Machine learning and signal processing methods for anomaly detection in predictive maintenance (Project outline)

Yapi Donatien Achou

April 10, 2019

Contents

1 Introduction

Since the industrial revolution, engines and machines have been the driving force for economical growth across industries such as automotive, airline, oil and gaze, to name a few. However, machines are prone to failure and must be monitor and maintain regularly to avoid catastrophic failure leading to significant financial and human lost. To mitigate equipment and machines proclivity toward failure and the associated cost, a process called predictive maintenance has been developed within the industrial community. Predictive maintenance for machines and industrial equipments can be defined as a maintenance philosophy or more generally a framework with a set of methods used to predict and prevent machine failure in order to avoid unexpected downtime and reduce related human and financial cost. This maintenance philosophy, when correctly implemented, increases machine life time, and reduces maintenance cost [reference].

As a framework, predictive maintenance has 4 levels [reference]. In level 1, visual inspection of machines or equipments are performed in order to assess any damage. In addition, equipment must fail before they are replaced, which can incur a high production cost. In level 2 and 3, machines characteristics such as vibration, temperature, electrical current, voltage etc are monitored continually or periodically depending on their criticality. This is called condition monitoring. In condition monitoring, the goal is to detect any change in machine normal behaviour, in order to detect failure as early as possible and schedule maintenance accordantly. Maintenance actions are performed periodically regardless of machine health condition. In level 4, big data and machine learning are the main driving forces in detecting failure and planning maintenance. At this level, maintenance action are not performed periodically but are planned according to machine health condition derived from the application of anomaly detection techniques. This reduce unnecessary maintenance action and significantly cut down maintenance cost as well as increasing machine life time.

According to PriceWaterhouseCoopers (PWC), one of the four largest auditing and consulting company in the world, a survey from 280 companies in Belgium, Germany and the Nederland, revealed that only 11% of companies have reached level 4 [4]. The application of level 4 requires collecting, saving and analysing large amount of data, from which maintenance decision can be made. Anomaly detection methods and more generally supervised and unsupervised learning methods are used in the analysis phase.

In supervised learning, based on available failure data from defect machine, a learning algorithm is trained to recognise the failure pattern in the data. This is sometime achieved by fitting the algorithm parameters to the data, which result in a model called a classifier or a regressor. A classifier is a model derived from a classification algorithm while a regressor is a model derived from a regression algorithm.

In machine learning, the failure data is called a labeled data because we can assign a categorical label such as fail or a numerical label such as 1 to specify the condition of the machine through the data measuring its characteristics. In the absence of labeled data, unsupervised learning methods can be apply to detect pattern in the data without prior knowledge to classify data regions as anomalous or not. One such technique is clustering algorithm where the input data is separated into subregions.

The procedure of classifying a data region as anomalous or not is called anomaly detection. An anomaly is defined as a pattern in the data, that does not conform to expected normal behaviour [3]. A general anomaly detection strategy will first detect normal behaviour, secondly set a boundary around the normal behaviour and finally declare any data out of the boundary as anomaly.

Several factor make this general approach of anomaly detection challenging: The notion of anomaly is different for different application domain and not all application domain have enough labeled data to train a supervise learning algorithm [3]. An extensive survey from [3] revealed that the majority of research have been focussing on simple anomalies while most application domain are faced with complex anomalies. There are mainly three types of anomalies: Point anomalies which are simple anomalies, contextual and collective anomalies which are complex anomalies [3]. When one data point in a time series is anomalous with respect to the other data points, we have a point anomaly. In a contextual anomaly, a data instance is anomalous relative to a context. For example the vibration of a machine might be very high if the load increase suddenly, and decrease when the load goes back to normal. But if the vibration increases monotonously regardless of the load, then we have a contextual anomaly. If a collection of related data instance is anomalous with respect to the entire data set, it is termed a collective anomaly [3].

In rotating machines, more than 40 % of anomalies can be attributed to bearing defect [references]. Figure 1 shows the failure statistics for rotating machines.
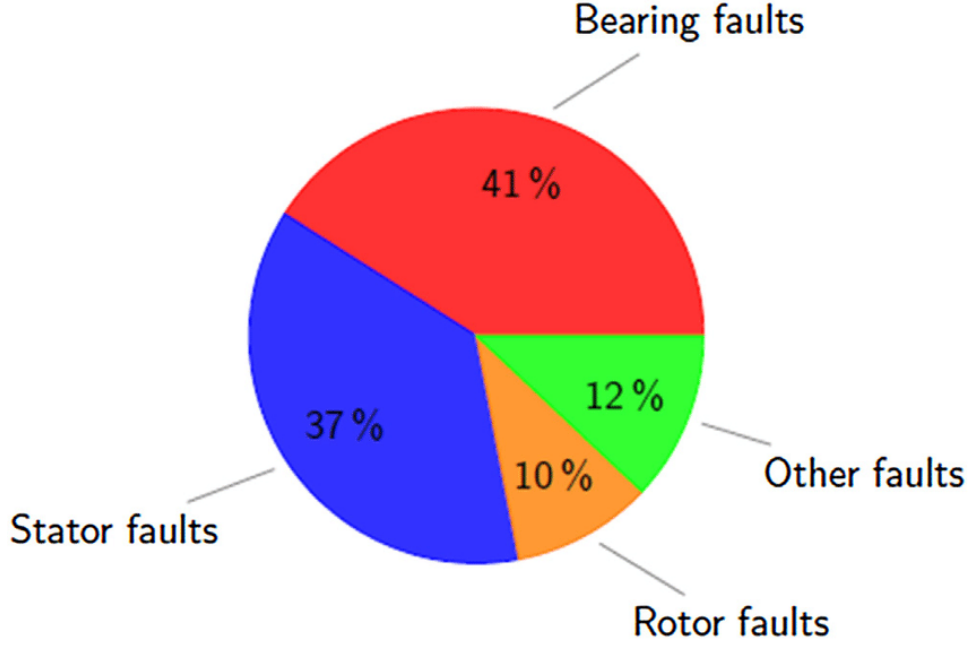
Figure 1: Defect statistics for rotating machines, taken from [reference]

In this project we present a mixed methodology to detect and predict bearing defects. The methodology consists of using signal processing for feature generation and data labelling, and machine learning for defects classification and failure prediction. For a given input dataset, the dataset is decomposed into it subcomponents or basis components. The basis components also called features or attributes are further used as input of a supervised learning algorithm for defect classification and failure prediction.

