Postprint: Rolling Bearing Fault Identification Algorithm Based on Shift-Invariant Dictionary Learning and Sparse Coding

Authors: Qu Jianling, Yu Lu, peak, Tian Yanping, Li Yan

Date: 2018-05-20T00:00:00+00:00

Abstract

To address the issue of excessive dependence on expert prior knowledge in existing rotating machinery fault identification algorithms, an adaptive fault identification algorithm based on Shift-Invariant Dictionary Learning and Sparse Coding (SIDL-SC) is proposed. First, vibration signals under different fault conditions are segmented and subjected to smoothing preprocessing to reduce data processing complexity. Next, a shift-invariant dictionary learning algorithm incorporating adaptive penalty factors is employed to extract shift-invariant basis functions for different fault states. Then, an efficient feature sign search algorithm is utilized to solve for the sparse coefficients of the signal to be identified under different basis functions, thereby achieving reconstruction of the signal. Finally, the reconstruction residual serves as the criterion for identifying the fault state of the signal. Experimental results on rolling bearing vibration databases and measured aero-engine vibration signals demonstrate that the algorithm achieves higher fault identification accuracy compared to existing algorithms and exhibits strong feasibility in practical applications.

Full Text

Preamble

Title: Fault Recognition Algorithm for Rolling Bearings Based on Shift Invariant Dictionary Learning and Sparse Coding

Authors: Qu Jianling[1], Yu Lu[1]†, Gao Feng[1], Tian Yanping[1], Li Yan[2]
[1] Qingdao Branch of Naval Aviation University, Qingdao, Shandong 266041, China
[2] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

Abstract: Existing fault recognition algorithms for rotating machinery rely excessively on expert prior knowledge. To address this limitation, we propose an adaptive fault recognition algorithm based on Shift Invariant Dictionary Learning and Sparse Coding (SIDL-SC). The algorithm first segments and smooths vibration signals under different fault conditions to reduce data processing complexity. It then employs a shift invariant dictionary learning algorithm with an adaptive penalty factor to extract shift invariant basis functions for different fault states. Subsequently, an efficient feature sign search algorithm solves for the sparse coefficients of the signal to be recognized under different basis functions to achieve signal reconstruction. Finally, reconstruction residuals serve as the criterion for identifying the fault state of the signal. Experimental results on rolling bearing vibration databases and measured aero-engine vibration signals demonstrate that the proposed algorithm achieves higher fault recognition accuracy than existing methods and exhibits strong feasibility in practical applications.

Keywords: shift invariant dictionary learning; sparse coding; feature sign search; vibration signal; fault diagnosis

0 Introduction

Condition monitoring and fault diagnosis of mechanical equipment are crucial for improving production efficiency and preventing accidents. In mechanical fault diagnosis, effective extraction of fault features from monitoring data is essential for fault pattern recognition. However, diverse operating conditions and complex mechanical structures make feature extraction challenging. Time-domain and frequency-domain feature extraction algorithms rely heavily on manual feature selection and often require domain expert knowledge to perform effectively.

Researchers have conducted extensive work on mechanical vibration fault feature extraction. Harmouche et al. [3] utilized spectrum analysis to study characteristic frequencies and envelope spectra under different bearing faults, constructing linear classifiers for fault recognition. However, spectrum-based analysis methods still depend on manual experience and lack effective data-driven mining. Wang et al. [4] proposed a bearing fault diagnosis method based on Empirical Mode Decomposition (EMD) and grey relational theory, which decomposes vibration signals into Intrinsic Mode Functions (IMFs) and establishes a grey relational model between IMF energy distribution and bearing fault states. Nevertheless, EMD suffers from mode mixing, which may interfere with energy distribution analysis across different signal states. Cai et al. [5] proposed a rolling bearing fault diagnosis method based on higher-order statistics, which essentially restores the signal's higher-order spectrum to a power spectrum for fault feature extraction. However, higher-order statistics are significantly affected by singular values, leading to unstable algorithm performance.

