**Impulse Feature Extraction of Bearing Faults Based on Convolutive Nonnegative Matrix Factorization**

LIN LIANG\(^{1,2}\), LEI SHAN\(^{1}\), FEI LIU\(^{3}\), MAO LIN LI\(^{3}\), BEN NIU\(^{1}\), AND GUANGHUA XU\(^{1,4}\)

\(^{1}\)School of Mechanical Engineering, Xi’an Jiaotong University, Xi’an 710049, China
\(^{2}\)Key Laboratory of Education Ministry for Modern Design and Rotor-Bearing System, Xi’an Jiaotong University, Xi’an 710049, China
\(^{3}\)Engineering Workshop, Xi’an Jiaotong University, Xi’an 710049, China
\(^{4}\)State Key Laboratory for Manufacturing Systems Engineering, Xi’an Jiaotong University, Xi’an 710054, China

Corresponding author: Fei Liu (fayelau@xjtu.edu.cn)

This work was supported in part by the Natural Science Foundation of Shaanxi Province, China, under Grant 2020JM-050, in part by the National Natural Science Foundation of China under Grant 51575438, in part by the China Postdoctoral Science Foundation under Grant 2017M623159 and Grant 2018T111046, in part by the Postdoctoral Science Foundation of Shaanxi Province under Grant 2017BSHEDZZ68, and in part by the Fundamental Research Funds for the Xi’an Jiaotong University under Grant XJJ2018047.

**ABSTRACT**

The detection of bearing faults is of great significance for the stable operation of rotating machinery. Apart from detection by analysing the vibration response, another powerful strategy is to extract the periodic impulse excitation directly induced by faults, which efficiently eliminates the influence of the transmission path and noise. Typical methods of this strategy include maximum correlated kurtosis deconvolution (MCKD) and multipoint optimal minimum entropy deconvolution (MOMEDA). However, these deconvolution methods based on maximizing a certain measurement index are still insufficient at finding the correct fault period directly because of the interference of noise components. To effectively extract the periodic impulse excitation from the vibration response, a new impulse feature extraction method from the vibration spectrogram based on convex hull convolutive nonnegative matrix factorization (CH-CNMF) is proposed. As the spectrogram intuitively reveals the time and frequency information of the impulse response generated by the fault excitation, according to the decomposition characteristics of CH-CNMF, the time-frequency structure of the impulse response is represented by the basis tensor, while the weight matrix corresponds to the impulse excitation. Meanwhile, autocorrelation is adopted to enhance the periodic impulse excitation. Finally, based on power spectral entropy and first-order correlated kurtosis, the optimal periodic pulse can be selected from the autocorrelation curves of the weight matrix. Both numerical simulation and experimental verifications on bearings indicate that the proposed method can eliminate the influence of random shock excitations and directly attain the periodic impulse for the source of the bearing fault, and that its extraction effectiveness outperforms MOMEDA.

**INDEX TERMS** Feature extraction, time-frequency distribution, nonnegative matrix factorization, deconvolution.

**I. INTRODUCTION**

The rolling bearing is one of the most widely used universal parts in various rotating machines. When there is local damage on the surface of a bearing element, it will strike the surface of other elements during rotary motions, generating an impulse force. This impulse will arouse the high-frequency natural vibration of the components of the bearing system, resulting in a periodic impulse response due to the bearing rotation. Therefore, the period of the impulse excitation corresponds to the frequency of the fault feature. Extracting this period of impulse excitation is a key component of bearing fault diagnosis. Specifically, the impulse of incipient faults is relatively weak and often buried in heavy background noise. Moreover, an abnormal interference component might mislead the classic response band optimization techniques in the wrong direction, ignoring the information related to the actual fault [1]. Methods such as spectrum analysis [2], [3], cepstral analysis [4], [5], envelope analysis [6], [7], and time-frequency analysis [8], [9] usually focus on finding the period...
of the attenuation response, which is often obscured by the transmission path, noise and other factors.

Essentially, the collected vibration signal is a convolution of the impulse excitation and the transfer function, so some scholars employ deconvolution as an effective way to extract the impulse excitation. Most of the existing deconvolution technologies are based on a certain measurement index, taking the maximum index as the optimization target and constructing an inverse filter to eliminate the convolution modulation, thus attaining impulse excitation. Endo et al. [10] combined the minimum entropy deconvolution technique (MED) [11] with an autoregressive model, and proposed the autoregressive minimum entropy deconvolution (ARMED) technique. Sawalhi et al. [3] extended ARMED to the identification of rolling bearing faults in combination with spectral kurtosis. However, the deconvoluted feature information appears as a single large-value impulse or irregular discrete impulse, which is a serious flaw in the judgement of the fault characteristics. Since then, McDonald et al. [12] proposed the maximum correlated kurtosis deconvolution (MCKD) technique to compensate for the shortcomings by absorbing the periodic pattern of the signal after deconvolution. However, the MCKD method is dependent on the period information of the impulse source being estimated in advance, and it often requires resetting the sampling manually according to the sampling process. Inspired by the MED technique, Miao et al. [13] proposed the sparse maximum harmonics-to-noise-ratio deconvolution technique, but this method still requires accurate input parameters to achieve an optimal result. To overcome the shortcomings of these methods, McDonald and Zhao [14] developed the multipoint optimal minimum entropy deconvolution (MOMEDA) technique that does not require the periodic information of the fault source. All of the above deconvolution algorithms, which use the statistical index as the objective function to optimize the solution, often determine an incorrect fault period because of noise interference.

