration numbers are set to 100.

In all experiments, the signal $\mathbf{x} \in \mathbb{R}^{3000}$ or $\mathbb{C}^{3000}$ to be recovered is k sparse signal in the real or complex case. When the sparse phase retrieval adopts the real-valued Gaussian model, the sparse signal is $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_{3000})$ and the sensing vector is $\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I}_{3000}), i = 1, \cdots, m$. When sparse phase retrieval adopts the complex Gaussian model, the sparse signal is $\mathbf{x} \sim \mathscr{C}\mathcal{N}(0, I_{3000}) = \mathcal{N}(0, 1/2I_{3000}) + j\mathcal{N}(0, 1/2I_{3000})$ and the sensing vector is $\mathbf{a}_i = \mathcal{N}(0, 1/2\mathbf{I}_{3000}) + j\mathcal{N}(0, 1/2\mathbf{I}_{3000}), i = 1, \cdots, m$. In addition, other parameters of the SPSAF algorithm are selected according to experience: learning rate $\mu = 0.8$ in the real case or $\mu = 1$ in the complex case, smoothing parameter $p = 2$.

We use relative error as our performance index, that is,

$$\text{Relative error: } = \frac{\operatorname{dist}(\mathbf{z}, \mathbf{x})}{\|\mathbf{x}\|_2}, \tag{34}$$

where $\operatorname{dist}(\mathbf{z}, \mathbf{x})$ is the Euclidean distance from the estimated value $\mathbf{z}$ to the real solution $\mathbf{x}$. When the relative error is less than 10^{-5}, it can be considered that the original sparse signal has been successfully recovered. Namely, the current experiment is successful.

4.1. Recovery Success Rate with Unknown Sparsity k. In some practical applications, the sparsity k of the signal may be unknown, so it is necessary to test the running state of the algorithm in the case of unknown sparsity, namely, the algorithm's stability. We define $\tilde{k}$ as the estimation of sparsity k and set it to the upper bound of sparsity level in theory. Theorem 1 in [16] proves that when m is approximately equal to n, the upper limit of the sparsity level in this paper is about $\lceil \sqrt{n} \rceil = 55$.

When the phase retrieval model is in the real case, the recovery success rates of SPSAF, SPARTA, and CRAF algorithms are compared as shown in Figure 1, where the number of measurements m increases from 400 to 2400. Notably, these curves show that SPSAF has higher precision recovery performance than other comparison algorithms. When the sparsity k is unknown, the SPSAF algorithm shows higher stability, and the recovery rate reaches more than 90% when $m = 1000$ and achieves accurate recovery when $m = 1400$.

When the phase retrieval model is in the complex case, the recovery success rates of SPSAF, SPARTA, and CRAF algorithms are shown in Figure 2, where the number of measurements m ranges from 800 to 2800. It can be found from Figure 2 that the SPSAF algorithm has a weak advantage in recovery ability compared with other algorithms. When the actual sparsity $k = 30$, the estimated sparsity $\tilde{k} = 55$; the SPSAF algorithm is stable. It is worth emphasizing that when the number of measurements $m = 500$, the SPSAF algorithm first reaches more than 80%, showing excellent recovery ability.

4.2. Recovery Success Rate with Known Sparsity k. In this section, we compare the signal recovery success rates of each algorithm when the sparsity k is known. The comparison of the algorithms in the real case is shown in Figure 3, where the number of measurements m increases from 400 to 2400. It can be seen from Figure 3 that the recovery performance of the SPSAF algorithm is significantly higher than the other two algorithms, showing excellent recovery ability. In addition, it is worth noting that compared with Figures 1 and 3, it can be found that the recovery performance of the SPSAF algorithm significantly improved when the sparse prior is known.

Figure 4 depicts the success rate of signal recovery in the complex case. We set the distribution of the number of measurements m from 800 to 2800, where the prior sparsity $k = 30$ is known. It can be seen from Figure 4 that the SPSAF algorithm is significantly better than the other two algorithms. Moreover, by comparing Figures 2 and 4, we find

that the recovery degree of the algorithm has been slightly improved when the sparsity k is known.

Figure 5 depicts the relationship between the success rate of signal recovery and the sparsity k in the real case, where the step size parameters of each algorithm take the optimal value. It can be seen that with the increasing sparsity k, the SPSAF algorithm can still accurately restore the signal at the sparsity $k = 50$ and can still maintain a success rate of more than 95% at $k = 100$. In comparison, the success rates of the CRAF algorithm and SPARTA algorithm with the same sparsity are reduced to about 80%.

4.3. Comparison of Convergence Consumption. This section compares the SPSAF algorithm with other algorithms in terms of convergence speed and time cost. We adopt the real Gaussian model when the number of measurements is $m = 1600$ or the complex Gaussian model when $m = 2800$, respectively. The prior sparsity of all models is $k = 30$, and the relative error of the algorithm is less than 10^{-15} which is regarded as an accurate recovery. The convergence curves in the real or complex case are shown in Figures 6 and 7. It can be seen from these figures that the number of iterations required for the convergence of the SPSAF algorithm is significantly less than that of the other two algorithms.

On the other hand, Table 1 compares convergence speed and time cost under the noise-free Gaussian model, where the coarse font is the current optimal value. It can be seen from the results that the convergence speed and time cost of the SPSAF algorithm are better than those of other algorithms in both real and complex cases, which show excellent performance.

4.4. Noise Robustness. In order to prove the stability of the SPSAF algorithm in the presence of additional noise, we plot the relative mean square error (MSE) as a function of the signal-to-noise ratio (SNR) value in dB. Amplitude measurements with noise can be expressed as

$$\psi_i = \left|\mathbf{a}_i^T \mathbf{x}\right| + \eta_i, \tag{35}$$

where the size of $\eta_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ is given by

$$\text{SNR}: = 10 \log_{10} \sum_{i=1}^{m} \frac{\left|\mathbf{a}_i^T \mathbf{x}\right|^2}{m\sigma^2}. \tag{36}$$

In this section, we adopt $m = 1600$ real-valued Gaussian model with sparse prior $k = 30$. Figure 8 depicts the relationship between the average relative error of the three algorithms when $\sigma^2 = 3$. It can be seen that the SPSAF algorithm provides the most accurate estimation under the noise addition.