In recent years, dictionary learning-based feature extraction algorithms have become a research hotspot in machine learning. The core idea is to construct a sparse transform domain from raw data, enabling sparse projection of signals to be recognized. Wright et al. [6] proposed a face recognition algorithm based on Sparse Representation Classification (SRC), establishing the first connection between sparse representation theory and pattern recognition. This enables sparse representation features to serve as feature vectors for accurate pattern recognition, offering new insights for fault pattern identification.

Dictionary construction typically follows two approaches: analytical dictionary construction and adaptive dictionary construction. Analytical methods leverage prior knowledge about raw data to achieve sparsification, often using orthogonal transform bases to construct overcomplete dictionaries. Vibration data commonly employs Discrete Cosine Transform [7] (DFT) and Fast Fourier Transform [8] (FFT) as orthogonal bases. While these orthogonal bases have well-established algorithmic frameworks, they lack flexibility and cannot guarantee optimal sparsity for corresponding data types. Certain signals, particularly those with wide time-frequency variations, remain non-sparse or insufficiently sparse under these bases, affecting data reconstruction accuracy [9]. Adaptive dictionary construction methods use raw data as training samples to construct overcomplete dictionaries through algorithms like K-Singular Value Decomposition [10] (K-SVD). Dictionaries constructed this way often achieve greater sparsity for representing original data. From a pattern recognition perspective, sparser features typically yield better classification performance.

For rotating machinery, characteristic features appear cyclically and repeatedly. Therefore, we propose introducing shift invariant dictionary learning into rotating machinery fault recognition to capture essential features with time-shift invariance from raw data. These features can be shifted to construct overcomplete dictionaries for effective identification of mechanical equipment fault states.

1.1 Shift-Invariant Feature Self-Learning Algorithm

In 2006, Smith [11] first proposed the Shift Invariant Sparse Coding (SISC) algorithm in *Nature* and successfully applied it to acoustic signal feature extraction. Conceptually, SISC treats the temporal shift of a feature pattern in a time series as an independent event, and the learned basis functions exhibit shift invariance relative to time. Regarding dictionary atoms, the basis functions learned by SISC represent essential features that repeatedly appear in signals, with each dictionary atom reflecting a specific characteristic of the signal. Consequently, the excellent shift-invariant properties of SISC have led to its application in image processing [12] and signal processing [13].

SISC uses a dictionary D containing M basis functions $\{d_m\}_{m=1}^{M}$, where each basis function $d_m \in \mathbb{R}^P$ has length P. Any input signal $y_i \in \mathbb{R}^N$ can be represented as a convolution sum of these basis functions and their corresponding coefficients. The basic mathematical model is:

$$y_i = \sum_{m=1}^{M} d_m * x_{im}$$

$$\min_{d,x} \sum_{i=1}^{N} \|y_i - \sum_{m=1}^{M} d_m * x_{im}\|_2^2 + \beta \sum_{i=1}^{N} \sum_{m=1}^{M} \|x_{im}\|_1$$

$$\text{s.t. } \|d_m\|_2^2 \leq 1, \forall m \in \{1, \ldots, M\}$$

where the sparse penalty term controls the sparsity level. Typically, this is set as a fixed value. However, due to varying training sample lengths, a fixed value often fails to achieve optimal solutions. Therefore, we propose an improvement to adapt to different training sample lengths for better sparsity control. Let $\beta = \frac{\lambda}{L}$, where L is the training sample length and λ is defined as the sparsity scale (typically $\lambda \in [0.01, 0.1]$), thereby achieving an improved sparse penalty term.

The dictionary learning stage involves simultaneous optimization of basis functions d and coefficients x, which constitutes a non-convex problem that cannot yield stable solutions. Therefore, drawing inspiration from the alternating optimization strategy in sparse coding, we update d and x iteratively until the objective function converges. The optimization process comprises two parts: coefficient solving and dictionary learning.

