Bearing fault diagnosis using the Kolmogorov-Smirnov test on frequency features extracted using the Goertzel algorithm

Daniel Cordoneanu

1Department of Mechatronics and Precision Mechanics, University “Politehnica” of Bucharest, 313, Splaiul Independenței, Bucharest-060042
daniel.cordoneanu@stud.mec.upb.ro

Abstract. Fault diagnosis has been a field of interest during the last period, especially for predictive maintenance, given the advances in artificial intelligence and the available state of the art algorithms for machine learning, in a great number of libraries. In the industrial sector, fault diagnosis plays a very important role in order to avoid as much as possible downtime. Usually, rotating motors are involved in the actuation of the machines used in industry; therefore, bearings are an important part of the kinematic chain. The faults of the defective bearings can be detected within the frequency spectrum, at particular frequencies, which can be mathematically computed on the geometrical bases. Moreover, the use of the Goertzel algorithm allows the extraction of the specific features for these frequencies and their harmonics. The paper proposes a new approach of these features use for creating an initial data distribution, in order to be compared with future acquired data by use of the Kolmogorov-Smirnov test, as a good-fitness test for the new data.

1. Introduction
In the fault diagnosis domain machine learning algorithms get more and more important as more complex systems are built and nonlinearities of the mathematical model are harder to solve. Detecting an early fault can save money and downtime of the system. In a system composed of a motor and a moving mass, as Natu describes in [1], the number of faults that can appear are distributed as follows: 40% bearing faults, 38% stator faults, 10% rotor faults, others 12%. As it can be seen, almost half of the faults appear in the bearings, so it is of high importance that these kinematic parts are monitored in order to keep the machine in functional parameters as much as possible. Features based on frequency and spectrum analysis are described in many algorithms in literature. In [2], MFCC (mel-frequency cepstral coefficients) are used with a gaussian mixture model and kurtosis to diagnose faults in bearings. Kurtosis is also used in [3] alongside envelope spectrum analysis. In [4], Nabhan et al. describe some fault detection techniques for ball bearings; it is shown that the vibration and spectrum analysis techniques are very useful tools for fault diagnosis in rolling bearings. In addition to these frequency features, there are numerous ways in which they can be used in algorithms. Statistical algorithms are one type of methods to distinguish between functional or defect mechanisms. In literature, the Kolmogorov-Smirnov statistical test (K-S test) was used to diagnose faults in bearings as in [5], where the authors would get features and create distributions for faulty bearings and for functional bearings, then comparing new data to these distribution. Also, this test was used to successfully monitor gearboxes as in [6] and [7].
In this paper, a new approach will be presented where the test is used to compare distributions created based on Goertzel-extracted features from functional bearings. In this case, it will be an algorithm that’s based on anomaly detection.

2. Extracted features for bearings fault frequencies and harmonics

2.1. Bearing fault frequencies

Ball bearings are rotating elements which enable the movement by greatly reducing friction and by handling stress coming from the linked kinematic parts. The geometric variables of a bearing can be used to compute the frequencies at which different parts of the bearing (inner ring, outer ring), when damaged, can get increase the amplitude. At these frequencies and their harmonics, when a fault appears, peaks start to be present. As the fault increases in size and affects more and more the mechanical part, the energy moves to the higher frequencies, showing peaks in higher harmonics of the pre-computed values.