The signal processing methods used are Fourier transform, wavelet transform and Hilbert Huang transform. We focus on ensemble learning and feed forward neural network for classification. Furthermore we show that the back-propagation process in the feed forward neural network can be modelled by an ordinary differential equation, whose solution represents the path of the hidden and output layer weights.

This thesis is structured into six parts. To put the current work in prospective, chapter two outlines preview works, where machine learning and/or signal processing have been applied to anomaly detection for machines fault detection. Since the latter is the main goal, chapter three presents a general anomaly detection methodology, followed by a proposed methodology for predictive maintenance. The main methods used for anomaly detection, that is signal processing and machine learning are presented in chapter four and five respectively. The capstone of this work, to say the least, is revealed in chapter six in a form of three case studies, where we apply our proposed methodology to solve up to date anomaly

detection problems encounter in a wide range of industries. As in any work, we conclude this thesis by summarising what have been done. The summary includes the strengths and weaknesses of our methodology, as well as what could be done to improve upon it.

2 Literature review (TO DO!!))

3 Anomaly detection methodology and methods (TO DO!!)

3.1 General anomaly detection methodology (TO DO!!)

3.2 A proposed methodology for predictive maintenance (TO DO!!)

4 Signal processing methods

4.1 Overview

In this section, we present three signal processing methods to generate new features also called attributes, for anomaly detection. In the context of this work, a feature or attribute is a numerical sequence that represents the state of a system. For example, voltage and vibration measurement taken from a motor over time are time series that we call features. These two features contain information about the state of the motor, the latter being the system in this case. By state we mean the current health of the motor. The health of the motor is represented by a health index which is a numerical value that quantifies the overall condition of the motor.

Generating new features from existing one, is a common procedure for anomaly detection. For example, by using Fourier transform, we can generate the frequency spectrum of a vibration time series. The frequency spectrum is the set of all frequencies from each vibration component of the original vibration time series. The importance of the frequency spectrum, lies in the fact that if a motor or machine is anomalous, the anomaly will generate an extra component in the vibration time signal, and the corresponding anomalous frequency will be visible in the frequency spectrum. Due to Fourier transform limitations, other signal processing techniques such as wavelet and Hilbert Huang transform are alternatives for generating new features to deal with more complex time series for anomaly detection.

In the rest of this section, we give an exposé of three signal processing methods, namely: Fourier transform, wavelet transform and Hilbert Huang transform. For each transform, we cover the theoretical back bone in terms of mathematical constructs such as basis, vector space, orthogonality, existence, uniqueness, to name a few. Furthermore, these mathematical constructs will set the limitations as well as the strengths of each transform. Following the theoretical set up is a concrete application, where we generate new features and show how they can be used for anomaly detection.

4.2   Fourier Analysis

4.2.1   Theory

Fourier analysis is concerned with the general problem of functions approximation. The basic ingredients required to approximate a function are: a vector space, a basis, which is a subspace of the vector space and a mathematical operation or more generally a function such as an inner product that maps two vectors to a real number. If a vector space has an inner product, we say that the vector space is an inner product space.

The function approximation process goes like this: We select an arbitrary function f. We pick an appropriate vector space which we call V, such that $f \in V$. We define a subspace V_0 of the vector space V and construct an inner product on V_0, if it does not exist. Furthermore, we fine an appropriate basis of V_0. A basis of V_0 is a set of linearly independent vectors say φ_j in V_0, conveniently denoted by $\{\varphi_j\}_{j=0}^n$ that span V_0. This means that any vector in V_0 can be written as a linear combination of the basis vectors. Once we have all this in place, then the best approximation of the function f is its orthogonal projection in the inner product space V_0. Figure 2 shows an illustration of function approximation.
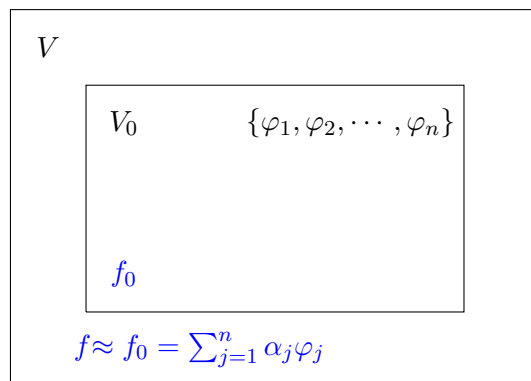

Figure 2: Illustration of a generic function approximation process of a function f into a subspace V_0 of a vector space V. φ_j are the basis function and α_j are real numbers for $j = 1, \cdots, n$.

Let have a closer look at each ingredient and define further properties that make Fourier analysis an appropriate tool for function approximation. We shall shortly define the concepts of vector space, basis, inner product, orthogonality and additional properties that are derived from them.

Definition 1 *A vector space V over a field F is a set with two operations $(\cdot, +)$ satisfying*

   1. Associativity: $(u+v)+w = u+(v+w)$, for any u,v, $w \in V$

6

2. *Commutativity: u+v = v+u, for all u,v ∈ V*

3. *Inverse: for any u ∈ V there exists an element -u ∈ V such that u + (-u) = 0*

4. *Identity element: $1 \cdot u = u$ where 1 is the multiplicative identity in he field F.*

5. *Distributivity: $a \cdot (u + v) = a \cdot u + a \cdot v$, for any $a \in F$, and $u, v \in V$*

6. *Compatibility $a \cdot (b \cdot u) = (a \cdot b) \cdot u$, for any $a,b \in V$ and $u \in V$*

7. *Scalar distributivity $(a + b) \cdot u = a \cdot u + b \cdot u$, for any $a,b \in F$, and $u \in V$*

The elements of the vector space V are called vectors and the element of the field F are call scalars.