In the coefficient solving stage, with dictionary D fixed, we solve for sparse coefficients x. The model in Equation (1) decomposes into M independent optimization problems. Since coefficients solved for each input signal y_i are independent of other inputs, we consider the coefficient solving optimization problem for a single input.

We employ the efficient Batch-OMP algorithm [15] to find the best matching atom in dictionary D, expressed as:

$$\min_{x_{ik}} \|y_i - \sum_{k} d_k * x_{ik}\|_2^2 \text{ s.t. } \|x_{ik}\|_0 \leq T$$

In the dictionary learning stage, with coefficients x fixed, we solve for dictionary matrix D. The original optimization problem transforms into a constrained optimization problem:

$$\min_{d} \sum_{i} \|y_i - \sum_{j} d_j * x_{ij}\|_2^2 \text{ s.t. } \|d_j\|_2^2 \leq 1$$

Directly solving this equation involves enormous computational complexity. In the dictionary learning stage, we draw on time-frequency domain conversion concepts, converting computationally intensive convolution operations in the

time domain to the frequency domain. Multiplication operations in the frequency domain replace convolution operations in the time domain, reducing computational complexity from $O(N^2)$ to $O(N \log N)$ and improving efficiency. According to Parseval's theorem, the optimization can be converted into the following problem [14]:

$$\min_{\hat{d}} \sum_i \|\hat{y}_i - \sum_j \hat{d}_j \hat{x}_{ij}\|_2^2 + c \sum_j \|\hat{d}_j\|_2^2$$

where c is a constant, $\hat{d}_j$ is the discrete Fourier transform of basis function d_j, and $\hat{x}_{ij}$ and $\hat{y}_i$ are the discrete Fourier transforms of x_{ij} and y_i, respectively. Using the Lagrange multiplier method to solve this problem decomposes it into a sum of quadratic terms. For each frequency f in the frequency domain, we construct the Lagrange function:

$$\mathcal{L}(\hat{d}, \lambda) = \sum_i \|\hat{y}_i(f) - \sum_j \hat{d}_j(f) \hat{x}_{ij}(f)\|_2^2 + \Lambda \sum_j (\|\hat{d}_j(f)\|_2^2 - 1)$$

where λ is the dual variable and I is the identity vector. Since $\hat{d}_j(f)$ is a complex function, we convert the complex function to a real function for solving using complex formulas. The optimization parameters can be solved using Newton's method, and substituting into the objective achieves iterative updating of the basis.

Therefore, the SIDL-SC algorithm implementation consists of four parts: data preprocessing, shift-invariant dictionary learning, reconstruction coefficient solving, and residual calculation. The specific implementation steps are shown in Figure 1 [Figure 1: see original paper].

1.2 Coefficient Solving Algorithm Based on Feature Sign Search

Considering that the Batch-OMP algorithm requires presetting sparsity and other parameters, which increases computational load to some extent, improper parameter selection may lead to slow convergence or even failure to converge. Since the sparsity of test samples is generally unknown, this paper introduces the Feature Sign Search (FS) algorithm in the test sample identification stage to solve for reconstruction coefficients of the signal to be recognized on different basis functions. The FS algorithm, proposed by Lee [16], guesses coefficient signs to divide coefficient components into a feasible set (Φ) and a zero set (Θ). It continuously searches for the component with the maximum gradient change in the zero set and adds its index i to the feasible set until convergence. The specific steps are as follows:

a) Initialize $x = 0$, $\theta = \text{sign}(-A^T y)$, where θ represents the coefficient component sign.

b) Search among coefficient components that are zero for the component with the maximum gradient change and its index i, and add it to the feasible set.

c) Select submatrix $\hat{A}$ from matrix A, where $\hat{A}$ contains only column vectors corresponding to indices in the feasible set. Similarly, select sub-coefficient vectors $\hat{x}$ and $\hat{\theta}$ corresponding to $\hat{A}$, and solve the unconstrained optimization equation to obtain the optimal solution:

$$\min_{\hat{x}} \|y - \hat{A}\hat{x}\|_2^2 + \gamma \hat{\theta}^T \hat{x}$$

Perform line search between $\hat{x}$ and $\hat{x}_{\text{new}}$, find all coefficient vectors with sign changes, compare their objective functions with that at $\hat{x}_{\text{new}}$, and update $\hat{x}$ to the coefficient vector that minimizes the objective function. Remove indices i from the feasible set where coefficient components in $\hat{x}_{\text{new}}$ are zero, and update the sign set θ.