The frequencies for the fault diagnosis of a bearing are:

- \( F_{or} \) – outer ring frequency
- \( F_{ir} \) – inner ring frequency
- \( F_b \) – ball frequency
- \( F_c \) – cage frequency

They can be calculated as follows [8]:

\[
F_{ord} = \frac{nf_{\text{rotation}}}{120} \left( 1 - \frac{B_d \cos(\beta)}{P_d} \right) 
\]

(1)

\[
F_{ird} = \frac{nf_{\text{rotation}}}{120} \left( 1 + \frac{B_d \cos(\beta)}{P_d} \right) 
\]

(2)

\[
F_{pd} = \frac{nf_{\text{rotation}}}{60} \frac{P_d}{B_d} \left( 1 - \left( \frac{B_d \cos(\beta)}{P_d} \right)^2 \right) 
\]

(3)

\[
F_c = \frac{f_{\text{rotation}}}{120} \left( 1 - \frac{B_d \cos(\beta)}{P_d} \right) 
\]

(4)

Where

- \( n \) is the number of balls
- \( B_d \) is the ball diameter
- \( \beta \) is the contact angle
- \( f_{\text{rotation}} \) is the rotation frequency of the shaft in rotations per minute (RPM)
- \( P_d \) is the pitch diameter

2.2. Extracting features based on bearing fault frequencies and Goertzel algorithm

In [9], the authors present a way to extract features based on the Goertzel algorithm for monitoring only these frequencies and their harmonics. These features are represented by the Discrete Fourier Transform (DFT) magnitudes at the specific fault frequencies of the bearing.

Each of the fault frequencies and their multiples can be considered as a random variable with its own distribution. The sample points of the distribution would be represented by each DFT magnitude computed with the Goertzel algorithm for multiple time windows. For example, figure 1 represents the
histograms of the distributions for the outer ring defect frequency when the bearing is functional and when the outer ring is damaged.

Both distributions are exponential, but they have a different decaying factor ($\lambda$) as the exponential distribution is characterized by the probability density function (PDF):

$$f(x) = \lambda e^{-\lambda x}, \ x > 0$$ (5)

The probability properties of this distribution will not be discussed since they are of little importance. What it’s important is that for every frequency the initial distributions when the system is functional will be extracted. By having initial distributions, it will be possible to apply the K-S test under the null hypothesis that the new data is part of the already known distribution.

3. The Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (K-S test) is a statistical test to decide if a sample comes from a specific distribution by comparing one-dimensional probability distributions. It is a non-parametric test (it can be applied on data that doesn’t have a gaussian distribution) which can compare a sample with a theoretical distribution (one-sample K-S test) or compare two samples (two-sample K-S test). It can also be used as a goodness of fit test (test that shows how well sample data fit a certain distribution). This test uses the null hypothesis testing to an alternative hypothesis by using a statistic known as D-stat. It can also give a p-value which by using the p-test, if it’s less than 0.05, rejects the null hypothesis (meaning a significance level of 5%).

The test can be described by the following:
- $H_0$ – The data follows a specified distribution
- $H_1$ – The data doesn’t follow the specified distribution
- D-stat – the statistic that is given by the test
- $\alpha$ – the significance level – in this case 0.05
- The critical value for D

The test will reject or not the null hypothesis ($H_0$) based on whether the calculated D-stat is greater than the critical value (which is usually obtained from a table based on several parameters). The critical value for a two-sample test with a 95% confidence interval (correspondent for a significance interval of 0.05) can be approximated as [10]:

![Figure 1. Histograms of the distributions of the outer ring fault frequency for a functional bearing and for a faulty bearing with the outer ring damaged.](image)
\[ D_{\text{crit},0.05} = 1.36(n^{-1}_x + n^{-1}_y)^{1/2} \]  

where \( n_x \) and \( n_y \) are the number of samples for each distribution.

The D-stat can be calculated as follows:

\[ D = \max_{1 \leq i \leq N} \left[ \frac{F(Y_i) - \frac{i - 1}{N}}{\frac{i}{N} - F(Y_i)} \right] \]  

Where \( F \) is the Empirical Cumulative Distribution Function (ECDF) of a continuous distribution (normal, exponential, etc.). \( D \) can also be described as the absolute distance between the new sample ECDFs and the theoretical cumulative distribution function (CDF).

