PAPER • OPEN ACCESS

SVD principle analysis and fault diagnosis for bearings based on the correlation coefficient

View the article online for updates and enhancements.

You may also like

SVD principle analysis and fault diagnosis for bearings based on the correlation coefficient

Zijian Qiao[1,2] and Zhengrong Pan[1,2]

[1] College of Electrical and Information Engineering, Lanzhou University of Technology, Lan Zhou 730050, People's Republic of China
[2] Gansu Province Key Laboratory for Manufacturing Informationization Systems, Lanzhou University of Technology, Lan Zhou 730050, People's Republic of China

E-mail: zijianqiao@hotmail.com

CrossMark

Abstract

Aiming at solving the existing sharp problems by using singular value decomposition (SVD) in the fault diagnosis of rolling bearings, such as the determination of the delay step k for creating the Hankel matrix and selection of effective singular values, the present study proposes a novel adaptive SVD method for fault feature detection based on the correlation coefficient by analyzing the principles of the SVD method. This proposed method achieves not only the optimal determination of the delay step k by means of the absolute value r_k of the autocorrelation function sequence of the collected vibration signal, but also the adaptive selection of effective singular values using the index ρ corresponding to useful component signals including weak fault information to detect weak fault signals for rolling bearings, especially weak impulse signals. The effectiveness of this method has been verified by contrastive results between the proposed method and traditional SVD, even using the wavelet-based method through simulated experiments. Finally, the proposed method has been applied to fault diagnosis for a deep-groove ball bearing in which a single point fault located on either the inner or outer race of rolling bearings is obtained successfully. Therefore, it can be stated that the proposed method is of great practical value in engineering applications.

Keywords: SVD, Hankel matrix, impulse signal detection, mechanical fault diagnosis, rolling bearings

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

1. Introduction

Rolling bearings are one of the most common classes of mechanical elements and play an important role in industrial applications. They generally operate in tough working environments and are easily subject to failures, which may cause machinery to break down and decrease machinery service performance such as manufacturing quality, operation safety, etc [1–3]. Therefore, increasing reliability with respect to possible faults has attracted considerable interest in mechanical

fault diagnosis in recent years [4, 5]. Finding adaptively effective signal-processing techniques to analyze vibration signals and to detect fault features has become a key problem in mechanical fault diagnosis. Meanwhile, it is also a challenge to propose and apply effective signal-processing techniques for extracting the crucial fault information from the collected vibration signals.

Currrently, traditional signal-processing technologies, including time-domain and frequency-domain analysis, are applied in mechanical fault diagnosis, such as wavelet

transform (WT) [6, 7], ensemble empirical mode decomposition (EEMD) [8], singular value decomposition (SVD) [9], etc. The essence of these signal processing techniques is restraining or eliminating noise. Among these signal-processing methods, however, SVD has exhibited a very good performance which is widely used to extract the fault features of rolling bearings in mechanical equipment, especially impulse features. Zhao *et al* [10] pointed out the similar mechanism of signal processing between SVD and WT, which is analyzed from the basis vector space angle and the characteristic of the Hankel matrix. Liu [11] presented a method of detecting abrupt information from the vibration signal, which uses SVD based on the Hankel matrix created by time series to extract early rub-impact faults between rotor and stator in rotating machinery. Kang Myeongsu *et al* [12] proposed SVD-based feature extraction method for fault classification of an induction motor, whose classification accuracy using a support vector machine (SVM) approach is very high. Subsequently, Brenner *et al* [13] investigated capability of extracting fault features from the flight data, and the comparisons between SVD and transformed-SVD detection results have been made. Although the above-mentioned literature has confirmed that the SVD method is an efficient and reliable tool for automated on-line analysis and mechanical fault diagnosis, there is very little literature to research the effect of different delay step k values for detection results and application of SVD in impulse signal detection. On the basis of analyzing the literature, two major problems restricting the application of SVD for impulse signal detection and mechanical fault diagnosis are summarized, especially in the field of rotating machinery. One is the determination of the delay step k for creating the Hankel matrix. A different delay step k can produce different reconstructed signals obtained by SVD. However, in nearly all of the research that we can see, the delay step k is subjectively fixed as a constant, generally 1, which not only loses generality but also does not consider the influence of different values of the delay step k on the reconstructed signal. According to the creation principle for the Hankel matrix, one can see that a smaller delay step can make the cross-correlation higher between two adjacent row vectors of Hankel matrix, causing information redundancy, whereas with larger k there is smaller information redundancy but many more data points are required for creating a Hankel matrix of the same dimension [14]. In addition, if the delay step is too small, then its Hankel matrix is a kind of ill-posed matrix [15], which can cause the reconstructed signal to be inaccurate due to the solution being approximate. Thus, it is necessary to research the effect of different delay steps for the reconstructed signal in order to propose a novel method of determining a proper delay step.

The other problem concerns how to select effective singular values to obtain the optimal reconstructed signal and detect fault features. Now for this problem, methods such as the difference spectrum of singular values (DSSV) [9, 16], median value of singular values, and mean value of singular values

can be employed to select effective singular values for getting the reconstructed signal. But these indices merely consider the magnitudes of singular values in terms of individual or global magnitude, and take no consideration of the contribution rate of each singular value to the original signal. These selection methods can cause weak fault features corresponding to smaller singular values to be eliminated and removed; conversely, strong disturbance information corresponding to greater singular values will remain, eventually causing difficulties in detection of fault features from the reconstructed signal. Therefore, it is vital for extracting mechanical fault features from the reconstructed signal to select suitable singular values.

In view of the above-mentioned problems which are of serious concern, the present study investigates SVD technology for mechanical fault feature detection and proposes an adaptively novel SVD method based on the absolute value of the autocorrelation function sequence to determine the delay step k for creating a proper Hankel matrix. On the basis of a large number of experiments, an effective measurement index r_k is proposed; by means of the minimum delay step k when $r_k < \varepsilon$ (where ε is an experimental value), this method can implement an effective determination of the optimal delay step k. Then, in order to make good the disadvantages of traditional selection methods, a difference spectrum algorithm based on a normalized correlation coefficient is proposed in the consideration of the contribution rate of every component signal or singular value for the original signal because there is a one-to-one relationship between component signals and singular values. Simulated experiments and engineering applications demonstrate that the proposed determination and selection method is effective in detecting weak impulse signals and fault features of rolling bearings.

The remaining part of this paper is organized as follows. Section 2 introduces the SVD algorithm and existing problems, and the signal decomposition principle of Hankel matrix-based SVD and the essence of component signals obtained by SVD are studied. Then the influence of the delay step k for the reconstructed signal and singular values distribution is illustrated through simulated fault experiments, and a form of adaptive determination method for the optimal delay step is proposed in section 3. In section 4 the correlation coefficient singular value decomposition (CCSVD) method of adaptively selecting effective singular values for getting the optimal reconstructed signal is proposed; this method can correctly detect impulse features and fault information that are submerged in strong ambient noise. In section 5 the fault features of the inner and outer races of rolling bearings are detected using the proposed CCSVD method. Finally the conclusions are given in section 6.

2. Principle analysis of SVD method

For a collected discrete signal $\vec{X} = [x(1), x(2), \ldots, x(N)]$, the Hankel matrix can be created using this signal as follows:

$$\overline{A} = \begin{bmatrix} x(1) & x(2) & \dots & x(n) \\ x(1 \times k + 1) & x(1 \times k + 2) & \dots & x(1 \times k + n) \\ \vdots & \vdots & \vdots & \vdots \\ x((m-1) \times k + 1) & x((m-1) \times k + 2) & \dots & x((m-1) \times k + n) \end{bmatrix}_{m \times n} \tag{1}$$

where $1 < n < N$, spectifically $n = N - (m-1) \times k$, and k is a constant integer called the delay step which generally is 1; then $\overline{A} \in \overline{R}^{m \times n}$.

The definition of singular value decomposition (SVD) [17–19] is as follows: for a matrix $\overline{A} \in \overline{R}^{m \times n}$, two orthogonal matrixes $\overline{U} = [\vec{u}_1, \vec{u}_2, \dots, \vec{u}_m] \in \overline{R}^{m \times m}$ and $\overline{V} = [\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n] \in \overline{R}^{n \times n}$ are guaranteed to exist that satisfy the following equation:

$$\overline{A} = \overline{U} \overline{\Sigma} \overline{V} \tag{2}$$

where $\overline{\Sigma} = [\text{diag}(\sigma_1, \sigma_2, \dots, \sigma_q), \overline{O}]$ or its transposition, determined by $m < n$ or $m > n$, $\overline{\Sigma} \in \overline{R}^{m \times n}$, while $\overline{O}$ is the zero matrix, $q = \min(m, n)$, and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_q > 0$. These $\sigma_i (i = 1, 2, \dots, q)$ are the singular values of matrix $\overline{A}$.

In order to implement the decomposition of a signal using SVD, equation (2) should be converted into the form of column vectors $\vec{u}_i$ and $\vec{v}_i$:

$$\overline{A} = \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{v}_2^T + \dots + \sigma_q \vec{u}_q \vec{v}_q^T \tag{3}$$

where $\vec{u}_i \in \overline{R}^{m \times 1}$, $\vec{v}_i \in \overline{R}^{n \times 1}$, $i = 1, 2, \dots, q$, and $q = \min(m, n)$. Based on the SVD principle, the vectors $\vec{u}_i$ are mutually orthonormal and they form an orthonormal basis of m-dimensional space; the vectors $\vec{v}_i$ are also orthonormal to one another and they form the orthonormal basis of n-dimensional space [20, 21].