We also describe the relative error as a function of SNR with different m/n to verify the stability of the SPSAF algorithm. Figure 9 depicts the experimental results. We use a real-valued Gaussian model with sparse prior $k = 30$, where SNR is between 5 dB and 50 dB. It can be seen that the SPSAF algorithm decreases approximately linearly with the increase of SNR in both real and complex cases, which proves that the proposed SPSAF algorithm is stable in noisy environments.

5. Conclusion

In this paper, we proposed the SPSAF algorithm to solve the problems of sparse phase retrieval. The proposed SPSAF algorithm is an amplitude-based nonconvex sparse smoothing phase retrieval algorithm divided into two stages: initialization and gradient refinement. The complexity of our algorithm is reduced by smoothing the phase retrieval loss function to avoid the truncation or weighting of the gradient. The SPSAF algorithm first estimates the support of the original signal by a reasonable rule, obtains the initial estimation by a carefully designed initialization method based on the support, and finally obtains the sparse estimation by a gradient descent method based on the hard threshold. Numerical experiments show that the SPSAF algorithm has significantly improved recovery performance and speed compared with the existing typical algorithms and has good robustness.

Data Availability

All research data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the Natural Science Research Project of Fuyang Normal University (2021FSKJ02ZD), Natural Science Research Project of Anhui Province (KJ2020A0539, KJ2021A0662, and KJ2021A0682), Science Research and Innovation Team of Fuyang Normal University (kytd202004 and TDJC2021008), and Fuyang Normal University Young Talents Key Project (rcxm202006 and rcxm202004).

References

[1] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "Phase retrieval: an overview of recent developments," *Optical Compressive Imaging*, pp. 279–312, 2016.

[2] X. Wang, D. Meng, M. Huang, and L. Wan, "Reweighted regularized sparse recovery for DOA estimation with unknown mutual coupling," *IEEE Communications Letters*, vol. 23, no. 2, pp. 290–293, 2019.

[3] X. Wang, M. Huang, and L. Wan, "Joint 2D-DOD and 2D-DOA estimation for coprime EMVS–MIMO radar," *Circuits, Systems, and Signal Processing*, vol. 40, no. 6, pp. 2950–2966, 2021.

[4] X. Wang, L. Wan, M. Huang, C. Shen, Z. Han, and T. Zhu, "Low-complexity channel estimation for circular and noncircular signals in virtual MIMO vehicle communication

systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 3916–3928, 2020.

[5] J. Zeng, J. Sun, G. Gui et al., "Downlink CSI feedback algorithm with deep transfer learning for FDD massive MIMO systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 4, pp. 1253–1265, 2021.

[6] H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization," *Journal of the Optical Society of America A Optics Image Science & Vision*, vol. 19, no. 7, pp. 1334–1345, 2002.

[7] Y. M. Lu and M. Vetterli, "Sparse spectral factorization: unicity and reconstruction algorithms," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5976–5979, Prague, Czech Republic, May 2011.

[8] H. Ohlsson and Y. C. Eldar, "On conditions for uniqueness in sparse phase retrieval," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1841–1845, Florence, Italy, May 2014.

[9] K. Jaganathan, S. Oymak, and B. Hassibi, "Recovery of sparse 1-D signals from the magnitudes of their Fourier transform," in *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 1473–1477, Cambridge, MA, USA, July 2012.

[10] X. Song, J. Wang, J. Wang et al., "SALDR: joint self-attention learning and dense refine for massive MIMO CSI feedback with multiple compression ratio," *IEEE Wireless Communications Letters*, vol. 10, no. 9, pp. 1899–1903, 2021.

[11] J. Wang, G. Gui, T. Ohtsuki, B. Adebisi, H. Gacanin, and H. Sari, "Compressive sampled CSI feedback method based on deep learning for FDD massive MIMO systems," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5873–5885, 2021.

[12] M. L. Moravec, J. K. Romberg, and R. G. Baraniuk, "Compressive phase retrieval," *Wavelets XII*, vol. 6701, pp. 712–722, 2007.

[13] Y. Shechtman, Y. C. Eldar, A. Szameit, and M. Segev, "Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing," *Optics Express*, vol. 19, no. 16, pp. 14807–14822, 2011.

[14] H. Ohlsson, A. Yang, R. Dong, and S. Sastry, "CPRL–an extension of compressive sensing to the phase retrieval problem," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1367–1375, 2012.

[15] S. Bahmani and J. Romberg, "Phase retrieval meets statistical learning theory: a flexible convex relaxation," *Artificial Intelligence and Statistics*, vol. 11, no. 2, pp. 252–260, 2017.

[16] P. Hand and V. Voroninski, "Compressed sensing from phaseless gaussian measurements via linear programming in the natural parameter space," 2016, http://arxiv.org/abs/1611.05985.

[17] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 18, pp. 4814–4826, 2015.

[18] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 1043–1055, 2015.

[19] Y. Shechtman, A. Beck, and Y. C. Eldar, "GESPAR: efficient phase retrieval of sparse signals," *IEEE Transactions on Signal Processing*, vol. 62, no. 4, pp. 928–938, 2014.

[20] T. T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow," *The Annals of Statistics*, vol. 44, no. 5, pp. 2221–2251, 2016.

[21] T. Qiu and D. P. Palomar, "Undersampled sparse phase retrieval via majorization-minimization," *IEEE Transactions on Signal Processing*, vol. 65, no. 22, pp. 5957–5969, 2017.

[22] G. Wang, L. Zhang, G. B. Giannakis, M. Akçakaya, and J. Chen, "Sparse phase retrieval via truncated amplitude flow," *IEEE Transactions on Signal Processing*, vol. 66, no. 2, pp. 479–491, 2018.

[23] L. Zhang, G. Wang, G. B. Giannakis, and J. Chen, "Compressive phase retrieval via reweighted amplitude flow," *IEEE Transactions on Signal Processing*, vol. 66, no. 19, pp. 5029–5040, 2018.

[24] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM Journal on Mathematical Analysis*, vol. 45, no. 5, pp. 3019–3033, 2013.

[25] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 773–794, 2018.