If the CDF of a random variable \( X \) is expressed as:

\[ F(x) = P(X \leq x) \]  

Then for a set of observations \( x_n \), the ECDF is:

\[ F_{\text{obs}}(x) = \frac{\# \text{observations} < x}{\# \text{observations}} \]  

If the observations are ordered as \( y_i \), then:

\[ F_{\text{obs}}(y_i) = \frac{i}{N} \]  

Once these variables have been calculated, the D-stat can be approximated, and the null hypothesis can be rejected in favor of the alternative hypothesis. The proposed algorithm tries to compare a distribution of newly recorded data against a pre-recorded distribution that is created using the features extracted from a functional bearing’s vibration data.

4. Using the K-S test for fault diagnosis

In the following section, a fault detection algorithm will be presented. The extracted Goertzel features using the algorithm from [9] will be processed as distributions. For each type of fault there will be available \( m \) features (represented by the base fault frequency and its multiples). For each feature, a distribution created from the functional bearing data will be used as comparison term for applying the K-S test against new sampled data. Based on the results of the K-S test, a fault detection score will be extracted and based on multiple samples a decision will be made.

4.1. Proposed algorithm

The steps of the proposed algorithm are as follows:

1. Find the rotation frequencies in RPM of the system that moves the kinematic part which has the monitored bearing – the most used rotation frequencies are enough; if a bearing has a fault, it will show up at any speed
2. For each rotation frequency, extract the features using the algorithm described in [9] by using data acquisition on a functional bearing (Shannon’s theorem should be considered when acquiring vibration data)
3. Store the features into a matrix for each of the possible faults. A matrix will have the dimensions \( m \) by \( n \) where \( m \) is the number of samples and \( n \) the number of frequencies
4. Record new data for the same time as for the training and apply step 2 and 3
5. For each possible fault, apply the K-S test where the data is represented by each column in the matrix
6. The probability that a fault is detected on the corresponding part for that sample is given by:
\[ P_{\text{sample}} = \frac{\# \text{rejected null hypotheses}}{\# \text{frequencies}} \]  

(11)

7. Given \( n \) samples, one could even apply conditional probability by using the law of total probability, which states that if several disjoint measurable events take place, then an event \( A \) of the same probability space would have the probability of:

\[ P(A) = \sum_{i} P(A \mid B_i)P(B_i) \]  

(12)

8. Considering \( A \) as being the event where the bearing has a fault in that specific component and \( B_i \) the event of the bearing having a fault present in the recorded data of the sample, the probability that \( A \) has happened given \( B_i \) is equal to the probability of \( B_i \) over the entire sample space:

\[ P(A \mid B_i) = \frac{P(B_i)}{n} \]  

(13)

In figure 2 a flowchart of the algorithm is presented. The learning phase can be considered as a batch learning, and the features will be stored for future use.

\[ P_{\text{sample}} = \frac{\# \text{rejected null hypotheses}}{\# \text{frequencies}} \]  

Figure 2. Flowchart of the algorithm using extracted features with Goertzel algorithm and applying the K-S test on them.
4.2. Experimentation
To check the algorithm’s output, data coming from Rockwell Science Center (available in [11]) was used. The DFT magnitudes for the frequencies and their harmonics are extracted using the algorithm described in [9]. This data was taken from accelerometers mounted on bearings linked to a 2 hp Reliance Electric motor. Data is available for the functional bearings and for faults made on the outer ring, inner ring and bearing balls. Knowing the type of bearing (6205-2RS JEM SKF, deep groove ball bearing) and the motor speed (1797 RPM), the fault frequencies were computed for the rotating speed (see table 1).

Table 1. Bearing fault frequencies in Hz for 1797 RPM.

|     | \( F_{or} \) | \( F_{ir} \) | \( F_{b} \) | \( F_{c} \) |
|-----|-------------|-------------|-------------|-------------|
|     | 107.36      | 162.18      | 141.16      | 11.92       |

Since in the data there is not cage fault recorded, \( F_c \) will be ignored in computing the K-S score. Given this data, a script was created in Octave which would separate the data into small windows of 1200 points (equivalent of 166 milliseconds for a sampling rate of 12 kHz). The windows would then be split intro training data (60%), cross-validation data (20%) and test data (20%). These steps were done for each type of present fault. Once the features are extracted, they are exported into .csv files.

