IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. 31 (2020) 055003 (17pp) https://doi.org/10.1088/1361-6501/ab62c8

Sparsity-based fractional spline wavelet denoising via overlapping group shrinkage with non-convex regularization and convex optimization for bearing fault diagnosis

Lei Wang[1], Xin Zhang[2,5], Zhiwen Liu[3] and Jinglin Wang[4]

[1] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, People's Republic of China
[2] School of Aeronautics and Astronautics, Sichuan University, Chengdu, Sichuan 610065, People's Republic of China
[3] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, People's Republic of China
[4] Aviation Key Laboratory of Science and Technology on Fault Diagnosis and Health Management, AVIC Shanghai Aeronautical Measurement-Controlling Research Institute, Shanghai 201601, People's Republic of China

E-mail: wang_llei@163.com and zhangx_15@126.com

Received 4 August 2019, revised 2 December 2019
Accepted for publication 17 December 2019
Published 31 January 2020

Abstract

Vibration monitoring has become a relatively common method for the fault diagnosis of rolling bearings in recent years. Vibration-based fault diagnosis technologies essentially rely upon accurate fault signature extraction from noisy vibration signals. To this end, this paper presents a sparsity-assisted denoising method using the fractional spline wavelet transform (FrSWT) and overlapping group shrinkage (OGS) with non-convex regularization and convex optimization (OGSNCRCO) to diagnose bearing faults. The FrSWT is used to promote sparsity of the wavelet coefficients for bearing fault signals. The OGSNCRCO balances data consistency and sparsity by using non-convex regularization and convex optimization, and is applied to shrink the wavelet coefficients of noisy signals in the fractional spline wavelet domain to extract bearing fault signals. Simulations and experiments validate the efficiency of the proposed method for fault diagnosis of rolling bearings. The results of analysis suggest that the proposed method is superior to other state-of-art techniques also commonly used for extracting fault signatures of rolling bearings, including the L1-norm denoising method (FrSWT with L1-norm regularization) and spectral kurtosis method.

Keywords: fault diagnosis, rolling bearings, fractional spline wavelet transform, overlapping group shrinkage, non-convex regularization, convex optimization

(Some figures may appear in colour only in the online journal)

1. Introduction

Rolling bearings are indispensable components of most power transmission systems. However, owing to harsh operating conditions, rolling bearings are prone to faults, which often decrease the performance of a system and cause other components, or even the whole system, to fail [1–4]. Therefore, incipient bearing fault detection and diagnosis are crucial to prevent system failures and ensure safe operation of the system. Since vibration signals can directly represent the dynamic behavior of bearings and are sensitive to structural changes, vibration-based techniques are one of

[5] Author to whom any correspondence should be addressed.

the most effective and prevailing methods for bearing fault diagnosis [5–8].

However, fault signals are often submerged in strong noise when incipient faults occur in bearings. Accurate recovery of incipient bearing fault signal from a measured vibration signal corrupted by noise remains a challenge in fault diagnosis for rolling bearings. Many signal processing methods have been introduced to diagnose bearing faults, including empirical mode decomposition (EMD) [9–11], variational mode decomposition (VMD) [12–14], spectral kurtosis (SK) [15, 16] and wavelet transform (WT) [17–19]. The EMD automatically decomposes a signal into intrinsic mode functions without presupposed basis function and reserves detail of the signal. The VMD turns signal decomposition into construct and solve a constrained variational problem and has robust performance against noise. The SK identifies the location of bearing fault signal in the frequency domain (center frequency and frequency band) based on the sensitivity of kurtosis to transient signal. After the raw signal is processed by the band-pass filter of which the parameters are determined by SK, fault features are extracted by applying the Hilbert transform or fast Fourier transform (FFT) to the filtered signal. The aforementioned methods are band bandpass filters in essence, which reserve bearing fault signal in specific frequency band and remove noise out of the frequency band. However, the in-scale noise remains in the reconstructed signal.

Wavelet denoising methods utilize the differences between the distributions of noise and bearing fault signal in time–frequency domain, and can reserve more signal of interest while reducing more noise. The WT maps signal to time–frequency space by computing inner products between signal and prototype wavelet basis. Bearing fault signal is represented by WT as a few large-amplitude wavelet coefficients, that is to say, sparse representations. Since noise is characterized by small-amplitude coefficients distributed throughout the wavelet domain, bearing fault signal with large-amplitude wavelet coefficients can be identified by using threshold shrinkage [20] or sparse regularization [21]. Therefore, in some ways, the performance of wavelet denoising methods is determined by the sparsity of the wavelet coefficients and coefficient shrinkage algorithm. For sparsity, the wavelet basis should be appropriately chosen to match the signal of interest [22]. Spline wavelets played an indispensable role in the early formation and development of the WT theory [23–25] and were also applied quite early for fault diagnosis [26]. Fractional spline wavelets are the generalized version of conventional spline wavelet with continuous order parameter, denoted by α [27–29]. Katunin [30] employed 2D fractional spline wavelets of which the parameters are set by using optimization algorithm to detect and localize damage in composite plates. Yang [31] combined fractional spline wavelet transform (FrSWT) and principal component analysis for image fusion and enhanced the spatial details of multi-spectral images. Shen [32] applied FrSWT to decompose the measured fault signal of gear system and selected one level of detail coefficients to reconstruct the gear fault signal. As a result, the in-scale noise remained in the obtained signal. Fractional spline wavelets have a fractional approximation order and are promising for analysis of the signals from complex systems. Moreover, the FrSWT is a fully-discrete self-inverting transform that can be efficiently implemented by applying the radix-2 FFT. Therefore, the FrSWT is an effective tool that can be adopted to diagnose bearing faults.

The overlapping group shrinkage (OGS) algorithm addresses the problem of recovering signal x, from an observation y, represented by

$$y(i) = x(i) + w(i),\ i \in Z = \{0, \ldots, N-1\} \tag{1}$$

where w represents Gaussian white noise and x is a group-sparse vector. The OGS assumes that large-amplitude values of x tend to form clusters (or groups) and are further overlapping groups. Due to the overlapping groups, OGS is shift-invariant and can prevent blocking artifacts [33]. Deng [34] proposed the adaptive OGS in which the regularization parameter is determined by the Bayesian framework to shrink the first-order difference of the heart sound signal. Zhao [35] adopted translation invariant wavelet transform and OGS to denoising electrocardiogram signal. Owing to the group-sparse characteristics of bearing fault signals in wavelet domain, the OGS was successfully applied to bearing fault diagnosis [36]. The aforementioned literature utilized convex regularization in the OGS optimization. Although the convex regularization makes the optimization problem easily solved by the convex optimization theory, it tends to underestimate large-amplitude components of the solution [37]. Chen and Selesnick [38] adopted non-convex sparsity regularization to OGS algorithm which meanwhile keeps the total cost function strictly convex. Non-convex regularization can yield a sparser solution than convex regularization for a given residual energy. On the other hand, strict convexity of the total cost function ensures a unique minimizer, which can be obtained by robust convex optimization methods.

In this paper, a sparsity-assisted denoising method using FrSWT and OGS with non-convex regularization and convex optimization (OGSNCRCO) is proposed for bearing fault diagnosis. First, the FrSWT is adopted to represent vibration signals collected from working bearings in a sparse way. Then, the wavelet coefficients are shrunken using OGSNCRCO, which uses non-convex regularization while retains the convexity of the total cost function. Finally, bearing fault signals are reconstructed via the inverse FrSWT and diagnosis results are deduced from the information of the reconstructed signals. Another contribution of this paper is the adaptive selection strategy of the regularization parameter of OGSNCRCO in different decomposition levels of FrSWT.

The structure of the paper is organized as follows. The FrSWT theory is introduced and OGSNCRCO is described in section 2. Furthermore, we demonstrate the group-sparsity of periodic transient signal in the fractional spline wavelet domain and propose a fault diagnosis method for bearings using FrSWT and OGSNCRCO. Section 3 presents the analysis results of a simulated signal, and section 4 is devoted to the experimental analyses for bearing fault diagnosis using the proposed method. Finally, section 5 gives some concluding remarks.

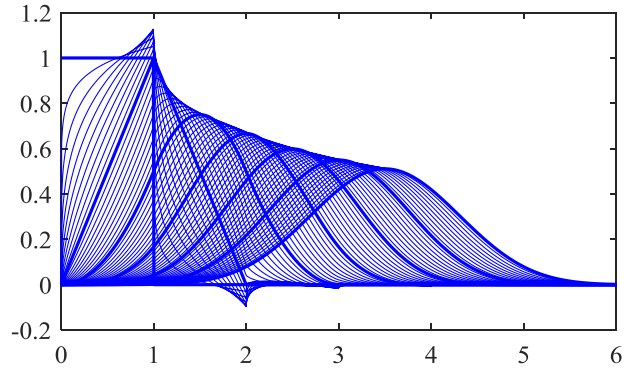

Figure 1. The fractional causal B-splines of order $0 \leqslant \alpha \leqslant 6$ with the step of 0.1.

2. Methodology

2.1. FrSWT

2.1.1. Fractional splines. Unser and Blu [29] extended fractional B-splines from Schoenberg's spline family with integer order to all fractional degrees $\alpha > -1$. Taking the $(\alpha+1)$th fractional difference for the one-sided power function, the fractional causal B-splines can be defined as

$$\begin{aligned}\beta_+^\alpha(x) &= \frac{1}{\Gamma(\alpha+1)}\Delta_+^{\alpha+1}x_+^\alpha \\ &= \frac{1}{\Gamma(\alpha+1)}\sum_{k=0}^{\infty}(-1)^k\binom{\alpha+1}{k}(x-k)_+^\alpha\end{aligned} \tag{2}$$

where k represents the integer greater than or equal to zero, and $\Gamma(\alpha+1)$ denotes Euler's Gamma function expressed as

$$\Gamma(\alpha+1) = \int_0^{+\infty} x^\alpha e^{-x}\mathrm{d}x \tag{3}$$

the one-sided power function is defined as

$$x_+^\alpha = \begin{cases} x^\alpha, & x \geqslant 0 \\ 0, & \text{otherwise}\end{cases} \tag{4}$$

the fractional forward finite difference operator of order α is written as

$$\Delta_+^\alpha f(x) = \sum_{k=0}^{\infty}(-1)^k\binom{\alpha}{k}f(x-k) \tag{5}$$

and the generalized fractional binomial coefficients are expressed as

$$\binom{\alpha}{k} = \frac{\Gamma(\alpha+1)}{\Gamma(k+1)\Gamma(\alpha-k+1)}. \tag{6}$$

Fractional B-splines of the order $0 \leqslant \alpha \leqslant 6$ with a step of 0.1 are shown in figure 1. The fractional B-splines are obtained by interpolating the conventional polynomial B-splines at each integer (thicker lines). Similar to causal B-splines, anti-causal B-splines of degree α is given by

$$\beta_-^\alpha(x) = \frac{1}{\Gamma(\alpha+1)}\Delta_-^{\alpha+1}x_-^\alpha = \beta_+^\alpha(-x) \tag{7}$$

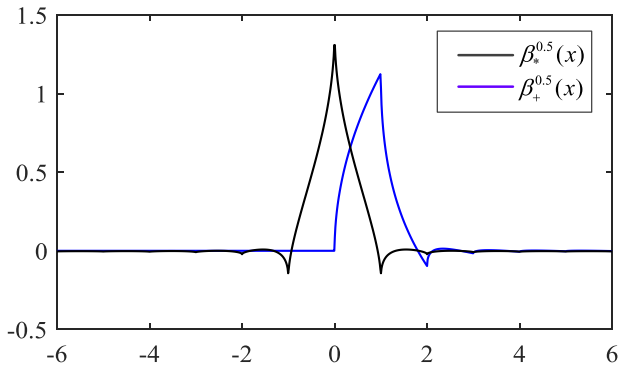

Figure 2. Plot of $\beta_+^{0.5}(x)$ and $\beta_*^{0.5}(x)$.

where $x_-^\alpha = (-x)_+^\alpha$, and Δ_-^α denotes the fractional backward finite difference operator, given by

$$\Delta_-^\alpha f(x) = \sum_{k=0}^{\infty}(-1)^k\binom{\alpha}{k}f(x+k). \tag{8}$$

The centered fractional B-splines of degree α are defined as

$$\beta_*^\alpha = \beta_+^{\frac{\alpha-1}{2}} * \beta_-^{\frac{\alpha-1}{2}} \tag{9}$$

where $*$ denotes convolution operator and β_*^α can be calculated as

$$\beta_*^\alpha(x) = \frac{1}{\Gamma(\alpha+1)}\Delta_*^\alpha|x|_*^\alpha = \frac{1}{\Gamma(\alpha+1)}\sum_{k=Z}(-1)^k\left|\binom{\alpha+1}{k}\right||x-k|_*^\alpha \tag{10}$$

where

$$|x|_*^\alpha = \begin{cases}\frac{|x|^\alpha}{-2\sin(\pi\alpha/2)}, & \alpha \neq 2m \text{ (not even)} \\ \frac{x^{2m}\log|x|}{(-1)^{1+m}\pi}, & \alpha = 2m \text{ (even)}\end{cases} \tag{11}$$

$$\left|\binom{\alpha}{k}\right| = \binom{\alpha}{k+\alpha/2} \tag{12}$$

and $\Delta_*^\alpha \overset{\text{Fourier}}{\longleftrightarrow} \left|1-e^{-j\omega}\right|^\alpha$ denotes the symmetrized version of the fractional finite difference operator.