[26] L.-H. Yeh, J. Dong, J. Zhong et al., "Experimental robustness of Fourier ptychography phase retrieval algorithms," *Optics Express*, vol. 23, no. 26, pp. 33214–33240, 2015.

[27] Y. Wang and Z. Xu, "Phase retrieval for sparse signals," *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 531–544, 2014.

[28] M. Akçakaya and V. Tarokh, "Sparse signal recovery from a mixture of linear and magnitude-only measurements," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1220–1223, 2015.

[29] M. Iwen, A. Viswanathan, and Y. Wang, "Robust sparse phase retrieval made easy," vol. 42, no. 1, pp. 135–142, 2017.

[30] Y. C. Eldar and S. Mendelson, "Phase retrieval: stability and recovery guarantees," *Applied and Computational Harmonic Analysis*, vol. 36, no. 3, pp. 473–494, 2014.

[31] P. M. Pardalos and S. A. Vavasis, "Quadratic programming with one negative eigenvalue is NP-hard," *Journal of Global Optimization*, vol. 1, no. 1, pp. 15–22, 1991.

[32] Y. Li, G. Gui, and X. Cheng, "From group sparse coding to rank minimization: a novel denoising model for low-level image restoration," *Signal Processing*, vol. 176, p. 107655, 2020.

[33] Y. Li, L. Liu, Y. Zhao, X. Cheng, and G. Gui, "Nonconvex nonsmooth low-rank minimization for generalized image compressed sensing via group sparse representation," *Journal of the Franklin Institute*, vol. 357, no. 10, pp. 6370–6405, 2020.

[34] D. Needell and J. A. Tropp, "CoSaMP: iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[35] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[36] J. C. Duchi and F. Ruan, "Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval," *Information and Inference*, vol. 8, no. 3, pp. 471–529, 2019.

[37] S. Pinilla, J. Bacca, and H. Arguello, "Phase retrieval algorithm via nonconvex minimization using a smoothing function," *IEEE Transactions on Signal Processing*, vol. 66, no. 17, pp. 4574–4584, 2018.

[38] B. Gao, X. Sun, Y. Wang, and Z. Xu, "Perturbed amplitude flow for phase retrieval," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5427–5440, 2020.

[39] Q. Luo, H. Wang, and S. Lin, "Phase retrieval via smoothed amplitude flow," *Signal Processing*, vol. 177, p. 107719, 2020.

[40] Z. Xiao, Y. Wang, and G. Gui, "Smoothed amplitude flow-based phase retrieval algorithm," *Journal of the Franklin Institute*, vol. 358, no. 14, pp. 7270–7285, 2021.

[41] G. Wang, G. B. Giannakis, Y. Saad, and J. Chen, "Phase retrieval via reweighted amplitude flow," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2818–2833, 2018.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3652936, 7 pages
https://doi.org/10.1155/2022/3652936

Research Article

Fast and Accurate Approach for DOA Estimation of Coherent Signals

Yuhong Liu,[1] **Hui Cao**,[1,2] **Yuntao Wu**,[1] **and Kehao Wang**[2]

[1]*Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology), Wuhan 430205, China*
[2]*School of Information Engineering, Wuhan University of Technology, Luoshi Road, Wuhan 430070, China*

Correspondence should be addressed to Hui Cao; tsaohui@163.com

Received 21 February 2022; Revised 7 April 2022; Accepted 15 April 2022; Published 13 May 2022

Academic Editor: Abdul Basit

An algorithm for estimation of direction of arrival (DOA) based on QR decomposition is proposed for coherent signals in this paper. When coherent is present, the rank of the signal covariance matrix is generally less than the signal number, making the estimation of the signal or noise subspace inaccurate. Therefore, we need to eliminate the spatial covariance matrix rank loss. According to the idea of matrix reconstruction, we try to construct three different data matrices, one is the signal covariance matrix, the other is the eigenvector reconstruction matrix of the signal covariance matrix, and the other is reconstructed matrix with the addition of spatial smoothing technology. Based on the resulting data matrix, whereafter, QR decomposition with the iteratively weighted least squares (IWLS) as solver is proposed to reduce the computational complexity of DOA estimation. Compared with other existing algorithms, the simulation results show that our method has high accuracy and great increase in the computational efficiency.

1. Introduction

In recent years, people have invested a lot of energy in researching high-resolution technology to estimate the angle of arrival of sources on linear arrays in wireless communications, including astronomy, radar, smart grid, and sonar [1–7]. Algorithms like MUSIC [8] and ESPRIT [9] are based on subspace decomposition to solve uncorrelated signals with good performance. And another classic subspace algorithm propagation operator (propagator method, PM) [10] used a series of linear operations to obtain a noise subspace orthogonal to the steering vector. The disadvantage is that the angle estimation accuracy is relatively poor under low signal-to-noise ratio (SNR). The prerequisite of these algorithms is incoherent signals. However, as the space environment becomes more and more complex, the signals received by the array are often mixed with coherent signals. These methods will encounter difficulties, such as multi-path propagation signals and co-frequency interference signals.

In response to this problem, scholars have proposed a series of algorithms, such as forward spatial smooth (FOSS) algorithm [11] splits the array into sub-arrays, and then superimposes each other to eliminate the coherence of the signal. But FOSS algorithm from the structure of the covariance matrix from the above point of view, forward smoothing does not make full use of all the information of the sample covariance matrix. In order to eliminate this shortcoming, while increasing the information utilization, Pillai and Kwon [12] proposed forward-backward spatial smoothing (FBSS), ignoring the cross-correlation information at both ends of the anti-diagonal line. Next, an improved spatial smoothing technique is proposed in [13] to use the statistics of the entire sample covariance matrix as much as possible. Matrix reconstruction technology [14, 15] is another type of method used to solve the problem of coherent signal sources, [15] proposed an eigenvector method (EVM), by selecting the largest feature value corresponding eigenvectors to construct a full-rank forward and backward

matrix. [16, 17] proposed a singular value decomposition (SVD) algorithm, while [18, 19] achieved decoherence by constructing the data receiving matrix into a Toeplitz matrix.