Since the time-frequency distribution (TFD) can simultaneously exhibit the time and frequency distribution of the vibration response, it provides a favourable condition for the deconvolution of the impulse response. Although Wodecki et al. [15] and Liang et al. [16] both applied nonnegative matrix factorization (NMF) on TFD to extract fault features, the decomposed basis matrix does not consider the time-varying characteristic of the signal, and it is difficult to extract the impulse excitation of the fault source. However, in speech signal processing, Smaragdis [17] used a convolution model to further extend the NMF to a convolutive NMF (CNMF) and employed it to extract the local time-frequency basis. O’grady and Pearlmutter [18] also applied CNMF to audio spectra, successfully forming an over-complete basis for the speech signal and achieving supervised monophonic speech separation. Smaragdis and O’grady used the CNMF model to extract the variable frequency components of speech signals based on the deconvolution characteristic of CNMF. By virtue of the deconvolution process of the signal, the CNMF model can integrate the time information into the basis matrix. Thus, when applying CNMF on the TFD of a vibration response, the decomposed time-frequency bases represent the impulse response, and the weight vector is the impulse excitation of the fault source, which corresponds to the time location of each impulse response. Therefore, CNMF provides an effective solution to deconvolute the impulse source signal directly from the observed signal, and the susceptibility of the kurtosis statistical indicator to irregular impulses can be effectively eliminated.

Based on the deconvolution characteristics of CNMF, this paper proposes an impulse feature extraction method for bearing signals with localized faults: after decomposing the TFD of the signal by convex-hull convolutive NMF (CH-CNMF), the obtained basis tensor reflects the time-frequency bases of the impulse response, and the weight matrix contains the location of the corresponding impulse excitation. Auto-correlation processing is also implemented on the weight matrix to enhance the periodic components within the signal. Considering the periodicity of the impulse excitation caused by failures, the optimal impulse excitation can be selected by combining the power spectral entropy (PSE) and the first-order correlated kurtosis ($cK_1$). The major innovation of this paper is the proposal of a new deconvolution strategy that directly extracts the time-frequency structure of the impulse from the TFD, without constructing an inverse filter for deconvolution based on a single index. This method fundamentally addresses the defect that a filter based on a single index is susceptible to interference, which enables obtaining more accurate results.

The rest of this paper is organized as follows. Section 2 gives a brief description of NMF and the characteristics of performing NMF on TFD, and then describes the CNMF algorithm and outlines the characteristics of applying CH-CNMF to deconvolute the impulse excitation. Section 3 introduces the basic principle and procedure of the proposed method. Section 4 verifies the effectiveness using a simulated signal and actual fault bearing data. Conclusions are drawn in Section 5.

II. PRINCIPLE OF NMF AND CNMF

A. NMF AND DECOMPOSITION ON TFD

The conventional basic nonnegative matrix factorization [19] decomposes a given nonnegative matrix $V = [v_1 \ v_2 \ \cdots \ v_N] \in R^{m \times N}$ into the product of two nonnegative matrices $W = [w_1 \ w_2 \ \cdots \ w_R] \in R^{N \times r}$ and $H = [h_1 \ h_2 \ \cdots \ h_r]^T \in R^{r \times m}$, such that the decomposition is represented as:

$$V_{n \times m} \approx W_{n \times r}H_{r \times m}$$

$$s.t. \ W \geq 0, \ H \geq 0,$$ (1)

where $W$ and $H$ are called the basis and weight, respectively, and $r$ is the low rank of the embedding space. Normally, $(m + n)r < mn$. 

L. Liang et al.: Impulse Feature Extraction of Bearing Faults Based on CNMF
The traditional NMF algorithm has the shortcoming of slow convergence speed [20]. To improve the performance of NMF, Ding et al. [21] proposed the Convex NMF algorithm. The \( W \) in Convex NMF is obtained by the linear combination of the original matrix \( V \). Set \( G \) as the combination coefficient, then:

\[
V = VGH, \tag{2}
\]

where \( \sum g_i = 1 \) and \( g_i \geq 0 \). Compared with traditional NMF, Convex NMF provides more practical results and is more conducive to finding a sparse solution [21].

Because Convex NMF does not consider the influence of the weight matrix \( H \), Thurau et al. [22] imposed another constraint on this basis (named Convex-Hull NMF), and Equation (2) can be rewritten by:

\[
V = VGH, \tag{3}
\]

If \( X = VG \) represents all vertices on a convex hull \( C(V) \), there must be such an \( H \) such that \( V = XH \), suggesting that the cost function \( J = \| V - XH \|^2 \) can be 0. As the amount of data must be much larger than the number of \( C(V) \) vertices, fewer convex hull vertices can achieve the optimization goal. Therefore, not only in the convergence speed but also in the result representation, the result of CH-NMF is superior to that of Convex NMF and traditional NMF [22].

As a traditional nonnegative matrix, the time-frequency distribution can be extracted by a short-time Fourier transform (STFT)

\[
STFT(t, f) = \sum_{k=0}^{N-1} x(t)w(t - k)e^{-j2\pi k/N}, \tag{4}
\]

where \( w(t - \tau) \) is a shifted window. Here, the time-frequency distribution can be represented as \( V = |STFT(t, f)| \). When NMF is applied to \( V \), each basis vector \( v_i \) in \( W \) can be regarded as a cluster in the frequency domain, and each vector \( h_i \) in \( H \) is a cluster in the time domain. Thus, the representation of \( V \) with \( w_i \) and \( h_i \) can be given as [23]:

\[
V = w_i h_1 + w_2 h_2 + \cdots + w_m h_r + E, \tag{5}
\]

where \( E \) represents error or noise. Liang et al. [16] used two intermittent sinusoidal signals of different frequencies to illustrate the clustering and correspondence. The decomposition is shown in Fig.1. Both \( w_1 \) and \( h_1 \) describe the signal of frequency \( f_1 \), and both \( w_2 \) and \( h_2 \) describe the signal of frequency \( f_1 \). The peak value of \( w_1 \) corresponds to the frequency of 2200 Hz, and \( w_2 \) peaks at the frequency of 1000 Hz. As the distribution of the signal in the time domain, \( h_1 \) corresponds to the time envelope of the frequency \( f_2 \) while \( h_2 \) corresponds to the time envelope of the frequency \( f_1 \).

Obviously, the frequency and time distribution of the vibration response in the TFD can be extracted by NMF. For the obtained basis matrix, because the basis vector only reflects the frequency and has no time structure, the variable frequency or the vibration response distribution of the time-varying signal is not reflected.

**FIGURE 1.** NMF on the spectrogram of the simulated signal \((r = 2)\). The simulated signal is mixed with two noncontinuous sinusoidal components of \( f_1 = 1000 \) Hz and \( f_2 = 2300 \) Hz, where the component of \( f_1 \) contains three nonzero segments and the component of \( f_2 \) contains four nonzero segments.