Let $\overline{A}_i = \sigma_i \vec{u}_i \vec{v}_i^T$, then $\overline{A}_i \in \overline{R}^{m \times n}$ also. Supposing that $\vec{P}_{i,1}$ is the first row vector of $\overline{A}_i$, and $\overline{H}_{i,(n-k+1,\dots,n)}$ are k column vectors in the last k columns of matrix $\overline{A}_i$, as shown in figure 1, according to the creation principle for the Hankel matrix; if $\vec{P}_{i,1}$ and all the row vectors of matrix $\overline{H}_{i,(n-k+1,\dots,n)}$ are linked together in a given form as exhibited in figure 2, then a SVD component signal $\vec{P}_i$ can be obtained, which can be expressed as the vector form

$$\vec{P}_i = (\vec{P}_{i,1}, \overline{H}_{i,(n-k+1,\dots,n)}^1, \dots, \overline{H}_{i,(n-k+1,\dots,n)}^{m-1}) \tag{4}$$

where $\overline{H}_{i,(n-k+1,\dots,n)}^j (j = 1, 2, \dots, m-1)$ is the jth row vector of matrix $\overline{H}_{i,(n-k+1,\dots,n)}$, $\vec{P}_{i,1} \in \overline{R}^{1 \times n}$, and $\overline{H}_{i,(n-k+1,\dots,n)} \in \overline{R}^{(m-1) \times k}$:

All the component signals formed by $\overline{A}_i (i = 1, 2, \dots, q)$ make up one kind of decomposition for the original signal $\overline{X}$. To research what these component signals reflect in nature, first we might as well divide the component signal $\vec{P}_i$ into two segments, as illustrated in figure 2, in which the initial segment is $\vec{P}_{i,1}$, is the first row vector of $\overline{A}_i$, while the terminal segment is the sequential connection of $m - 1$ row vectors of matrix $\overline{H}_{i,(n-k+1,\dots,n)}$, where $\overline{H}_{i,(n-k+1,\dots,n)}$ is formed by the last k column vectors of $\overline{A}_i$ and the connection method is drawn by using the dotted line with arrows. It can be seen that if the delay step k is equal to 1, the terminal segment of component

signal $\vec{P}_i$ is the last column vector of matrix $\overline{H}_{i,(n-k+1,\dots,n)}$, highlighting a prevalent problem in which the effect of the delay step has been implicitly ignored. However, to extensively research the essence of the component signal, we will consider that k is not necessarily equal to 1; in this situation the terminal segment is a successive arrangement of the row vectors of matrix $\overline{H}_{i,(n-k+1,\dots,n)}$.

Suppose that the Hankel matrix $\overline{A}$ with delay step k created by the original signal is expressed by row vectors $\overline{X}_1, \overline{X}_2, \dots, \overline{X}_m, \overline{X}_m \in \overline{R}^{1 \times n}$. As is known from the creation principle for $\overline{A}$, the first vector $\overline{X}_1$ is the initial segment of the original signal and its projective coefficient on the basis vector $\vec{v}_i$ of n-dimensional space can be computed from equation (2) as follows:

$$\overline{X}_1 \vec{v}_i = \sigma_i u_{i1} \tag{5}$$

where u_{i1} is the first element of basis vector $\vec{u}_i$.

$\vec{P}_{i,1}$, which is the initial segment of component signal $\vec{P}_i$ and also the first row vector of matrix $\overline{A}_i$, can also be calculated according to the definition of $\overline{A}_i$ as below:

$$\vec{P}_{i,1} = \sigma_i u_{i1} \vec{v}_i^T. \tag{6}$$

One can easily see that $\vec{P}_{i,1}$ is the product of basis vector $\vec{v}_i$ and projective coefficient of $\overline{X}_1$ on this basis vector, in which the projective coefficient $\sigma_i u_{i1}$ decides the magnitude of $\vec{P}_{i,1}$, and vector $\vec{v}_i$ determines the direction of $\vec{P}_{i,1}$. While $\overline{X}_1$ is the initial segment of the original signal, so obviously $\vec{P}_{i,1}$ is actually the projection of the initial segment of the original signal on the vector $\vec{v}_i$, which is the ith basis vector of n-dimensional space, and this relationship is illustrated in figure 3(a).

Similarly, supposing that the Hankel matrix $\overline{A}$ created by the original signal is described by the column vectors $\overline{Y}_1, \overline{Y}_2, \dots, \overline{Y}_n, \overline{Y}_n \in R^{m \times 1}$, as is known from the construction principle of $\overline{A}$, these column vectors $\overline{Y}_{n-k+1}, \dots, \overline{Y}_n$ are the terminal segment of the original signal, which is not a simple transposition of these column vectors $\overline{Y}_{n-k+1}, \dots, \overline{Y}_n$ but the sequential connection of row vectors of the arranged matrix by them and projective coefficients of column vectors $\overline{Y}_{n-k+1}, \dots, \overline{Y}_n$ on the basis vector $\vec{u}_i$ of m-dimensional space can also be computed from equation (2):

$$\begin{cases} \overline{Y}_{n-k+1}^T \vec{u}_i = \sigma_i v_{i(n-k+1)} \\ \quad \vdots \\ \overline{Y}_n^T \vec{u}_i = \sigma_i v_{in} \end{cases} \tag{7}$$

where v_{in} is the nth element of basis vector $\vec{v}_i$.

The vectors $\overline{L}_{i,n-k+1}^T, \dots, \overline{L}_{i,n}^T$, which constitute the terminal segment of component signal $\vec{P}_i$ and also the last k column vectors of matrix $\overline{A}_i$, can be computed according to the definition of $\overline{A}_i$ as follows:

$$\vec{A}_i = \begin{bmatrix} x_i(1) & x_i(2) & \cdots & x_i(n-k+1) & \cdots & x_i(n) \\ x_i(k+1) & x_i(k+2) & \cdots & x_i(n+1) & \cdots & x_i(k+n) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_i((m-1)k+1) & x_i((m-1)k+2) & \cdots & x_i((m-1)k+n-k+1) & \cdots & x_i(N) \end{bmatrix} \vec{P}_{i,1}$$

$$\overline{H}_{i,(n-k+1,\ldots,n)}$$

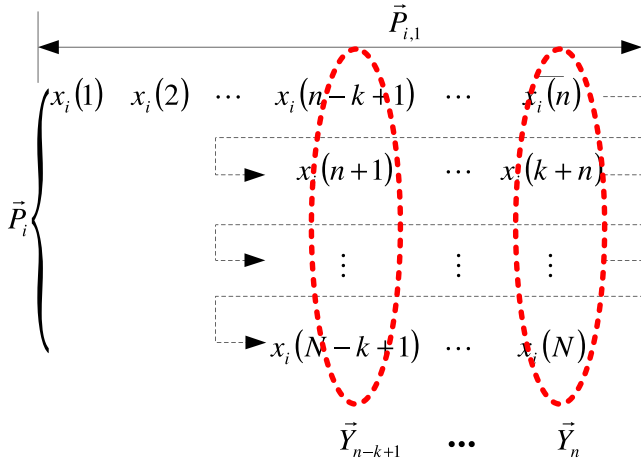**Figure 1.** The principle for forming the component signal $\vec{P}_i$ when the Hankel matrix is used.

Figure 2. The initial and terminal segments of component signal $\vec{P}_i$.

$$\begin{cases} \overrightarrow{L}_{i,n-k+1}^T = \sigma_i v_{i(n-k+1)} \vec{u}_i^T \\ \qquad\qquad \vdots \\ \overrightarrow{L}_{i,n}^T = \sigma_i v_{in} \vec{u}_i^T. \end{cases} \tag{8}$$

It can also be seen that $\overrightarrow{L}_{i,n-k+1}^T, \ldots, \overrightarrow{L}_{i,n}^T$ are the products of basis vectors $\vec{u}_i$ and projective coefficients of $\vec{Y}_{n-k+1}, \ldots, \vec{Y}_n$ on these basis vectors respectively, with the projective coefficients $\sigma_i v_{i(n-k+1)}, \ldots, \sigma_i v_{in}$ determining the magnitudes of $\overrightarrow{L}_{i,n-k+1}^T, \ldots, \overrightarrow{L}_{i,n}^T$, while the vectors $\vec{u}_i$ determine the directions of $\overrightarrow{L}_{i,n-k+1}^T, \ldots, \overrightarrow{L}_{i,n}^T$. $\vec{Y}_{n-k+1}, \ldots, \vec{Y}_n$ are the terminal segment of original signal, so $\overrightarrow{L}_{i,n-k+1}^T, \ldots, \overrightarrow{L}_{i,n}^T$ are actually the projections of the terminal segment of the original signal on the vector $\vec{u}_i$, which is the ith basis vector of m-dimensional space, and this relationship is illustrated in figure 3(b). So from the above analysis we can see that under different delay step values the nature of singular value decomposition based Hankel matrix is in fact to decompose the signal into m-dimensional and n-dimensional spaces.