Most of the above methods or methods developed on this basis to combat the DOA estimation problem of coherent sources are at the expense of array aperture, that is, the degree of freedom of the array will be reduced. This makes the performance of the subspace algorithm unable to reach the Cramer-Rao lower bound even under the condition of high SNR. In order to remedy the above problems, a method based on principal-singular-vector utilization for modal analysis (PUMA) [20–22] was proposed, but when the incident signal is completely coherent, the performance deteriorates severely. The algorithm goal for DOA estimation is able to provide reliable azimuth information in a limited sample and low signal-to-noise ratio environment. Many existing subspace algorithms have severe thresholding effects under such conditions.

A QR-IWLS method was proposed in [23] by applying QR decomposition to the parameter estimation of two-dimensional complex-valued sinusoidal signals under additive white Gaussian noise with the use of the rank and linear prediction properties of the noise-free data matrix; then, an iteratively weighted least squares (IWLS) algorithm is used to estimate the linear prediction coefficients. On the basis, [24] proposed a fast rank revealing technique for frequency estimation of complex sinusoids in a noisy environment, and a two-stage QR decomposition frequency estimation method using weighted least squares (WLS) as the solver.

Inspired by the literatures [23, 24], in this paper, we propose a fast and low-complexity DOA estimation method based on subspace decomposition. First, the method uses the steering vector to linearly express and constructs an equivalent covariance matrix under Gaussian noise based on the covariance matrix of the received information of the data. Second, the proposed method uses QR decomposition to replace the computationally intensive EVD/SVD, and uses the first row of the R matrix to reconstruct the equivalent covariance matrix Y again. In addition, we also try to use forward and backward method to construct the data matrix.

The rest of the paper is organized as follows. In the Section of System Model, the signal model for DOA estimation of coherent signal is given. The algorithm is proposed in the Section of Proposed Approach. In the Section of Simulation Results, we perform the simulation to evaluate the accuracy and efficiency of the proposed approach. Finally, conclusions are drawn in the Section of Conclusion.

2. System Model

In this work, we consider P far-field narrow-band sources from directions $\theta_i (i = 1, 2, \cdots, P)$ impinging on a uniform linear array (ULA) having $M (M > P)$ isotropic sensors with halfwavelength interspacing as shown in Figure 1 [18].

The first L signals are fully coherent, while the other $P - L$ signals are uncorrelated and independent of the first ones. We take $s_1(t)$ as a reference, and the lth coherent signal is:

$$s_l(t) = \beta_l s_1(t) = \rho_l e^{j\Delta\phi_l} s_1(t), l = 1, \cdots, L, \tag{1}$$

where $\beta_l = \rho_l e^{j\Delta\phi_l}$, ρ_l is the amplitude fading factor, and $\Delta\phi_l$ denotes the phase difference of $s_l(t)$ related to $s_1(t)$. Let the sensor 1 be the reference, and the mth array output at time t is:

$$\begin{aligned} x_m(t) &= \sum_{i=1}^{P} s_i(t) e^{(-j2\pi md \sin\theta_i)/\lambda} + n_m(t) \\ &= s_1(t) \sum_{i=1}^{L} \beta_i e^{(-j2\pi md \sin\theta_i)/\lambda} \\ &\quad + \sum_{i=L+1}^{P} s_i(t) e^{(-j2\pi md \sin\theta_i)/\lambda} + n_m(t), \end{aligned} \tag{2}$$

where $s_i(t)$ and $n_m(t)$ are the complex envelop of the ith signal and the additive white noise at the mth element $(m = 1, \cdots, M)$. $d = \lambda/2$ is the interspacing, and λ is the carrier wavelength. The final array received signal vector is:

$$x(t) = [x_1(t), \cdots, x_M(t)]^T = \mathbf{A}(\theta) s(t) + \mathbf{n}(t), \tag{3}$$

where $(t) = [s_1(t), \cdots, s_P(t)]^T$ denotes the source vector. $\boldsymbol{n}(t) = [n_1(t), \cdots, n_M(t)]^T$ is the complex Gaussian noise vector. $\boldsymbol{A}(\theta) = [\mathrm{a}(\theta_1), \cdots, \mathrm{a}(\theta_P)]$ is the array steering matrix with $a(\theta_i) = [e^{-j2\pi d \sin\theta_i/\lambda}, e^{-j2\pi 2d \sin\theta_i/\lambda}, \cdots, e^{-j2\pi Md \sin\theta_i/\lambda}]^T$ $(\cdot)^T$ is the transpose.

3. Proposed Approach

In the above received signal vector $x(t)$, when the number of snapshots is N, one can evaluate the output covariance matrix:

$$\mathbf{R}_{xx} = \frac{1}{N} \sum_{t=1}^{N} x(t) x^H(t), \tag{4}$$

where $(\cdot)^H$ is the conjugate transpose. We can perform the EVD/SVD on the covariance matrix $\mathbf{R}_{xx}$ to obtain the signal subspace and use MUSIC or other subspace-based method to estimate the DOA. However, the sources are completely coherent, namely, $L = P$. The rank of sources covariance matrix is one. After the EVD of $\mathbf{R}_{xx}$, the dimension of the signal subspace is less than the rank of the array manifold $\mathbf{A}(\theta)$, which leads to the consequence that the steering vector is no longer orthogonal to the noise subspace and makes failure of subspace algorithm. According to [25], we assume the noise is temporal white and the noise covariance matrix denoted by $\mathbf{R}_n$, is full-rank. As the steering vectors span the signal subspace, so we have linear representation form as:

$$\mathbf{R}_n \mathbf{e}_k = \sum_{n=1}^{P} \alpha_k(n) a(\theta_n), \tag{5}$$

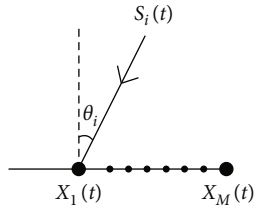

Figure 1: Received signal model of uniform linear array.

where $\mathbf{e}_k(1 \le k \le K)$ is an eigenvector of the received signal covariance matrix (the first K eigenvectors corresponding to first eigenvalues in decreasing order), and $\alpha_k(n)$ is a linear combination factor. When the noise covariance is an identity matrix, (5) is simplified as:

$$\mathbf{e}_k = \sum_{n=1}^{P} \alpha_k(n)\mathbf{a}(\theta_n). \tag{6}$$

For completely coherent case, namely, $K = 1$, the above equation is reduced to:

$$\mathbf{e}_1 = \sum_{n=1}^{P} \alpha_1(n)\mathbf{a}(\theta_n). \tag{7}$$

It indicates that the largest eigenvalue contains all the signal information. So the vector $\mathbf{e}_1$ can be used to reconstruct an equivalence covariance matrix $\boldsymbol{Y}$, constructed as:

$$\boldsymbol{Y} = \begin{bmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,g} \\ e_{1,2} & e_{1,3} & \cdots & e_{1,g+1} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1,m} & e_{1,m+1} & \cdots & e_{1,M} \end{bmatrix}, \tag{8}$$

where $e_{1,m}$ is the element of $\mathbf{e}_1$, $m = M - g + 1$, $m > P$, $g > P$.