d) (a) If condition (b) holds, then the optimal solution is obtained; otherwise, return to step c). (b) If $\forall i \notin \Phi$, $|\frac{\partial}{\partial x_i}\|y - A\hat{x}\|_2^2| \leq \gamma$, and $\forall i \in \Phi$, $\text{sign}(\frac{\partial}{\partial x_i}\|y - A\hat{x}\|_2^2) = \theta_i$, then the optimal solution is obtained; otherwise, return to step b).

1.3 Fault Recognition Algorithm Based on Reconstruction Residual

Theoretically, the closer the basis function is to the true state of the signal to be recognized, the smaller the error between the reconstructed signal and the original signal. Therefore, we consider using reconstruction residual as the criterion for fault recognition. Using the FS algorithm, we solve for the sparse coefficients $\hat{x}_j$ of the signal to be recognized y under shift-invariant basis functions of each state. The reconstruction is performed using the following equation to solve for basis functions and their corresponding sparse coefficients:

$$\hat{y} = \sum_{j=1}^{M} d_j * \hat{x}_j$$

2.1 Experimental Setup

This paper selects the open bearing vibration database from Case Western Reserve University for experiments [17]. Single-point faults were set on the bearing outer ring, inner ring, and rolling elements using electric discharge machining technology, with fault diameters of 0.18 mm, 0.36 mm, and 0.54 mm, respectively. We divide different fault data into two groups as shown in Table 1 . Experiment 1 uses samples from Dataset A for training and testing, while Experiment 2 employs the basis functions learned from Experiment 1 to test

samples under 1-2 hp loads, thereby evaluating the generalization capability of the proposed algorithm.

2.2.1 Sparse Scale Parameter

In shift-invariant dictionary learning, we adopt an adaptive sparse scale parameter β to control the proportion of sparse coefficients in the objective function, preventing coefficients from becoming too large while basis functions become too small. In practice, larger β values yield faster convergence but higher reconstruction errors. Figure 2(a) compares reconstruction residuals of the original signal under different β values using inner ring 18mm fault samples from Dataset A as an example. The figure shows that lower sparse scales produce lower reconstruction errors, while reconstruction errors increase significantly as the sparse scale grows. However, from a computational efficiency perspective, smaller β values result in larger computational load per iteration, leading to slower convergence. Therefore, balancing reconstruction error and computational time efficiency, a sparse scale of 0.025 is appropriate.

2.2.2 Basis Function Length Selection

In shift-invariant dictionary learning, the selection of basis function length affects sample representation. For rolling bearings, the learned basis functions should contain at least one complete impact. From actual measurements, the average interval between two adjacent peaks in rolling bearing signals is approximately 76. Using experimental samples from the previous section, Figure 2 Figure 2: see original paper shows the sample recognition rate under different basis function lengths. The figure indicates that basis functions achieve higher recognition rates at lengths of 70-90. When basis function length is too short, each shift-invariant basis function cannot learn a complete signal impact, preventing optimal accuracy. Overly long basis functions may introduce information redundancy, which is detrimental to subsequent FS-based reconstruction. Therefore, we select a basis function length of 80.