Figure 2 shows a plot of fractional B-splines of order $\alpha = 0.5$ ($\beta_+^{0.5}$ and $\beta_*^{0.5}$). It can be seen that the support of $\beta_+^{0.5}$ belongs to $\mathbb{R}^+$ and $\beta_*^{0.5}$ is symmetric. The Fourier transform of the causal and centered fractional splines can be calculated as follows:

$$\hat{\beta}_+^\alpha(\omega) = \left(\frac{1-e^{-j\omega}}{j\omega}\right)^{\alpha+1} \tag{13}$$

$$\hat{\beta}_*^\alpha(\omega) = \left|\frac{\sin(\omega/2)}{\omega/2}\right|^{\alpha+1}. \tag{14}$$

2.1.2. Implement of FrSWT. The lack of compact support is the only disadvantage of the fractional splines when used to construct the wavelet basis. Fortunately, fractional splines decay fast enough to have an effective bounded support. In fact, fractional splines satisfy all the requirements for

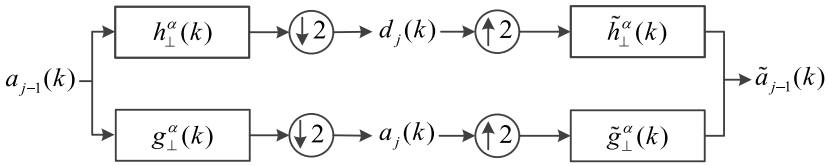

Figure 3. The analysis and synthesis filter-bank for the implement of FrSWT. ($h_{\perp}^{\alpha}(k)$, $g_{\perp}^{\alpha}(k)$, $\tilde{h}_{\perp}^{\alpha}(k)$ and $\tilde{g}_{\perp}^{\alpha}(k)$ denote the high-pass and low-pass filter coefficients of analysis and synthesis filter banks, respectively. $a_{j-1}(k)$ is approximation coefficients in scale $(j-1)$, $\tilde{a}_{j-1}(k)$ is reconstructed approximation coefficients in scale $(j-1)$, $a_j(k)$ is approximation coefficients in scale j and $d_j(k)$ is detail coefficients in scale j).

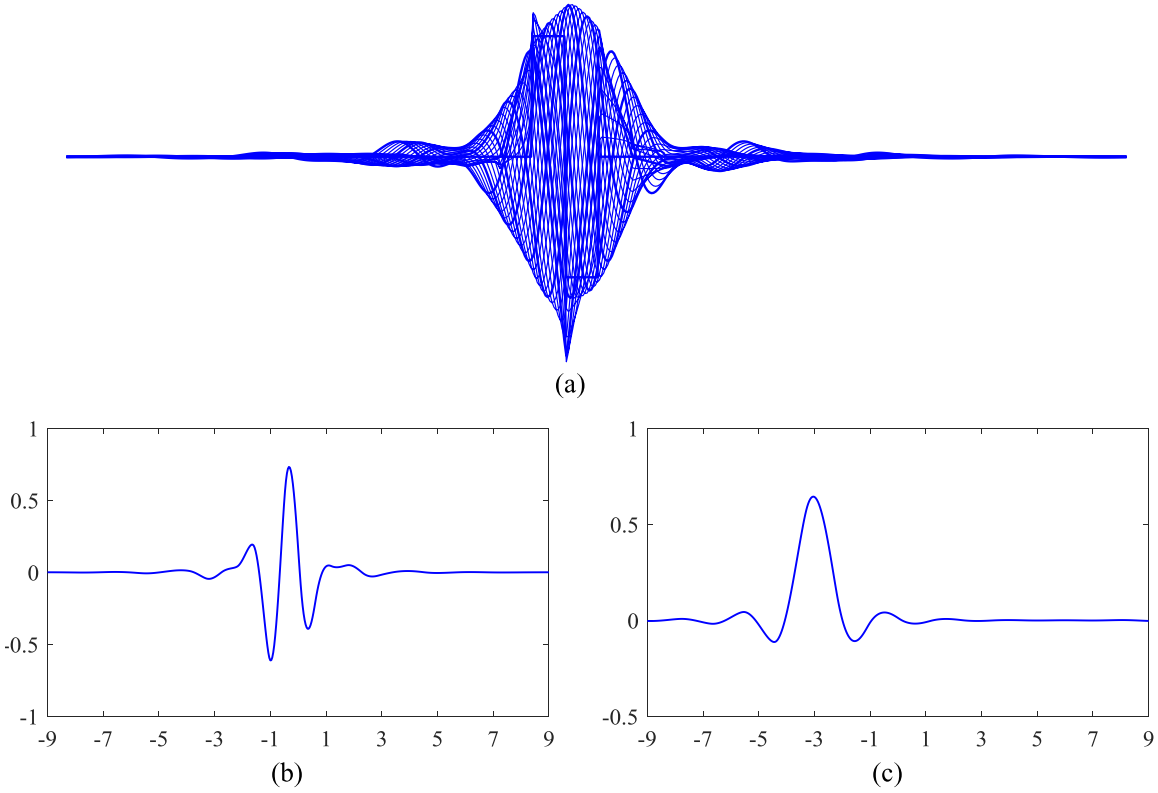

Figure 4. (a) The orthogonal fractional causal spline wavelet of order $0 \leqslant \alpha \leqslant 6$ with the step of 0.2; (b) the wavelet function of order $\alpha = 2.5$; (c) The scale function of order $\alpha = 2.5$.

constructing wavelet basis for $\alpha > -0.5$, particularly, the two-scale relation

$$\beta^{\alpha}(x/2) = \sum_{k \in Z} h^{\alpha}(k)\beta^{\alpha}(x-k) \tag{15}$$

where the refinement filters $h^{\alpha}(k)$ can denote $h_{+}^{\alpha}(k)$, $h_{-}^{\alpha}(k)$ and $h_{*}^{\alpha}(k)$, expressed as

$$h_{+}^{\alpha}(k) = \frac{1}{2^{\alpha}}\binom{\alpha+1}{k} \leftrightarrow \hat{h}_{+}^{\alpha}(\omega) = 2\left(\frac{1+e^{-j\omega}}{2}\right)^{\alpha+1} \tag{16}$$

$$h_{*}^{\alpha}(k) = \frac{1}{2^{\alpha}}\left|\begin{matrix}\alpha+1\\ k\end{matrix}\right| \leftrightarrow \hat{h}_{*}^{\alpha}(\omega) = 2\left|\frac{1+e^{-j\omega}}{2}\right|^{\alpha+1} \tag{17}$$

and $h_{-}^{\alpha}(k) = h_{+}^{\alpha}(-k)$, $\hat{h}_{-}^{\alpha}(\omega) = \hat{h}_{+}^{\alpha}(-\omega)$.

By orthonormalizing the fractional splines via the general approach in [39], the orthogonal scaling function is expressed as

$$\phi(x) = \sum_{k \in Z} (a_{\varphi}^{\alpha}(k))^{-1/2}\beta^{\alpha}(x-k) \tag{18}$$

where $(a_{\varphi}^{\alpha}(k))^{-1/2}$ is the inverse convolution square-root of the fractional spline auto-correlation sequence, $a_{\varphi}^{\alpha}(k)$. $a_{\varphi}^{\alpha}(k)$ and its Fourier transform are given by

$$a_{\varphi}^{\alpha}(k) = \langle \beta^{\alpha}(x), \beta^{\alpha}(x-k)\rangle = \beta_{*}^{2\alpha+1}(k) \tag{19}$$

$$A_{\varphi}^{\alpha}(\omega) = \sum_{n \in Z} \beta_{*}^{2\alpha+1}(n)e^{-j\omega n}. \tag{20}$$

The corresponding two-scale relation is

$$\phi(x/2) = \sum_{k \in Z} h_{\perp}^{\alpha}(k)\phi(x-k). \tag{21}$$

Using Mallat's pyramid algorithm, the low-pass and high-pass filters can be written as

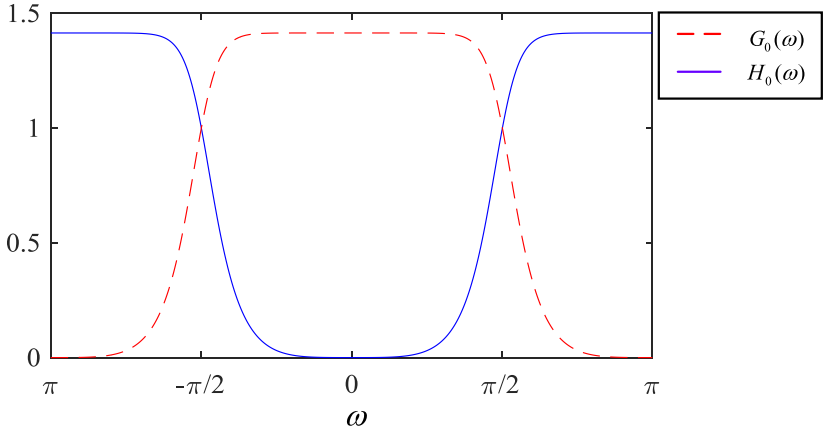

Figure 5. Frequency responses of the low-pass $G_0(\omega)$ and high-pass filters $H_0(\omega)$ for $\alpha = 2.5$.

$$H_{\perp}^{\alpha}(\omega) = \hat{h}^{\alpha}(\omega)\sqrt{\frac{A_{\varphi}^{\alpha}(\omega)}{A_{\varphi}^{\alpha}(2\omega)}} \tag{22}$$

and

$$G_{\perp}^{\alpha}(\omega) = e^{-j\omega}\overline{H_{\perp}^{\alpha}(\omega+\pi)}, \tag{23}$$

where $\hat{h}^{\alpha}(\omega)$ can be expressed as $\hat{h}_{+}^{\alpha}(\omega)$, $\hat{h}_{*}^{\alpha}(\omega)$ and $\hat{h}_{-}^{\alpha}(\omega)$.