Here, QR decomposition is used instead of EVD/SVD to reduce the complexity of calculation. Then, the first row of the matrix $\boldsymbol{R}$ is used to construct a equivalence covariance matrix $\boldsymbol{Y}$, and $e_{1,m}$ is the elements of $\mathbf{r}_1$, $\boldsymbol{R} = [\mathbf{r}_1, \cdots, \mathbf{r}_M]^T$.

Finally, we apply the QR decomposition again to the constructed matrix Y and estimate the DOA with the use of the QR-IWLS method [23].

On the other hand, as we know, the performance of the forward and backward method [12] is much better than that only using the constructed $\boldsymbol{Y}$, which can also be used to solve the coherent situation. So in this paper, we also try this similar way; the backward matrix is:

$$\mathbf{Y}_b = \mathbf{J}_m \boldsymbol{Y}^* \mathbf{J}_q, \tag{9}$$

where both of $\mathbf{J}_m$ and $\mathbf{J}_q$ are anti-diagonal eye matrices, and $(\cdot)^*$ denotes the conjugate operation, then we obtain the equivalence covariance matrix:

$$\mathbf{Y}_r = \frac{1}{2}(\mathbf{Y} + \mathbf{Y}_b). \tag{10}$$

Then, we also perform the QR factorization on $\mathbf{Y}_r$ to obtain the signal subspace. And next the QR-IWLS method is used to estimate the DOA.

For the above analysis, the QR-IWLS algorithm steps are summarized as follows:

Step 1: Calculate the covariance matrix $\mathbf{R}_{xx}$ of the array received signal data by (4).

Step 2: Perform QR decomposition on $\mathbf{R}_{xx}$, take the eigenvector $\mathbf{e}_k$ corresponding to the largest eigenvalue.

Step 3: Reconstruction the matrix $\boldsymbol{Y}$, $\mathbf{Y}_r$ by (8), (10).

Step 4: Apply the QR decomposition to the $\mathbf{R}_{xx}$, $\boldsymbol{Y}$, $\boldsymbol{Y}_r$, respectively, to obtain the signal subspace.

Step 5: Use QR-IWLS to estimate DOAs of signal.

To sum up, we construct three different data matrices: covariance matrix (QR-IWLS-matrix), eigenvector reconstruction matrix using signal covariance (QR-IWLS-rematrix), and reconstruction matrix by applying smoothing technique (QR-IWLS-fbmatrix); then, we use QR-IWLS method to estimate DOA. The simulation experiments are carried out under different SNR, different snapshots, and different array numbers, as shown in Figures 2–4. The results show that as the SNR increases, QR-IWLS-rematrix is better than the other two methods, and QR-IWLS-fbmatrix performs poorly. So in the next section, we choose QR-IWLS-matrix and QR-IWLS-rematrix for experiments.

4. Simulation Results

In the simulation part, computer simulations are conducted to evaluate the performance of the QR-IWLS method in white Gaussian noise, compared with MUSIC [17] and PUMA [22] algorithms. In the experiment, we consider a ULA composed of different numbers of isotropic sensors and the element spacing is equal to half a wavelength, if not mentioned, two narrow-band far-field signal sources with angles of 5° and 75°. Furthermore, all results are averaged after 200 independent Monte Carlo experiments. The mean square error (MSE) is used to evaluate and analyze the performance of the algorithm, which is computed by $\mathrm{MSE} = 1/200\sum_{i=1}^{G}[(\theta_i - \widehat{\theta}_i)^2]$, where θ_i is the actual value, and $\widehat{\theta}_i$ is the measured value, respectively. The simulation is performed on the MATLAB R2019b on a computer with 16 GB of RAM and the 64-bit Windows 11 operating system.

In our first test, the number of snapshots N is set to 512, and the number of sensors M is 18. We investigate the MSE of the proposed method versus the SNR, as shown in Figure 5. The proposed method achieves DOA estimation performance better than the MUSIC method and QR-IWLS-matrix method, and has comparable performance with the PUMA algorithm.

In the second test, we set the SNR to 10 dB and the number of sensors M to 18. We compare the performance of each method in the case of different snapshots. As seen in Figure 6, with the number of snapshots increases, the performance of MUSIC method is severely damaged due to the

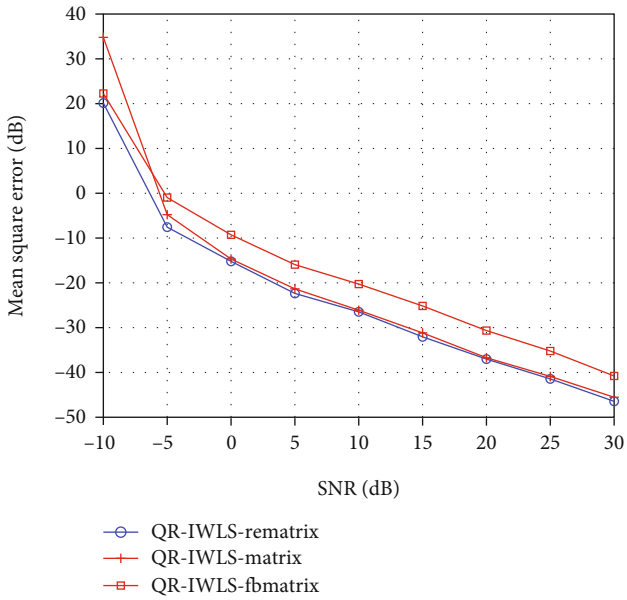

Figure 2: MSE versus SNR for three different data matrices.