### B. CNMF AND EXTRACTION OF TIME-FREQUENCY BASIS

To absorb the time information into the spectral basis matrix, it is necessary to consider the potential correlation among adjacent columns of the input time-frequency matrix \( V \). Therefore, Smaragdis [17] employs the convolution model to extend the basic NMF to CNMF, and decomposes the input matrix into the sum of the continuous basis matrix sequence \( W_i \) and the corresponding coefficient matrix \( H \):

\[
V \approx \sum_{i=0}^{T-1} W_i H, \tag{6}
\]

where \( T \) is the length of the shifted sequence, \( i \rightarrow (\bullet) \) is an operator that shifts the column to the right by \( i \) bits, and the left shift portion is padded with zeros.

Based on CNMF, Vaz et al. [24] proposed the Convex-Hull CNMF method (CH-CNMF), which attempts to find \( K \) time components in \( V \) at any time \( t \), and the decomposition can be defined by

\[
V \approx S \sum_{i=0}^{T-1} G(t) H, \tag{7}
\]

where \( S \in \mathbb{R}^{m \times p} \) are \( p \) vertices of \( C(V) \), and \( G \in \mathbb{R}^{T \times K} \) is the combination coefficient. Each time series \( W(t) = SG(t) \in \mathbb{R}^{m \times K} \) is the basis matrix obtained by the decomposition of \( V \), and \( H \in \mathbb{R}^{K \times n} \) is the weight matrix.

To improve the performance of CH-CNMF, the sparse constraint is added to the weight matrix \( H \), and the following
cost function is constructed by:

$$f (\hat{G}, \hat{H}) = \min \left\| V - S \sum_{t=0}^{T-1} G(t) H^T \right\|_F^2 + \lambda \|H\|_1$$

s.t. \(\|g_x(t)\|_1 = 1, \ G(t) \geq 0, \ H \geq 0, \) (8)

where \(\lambda\) is the sparsity-regularization parameter. The sparsity constraint improves the interpretability of the basis vector because the weight vector is closer to the sparse impulse signal. By updating \(G\) and \(H\) alternately by iteration until they reach the predetermined convergence result or iteration steps, the update rules are as follows:

$$G(t) \leftarrow G(t) \odot \frac{([S^T V]^+ + [S^T S]^T - F) H^T}{([S^T V]^T + [S^T S]^T + F) H^T},$$

$$H \leftarrow H \odot \frac{\sum_{t=1}^{T-1} G(t) ([S^T V]^+ I_n + [S^T S]^T - F)}{\sum_{t=0}^{T-1} G(t) ([S^T V]^T - I_n + [S^T S]^T + F) + \lambda}$$

(9) (10)

where \(F = \sum_{t=0}^{T-1} G(t) H^T, t \in \{0, \cdots, T - 1\}, I_n\) is the identity matrix of \(n \times n,\) and \([A]^+ = 0.5(|A| + A)\) and \([A]^− = 0.5(|A| - A)\) represent the positive and negative components of matrix \(A,\) respectively.

Inconsistent with NMF, the weight matrix extracted by CH-CNMF corresponds to the onset time of each variable frequency component, which indicates that CH-CNMF is a deconvolution process. Moreover, although \(w_1\) contains an interference component, there is no clear performance in \(h_1,\) which implies that the individual interference signals have less impact on the weight matrix structure.

### C. DECONVOLUTION OF VIBRATION RESPONSE BY CNMF

When a local fault occurs in the rolling bearing, periodic impulse excitation \(h(t)\) will occur under the rotation of the bearing [25]. These excitations are convoluted by the response of the bearing system and are reflected in the vibration signal as a periodic attenuation waveform. Therefore, the impulse response \(x(t)\) can be expressed as:

$$x(t) = f(t) * h(t),$$

(11)

where \(f(t)\) represents the transfer function. Usually, \(f(t) = e^{-\xi t}A \cos(2\pi f_r t + \phi),\) where \(\xi\) is the attenuation factor, \(A\) is the initial amplitude, \(\phi\) is the initial phase, and \(f_r\) denotes the natural frequency. The response mechanism of the rolling bearing can be expressed in a graphical form as shown in Fig.3.

Then the observed signal \(y(t)\) can be expressed as:

$$y(t) = x(t) + e(t) = f(t) * h(t) + e(t),$$

(12)

where \(e(t)\) denotes noise and other interference components. For the signal in Fig.3, the carrier frequency of the periodic attenuation response corresponds to the natural frequency of the bearing system, and the periodicity of the impulse corresponds to the frequency associated with the fault. Thus, the deconvolution of the vibration response is efficacious in extracting the impulse excitation \(h(t)\).

When we apply CH-CNMF to the periodic vibration response signal (as shown in Fig.4), the decomposed \(w_1\) is the time-frequency distribution of a single impulse response, and the weight vector \(h_1\) is the impulse excitation \(h(t),\) which corresponds to the time location of each impulse. These results suggest that CH-CNMF provides an effective method to deconvolute the impulse source \(h(t)\) directly from the observed signal \(y(t).\)

### III. IMPULSE FEATURE EXTRACTION WITH DECONVOLUTION

Suppose that there is a time domain signal:

$$y(t) = x(t) + e(t) = f(t) * \sum_{k=1}^{N} A_k \delta(t - kT_0) + e(t),$$

(13)

where \(A_k\) is the amplitude of the impulse, \(\delta(t)\) is the Dirac delta function, and \(T_0\) is the period of the impulse. The Fourier transform of \(y(t)\) is:

$$Y(f) = F(f) \sum_{k=1}^{N} A_k e^{2\pi k T_0 f} + E(f),$$

(14)

To illustrate the decomposition result of CH-CNMF, we select the periodic swept-frequency cosine (chirp) signal for analysis, and the input signal is a combination of two periodic chirp signals. The lower frequency signal increases from 100 Hz to 150 Hz, the higher frequency component decreases from 400 Hz to 300 Hz, and the interference component is added at 1.07 s. With \(K = 2\) and \(T = 0.2\) s, the decomposition result is presented in Fig.2. It can be seen that the basis \(W\) extracted by CH-CNMF can effectively reflect the local time-frequency basis in the non-stationary signal.