In order to further analyze the characteristics of this kind of decomposition, assuming $\vec{A}_i$ is described by row vectors $\vec{P}_{i,1}, \ldots, \vec{P}_{i,m} \in \overline{R}^{1 \times n}$, while $\vec{A}$ is expressed by row vectors $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_m, \vec{X}_m \in \overline{R}^{1 \times n}$. According to equation (3), it is clear that each row vector of $\vec{A}$ equals the sum of corresponding row vectors in all $\vec{A}_i (i=1,2,\ldots,q)$, therefore we can obtain

$$\vec{X}_1 = \vec{P}_{1,1} + \vec{P}_{2,1} + \cdots + \vec{P}_{q,1}. \tag{9}$$

The matrix $\overline{H}_{i,(n-k+1,\ldots,n)}$ lies within $\vec{A}_i$, where $\overline{H}_{i,(n-k+1,\ldots,n)} \in \overline{R}^{(m-1)\times k}$. Supposing that the corresponding column vector in $\vec{A}$ is $\vec{I}_n$, where $\vec{I}_n \in \overline{R}^{(m-1)\times k}$, we can also see that $\vec{I}_n$ equals the sum of the corresponding column vectors $\overline{H}_{i,(n-k+1,\ldots,n)}$ in all $\vec{A}_i (i=1,2,\ldots,q)$. Obviously, their transposition can also meet this relationship, i.e.

$$\vec{I}_n^T = \overline{H}_{1,(n-k+1,\ldots,n)}^T + \overline{H}_{2,(n-k+1,\ldots,n)}^T + \cdots + \overline{H}_{q,(n-k+1,\ldots,n)}^T. \tag{10}$$

Based on the creation principle for the Hankel matrix, the original signal $\vec{X}$ can be described as the vector form $\vec{X} = (\vec{X}_1, \text{sub}(\vec{I}_n^T))$, where the operator 'sub' sequentially takes the column vectors of matrix $\vec{I}_n^T$, so the component signal $\vec{P}_i$ can also be described as the vector form $\vec{P}_i = (\vec{P}_{i,1}, \text{sub}(\overline{H}_{(n-k+1,\ldots,n)}^T))$. Then the sum of all these component signals is

$$\vec{P}_1 + \vec{P}_2 + \cdots + \vec{P}_q = (\vec{P}_{1,1} + \vec{P}_{2,1} + \cdots + \vec{P}_{q,1}, \text{sub}(\vec{I}_n^T)). \tag{11}$$

Based on equations (9) and (10), the right side of the above formula can be written as $(\vec{X}_1, \text{sub}(\vec{I}_n^T))$, so

$$\vec{P}_1 + \vec{P}_2 + \cdots + \vec{P}_q = \vec{X}. \tag{12}$$

Under different delay steps k, from equation (12) it can be seen that when the Hankel matrix is used, the component signals obtained by SVD can form a simple linear superposition for the original signal. This conclusion extends the range of the studies in [22–24], in which the delay step k is equal to 1. The advantage of this linear superposition is that the isolation of one component signal from original signal corresponds to simply subtracting this component from the original signal, and this subtraction computation will make the isolated component signal keep its phase the same as in the original signal; thus, there is no phase shift in the isolated component signal. Equation (12) is also the reconstruction formula for Hankel matrix-based SVD, and it is important that several effective component signals [25–27] can be simply added together to extract the feature information of the original signal.

3. Determination method of delay step *k*

To solve the above-mentioned problem, an effective method is proposed to create the proper Hankel matrix. For a collected discrete vibration signal $\vec{X} = [x(1), x(2), \ldots, x(N)]$, the autocorrelation function sequence is calculated using equation (13):
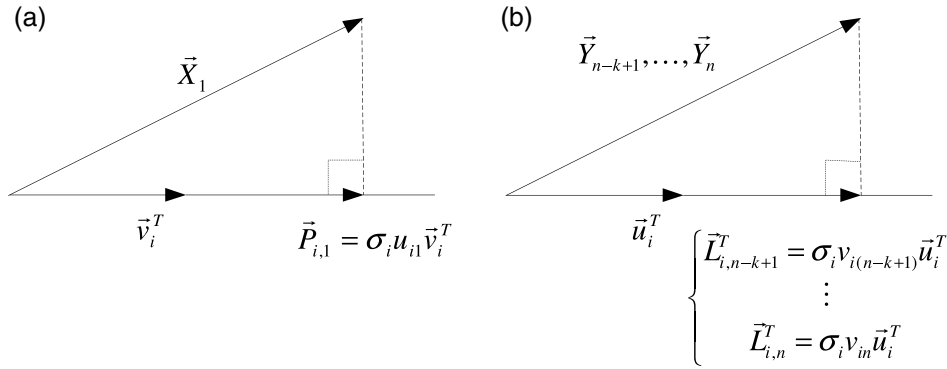
Figure 3. The essence of component signal $\vec{P_i}$: (a) essence of the initial segment of $\vec{P_i}$; (b) essence of the terminal segment of $\vec{P_i}$.
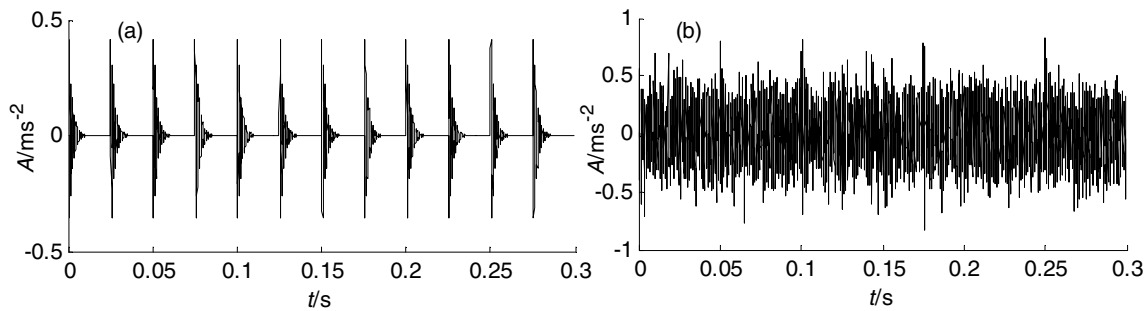

Figure 4. The bearing fault signal: (a) simulated impulse signal; (b) simulated fault signal.

$$r_k = \frac{1}{N}\text{abs}\left(\sum_{i=1}^{N} x(i)x(i+k)\right), k = 0, 1, \ldots, N. \qquad (13)$$

Here, we consider positive and negative correlations, so the absolute value of the autocorrelation function sequence is identified as index r_k, which can reflect the cross-correlation between two adjacent row vectors of Hankel matrix. The greater the value of r_k, the greater the correlation between two adjacent row vectors of the Hankel matrix. Through a number of experiments of different impulse signals and signal-to-noise ratios (SNRs), impulse signal amplitudes generated by one impulse fault are different, and the SNR in the simulated fault signal is different. We find that the delay step k is satisfactory when the index r_k is initially less than 0.1, which is an experimental value. The value weakens the correlation between two adjacent row vectors to reduce information redundancy and eliminate the problem of an ill-posed matrix [15]. Likewise, suitable data length for creating the Hankel matrix is required. In addition, the SNR is also relatively higher. This is also a compromise between information redundancy and data length required.

In order to verify the performance of proposed determination method, an impulse signal in [28] is employed to simulate bearing fault impulse, shown in figure 4(a). Gaussian white noise with SNR −9.6461 dB is added into the periodic impulse signal to obtain the simulated fault signal of rolling bearings, which can be shown in figure 4(b) where the sampling frequency is 20 000 Hz. The periodic impulse signal submerged in the surrounding noise can effectively simulate the tough operating environment of mechanical equipments. Thus, it has practical significance that the simulated fault signal can
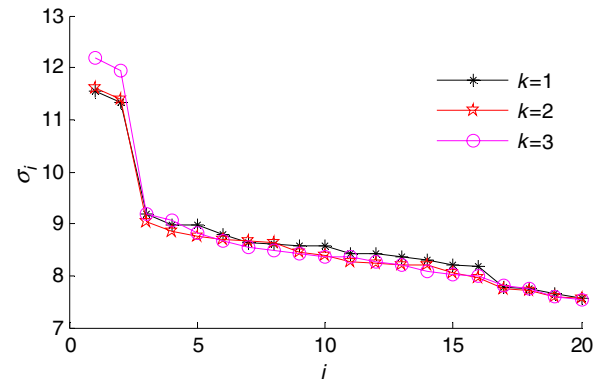

Figure 5. Singular values of simulated fault signals for different k.

be employed to research the effect of delay step k for detecting fault features by using SVD method.