Figure 3 shows a block diagram of the analysis and synthesis filter banks for the implement of FrSWT. The FrSWT can be implemented by recursively applying the two-channel analysis filter bank on the low-pass channel output. The orthonormal fractional causal spline wavelets are plotted in figure 4. The shape of the wavelet is very similar to the impact signal induced by bearing fault, which is often modeled as a single-side impact signal. Frequency responses of the low-pass and high-pass filters are displayed in figure 5 ($\alpha = 2.5$). As α increases, the filters tend towards ideal low-pass and high-pass filters.

2.2. OGSNCRCO

In sparse regularization, shrinkage (or thresholding) functions are determined by solving the optimization problem, as follows:

$$x^{*} = \arg\min_{x}\left\{F(x) = \frac{1}{2}\|y-x\|_2^2 + \lambda R(x)\right\} \tag{24}$$

where x can either represent the wavelet coefficients of a signal or the signal itself, provided the sparsity is properly presented; the term $\frac{1}{2}\|y-x\|_2^2$ is to ensure the consistency of the solution x^{*} to x; $R(x)$ is the regularization (or penalty) term which guarantees the sparsity of the solution; $\lambda > 0$ is the regularization parameter that balances consistency and sparsity of the solution.

Most shrinkage functions can be regarded as solutions to the problem of equation (24). In particular, the solution of the problem of equation (24) when $R(x) = \|x\|_1$ (L1 norm), is known as the soft threshold function. In the OGS algorithm with convex regularization, $R(x)$ is in the form of $R(x) = \sum_i \left[\sum_{j=0}^{K-1}|x(i+j)|^2\right]^{1/2}$ ($\{i = 1, 2, \ldots N\}$ denotes the group index, and $\{j = 0, 2, \ldots K-1\}$ denotes the coefficient index within the group i), which can capture the overlapping group sparsity of the data. However, the convex penalty tends to underestimate large-amplitude components of the sparse solution, which is usually related to signal of interest. The OGSNCRCO can prevent underestimation like convex penalties and simultaneously maintain the convexity of the total cost function.

2.2.1. OGSNCRCO.

The cost function of the OGS algorithm with group size K takes the general form of (with convex or non-convex penalty):

$$F(x) = \frac{1}{2}\|y-x\|_2^2 + \lambda\sum_{i\in Z}\phi(\|x_{i,K}\|_2; a) \tag{25}$$

where $x_{i,K} = [x(i), \ldots, x(i+K-1)] \in \mathbb{R}^K$ and $\|x_{i,K}\|_2 = \left[\sum_{j=0}^{K-1} x_{i+j}^2\right]^{1/2}$. To retain the convexity of the total cost function while using non-convex regularization, the sparsity promoting regularization function is assumed to satisfy the following properties:

(1) $\phi(-x; a) = \phi(x; a)$ (even symmetric).
(2) $\phi(x; a)$ is a continuous function on $\mathbb{R}$.
(3) $\phi(x; a)$ is quadratic differentiable on $\mathbb{R}\backslash\{0\}$ where $\mathbb{R}\backslash\{0\} = \{x \in \mathbb{R},\ x \notin 0\}$.
(4) $\phi'(x; a) > 0,\ \forall x > 0$ (increasing on $\mathbb{R}_{+}$).
(5) $\phi'(0^{+}; a) = 1$ (unit slope at 0).
(6) $\phi''(x; a) \leqslant 0,\ x > 0$ (concave on $\mathbb{R}_{+}$).
(7) $\phi''(0^{+}; a) \leqslant \phi''(x; a),\ \forall x > 0$.
(8) $\phi''(0^{+}; a)$ is finite.

The scalar parameter $a > 0$ controls the concavity of the function. The cost function, F (in equation (25)), is proven to be strictly convex if [37]

$$0 < a < \frac{1}{\lambda K}. \tag{26}$$

Table 1 lists some common regularization functions that satisfy the above conditions, and the corresponding functions $\phi'(u)/u$. The functions $\phi'(u)/u$ have very similar functional forms. As $a \to 0$, other three penalty functions ($\phi_{\log}$, ϕ_{atan} and ϕ_{rat}) tend to the absolute value function, $\phi_{\text{abs}}(x) = |x|$. Figure 6 displays the four penalty functions when $a = 0.2$.

Table 1. Sparsity-promoting penalty functions.

Penalty	$\phi(u)$	$\phi'(u)/u$
abs	$\lvert u\rvert$	$\frac{1}{\lvert u\rvert}$
log	$\frac{1}{a}\log(1+a\lvert u\rvert)$	$\frac{1}{\lvert u\rvert(1+a\lvert u\rvert)}$
atan	$\frac{2}{a\sqrt{3}}(\tan^{-1}(\frac{1+2a\lvert u\rvert}{\sqrt{3}})-\frac{\pi}{6})$	$\frac{1}{\lvert u\rvert(1+a\lvert u\rvert+a^2\lvert u\rvert^2)}$
rat	$\frac{\lvert u\rvert}{1+a\lvert u\rvert/2}$	$\frac{1}{\lvert u\rvert(1+a\lvert u\rvert)^2}$

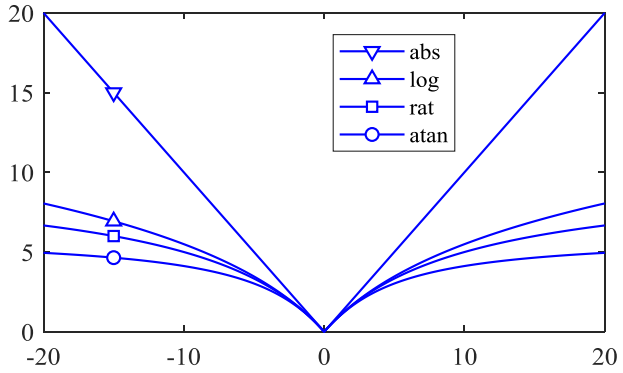

Figure 6. Some sparsity promoting penalty functions satisfying the assumptions in section 2.2.1.

For the same scalar parameter a, the arctangent penalty ϕ_{atan} exhibits the most 'concavity' among the four given penalty functions. In other words, ϕ_{atan} can induce the strongest sparsity and thus is used in this paper.

2.2.2. Majorization-minimization method. Owing to the strict convexity of the cost function, the majorization-minimization (MM) algorithm can be used to obtain the unique minimizer in OGSNCRCO. The MM method solves the complicated minimization problem by replacing it with a sequence of simpler problems [40]. More specifically, consider the optimization problem in equation (25). Using the MM procedure, the problem can be described as an iteration of

$$x^{(p+1)} = \arg\min_x Q(x, x^{(p)}) \tag{27}$$

where p is the iteration index of the function, and $Q: \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is the majorizer (upper bound) of F, which should satisfy

$$Q(x,v) \geqslant F(x),\ \forall x \in \mathbb{R}^N, \forall v \in \mathbb{R}^N \backslash \{0\} \tag{28}$$

$$Q(v,v) = F(v), \forall v \in \mathbb{R}^N \backslash \{0\}. \tag{29}$$

To solve equation (25), Q can be expressed as

$$Q(x,v) = \frac{1}{2}\|y-x\|_2^2 + \frac{\lambda}{2}\sum_i r(i;v)x^2(i) + C \tag{30}$$

where C is independent of x and $r: \mathbb{Z} \times \mathbb{R}^K \to \mathbb{R}$ is defined as

$$r(i;v) = \sum_{j=0}^{K-1} \frac{\phi'(\|v_{i-j,K}\|_2)}{\|v_{i-j,K}\|_2}. \tag{31}$$

Using equation (30) in the MM iteration, equation (27) can be rewritten as

Table 2. OGSNCRCO algorithm with penalty ϕ.

Input: $y \in \mathbb{R}^N,\ \lambda > 0,\ K,\ \phi, Nit.$
Initialize: $x^{(0)} = y$
$S = \{i \in Z_N \,|\, y(i) \neq 0\}$
For $p = 1{:}Nit$
$a(i) = \left[\sum_{j=0}^{K-1} \left|x^{(p-1)}(i+j)\right|^2\right]^{1/2}$, $b(i) = \frac{\phi'(a(i))}{a(i)}$,
$r(i) = \sum_{j=0}^{K-1} b(i-j)$, $x^{(p)}(i) = \frac{y(i)}{1+\lambda r(i)}$, $i \in S$.
End **for**
Return x.

$$x^{(p+1)}(i) = \frac{y(i)}{1+\lambda r(i; x^{(p)})},\ i \in \mathbb{Z}_N \tag{32}$$

where r is obtained from equation (31). The OGSNCRCO algorithm is described in table 2. The convergence of OGSNCRCO is guaranteed by the MM principle.

2.3. Proposed sparsity-assisted denoising scheme

Bearing fault signals are always modeled as pseudo-cyclostationary transient signals which exhibit typical group-sparsity [41, 42]. In this section, we propose a sparsity-assisted denoising scheme to diagnose bearing faults based on FrSWT and OGSNCRCO.

2.3.1. Group-sparsity of periodic transient signal in the fractional spline wavelet domain. For the OGSNCRCO algorithm, sparse representation of the target signal has a significant impact on the denoising results. In this way, a representation that maximizes sparsity of the target signal is desirable. The WT is a well-known tool for sparse representation. Orthogonal fractional spline wavelets are continuous versions of conventional orthogonal B-spline wavelets (i.e. Battle–Lemarie wavelets). Continuous-order α can control the vanishing moment of fractional spline wavelets. Orthogonal fractional spline wavelets essentially behave like fractional differentiators for non-integer α. To illustrate the sparse representations of bearing fault signal using FrSWT, consider the following simulated transient signal:

$$x(t) = \sum_{l=0}^{L-1} h(t - lT - T_0) \tag{33}$$

$$h(t) = \begin{cases} Ae^{-2\pi\zeta f_n t}\sin(2\pi f_n\sqrt{1-\zeta^2}t),\ t \geqslant 0 \\ 0,\ t < 0 \end{cases} \tag{34}$$

where amplitude $A = 1$, damping ratio $\zeta = 0.1$, time center constant $T_0 = 0$, natural frequency $f_n = 1000$, transient number $L = 21$, and transient period $T = 0.02s$.

The waveform of the simulated signal with a sampling rate of 20 kHz and 8196 points is depicted in figure 7. Figure 8 illustrates approximation coefficients and detail coefficients of the simulated bearing fault signal decomposed by FrSWT. The FrSWT parameters are set to $a = 2.6$ and $J = 10$. Note that detail coefficients present better sparsity than the original

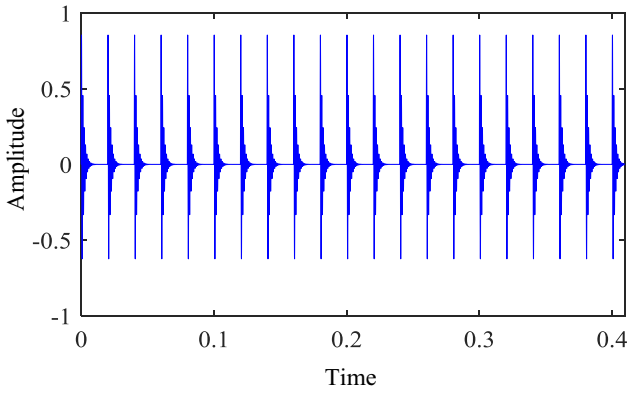

Figure 7. The waveform of the simulation bearing fault characteristic signal without noise.