A script in R will do the K-S test for different frequencies for each fault by first loading the data from the .csv files and then going through each fault and frequency.

4.3. Results
The following figures represent the computed ECDF for the features extracted using the algorithm in [9]. They have been generated by the R script only for the fault frequencies in table 1: \( F_{or}, F_{ir}, F_{b} \).

![Figure 3. ECDFs of the distributions of the outer ring fault frequency for a functional bearing and for a faulty bearing with the outer ring damaged.](image)
Figure 4. ECDFs of the distributions of the inner ring fault frequency for a functional bearing and for a faulty bearing with the inner ring damaged.

Figure 5. ECDFs of the distributions of the ball fault frequency for a functional bearing and for a faulty bearing with the one of the balls damaged.

Based on the differences that can be seen in each of the figures 3, 4 and 5, one can tell that the faulty bearing data distribution is not part of the same distribution as the data recorded for the functional bearing. The K-S test can also easily reject the null hypothesis as described in chapter 3 based on the D-statistic.
The script written in R, besides going over each feature and doing the K-S test (by using the ks.test function [12]), it also calculates the critical D value and the D-stat for each feature. In the following tables the results based on the D-stat are presented alongside the calculated critical D value for each type of bearing fault. Tables 3, 4 and 5 are for the bearing fault frequencies at which the amplitude increases for a fault present in the inner ring, outer ring or one of the balls respectively.

| Nr. feature | D-stat for outer ring fault frequency | D_critical for outer ring fault frequency |
|-------------|--------------------------------------|------------------------------------------|
| 1           | 0.7639767                             |                                          |
| 2           | 0.3739264                             |                                          |
| 3           | 0.2757252                             |                                          |
| 4           | 0.9917355                             |                                          |
| 5           | 0.9525199                             | 0.1828094                                |
| 6           | 0.5648193                             |                                          |
| 7           | 0.4141144                             |                                          |
| 8           | 0.1521633                             |                                          |
| 9           | 0.5261708                             |                                          |
| 10          | 0.2399125                             |                                          |

| Nr. feature | D-stat for inner ring fault frequency | D_critical for inner ring fault frequency |
|-------------|--------------------------------------|------------------------------------------|
| 1           | 0.4799060                             |                                          |
| 2           | 0.3806514                             |                                          |
| 3           | 0.4998380                             |                                          |
| 4           | 0.5794037                             |                                          |
| 5           | 0.4731810                             | 0.1828094                                |
| 6           | 0.7681899                             |                                          |
| 7           | 0.9803922                             |                                          |
| 8           | 0.9638632                             |                                          |
| 9           | 0.9705882                             |                                          |
| 10          | 0.9509804                             |                                          |

| Nr. feature | D-stat for ball fault frequency | D_critical for ball fault frequency |
|-------------|---------------------------------|-----------------------------------|
| 1           | 0.9144668                       |                                   |
| 2           | 0.7031212                       |                                   |
| 3           | 0.9543449                       |                                   |
| 4           | 0.7896975                       |                                   |
| 5           | 0.1461927                       | 0.1823273                        |
| 6           | 0.5820428                       |                                   |
| 7           | 0.2888550                       |                                   |
| 8           | 0.3361951                       |                                   |
| 9           | 0.2758565                       |                                   |
| 10          | 0.6847469                       |                                   |