An example of an inner product space is the L^2 space which encompasses square integrable functions. By definition, the set of all squared integrable functions on an interval say $[0,T]$ is denoted by $L^2[0,T]$. Any real function f in $L^2[0,T]$ satisfies

$$\int_0^T f(t)^2 \mathrm{d}t < \infty. \tag{1}$$

For two functions f and g in $L^2[0,T]$, we define their inner product as

$$\langle\, f, g \rangle = \frac{1}{T} \int_0^T f(t)g(t)\mathrm{d}t. \tag{2}$$

Equation (1) says that the total energy of f is finite. This suggest that the L^2 space is important for real world applications such as signal processing. In fact, we will see later in more detail that any function in L^2 has a Fourier expansion given in terms of linear combination of sinusoidal basis functions.

A legitimate question that comes in mind is, how are vectors organised in a vector space. One way to organise vectors in a vector space is to define them in terms of special set of vectors called basis, having well defined properties. Each vector will then be uniquely expressed in terms of the basis vectors.

Definition 2 *A basis is a set of vectors $\varphi_i, i = 1, \cdots, n$, covenantally written as*

$$\{\varphi_1, \cdots, \varphi_n\},$$

that spans the whole vector space and is linearly independent. Spanning the whole space means that for any vector $u \in V$ there exist unique scalars a_i where $i = 1, \cdots, n$, such that

$$u = a_1\varphi_1 + \cdots + a_n\varphi_n,$$

and linearly independent means that if

$$a_1\varphi_1 + \cdots + a_n\varphi_n = 0,$$

then

$$a_1 = \cdots = a_n = 0$$

With a given basis we can uniquely approximate a vector by decomposing its in terms of the basis vectors. However, a vector space can have several basis, giving rise to several approximations of the same vector. In mathematics and in science in general, one is often concerned with the optimum solution or the unique solution. In terms of an application, uniqueness allows one to pick the relevant physical solution. It turns out that, regardless of the basis, the orthogonal projection of a vector is its optimum approximation. Orthogonal projection is achieved by a special operation called inner product. Let formally define an inner product and show how its define orthogonality.

Definition 3 *An inner product on a real vector space V is a function*

$$\langle\,,\rangle : V \times V \to \mathbb{R}$$

that satisfies the following properties $\forall u, v \in V$

1. *Positivity: $\langle u, v \rangle > 0$,*

2. *Symmetry: $\langle u, v \rangle > 0 = \langle u, v \rangle$*

A vector space with an inner product is called an inner product space.

$$\vdots$$
$$\vdots$$

(NEED TO EXPAND MORE HERE AFTER THE DEFINITION))

Definition 4 *Suppose V is an inner product space.*

- *The vectors u and v are orthogonal if $\langle u, v \rangle = 0$*

- *A finite set of vectors $\{v_j\}_{j=1}^n$ are orthonormal if each vector have unit length with respect to a norm $||,||$, that is*
$$||v_j|| = 1$$
and v_i and v_j are orthogonal for $i \neq j$

$$\vdots$$
$$\vdots$$

(NEED TO EXPAND MORE HERE AFTER THE DEFINITION)

8

Definition 5 *Suppose V_0 is a finite-dimensional subspace of an inner product space V. For any vector $v \in V$, the orthogonal projection of v into V_0 is the unique vector $v_0 \in V_0$ that is closest to v, that is,*

$$||v - v_0|| = \min_{\forall w \in V_0} ||v - w||, \tag{3}$$

$\vdots$
$\vdots$

(NEED TO EXPAND MORE HERE AFTER THE DEFINITION)

Theorem 1 *Suppose V_0 is a n-dimensional subspace of an inner product space V and let $\{\varphi_j\}_{j=1}^n$ be an orthonormal basis for V_0, then the orthogonal projection of a vector $v \in V$ onto V_0 is given by*

$$v = \sum_{j=1}^{n} \alpha_j \varphi_j, \quad where, \quad \alpha_j = \langle\, v, \varphi_j \rangle \in \mathbb{R} \tag{4}$$

Proof 1 *(Need a proof here)*

From Theorem 1, We saw that a vector can be decomposed in terms of basis functions by orthogonal projection. The orthogonal projection ensures that the decomposition is optimum with respect to a chosen basis. In Fourier analysis, the basis functions are trigonometric functions of period 2π.

Theorem 2 *The set of functions*

$$\left\{ \cdots, \frac{\cos(2x)}{\sqrt{\pi}}, \frac{\cos(x)}{\sqrt{\pi}}, \frac{1}{\sqrt{\pi}}, \frac{\sin(x)}{\sqrt{\pi}}, \frac{\sin(2x)}{\sqrt{\pi}}, \cdots \right\} \tag{5}$$

is an orthonormal set in $L^2([-\pi, \pi])$ space

Proof 2 *(Need a proof here)*

(NEED TO EXPAND MORE HERE AFTER THE THEOREM)

Definition 6 *The Fourier series of a periodic function g is given by the expansion*

$$g(x) = a_0 + \sum_{k=1}^{n} a_k \cos(kx) + b_k \sin(kx) \tag{6}$$

where $a_k, k = 0, \cdots, n$ are called the Fourier coefficients of the function g

$\vdots$
$\vdots$

(NEED TO EXPAND MORE HERE AFTER THE DEFINITION)

9

Theorem 3 *if*

$$g(x) = a_0 + \sum_{k=1}^{n} a_k \cos(kx) + b_k \sin(kx) \tag{7}$$

then

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x) \mathrm{d}x \tag{8}$$

$$a_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x) \cos(kx) \mathrm{d}x \tag{9}$$

$$b_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x) \sin(kx) \mathrm{d}x \tag{10}$$

⋮
⋮

Proof 3 *(need a proof here) We use the orthogonality property to derive the Fourier coefficients.*

⋮
⋮

   TO BE CONTINUED

4.2.2   Application

In this approach we use Fourier transform to detect defect frequency such outer race defect frequency and inner race defect frequency in the envelop spectrum. In this case the envelop spectrum is the new feature generated from the original vibration data. The bearing type used are Rexnor ZA-2115. For this type of bearing, the constant defect frequency (cdf) for ball pass frequency outer race defect is 0.1182 and the constant defect frequency for ball pass frequency inner race defect is 0.1484 [reference]. According to Rexnord product engineering group [reference], to find the defect frequency in Hz we multiply the constant defect frequency (cdf) by the rotational speed of the bearing (here 2000 RPM), to find the defect frequency in Hz. Note that the defect frequency is computed from the geometry of the bearing, which means that for a specific bearing, this is a constant value.