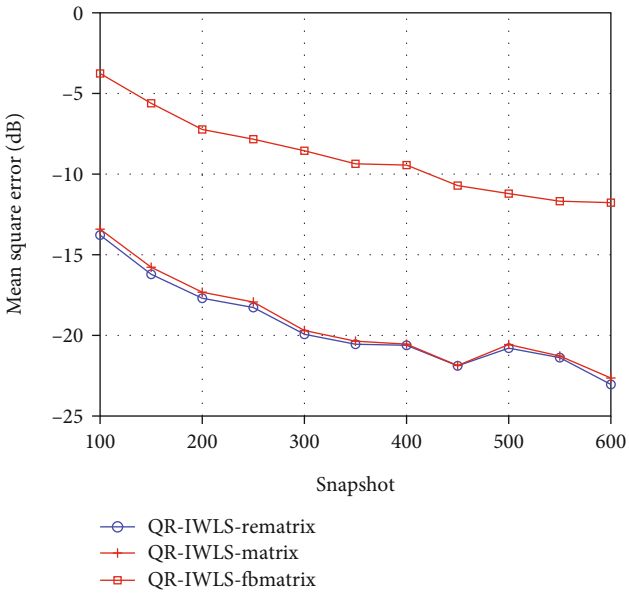

Figure 3: MSE versus snapshots for three different data matrices.

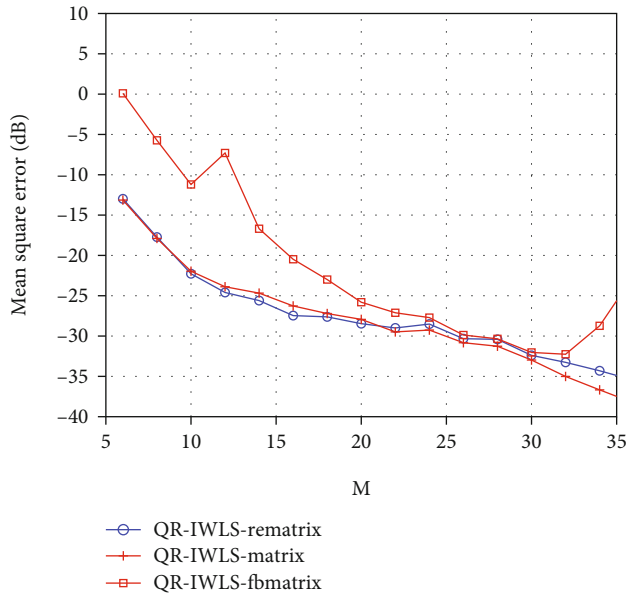

Figure 4: MSE versus array number for three different data matrices.

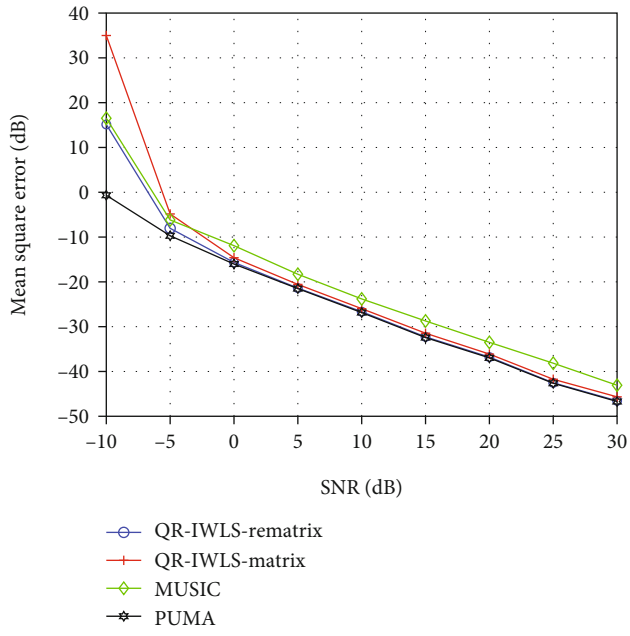

Figure 5: MSE versus SNR.

coherent of signals. The figure also confirms that for the QR-IWLS-matrix method and the QR-IWLS-rematrix method, the error is much smaller than other methods. Among them, the QR-IWLS-rematrix method has the best performance.

In the third experiment, the MSE versus the number of sensors is investigated. The number of sensors M is varied from 5 to 35, the SNR is 10 dB, and the number of snapshots is 512. As shown in Figure 7, the proposed method achieves comparable performance with the PUMA method and it is superior to the MUSIC algorithms.

Next, we investigate the relationship between success probability and SNR; the results are plotted in Figure 8, where the ratio of the number of successful runs to the total number of independent runs is calculated as the probability of success. We assumed that the number of DOAs is 2, when $\max |\theta_i - \widehat{\theta}_i| \leq |\theta_2 - \theta_1|/2$. It is concluded that all methods achieve 100% success at SNR ≥ 5 dB, and our QR-IWLS-rematrix method and the PUMA method achieve the highest resolution probability.

Finally, the computational complexity of the algorithm is compared, where the computer runtime is shown in Figure 9 and the complexity analysis is shown in Table 1. For

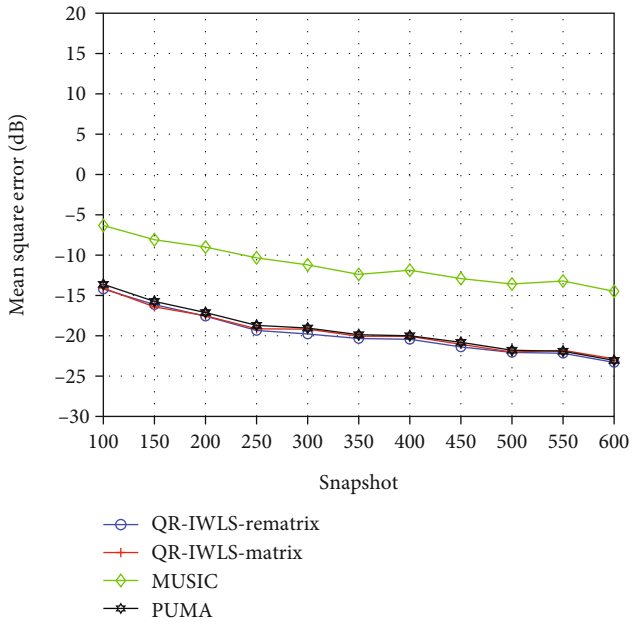

Figure 6: MSE versus snapshots number.