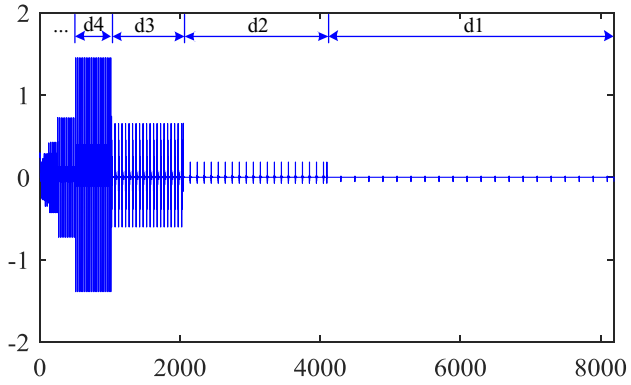

Figure 8. The sparse representations of the simulation signal in fractional spline wavelet domain.

signal in the time domain. The detail coefficients in scales 1, 2, 7, 8 and 9 are small, and the energy of the signal is concentrated in scales 3–6. From this, it follows that the coefficients of the simulated bearing fault signal in the fractional spline wavelet domain are sparse in both time and scale. Furthermore, note that the coefficients are not only sparse but also present grouping property in the time scale.

2.3.2. Denoising algorithm based on FrSWT and OGSNCRCO. Considering group sparsity of bearing fault signals in the fractional spline wavelet domain, the OGSNCRCO algorithm introduced in section 2.2 can be applied to shrink the wavelet coefficients of noisy signals collected from rolling bearings. A flow chart of the proposed method for bearing fault diagnosis is described in figure 9. The procedure can be carried out using the following steps:

(1) Measure vibration acceleration signal of working rolling bearing via a data acquisition system.
(2) Decompose the measured signal using FrSWT to obtain detail coefficients and approximation coefficients.
(3) Use OGSNCRCO to shrink the detail coefficients. The approximation coefficients are zeroed.
(4) Reconstruct the signal from the shrunken detail coefficients and approximation coefficients by using the inverse FrSWT. Bearing fault features can be identified from the reconstructed signal.

According to figure 8, detail coefficients in different scales contain different quantities of bearing fault signal. Considering that the kurtosis is sensitive to bearing fault signal, the reciprocal of the kurtosis is used to weight the regularization parameter in the OGSNCRCO. Thus, the regularization parameters in different scales are set to $\lambda_s = C/\mathrm{Kurt}_s$, where constant C is related to signal-to-noise ratio (SNR) of the signal and Kurt_s is the kurtosis of the wavelet coefficients in scale s. When the detail coefficients contain more bearing fault signal, the kurtosis is larger, λ_s is smaller and more coefficients are reserved.

3. Simulation validation

We provide a simulation study to illustrate the validation of the proposed method for periodic transient signal extraction in this section. To further validate the advantage of the proposed method, the results are compared with the results of the L1-norm denoising method (FrSWT with L1-norm regularization).

A noisy simulated signal is considered as follows:

$$y(t) = x(t) + n(t) + c(t) \tag{35}$$

where $x(t)$ is the bearing fault signal described in equations (33) and (34). $n(t)$ is Gaussian white noise with a standard deviation $\sigma = 0.5$. $c(t) = A_1 * \cos(2\pi f_1 t)$ simulates the harmonic component in bearing vibration signal where $A_1 = 0.5$ and $f_1 = 4000$ Hz. Figures 10(a)–(c) display its waveform in time domain, envelope spectrum and time–frequency distribution, respectively. The time–frequency distribution of the periodic transient signal without noise is presented in figure 10(d). It can be seen that the periodic transient signal is fully submerged in the noise. The SNR of the mixed signal is -12.58 dB. Besides, the kurtosis of the signal is 2.90, close to the value of Gaussian white noise.

The proposed sparsity-assisted denoising method is adopted to analyze the simulated signal $y(t)$ and extract the periodic transients from the noisy signal. The waveform, time–frequency distribution and envelope spectrum of the denoised signal are depicted in figures 11(a)–(c), respectively. Successive periodic transients with period $t \approx 0.02$ s are clearly observed in figures 11(a) and (c). Additionally, the envelope frequency of the transients $f \approx 50$ Hz and its multiple frequencies are able to be found in figure 11(b). For comparison, the L1-norm denoising method is applied to process the same simulated signal and the analysis results are displayed in figure 12. The signal denoised by L1-norm method still contains considerable noise and the periodic transients are difficult to identify in figure 12.

In summary, the proposed method preserves more features of periodic transients while the reconstructed signal of the L1-norm method is still contaminated by considerable noise.

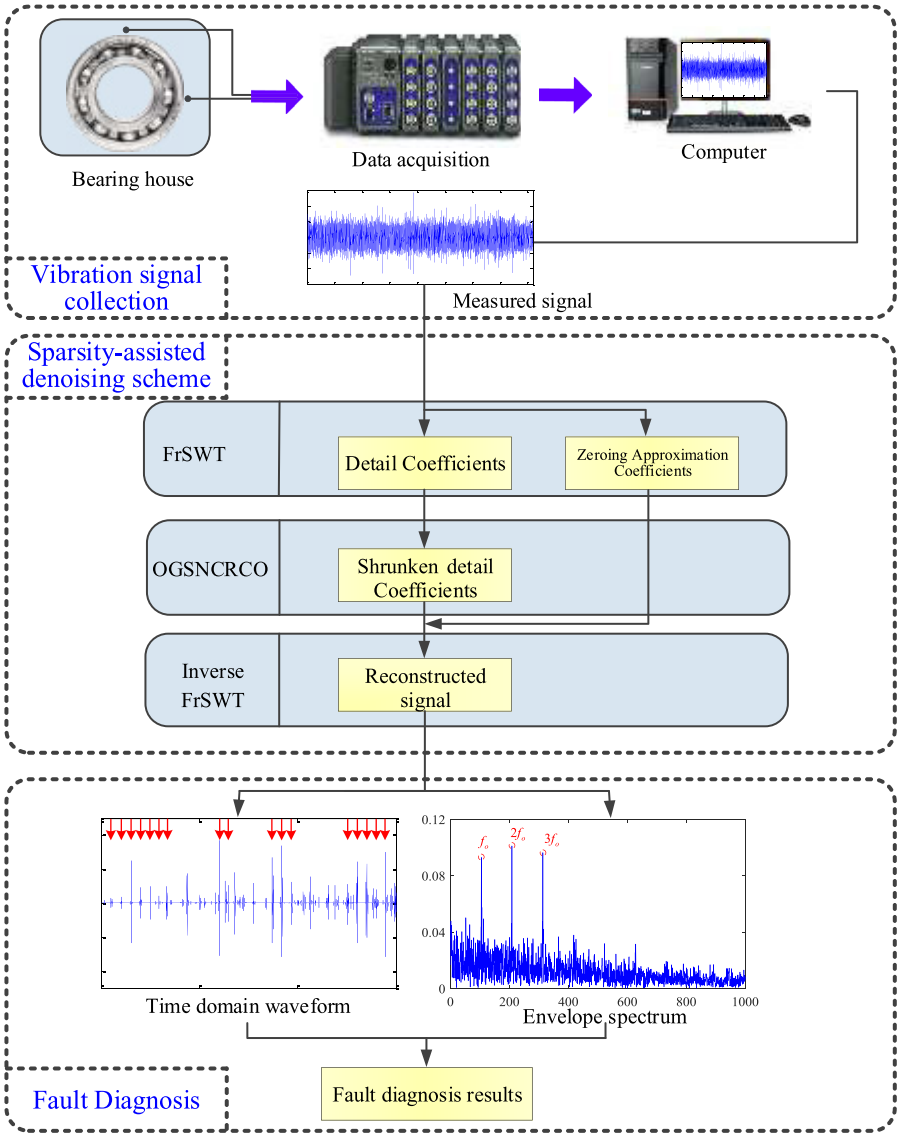

Figure 9. Flow chart of the proposed method for bearing fault diagnosis.

The envelope spectrum obtained by the proposed method clearly reveals the envelope frequency of the periodic transients with more harmonic frequencies, larger amplitudes, and fewer interference frequencies. Moreover, the time–frequency distribution of the denoised signal is more consistent with that of the original simulated signal displayed in figure 10(d). The L1-norm method reserves more noise and less bearing fault signal. Thus, the comparison demonstrates that the proposed method more effectively extracts the periodic transient signal than the L1-norm denoising method.

4. Experimental case studies

In section 3, we validated the efficiency of the proposed method for extracting periodic transient signal. In this section, we adopt the proposed method to analyze two vibration signals measured from a bearing with an implanted fault and a run-to-failure bearing to diagnose the bearing faults. To further verify the superiority, the proposed method is compared with two methods namely the L1-norm denoising method and the SK method.

4.1. Case study 1

4.1.1. Data acquisition. The bearing vibration signal is collected from an engine rotor-bearing-casing fault simulator [43], as described in figure 13(a). The experimental system is composed of a rotor-bearing-casing system, a miniature accelerometer of type 4508 from Brüel & Kjær Sound & Vibration Measurement A/S, and a data acquisition module of type NI9234. Figure 13(b) presents the bearing with implanted outer race fault, which is installed at the end of the shaft. The vibration signal is acquired by an accelerometer installed on the top of the bearing block. Table 3 lists the parameters of the test bearing to be diagnosed. The data is sampled at 10 kHz, and the length of the signal to be analyzed is 32768 points. The bearing fault characteristic frequencies are described in table 4.

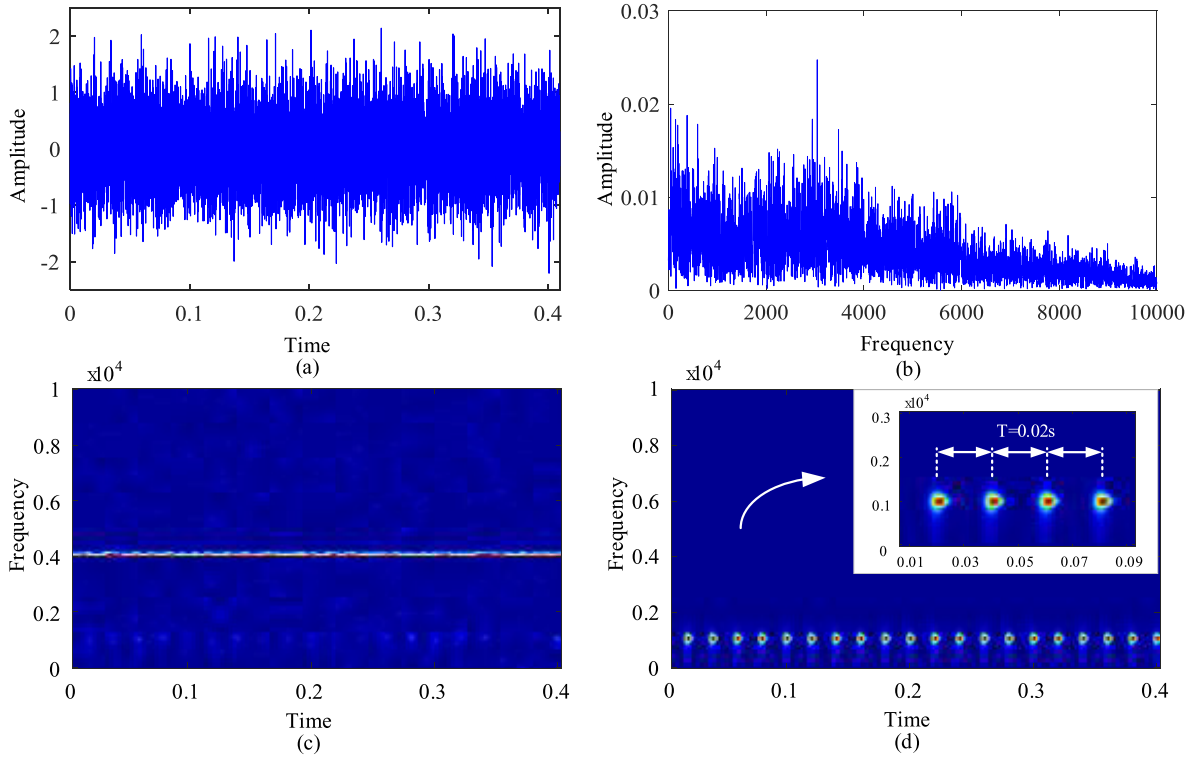

Figure 10. (a) Waveform of the mixed-signal, (b) its envelope spectrum, (c) its time-frequency distribution and (d) time-frequency distribution of the signal without noise.

