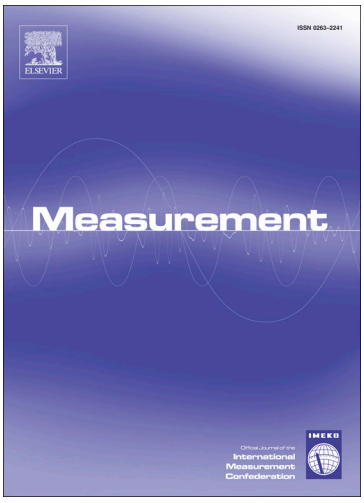
Semi-automated diagnosis of bearing faults based on a hidden Markov model of the vibration signals

Ge Xin, Nacer Hamzaoui, Jérôme Antoni



Semi-automated diagnosis of bearing faults based on a hidden Markov model of the vibration signals

Ge Xin, Nacer Hamzaoui, Jerome Antoni

PII: S0263-2241(18)30429-9
DOI: https://doi.org/10.1016/j.measurement.2018.05.040
Reference: MEASUR 5541

To appear in: *Measurement*

Received Date: 15 December 2017
Revised Date: 4 April 2018
Accepted Date: 9 May 2018

Please cite this article as: G. Xin, N. Hamzaoui, J. Antoni, Semi-automated diagnosis of bearing faults based on a hidden Markov model of the vibration signals, *Measurement* (2018), doi: https://doi.org/10.1016/j.measurement.2018.05.040

Semi-automated diagnosis of bearing faults based on a hidden Markov model of the vibration signals

Ge Xin[a,b], NacerHamzaoui[b], Jerome Antoni[b,*]

[a] School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, PR China

[b] Laboratoire Vibrations Acoustique, Univ Lyon, INSA-Lyon, LVA EA677, F-69621 Villeurbanne, France

Abstract

Due to its practical importance, the diagnosis of rolling element bearing has attracted constant interest in the scientific community. At the incipient stage of a failure, the measured vibration signal typically consists of a series of repetitive transients immerged in background noise. Although they are usually carried in high frequency bands due to the high stiffness of bearings, they are fairly weak compared with surrounding noise and other interfering signals. In addition, taking random slips and fluctuations into account, the transients produced by impacts are not strictly periodic but rather tend to be random cyclostationary. This makes the diagnosis of rolling element bearing quite challenging and, consequently, various signal processing techniques have been developed for either the detection, the identification or the extraction of the fault, whose combination asks for a high level of expertise of the user. The aim of this paper is to address all these objectives at once, in the same algorithm, by proposing a semi-automated method that requires the setting of only one parameter. It is rooted on a probabilistic model, in the form of a mixture of Gaussians, endowed with a hidden variable that indicates the occurrence of impacts. The method is shown to be optimal for detection in the Neyman-Pearson sense, it returns an envelope spectrum comparable to the best that can be obtained by other means – which often require a careful pre-filtering step – from which fault frequencies can be identified, and it eventually returns the fault signal from which subsequent severity/health indicators can be computed. There is almost no demand on the user's expertise (apart from setting the frequency resolution), even though the method does not address the decision part. The performance is investigated on synthetic signals and its robustness is also verified on several vibration signals measured on test-rigs. Results are found superior or at least equivalent to those of the conventional semi-automated method based on the fast kurtogram in combination with the envelope analysis.

Keywords: Automated diagnosis, repetitive transients detection, repetitive transients extraction, bearing faults, hidden Markov model, mixture of Gaussians, cyclostationary signals, non-stationary operating conditions.

Highlights:

A mixture of Gaussians model is introduced for rolling element bearing vibrations.

The model allows semi-automated diagnosis of bearing faults without need for pre-processing.

It is optimal in the Neyman-Pearson sense for detecting repetitive transients.

It allows full-band reconstruction of transients in the time domain.

It applies under general assumptions, including time-varying operating conditions.

[*]Email address : jerome.antoni@insa-lyon.fr (J. Antoni), xinge@bjtu.edu.cn (G. Xin).

1. Introduction

Rolling element bearings are common but fragile components used in various types of mechanical systems, and thus a foremost cause of machinery breakdown. Typical bearing defects are caused by crack, breakage, spall or uneven wear (pitting, scuffing, abrasion, erosion), often located on the matting surface of the inner race, the outer race or the rolling elements. As the rolling elements strike a defect, a series of vibration transients occur at a specific rate corresponding to the bearing "characteristic frequency" or "fault frequency". At the incipient stage, each transient resembles a damped impulse response with frequency content specified by the excited structural resonances. Due to the load distribution, the series of transients are additionally amplitude modulated by the passing period into and out of the load zone. Bearing fault signatures have been well-investigated and their characteristic frequencies are listed from given kinematic and geometric parameters [1-2]. Nevertheless, bearing fault signals are often masked by numerous extraneous sources of vibration, simply referred to hereafter as "background noise", such that signal processing is necessary to recover the diagnostics information. Another difficulty arises from machines which do not operate in stationary conditions, such as wind turbines or crushers, which ask for even more advanced data processing.

The diagnosis of rolling element bearings – as understood in the present paper – includes three main steps, whose chronological order actually coincides with their degree of difficulty: 1) detection of the fault, 2) identification of the fault and 3) extraction of the fault signal. Since they correspond to different goals, they are usually addressed by different signal processing tools.

The detection of the fault is usually formulated as a problem of detecting the presence of a weak signal in background noise, ideally in the form of statistical test. In the case of bearing faults, the signal has a complex statistical structure – it is actually cyclostationary, i.e. non-stationary random with periodic statistics [3] – which sometime makes this task challenging. This step has nourished a vast literature on a variety of methods which nowadays probably culminate to their asymptote in terms of performance. Some typical tools are time-frequency analysis[4-6], wavelet analysis [7-9], the stochastic resonance [10,11], morphological analysis [12], the spectral kurtosis and the fast kurtogram [13], the spectral correlation [14,15], sparsity [16], etc. It must be said that in most of these methods the detection of a fault is done visually; few of them actually tackle the problem as a statistical test in terms of probabilities of detection and of false alarm.

Next, the identification of the fault mainly consists in estimating the fault frequency and associating it to a given component in the machine. The prevailing method in the modern literature is surely the squared envelope spectrum (SES). The envelope spectrum is a mean to demodulate a non-stationary signal – of possible random nature, in particular cyclostationary– and identify periodic modulations related with the bearing characteristic frequencies [17]. In its most frequent version, the envelope is estimated by squaring the signal – or better, the modulus of the analytical signal – and its Fourier spectrum is then computed. The fact that the SES contains most of the diagnostic information can be justified theoretically by modelling the bearing fault signals as cyclostationary [18]. It is noteworthy that, in the cyclostationary framework, the SES can also be introduced as a statistics for fault detection [19]. However, the SES of the raw signal is rarely a good diagnostic indicator when used without pre-processing, because it is highly sensitive to the presence of background noise and other interfering components [18-22]. The conventional way to cope with this difficulty is to determine an optimal frequency band which maximizes the signal-to-noise ratio (SNR) – i.e. the energy of the transients with respect to that of the background noise – before enveloping (incidentally, this is also theoretically required to properly define the "envelope" of a signal [23]). This issue has been well addressed and has led to the use of the SES in combination with different pre-processing tools apt to identify the optimal demodulation band, for instance rooted on the spectral correlation or, most frequently, on the spectral kurtosis – computed with the fast kurtogram– as an effective measure of the 'impulsiveness' hidden in a signal [1, 24, 25].The use of the kurtosis at this stage is further supported by the fact that it is related to the sum of the peaks in the SES divided by the zero-

frequency SES [26]. A recent extension of the kurtogram, with similar goal but refined properties, is the infogram [27-29]. Other approaches – possibly used in conjunction with the latter ones– are based on first denoising the signals by various techniques before computing the SES [30, 31].

Although not strictly necessary for diagnosis, the final step consisting in extracting the fault signal is useful for assessing the severity of the fault and for designing health indicators for trending and for prognostics. From the signal processing point of view, this is yet a different task. One way to formulate the problem is as a blind deconvolution task, where the impulses due to the contact force on the defect are to be recovered from the measurements without knowledge of the impulse response [32-34]. A different and probably less ambitious approach is to recover the transients (i.e. the responses to the impulses) as if they had been observed in the absence of background noise [35-40].The former solutions actually require the use of sophisticated signal processing methods. A suboptimal solution justified by its simplicity is to extract an estimate of the fault signal by bandpass filtering with the optimal filter used in the computation of the SES.

This brief tour of the state-of-the-art shows that bearing diagnosis often involve a combination of different signal processing methods which require a high level of expertise from the user. As a consequence, one of the current challenges is to develop a complete standalone diagnosis methodology that can be run without (or very limited) intervention of the user. This is referred to as automated diagnosis, or "semi-automated" to highlight the fact that even though the processing is fully automatic, the final decision is still taken by the end-user upon interpretation of the results. (Semi-) automated diagnosis of rolling element bearings has been addressed from two main points of view. The first one is by means of classification algorithms borrowed from the domains of data mining and machine learning. Although a large body of literature exists on the subject [41-47], it seems that the majority of the proposed solutions are confined to apply to a specific installation – on which a classifier has been trained –and cannot be transposed to data recorded on other systems which have not been incorporated in the learning set. The second point of view proceeds from the signal processing side and ambitions a greater generality than the first one, yet it also appears much more challenging. There is indeed a very limited number of publications on the subject and they rarely address all the aspects of the diagnosis methodology [48-54].One methodology which has proved quite successful is described in Refs. [1, 2, 55] and has been further elaborated in Refs. [56, 57]. It consists in first pre-processing the signal, typically by whitening, in order to enhance the presence of low-energy transients. Next, an optimal band for demodulation is found from the fast kurtogram, from which the SES is computed. Finally, the fault signal is estimated by bandpass filtering in the optimal band returned by the kurtogram. In its trimmest version (where whitening is achieved by cepstral editing [56]), the whole process requires only one parameter to set which is the decomposition depth of the kurtogram. In many instances, the latter can actually be set by default to a predefined value. More sophisticated versions essentially improve the pre-processing step by using other whitening schemes [58, 59] or blind deconvolution [60]. Eventually, statistical tests can be designed on the so-obtained SES by following the lines of Refs. [14, 61-63] in order to automatically detect a fault with a given probability of false alarm.

The aim of the present work is to proceed with the objective of semi-automated diagnosis by using more advanced signal processing tools, while still simplifying the pre-processing step and keeping the number of tuning parameters to its minimum. A solution is proposed which addresses at once, in the same algorithm, the three goals of fault detection, fault identification and fault extraction. It is rooted on a short-time-Fourier-transform (STFT) representation of the signal and therefore requires only one parameter setting corresponding the spectral resolution. As compared to the aforementioned methodology (whitening + fast kurtogram + SES + filtration), it has the advantage of not requiring any pre-processing (e.g. before computing the STFT), of being optimal for detection – in the Neyman-Pearson sense – and of extracting a full-band version of the fault signal. It is also interesting as such since it achieves similar goals from a different way of processing the data and therefore offers a methodological diversity that is often desirable to improve the robustness of a diagnostic system.

The use of the STFT is justified because it captures well the time-frequency structure of a series of transients produced by a bearing fault. Since the objective is not to arrive at a *visual* detection, there is no critical limitation to expect from the uncertainty principle – e.g. as compared to other time-frequency representations, with better time-frequency localization, which have been advocated for diagnosis [5] – as long as the STFT returns the whole information contained in the analyzed signal, that is as long as it is *invertible*. In addition, the STFT is a linear transform that can be computed by means of very efficient algorithms and is therefore well suited to semi-automated diagnosis. The novelty of the present work is to endow the STFT coefficients of a bearing signal with a probabilistic model which switches between states: one where only background noise is observed and one where it is superposed with the occurrence of a transient. The label of the state is represented by a random latent variable which is encoded into a hidden Markov model (HMM). Since the STFT coefficients quickly tend in distribution to a complex-valued Gaussian in virtue of the Central Limit theorem [64], this is eventually formulated as a Gaussian mixture model. It thus happens that the value of the latent variable is related to a generalized likelihood ratio (LLR), from which a detection test can be designed that is optimal in the Neyman-Pearson sense – i.e. which maximizes the probability of detection of a fault for a given probability of false alarm. Besides, the Fourier transform of the LLR provides a spectrum which is in all point similar to the SES, yet obtained automatically without need for careful pre-filtering. Finally, after the GMM has been identified, the corresponding faulty state can be extracted as a byproduct and the corresponding time signal recovered by inverting the STFT. It is highlighted that the fault signal is extracted in full-band, contrary to the output of the fast-kurtogram. Ultimately, it is shown that the proposed approach is general enough to deal with non-stationary operating conditions.

It is noteworthy that HMMs have been proposed in 1970s as statistical modelsfor time series and so far they have been applied in a wide range of fields including speech recognition, computer vision, pattern recognition and many other areas. Numerous works based on HMMs have been reported in fault diagnosis during the last decade [65-74], however they are mainly concerned with the use of HMMs as classifiers [65-68, 74-75]and not for modelling the vibration signal itself as proposed in the present work.

The paper is organized as follows. Section 2 first introduces the probabilistic model based on the STFT decomposition and next explains the inference of the model parameters by means of the EM (expected-maximization) algorithm. Section 3 provides the semi-automated diagnosis methodology that addresses the issues of optimal fault detection by designing a generalized likelihood ratio test (GLRT), of fault identification by means of the Fourier spectrum of the LLR and of fault extraction by means of a time-varying filter obtained from inverting the STFT. Section 4 addresses the important issue of parameter settings and of algorithm initialization which is hereafter verified by synthetic signals in section 5. Finally, section 6 validates the proposed methodology on several vibration signals and compare it with the reference methodology based on the (whitening + fast kurtogram + SES + filtration) sequence.

2. Probabilistic model

This section introduces the probabilistic model, its corresponding assumptions, and the inference of its parameters by the EM algorithm.

2.1. Signal model and STFT decomposition

Let $y(t)$ denote the measured signal in the time domain, $x(t)$ the part that contains the diagnostic information and $n(t)$ the background noise. By definition, the "informative signal" $x(t)$ and the "noise" $n(t)$ are assumed mutually independent. The measured signal is thus expressed by the additive model

$$y(t) = x(t) + n(t). \tag{1}$$

Although the background noise actually comprises multiple components, it intervenes in model (1) as a global noise process, (e.g. with an equivalent covariance matrix). It is noteworthy that its exact probability distribution is not needed at this stage, yet a fair (and widely accepted) assumption is to model it as stationary. In contrast, since $x(t)$ represents the fault signal in its early stage, it is well modelled by a series of impacts that repetitively excite resonances of the bearing and of its receiving structure, thus leading to successive damped impulse responses. Having a localized signature both in time and in frequency, such transients are well captured in a time-frequency decomposition, on the contrary to the stationary background noise $n(t)$which is spread all over the time-frequency plane. Although several time-frequency decompositions are possible, the proposed approach only requires an invertible one. The STFT meets this property while being associated with efficient algorithmic implementations. In addition, due to the Central Limit Theorem applied to the DFT (discrete Fourier Transform), the coefficients of the STFT quickly tend in distribution to a complex-valued Gaussian [64], an assumption that will substantially simplify the probabilistic model introduced hereafter. It is noteworthy that the aim is to *decompose* the signal without loss of information in the time-frequency domain in which it will be processed; therefore, considerations such as the uncertainty principle will not matter because they are relevant to time-frequency representation (i.e. *visual* analysis), which is not of concern here.

The STFT of signal $y(t)$ over a time interval of length N_w is defined as

$$Y(i,f_b) = \sum_{m=0}^{N_w-1} w[m]\cdot y[iR+m]\cdot \mathrm{e}^{-j2\pi f_b \frac{iR+m}{F_s}} \quad (2)$$

where $w[m]$ denotes a positive and smooth N_w-long data-window which truncates a segment of the L-long signal $y(t)$ at time datum i ($i = 1, \ldots, N$, $N = \mathrm{floor}[(L-N_w)/R+1]$) with window shift R ($1 < R < N_w$) and where $f_b = b\cdot\Delta f$ denotes the frequency (from 0 to $F_s/2$) with frequency resolution$\Delta f = F_s/N_w$and bin index$b$ = 1, …, N_f with $N_f = N_w/2+1$.

It is noteworthy that $Y(i,f_b)$ is related to the "instantaneous complex envelope" of the signal described in both time and frequency. More precisely, its squared magnitude reflects the energy flow which is mapped by time index i and frequency f_b centered in a narrow frequency band Δf [76].

The next subsection introduces a two-state HMM to account for the different probability distributions of the STFT coefficients depending on whether a transient is present or not.

2.2. Hidden Markov model

Hereafter, the STFT coefficients $Y(i,f_b)$are collected for all frequency bins at a given time instant *i* in a column vector $\mathbf{Y}(i) = \left[Y(i,f_1)\ldots Y\left(i,f_{N_f}\right)\right]^T$. Next, $\mathbf{Y}(i)$ is represented by a linear combination of K components – whose events are assumed mutually exclusive – denoted by $\mathbf{X}^k(i) = \left[X^k(i,f_1)\ldots X^k\left(i,f_{N_f}\right)\right]^T$ and contaminated by a noisy component $\mathbf{N}(i)$. All components in the model are allowed to have different probability distributions and are controlled by a vector of latent variables, $\boldsymbol{\zeta}(i) = [\zeta^1(i)\ldots\zeta^K(i)]^T$, each of which acting as a switch taking only values 0 or 1. Thus, the proposed model reads

$$\mathbf{Y}(i) = \mathbf{X}(i)\boldsymbol{\zeta}(i) + \mathbf{N}(i) \quad (3)$$

where $\mathbf{X}(i) = [\mathbf{X}^1(i)\ldots\mathbf{X}^K(i)]$ is a matrix consisting of K column vectors $\mathbf{X}^k(i)$ and $\mathbf{N}(i) = \left[N(i,f_1)\ldots N\left(i,f_{N_f}\right)\right]^T$.

a) *One component case*

Let us first consider the simplest model with only one component of interest, that is

$$\mathbf{Y}(i) = \zeta^1(i)\mathbf{X}^1(i) + \mathbf{N}(i). \tag{4}$$

Since the latent variable $\zeta^1(i)$ can take only two values, 0 or 1, with a given probability, say π, it follows (by definition)the Bernoulli distribution,$\zeta^1(i) \sim Bernoulli(\pi)$:

$$\begin{cases} p(\zeta^1(i) = 0|\pi) = 1 - \pi \\ p(\zeta^1(i) = 1|\pi) = \pi \end{cases}. \tag{5}$$

Here $\zeta^1(i) = 0$ means the presence of noise only, i.e. "State 0: $\mathbf{Y}(i) = \mathbf{N}(i)$", whereas$\zeta^1(i) = 1$ indicates the presence of noise *and* the signal of interest, i.e. "State 1: $\mathbf{Y}(i) = p(\mathbf{Y}(i)|\zeta^1(i) = 1, \mathbf{C}_n, \mathbf{C}_x^1)p(\zeta^1(i) = 1|\pi)\mathbf{X}^1(i) + \mathbf{N}(i)$". This is recognized as a HMM with two hidden states.