Figure 3 shows the geometry of a bearing, from which bearing defect frequency can be computed. Figure 4 illustrates how Fourier transform can decompose vibration data into its frequency components. Each frequency component which are trigonometric functions are characterised by constant frequency and amplitude. In a defect bearing, the defect frequency will be coupled with a relatively high amplitude.
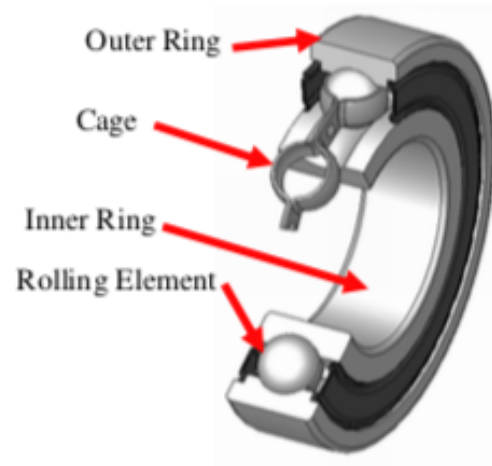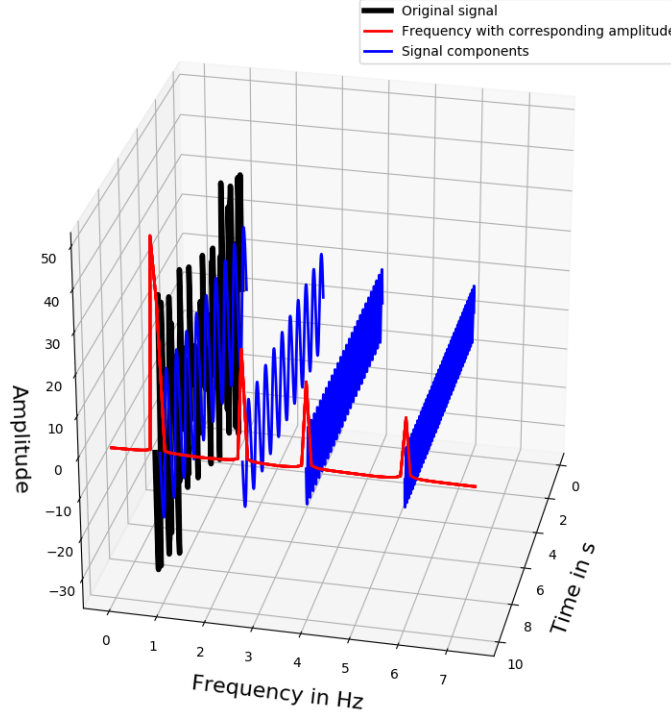
Figure 3: Geometry of a bearing

Figure 4: Illustration of using Fourier transform to decompose a time signal to its frequency components

Next, we compute the envelop spectrum for each vibration signal by using the FFT method (Fast Fourier Transform). The envelop spectrum is used to detect early sign of failure. Recall that a vibration signal can be decomposed into its sub components, where each sub component is characterised by its frequency and its amplitude. Early sign of failure can be seen in the high frequency low amplitude sub component. As the failure becomes more pronounced, it becomes visible in the low frequency high amplitude sub component. Figure 5 shows the raw vibration signal, the corresponding envelop spectrum and the ball pass frequency outer race defect.
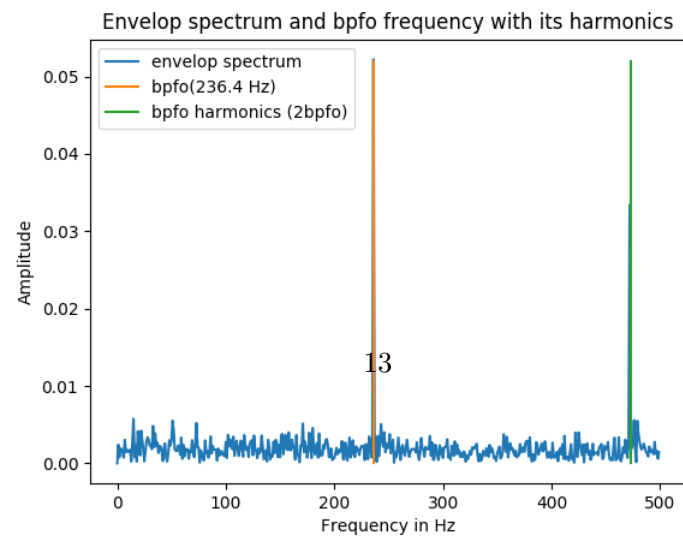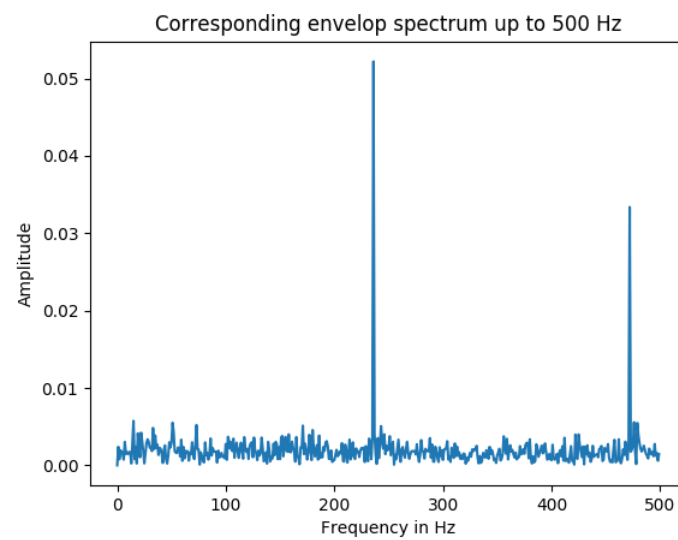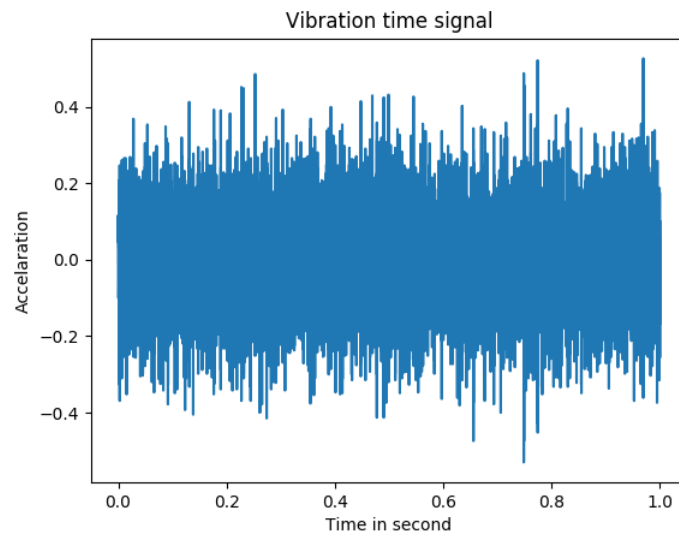
Figure 5: Vibration time signal