**FIGURE 2. CH-CNMF on the spectrogram of the input signal.**

The local time-frequency basis in the non-stationary signal.

Inconsistent with NMF, the weight matrix extracted by CH-CNMF corresponds to the onset time of each variable frequency component, which indicates that CH-CNMF is a deconvolution process. Moreover, although \(w_1\) contains an interference component, there is no clear performance in \(h_1,\) which implies that the individual interference signals have less impact on the weight matrix structure.
where $Y(f)$, $F(f)$, and $E(f)$ are the Fourier transforms of $y(t)$, $f(t)$, and $e(t)$, respectively. To extract this impulse signal, the following main steps are proposed.

**A. DECOMPOSITION OF VIBRATION RESPONSE**

With the determined decomposition dimension $K$ and time $T$, the time-frequency distribution $V = |STFT_y(t, f)|$ is decomposed by CH-CNMF. The parameter $K$, a common empirical value, has not been strictly defined in previous studies. However, to completely separate the fault signal, we introduce a limitation, $a \leq K \leq a + 3$, to reduce overlap among resonance bands, where $a$ is the number of frequency resonance regions with larger amplitude or known frequency components. Appropriately increasing the value of $K$ can separate the frequency components and avoid the aliasing of resonance peaks. It is difficult to extract the response signal with too small a value of $T$, and too large a value increases the spatial complexity. The time $T$ should be guaranteed to contain at least most of the impulse attenuation time $t_s$. According to the impulse response mechanism of the rolling bearing, if the time taken for the response amplitude to reach $\varepsilon$\% of the initial amplitude is $t_s$, then there will be:

$$
\eta = \frac{Ae^{-\zeta r T_0}}{Ae^{-\zeta r (T_0 + t_s)}} = e^{-\zeta r t_s} = \varepsilon \%
$$

(15)

From Equation (15), we can obtain $t_s = -\frac{\ln \varepsilon \%}{\zeta r}$. However, $\zeta r$ is difficult to determine or even unknown for more complex signals, so it is difficult to use $t_s$ to determine $T$. In fact, as long as $T \geq t_s$, the decomposition time dimension already contains at least one impulse attenuation response (as shown in Fig.5), and it is possible to extract the fault information. Although it is difficult to determine $t_s$ directly, it is feasible to predict $T_0$, and $T_0$ must contain an impulse attenuation response, so $T$ is set to be near $T_0$ in this paper.

**B. ENHANCEMENT OF IMPULSE FEATURE**

By virtue of the decomposition, the basis tensor $W$ encompasses matrices $w_i$ that contain the time-frequency basis of the impulse response $f(t)$. The corresponding weight vector $h_i$ will also have the impulse excitation $\sum_{k=1}^{N} A_k \delta (t - kT_0)$. 

---

**FIGURE 3.** The impulse response schematic diagram, where $T_0$ represents the period of the impulse, reflecting the fault feature frequency.

**FIGURE 4.** (a) The waveform of the simulated signal; (b) CH-CNMF on the time-frequency distribution.

**FIGURE 5.** The vibration response signal.
The weight vector $h_i$ can be expressed as:

$$h_i(t) = \sum_{k=1}^{N} B_k \delta(t - kT_0) + n(t),$$  \hspace{1cm} (16)

where $B_k$ represents the amplitude of the impulse, $B_k \neq A_k$, and $n(t)$ represents noise.

To diminish noise and make $h_i(t)$ closer to $h(t)$, bandpass filtering with an approximate passband ranging from $2f_r$ to $2f_h$ is adopted to remove the lower frequency component, especially the rotation frequency, and the high-frequency interference, where $f_r$ is the rotational frequency of the shaft and $f_h$ is the fault characteristic frequency. In case the fault type is unknown, it is recommended to set $f_h$ to the maximum characteristic value of the inner ring fault to contain more information of various fault types. Thus, the filtered signal is:

$$h_{if}(t) = \sum_{k=1}^{N} C_k \delta(t - kT_0) + n(t).$$  \hspace{1cm} (17)

The autocorrelation has the benefit of removing the uncorrelated components of the signal, i.e. noise and random impulsive contents, that are unrelated to any specific bearing fault [26]. Furthermore, the periodic excitation in $h$ is enhanced and higher periodic peaks in the autocorrelation suggest more pronounced periodicity of the signal.

**C. OPTIMAL SELECTION OF IMPULSE FEATURE**

Since impulse excitation usually generates vibrations with multiple natural frequencies, the decomposition dimension generally satisfies $K > a$ to ensure complete decomposition, which implies that more than one of the $K$ autocorrelation curves contains the impulse source information. Therefore, it is necessary to determine the optimal autocorrelation curve that best reflects the fault impulse signal. The optimal objective is to select the curve with the simplest structure and the most prominent impulse period $T_0$.

First, the power spectral entropy (PSE) is used to evaluate the spectral structure of the signal. For the autocorrelation

$$R_i^f(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_{0}^{T} h_i^f(t)h_i^f(t \pm \tau) dt.$$  \hspace{1cm} (18)

The autocorrelation has the benefit of removing the uncorrelated components of the signal, i.e. noise and random impulsive contents, that are unrelated to any specific bearing fault [26]. Furthermore, the periodic excitation in $h$ is enhanced and higher periodic peaks in the autocorrelation suggest more pronounced periodicity of the signal.
curve $R_h^i(\tau)$ of each weight vector, its Fourier transform is

$$H(f) = \text{FT}(R_h^i(\tau)).$$  \hspace{1cm} (19)

Then its power spectrum is

$$S(f) = \frac{1}{N} |H(f)|^2.$$ \hspace{1cm} (20)

$S = \{S_1, S_2, \cdots S_p\}$ is used as the division of the power spectral energy. The information entropy of the response, i.e., the power spectral entropy is defined as:

$$E_f = -\sum_{i=1}^{p} q_i \ln q_i,$$ \hspace{1cm} (21)

where $q_i = \frac{s_i}{\text{sum}(S)}$ represents the percentage of the power spectrum of the i-th segment in the entire spectrum. Obviously, the power spectral entropy index characterizes the spectral structure of the autocorrelation. The more uniform the energy distribution in the whole frequency component is, the more complex the signal is, the greater the uncertainty is, and the larger the power spectral entropy is [27].