2.2.3 Training Sample Length Selection

Experiments reveal that different training sample lengths significantly impact fault recognition. Using samples from classes 1, 2, 5, and 8 under 0 hp load as examples for training and testing, with 100 samples randomly selected from each fault state for testing, the results are shown in Figure 3 [Figure 3: see original paper]. The figure clearly shows that when training sample lengths are 128 and 256, the reconstruction of class 8 faults exhibits overlapping reconstruction errors from basis functions learned from classes 2, 5, and 8, as all three can reconstruct class 8 with low residuals. This overlap complicates correct sample identification. In contrast, input sample lengths of 512 and 1024 better avoid this phenomenon. Comparing original signals reveals that training sample lengths of 128 and 256 contain 1 and 3 complete impacts, respectively,

while lengths of 512 and 1024 contain more complete impacts. From a basis function self-learning perspective, more impacts enable basis functions to learn more distinctive features, which is more conducive to subsequent reconstruction and classification. However, increased training sample length leads to longer learning time per batch. Therefore, we select a training sample length of 512.

2.3 Experimental Results

Based on the experimental discussions in the previous section, the final parameter values for the SIDL-SC algorithm are determined as shown in Table 2 . Figure 4 [Figure 4: see original paper] displays some learned shift-invariant basis functions. In Experiment 1, 200 samples were randomly selected from each state in Dataset A for testing, with results shown in Table 3 .

Table 3 demonstrates that the proposed method performs well on Dataset A, with most samples achieving 100% recognition accuracy and only a few samples having 1-2 misclassifications. As the dataset sample size increases, the algorithm can achieve higher recognition accuracy.

2.4 Comparison with Traditional Methods

To demonstrate the superiority of the SIDL-SC algorithm in bearing fault recognition, we conducted comparative experiments with algorithms from other researchers. Liu [18] also utilized dictionary learning and sparse coding for bearing fault recognition but did not explore parameter selection issues, instead constructing a redundant dictionary from all basis functions and using an LDA classifier for fault classification. Chen [19] extracted time and frequency domain features such as standard deviation, skewness, and kurtosis from vibration signals and fed them into a convolutional neural network classifier using the LeNet5 model for fault recognition. Both methods, along with our algorithm, were trained on samples under 0 hp and tested on samples under various loads in Dataset B, with average results from 100 experiments shown in Figure 5 [Figure 5: see original paper].

The experimental results indicate that Chen's algorithm using traditional time-frequency domain features has relatively low recognition rates. While Liu's algorithm achieves high recognition rates on individual samples, its overall stability is inferior to our proposed algorithm.

2.5 Aero-Engine Field Experiments

To verify the feasibility of our algorithm in actual equipment, we applied it to measured vibration signals from a certain type of aero-engine. The experimental data were collected from multiple test runs of this helicopter turboshaft engine under four states: normal condition, rotor unbalance fault, rotor looseness fault, and rubbing fault. Equal-arc sampling was employed with 128 data points

collected per cycle. Typical vibration signals are shown in Figure 6 [Figure 6: see original paper].

Using the SIDL-SC algorithm for shift-invariant dictionary learning on vibration signals under different states, the parameters were set as follows: training sample length 256, basis function length 50, sparse scale 0.01, number of basis functions 2, with other parameters identical to Table 3. The learned shift-invariant basis functions are shown in Figure 7 [Figure 7: see original paper].

The figure shows that basis functions under normal state have relatively smooth waveforms, while those learned under fault states contain distinct characteristics unique to each fault type. Recognition experiments were repeated 100 times for each state, with recognition accuracy as the evaluation metric. Results are shown in Table 4 and compared with methods from references [18] and [19]. The experimental results indicate that due to waveform similarities between rotor unbalance fault, rotor looseness fault, and normal state, the algorithm's recognition rate is lower than that on the CWRU dataset. However, compared with the two aforementioned algorithms, our algorithm still maintains higher fault recognition rates, further verifying its feasibility.

3 Conclusion

SIDL-SC is a vibration signal fault recognition algorithm combining shift-invariant dictionary learning and sparse coding, particularly demonstrating high recognition accuracy for fault identification under multiple operating conditions. The method uses raw vibration data as training samples and captures shift-invariant basis functions reflecting essential characteristics of different states through feature self-learning. Fault categories are determined based on the minimum reconstruction residual criterion. Additionally, we optimize the penalty factor in shift-invariant dictionary learning by using sparsity scale as a variable, effectively solving the penalty factor setting problem under different training sample lengths. Using the CWRU vibration database, we thoroughly investigated the impact of different parameter selections on experimental results and selected optimal parameters as final experimental variables. Furthermore, we applied our algorithm to measured aero-engine vibration signals and compared it with other researchers' algorithms. The results demonstrate that our algorithm possesses stronger generalization capability and can be better applied to practical vibration data processing.