Let us now introduce the likelihood function,

$$p(\mathbf{Y}(i)|\zeta^1(i), \mathbf{C}_n, \mathbf{C}_x^1) \sim \mathcal{CN}(\mathbf{Y}(i); \mathbf{0}, \mathbf{C}_n + \zeta^1(i)\mathbf{C}_x^1), \tag{6}$$

where $\mathcal{CN}(\mathbf{Y}; \boldsymbol{\mu}, \mathbf{C}) = \exp(-\mathbf{Y}(i)^H\mathbf{C}^{-1}\mathbf{Y}(i))/(\pi^{N_f}|\mathbf{C}|)$ denotes the circular-symmetric complex normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$which naturally arrives as a result of the Central Limit theorem applied to the STFT coefficients$\mathbf{Y}$[64]. Without loss of generality, it is assumed that $\boldsymbol{\mu} = \mathbf{0}$ (as obtained after first centering the signal). Although the covariance matrix $\mathbf{C}$ will be full in general, it has been observed in numerous cases that a diagonal structure provides an excellent approximation. Diagonal covariance matrices will therefore be assumed from now onwards, with the advantage of considerably limiting the number of unknowns in the model as well as providing a clear physical interpretation to the diagonal elements in terms of the instantaneous spectral density of signal *y*(*t*) (at time *i*).

Meanwhile, all the unknown parameters of the proposed model are denoted as $\boldsymbol{\theta} = \{\mathbf{C}_n, \mathbf{C}_x^1, \zeta^1(i), \pi\}$. It is highlighted that the latent variables $\boldsymbol{\zeta}(i)$ are hidden in the sense that they are not observed directly. Assuming independent segments in the STFT, the complete log-likelihood function is evaluated from Eq. (6) as

$$\log L_C\,(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log\,(p(\mathbf{Y}(i)|\boldsymbol{\theta})). \tag{7}$$

Developing further, one has

$$\sum_{i=1}^{N} \log\left(p(\mathbf{Y}(i)|\zeta^1(i) = 0, \mathbf{C}_n)p(\zeta^1(i) = 0|\pi) + p(\mathbf{Y}(i)|\zeta^1(i) = 1, \mathbf{C}_n, \mathbf{C}_x^1)p(\zeta^1(i) = 1|\pi)\right) \tag{8}$$

$$= \sum_{i=1}^{N} \log((1 - \pi) \times \mathcal{CN}(\mathbf{0}, \mathbf{C}_n) + \pi \times \mathcal{CN}(\mathbf{0}, \mathbf{C}_n + \mathbf{C}_x^1)) \tag{9}$$

where it has been assumed that all states are apriori equally probable.

The parameters $\boldsymbol{\theta}$ are next estimated by maximizing the above likelihood function. In theory, this completely solves the problem since the estimated latent variable $\boldsymbol{\zeta}(i)$ will then return the times of occurrence of the impacts on the faults and therefore the bearing characteristic frequency. Since it is difficult to find a closed-form solution, the EM algorithm [77] is used as an iterative method to find the maximum likelihood estimates. The EM algorithm makes use of the following quantities.

First, the posteriori probability distribution of the latent variable is formed as

$$p(\zeta^1(i)|\mathbf{Y}(i), \mathbf{C}_n, \mathbf{C}_x^1, \pi) = \frac{p(\mathbf{Y}(i)|\zeta^1(i), \mathbf{C}_n, \mathbf{C}_x^1)p(\zeta^1(i)|\pi)}{p(\mathbf{Y}(i)|\mathbf{C}_n, \mathbf{C}_x^1, \pi)}. \tag{10}$$

According to Eq. (10), the expectation of the latent variable $\zeta^1(i)$ given the measurement is then computed as

$$\boldsymbol{E}\{\zeta^1(i)|\mathbf{Y}(i),\mathbf{C}_n,\mathbf{C}_x^1,\pi\}=\frac{\pi\mathcal{CN}(\mathbf{0},\mathbf{C}_{x+n})}{(1-\pi)\mathcal{CN}(\mathbf{0},\mathbf{C}_n)+\pi\mathcal{CN}(\mathbf{0},\mathbf{C}_{x+n})} \tag{11}$$

where $\mathbf{C}_{\boldsymbol{x+n}}$ denotes $\mathbf{C}_{\boldsymbol{n}}+\mathbf{C}_{\boldsymbol{x}}^1$. After simple arrangement, this is expressed as

$$\boldsymbol{E}\{\zeta^1(i)|\mathbf{Y}(i),\mathbf{C}_n,\mathbf{C}_x^1,\pi\}=\frac{1}{1+\frac{1-\pi}{\pi}\times\mathrm{e}^{-\mathrm{LLR}(i)}} \tag{12}$$

where LLR denotes the natural logarithm of the likelihood ratio between the component of interest and the noise as given by

$$\mathrm{LLR}(i)=\mathbf{Y}(i)^H\left(\mathbf{C}_n{}^{-1}-\mathbf{C}_{x+n}{}^{-1}\right)\mathbf{Y}(i)-\log\frac{|\mathbf{C}_{x+n}|}{|\mathbf{C}_n|}. \tag{13}$$

The EM algorithm is summarized in Table 1.It is noted that it involves two parameters: $k_{\max}$for the maximum number of iterations and $\mathrm{STOP}-\mathrm{CRIT}$for the expected relative tolerance between $\hat{\mathbf{C}}_n^{[k+1]}$ and $\hat{\mathbf{C}}_n^{[k]}$. These can be easily set by default. The estimation of the covariance of the signal of interest is finally obtained as $\hat{\mathbf{C}}_x^1=\left(\hat{\mathbf{C}}_{x+n}-\hat{\mathbf{C}}_n\right)_+$where operator $(\ldots)_+$ keeps only the positive eigenvalues of a matrix (here the positive elements of a diagonal matrix).

With these estimated parameters, an automatic fault detection scheme and a time-varying filter for filtering out the signal of interest are proposed in the next section.

Table 1 : Explicit steps of the EM algorithm to infer the parameters in the HMM.

Input: $y(t) \in \mathbb{R}^L$, N_w, R, $N = floor[(L - N_w)/R + 1]$;

F_s, $k_{\max}$, STOP-CRIT.

Initialization: $Y(i, f_b) \in \mathbb{R}^{N \times N_f}$;

$\hat{\mathbf{C}}_{\boldsymbol{n}}{}^{[0]}$, $\hat{\mathbf{C}}_{\boldsymbol{x+n}}{}^{[0]}$, $\hat{\pi}^{[0]}$, see $Eqs.(33) \sim (36)$;

$k \longleftarrow 0$.

Repeat: STOP-CRIT or $(k < k_{\max})$

E-step: For each i, set

$$\mathrm{LLR}(i)^{[\mathrm{k}+1]} := \log \frac{|\hat{\mathbf{C}}_{\boldsymbol{n}}{}^{[\mathrm{k}]}|}{|\hat{\mathbf{C}}_{\boldsymbol{x+n}}{}^{[\mathrm{k}]}|} e^{\mathbf{Y}(i)^H (\hat{\mathbf{C}}_{\boldsymbol{n}}{}^{-1[\mathrm{k}]} - \hat{\mathbf{C}}_{\boldsymbol{x+n}}{}^{-1[\mathrm{k}]}) \mathbf{Y}(i)};$$

$$\hat{\zeta}^1(i)^{[\mathrm{k}+1]} := \frac{1}{1 + \frac{1-\hat{\pi}^{[\mathrm{k}]}}{\hat{\pi}^{[\mathrm{k}]}} \times \mathrm{e}^{-\mathrm{LLR}(i)^{[\mathrm{k}+1]}}};$$

M-step: Update the parameters:

$$\hat{\pi}^{[\mathrm{k}+1]} := \frac{1}{N} \sum_{i=1}^{N} \hat{\zeta}^1(i)^{[\mathrm{k}+1]};$$

$$\hat{\mathbf{C}}_{\boldsymbol{n}}{}^{[\mathrm{k}+1]} := \frac{\sum_{i=1}^{N} (1 - \hat{\zeta}^1(i)^{[\mathrm{k}+1]}) \mathbf{Y}(i)^H \mathbf{Y}(i)}{\sum_{i=1}^{N} (1 - \hat{\zeta}^1(i)\}^{[\mathrm{k}+1]})};$$

$$\hat{\mathbf{C}}_{\boldsymbol{x+n}}{}^{[\mathrm{k}+1]} := \frac{\sum_{i=1}^{N} \hat{\zeta}^1(i)^{[\mathrm{k}+1]} \mathbf{Y}(i)^H \mathbf{Y}(i)}{\sum_{i=1}^{N} \hat{\zeta}^1(i)^{[\mathrm{k}+1]}};$$

$k \longleftarrow k + 1$.

Until convergence

Output: $\hat{\mathbf{C}}_{\boldsymbol{n}}{}^{[\mathrm{k}+1]}, \hat{\mathbf{C}}_{\boldsymbol{x+n}}{}^{[\mathrm{k}+1]}, \hat{\pi}^{[\mathrm{k}+1]}, \hat{\zeta}^1(i)^{[\mathrm{k}+1]}, \mathrm{LLR}(i)^{[\mathrm{k}+1]},$

see $Eqs.(11) \sim (13)$.

b) *K-component case*

Let us now consider the general case with $K > 1$components of interest, $\mathbf{X}^k(i), k = 1, \dots, K$. In this case, the probability that two (or more) components occur together will be assumed so small that such events will be disregarded(this may be seen as an extreme sparse representation where only one state is allowed at a time). Therefore, under the mutually exclusive assumption, the occurrence of the k^{th}state is defined as

$$A_k = \{\zeta^k(i) = 1;\ \zeta^l(i) = 0 \mid 1 \le l \le K;\ l \ne k\} \tag{14}$$

where $\zeta^k(i) = 1$ indicates the presence of the k^{th} signal $\mathbf{X}^k(i)$. The pure noise case – i.e. $A_{K+1} = \{\boldsymbol{\zeta}(i) = \mathbf{0}\}$ – is denoted as the $(K+1)^{th}$ state. Thus, there is a total of $K+1$ possible states in the model. Therefore, the marginal probability distribution treads

$$p\left(\mathbf{Y}(i)|\mathbf{C}_n, \mathbf{C}_x^k, \pi^k, k=1,\dots,K\right) = \sum_{k=1}^{K+1} p\left(\mathbf{Y}(i)\middle|A_k, \mathbf{C}_n, \mathbf{C}_x^k\right) p(A_k|\pi_k) \tag{15}$$

where the k^{th} latent variable follows the Bernoulli distribution, $\zeta^k(i) \sim Bernouilli(\pi^k)$:

$$\begin{cases} p\left(\zeta^k(i)=0\right) = 1-\pi^k \\ p\left(\zeta^k(i)=1\right) = \pi^k \end{cases}. \tag{16}$$

(Note that the latent variables are now dependent because of the mutually exclusive assumption.)

The posteriori probability distribution then reads

$$p\left(\boldsymbol{\zeta}(i)\middle|\mathbf{Y}(i), \mathbf{C}_n, \mathbf{C}_x^k, \pi^k, k=1,\dots,K\right) = \frac{p(\mathbf{Y}(i)|\boldsymbol{\zeta}(i), \mathbf{C}_n, \mathbf{C}_x^k) p(\boldsymbol{\zeta}(i)|\pi^k)}{p\left(\mathbf{Y}(i)|\mathbf{C}_n, \mathbf{C}_x^k, \pi^k\right)}. \tag{17}$$

Assuming mutually exclusive states, the expectation of the k^{th} latent variable is thus

$$\boldsymbol{E}\{\zeta^k(i)|\mathbf{Y}(i), \mathbf{C}_n, \mathbf{C}_x^k, \pi^k, k=1,\dots,K\} = \frac{\pi^k \mathcal{CN}(\mathbf{0}, \mathbf{C}_n + \mathbf{C}_x^k)}{\mathcal{CN}(\mathbf{0}, \mathbf{C}_n) \prod_{k=1}^{K}(1-\pi^k) + \sum_{k=1}^{K} \pi^k \mathcal{CN}(\mathbf{0}, \mathbf{C}_n + \mathbf{C}_x^k)}. \tag{18}$$

Figure 1 illustrates the situation with K = 2 components, which involves three states.

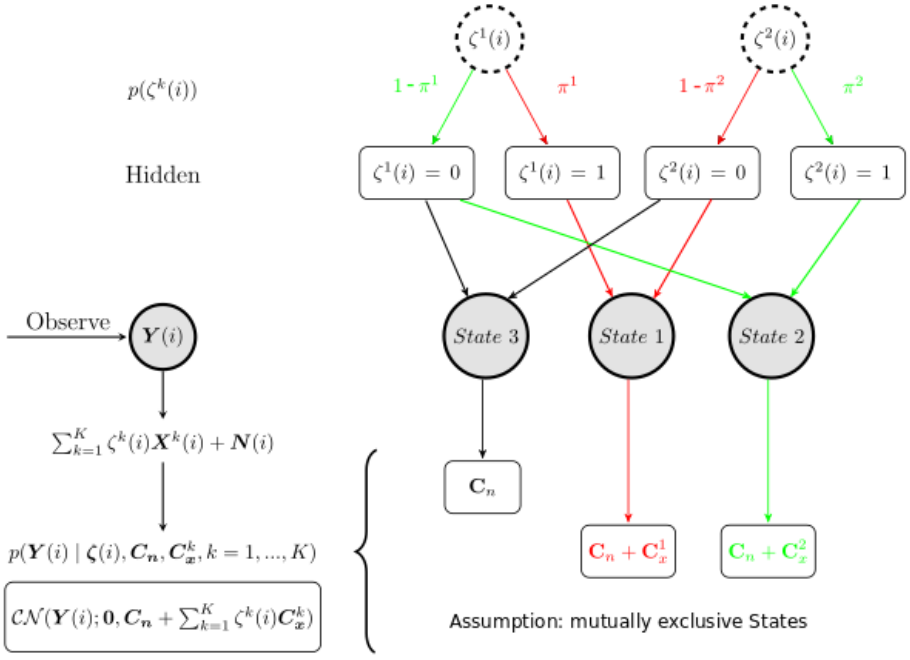


Figure 1. Graphical model in the case of K = 2 components.

3. Proposed methodology

This section now explains how the proposed HMM can be used for diagnosis. As explained in the introduction, the diagnosis tasks are divided into the detection, the identification and the extraction of a fault.

3.1. Detection

Being base on a Gaussian mixture model, the proposed probabilistic model easily lends itself to the formulation of a statistical detection test. This is formally stated by comparing two hypothesis, the null one, H_0, corresponding to the situation where no transient is present in the signal and the alternative one, H_1, where transients (and therefore a fault) are present. The test of hypothesis therefore literally reads

$$\begin{cases} H_0: "no\ fault\ is\ present" \\ H_1: "a\ fault\ is\ present" \end{cases} \tag{19}$$

or, in mathematical language,

$$\begin{cases} H_0: p(\mathbf{Y}(i), i = 1, \dots, N|H_0) \sim \prod_{i=1}^{N} \mathcal{CN}\left(\mathbf{Y}(i); 0, \boldsymbol{C}_n^{H_0}\right) \\ H_1: p(\mathbf{Y}(i), i = 1, \dots, N|H_1) \sim \prod_{i=1}^{N} (1-\pi) \cdot \mathcal{CN}\left(\mathbf{Y}(i); 0, \boldsymbol{C}_n^{H_1}\right) + \pi \cdot \mathcal{CN}\left(\mathbf{Y}(i); 0, \boldsymbol{C}_x + \boldsymbol{C}_n^{H_1}\right) \end{cases} \tag{20}.$$

An advantageous solution for testing H_1 against H_0 is by means of the generalized likelihood ratio test (GLRT). The principle is to take the ratio (or its logarithm) of the probability under H_1 to the probability under H_0– where all unknown parameters are replaced by their maximum likelihood estimates – and to reject the null hypothesis if the ratio happens to be greater than a certain value. The GLRT has the advantage of being optimal in the Neyman-Pearson sense, that is it maximizes the probability of detection (here of accepting H_1 when there is actually a fault) for a given probability of false alarm (accepting H_1 when there is no fault). Another definite advantage of the GLRT in the present case is to come with a theoretical value for the threshold against which to compare the ratio.

Using Eq. (20), the logarithm of the GLRT is

$$\ln\Lambda = \ln \frac{\prod_{i=1}^{N} (1-\hat{\pi}) \cdot \mathcal{CN}\left(\mathbf{Y}(i); 0, \widehat{\boldsymbol{C}}_n^{H_1}\right) + \hat{\pi} \cdot \mathcal{CN}\left(\mathbf{Y}(i); 0, \widehat{\boldsymbol{C}}_x + \widehat{\boldsymbol{C}}_n^{H_1}\right)}{\prod_{i=1}^{N} \mathcal{CN}\left(\mathbf{Y}(i); 0, \widehat{\boldsymbol{C}}_n^{H_0}\right)} \tag{21}$$

$$= \sum_{i=1}^{N} \ln\left((1-\hat{\pi}) \cdot \frac{e^{-\mathbf{Y}(i)^H \left(\widehat{\boldsymbol{C}}_n^{H_1}\right)^{-1} \mathbf{Y}(i)}}{\left|\widehat{\boldsymbol{C}}_n^{H_1}\right|} + \hat{\pi} \cdot \frac{e^{-\mathbf{Y}(i)^H \left(\widehat{\boldsymbol{C}}_x + \widehat{\boldsymbol{C}}_n^{H_1}\right)^{-1} \mathbf{Y}(i)}}{\left|\widehat{\boldsymbol{C}}_x + \widehat{\boldsymbol{C}}_n^{H_1}\right|} \right) + \sum_{i=1}^{N} \left(\mathbf{Y}(i)^H \left(\widehat{\boldsymbol{C}}_n^{H_0}\right)^{-1} \mathbf{Y}(i) + \ln\left|\widehat{\boldsymbol{C}}_n^{H_0}\right| \right)$$

where $\hat{\pi}$, $\widehat{\boldsymbol{C}}_n^{H_1}$ and $\widehat{\boldsymbol{C}}_x$ are the maximum likelihood estimates returned by the EM algorithm (see Table 1) and

$$\widehat{\boldsymbol{C}}_n^{H_0} = (1-\hat{\pi}) \cdot \widehat{\boldsymbol{C}}_n^{H_1} + \hat{\pi} \cdot \widehat{\boldsymbol{C}}_x \ . \tag{22}$$

Because the hypotheses H_0 and H_1 are nested, it can be shown from Wilk's theorem that, under H_0, twice the logarithm of the GLRT follows a Chi2 distribution with number of degrees of freedom equal to the difference between the number of unknown parameters under H_1 (1 for $\hat{\pi}$, N_f for $\widehat{\boldsymbol{C}}_n^{H_1}$ and N_f for $\widehat{\boldsymbol{C}}_x$) and

H_0 (N_f for $\widehat{\boldsymbol{C}}_n^{H_0}$), viz

$$\ln\Lambda \sim \chi^2_{1+N_f}. \tag{23}$$

Therefore, the test of hypothesis is

$$\text{Reject } H_0 \text{ if } \ln\Lambda > \chi^2_{1+N_f,1-\alpha}$$

at the risk α (i.e. probability of alarm), where $\chi^2_{1+N_f,1-\alpha}$ is the quantile of the Chi2 with probability 1-α.