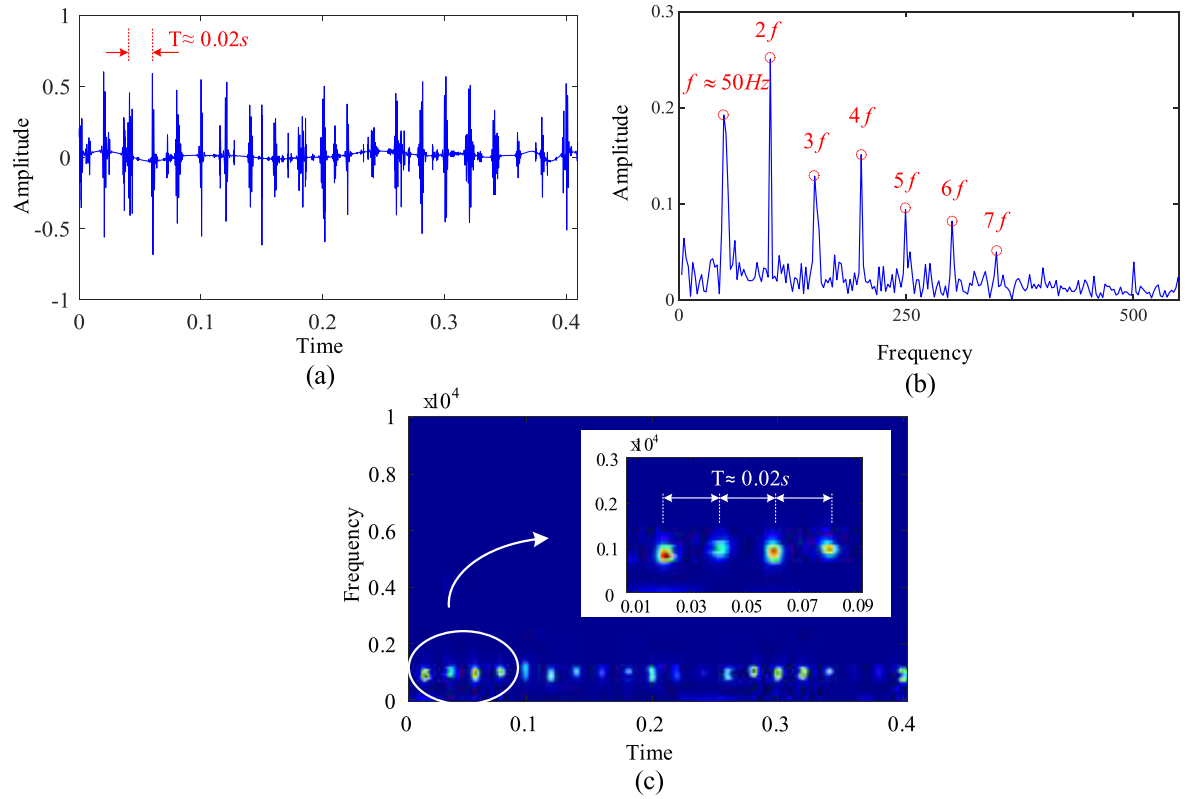

Figure 11. (a) Waveform of the denoised signal via the proposed method, (b) its envelope spectrum and (c) its time-frequency distribution.

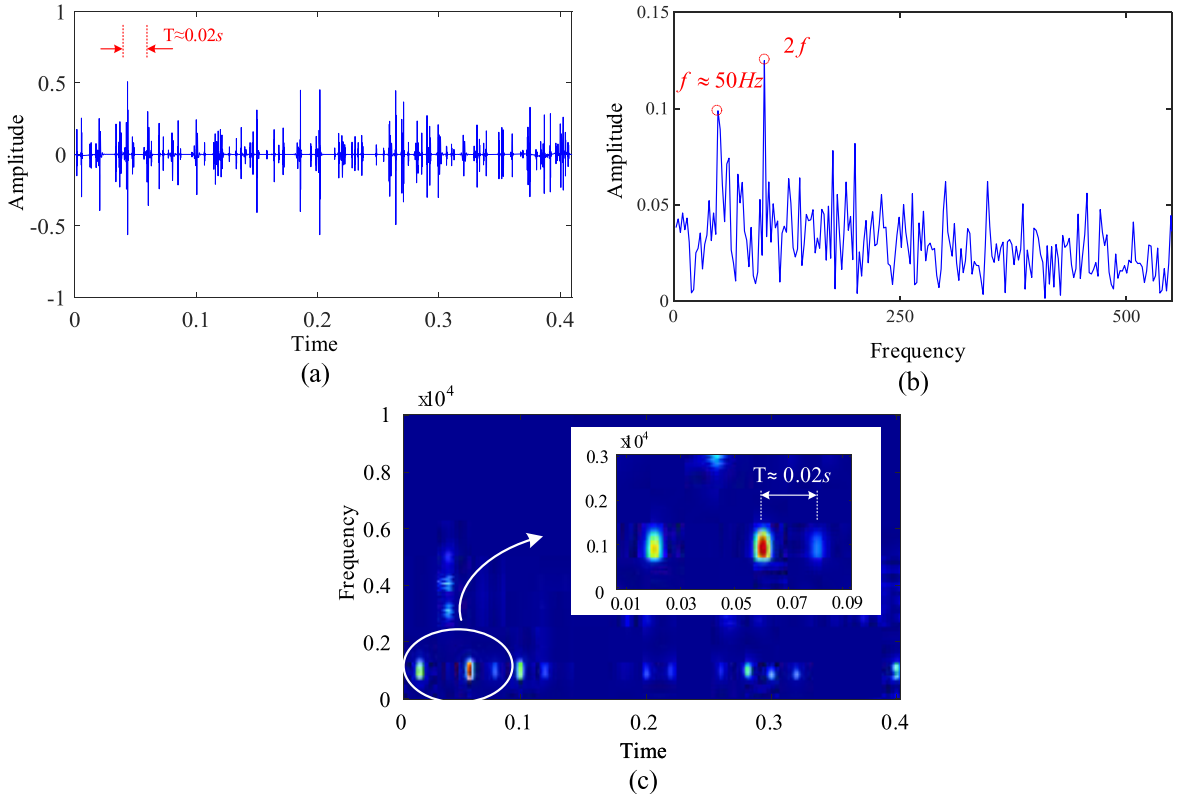

Figure 12. (a) Waveform of the denoised signal via L1-norm denoising method, (b) its envelope spectrum and (c) its time-frequency distribution.

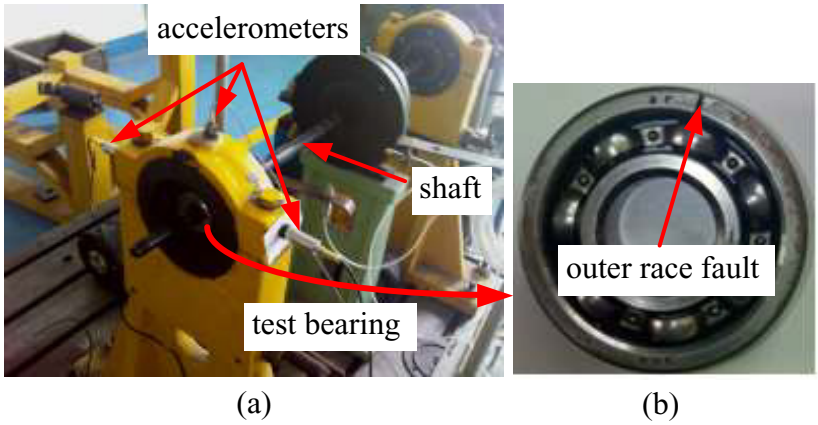

Figure 13. Case1: (a) the experiment rig and (b) the rolling bearing with an outer race fault.

Table 3. Bearing parameters in case 1 and the shaft rotation frequency.

Parameter	Roller diameter (d/mm)	Pitch diameter (D/mm)	Roller number (N)	Contact angle (α/°)	Shaft rotation frequency (f_r/Hz)
Value	9.6	36	7	0	34.25

Table 4. Bearing fault characteristic frequencies in case 1.

Inner race fault	Outer race fault	Roller fault	Cage fault
$f_i = 151.8$ Hz	$f_o = 87.9$ Hz	$f_b = 59.6$ Hz	$f_c = 12.6$ Hz

4.1.2. Results and analysis. Figure 14 shows the time domain and Fourier spectrum of the collected bearing vibration signal. Periodic transients are not able to be observed in the time domain, and the periodic impulse feature generated by the fault is almost fully submerged in noise.

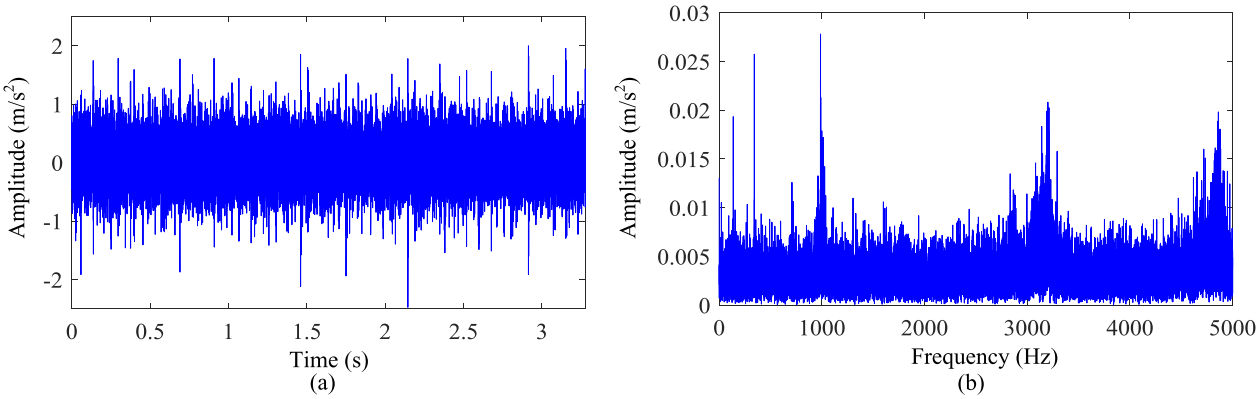

Figure 14. Case 1: (a) the time-domain waveform of the original signal and (b) its FFT spectrum.