To research the effect of different delay steps for detection capability, we consider situations in which k is 1, 2, and 3 respectively. Assuming that the simulated fault signal is long enough, a Hankel matrix with $m = 1700$ rows and $n = 20$ columns [10] is created using this simulated fault signal; 20 singular values can be obtained by the SVD method, as shown in figure 5. One can easily see that the first two singular values are relatively larger than any others, and there is a very large leap between the second and third singular value no matter what k is. This tells us that the characteristics of the impulse signal can be hidden in the first two singular values. With the gradual increase of k, however, the first two singular values start to rise slowly, and any other values may also descend slowly, with the gap between them becoming smaller and making the curve smoother. These changes can eliminate false peaks and
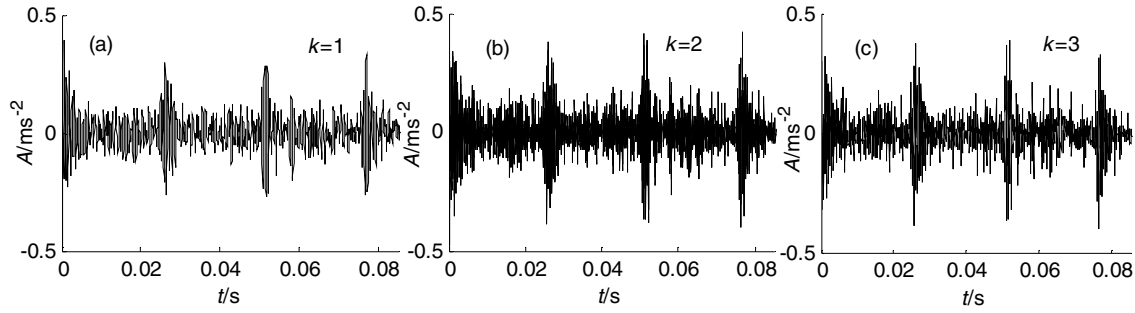
Figure 6. Reconstructed signal obtained by SVD for different *k*.

make true peaks higher in the DSSV. Traditionally speaking, the larger singular values can contain much more useful fault features, which should be effective ones [16] corresponding to the component signals having more fault information and less noise. Thus the above phenomena indicate that the effective singular values will be enhanced and the ineffective ones will be weakened with the increase of *k*. According to the DSSV, the first two singular values should be selected to reconstruct the original signal; the reconstructed signals with different delay steps are exhibited in figure 6. It is clear that the impulse signal can be extracted effectively and correctly in three figures, whose periods of 0.025 s can meet the given period in figure 4(a). Likewise, the surrounding noise is eliminated to a large extent, and by comparing the three figures one can easily see that the detection result at $k = 3$ has higher amplitude than the one at $k = 1$ and less noise than the one at $k = 2$. Relatively speaking, the detection performance at $k = 3$ is more satisfactory. In addition, the SNR between the three reconstructed signals and the simulated fault signal and data length required can be calculated respectively, as shown in table 1. With the increase of *k*, we can see that, first, the SNR may start to rise rapidly and then decline gradually, and second, the data length required for creating the Hankel matrix should be $n + (m - 1) \times k$, in which the larger *k* can cause a longer data length required for creating a Hankel matrix of the same dimension. Figure 6 and table 1 tell us that the optimal delay step should be 3 rather than 1 in terms of the best compromise.

Summarizing and analyzing the above three cases, first, we can clearly see that the energy of the original signal will be mainly concentrated in the first few component signals, which correspond to the first few largest singular values. Therefore, the first few component signals reflect the main skeleton of the original signal and achieve a similar effect to that of the approximation signal in wavelet transform. Meanwhile, detail features with low energy in the original signal will be isolated to the other component signals, which correspond to much smaller singular values. Second, there is no phase shift in component signals due to different delay steps *k*. In other words, the component signals obtained by the matrix-based SVD method have no phase shift no matter what *k* is. Finally, the processing and denoising effects of the reconstructed signal for different *k* are different.

The index r_k of the simulated fault signal is calculated using equation (13), as shown in figure 7. With the increase of *k*, we can see that the index r_k initially declines rapidly, then fluctuates slowly as the wave peaks appear. It is obvious that the

Table 1. Data length required and SNR for different *k*.

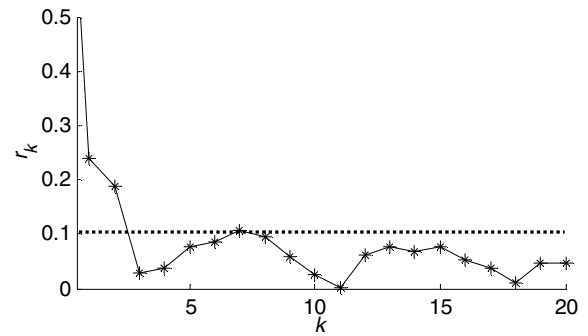| *k* (delay step)	*N* (data length required)	SNR (signal-to-noise ratio)
1	1719	− 5.2873
2	3418	− 4.0538
3	5117	− 1.7861
4	6876	− 2.1672

Figure 7. Index r_k of the simulated fault signal.

minimum of *k* is 3 under the condition $r_k < 0.1$, in other words, in which *k* is the *x*-axis value of the first left-hand point (under black dotted line). The value of *k* obtained by using the proposed determination method can meet the analysis results simulated above, in which the SNR between reconstructed signal and simulated fault signal is more satisfactory and the data length required is suitable. In essence, however, the determination method not only weakens the correlation between two adjacent row vectors of the Hankel matrix to reduce information redundancy and to eliminate the problem of an ill-posed matrix, but also allows a reasonable data length for creating the Hankel matrix. It is also an effective compromise between information redundancy and required data length.

4. Selection method of effective singular values

Up to now, no fixed selection method of effective singular values to reconstruct the original signal has been found. The existing methods such as DSSV, median value of singular values, and mean value of singular values merely consider the magnitude of singular values in terms of individual or global magnitude, and take no consideration of the contribution rate of each singular value for the original signal. The performance
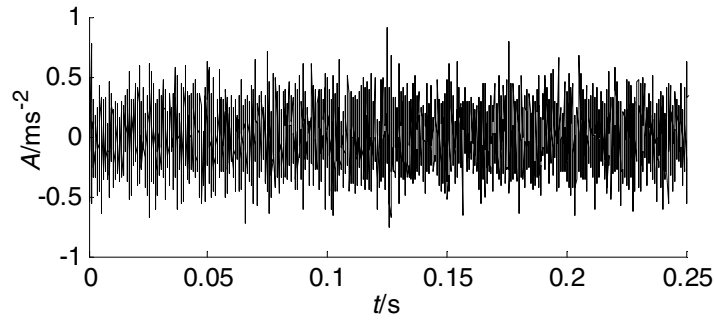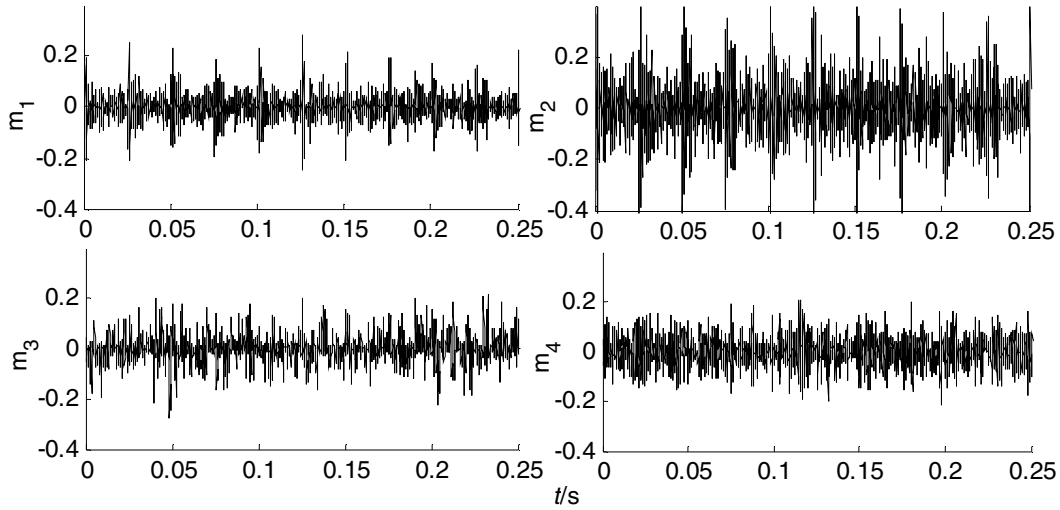
Figure 8. The simulated fault signal.

Figure 9. The component signals m_1, m_2, m_3, m_4 when $k = 3$.

of these traditional selection methods can be determined by the magnitude of individual or global singular values and gaps between each other, revealing two shortcomings: first, the loss of important fault features or significant noise remaining in the reconstructed signal. If the SNR is very low, then singular values corresponding to noise will be enlarged, causing a higher threshold. Thus the important features included in relatively small singular values will be lost [27]. If the SNR is relatively high, then the mean and median values of the singular values can be lessened to make the threshold lower. Therefore, much more noise may remain in the reconstructed signal, for example the median value method and mean value method of singular values; and second, there is the false peak value problem. Sometimes when dealing with signals with strong trend [18], such as aircraft engine health signals, one may hardly choose the right singular values by using the DSSV method due to the strong trend resulting in the false peak. To solve the above-mentioned problems, the CCSVD method is proposed in this section to select the effective singular values for obtaining the reconstructed signal. Assume that the original signal X is decomposed by SVD, and we can obtain a series of component signals $P_1, \ldots, P_q$, and calculate autocorrelation functions $R_X, R_{P_1}, R_{P_2}, \ldots, R_{P_q}$ using equation (14) as follows:

$$R(m) = \frac{1}{N} \sum_{i=1}^{N-1} x(i)x(i+m). \tag{14}$$

Then these correlation coefficients ρ may be computed among the autocorrelation function R_X and $R_{P_1}, R_{P_2}, \ldots, R_{P_q}$, and be normalized using equation (15):

$$\rho(j) = \frac{\displaystyle\sum_{i=1}^{2N-1} R_{P_j}(i)R_X(i)}{\sqrt{\displaystyle\sum_{i=1}^{2N-1} R_{P_j}^2(i) \sum_{i=1}^{2N-1} R_X^2(i)}} \tag{15}$$

where N is the length of the original signal, and $\rho(j), j = 1, 2, \ldots, q$ is the normalized correlation coefficient of jth component signal. Likewise, the difference spectrum of the normalized correlation coefficient is defined as

$$D_\rho(j) = \mathrm{abs}(\rho(j) - \rho(j+1)) \tag{16}$$

where $j = 1, 2, \ldots, q - 1$. The principle of selecting effective singular values for obtaining the reconstructed signal by using the index D_ρ is: if the maximum peak of D_ρ happens at $j = k$, then the first k singular values should be selected to reconstruct the original signal. But if the maximum peak of D_ρ happens at $j = 1$, then the first two singular values should be selected.