3.2. Identification

As explained in the introduction, the main tool to identify a bearing fault is the envelope spectrum due its ability to estimate the bearing characteristic frequencies. In the proposed HMM, the LLR in Eq. (13) may be interpreted as the likelihood that a transient is present at any time datum i and its variation in time reveals the fault frequency. Therefore, the Fourier transform of the LLR turns out a valid alternative to the SES. Other quantities also contain the information on the fault frequency such as the latent variables $\zeta(i)$ and the logarithm of the GLRT in Eq. (21), yet it has been verified in numerous examples that the spectrum of the LLR is more accurate than the former and just as good as the latter while involving a simpler expression. One shared advantage of the LLR and GLRT is that they can capture subtle modulations of the transient magnitudes (as would typically happen when the fault is subjected a non-uniform load distribution), which is less likely the case for the binary latent variable.

3.3. Fault extraction

As shown in this subsection, the fault signal can be reconstructed in full band based on the latent variables and the covariance matrices estimated in the HMM. First, let introduce the posterior probability distribution of the k^{th} signal of interest $\mathbf{X}^k(i)$ as

$$p\left(\mathbf{X}^k(i)|\mathbf{Y}(i),\widehat{\boldsymbol{\theta}}^k\right) \propto p\left(\mathbf{Y}(i)|\mathbf{X}^k(i),\widehat{\boldsymbol{\theta}}^k\right)p\left(\mathbf{X}^k(i)|\widehat{\boldsymbol{\theta}}^k\right) \tag{24}$$

The posterior probability density at a given frequency then reads

$$p\left(X^k(i,f_b)|Y(i,f_b),\widehat{\boldsymbol{\theta}}^k\right) = \frac{\mathrm{e}^{\frac{-\left|Y(i,f_b)-\zeta^k(i)X^k(i,f_b)\right|^2}{\hat{C}_n^k(f_b)}}\mathrm{e}^{\frac{-\left|X^k(i,f_b)\right|^2}{\hat{C}_x^k(f_b)}}}{\pi^2\hat{C}_n^k(f_b)\hat{C}_x^k(f_b)} \tag{25}$$

where $\hat{C}_n^k(f_b)$ and $\hat{C}_x^k(f_b)$ stand for the noise and signal variance of the k^{th} component frequency f_b, respectively. After some manipulations, Eq. (25) is expressed as

$$p\left(X^k(i,f_b)|Y(i,f_b),\widehat{\boldsymbol{\theta}}^k\right) = \frac{\mathrm{e}^{\frac{-\left|X^k(i,f_b)-\mu_x^k(f_b)\right|^2}{C_x^k(f_b)}}}{\pi C_x^k(f_b)} = \mathcal{CN}\left(X^k(i,f_b);\mu_x^k(f_b),C_x^k(f_b)\right) \tag{26}$$

with

$$\begin{cases} C_x^k(f_b) = \left(\frac{\zeta^k(i)^2}{\hat{C}_n^k(f_b)} + \frac{1}{\hat{C}_x^k(f_b)}\right)^{-1} \\ \mu_x^k(f_b) = \frac{C_x^k(f_b)}{\hat{C}_n^k(f_b)}\zeta^k(i)Y(i,f_b) \end{cases} \tag{27}.$$

Therefore the expectation of the k^{th} signal of interest $X^k(i,f_b)$ is

$$\boldsymbol{E}\{X^k(i,f_b)|Y(i,f_b),\widehat{\boldsymbol{\theta}}^k\} = \mu_x^k(f_b) = \frac{\zeta^k(i)}{\zeta^k(i)^2+\frac{\widehat{c}_n^k(f_b)}{\widehat{c}_x^k(f_b)}}Y(i,f_b). \tag{28}$$

Finally, the time signal $\hat{x}^k[n]$ is obtained from Eq. (28) by using the inverse STFT.

Two remarks are noteworthy. First, it is seen that Eq. (28) corresponds to a time-varying filter from which superior performance is expected than from a conventional time-invariant filter. Second, the standard Wiener filter appears as a particular case under the assumption of stationarity, that is

$$\boldsymbol{E}\{X^k(i,f_b)|Y(i,f_b)\} = \frac{1}{1+\frac{c_n^k(f_b)}{c_x^k(f_b)}}Y(i,f_b), \tag{29}$$

where the latent variable$\zeta(i) = 1$ for all time instants. In other words, Eq. (29) corresponds to the case where "State 1: $\mathbf{Y}(i) = \mathbf{X}(i) + \mathbf{N}(i)$" occurs only.

The detection, identification and extraction tasks will be extensively illustrated in the next sections.

4. Parameter setting, algorithm initialization and validation

This section is divided in two parts. The first one addresses the setting of the parameters of the STFT decomposition – which are the only ones not estimated in the proposed methodology – and in particular of the frequency resolution which is the only one that possibly requires an intervention by the user. The second part then addresses the initialization the EM algorithm by means of an effective data-driven approach.

4.1. Parameter setting

The parameters entering into the STFT are the window length N_w and the window shift R (see Eq. (2))

4.1.1. Window length N_w

The value of N_w directly controls the frequency resolution,

$$\Delta f = F_s/N_w, \tag{30}$$

which characterizes the carrier frequency. It is required to cover at least the duration T_I of a transient, which implies the condition

$$\Delta f < 1/T_I. \tag{31}$$

As the STFT is subjected to the uncertainty principle, $\Delta t \Delta f \geq 1$, the highest switching frequency of the latent variable, $\alpha = 1/\Delta t$, is bounded upward by Δf [78]. Therefore the available range of the latent variable is limited by

$$\alpha \leq \Delta f. \tag{32}$$

Therefore, N_w should be taken short to allow a high switching rate in Eq. (32), but long enough to satisfy Eq. (31), i.e. $F_s \cdot T_I < N_w \leq F_s \cdot \Delta t$ as illustrated in Fig. 2.

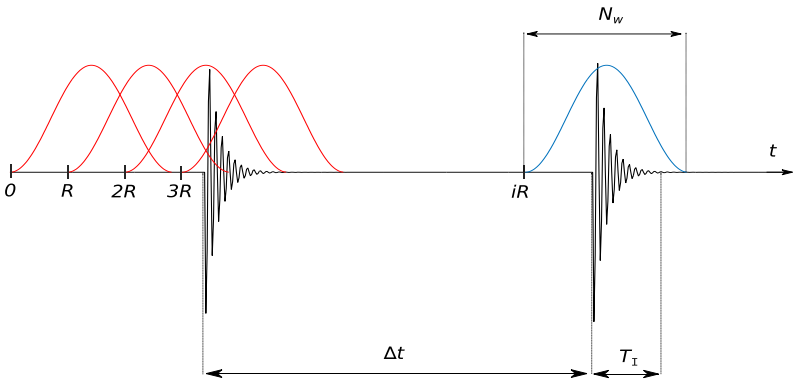


Figure 2: Illustration of how to select the window length N_w and shift R with respect to transient durations T_I and cycle Δt.

Other reasons for taking N_w small is to reduce the computation time required by the STFT and also to ensure sufficient segments for accurate parameter estimation.

It is noticed here that the rule for setting N_w also works in the special case where the interval between adjacent transients Δt is close to the transient duration T_I. In other words, it is robust enough to balance the trade-off between a fine resolution and a high switching frequency of the latent variable. These facts will be further experimentally verified in subsection 4.3.

4.1.2. Window shift R

There are two considerations for setting the window shift R:

- first, for the STFT to be invertible, it is recommended to take at least 75% overlap with a Hanning window,
- second, if inevitability is not required, R should be taken sufficiently small to keep enough diagnostic information while not increasing too much the computational cost and the dependence between adjacent segments; a typical choice is within 50% and 75% overlap with a Hanning window.

Therefore, the window shift can be easily set by default.

4.2. Initializing of the EM algorithm

The EM algorithm generally requires a good initialization for two reasons. The first ones is to avoid being trapped in possible local maxima of the likelihood and the second one is to achieve a fast convergence speed. A simple self-running solution is given hereafter to obtain good initial values for the covariance matrices of the two states in the HMM.

Initial estimates of the diagonal elements of the noise covariance matrix, $\hat{C}_{\boldsymbol{n}}^{[0]}(f_b)$, are obtained by taking the median value of the natural logarithm of the squared magnitude of the STFT coefficients $Y(i, f_b)$ with

respect to time instant i, i.e.

$$\hat{C}_n^{[0]}(f_b) = \exp(\text{median}\{\log|Y(i,f_b)|^2\}). \quad (33)$$

This estimator is based on the fact that State 0 is characterized by a higher probability than State 1 so that the median in the above equation is almost unaffected by the occurrences of the fault. Besides, the initial estimate of the covariance matrix in State 1 is based on the extreme values of $\log|Y(i,f_b)|^2$.Specifically, let define the "deviation"

$$\delta(i) = \frac{1}{N_f}\sum_{f_b=1}^{N_f}\log|Y(i,f_b)|^2 - \frac{1}{N_f}\sum_{f_b=1}^{N_f}\log\hat{C}_n(f_b)^{[0]} \quad (34)$$

and a threshold λequal to a fraction of the maximum of $\delta(i)$. Then, the sample set $I^{(1)}$is constructed by collected all indices i where $\delta(i)$ is found greater than λ. Hence, the diagonal elements of the covariance matrix in State 1 are estimated as

$$\hat{C}_{x+n}^{[0]}(f_b) = \frac{1}{N_1}\sum_{i\in I^{(1)}}|Y(i,f_b)|^2 \quad (35)$$

where N_1 is the cardinal of set $I^{(1)}$. Finally, $\hat{\mathbf{C}}_x^{[0]}$ is obtained as $\left(\hat{\mathbf{C}}_{x+n}^{[0]} - \hat{\mathbf{C}}_n^{[0]}\right)_+$.The corresponding probability is initialized to

$$\hat{\pi}^{[0]} = \frac{N-N_1}{N}. \quad (36)$$

It has been observed in numerous experiments that the proposed initializations are often quite close to the maximum likelihood estimates (global maximum)while allowing at the same time a fast convergence speed of the EM algorithm. This will be demonstrated in the next section.

5. Validation on synthetic signals

The section deals with a validation of the proposed methodology on synthetic signals.

5.1.Cases 1&2: demonstration of parameter selection

To demonstrate the performance of the proposed algorithm and its initialization, a synthetic signal is generated with a resonance frequency $f_0 = 0.15$ Hz, which is further modulated by a relatively high fault frequency $\alpha_0 = 1.25 \times 10^{-3}$ Hz ($T = 1/\alpha_0 = 800$ s, the sampling frequency is normalized as $F_s = 1$ Hz). The synthetic signal is described as:

$$y(t) = \sum_{j=-\infty}^{+\infty} h\big(t - jT - \tau_j\big)A_j + n(t) \quad (37)$$

$$H(z) = \frac{b_1 + b_2 \cdot z^{-1}}{a_1 + a_2 \cdot z^{-1} + a_3 \cdot z^{-2}} \quad (38)$$

where $\tau_j \sim \mathcal{N}(\mu_\tau = 0, \sigma_\tau = 0.02T)$ and $A_j \sim \mathcal{N}(\mu_A = 0, \sigma_A = 0.1)$ account for the uncertainties on the arrival time and on the magnitude of the j^{th} transient, respectively. The white noise $n(t)$ is set to a SNR of -6dB and the signal length is $L = 10^5$ samples. A second-order system is defined by Eq. (38), whose numerator and denominator coefficients are $\mathbf{b} = [-1, 1]$ and $\mathbf{a} = [1, -2\cos(2\pi f_0)\, r, r^2]$ with $r = 0.95$, respectively. Figure3 shows the spectrogram (magnitude of the STFT) of the raw signal whose record in time is displayed in Fig. 4 (a).The frequency resolution is set to$\Delta f = F_s/N_w = 1/2^7$Hz and a default value of 85% is used for overlap (R = 20).

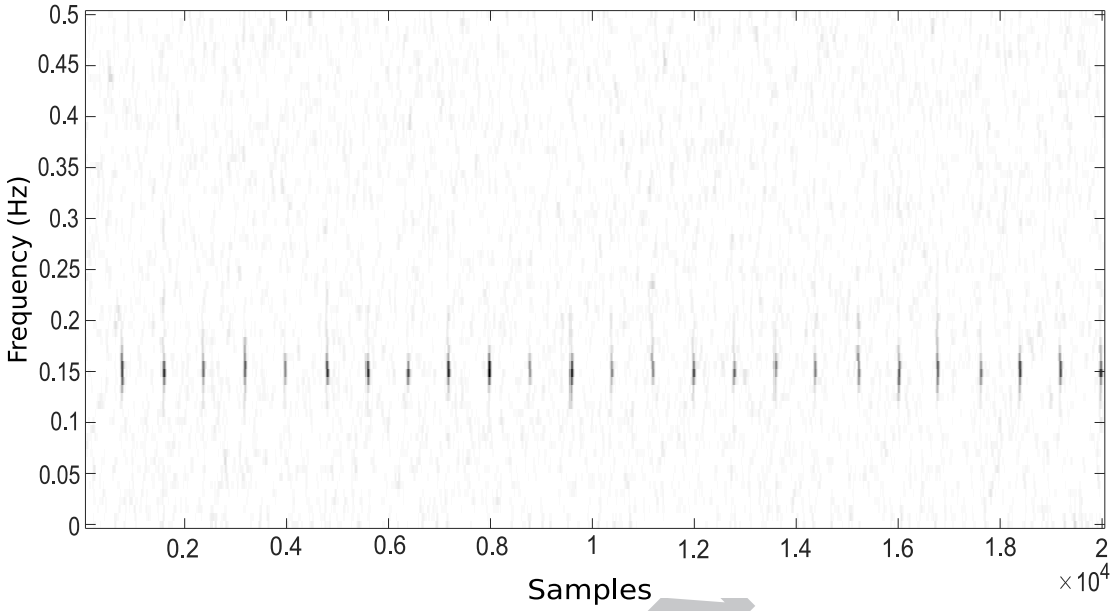


Figure 3. Spectrogram of the signal simulated in Case 1 with resonance frequency $f_0 = 0.15$ Hz, $r = 0.95$ and fault frequency $\alpha_0 = 1.25 \times 10^{-3}$ Hz ($T = 1/\alpha_0 = 800$ s, $N_w = 2^7$,$R = 20$ and SNR = -6dB).

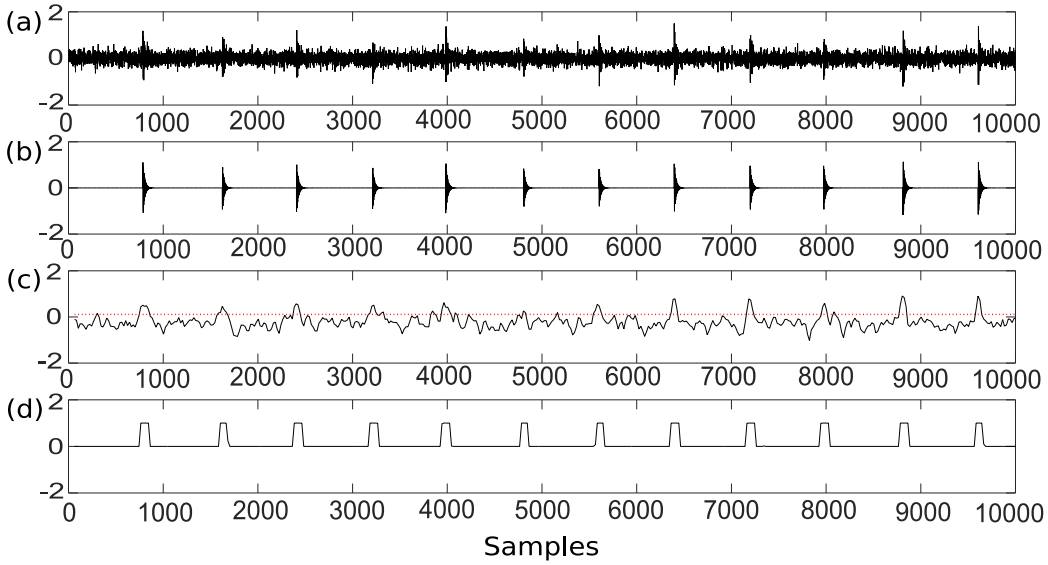


Figure 4. (a) Synthetic signal of Case 1with white noise (SNR = -6 dB). (b) Synthetic repetitive transients. (c) Deviation $\delta(i)$ (black line) with thresholdλ (red dotted line) set to $0.1 \times \max(\delta(i))$ with $\hat{\pi}^{[0]} = 0.112$.(d) Estimated latent variable $\hat{\zeta}(i)$.

Following Eqs. (33)-(36), one can initialize the parameters $\hat{\mathbf{C}}_n^{[0]}$, $\hat{\mathbf{C}}_x^{[0]}$ and $\hat{\pi}^{[0]}$ as shown in Fig. 5 (a).It is seen that the proposed initialization is simple and effective, even though the estimated spectrum of the signal of interest still contains a significant contribution from noise especially below 0.08 Hz. After

convergence of the EM algorithm, the estimation of the signal and noise spectra are close to the real values as can be seen in Fig. 5 (b). The very good estimation of the latent variable is verified in Fig. 4 (d).

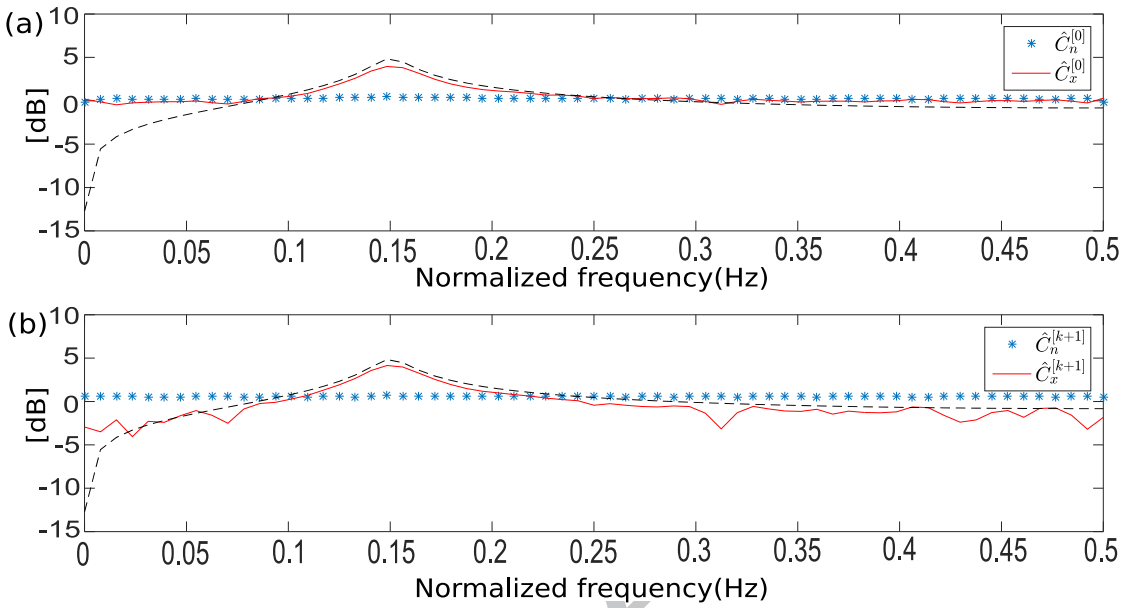


Figure 5. (a) Initialized spectra (diagonals of covariance matrices $\hat{\mathbf{C}}_n^{[0]}$ and $\hat{\mathbf{C}}_x^{[0]}$) (red line and blue asterisks) and (b) estimated ones from the EM algorithm (diagonals of covariance matrices $\hat{\mathbf{C}}_n^{[k+1]}$ and $\hat{\mathbf{C}}_x^{[k+1]}$) (red lines) together with the squared magnitude of the frequency response $H(z)$(black dashed line and blue asterisks).