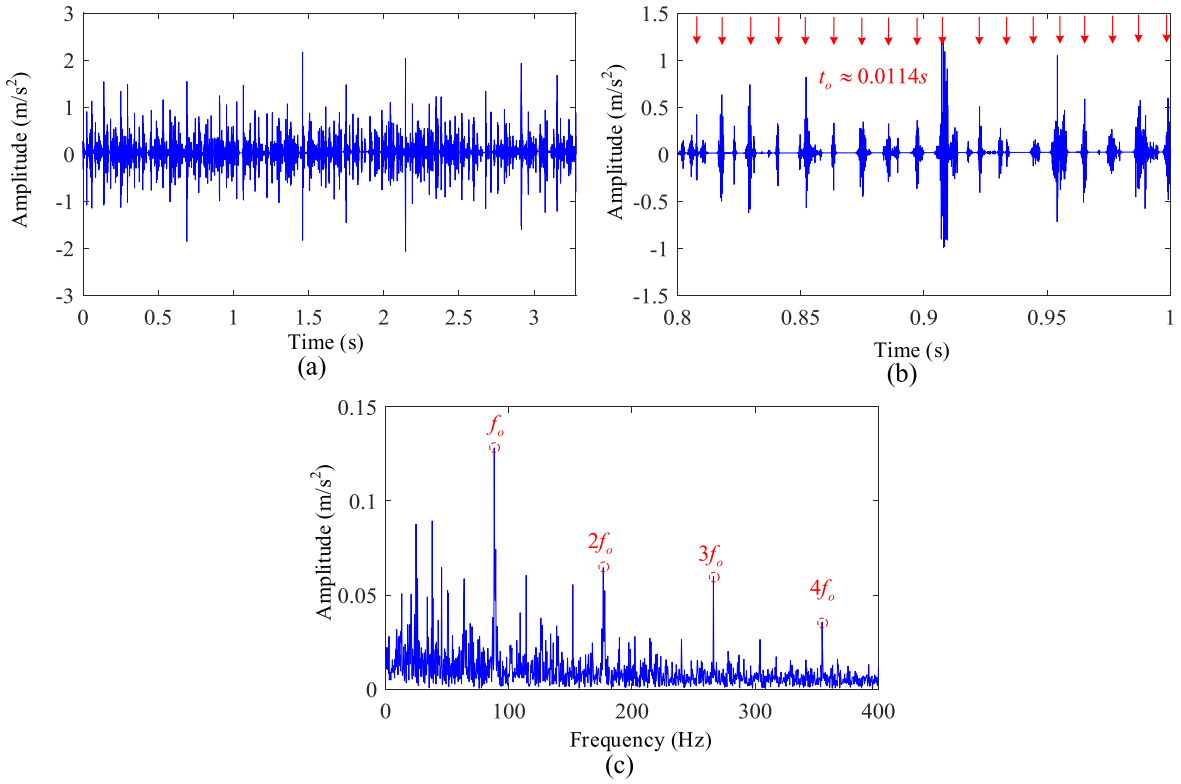

Figure 15. Denoised results via the proposed method in case 1: (a) the time-domain waveform, (b) its zoomed-in version and (c) its envelope spectrum.

The vibration signal is then analyzed by the proposed method. The analysis results are displayed in figure 15. Successive periodic transients are able to be clearly observed in the time domain of the denoised signal. The zoomed-in figure shows that period, $t_o \approx 0.0114$ s, is consistent with the ball pass period of the outer race fault. Meanwhile, the outer race fault characteristic frequency ($f_o = 87.9$ Hz) and its multiple frequencies are found in the envelop spectrum, as shown in figure 15(c).

For comparison, the L1-norm method and SK method are used to process the original signal. The analysis results are described in figures 16 and 17, respectively. It is easy to observe that the proposed method reserves more periodic transients related to the outer race fault while reducing noise, in comparison to the other two methods. Moreover, fault features are more clearly revealed in the envelope spectrum obtained by the proposed method, including large amplitudes of the outer race fault characteristic frequency and its multiple frequencies.

4.2. Case study 2

4.2.1. Data acquisition. The proposed method is studied by a run-to-failure bearing experiment for an aero engine application context in this case. Figures 18(a) and (b) present the experimental test rig and its structural diagram. An electric

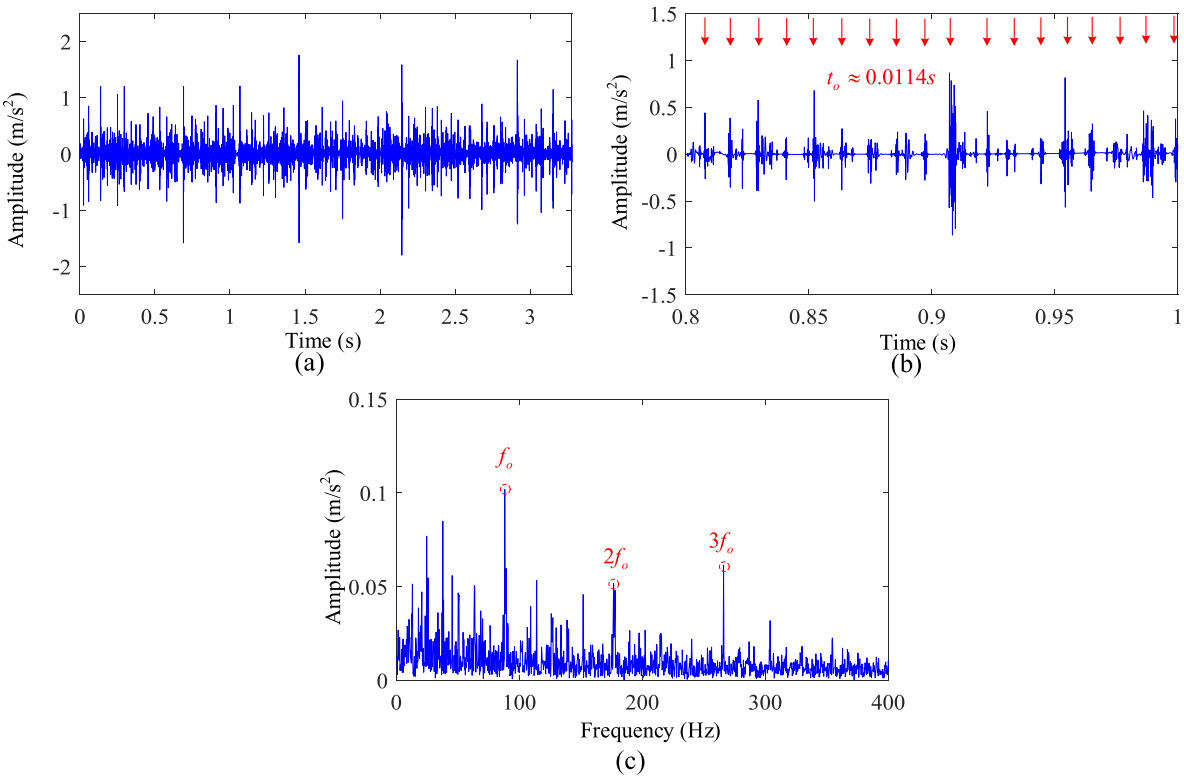

Figure 16. Denoised results via the L1-norm denoising method in case 1: (a) the time-domain waveform, (b) its zoomed-in version and (c) its envelope spectrum.

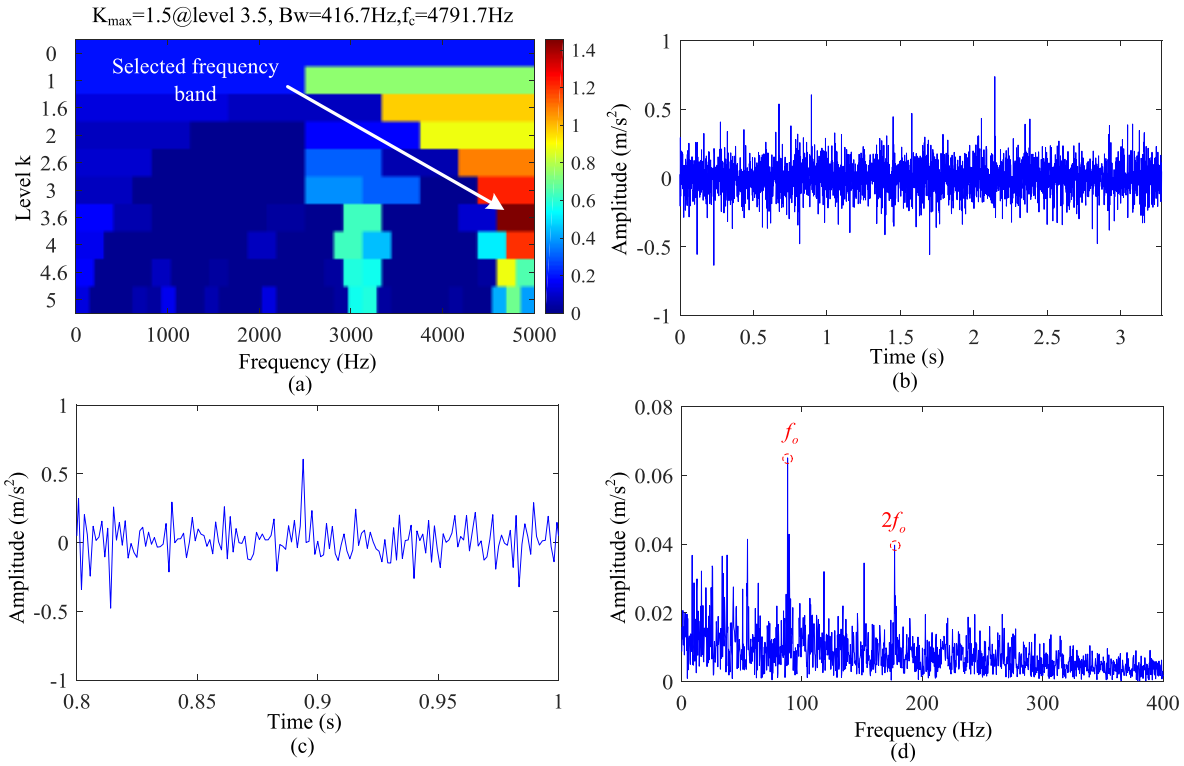

Figure 17. SK results in case 1: (a) the kurtogram, (b) the filtered signal via SK (c) its zoomed-in version and (d) its envelope spectrum.

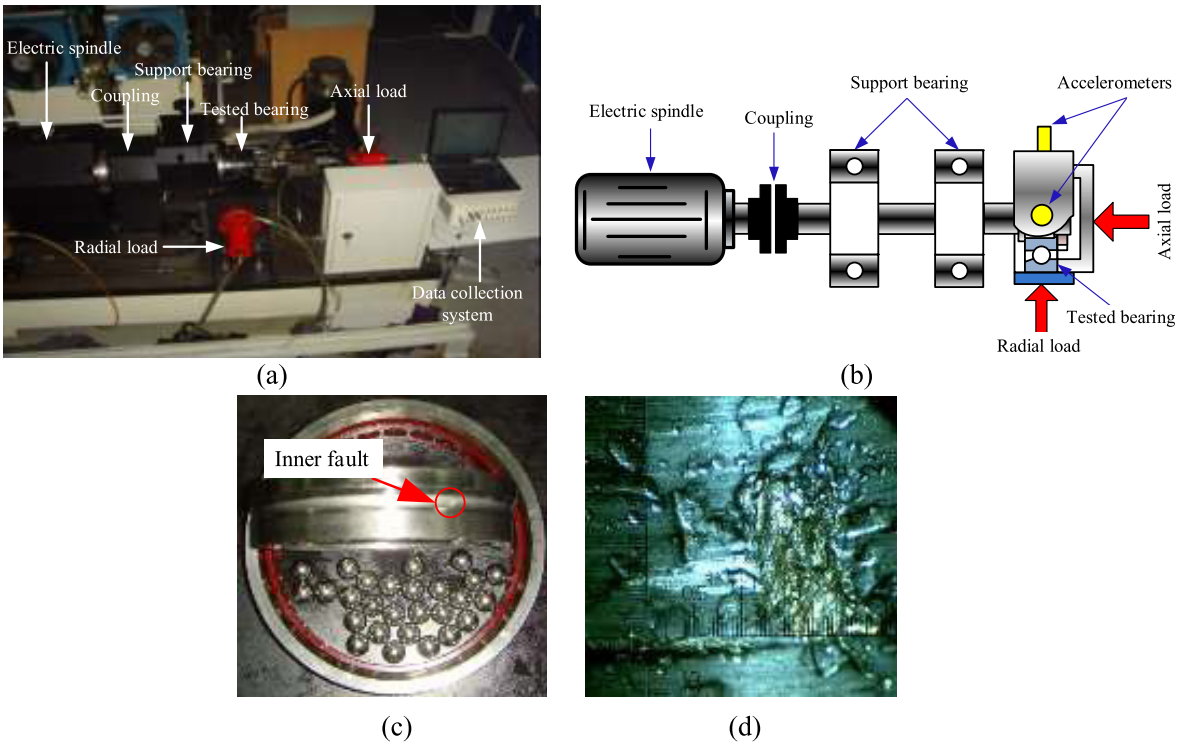

Figure 18. Case2: (a) the experiment rig, (b) the structure diagram of the test rig, (c) the rolling bearing with inner race fault and (d) the area of the spall.