In order to test and verify the proposed selection method, the simulated fault signal in section 3 is employed figure 8 with SNR of $-10\,$dB. One can see that the fault impulse features are completely submerged in the surrounding noise. The index r_k of the simulated fault signal can be calculated
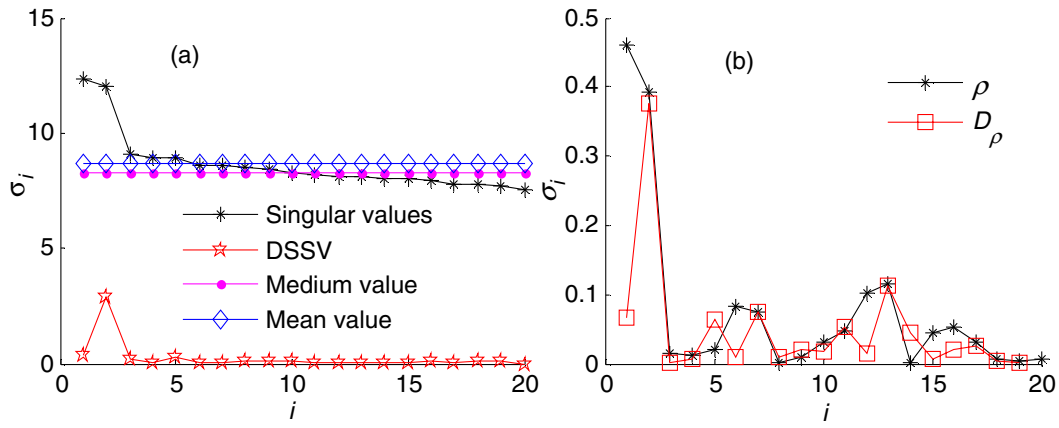
Figure 10. The selection indices: (a) the traditional selection methods; (b) proposed selection method.
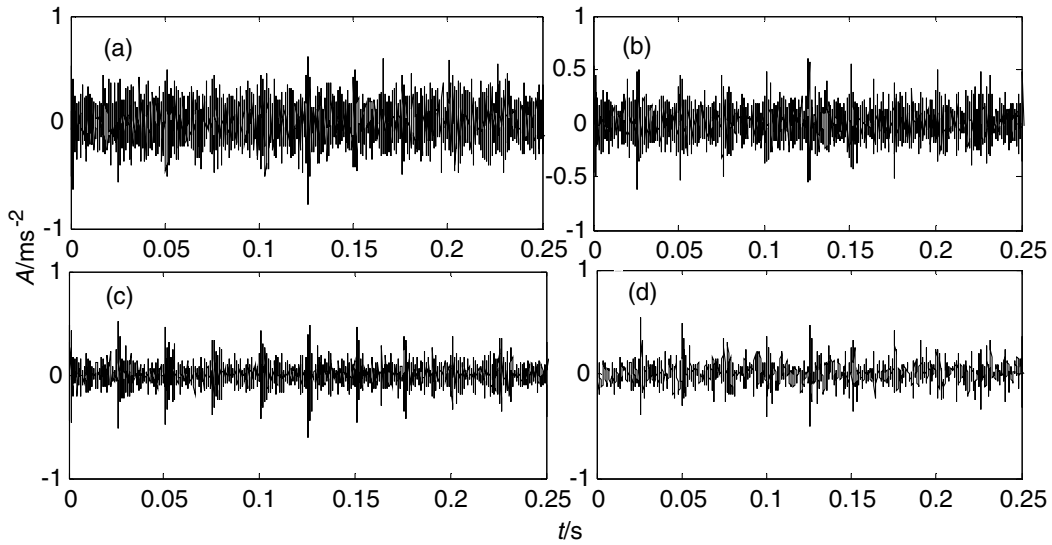

Figure 11. Comparison of five detection results: (a) median value method; (b) mean value method; (c) DSSV method and the proposed method; (d) wavelet-based method.

using the determination method proposed in section 3. We can obtain an optimal delay step $k = 3$ when r_k is initially less than 0.1, so the Hankel matrix can be created with $n = 20$ columns and $m = 1700$ rows based on $k = 3$. Then we can obtain the first 4 component signals, shown in figure 9. It is found that the periodic impulse features are exhibited in the component signals m_1 and m_2, but cannot be seen in other component signals. In other words the Hankel matrix with $k = 3$ can concentrate a majority of fault energy in the first two component signals.

Additionally, the values of ρ and D_ρ can also be computed according to equations (15) and (16), as shown in figure 10(b). It is obvious that the $\rho(1)$ and $\rho(2)$ are larger than any others, and the difference spectrum $D_\rho(i)$ can appear the highest spectrum peak at $i = 2$. According to the proposed CCSVD method, the first two singular values should be selected to obtain the reconstructed signal, which is exhibited in figure 11(c). Here we can see obviously periodic impulse features, and the surrounding noise has been weakened significantly. Compared with figures 8 and 11(c), we can also see that the impulse fault signal has been detected correctly by the CCSVD method, and there is also no phase shift in detection result. Although a small fraction of noise remains in the reconstructed signal

(detection result), it is not very important for us to examine and extract the periodic impulse features. To make a comparison with the traditional SVD methods, figure 10(a) exhibits the singular values curve of the simulated fault signal, the median value and the mean lines for them, and the DSSV. The curve of singular values in figure 10(a) and the curve of the normalized correlation coefficient ρ can be compared, and it can be seen that the first two greatest singular values in figure 10(a) correspond to the first two greater ρ values, while the different singular values with amplitudes close to each other in figure 10(a) reflect different magnitudes of ρ in figure 10(b). This tells us that different singular values of the same magnitudes have different contribution rates for the original signal. Thus, it is not correct when selecting effective singular values to consider merely the magnitudes of singular values. According to the selection principles of the median and mean value methods, we should choose these singular values above their lines to reconstruct the original signal, in which are first nine and first five singular values respectively. The reconstructed signals based on the median value and mean value methods can be obtained by SVD, shown in figures 11(a) and (b) respectively. Simultaneously, we can
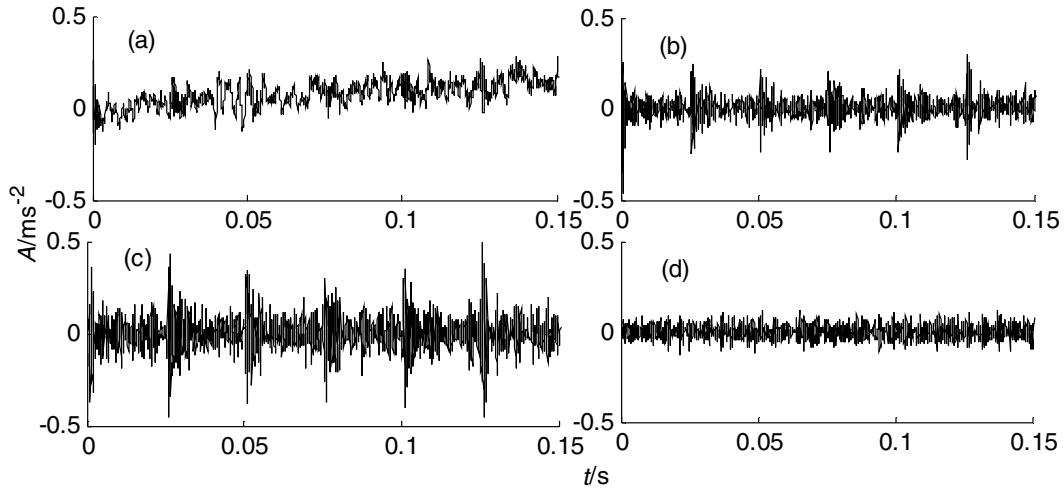
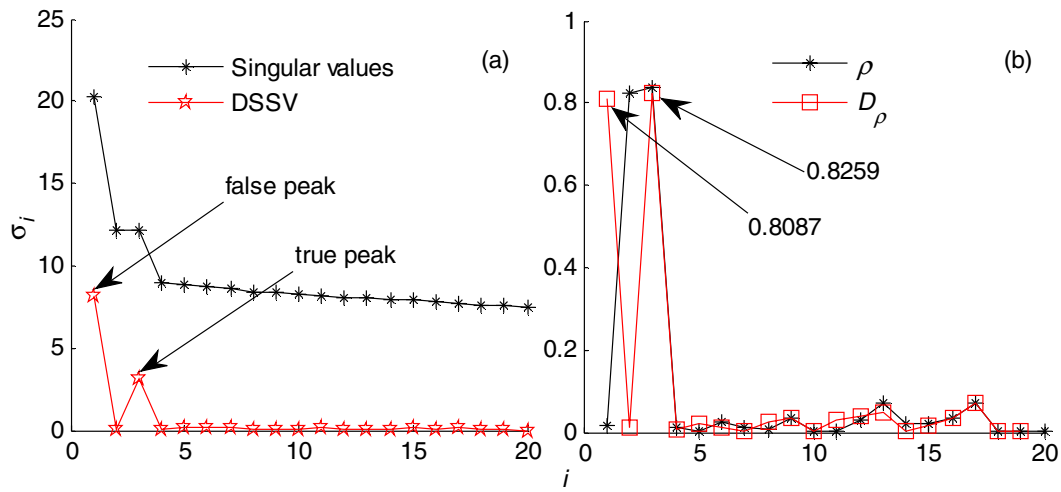**Figure 12.** The component signals m_1, m_2, m_3, m_4 when $k = 2$.