The detection capability of the HMM is now checked by means of the statistical test introduced in section 3.1. The same experience is repeated for several values of the signal-to-noise ratio (SNR) ranging from -20dB to 0dB and different values of the frequency resolutioncorresponding to $\Delta f = 1/N_w = 2^{-5}, 2^{-6}, \dots, 2^{-9}$. In each case, 500 hundreds realizations of the same signal (with different noise generations) are run. The signal length is 100,000 samples and the risk is set to 5%. The results are reported in Table 2. It is seen that the presence of the fault is detected for SNRs greater than or equal to -16dBfor all frequency resolution up to 2^{-8} (SNRs even lower than this can be reached when the transient are lightly damped), which proves a certain robustness with respect to the latter parameter. It reduces to -8dB when the frequency resolution decrease to 2^{-9} because it then becomes comparable to the fault frequency a_0. For frequency resolution smaller than a_0 the fault could not be detected, which is consistent with the discussion of subsection 4.1.1. The results are also displayed in Fig. 6, which demonstrates that the proposed GLRT grows exponentially above the statistical threshold as the SNR increases.

Table 2: parameters used in the detection test.

$\Delta f = 1/N_w$	2^{-5} = 1/32	2^{-6} = 1/64	2^{-7} = 1/128	2^{-8} = 1/256	2^{-9} = 1/512
$N_f = N_w/2 - 1$	15	31	63	127	255
N (STFT length)	19994	9994	4994	2558	1293
Threshold $\chi^2_{1+N_f,1-\alpha}$	13	23	42	78	147
SNR from which H_0 is rejected	-16dB	-16dB	-16dB	-16dB	-8dB

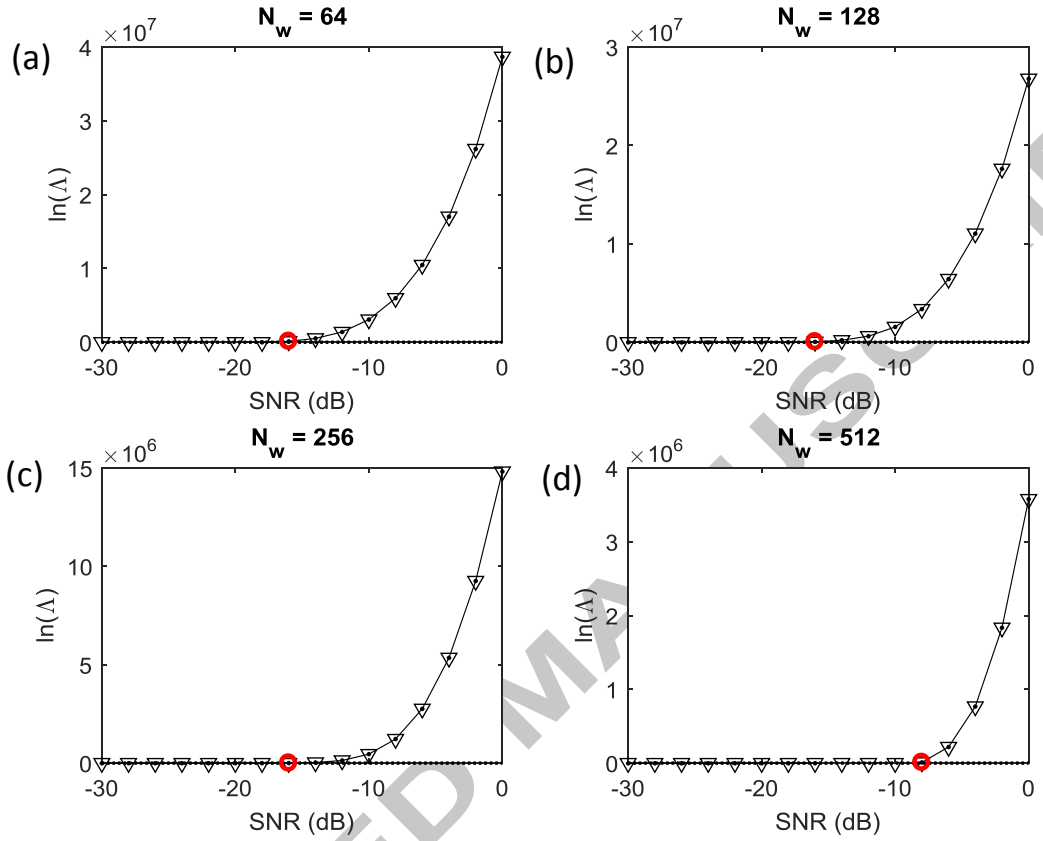


Figure 6. Evolution of the logarithm of the GLRT as a function of the SNR and the frequency resolution $\Delta f = 1/N_w$. The statistical threshold is marked by a dotted horizontal line corresponding to a risk of 5%. The red bullet indicates the SNR where the GLRT goes above the threshold, i.e. where the fault is detected.

The same experiment is now carried on in order to check the effect of the initialization of the EM algorithm. The main parameter that governs initialization is the threshold λ defined after Eq. (34). Different values are tested, i.e. $0.1 \times$, $0.2 \times$, $0.3 \times$, $0.4 \times \max(\delta(i))$, which are represented in Fig. 7 (a) together with the deviation $\delta(i)$. Figure 7 (b) displays the corresponding estimates of the transient spectrum in State 1 (i.e. the diagonal of covariance matrix $\hat{\mathbf{C}}_x^{[0]}$), which is probably the most difficult to obtain. It is seen that all initial estimates are very similar and also quite close the reference given by the squared magnitude of the frequency response $H(z)$. A closer look around the resonance in the band 0.1 Hz to 0.2 Hzactually shows that the initial estimation gradually improves when the threshold is increased.

The normalized root-mean-square error (RMSE) $\sqrt{\sum_{i=1}^{N}\left(\hat{\zeta}(i)^{[k+1]} - \hat{\zeta}(i)^{[k]}\right)^2 / \left(\sum_{i=1}^{N}\hat{\zeta}(i)^{[k]}\right)^2 / N}$ is displayed in Fig. 8 as a function of the iteration number k in the EM algorithm. It evidences that the convergence speed of EM algorithm also increases with the threshold. For information, thetotal computational time in the case $0.1 \times \max(\delta(i))$is 1.34sCPU (Central Processing Unit) with a PC with i7-3930K 3.20 GHz Processor.

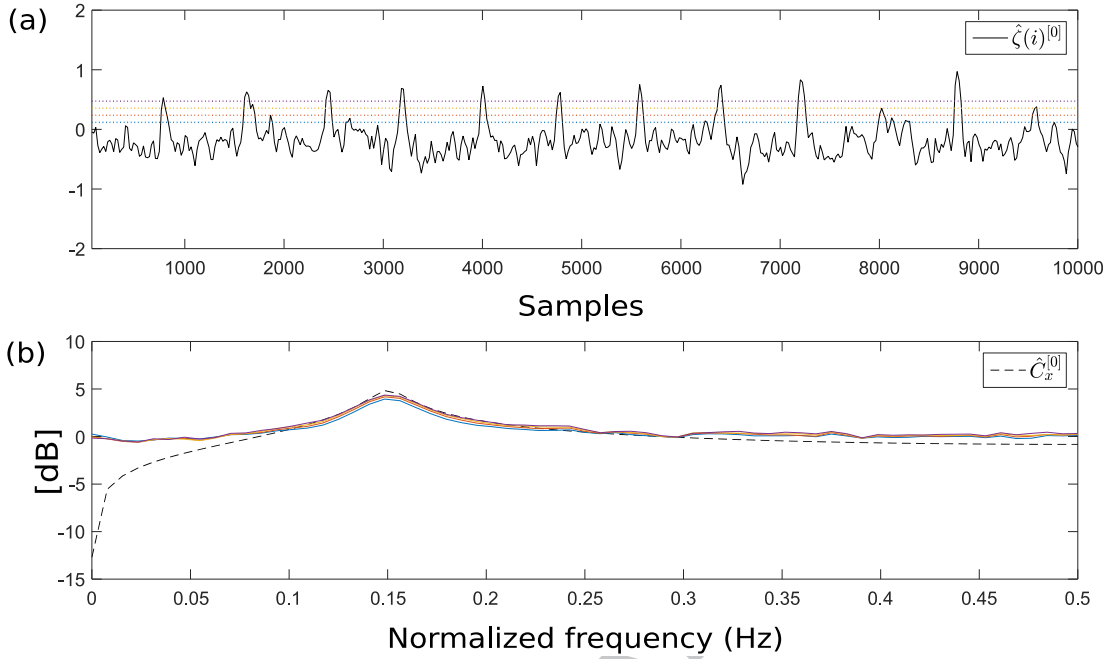


Figure 7. (a) Deviation $\delta(i)$(black line) with thresholdλ (colored dotted lines) set to $0.1\times$, $0.2\times$, $0.3\times$, $0.4\times$ $\max(\delta(i))$ and (b) corresponding initialized spectrum (diagonal of covariance matrix$\hat{\mathbf{C}}_x^{[0]}$) in colored lines together with the squared magnitude of the frequency response $H(z)$ (black dashed line and blue asteriks).

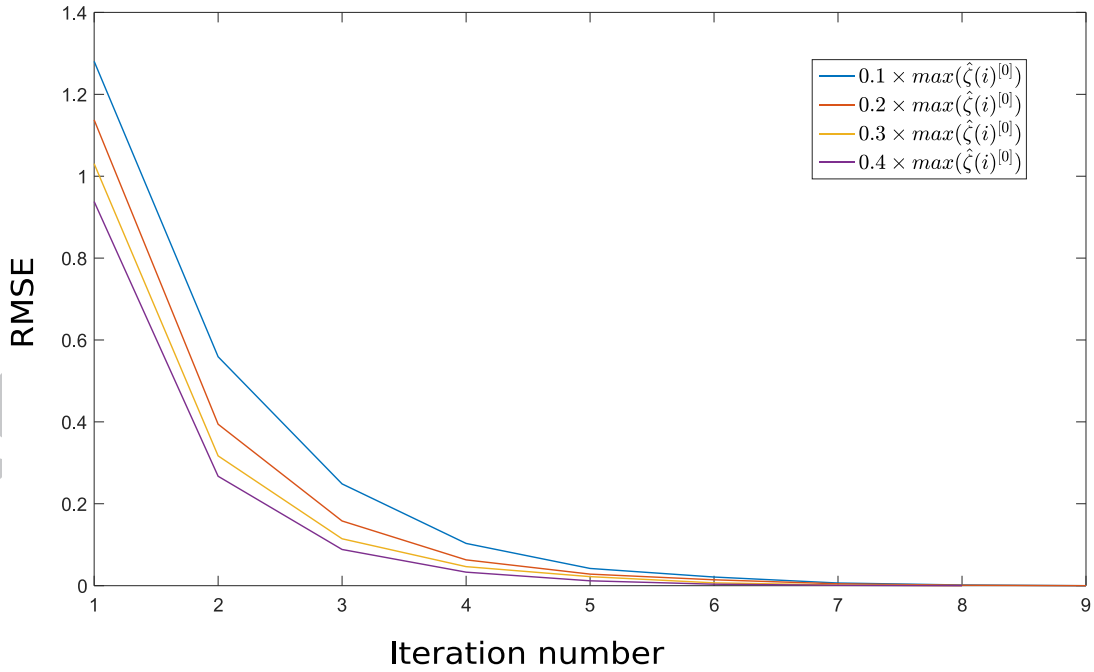


Figure 8. Normalized root-mean-square error (RMSE) on$\hat{\zeta}(i)^{[k]}$as a function of the iteration number of the EM algorithm ($k_{\max}=20$and $\text{STOP}-\text{CRIT}=10^{-4}on\hat{\zeta}(i)^{[k+1]}$).

In order to verify the robustness with respect to the setting of the frequency resolution, a second synthetic signal is generated and analyzed with a deliberately coarse value of Δf. The resonance frequency is $f_0 = 6 \times 10^{-3}$ Hz and the fault frequency $\alpha_0 = 1.25 \times 10^{-3}$ Hz. The frequency resolution is chosen as $N_w = 2^7$ ($\Delta f = 7.8 \times 10^{-3}$ Hz), which covers 6 integer multiples of α_0 Hz but cannot resolve the resonance frequency f_0. The corresponding limit on the detection of the fault frequency is $\alpha_{max} \leq \Delta f = 7.8 \times 10^{-3}$ Hz. All the other parameters are as in Case 1 (SNR = -6dB). Figure 9 shows the spectrogram of the raw signal. Since the modulation frequency is now close to the resonance one, this case encounters a trade-off between a fine spectral content and a large fault frequency range. After running the detection test, the GLRT $\ln\Lambda$ is found equal to 3703 which is much greater than the statistical threshold at the risk of 5%, $\chi^2_{1+N_f,1-\alpha} = 42$. Hence the fault is detected without ambiguity. Figure 10 (a) displays the estimated LLR, which accurately localizes the fault occurrences: the function sharply takes very large values when it identifies a transient. Besides, the estimated latent variable $\hat{\zeta}(i)$ locates exactly all the STFT segments that contains a fault occurrence, as shown in Fig. 10 (b). To further identify the fault type, the spectra of the LLR and of the latent variable are displayed in Fig. 11 together with the SES of the actual fault signal (note that all spectra are normalized to a unit maximum and that the limit $\alpha_{\max}$ is indicted by a vertical black dotted line). As explained in section 3, the spectrum of the LLR is preferred because it can capture modulations of the transient magnitudes which are not in the binary latent variable. It is indeed checked that the spectrum of the LLR in Fig. 11 (a) has a slightly larger extend than the spectrum of the latent variable in Fig. 11 (b). It is also noteworthy that the two spectra perfectly match the SES of the theoretical fault signal. Despite the coarse frequency resolution used in this case, this proves that the proposed method can still detect and identify the expected fault frequency with very good accuracy.

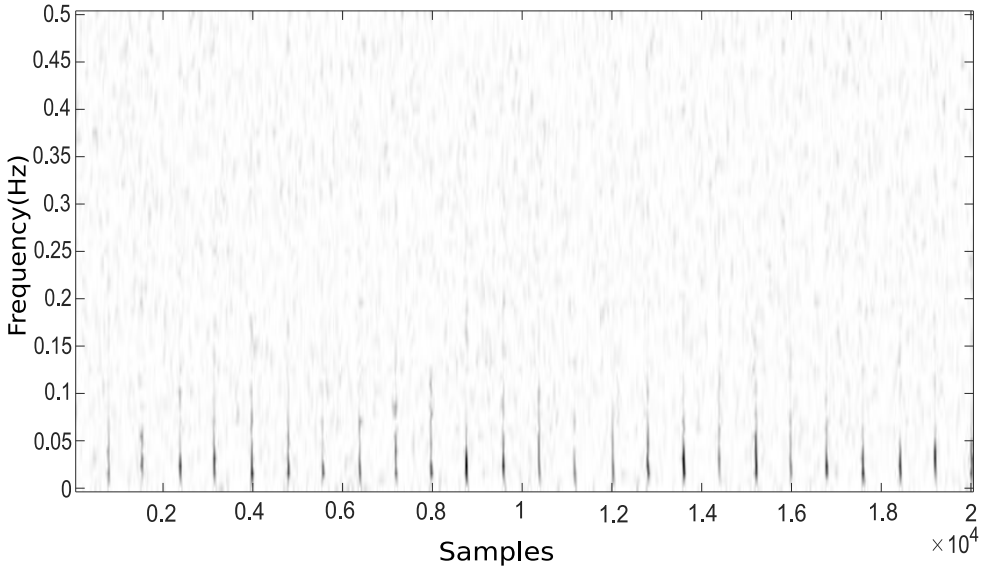


Figure 9. Spectrogram of the signal simulated in Case 2 with resonance frequency $f_0 = 6 \times 10^{-3}$ Hz, $r = 0.9$ and fault frequency $\alpha_0 = 1.25 \times 10^{-3}$ Hz (SNR = -6dB).

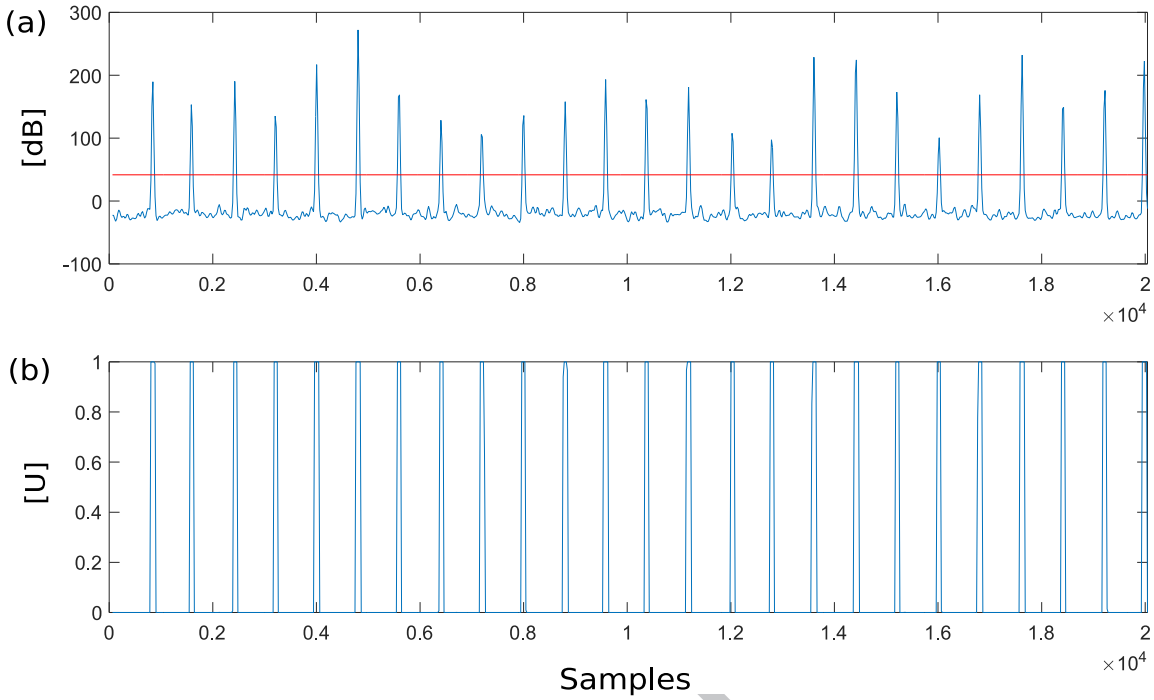


Figure 10(a) LLR(i)with the statistical threshold at the risk of 5%, $\chi^2_{1+N_f,1-\alpha} = 42$and (b) latent variable $\hat{\zeta}(i)$ in the time domain.