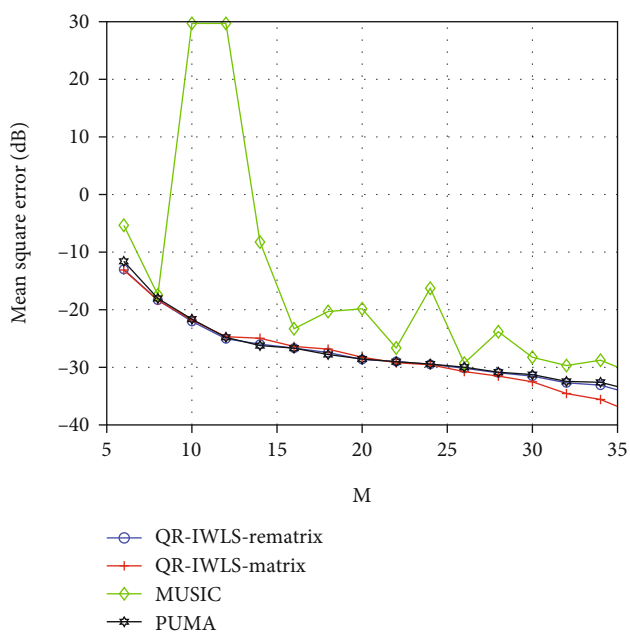

Figure 7: MSE versus array number.

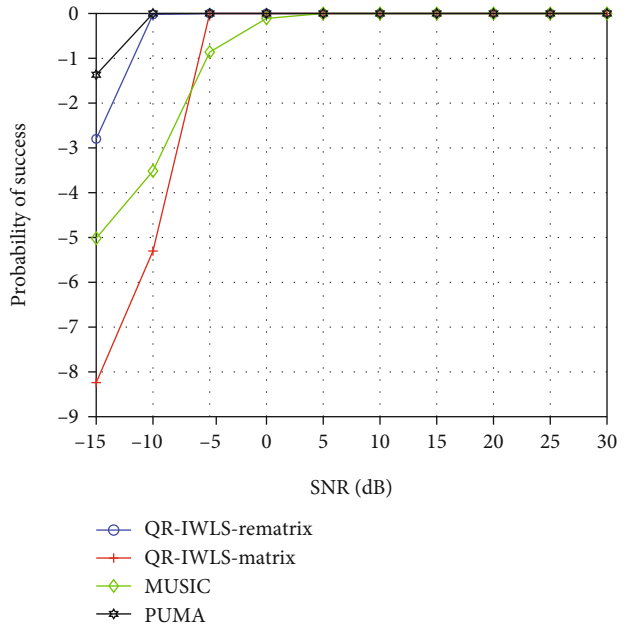

Figure 8: Probability of success.

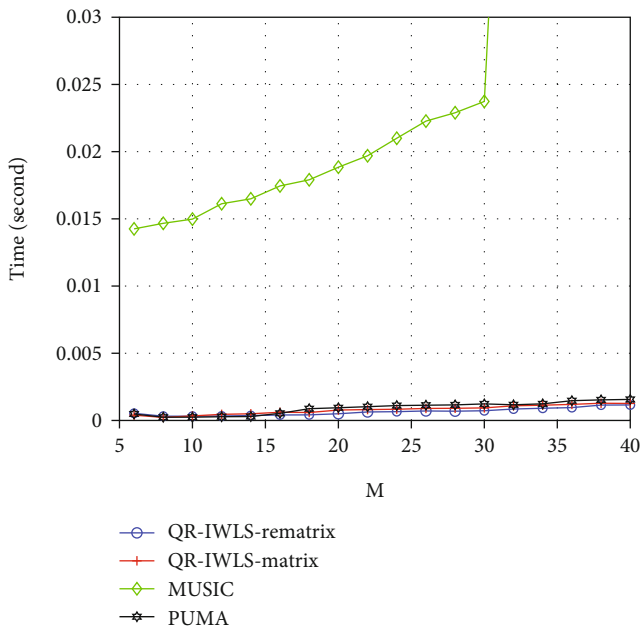

Figure 9: Time comparison versus array number.

simplicity, we only consider the main complexity of each method. Both simulation results and the complexity analysis demonstrate that the QR-IWLS-rematrix method has the lowest computational complexity among all methods, that is, under the conditions of corresponding SNR and sufficiently large number of sensors, the proposed method has a smaller MSE with less complexity compared with other mentioned methods.

In the above experiments we did, the setting parameters are large. To further investigate the performances of our proposed approaches, we simulate another experiment, that is, turn down the parameter setting. Fix $\theta_1 = 5^\circ$, and ranging θ_2 from 10° to 80° with $M = 10$ elements. As shown in Figure 10, our proposed algorithms have better results than the other two algorithms, and the performance of MUSIC algorithm is unstable and even fails when θ_2 is larger than 60°. Thus, our algorithm is persuasive and effective.

Table 1: The computational complexity comparison.

Algorithm	Computational complexity
QR-IWLS-rematrix	$O(-M^2N - MN + M^3)$
QR-IWLS-matrix	$O(-M^2N + M^3)$
PUMA	$O(M^2N + M^3)$
MUSIC	$O(M^2N + 2MN + M^3)$

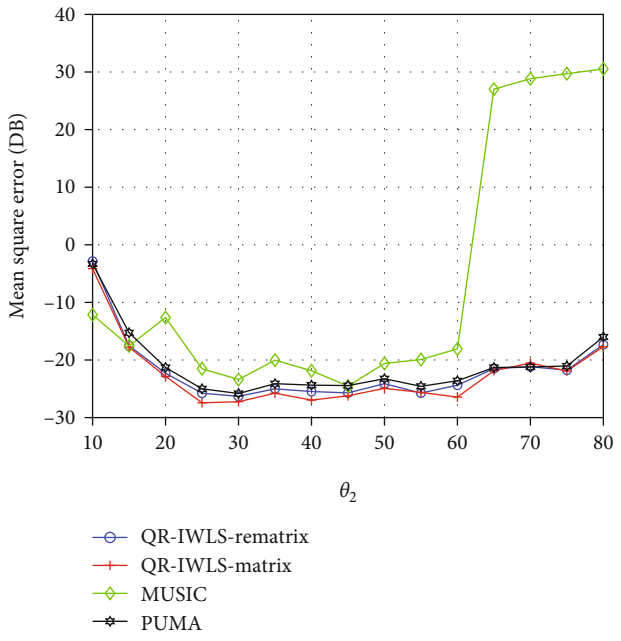

Figure 10: Fix θ_1 and change θ_2 with 10 elements.