By computing the envelop spectrum for each time signal and extracting the amplitude of a defect frequency we can visualise the severity of a defect. Figure 6 and **??** show the amplitude of outer race defect and inner race defect for four bearings and three bearings respectively. In Figure 6 we can observe that bearing number one is affected severely by ball pass frequency outer race defect. Figure 7 shows that bearing number three has ball pass frequency inner race defect.
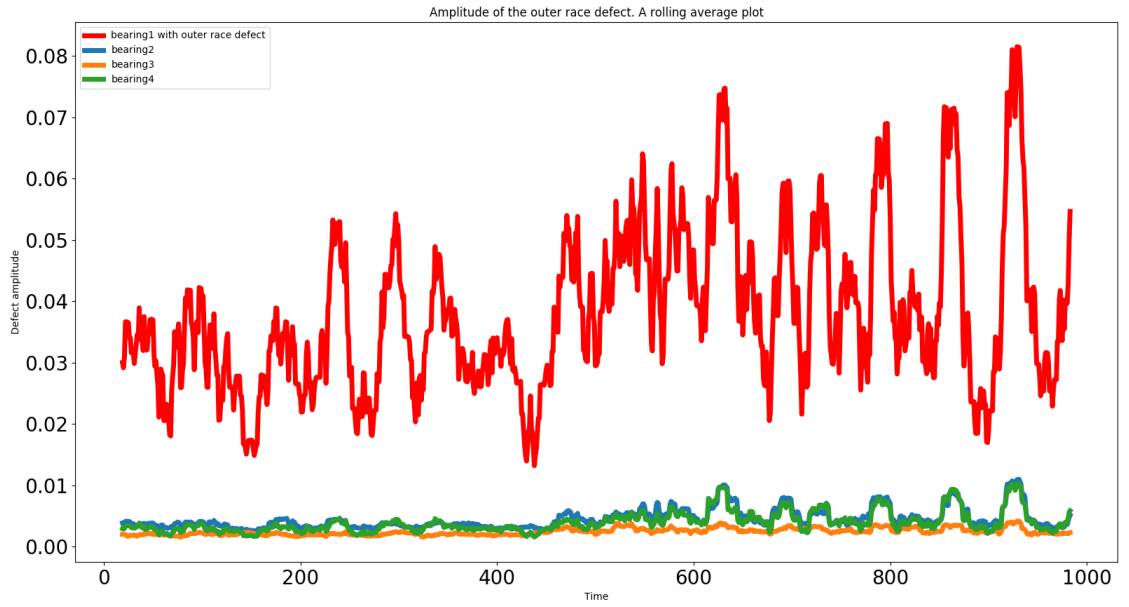
Figure 6: Ball pass frequency outer race detection
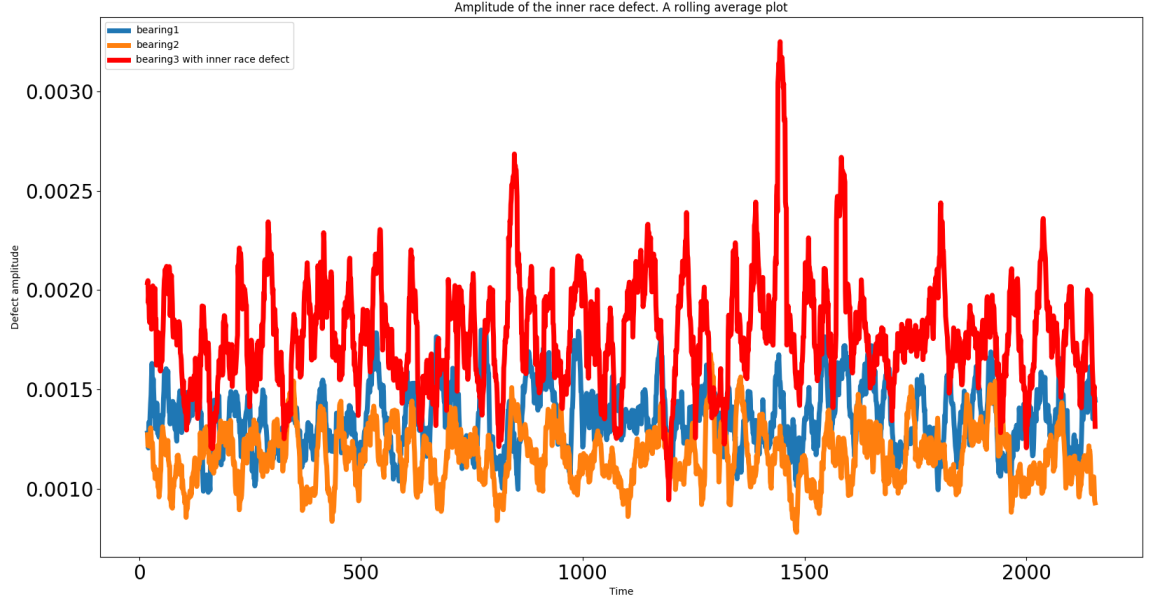
Figure 7: Ball pass frequency inner race detection