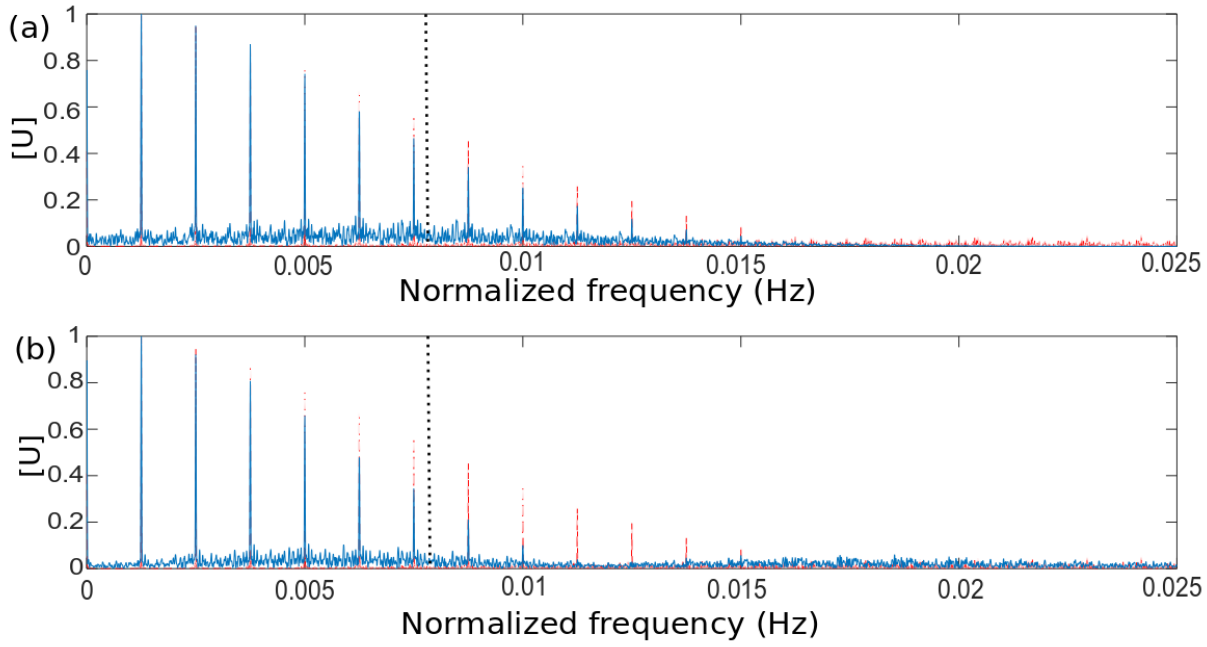


Figure 11: Spectra of the LLR(blue line) (a) and of the latent variable (b) superimposed with the SES of actual fault signal (red dotted line). The limit $\alpha_{\max} = 7.8 \times 10^{-3}$ Hz is indicted by a vertical black dotted line(Normalization to unit maximum value).

5.2. Case 3: Separation of a mixture of independent transients

This subsection demonstrates the potential of proposed HMM to deal with K = 2 simultaneous components in background noise. A synthetic signal with two different components is generated as shown in Fig. 12. The first component $\mathbf{X}^1(i)$ has a resonance frequency ${f_0}^1 = 0.25$ Hz with $r = 0.9$ and its cyclic frequency is${\alpha_0}^1 = 1.9 \times 10^{-3}$ Hz (${T_1} = 1/{\alpha_0}^1 = 530$ samples), whereas the second one$\mathbf{X}^2(i)$ has a resonance frequency ${f_0}^2 = 0.35$ Hz with $r = 0.7$ and cyclic frequency ${\alpha_0}^2 = 2.1 \times 10^{-3}$ Hz (${T_2} = 1/{\alpha_0}^2 = 470$ samples) – see Fig. 12 (a) and (b), respectively. All other parameters are set as in Case 1, with${\sigma_\tau}^{1,2} = 0.02{T_{1,2}}$ and ${\sigma_A}^{1,2} = 0.1$. The SNR is 0 dB.

The spectrogram of the measurement computed with $N_w = 2^6$ and $R = 23$ (overlapping ratio $R/N_w = 0.359$) is shown in Fig. 13. The estimated diagonals of the three covariance matrices are displayed in Fig. 14. It is seen that the spectra of all components are correctly identified. The corresponding latent variables $\zeta^1(i)$ and $\zeta^2(i)$ are displayed in Fig. 15. Compared with the reference signals, it is obvious that the times of occurrence of the two components have been correctly located.

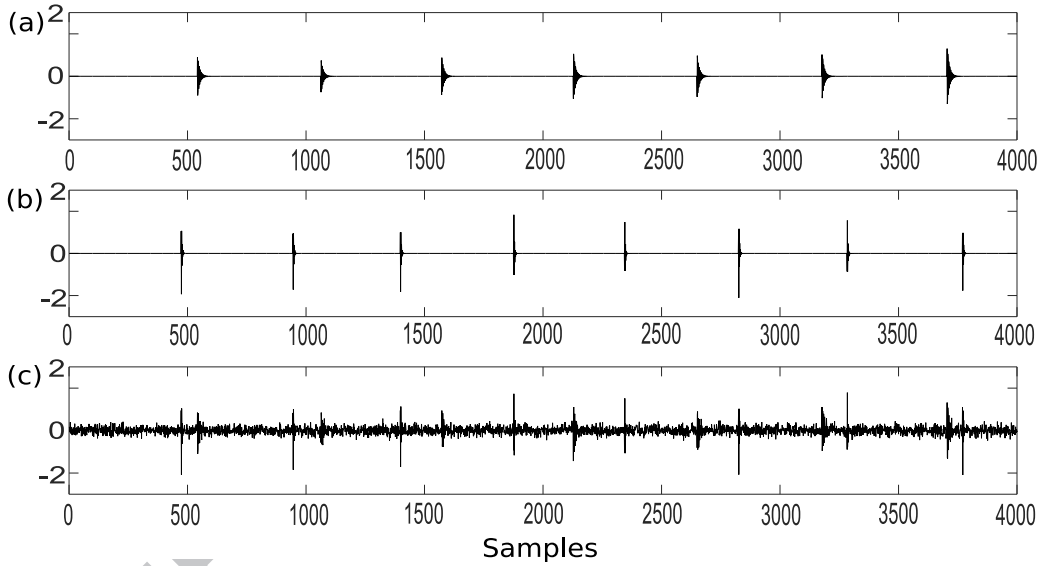


Figure 12(a) Component $\mathbf{X}^1(i)$ with period ${T_1} = 530$ samples (${\sigma_\tau}^1 = 0.02{T_1}$ and ${\sigma_A}^1 = 0.1$).(b) Component $\mathbf{X}^2(i)$ with period ${T_2} = 470$ samples (${\sigma_\tau}^2 = 0.02{T_2}$ and ${\sigma_A}^2 = 0.1$).(c) Noisy measurement (SNR =0 dB).

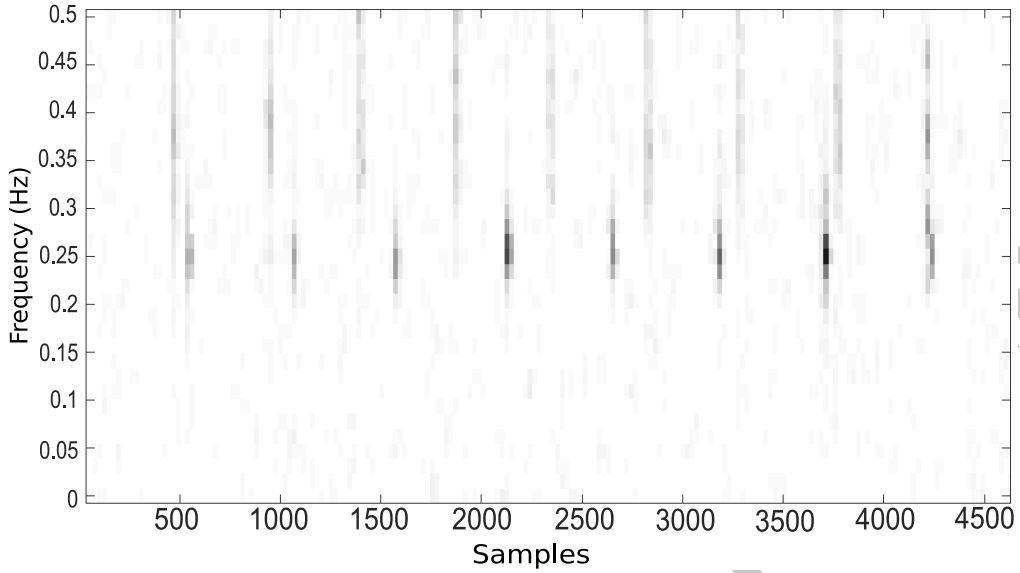


Figure 13. Spectrogram of the signal simulated in Case 3 with two components ($T_1 = 530$ samples and $T_2 = 470$ samples respectively).

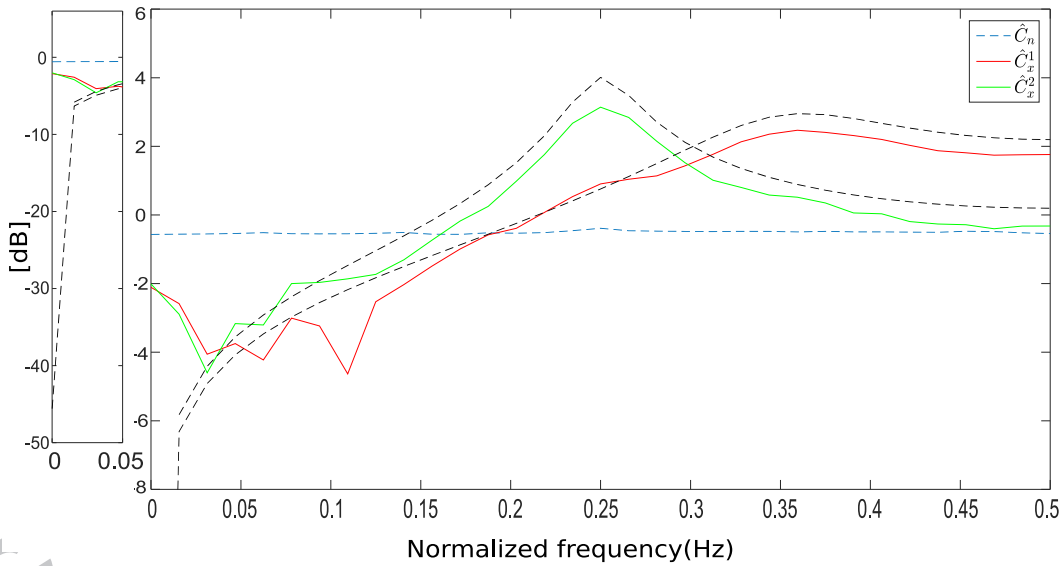


Figure 14. Estimated spectra of component $\mathbf{X}^1(i)$ (red solid line), component $\mathbf{X}^2(i)$ (green solid line) and noise (blue dashed line). The squared magnitudes of the two frequency responses are indicated by black dashed lines.

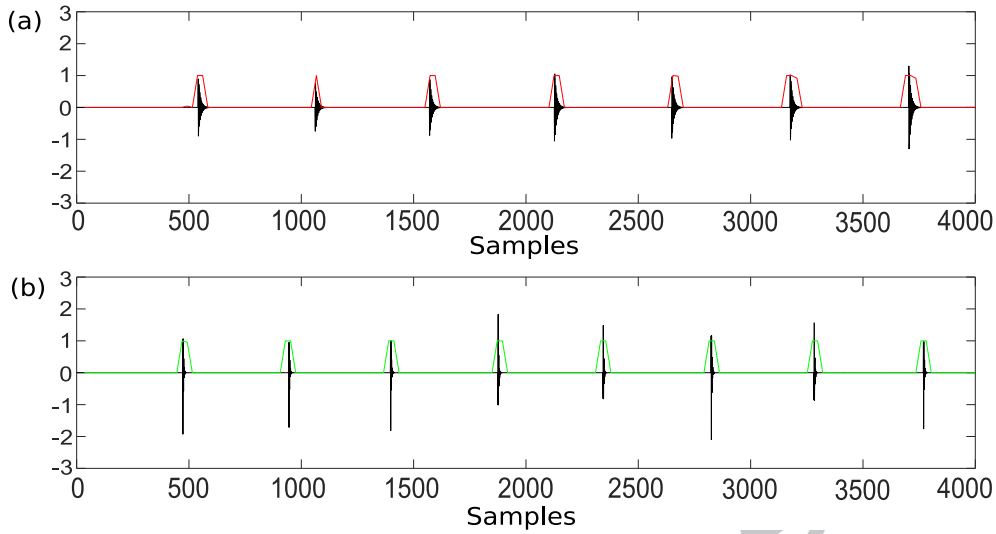


Figure 15. (a) Estimated latent variable $\zeta^1(i)$ (red solid line) together with the actual first component (black solid line). (b) Estimated latent variable $\zeta^2(i)$ (green solid line) together with the actual second component (black solid line).

Finally, Fig. 16 displays the reconstructed repetitive transients $\mathbf{X}^1(i)$ and $\mathbf{X}^2(i)$ as well as their summation. Very good reconstruction is obtained, which demonstrates the performance of the proposed algorithm.

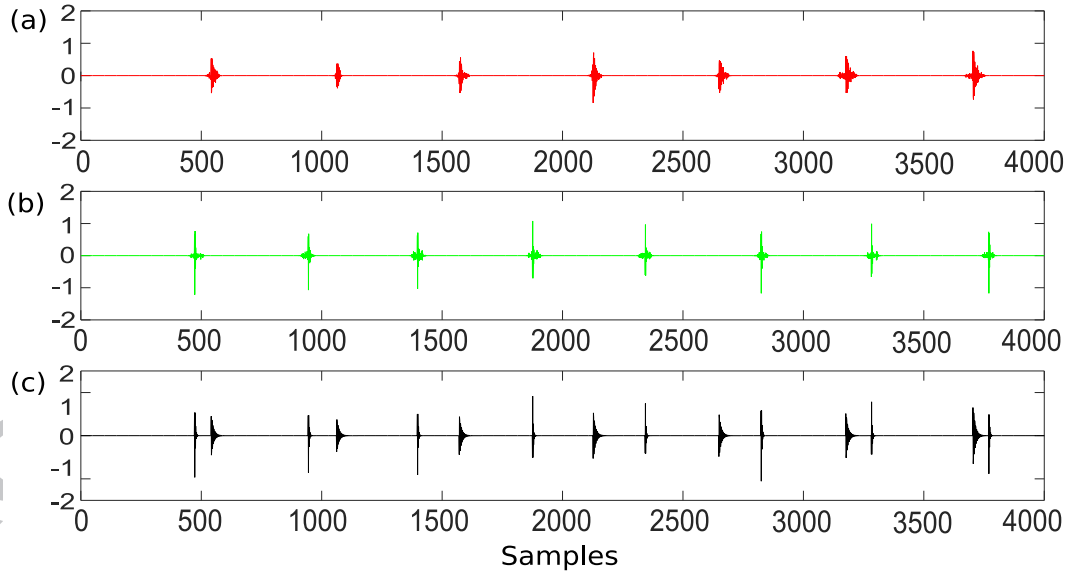


Figure 16. Reconstructed repetitive transients corresponding to (a) the first component $\mathbf{X}^1(i)$ with period $T_1 = 530$ samples and (b) the second component $\mathbf{X}^2(i)$ with period $T_2 = 470$ samples. (c) Summation of two components $\mathbf{X}^1(i)$ and $\mathbf{X}^2(i)$.

6. Validation on vibration signals

This section illustrates the application of the proposed methodology on actual vibration signals, where the three goals of detection, identification and fault extraction are again addressed separately. It is also demonstrated that the proposed methodology applies to time-varying operating conditions (data captured during a run-up), as often encountered in industrial applications. Comparison are made with the reference semi-automated diagnosis method described in the introduction where the signals are first whitened, then processed with the fast kurtogram in order to estimate an optimal frequency band for computing the SES and finally the fault is extracted by bandpass filtering.

6.1. Case 4: diagnosis of a ball fault

Three typical types of fault (i.e. inner race, outer race and ball fault) are investigated in a dataset from the Vibrations and Acoustics Laboratory of the University of New South Wales (Sydney) [78]. The test-rig is a one-stage gearbox with primary and secondary shafts supported by ball bearings. Since it is often more difficult to identify a ball defect, particularly at incipient stage, this case is tested here. The spectrogram of the raw signal is displayed in Fig. 17. It is seen that there exists non-stationary components in the high frequency band above 10 kHz, whereas the low frequency range is dominated by high energy components related to the gearbox vibrations.

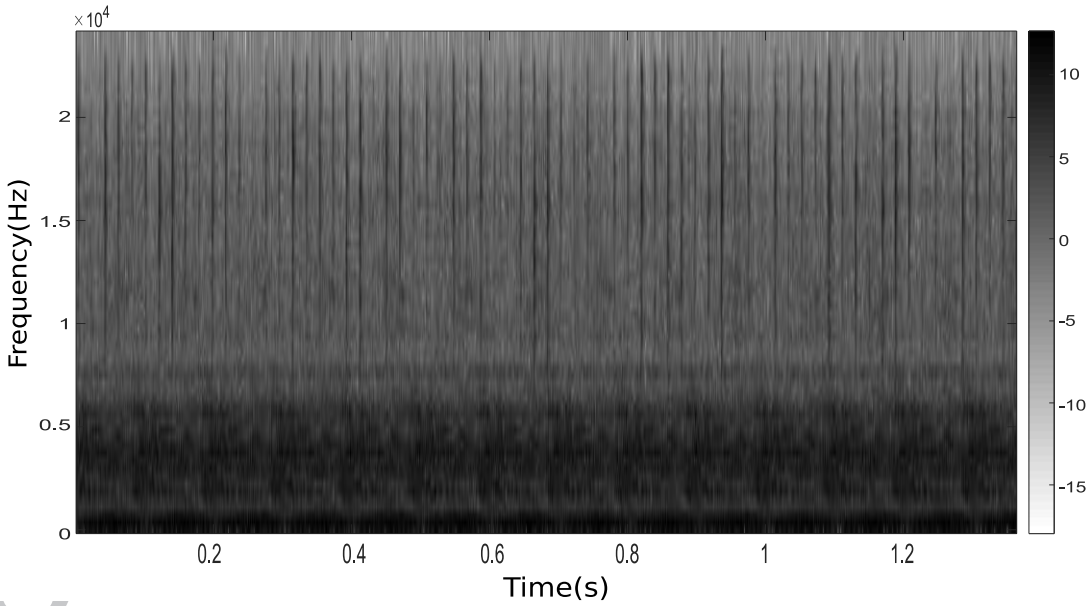


Figure 17. Spectrogram (logarithmic scale) of signal of Case 4 (frequency resolution $\Delta f = 375$ Hz).