References

[1] Zhu Huijie, Wang Xinqing, Rui Ting, et al. Application of sparse coding based on frequency domain signals in mechanical fault diagnosis [J]. Journal of Vibration and Shock, 2015, 34(12): 59-64.

[2] Janssens O, Slavkovikj V, Vervisch B, et al. Convolutional neural network based fault detection for rotating machinery [J]. Journal of Sound & Vibration,

2016, 377: 331-345.

[3] Harmouche J, Delpha C, Diallo D. Improved fault diagnosis of ball bearings based on the global spectrum of vibration signals [J]. IEEE Trans on Energy Conversion, 2015, 30(1): 376-383.

[4] Wang Luyan, Wang Qiang, Zhang Meijun, et al. Grey diagnosis method for rolling bearing faults based on EMD [J]. Journal of Vibration and Shock, 2014, 33(3): 197-202.

[5] Cai Jianhua, Hu Weiwen, Wang Xianchun. Rolling bearing fault diagnosis method based on higher-order statistics [J]. Journal of Vibration, Measurement & Diagnosis, 2013, 33(2): 298-301.

[6] Wright J, Yang A, Ganesh A, et al. Robust face recognition via sparse representation [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2009, 31(2): 210-227.

[7] Zhang Xinpeng, Hu Niaoqing, Cheng Zhe, et al. Vibration data repair method based on compressed sensing [J]. Acta Physica Sinica, 2014, 63(20): 115-124.

[8] Guo Liang, Gao Hongli, Huang Haifeng, et al. Time-varying signal compression technology based on compressed sensing theory [J]. Journal of Southwest Jiaotong University, 2015, 50(3): 511-516.

[9] Peng Xiangdong, Zhang Hua, Liu Jizhong. Body area network compressed sensing ECG reconstruction based on overcomplete dictionary [J]. Acta Automatica Sinica, 2014, 40(7): 1421-1432.

[10] Aharon M, Elad M, Bruckstein A M. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation [J]. IEEE Trans. on Signal Processing, 2006, 54(11): 4311-4322.

[11] Smith E C, Lewicki M S. Efficient auditory coding [J]. Nature, 2006, 439(7079): 978-982.

[12] Mørup M, Schmidt M N, Hansen L K. Shift invariant sparse coding of image and music data [R]. Technical Report, 2007.

[13] Tang Haifeng, Chen Jin, Dong Guangming. Sparse representation based latent components analysis for machinery weak fault detection [J]. Mechanical Systems & Signal Processing, 2014, 46(2): 373-388.

[14] Grosse R, Raina R, Kwong H, et al. Shift-invariant sparse coding for audio classification [C]// Proc of Conference on Uncertainty in AI. Vancouver: AUAI Press, 2007: 149-158.

[15] Rubinstein R, Zibulevsky M, Elad M. Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit [J]. CS Technion, 2011, 40.

[16] Honglak L, Battle A, Raina R, et al. Efficient sparse coding algorithms [C]// Advances in Neural Information Processing Systems. 2006: 801-808.

[17] http://csegroups.case.edu/bearingdatacenter/pages/download-data-file [EB/OL].

[18] Liu Haining, Liu Chengliang, Huang Yixiang. Adaptive feature extraction using sparse coding for machinery fault diagnosis [J]. Mechanical Systems and Signal Processing, 2011, 25(2): 558-574.

[19] Chen Zhiqiang, Li Chuan, Sanchez R V. Gearbox fault identification and classification with convolutional neural networks [J]. Shock & Vibration, 2015(2): 1-10.

Note: Figure translations are in progress. See original paper for figures.