4.3 Wavelet transform

4.3.1 Theory

4.3.2 Application

In the wavelet transform we generate two extra features from the vibration time signal, namely the discrete detailed coefficient cD and the approximate coefficient cA. The detail coefficient cD represents the hight frequency component of the vibration time signal and the approximate coefficient cA represents the low frequency component. For the mother wavelet we use Daubechies 20 or db20 shown in Figure 8

15
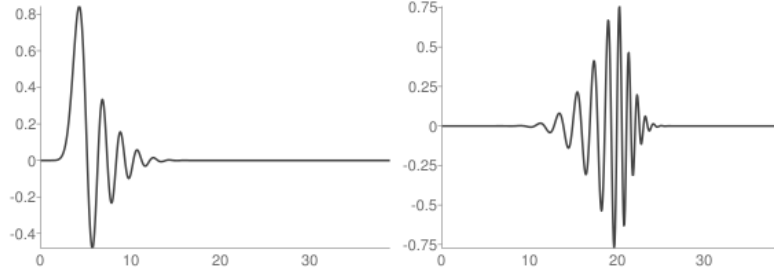
Figure 8: The scaling function φ (left) and the wavelet function Ψ (right)

Once we have the two extra features, we compute the dissimilarity between a reference sample and subsequent samples for each feature. This process generates a set of points (x, y) that represent the health index of each sample. From Figure 9 and 10, we can observe that bearing number four and bearing number three suffer from ball pass frequency outer race and ball pass frequency inner race defect respectively. We can also observe that a bearing can go through three main stages:

1. A healthy stage characterised by a low health index

2. A warning stage characterised by an increasing health index

3. An alarm stage characterised by a high health index
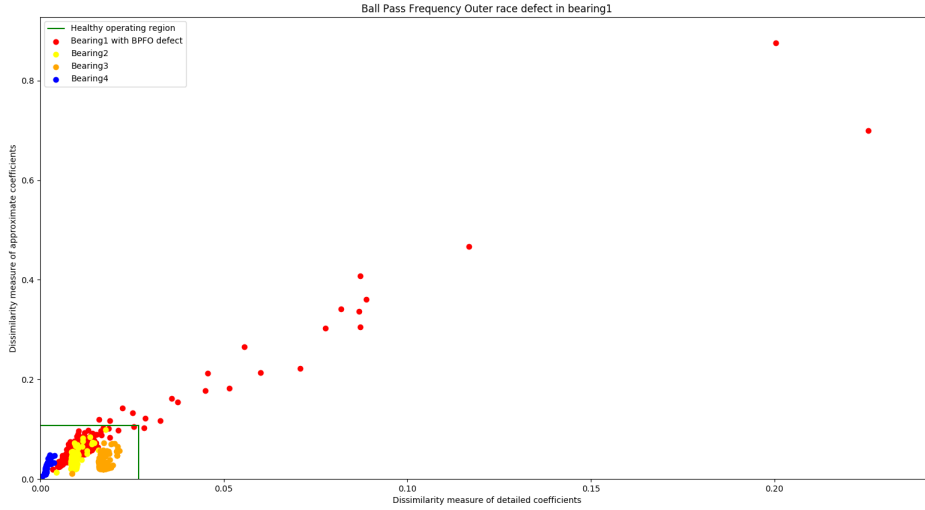
Figure 9: Ball past frequency outer race defect detection from wavelet transform
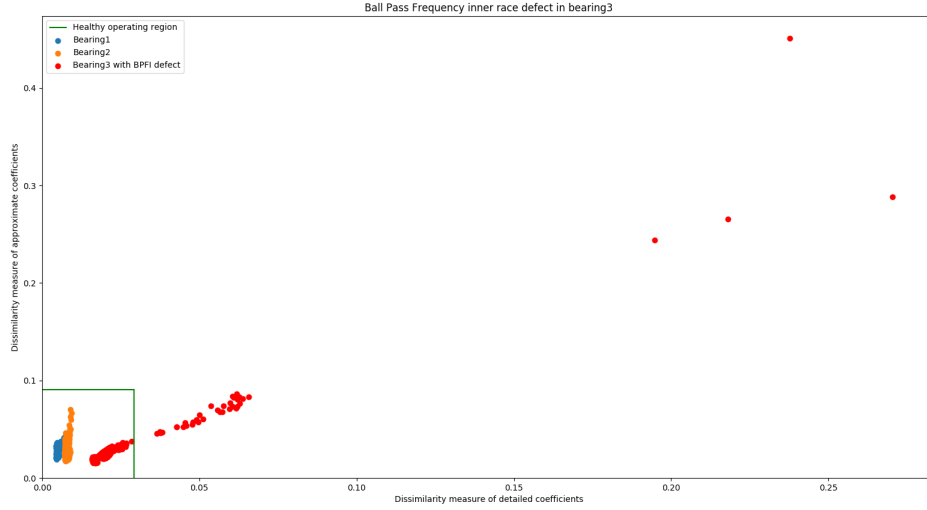
Figure 10: Ball past frequency outer race defect detection from wavelet transform

In the alarm stage, as the degradation becomes more severe, the distance between the points increases.

4.4   Hilbert Huang transform

4.4.1   Theory

The Hilbert-Huang transform is a data decomposition methods that consists of decomposing data in an adaptive fashion. Adaptivity means that rather than imposing an a priory basis such as trigonometric functions, a posteriori basis functions are derived from the data itself [1]. In doing so, the method deals better with nonlinearity and non stationarity which are inherently present in real world data.