The SES of the raw signal is displayed in Fig. 18. It is noted that there exists a relatively high value at the gearmesh frequency $f_g = 319.7$ Hz surrounded by shaft speed sidebands ($f_{\mathrm{rot}} = 10$ Hz) which originates from the gearbox. Clearly, the information of the bearing fault is completely masked by high-energy components from the gearbox in the SES. At this point, it is therefore important to resort to methods that automatically select frequency bands in the signal where the SNR is maximum. As stated in Ref. [78], the fast kurtogram has proved a powerful fourth-order spectral analysis tool for detecting and characterizing impulses in a signal. The fast kurtogram is applied here with $K = 7$ decomposition levels in a $1/3$-binary

tree. As seen in Fig. 19, there exists several local maxima in the kurtogram. They are coherent with the spectrogram of Fig. 17 which evidences a clear non-stationary activity above 10 kHz. All the dyads with very high kurtosis values have been checked to have similar complex envelopes. Therefore one maximum is taken at 71.83 whose corresponding to the frequency band [13500; 15000] Hz. The SES in that band is displayed in Fig. 20. It clearly reveals the even BSF (ball spin frequency) surrounded by modulation sidebands at cage speed (FTF).

Next, the HMM model is estimated with the parameter settings listed in Table . The detection test returns a value of 171, 160 for the GLRT to be compared to a statistical threshold of 42 with a risk of 5%, which clearly concludes to the presence of a fault. Next, the LLR spectrum is displayed in Fig. 21. Comparing with Fig. 20, the LLR spectrum better enhances the odd harmonics of the BSF than the SES from the kurtogram, even if the diagnostics information is very similar in both cases.

Table 3: Parameter settings in Case 4.

Sampling frequency F_s (Hz)	48000
Duration (s)	1.365
N_w	2^7
R	20
Rotation frequency – f_{rot} (Hz)	10
Ball spin frequency – BSF (Hz)	26.11
Fundamental train frequency – FTF (Hz)	4.08

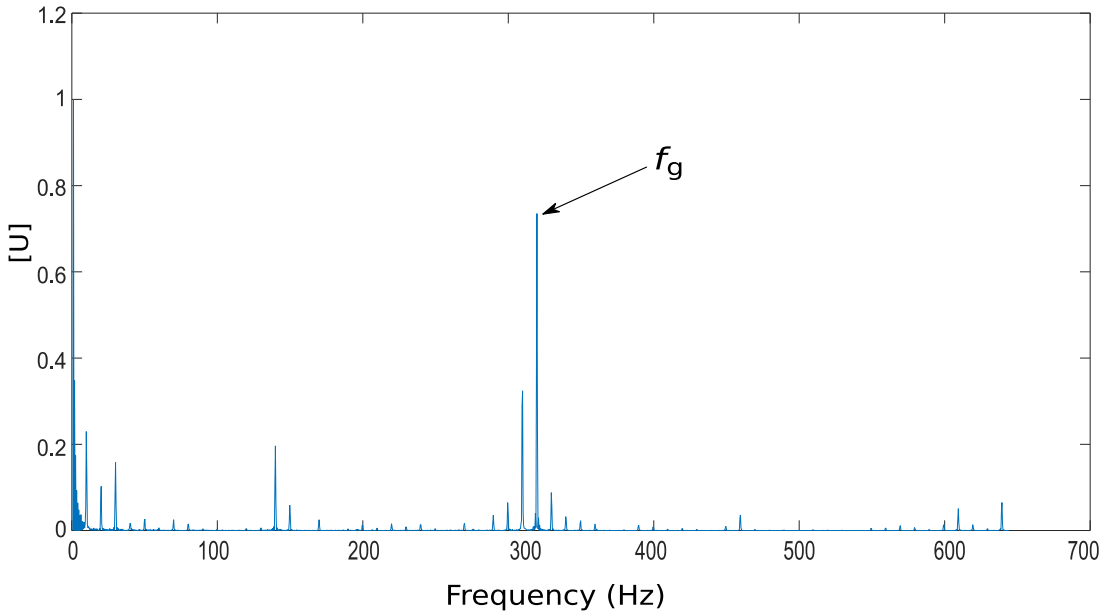


Figure 18: Squared envelope spectrum of the raw signal (normalized to unit maximum value).

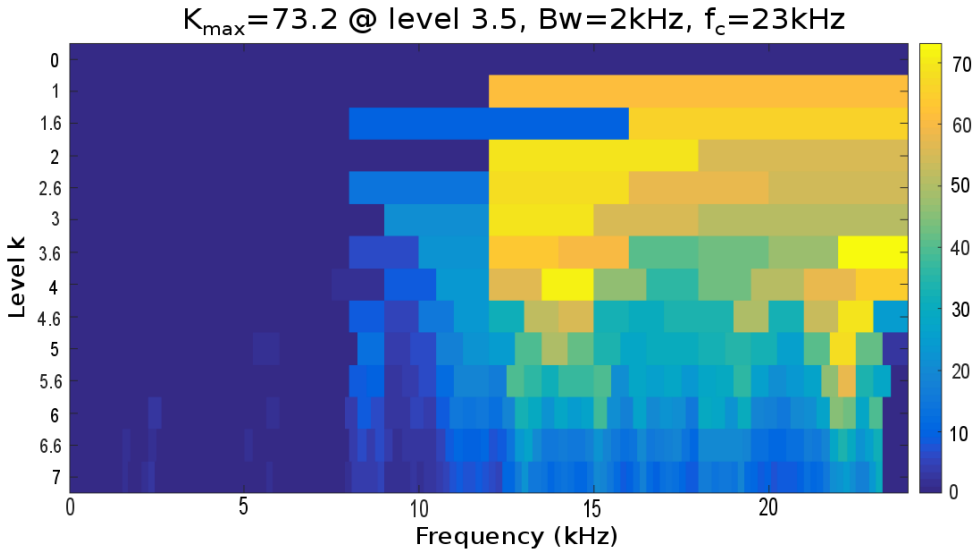


Figure 19. Kurtogram of signal of Case 4 computed over $K = 7$ levels with a 1/3-binary tree and a 8 coefficient prototype filter. Several local maxima are presented. One maximum is taken to the frequency band [13500; 15000] Hz.

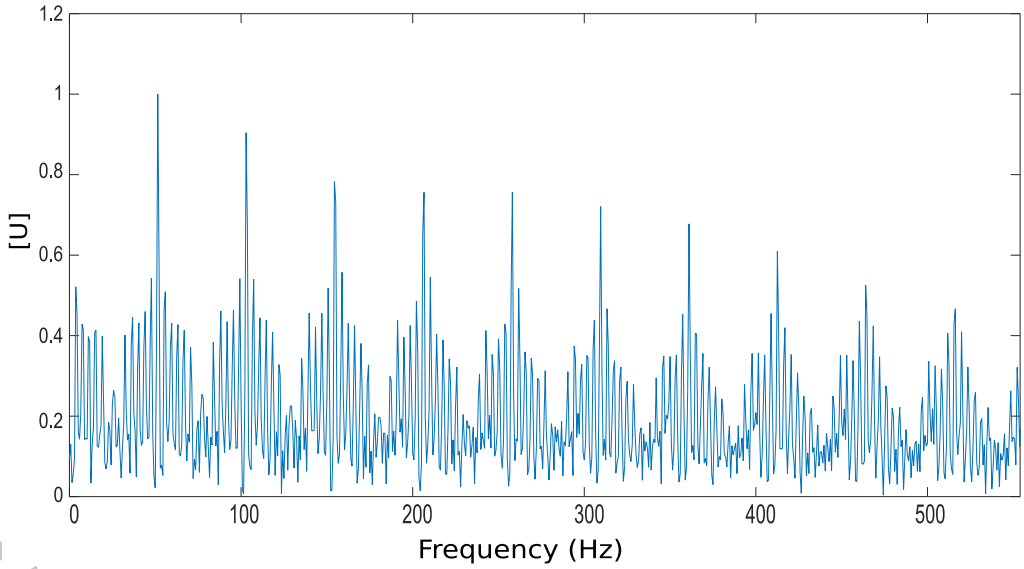


Figure 20. Squared envelope spectrum in frequency band [13500; 15000] Hz returned by the kurtogram (normalization to unit maximum value).

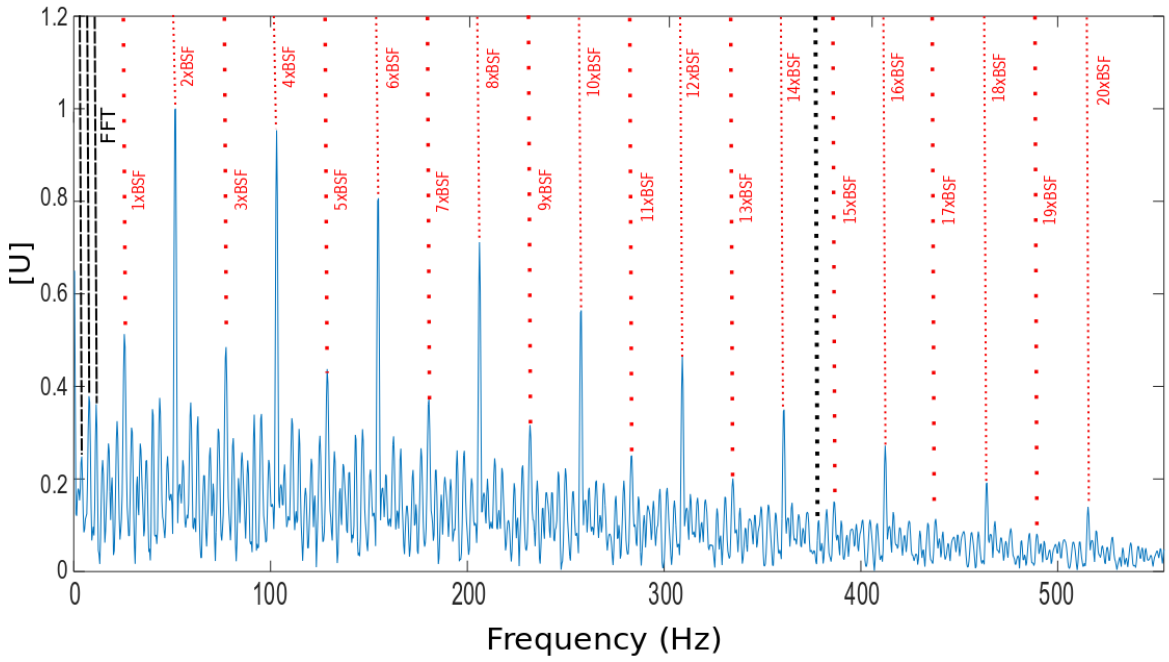


Figure 21: Spectrum of the LLR with markers at the theoretical fault frequency and harmonics; the limit $\alpha_{\max} = 375$ Hz is indicted by a vertical black dotted line(normalization to unit maximum value).

Before extracting the fault signal in the time domain, it is interesting to display the noise and signal spectra (diagonals of the covariance matrices $\hat{\mathbf{C}}_n$ and $\hat{\mathbf{C}}_x^1$). It is seen in Fig. 22 that the two spectra cross around 8 kHz, which is consistent with the two frequency bands identified in the spectrogram of Fig. 17. This reflects the fact that the high energy vibrations of the gearbox dominate the lower frequency range, whereas the repetitive transients dominate the higher frequency range.

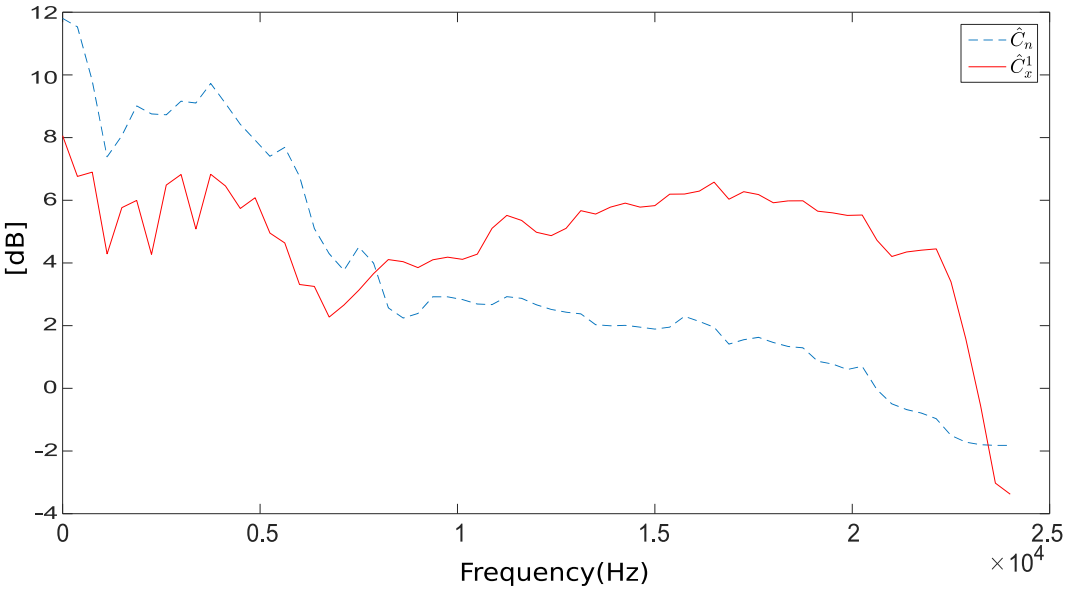


Figure 22. Diagonals of covariance matrices (frequency resolution $\Delta f = 375$ Hz); the spectrum of the fault signal is indicated by a red solid line and that of the noise by a blue dashed line.

The capability of the proposed HMM method to reconstruct the full-band fault signal is now demonstrated and compared with the band-pass results obtained from the kurtogram.Figure23 (a) displays 1.36 s of the vibration signal and Fig. 24 (a) an enlarged view in the vicinity of a transient. The band-pass filtered signal in band $[13500; 15000]$ Hz is displayed in Fig. 23 (b) and its enlarged view in Fig. 24 (b); it clearly evidences the presence of transients with maximum SNR. The reconstructed signal from the proposed HMM-based time-varying filter is displayed in Fig. 23 (c) and its enlarged view in Fig. 24 (c). As compared to the filtered signal based on the kurtogram, the reconstructed signal achieves an exact location of the transients with their full-band spectral content. This may be used advantageously to better characterize the fault signature, infer the fault dimension and spectral content, and possibly update trend models for prognostics.

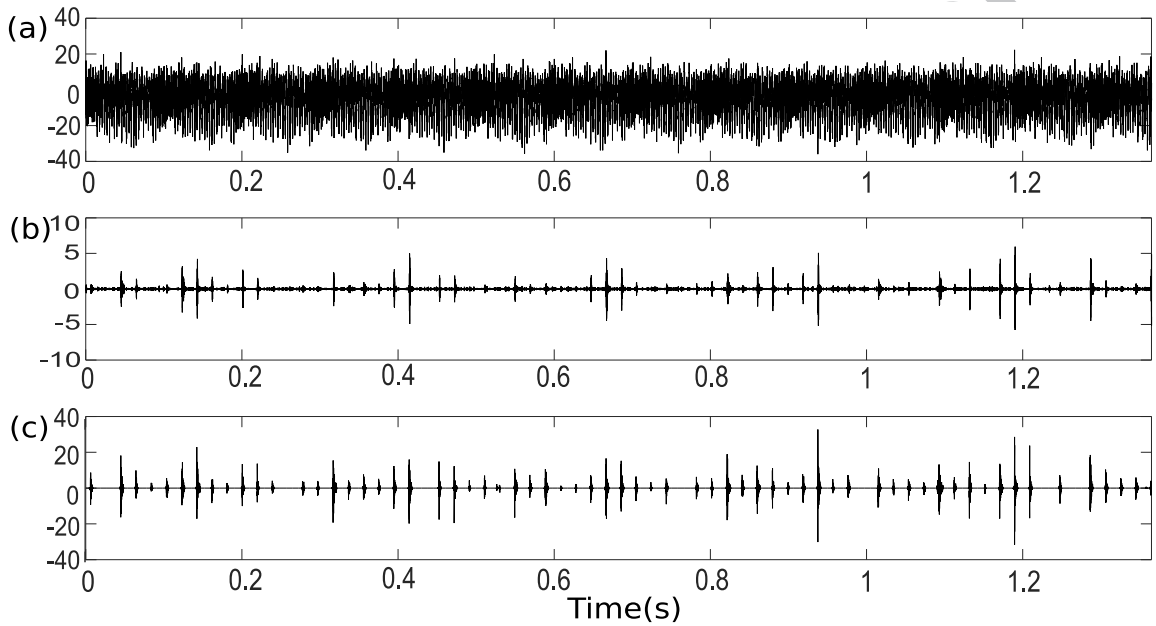


Figure 23(a) Vibration signal of Case 4 and (b) band-pass filtered signal in the frequency band $[13500; 15000]$Hz. (c) Full-band reconstructed fault signal from the proposed HMM-based time-varying filter.

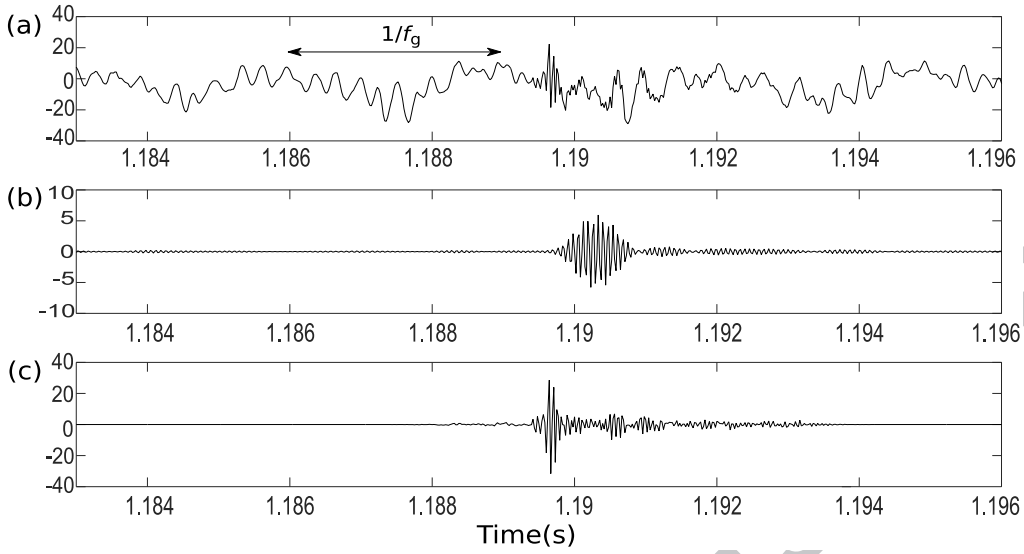


Figure 24(a) Enlarged view of vibration signal of Case 4 (with indication of the period corresponding to the peak f_g in Fig. 18) and (b) enlarged view of band-pass filtered signal in the frequency band [13500; 15000] Hz. (c) Enlarged view of the full-band reconstructed fault signal from the proposed HMM-based time-varying filter.