Second, the correlated spectral kurtosis is adopted to measure the impulse feature. In previous studies, spectral kurtosis was usually used to measure the impact signal. However, this kurtosis guidance strategy is susceptible to strong non-Gaussian noise and tends to highlight outliers that are not related to real faults. To improve this indicator, McDonald et al. [12] proposed the correlated kurtosis indicator. The correlation spectrum kurtosis of the first-order shift ($CK_1$) is

$$CK_1(T) = \left(\frac{\sum_{i=1}^{L} \left(y_i y_{i-T}\right)^2}{\left(\sum_{i=1}^{L} y_i^2\right)^2}\right)^\frac{1}{2},$$ \hspace{1cm} (22)

and the correlation spectrum kurtosis of the $M$-order shift ($CK_M$) is

$$CK_M(T) = \left(\frac{\sum_{i=1}^{L} \left(\sum_{m=0}^{M} y_{i-mT}\right)^2}{\left(\sum_{i=1}^{L} y_i^2\right)^{M+1}}\right)^\frac{1}{2},$$ \hspace{1cm} (23)

where $L$ is the length of the acquired signal sequence, $y_l = 0$ if $l \neq 1, 2, \cdots, L$, and $T$ is the estimated fault characteristic frequency that has been set when performing CH-CNMF. $CK_1$ characterizes the maximum value of periodic impulse over a particular period. Compared with other indexes used in fault diagnosis such as the Gini index [28] and kurtosis [13], $CK_1$ takes into account the well-known periodic characteristics of fault signals. The higher shift $CK$ indicates a larger sequence of continuous periodic pulses.

Considering the complexity of the signal, we expect that the ideal periodic impulse excitation should be as simple as possible and contain obvious periodicity. However, there is no index that can take both characteristics into account in the current research, so we combine PSE with $CK_1(T)$ to make a judgement. The smaller spectral entropy indicates that the spectral structure of the signal is simpler, and the larger $CK_1(T)$ indicates that the signal contains more impulse responses with the period $T$. Therefore, three weight vectors $\{h_a, h_b, h_c\}$ with the smaller PSE are chosen, and then the weight vector $h_{opt}$ with the largest $CK_1(T)$ among them is taken as the optimal object:

$$h_a, h_b, h_c = \min \text{PSE}(\{h_1, h_2, \cdots, h_K\}),$$

$$h_{opt} = \max CK_1(\{h_a, h_b, h_c\}).$$ \hspace{1cm} (24)

D. ALGORITHM

Based on the main steps above, we propose an impulse extraction method (CNMFD) for bearing fault diagnosis with deconvolution characteristics via CH-CNMF. The procedure of the algorithm is as follows:

**Algorithm CNMFD**

Input: signal $y$, decomposition dimension $K$, time $T$

Step1: Perform STFT on the input signal to obtain the time-frequency distribution $V$

Step2: Determine the decomposition dimension $K$ and time $T$, and perform CH-CNMF on $V$

Step3: Perform bandpass filtering on each weight vector

Step4: Perform autocorrelation on each filtered weight vector

Step5: Combine the PSE and the $CK_1(T)$ to select the optimal analysis object $h_i$

Step6: Perform spectrum analysis on the autocorrelation curve $R_h^i(\tau)$ to obtain the spectrum $H(f)$

Step7: Find the characteristic frequency and its multiplier, and determine the location of the fault

Step8: Determine whether $H(f)$ is clean. If there are many interference components, skip to Step 4 and reduce the passband range according to the obtained characteristic frequency in Step 7

IV. EXPERIMENTAL VERIFICATION

A. NUMERICAL EXPERIMENT

According to the vibration response mechanism of a rolling bearing, the following rolling bearing vibration model is given as [29]:

$$x(t) = \sum_{i=1}^{M_1} f_j(t - iT_{a}) + \sum_{j=1}^{M_2} f_j(t - jT_{d}) + \sum_{k=1}^{M_3} P_k \sin(2\pi f_k t + \beta_k) + n(t).$$ \hspace{1cm} (25)

where $T_{a}$ is the cycle of the failure pulses, $T_{a} = 1/f_a$, and $f_a$ is the fault characteristic frequency. $T_{d}$ describes the time interval between external impulses. Because of its unpredictable
nature, it is a random variable. $P_k, f_k, \beta_k$ correspond to the amplitude, frequency and phase of the discrete harmonics generated from the shaft, gear, etc.

Assuming that a local fault occurs in the outer ring of the bearing, the formant is at the natural frequency of 2200 Hz, and the modulation frequency $f_h$ is 80 Hz. In addition, three nonperiodic impulse attenuation components are included. The carrier frequencies are 1000 Hz, 2600 Hz and 3000 Hz, the fractional bandwidths are 0.1, 0.2, 0.3, and the phase angle is set to 0. Finally, Gaussian white noise with a signal-to-noise ratio (SNR) $= 0$ is added, and the vibration response signal is sampled at 10 240 Hz. The model of the simulated signal is described in Fig.7 and its waveform and spectrum are shown in Fig.8. Regardless of the waveform or the spectrum, it is difficult to directly observe the impulse response and the three natural frequency regions due to strong background noise.

As four interference frequency bands are found, the decomposition dimension $K$ is set to 5, and $T$ is set to $1/f_h = 12.5$ ms to cover one impulse response interval. The decomposition result of the TFD is shown in Fig.9, and Fig.10 gives the PSE and $CK_1(T)$ of each weight vector. It can be seen that the weight vectors with smaller PSE are $h_1, h_2$ and $h_5$. However, the dominant part of $w_5$ corresponding to $h_5$ corresponds to the random impulse at 1000 Hz and contains fewer frequency components at 2200 Hz in Fig.9. The vector with the maximum $CK_1(T)$ is $h_2$ among the three weight vectors with smaller PSE, and the $CK_1(T)$ of $h_1$ is very small. Therefore, we choose $h_2$ for subsequent analysis and list the result of $h_5$ for contrast.