This method gives an alternative approach of time-frequency-energy paradigm by using Hilbert spectral analysis and the so call empirical mode decomposition (EMD) to express the nonlinearity and the non stationary in data with instantaneous frequency and instantaneous amplitude [1].

The empirical mode decomposition (EMD) originated from the quest of functions that can be expressed by a time-frequency-amplitude expression, such that the frequency is physically meaningful. Consider a time series $x(t)$. Its Hilbert transform H(t) is given by

$$H(t) = \frac{1}{\pi} P \int_{\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau, \tag{11}$$

where P is the Cauchy principal value. The corresponding time-frequency-amplitude function of $x(t)$ is the analytical function

$$z(t) = x(t) + iy(t) = a(t)e^{i\theta(t)}, \tag{12}$$

where the instantaneous amplitude $a(t)$ and phase $\theta(t)$ can be computed by

$$a(t) = \sqrt{x(t)^2 + y(t)^2} \tag{13}$$

$$\theta(t) = \tan^{-1}\left(\frac{y(t)}{x(t)}\right). \tag{14}$$

Furthermore, the instantaneous frequency $w(t)$ can be derived from the phase $\theta(t)$ as

$$w(t) = \frac{d\theta}{dt}. \tag{15}$$

By setting

$$f(t) = \frac{y(t)}{x(t)},$$

the expression of the instantaneous amplitude $w(t)$ in (15) can be expanded as

$$w(t) = \frac{f'(t)}{1 + f(t)^2} = \frac{y'(t)x(t) - y(t)x'(t)}{x(t)(x(t) + y(t)^2)}. \tag{16}$$

The instantaneous frequency $w(t)$ using the Hilbert transform is not always physically meaning. For example for an arbitrarily function, the instantaneous physical frequency values should be positive. However this is not always the case.

For example if

$$f(x) = \cos(ct) + d \tag{17}$$

where c and d are constants, the instantaneous frequency is given by

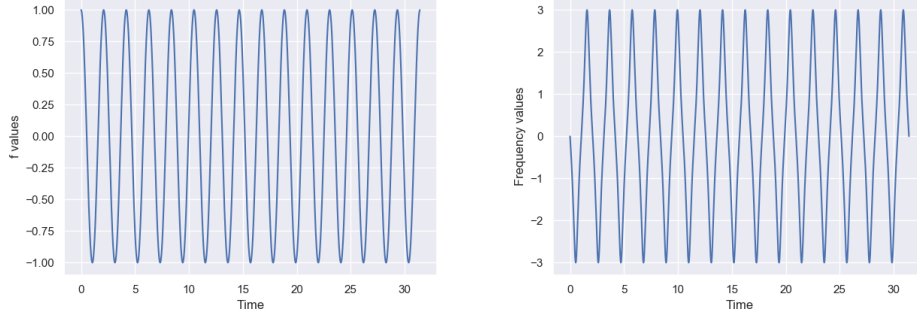$$w(t) = \frac{-c\sin(ct)}{1 + (\cos(ct) + d)^2} \tag{18}$$

Figure 11: function f and its corresponding frequency

From Figure 11, we see that the instantaneous frequency takes negative values, which is not physically meaning full.

To circumvent this, the Hilbert-Huang transform offers a methodology to obtain from an arbitrarily function or time series $x(t)$ a set of finite subcomponents whose instantaneous frequency are physically meaningful. This methodology let to the empirical mode decomposition.

The necessary condition for obtaining a physical frequency is that $x(t)$ satisfies the approximate local envelope symmetry condition [2].
This condition is expressed in the empirical mode decomposition (EMD) such that an arbitrarily time series $x(t)$ can be decomposed by a sifting process into intrinsic mode function c_i

$$x(t) = \sum_{i=1}^{n} c_i + r_n \tag{19}$$

where the c_i satisfies the approximate local envelope symmetry condition

$$SD_k = \frac{\sum_{t=0}^{T}}{\sum_{t=0}^{T}} < \epsilon \tag{20}$$

where ϵ is a small predefined real number.

4.4.2 Application for bearings fault detection

we consider a vibration signal with sample frequency of 20000Hz rotating speed of 2000 RPM
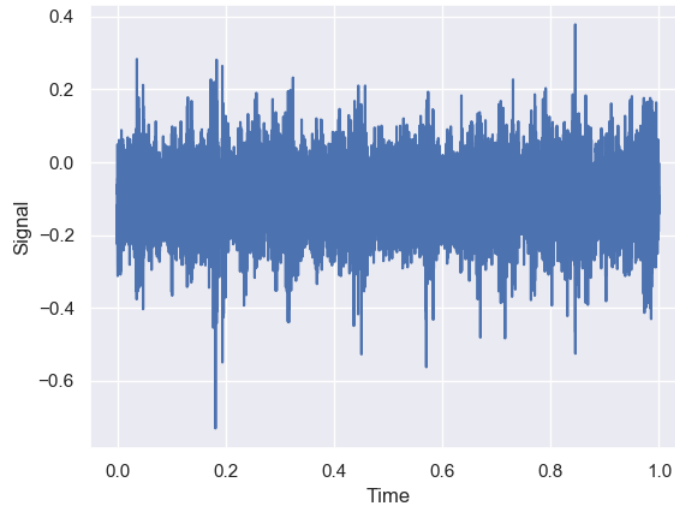
19

Figure 12: Vibration signal of 1 second snapshot

After applying the empirical mode decomposition on the vibration data from figure 12 we get sixteen intrinsic mode functions
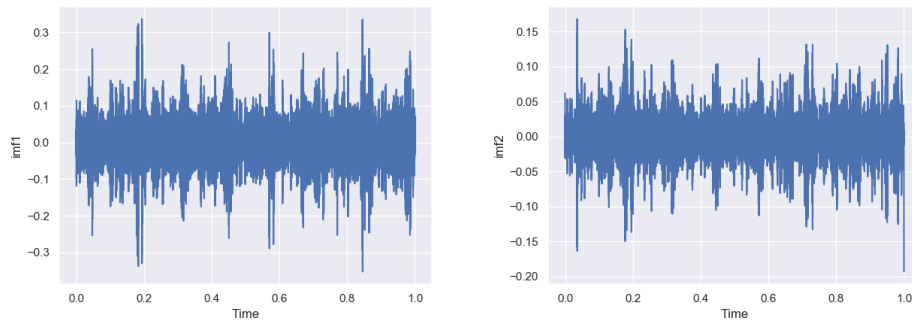