Table 5. Bearing parameters in case 1 and the shaft rotation frequency.

Parameter	Roller diameter (d/mm)	Pitch diameter (D/mm)	Roller number (N)	Contact angle (α/°)	Shaft rotation frequency (f_r/Hz)
Value	11	115	27	15	100

Table 6. Bearing fault characteristic frequencies in case 1.

Inner race fault	Outer race fault	Roller fault	Cage fault
$f_i = 1475$Hz	$f_o = 1225$ Hz	$f_b = 518$ Hz	$f_c = 45$Hz

motor drives the test rig with a rotating speed of 6000 rpm. An angular contact ball bearing of type H7018C is mounted on the right of the shaft. A hydraulic loading system adds an axial load of 2 kN and a radial load of 11 kN to the test bearing. An oil circulation system is used to adequately lubricate the test bearing. An accelerometer is installed in the vertical direction of the bearing housing due to space constraints. A data acquisition system collects vibration signal at 20 kHz sampling frequency. A spall with an area of 3 mm^2 formed in the inner race after about 146 working hours (see figures 18(c) and (d)). Table 5 describes the parameters of the bearing of type H7018C used in the experiment. The bearing fault characteristic frequencies are listed in table 6.

4.2.2. Results and analysis. The time domain and Fourier spectrum of the bearing vibration signal with a length of 32768 are described in figure 19. The proposed method is employed to analyze the bearing vibration signal. As presented in figure 20, the proposed method effectively extracts the periodic transient signal while eliminating noise. In the time interval marked by the red ellipse, transients with a period of $t_i \approx 0.0007$ s are observed, consistent with the inner race fault. Furthermore, since the fault occurs on the bearing inner race, periodic changes in the load result in a modulation of the transients, as presented in the magnified figure. The rotating frequency, inner race fault characteristic frequency, and their multiple frequencies are dominant in the envelope spectrum, which further validates the effectiveness of the proposed method. Sideband frequencies resulting from modulation of periodic changes in the load are also able to be found in the envelope spectrum.

Analysis results using the L1-norm method and SK method are shown in figures 21 and 22, respectively. Although the L1-norm denoising method and SK can reveal

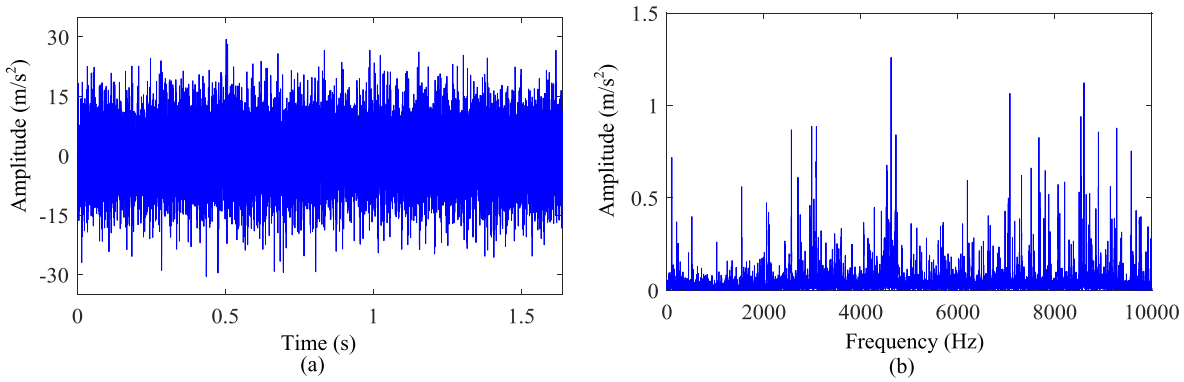

Figure 19. Case2: (a) the time-domain waveform of the original signal in case 2, (b) its FFT spectrum.

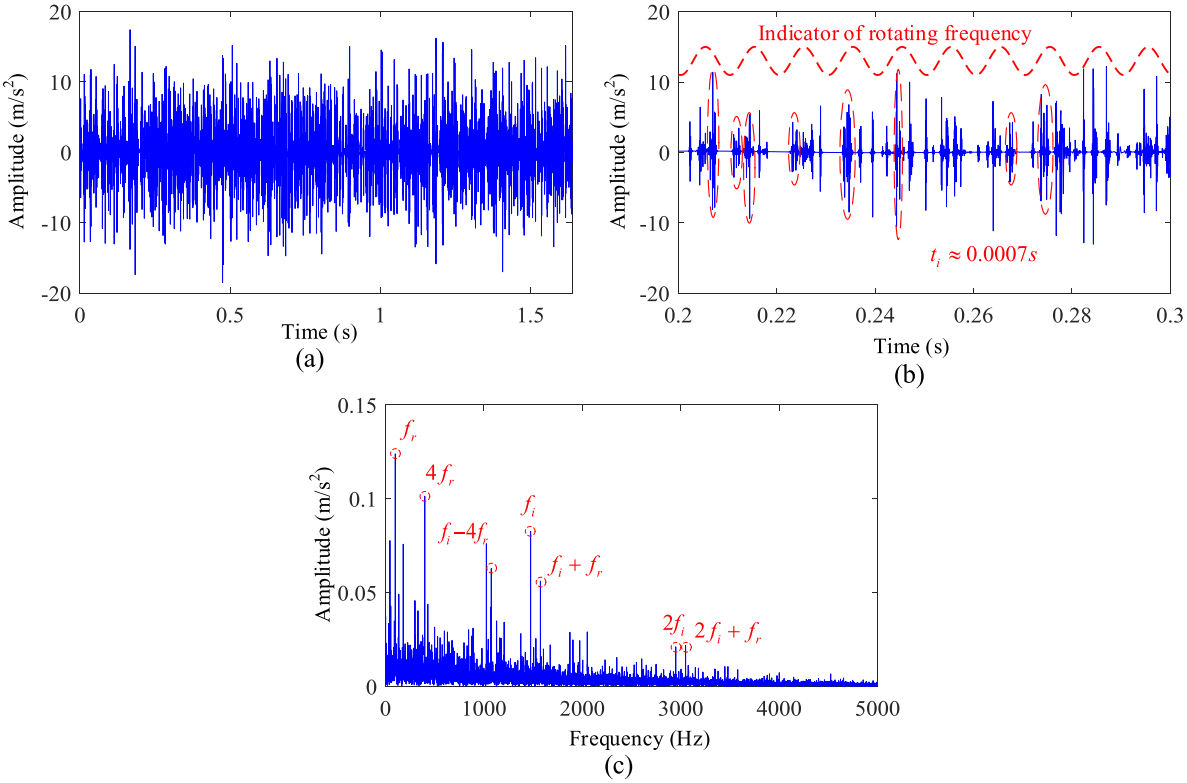

Figure 20. Denoised results via the proposed method in case 2: (a) the time-domain waveform, (b) its zoomed-in version and (c) its envelope spectrum.

the inner race fault, the proposed method shows much more clearly. From the denoised signals in the time domain, it can be observed that the proposed method preserves more transients related to the inner race fault. The modulation of the transients induced by the periodic changes of load is not observed in the SK-filtered signal. The envelope spectra also show that the signal energy aggregation in the rotating frequency and fault characteristic frequencies obtained by the proposed method is superior to those of L1-norm method and SK method.

5. Conclusions

In this paper, a sparsity-assisted denoising method to diagnose bearing faults based on FrSWT and OGSNCRCO is proposed. The FrSWT demonstrates robust performance in terms of sparse representation of bearing fault signals. Furthermore, the OGSNCRCO algorithm involves non-convex sparsity promoting regularization which meanwhile keeps the total cost function strictly convex. The proposed method can efficiently recover group-sparsity components from the wavelet

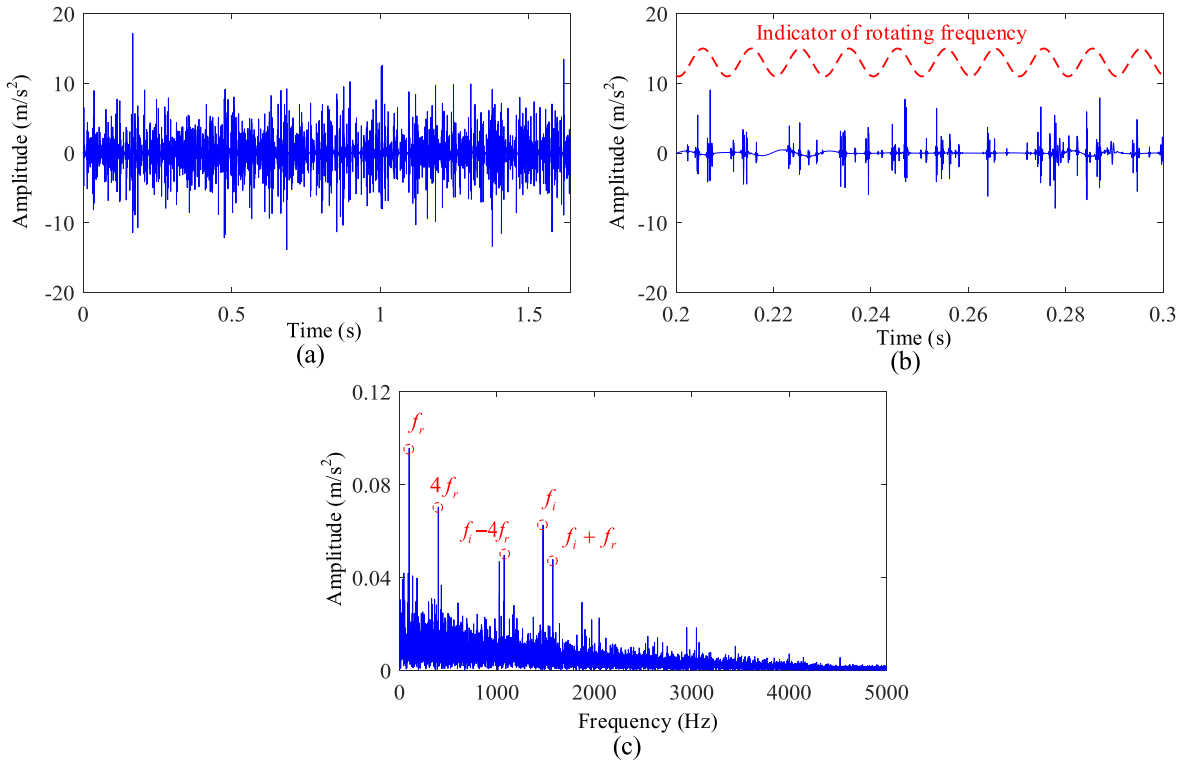

Figure 21. Denoised results via the L1-norm denoising method in case 2: (a) the time-domain waveform, (b) its zoomed-in version and (c) its envelope spectrum.