6.2. Cases5&6: diagnosis of bearing and gears

Signal recorded on another test rigare now considered. The test rig mainly consists of an electric asynchronous motor, a rotary encoder, 4 accelerometer sensors, a speed variator, a driving gear with 45 teeth, four bearings (3 healthy and 1 outer race fault) and two pinions (healthy and broken) – see Fig. 25.

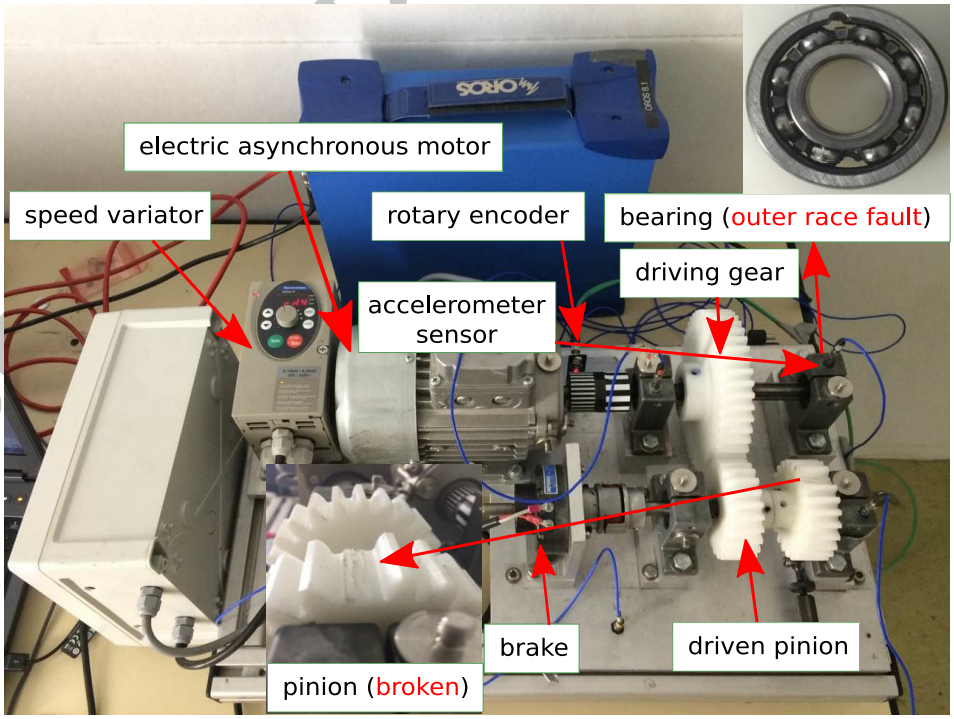


Figure 25. Test rig setup.

Two conditions are considered: 1) a damage bearing (outer race fault),and 2) a damage bearing (outer race fault) and a damaged gear together. It is highlighted here that the hole seen on the flank of the bearing in the upper right part of Fig. 25 is not the fault, but a through-bore used to seed a defect on the outer race. The actual size of the defect is about 2mm.

To demonstrate the effectiveness of the proposed method, two tests are considered hereafter: Case 5involvesa damage bearing (outer race fault) in cooperation with almost new gears and Case 6relates to the combination of the outer race fault and the broken pinion connected to the driving gear. The parameter settings for the HMM are listed in Table .

Table 4: Parameter settings in Case 5 and Case 6.

	Case 5	Case 6
Sampling frequency F_s (Hz)	51200	
Duration (s)	10	
N_w	2^7	
R	45	71
Main shaft rotation frequency – $f_{\mathrm{rot},1}$ (Hz)	$22.8 - 24.1$	$22.2 - 24.5$
Secondary shaft rotation frequency – $f_{\mathrm{rot},2}$ (Hz)	$1.875 \times f_{\mathrm{rot},1}$ $(42.8 - 45.2)$	$1.875 \times f_{\mathrm{rot},1}$ $(41.6 - 45.9)$

a) *Analysis of Case 5*

The detection test returns a value of 392,330 for the GLRT to be compared to a statistical threshold of 42 with a risk of 5%, which clearly concludes to the presence of a fault. Figure 26 shows the spectrum of the LLR which indicates the harmonic structure of the suspected fault frequencies BPFO (Ball Pass Frequency on the Outer Race). Meanwhile the repetitive transients (characterized by the fault frequency),and the residual noise are separated from the raw signal as shown in Fig. 27.

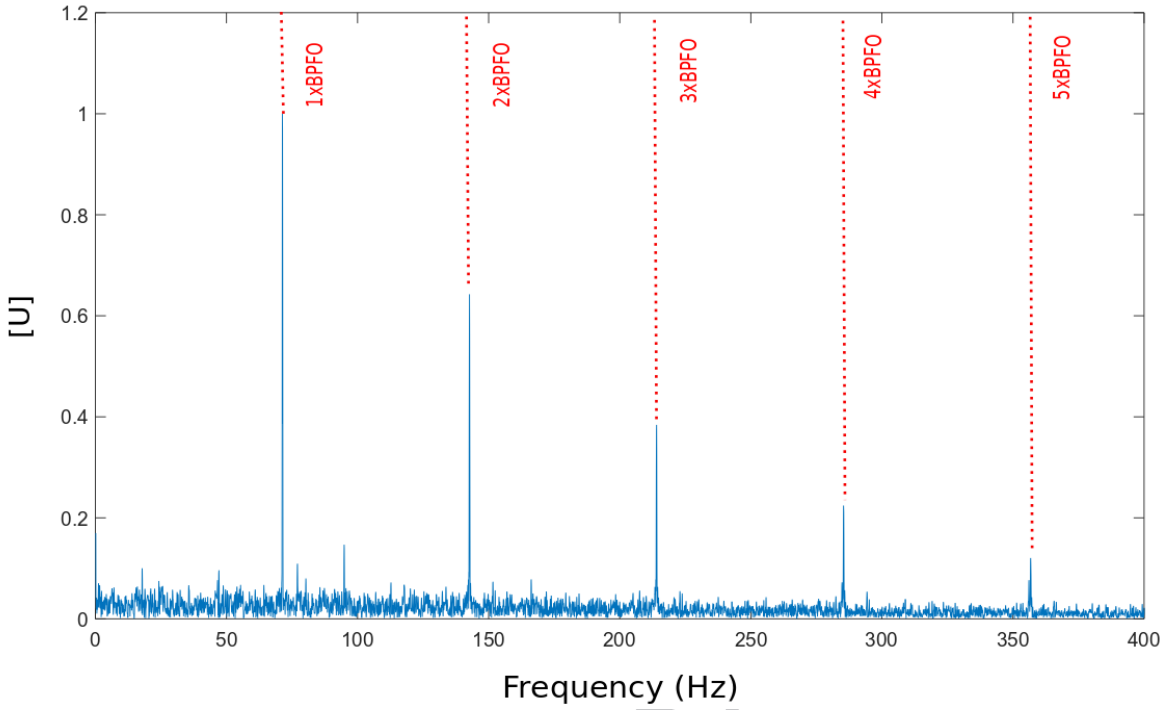


Figure 26: Spectrum of the LLR with markers at the suspected fault frequencies BPFOand its harmonics (normalization to unit maximum value).

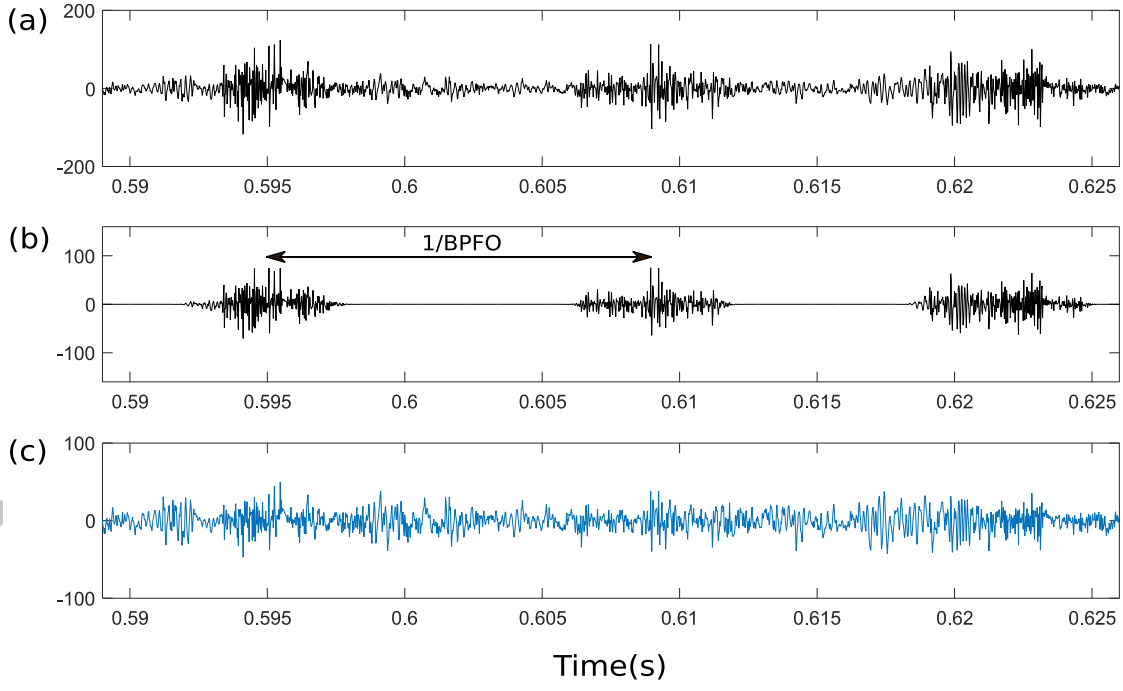


Figure 27(a) Vibration signal of Case 6 divided into (b)the full-band reconstructed fault signal $\hat{x}[n]$from the proposed HMM-based time-varying filter and (c) the noise (residual) signal ($\hat{n}[n] = y[n] - \hat{x}[n]$).

b) *Analysis of Case 6*

Using the same test rig and the same parameter settings, a compound source of vibration that contains gear and bearing faults together is diagnosed in Case 6. The detection test returns a value of 117, 640 for the GLRT to be compared to a statistical threshold of 42 with a risk of 5%, which concludes to the presence of a fault. Figure28 shows the spectrum of the LLR which reveals the harmonics of BPFO and the harmonics of the secondary shaft ($f_{\mathrm{rot},2}$), thus demonstrating the simultaneous presence of the outer race defect and of the gear defect. The extracted transients – signal $\hat{x}[n]$–and the residual noise – $\hat{n}[n]$ – are displayed in Fig. 29 and their respective SES's in Fig. 30. Since the bearing fault has a transient nature it is recovered in $\hat{x}[n]$ whereas the gear fault has a more stationary nature (with slight modulations) that is recovered in $\hat{n}[n]$.

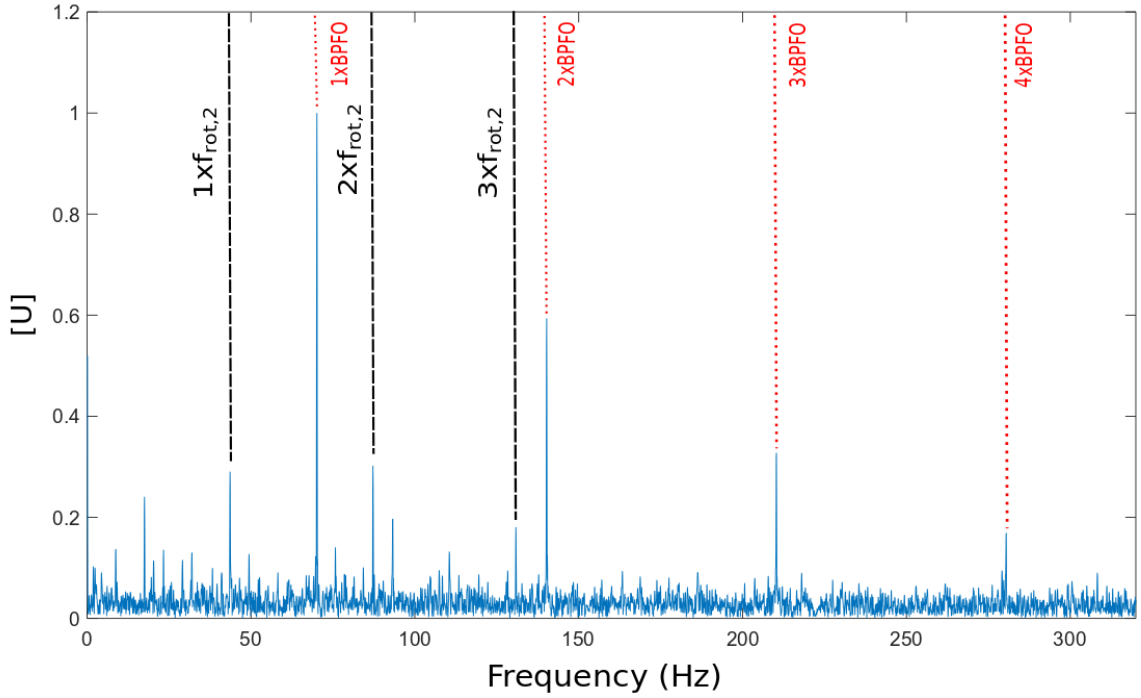


Figure 28: Spectrum of the LLR with markers at the suspected fault frequencies BPFO, its harmonics and $f_{\mathrm{rot},2}$ (normalization to unit maximum value).

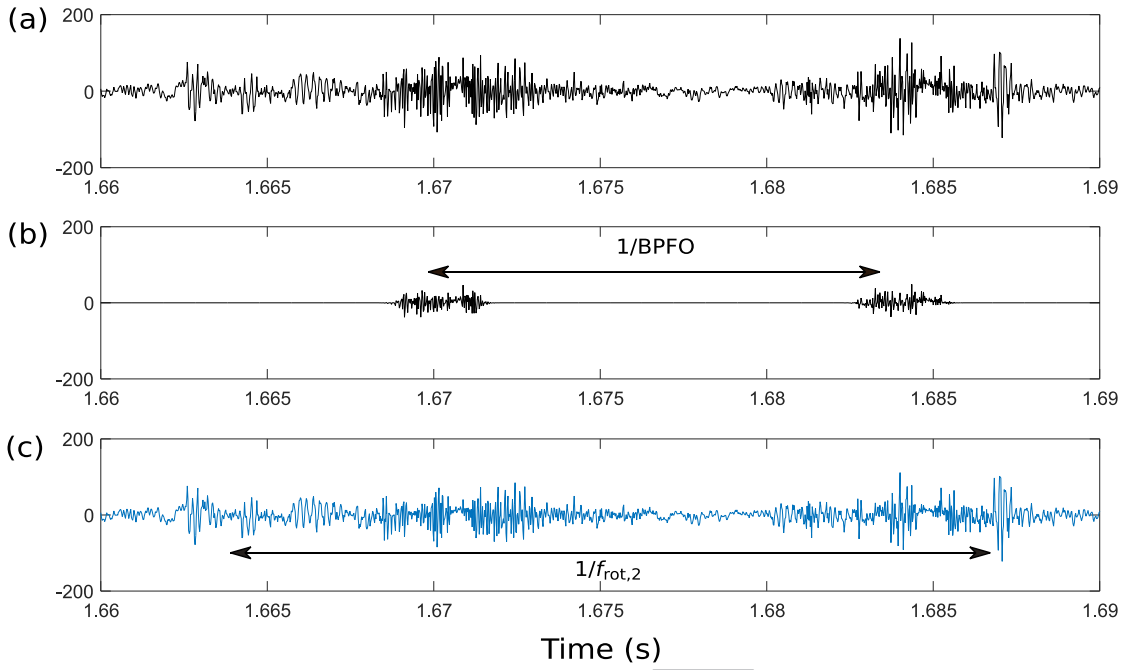


Figure 29 (a) Vibration signal of Case 6divided into (b) the full-band reconstructed fault signal $\hat{x}[n]$from the proposed HMM-based time-varying filter and (c) the noise (residual) signal ($\hat{n}[n] = y[n] - \hat{x}[n]$).

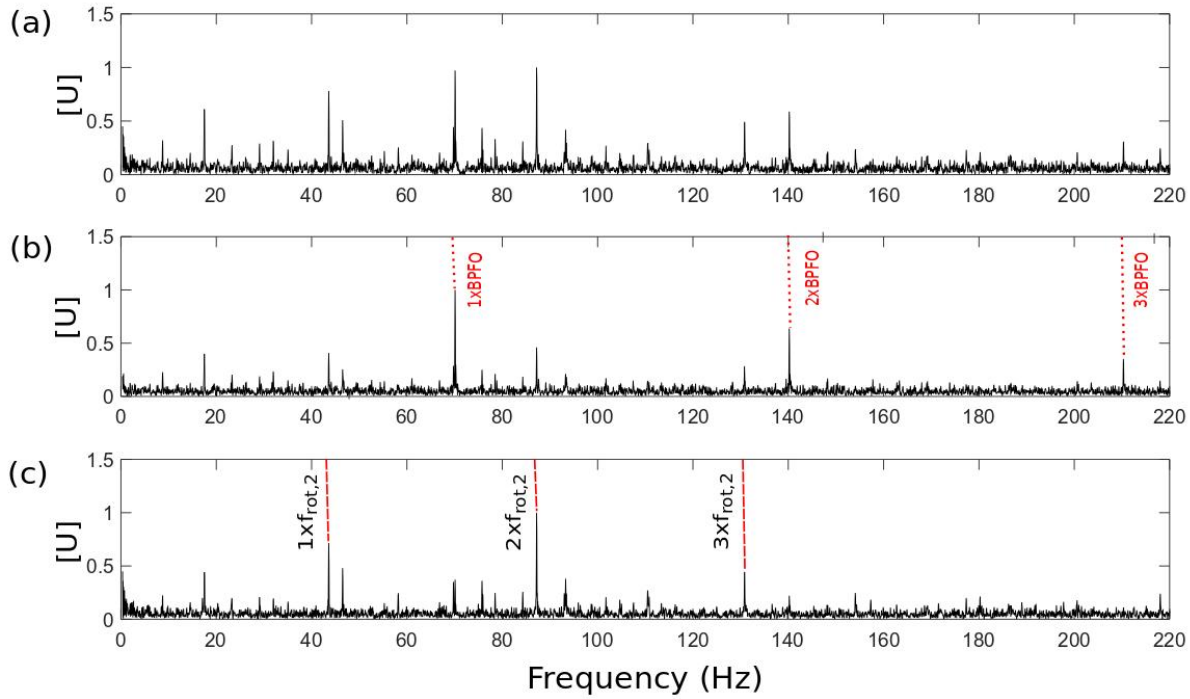


Figure 30. Squared envelope spectrum of (a) the vibration signal of Case 6, (b) the full-band recovered time signal and (c) the noise (residual) signal (normalization to unit maximum value).