5. Conclusion

In this paper, we proposed a QR decomposition-based method to handle DOA estimation of coherent signals. The coherent of the received signal is deduced by the reconstructing of a data matrix, whose elements come from the eigenvalues of a covariance matrix. The DOA is then estimated with the use of QR decomposition and IWLS-based method. The simulation results verify that the proposed approach has a comparable performance and less complexity compared with other advanced methods.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was jointly supported by the Hubei Key Laboratory of Intelligent Robot (HBIR202005), Natural Science Foundation of Hubei Province (2020CFB439), the Three Gorges Laboratory Open Fund of Hubei Province (SC215001), and the Natural Science Foundation of China (62172313).

References

[1] W. Sun, L. Huang, H. C. So, and J. Wang, "Orthogonal tubal rank-1 tensor pursuit for tensor completion," *Signal Processing*, vol. 157, pp. 213–224, 2019.

[2] W. Sun, Y. Huang, L. Huang, Q. Li, and J. Zhang, "ℓ2, p-correlation and robust matching pursuit for sparse approximation," *Digital Signal Processing*, vol. 104, article 102761, 2020.

[3] Q. Liu, G. Yuantao, and H. C. So, "DOA estimation in impulsive noise via low-rank matrix approximation and weakly convex optimization," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 6, pp. 3603–3616, 2019.

[4] Q. Liu, H. C. So, and G. Yuantao, "Off-grid DOA estimation with nonconvex regularization via joint sparse representation," *Signal Processing*, vol. 140, pp. 171–176, 2017.

[5] Q. Liu, X. Li, and H. Cao, "Two-dimensional localization: low-rank matrix completion with random sampling in massive MIMO system," *IEEE Systems Journal*, vol. 15, no. 3, pp. 3628–3631, 2021.

[6] Z. Wei, W. Wang, F. Dong, and Q. Liu, "Gridless one-bit direction-of-arrival estimation via atomic norm denoising," *IEEE Communications Letters*, vol. 24, no. 10, pp. 2177–2181, 2020.

[7] F. Wen, J. Shi, and Z. Zhang, "Generalized spatial smoothing in bistatic EMVS-MIMO radar," *Signal Processing*, vol. 193, article 108406, 2022.

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[9] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[10] S. Marcos, A. Marsal, and M. Benidir, "Performances analysis of the propagator method for source bearing estimation," in *Proceedings of ICASSP 1994. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. IV/237–IV/240, Adelaide, SA, Australia., 1994.

[11] T.-J. Shan, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation of coherent signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 806–811, 1985.

[12] S. U. Pillai and B. H. Kwon, "Forward/backward spatial smoothing techniques for coherent signal identification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 1, pp. 8–15, 1989.

[13] W. Du and R. L. Kirlin, "Improved spatial smoothing techniques for DOA estimation of coherent signals," *IEEE Transactions on Signal Processing*, vol. 39, no. 5, pp. 1208–1210, 1991.

[14] F.-J. Chen, S. Kwong, and C.-W. Kok, "ESPRIT-like two-dimensional DOA estimation for coherent signals," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 46, no. 3, pp. 1477–1484, 2010.

[15] Y.-H. Choi, "ESPRIT-based coherent source localization with forward and backward vectors," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 6416–6420, 2010.

[16] S.-W. Gao and Z. Bao, "Data-based matrix decomposition technique for high-resolution array processing of coherent signals," *Electronics Letters*, vol. 23, no. 12, pp. 643–645, 1987.

[17] J. Zhang, D. Huang, P. Huang, and J. Kang, "Estimation DOAs of the coherent sources based on SVD," in *2010 2nd International Conference on Signal Processing Systems*, vol. 3, pp. V3–174–V3–177, Dalian, China., 2010.

[18] F.-M. Han and X.-D. Zhang, "An ESPRIT-like algorithm for coherent DOA estimation," *IEEE Antennas and Wireless Propagation Letters*, vol. 4, pp. 443–446, 2005.

[19] L. Liu, Y. Gai, H. Wang, and G. Ding, "An improved ESPRIT-like algorithm for coherent signal and its application for 2-D DOA estimation," in *2006 7th International Symposium on Antennas, Propagation EM Theory*, pp. 1–4, Guilin, China., 2006.

[20] H. C. So, F. K. W. Chan, and W. Sun, "Subspace approach for fast and accurate single-tone frequency estimation," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 827–831, 2011.

[21] F. K. W. Chan, H. C. So, and W. Sun, "Subspace approach for two-dimensional parameter estimation of multiple damped sinusoids," *Signal Processing*, vol. 92, no. 9, pp. 2172–2179, 2012.

[22] Q. Cheng, L. Huang, M. Cao, J. Xie, and H. C. So, "PUMA: an improved realization of MODE for DOA estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 5, pp. 2128–2139, 2017.

[23] Q. Xiang, H. Cao, and H. C. So, "Fast and accurate estimator for two-dimensional cisoids based on QR decomposition," *Signal Processing*, vol. 120, pp. 553–561, 2016.

[24] H. Cao and Q. Liu, "Fast rank-revealing QR factorization for two-dimensional frequency estimation," *IEEE Communications Letters*, vol. 24, no. 6, pp. 1240–1243, 2020.

[25] Z.-J. Zhao, Y. Wang, and X. Chun-Yun, "DOA estimation of coherent signals based on improved SVD algorithm," in *2012 Second International Conference on Instrumentation, Measurement, Computer, Communication and Control*, pp. 524–528, Harbin, China., 2012.