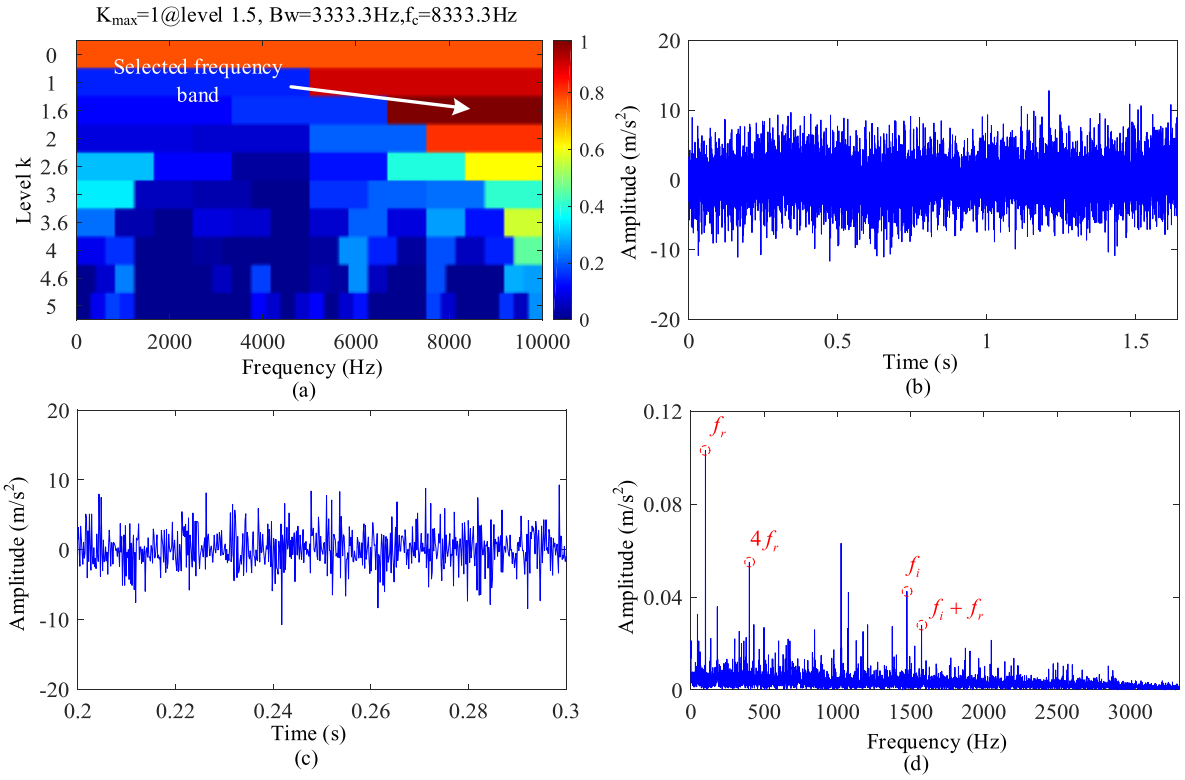

Figure 22. SK results in case 2: (a) the kurtogram, (b) the filtered signal via SK, (c) its zoomed-in version and (d) its envelope spectrum.

coefficients of noisy signals in the fractional spline wavelet domain. In addition, the performance of the proposed method for extracting periodic transient signals has been demonstrated by simulation study and application analyses based on two practical applications of bearing fault diagnosis. For bearing fault diagnosis, the proposed sparsity-assisted denoising method outperforms the state-of-art methods, including the L1-norm denoising method and SK method. The future research will concern how to design dictionary using the fractional spline wavelets and conduct sparse reconstructions for bearing fault signal.

Acknowledgments

This research was supported by the National Natural Science Foundation of China and the Civil Aviation Administration of China jointly funded project (Grant No. U1733107), the Aeronautical Science Foundation of China (Grant No. 20173319003), and the Fundamental Research Funds for the Central Universities of China (Grant No. YJ201662).

ORCID iDs

Xin Zhang https://orcid.org/0000-0003-1138-7637

References

[1] Lei Y and Zuo M J 2009 Fault diagnosis of rotating machinery using an improved HHT based on EEMD and sensitive IMFs *Meas. Sci. Technol.* **20** 125701
[2] Wang Y, Yang L, Xiang J, Yang J and He S 2017 A hybrid approach to fault diagnosis of roller bearings under variable speed conditions *Meas. Sci. Technol.* **28** 125104
[3] Xia P, Xu H, Lei M and Ma Z 2018 An improved stochastic resonance method with arbitrary stable-state matching in underdamped nonlinear systems with a periodic potential for incipient bearing fault diagnosis *Meas. Sci. Technol.* **29** 085002
[4] Fu W, Tan J, Zhang X, Chen T and Wang K 2019 Blind parameter identification of MAR model and mutation hybrid GWO-SCA optimized SVM for fault diagnosis of rotating machinery *Complexity* **2019** 3264969
[5] Wang L, Liu Z, Miao Q and Zhang X 2018 Time–frequency analysis based on ensemble local mean decomposition and fast kurtogram for rotating machinery fault diagnosis *Mech. Syst. Signal Process.* **103** 60–75
[6] Shao H, Jiang H, Zhang X and Niu M 2015 Rolling bearing fault diagnosis using an optimization deep belief network *Meas. Sci. Technol.* **26** 115002
[7] Wang L, Liu Z, Miao Q and Zhang X 2018 Complete ensemble local mean decomposition with adaptive noise and its application to fault diagnosis for rolling bearings *Mech. Syst. Signal Process.* **106** 24–39
[8] Hu B and Li B 2016 A new multiscale noise tuning stochastic resonance for enhanced fault diagnosis in wind turbine drivetrains *Meas. Sci. Technol.* **27** 025017
[9] Wang F, Deng G, Liu C, Su W, Han Q and Li H 2018 A deep feature extraction method for bearing fault diagnosis based on empirical mode decomposition and kernel function *Adv. Mech. Eng.* **10** 1–12
[10] Lei Y, Liu Z, Ouazri J and Lin J 2017 A fault diagnosis method of rolling element bearings based on CEEMDAN *Proc. Inst. Mech. Eng.* C **231** 1804–15
[11] Zhang X, Liu Z, Miao Q and Wang L 2018 An optimized time varying filtering based empirical mode decomposition method with grey wolf optimizer for machinery fault diagnosis *J. Sound Vib.* **418** 55–78
[12] Du X, Wen G, Liu D, Chen X, Zhang Y and Luo J 2019 Fractional iterative variational mode decomposition and its application in fault diagnosis of rotating machinery *Meas. Sci. Technol.* **30** 125009
[13] Zhang S, Wang Y, He S and Jiang Z 2016 Bearing fault diagnosis based on variational mode decomposition and total variation denoising *Meas. Sci. Technol.* **27** 075101
[14] Zhang X, Miao Q, Zhang H and Wang L 2018 A parameter-adaptive VMD method based on grasshopper optimization algorithm to analyze vibration signals from rotating machinery *Mech. Syst. Signal Process.* **108** 58–72
[15] Antoni J and Randall R B 2006 The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines *Mech. Syst. Signal Process.* **20** 308–31
[16] Antoni J 2007 Fast computation of the kurtogram for the detection of transient faults *Mech. Syst. Signal Process.* **21** 108–24
[17] Zhang X, Liu Z, Wang J and Wang J 2019 Time–frequency analysis for bearing fault diagnosis using multiple *Q*-factor Gabor wavelets *ISA Trans.* **87** 225–34
[18] Jiang H, Xia Y and Wang X 2013 Rolling bearing fault detection using an adaptive lifting multiwavelet packet with a 1 1/2 dimension spectrum *Meas. Sci. Technol.* **24** 125002
[19] Zhang X, Wang J, Liu Z and Wang J 2019 Weak feature enhancement in machinery fault diagnosis using empirical wavelet transform and an improved adaptive bistable stochastic resonance *ISA Trans.* **84** 283–95
[20] Liu Z, He Z, Guo W and Tang Z 2016 A hybrid fault diagnosis method based on second generation wavelet de-noising and local mean decomposition for rotating machinery *ISA Trans.* **2016** 211–20
[21] Wang S, Selesnick I, Cai G, Feng Y, Sui X and Chen X 2018 Nonconvex sparse regularization and convex optimization for bearing fault diagnosis *IEEE Trans. Ind. Electron.* **65** 7332–42
[22] Wang X, Zi Y and He Z 2009 Multiwavelet construction via an adaptive symmetric lifting scheme and its applications for rotating machinery fault diagnosis *Meas. Sci. Technol.* **20** 045103
[23] Mallat S G 1989 A theory for multiresolution signal decomposition: the wavelet representation *IEEE Trans. Pattern Anal. Mach. Intell.* **1989** 674–93
[24] Lemarié P G 1988 Ondelettes à localisation exponentielle *J. Math. Pures Appl.* **67** 227–36 (in French)
[25] Battle G 1988 A block spin construction of ondelettes Part II: the QFT connection *Commun. Math. Phys.* **114** 93–102
[26] Lin S T and McFadden P D 1997 Gear vibration analysis by B-spline wavelet-based linear wavelet transform *Mech. Syst. Signal Process.* **11** 603–9
[27] Unser M and Blu T 2000 Fractional splines and wavelets *SIAM Rev.* **42** 43–67
[28] Unser M A and Blu T 1991 Construction of fractional spline wavelet bases *Wavelet Appl. Signal Image Process. VII* **3813** 422–31
[29] Blu T and Unser M 2000 The fractional spline wavelet transform: definition end implementation ICASSP *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*
[30] Katunin A and Przystałka P 2014 Damage assessment in composite plates using fractional wavelet transform of modal shapes with optimized selection of spatial wavelets *Eng. Appl. Artif. Intell.* **30** 73–85
[31] Yang W and Gong Y 2012 Multi-spectral and panchromatic images fusion based on PCA and fractional spline wavelet *Int. J. Remote Sens.* **33** 7060–74

[32] Shen Y, Yang S and Ma B 2006 Application of fractional spline wavelet in detection of abrupt information from fault gear system *Key Eng. Mater.* **324** 371–4
[33] Chen P Y and Selesnick I W 2014 Translation-invariant shrinkage/thresholding of group sparse signals *Signal Process.* **94** 476–89
[34] Deng S W and Han J Q 2018 Adaptive overlapping-group sparse denoising for heart sound signals *Biomed. Signal Process. Control* **40** 49–57
[35] Publishing I, Zhao Z, Lv M, Zhang X, Du J and Zheng M 2014 ECG de-noising based on translation invariant wavelet transform and overlapping group shrinkage *Sens. Transducers* **177** 54–62
[36] He W and Zi Y 2014 Sparsity-assisted signal representation for rotating machinery fault diagnosis using the tunable *Q*-factor wavelet transform with overlapping group shrinkage *Int. Conf. on Wavelet Analysis and Pattern Recognition*
[37] Ding Y and Selesnick I W 2015 Artifact-free wavelet denoising: non-convex sparse regularization, convex optimization *IEEE Signal Process. Lett.* **22** 1364–8
[38] Chen P Y and Selesnick I W 2014 Group-sparse signal denoising: non-convex regularization, convex optimization *IEEE Trans. Signal Process.* **62** 3464–78
[39] Aldroubi A and Unser M 1992 Families of wavelet transforms in connection with shannon's sampling theory and the gabor transform *Wavelets* **1992** 509–28
[40] Figueiredo M A T, Bioucas-Dias J M and Nowak R D 2007 Majorization-minimization algorithms for wavelet-based image restoration *IEEE Trans. Image Process.* **16** 2980–91
[41] Zhang X, Liu Z, Miao Q and Wang L 2018 Bearing fault diagnosis using a whale optimization algorithm-optimized orthogonal matching pursuit with a combined time–frequency atom dictionary *Mech. Syst. Signal Process.* **107** 29–42
[42] Cai G, Chen X and He Z 2013 Sparsity-enabled signal decomposition using tunable *Q*-factor wavelet transform for fault feature extraction of gearbox *Mech. Syst. Signal Process.* **41** 34–3
[43] NUAA 2016 Nanjing University of Aeronautics and Astronautics, research laboratory of intelligent diagnosis and expert system data download http://ides.nuaa.edu.cn (Accessed: 17 May 2018)