Since only one sample of fault recorded data was used to test the algorithm, the probability of the bearing having a fault for a component is equal to the probability of that sample.
The K-S test rejects the null hypothesis that the new data is part of the same distribution as the pre-recorded data if \( D_{\text{stat}} > D_{\text{critical}} \). Based on the data displayed in the table 4, 5 and 6 the probabilities of a fault being present for a certain part can be calculated. These probabilities are also computed by the script successfully and the algorithm detects the faults in all the damaged components. The results are available in table 5 where:

- \( P_{\text{ord}} \) is the probability of an outer ring defect
- \( P_{\text{ird}} \) is the probability of an inner ring defect
- \( P_{\text{bd}} \) is the probability of a ball defect

| \( P_{\text{ord}} \) | \( P_{\text{ird}} \) | \( P_{\text{bd}} \) |
|-------------------|-------------------|-------------------|
| 90%               | 100%              | 90%               |

5. Conclusions and future work
Given the results in table 5, the algorithm successfully detects the different distributions in the monitored frequencies.

5.1. Advantages and disadvantages
The main advantages of this algorithm are:

- Use of a way to extract spectral features which has higher computational efficiency than the FFT given the needed frequencies
- No loss function and no need to find a global minimum of a function
- It is linear in space and time with the number of recorded samples
- The features do not need to be processed immediately as there can be time spent between the acquisition and the processing in which the processor is free

Besides the advantages, some disadvantages of this algorithm would be present especially if it needs to be implemented on devices with lower processing power or embedded devices:

- The need to store the data as matrices, which could take a lot of space
- The fault is not detected immediately as it appears

5.2. Future work
As for a future development, it would be interesting to test how this algorithm can be used to extract useful information about which harmonic is more prone to a different distribution in different states of a fault.

6. References
[1] M. Natu, “Bearing Fault Analysis using Frequency and Wavelet Techniques,” Int. J. Innov. Manag. Technol., 2012.
[2] F. V. Nelwamondo and T. Marwala, “Faults detection using Gaussian mixture models, mel-frequency cepstral coefficients and kurtosis,” in Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, 2007.
[3] S. Wan, X. Zhang, and L. Dou, “Compound fault diagnosis of bearings using improved fast spectral kurtosis with VMD,” J. Mech. Sci. Technol., 2018.
[4] A. Nabhan, N. M. Ghazaly, A. Samy, and M. M.O, “Bearing fault detection techniques - a review,” Turkish J. Eng. Sci. Technol., 2015.
[5] C. Kar and A. R. Mohanty, “Application of KS test in ball bearing fault diagnosis,” J. Sound Vib., 2004.
[6] C. Kar and A. R. Mohanty, “Multistage gearbox condition monitoring using motor current signature analysis and Kolmogorov-Smirnov test,” J. Sound Vib., 2006.
[7] X. Wang and Z. Wan, “Using Kolmogorov-Smirnov test on diagnosis of gear faults under fluctuating load conditions,” Zhongguo Jixie Gongcheng/China Mech. Eng., 2009.
[8] SKF, “Bearings Spectrum Analysis,” 2012.
[9] D. Cordoneanu and C. Nițu, “An Approach of Extracting Features for Fault Diagnosis in Bearings Using the Goertzel Algorithm,” in Lecture Notes in Networks and Systems, 2020.
[10] “Critical Values for the Two-sample Kolmogorov-Smirnov test (2-sided).” [Online]. Available: http://sparky.rice.edu/astr360/kstest.pdf.
[11] “Bearing Data Center.” [Online]. Available: http://csegroups.case.edu/bearingdatacenter/home. [ Accessed: 02-Jun-2019].
[12] “R KS test documentation.” [Online]. Available: https://www.rdocumentation.org/packages/dgof/versions/1.2/topics/ks.test.

Acknowledgments
This work has been funded by the European Social Fund from the Sectoral Operational Programme Human Capital 2014-2020, through the Financial Agreement with the title “Scholarships for entrepreneurial education among doctoral students and postdoctoral researchers (Be Antreprenor!)”, Contract no. 51680/09.07.2019 - SMIS code: 124539.