According to the criterion that the filter band should be set to approximately $2f_r - 2f_h$, we round it to approximately 60 - 200 Hz to facilitate the actual filter setting. The autocorrelation curve of the filtered weight vector and its spectrum are given in Fig.11. For $h_2$, the adjacent peak interval is 12.5 ms in the autocorrelation curve, which clearly shows the location of the fault impulse, and its spectrum shows the obvious fault frequency $f_h$. In comparison, although $h_5$ captures the frequency of fault impulses in the waveform and spectrum, it has a smaller amplitude and more interference components than $h_2$ because the main frequency of $h_5$ points to aperiodic pulses at 1000 Hz.

Using the NMF method proposed by Liang et al. [16] to decompose the spectrogram, the results obtained are shown in Fig.12. The basis vector $w_5$ strikes at the fault carrier frequency, that is, $h_5$ corresponds to the envelope of the fault signal in the time domain. Fig.13 shows the time-domain signals extracted by CH-CNMF and NMF respectively.
Results extracted by these two methods are basically consistent with the position of the theoretical pulse. Due to the difference in the solution goal, the relative magnitudes of the two solutions are not exactly the same. For example, the impulse can be clearly observed by CH-CNMF at 0.038 s, but it is difficult to observe the existence of the impact envelope by NMF. Meanwhile, the envelope wave has difficulty directly observing the location of impulses due to noise interference. In comparison, the impulse features extracted by CH-CNMF are simpler and more intuitive. However, there are missing amplitudes or interference amplitudes in some places that are greatly affected by noise, which is the main reason for further optimization by autocorrelation analysis.

For comparison, the impulse source extracted using the MOMEDA algorithm is shown in Fig.14. The curve of multipoint kurtosis with samples is shown in Fig.14(a), and the filtered signal has the largest spectral kurtosis when samples = 2. The result extracted by samples = 2 is shown in Fig.14(b), (c). Due to the influence of interference noise, the maximum kurtosis does not point to the periodic pulse, and the obtained result does not reflect the simulated fault.

In fact, according to the fault frequency, the fault information should be best reflected by setting samples = 1/f_h * f_s = 128, and the extracted result is shown in Fig.15(a), (b).
Compared with the extracted impulse in Figs. 11(a) and (b), although the fault frequency is already known, the extracted impulse source signal by MOMEDA still contains more noise and interference components.

B. OUTER RACE FAULTS

The bearing life test data [30] provided by the Intelligent Maintenance System Center of Cincinnati University are used to verify the method, and illustrate the effect of the proposed method. As shown in Fig.16, four bearings are mounted on the same shaft and the two bearings in the middle are subjected to a radial pressure of 27 kN. The vibration data are acquired by an acceleration sensor at a sampling frequency of 20,000 Hz every 10 minutes. Here we choose the data of outer ring bearing failure on February 16, 2004. According to the bearing parameters, when the rotational frequency \( f_r = 33.3 \) Hz, the outer ring fault characteristic frequency \( f_o = 236.4 \) Hz.

The waveform and spectrum of the data are shown in Fig.17. Since the bearing is in the early stage of the fault, there is no periodic impulse in the time domain signal. In the spectrum shown in Fig. 17(b), not only are the low frequency components complex, there are also multi-resonance bands, and the main resonance band frequency is distributed between 3000 Hz and 5000 Hz.

Since each resonance band can correspond to the characteristics of the impulse response, it is not necessary to find out all the resonance bands. In Fig.17(b), the spectrum contains
four main resonance bands, so $K$ is set to 5 to ensure effective separation. $T$ is approximately $1/f_o = 4.2$ ms (approximately 85 points), so we set $T = 85f_s = 4.25$ ms to decompose the time-frequency distribution of the signal by CH-CNMF.

Fig. 18 and 19 show the decomposed result and the PSE and $CK_1$ curve of each weight vector, respectively. In Fig.19, $h_1$, $h_2$, and $h_3$ have smaller power spectral entropy, that is, they have simple spectral structures. Among them, $h_2$ has the largest $CK_1$, so we select $h_2$ as the analysis object.

In Fig. 20, we compare the spectrograms, $w_2$, $h_2$, and $h_4$. It can be seen that $w_2$ is prominent with an impulse response around 4480 Hz. Most peaks of $h_2$ reflect the location of each impulse corresponding to $w_2$ in the time domain. $h_4$ also has
many peaks corresponding to this frequency range. Due to
the complexity of the actual system structure, the shape of the
defect, and noise, the actual impulse response waveform may
have a certain change, which is difficult to keep consistent
with the theoretical shape in Fig.4. Therefore, the impulse
response corresponding to the same time-frequency basis will
be decomposed into the same $h_i$, and individual changed
impulse responses will be decomposed into other weight vec-
tors, for which the corresponding impulse excitation position
in $h_i$ may be missing, and the positions can be restored by the
correlation function.

To eliminate the frequency modulation and other compo-
nent interference, the bandwidth of the bandpass filter is from
$2f_r$ to $2f_o$, and the parameters of the filter are adjusted to in-
teger for simple calculation. Thus, since $2f_r$ and $2f_o$ are equal
to 66.6 Hz and 472.8 Hz respectively, the cutoff frequencies
of pass-band filter are set to 100 Hz and 500 Hz. Finally,
the autocorrelation curve and its spectrum corresponding to
$h_2$ are shown in Fig.21(a), (b). The autocorrelation curve has
no other interference components, and the interval between
adjacent peaks is approximately 4.2 ms, corresponding to a
fault frequency of 236.4 Hz. The frequency component of
fault impulses can be observed directly in the spectrum of the
autocorrelation curve $R_2^2(\tau)$.

In addition, the MOMEDA algorithm is adopted to extract
the fault impulse and the result is shown in Fig. 22. For
Fig.22(a), the filtered signal has the largest spectral kurtosis
when samples = 12, which is highly likely to be affected by
interference noise. The result also shows that the filtered sig-
nal is independent of the fault characteristics. Considering the
frequency of the fault feature, when samples = $1/f_o \times f_s = 85$,
the MOMEDA algorithm will obtain the best result which is
shown in Fig.22(c). However, if the parameter is set manually,
the value of blind deconvolution is lost. Meanwhile, there are more noise and interference components in the waveform. The result shows that the proposed method CNMFD can find the correct fault cycle, and the extracted impulse signal is cleaner, which has better deconvolution ability than that of MOMEDA.