Figure 13. Two cases: (a) singular values and DSSV; (b) the contribution rate ρ and its difference spectrum D_ρ.

observe that the maximum peaks in the DSSV curve and of D_ρ both occur at $i = 2$. According to the selection principles of the DSSV and CCSVD methods, we should choose first two singular values to reconstruct the original signal, and the reconstructed signals are shown in figure 11(c). To make a sufficient comparison, the detection result using a wavelet-based method is shown in figure 11(d), using the threshold function 'rigrsure', 3-layer decomposition, and wavelet base db3. By comparing the five detection results, it can be found that the detection result of the medianm value method in figure 11(a) has too much noise to observe the impulse features, while the mean value method in figure 11(b) can extract the impulse features but they are very weak. The wavelet-based method in figure 11(d) eliminates the excessive noise which causes the loss and damage of the impulse faults, and it is disadvantagous for us to observe the periodic impulse faults. However, the performances of CCSVD and DSSV methods are same and their detection results also relatively more satisfactory in all detection results. Thus, the above detection results prove that the proposed CCSVD method can properly select effective singular values to obtain the reconstructed signal, which only retains much fault information but also weakens the

surrounding noise effectively. Of its nature, the proposed CCSVD method can select effective singular values based not on magnitudes of singular values themselves but on their contribution rates for the original signal, which can overcome the two above-mentioned shortcomings.

Through the above comparisons, one can see that the detection results of CCSVD and DSSV are the same for the impulse signal combined with Gaussian noise. To further prove the performance of the CCSVD method, suppose that there is a strong trend in the simulated fault signal [18], such as in an aircraft engine health signal. The strong trend can reflect and simulate the gradual increasing trend of a mechanical fault with time; therefore, a strong time (t) trend can be added into the fault signal simulated above. Using the determination method in section 3, we obtain the delay step as 2 and a Hankel matrix with 1700 rows and 20 columns can be created when $k = 2$. Then, the first 4 component signals can be obtained by SVD, which are exhibited in figure 12. It is clear that component signal m_1 merely contains the prevailing trend of the fault signal, while the impulse fault information is hidden in m_2 and m_3. This tells us that the energy of the strong trend can be mainly concentrated in the first component
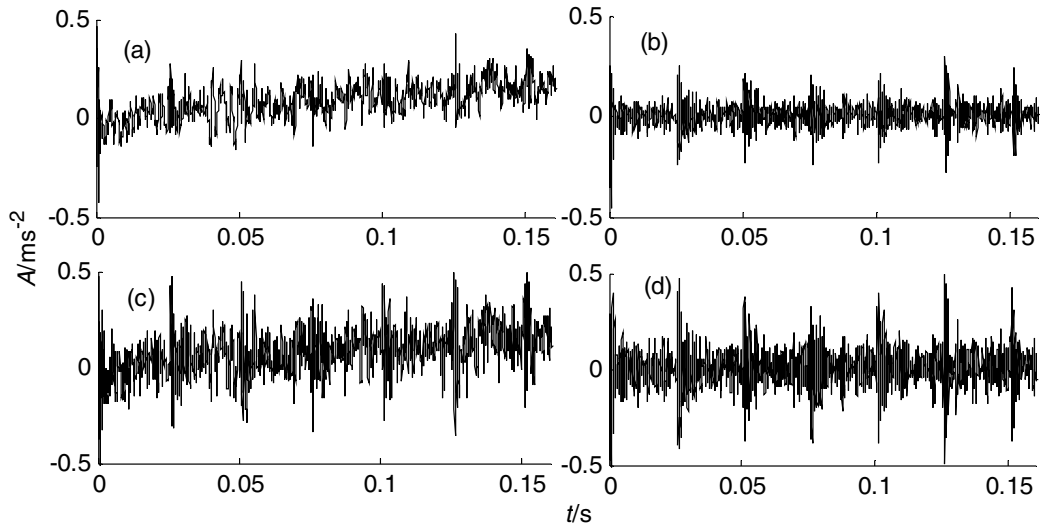
Figure 14. The detection results: (a) DSSV method; (b) strong trend subtracted using DSSV (a); (c) proposed CCSVD method; (d) strong trend subtracted using proposed method.
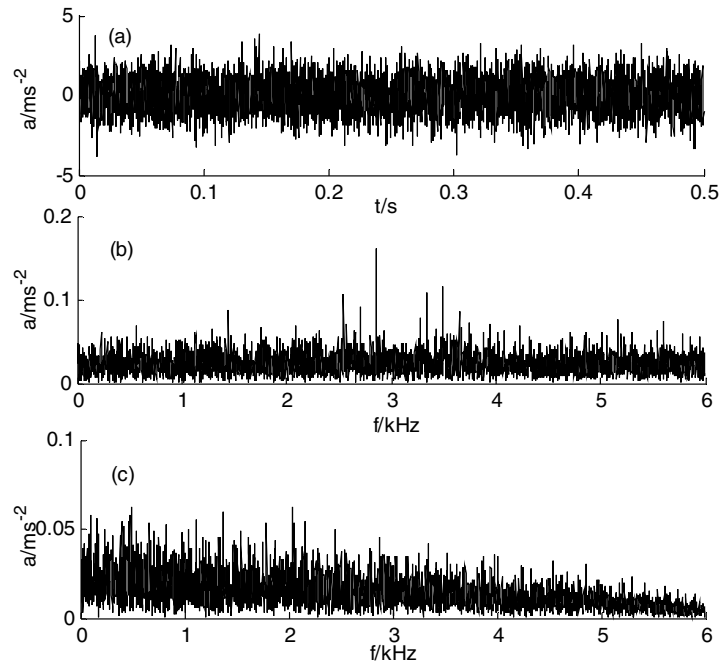

Figure 15. Fault signal of inner race: (a) time-domain waveform; (b) frequency spectrum; (c) envelope spectrum.

signal, making the first singular value greater. Therefore, the strong trend can make the gap greater between the first two singular values as shown in figure 13(a), which can cause the appearance of false peaks in the DSSV. According to the selection principle of DSSV, if the maximum peak happens at $i = 1$, then we should choose the first two singular values to reconstruct the original signal; the reconstructed signal is shown in figure 14(a). Meanwhile, the detection result subtracted the component signal m_1 which includes the strong trend exhibited in figure 14(b). According to the selection principle of CCSVD, from figure 13(b) one can see that the maximum peak of D_ρ happens at $i = 3$ and the false peak from the DSSV method can be eliminated and avoided. Thus, we should choose the first three singular values to get the reconstructed signal, as shown in figure 14(c) and the reconstructed

signal with the strong trend m_1 subtracted can be exhibited in figure 14(d). By comparing the detection results, it is clear that the proposed method can correctly locate the peak and select effective singular values to reconstruct the original signal. In addition, the proposed method has overcome the shortcoming of the false peak from the DSSV method.

Through two experimental results for impulse signal without and with strong trend, the detection results have proved fully that the proposed method and DSSV have same performance in the processing of a signal without strong trend, better than any other traditional methods. Likewise, compared with the wavelet-based method, it is found that the SVD method has better impulse signal detection capability . Finally, a comparison has been made in impulse signal detection with strong trend, and the experimental results have
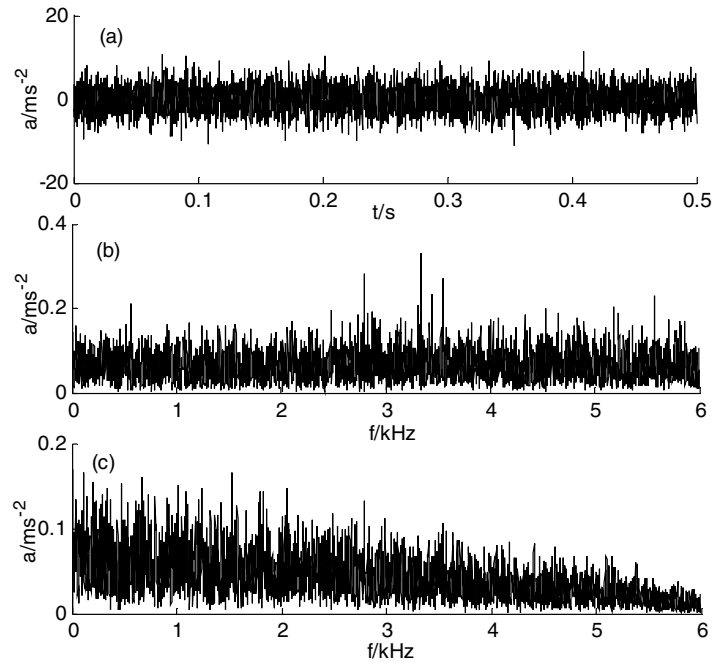
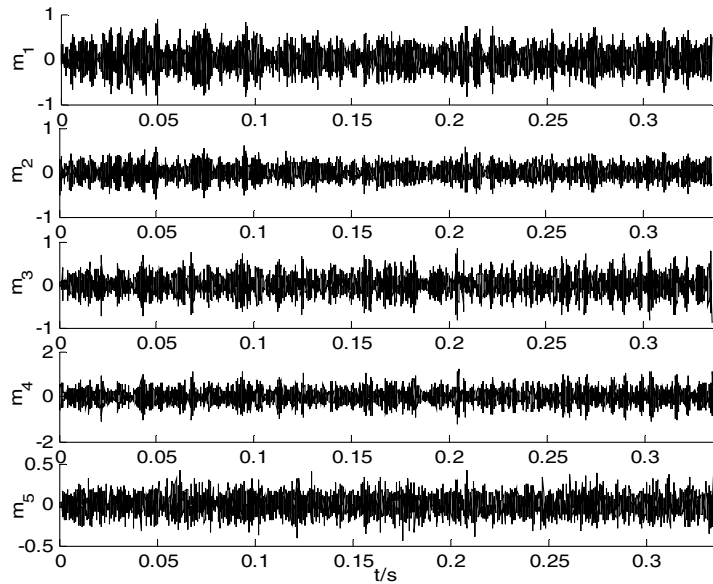**Figure 16.** Fault signal of outer race: (a) time-domain waveform; (b) frequency spectrum; (c) envelope spectrum.