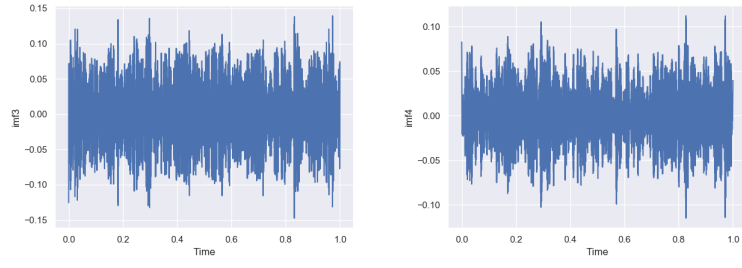Figure 13: 1th and 2nd intrinsic mode function (imf)

20

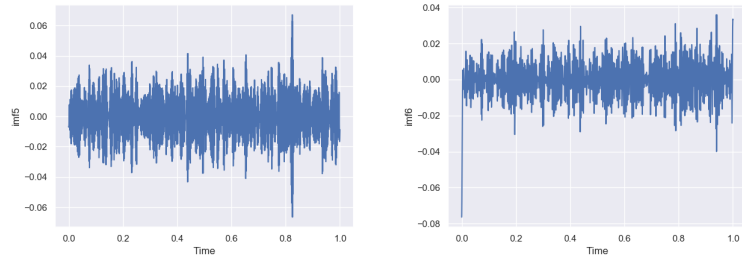Figure 14: 3rd and 4th intrinsic mode function (imf)

Figure 15: 5th and 6th intrinsic mode function (imf)
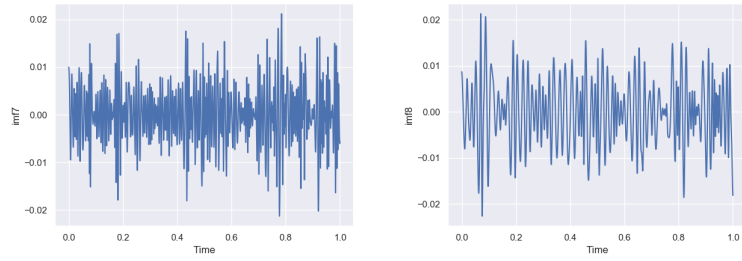

Figure 16: 7th and 8th intrinsic mode function (imf)
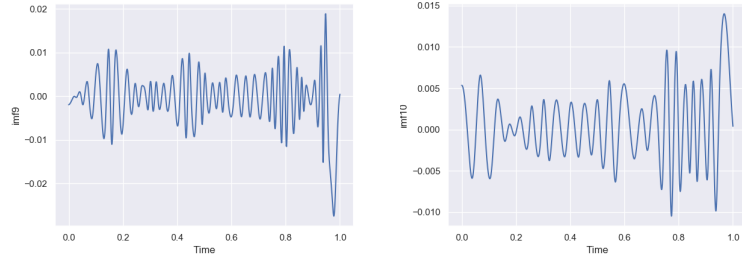
21

Figure 17: 9th and 10th intrinsic mode function (imf)
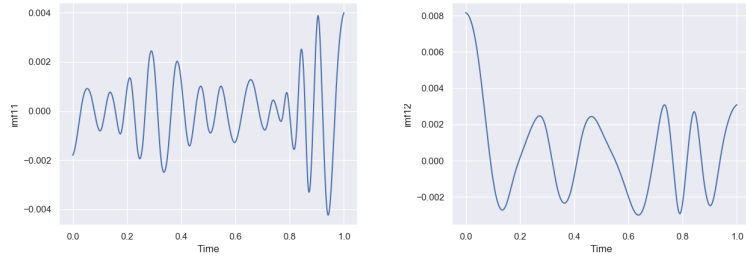

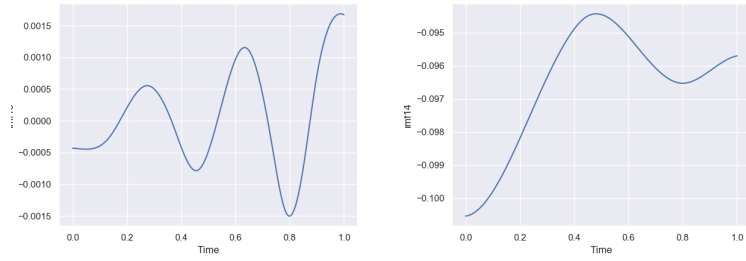Figure 18: 11th and 12th intrinsic mode function (imf)

Figure 19: 13th and 14th intrinsic mode function (imf)

5 Machine learning methods (TO DO!!)

5.1 Overview (TO DO!!)

6 Application of signal processing and machine learning: Case study (TO DO!!)

6.1 Overview (TO DO!!)

7 Conclusion (TO DO!!)

References

[1] Norden E. Huang and Zhaohua Wu *A review on Hilbert-Huang Transform: Methods and its Application to Geophysical Studies.* Review of Geophysics, 2008

[2] Norden E. Huang, Zheng Shen, Steven R. Long, Manlic C. Wu, Hsing H.Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung and Henry H. Liu *The Empirical mode decomposition and the Hilbert Spectrum for nonlinear and non-stationary time series analysis* Proc. R. Soc. Lond. A (1998) 455, 903:995.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. *Anomaly detection: A survey* Proc. R. Soc. Lond. A (1998) 455, 903:995.

[4] Michel Mulders and Mark Haarman *Predictive maintenance 4.0. Predict the unpredictable, June 2017*

[5] Hui Li, Yuping Zhang and Haiqi Zheng *Hilbert-Huang transform and marginal spectrum for detecting and diagnosis of localized defects in roller bearings* Journal of Mechanical Technology, 2007