C. INNER RACE FAULTS
The other experiment is carried out on the rolling bearing test bench. The test bench is mainly composed of a governor, a DC motor, a support box and a radial loading device. The basic structure is shown in Fig.23. A certain radial load is applied to the 6308 rolling bearing, and the accelerometer is installed in the vertical direction of the motor drive end bearing housing. In the experiment, the bearing with an inner ring scratch is used for analysis. The data are collected by the OR35 data acquisition system, and the sampling frequency is 12 800 Hz. When the rotation frequency is set to

FIGURE 22. (a) Kurtosis curve with sampling period; (b) impulse waveform at samples = 12 automatically; (c) impulse waveform at samples = 85 artificially.

FIGURE 23. The rolling bearing test bench.

FIGURE 24. Inner ring fault bearing: (a) waveform of the fault signal; (b) spectrum of the fault signal.
1050 r/min ($f_r = 17.5$ Hz), the fault characteristic frequency $f_i$ is 86.24 Hz accordingly. The waveform and spectrum of the acquired signal are shown in Fig. 24. Due to the minor fault, it is difficult to observe the periodic impulse response directly from the waveform. The spectrum contains a large number of resonance peaks, and it is difficult to find the existence of the fault.

As the spectrum contains approximately four main frequency peaks, $K$ is set to 5 to ensure complete decomposition. According to the fault frequency, $T = 1/f_i = 11.6$ ms is taken to obtain at least one time-frequency structure. The decomposition result of TFD is shown in Fig. 25. To measure the spectral structure and correlated kurtosis, the PSE and $CK_1$ of each weight vector are calculated, which are presented in Fig. 26. In Fig. 26, $h_1$, $h_2$, and $h_3$ have smaller power spectral entropy and $h_2$ has the largest $CK_1$ among the three, which means that $h_2$ may be the closest to the fault impulse signal. Moreover, $h_2$ is analysed between 50 Hz and 180 Hz to eliminate the frequency modulation and other interference, and the autocorrelation of the filter signal and its spectrum are shown in Figs. 27(a) and (b). Although most of the adjacent peak intervals are 11.7 ms in the autocorrelation curve, which is consistent with the fault frequency, due to the influence of the interference component, the extracted impulse contains some irrelevant peaks. At the same time, the spectrum in Fig. 27(b) contains some interference peaks, and it does not affect the judgement of the fault frequency.

The impulse source signal extracted by MOMEDA is also given in Fig. 28 for comparison. If we attain results automatically from the original algorithm, the filtered signal has the largest spectral kurtosis when samples $= 2$, which is most likely affected by the interference noise. Based on
the actual fault frequency, if we artificially set samples = \(1/f_o \times f_s = 148\) by the fault characteristic frequency, this should achieve the best result. Although the correct result can be obtained by setting samples = 148, the extracted result contains a large amount of noise in Fig.28(c). The impulse signal extracted by accurate sampling length is not as clean as the result of the proposed method in the waveform, reflecting that the proposed method has a better deconvolution effect.

V. CONCLUSION

Directly extracting an impulse excitation from a vibration signal is an efficacious approach to bearing fault diagnosis. Existing methods, e.g., MED, MCKD, and MOMEDA. MCKD requires priori knowledge of the fault period, which considerably limits fault diagnosis. MOMEDA takes the kurtosis indicator to construct the filter and tends to find the wrong period for complex signals due to the existence of interference components and noise. Based on the time-frequency distribution of the impulse response signal, this paper proposes a deconvolution method for extracting impulse excitation using Convolution NMF. Time-frequency distributions can effectively characterize the localized features of the impulse response, including the resonance frequency band and the attenuation response. Convolution NMF is employed to decompose the impulse response. Meanwhile, an autocorrelation function is adopted to further enhance the impulse feature in the weight vectors. Compared with MOMEDA, the simulation results and practical experiments verify that the proposed method circumvents the incorrect orientation of the kurtosis indicator and successfully extracts pure and accurate impulse features.

REFERENCES

[1] Y. Miao, M. Zhao, and J. Lin, “Identification of mechanical compound-fault based on the improved parameter-adaptive variational mode decomposition,” ISA Trans., vol. 84, pp. 82–95, Jan. 2019. doi: 10.1016/j.isatra.2018.10.008.

[2] S. Ericsson, N. Grip, E. Johansson, L.-E. Persson, R. Sjöberg, and J.-O. Strömberg, “Towards automatic detection of local bearing defects in rotating machines,” Mech. Syst. Signal Process., vol. 19, no. 3, pp. 509–535, May 2005. doi: 10.1016/j.ymssp.2003.12.004.

[3] N. Sawalhi, R. B. Randall, and H. Endo, “The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis,” Mech. Syst. Signal Process., vol. 21, no. 6, pp. 2616–2633, Aug. 2007. doi: 10.1016/j.ymssp.2006.12.002.

[4] C.-S. Park, Y.-C. Choi, and Y.-H. Kim, “Early fault detection in automotive ball bearings using the minimum variance cepstrum,” Mech. Syst. Signal Process., vol. 38, no. 2, pp. 534–548, Jul. 2013. doi: 10.1016/j.ymssp.2013.02.017.

[5] Y.-C. Choi and Y.-H. Kim, “Fault detection in a ball bearing system using minimum variance cepstrum,” Meas. Sci. Technol., vol. 18, no. 5, pp. 1433–1440, May 2007. doi: 10.1088/0957-0233/18/5/031.

[6] Y. Ming, J. Chen, and G. Dong, “Weak fault feature extraction of rolling bearing based on cyclic Wiener filter and envelope spectrum,” Mech. Syst. Signal Process., vol. 25, no. 5, pp. 1773–1785, Jul. 2011. doi: 10.1016/j.ymssp.2010.12.002.

[7] Y.-H. Kim, W.-S. Cheung, and Y. K. Kwak, “Complex envelope of vibration signature of ball bearing system and its application to fault detection,” in Proc. 4th Int. Congr. Exp. Mech., 1988, pp. 1067–1071.