Figure 17. Component signals $m_1, m_2, \ldots, m_5$ of inner race by SVD where $k = 1$.

testified that the proposed method can overcome the effect and disturbance of a false peak and correctly select effective singular values to get the optimal reconstructed signal.

5. Engineering applications

To verify the proposed diagnostic method, the test platform consists of a 2 hp motor, a torque transducer/encoder, a dynamometer, and control electronics. The test bearings which are 6205-2RS JEM SKF, deep-groove ball bearings, support the motor shaft. Single point faults were introduced to the test bearings using electrodischarge machining with fault diameters of 7 mm in both the inner and outer races. Vibration signal

was collected using accelerometers, which were attached to the housing with magnetic bases. Digital data were collected at 12 000 samples per second. We consider the instability in the beginning of the data, so data points between 1000 and 7000 of the fault signals from the inner and outer races are selected to analyze for extracting the fault features, shown in figures 15 and 16, where figures 15(a) and 16(a) are the fault signal time-domain waveforms of the inner and outer race respectively, figures 15(b) and 16(b) their frequency spectra, and figures 15(c) and 16(c) their envelope spectra. According to fault characteristic frequency theory for rolling bearings, one can determine that the fault frequencies of the inner and outer race on the rolling bearing are 157.94 Hz and 104.56 Hz respectively, but it is very difficult to see the fault frequency
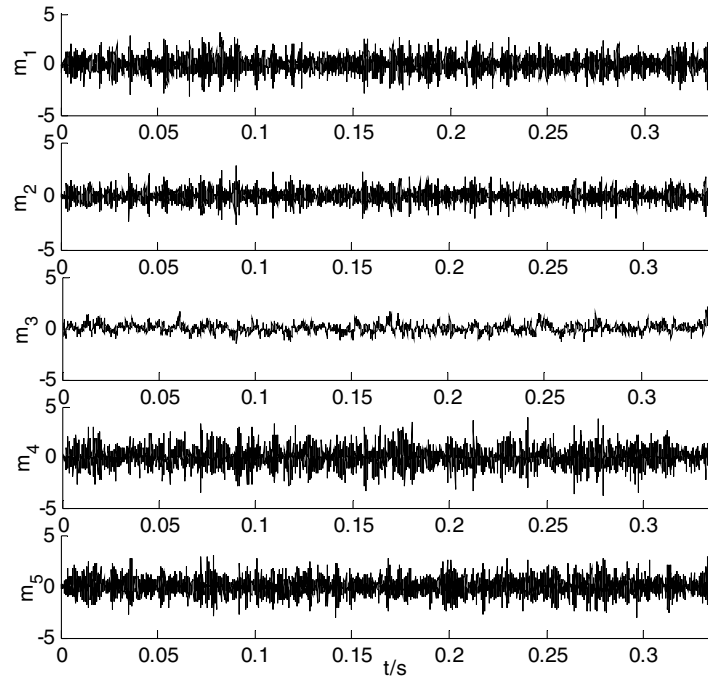
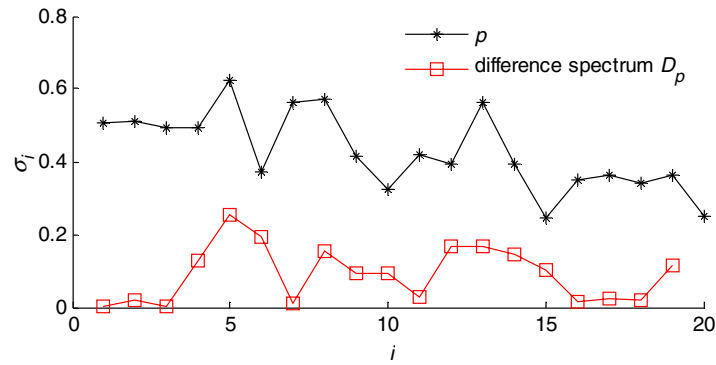**Figure 18.** Component signals $m_1, m_2, \ldots, m_5$ of the outer race by SVD where $k = 2$.

Figure 19. Normalized correlation coefficients and its difference spectrum for the inner race.
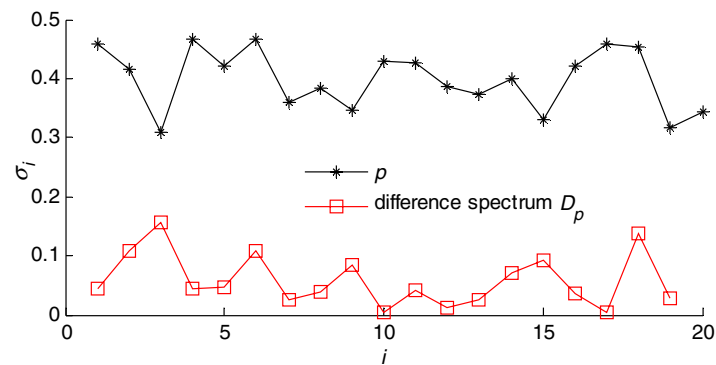

Figure 20. Normalized correlation coefficients and its difference spectrum for the outer race.

from figures 15 and 16. The fault information is completely submerged in the strong surrounding noise.

In order to detect the fault information, however, the proposed method in this paper is employed to process the fault signal of inner race and outer race of rolling bearings. The autocorrelation function sequence r_k can be calculated by using equation (13), where the minimum of delay step k are 1 (inner race) and 2 (outer race) respectively under the condition $r_k < 0.1$. Then, the Hankel matrix with $n = 20$ columns and $m = 4000$ rows is created by using equation (1) and the first five component signals $m_1, m_2, \ldots, m_5$ can be obtained by SVD, as shown in figures 17 and 18. In figure 17, we can see
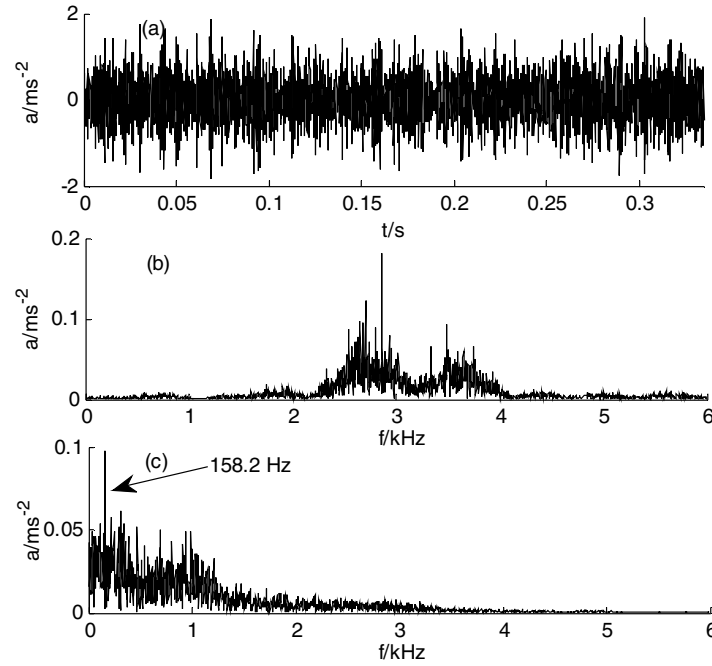
Figure 21. Reconstructed fault signal of the inner race: (a) time-domain waveform; (b) frequency spectrum; (c) envelope spectrum.
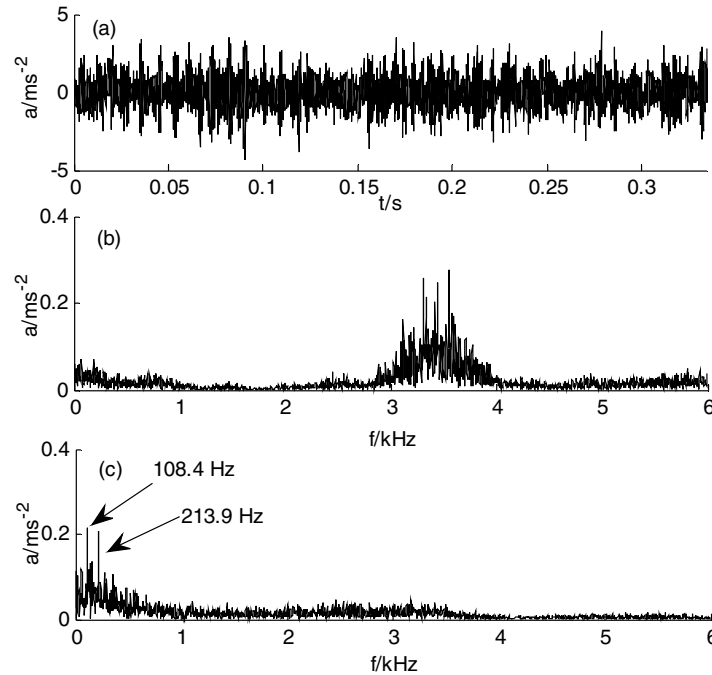