[8] Z. Feng, M. Liang, and F. Chu, “Recent advances in time–frequency analysis methods for machinery fault diagnosis: A review with application examples,” Mech. Syst. Signal Process., vol. 38, pp. 165–205, Jul. 2013. doi: 10.1016/j.ymssp.2013.01.017.

[9] Z. K. Peng and F. L. Chu, “Application of the wavelet transform in machine condition monitoring and fault diagnostics: A review with bibliography,” Mech. Syst. Signal Process., vol. 18, pp. 199–221, Mar. 2004. doi: 10.1016/S0888-3270(03)00075-X.

[10] H. Endo and R. B. Randall, “Enhancement of autoregressive model based gear tooth fault detection technique by the use of minimum entropy deconvolution filter,” Mech. Syst. Signal Process., vol. 21, no. 2, pp. 906–919, Feb. 2007. doi: 10.1016/j.ymssp.2006.02.002.

[11] R. A. Wiggins, “Minimum entropy deconvolution,” Geophysics, vol. 16, pp. 12–35, Apr. 1978. doi: 10.1190/1.1440424.
[12] G. L. McDonald, Q. Zhao, and M. J. Zuo, “Maximum correlated kurtosis deconvolution and application on gear tooth chip fault detection,” Mech. Syst. Signal Process., vol. 33, pp. 237–255, Nov. 2012, doi:10.1016/j.ymssp.2012.06.010.

[13] Y. Miao, M. Zhao, J. Lin, and X. Xu, “Sparse maximum harmonics-to-noise-ratio deconvolution for weak fault signature detection in bearings,” Meas. Sci. Technol., vol. 27, no. 10, Oct. 2016, Art. no. 105004, doi:10.1088/0957-0233/27/10/105004.

[14] G. L. McDonald and Q. Zhao, “Multipoint optimal minimum entropy deconvolution and convolution fix: Application to vibration fault detection,” Mech. Syst. Signal Process., vol. 82, pp. 461–477, Jan. 2017, doi:10.1016/j.ymssp.2016.05.036.

[15] J. Wodecki, P. Krucezk, A. Wylomanska, A. Bartkowiak, and R. Zimroz, “Novel method of informative frequency band selection for vibration signal using nonnegative matrix factorization of short-time Fourier transform,” in Proc. IEEE 11th Int. Symp. Dig. for Electr. Mach., Power Electron. Drives (SDEMPED), Aug. 2017, pp. 129–133, doi:10.1109/SDEMPED.2017.8062345.

[16] L. Liang, L. Shan, F. Liu, B. Niu, and G. Xu, “Sparse envelope spectra for feature extraction of bearing faults based on NMF,” Appl. Sci., vol. 9, no. 4, p. 755, Feb. 2019, doi:10.3390/app09040755.

[17] P. Smaragdis, “Non-negative matrix factor deconvolution: extraction of multiple sound sources from monophonic inputs,” in Independent Component Analysis and Blind Signal Separation, vol. 3195, 2010, pp. 494–499, doi:10.1007/978-3-540-30110-3_63.

[18] D. O’Grady and B. A. Pearlmutter, “Discovering convolutive speech phones using sparseness and non-negativity,” in Independent Component Analysis and Signal Separation. Springer, 2007, pp. 520–527, doi:10.1007/978-3-540-74494-8_65.

[19] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 787–791, Oct. 1999, doi:10.1038/45655.

[20] D. Kim, S. Sra, and I. S. Dhillon, “Fast projection-based methods for the least squares nonnegative matrix approximation problem,” Stat. Anal. Data Mining, vol. 1, no. 1, pp. 38–51, Feb. 2008, doi:10.1002/sam.1014.

[21] C. Ding, T. Li, and M. I. Jordan, “Convex and semi-non-negative matrix factorizations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 1, pp. 45–55, Jan. 2010, doi:10.1109/TPAMI.2008.277.

[22] G. Thurau, K. Kersing, M. Wabha, and C. Baeckhage, “Convex non-negative matrix factorization for massive datasets,” Knowl. Inf. Syst., vol. 29, no. 2, pp. 457–478, Nov. 2011, doi:10.1007/s10115-010-0352-6.

[23] A. Cichocki, R. Zdunek, and S. Amari, Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. John Hoboken, NJ, USA: Wiley, 2008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4408452.

[24] C. Vaz, A. Toutou, and S. S. Narayanan, “Convex hull convolutive non-negative matrix factorization for uncovering temporal patterns in multivariate time-series data,” in Proc. Ann. Conv. Int. Speech Commun. Assoc. (INTERSPEECH), 2016, pp. 963–967, doi:10.21437/Interspeech.2016-571.

[25] W. Zhang, D. Yu, and X. Chen, “Fault diagnosis of rolling bearings based on resonance-based sparse signal decomposition and energy operator demodulating,” Proc. CSEE, vol. 33, no. 20, pp. 111–118, 2013.

[26] A. Moshefizadeh and A. Fasana, “The autogram: An effective approach for selecting the optimal demodulation band in rolling element bearings diagnosis,” Mech. Syst. Signal Process., vol. 105, pp. 294–318, May 2018, doi:10.1016/j.ymssp.2017.12.009.

[27] T. Shen, “Extracting information entropy features for rotating machinery vibration signals,” Chin. J. Mech. Eng., vol. 37, no. 6, pp. 94–98, 2009, doi:10.3901/jme.2001.06.094.

[28] Y. Miao, M. Zhao, and J. Lin, “Improvement of kurtosis-guided-grams via gini index for bearing fault feature identification,” Meas. Sci. Technol., vol. 28, no. 12, Dec. 2017, Art. no. 125001, doi:10.1088/1361-6501/aa8a37.

[29] 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,” Mech. Syst. Signal Process., vol. 92, pp. 173–195, Aug. 2017, doi:10.1016/j.ymssp.2017.01.033.

[30] H. Qiu, J. Lee, J. Lin, and G. Yu, “Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics,” J. Sound Vibrat., vol. 289, nos. 4–5, pp. 1066–1090, Feb. 2006, doi:10.1016/j.jsv.2005.03.007.