Figure 22. Reconstructed fault signal of the outer race: (a) time-domain waveform; (b) frequency spectrum; (c) envelope spectrum.

the strong impulse features in component signals m_1, m_2, m_3, and m_4, and the component signal m_5 has the modulated information to some extent. In figure 18, the component signals m_1 and m_2 have the strong impulse features, and the component signal m_3 has a little modulated information. However, in the component signals m_4 and m_5 there are also the weak impulse features, but the surrounding noise is also very strong, and relatively speaking the impulse features are very weak. Finally, the normalized correlation coefficients and difference spectra of the inner and outer race can be obtained according to equation (15) and (16) respectively, as shown in figures 19

and 20, and it is clear that the highest peak of $D_\rho(i)$ happens at $i = 5$ and $i = 3$ respectively. Based on the proposed CCSVD method we should respectively select the first 5 and 3 singular values (inner race and outer race) to reconstruct the fault signal, as shown in figures 21 and 22, in which figures 21(a) and 22(a) show the time-domain waveforms of reconstructed inner and outer race fault signal, figures 21(b) and 22(b) their frequency spectra, and figures 21(c) and 22(c) their envelope spectra. One can see the weak impulse features for the inner race from figure 21(a), where the surrounding noise has been weakened effectively. At the same time, from its frequency

13

spectrum in figure 21(b), we can observe that the frequency energy is concentrated in the frequency band between 2400 Hz and 4000 Hz. However, the frequency band of the noise is uniformly distributed in the white frequency band; in other words, the larger part of the noise is eliminated by the proposed CCSVD method. Finally the fault frequency is found to be 158.2 Hz (theoretical value is 157.94 Hz) from figure 21(c) and its amplitude is about 0.1, while the frequencies of noise and disturbance signal are concentrated in between 0 Hz and 1500 Hz, and their amplitudes are lower than 0.05. Compared with figures 15 and 21, we can see that the fault features are detected accurately and the surrounding noise is weakened effectively. Likewise, from the detection results of the outer race we can also see that the frequencies in figure 22(b) are concentrated in the frequency band between 3000 Hz and 4000 Hz and the frequencies in the other frequency bands are very weak. Comparing the envelope spectra in figures 16(c) and 22(c), it is clear that fault frequency of the outer race is 108.4 Hz (theoretical value is 104.56 Hz) and its amplitude is about 0.22. Additionally, it can be seen clearly that the double frequency is 213.9 Hz (theoretical value is 209.12 Hz). Therefore, the proposed method in this study has a better processing capability for weak impulse signal and a specific practical value in engineering applications.

6. Conclusions

The signal decomposition principle of Hankel matrix-based SVD and the essence of component signals obtained by this method are studied. Meanwhile the mechanism of SVD method is analyzed from the basis vector space angle and characteristics of the Hankel matrix. By theoretical analysis and signal processing examples, the following conclusions can be drawn:

1. A signal can be decomposed into the linear sum of component signals by Hankel matrix-based SVD no matter what k is, and these component signals physically reflect the projections of original signal on the orthonormal bases of m-dimensional and n-dimensional spaces.
2. The energy of the original signal will be mainly concentrated on the first several component signals for the structure characteristics of the Hankel matrix in itself, which can be corresponding to the first several singular values.
3. With the change of delay step k the SNR is different between the reconstructed signal and the original signal, but there is no phase shift in all component signals. A Hankel matrix with n columns and m rows is created, in which n and m are fixed, and the data length required is $n + (m - 1) \times k$ which grows with increasing k. Thus the proposed determination method is an effective compromise between SNR, information redundancy, and data length required.
4. The CCSVD method is proposed in section 4. This proposed method can eliminate the false peak in processing an impulse signal with strong trend and enhance the SNR in the reconstructed signal. Finally, it is applied to make fault diagnosis of rolling bearings. The experiments verify this proposed method is effective and accurate.

References

[1] Lei Y *et al* 2013 A review on empirical mode decomposition in fault diagnosis of rotating machinery *Mech. Syst. Signal Process.* **35** 108–26
[2] Lei Y *et al* 2012 Fault detection of planetary gearboxes using new diagnostic parameters *Meas. Sci. Technol.* **23** 055605
[3] Bottasso C L, Cacciola S and Iriarte X 2014 Calibration of wind turbine lifting line models from rotor loads *J. Wind Eng. Indust. Aerodyn.* **124** 29–45
[4] Zhu S, Qiao Z and Yang Z 2014 An improved method for the extraction of weak signal based on SVD and EMD *Meas. Control Technol.* **33** 60–2 (in Chinese)
[5] Fan X and Zuo M J 2008 Machine fault feature extraction based on intrinsic mode functions *Meas. Sci. Technol.* **19** 045105
[6] Gao L *et al* 2011 Roller bearing fault diagnosis based on nonlinear redundant lifting wavelet packet analysis *Sensors* **11** 260–77
[7] Li B *et al* 2011 Feature extraction for rolling element bearing fault diagnosis utilizing generalized S transform and 2D non-negative matrix factorization *J. Sound Vib.* **330** 2388–99
[8] Lei Y *et al* 2013 Fault diagnosis of rotating machinery based on an adaptive ensemble empirical mode decomposition *Sensors* **13** 16950–64
[9] Zhao X, Ye B and Chen T 2010 Difference spectrum theory of singular value and its application to the fault diagnosis of headstock lathe *Chin. J. Mech. Eng.* **1** 100–8 (in Chinese)
[10] Zhao X and Ye B 2009 Similarity of signal processing effect between Hankel matrix-based SVD and wavelet transform and its mechanism analysis *Mech. Syst. Signal Process.* **23** 1062–75
[11] Liu X 2002 New method of detecting abrupt information based on singular value decomposition and its application *Chin. J. Mech. Eng.* **6** 102–5 (in Chinese)
[12] Kang M and Kim J 2013 Singular value decomposition based feature extraction approaches for classifying faults of induction motors *Mech. Syst. Signal Process.* **41** 348–56
[13] Brenner M 2003 Non-stationary dynamics data analysis with wavelet-SVD filtering *Mech. Syst. Signal Process.* **17** 765–86
[14] Wu Y J, Chen E L and Shen Y J 2007 An improved method of detecting modulated gear fault characteristic based on singularity value decomposition *J. Shijiazhuang Railway Institute* **21** 530–4 (in Chinese)
[15] Morigi S, Reichel L and Sgallari F 2006 A truncated projected SVD method for linear discrete ill-posed problems *Numer. Algorithms* **43** 197–213
[16] Zhao X and Ye B 2011 Selection of effective singular values using difference spectrum and its application to fault diagnosis of headstock *Mech. Syst. Signal Process.* **25** 1617–31
[17] Hu H *et al* 2014 Incorporation of perceptually adaptive QIM with singular value decomposition for blind audio watermarking *EURASIP J. Adv. Signal Process.* **2014** 1–9
[18] Lei D and Zhong S 2013 Aircraft engine health signal denoising based on singular value decomposition and empirical mode decomposition methods *J. Jilin Univ.* **3** 764–70 (Engineering and Technology Edition) (in Chinese)

[19] Cong F *et al* 2013 Short-time matrix series based singular value decomposition for rolling bearing fault diagnosis *Mech. Syst. Signal Process.* **34** 218–30

[20] Banerjee M and Pal N R 2014 Feature selection with SVD entropy: some modification and extension *Inf. Sci.* **264** 118–34

[21] Marin L, Karageorghis A and Lesnic D 2015 A numerical study of the SVD-MFS solution of inverse boundary value problems in 2D steady-state linear thermoelasticity *Number Methods Partial Differ. Equitation* **31** 168–201

[22] Golub G H and Loan C F 1996 *Matrix Computations* 3rd edn (Baltimore, MD: Johns Hopkins University Press)

[23] Zhu Q *et al* 2005 Study on noise reduction in singular value decomposition based on structural risk minimization *J. Vib. Eng.* **18** 204–7 (in Chinese)

[24] Shin K *et al* 2003 Optimal autoregressive modeling of a measure noisy deterministic signal using singular value decomposition *Mech. Syst. Signal Process.* **17** 423–32

[25] Chiementin X *et al* 2012 Performance of wavelet denoising in vibration analysis: highlighting *J. Vib. Control* **18** 850–8

[26] Milford D and Sandell M 2014 Singular value decomposition using an array of CORDIC processors *Signal Process.* **102** 163–70

[27] Pan Z and Qiao Z 2014 Feature extraction based on improved SVD denoising and spectral kurtosis in early fault diagnosis of rolling element bearings *Proc. 5th Int. Symp. on Test Automation and Instrumentation* vol 2014 pp 14–21

[28] Randall R B and Antoni J 2011 Rolling element bearing diagnostics—a tutorial *Mech. Syst. Signal Process.* **25** 485–520