6.3.Case 7: diagnosis of bearing in non-stationary operating conditions

This subsection illustrates the potential of the proposed methodology in variable operating conditions. A runup of $15\,s$ from $2\,Hz$ to $25\,Hz$ has been manually produced with the test rig of Fig. 25. The instantaneous speed is displayed in Fig. 31 (a) and the corresponding acceleration signal in Fig. 31 (b). The latter undergoes speed-dependent variations in magnitude and phase. Since it places no constraint on the distribution of the time instants of the impacts, the HMM can in theory cope with machine signals recorded in time-varying operating conditions. However, in order to account for the speed-dependent variation of the probability distribution of the states, the whole signal is divided into consecutive speed segments of 12 rotations as suggested in Ref. [79]– see Fig. 31 (c). Then, the parameters of the HMM are estimated separately on each segment. The probability of State 1 in the five operating conditions takes the following estimates:$\hat{\pi}$ =0.28, 0.30, 0.30, 0.33 and 0.37.

The detection test returns values of consecutive speed segments, i.e. 749,520, 195,640, 161,570, 155,220 and 339,850 for the GLRT to be compared to a statistical threshold of 42 with a risk of 5%, which concludes to the presence of a fault.

Figs. 32-33 display the raw signal on a few selected segments together with the corresponding reconstructed transients. It is seen that the transients are all well identified, although their behavior slightly changes with the rotation speed: the 4th segment corresponding to a speed around 20Hz exhibits the cleanest signature of the fault.

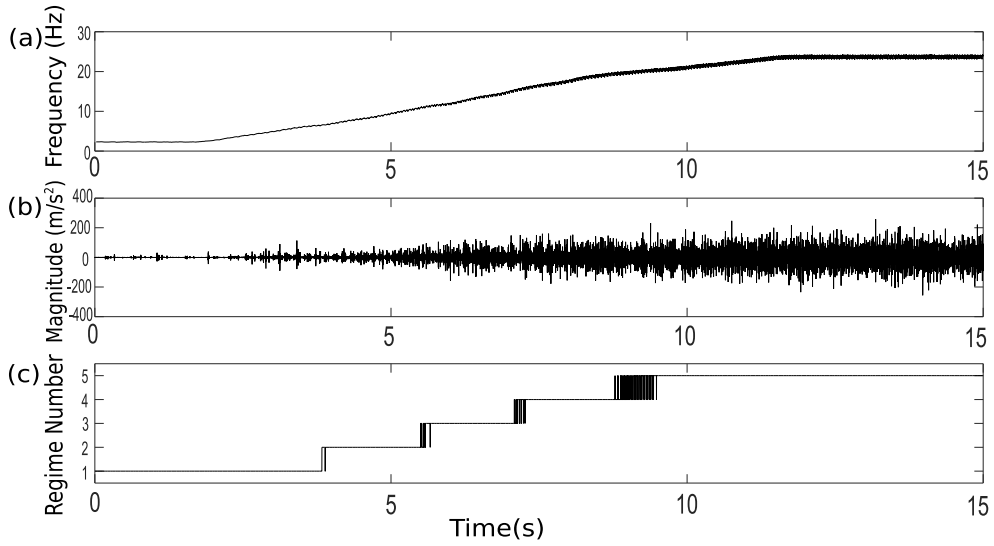


Figure 31.(a) Estimated instantaneous speed of signal in Case 7 and (b) its corresponding acceleration signal which undergoes speed-dependent magnitude modulation.(c) Division of the estimated instantaneous speed in 5 operating conditions.

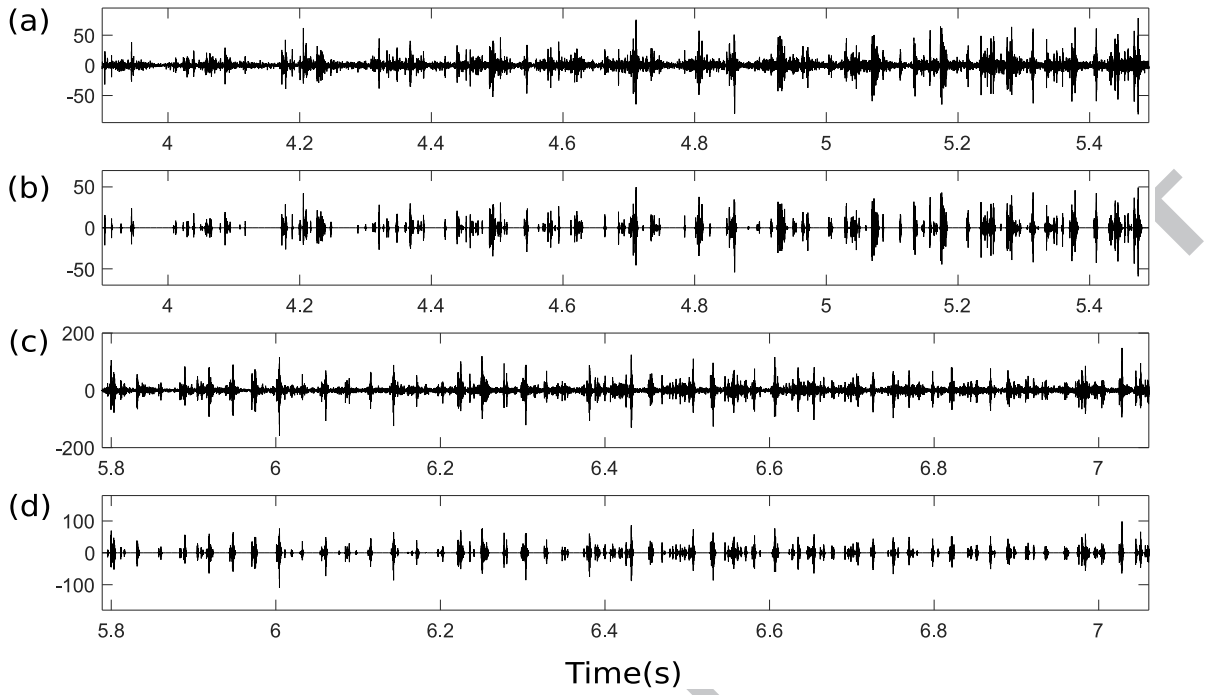


Figure 32. (a) Raw signal in operating condition No. 2 and (b) the corresponding reconstructed transients from the HMM-based time-varying filter; (c) raw signal in operating condition No. 3 and (d) the corresponding reconstructed transients from the HMM-based time-varying filter.

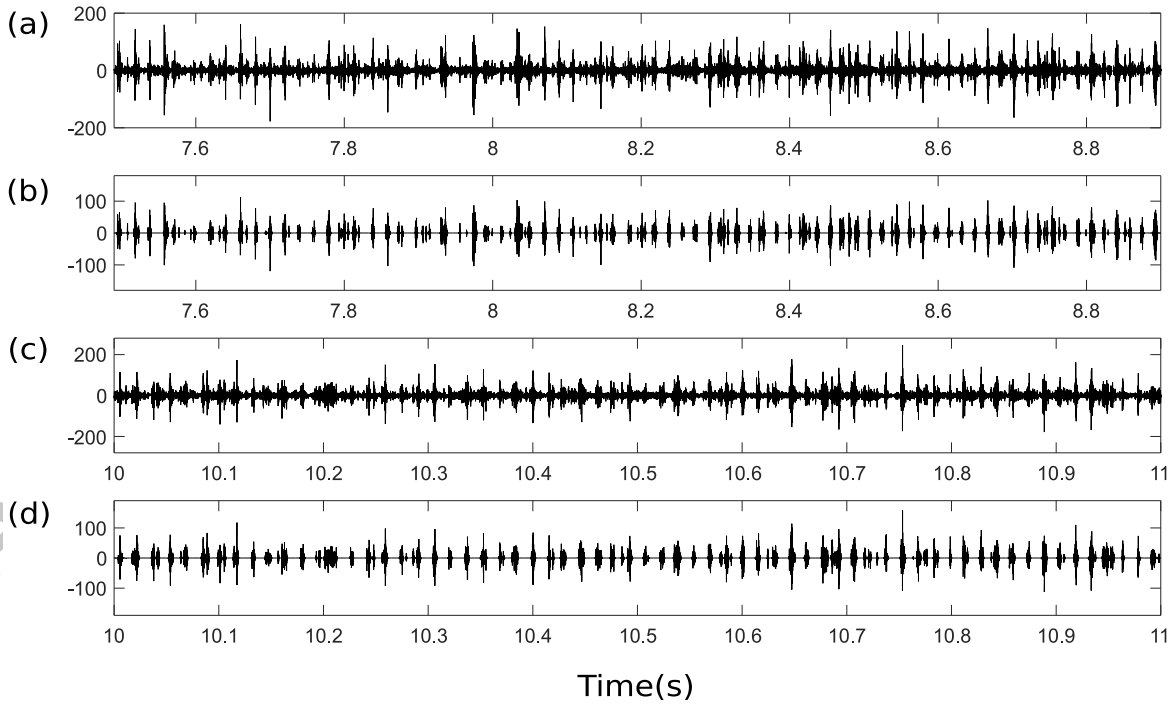


Figure 33. (a) Raw signal in operating condition No. 4 and (b) the corresponding reconstructed transients from the HMM-based time-varying filter; (c) raw signal in operating condition No. 5 and (d) the corresponding reconstructed transients from the HMM-based time-varying filter.

The identification of the fault characteristic frequency requires specific processing under non-stationary operating conditions. Since the defect impacts occur periodically with respect to the angular position, its frequency is to be computed in the order domain [80]. Therefore, the phase-corrected STFT in Eq. (2) have been resampled (using cubic splines interpolation) from the time to the angular domain, while maintaining a constant spectral bandwidth. The order spectrum of the LLR has then been computed on the resampled data. As seen in Fig. 34, it clearly reveals the presence of the Ball Pass Order on the Outer race (BPOO).

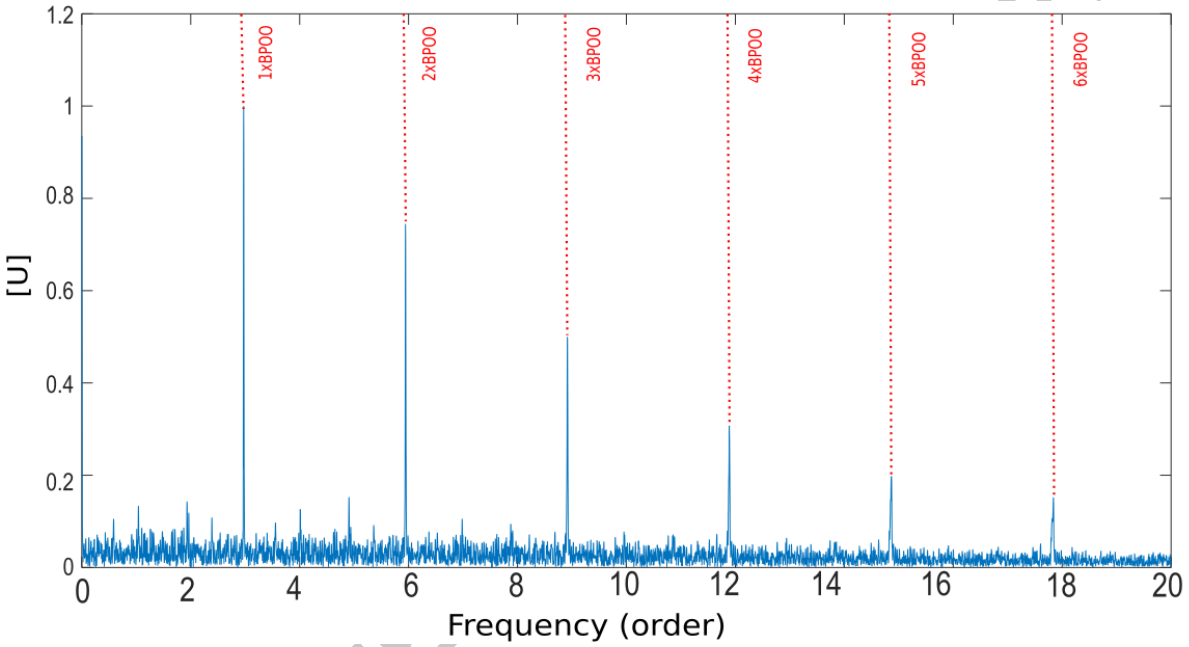


Figure 34. Order spectrum of LLR with markers at the suspected fault frequency BPOO and its harmonics (normalization to unit maximum value).

6.4. Case 8: diagnosis in the presence of multiple components

In order to demonstrate the performance of the HMM in the case of multiple-components, the dataset supported by the Department of Mechanical Engineering of Curtin University (Bentley) and made available online(http://data-acoustics.com/measurements/bearing-faults/bearing-1/) has been used. It corresponds to radial vibration measurements taken on the bearing housing of the SpectraQuest Machinery Fault Simulator test rig with a known outer race bearing fault. This case is interesting since there exists two different probabilities of states, as shown in Fig. 35. Their corresponding spectrograms are displayed in Fig. 36. Table presents the parameter settings used in Case 8.

Table 5: Parameter settings in Case 8.

Sampling frequency F_s (Hz)	51200
Duration (s)	10
N_w	2^7

R	45
Rotation frequency – f_{rot} (Hz)	29
Ball pass frequency, outer race – BPFO (Hz)	103.6
Ball pass frequency, inner race – BPFI (Hz)	157.4
Fundamental train frequency – FTF (Hz)	11.5
Ball (roller) spin frequency – BSF (Hz)	67.3

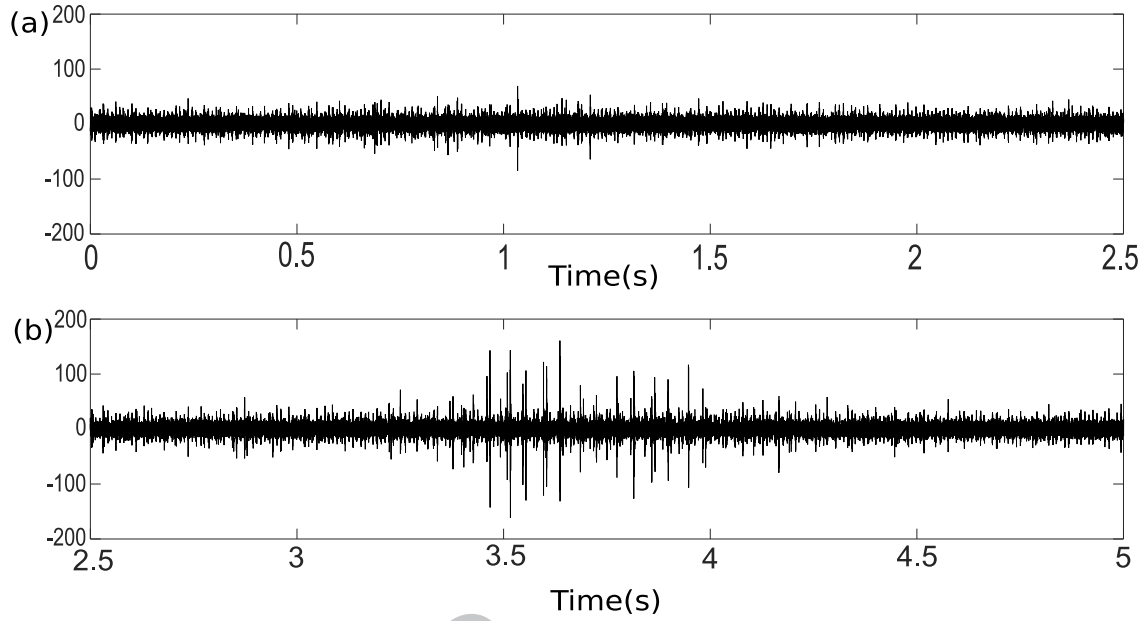


Figure 35. Measured signal(a) from 0 to 2.5s and (b) from 2.5 to 5s withevidence of interfering components in time interval[3.4 4]s.

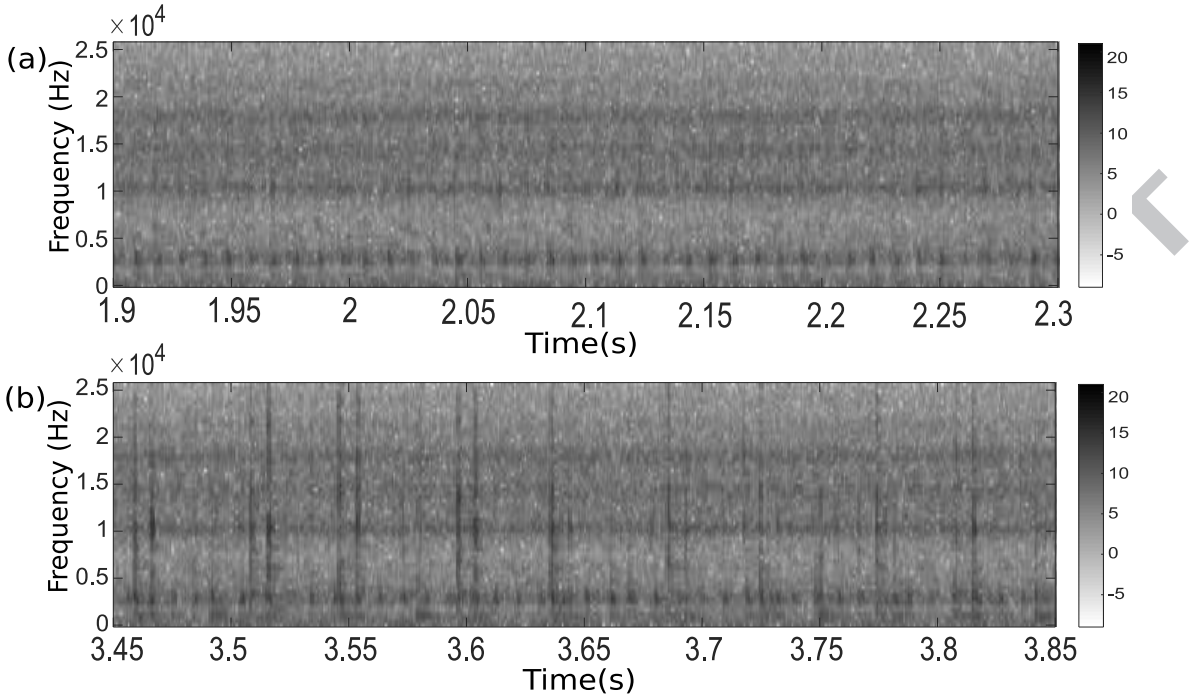


Figure 36. Spectrogram (logarithmic scale) of signal of Case 8 with evidence of two states: (a) global distribution with spectral content in band [1.8 3.8] kHz and (b) local distribution with spectral content in band [1.8 3.8] kHz and [9 11] kHz.

It is seen that two families of transients occur with different frequency contents. The proposed multiple-component model introduced in Section 2.2 has thus been used with K = 2. The estimated probabilities are $\hat{\pi}_3 =$ 0.634 (noise only), $\hat{\pi}_1 = 0.029$ and $\hat{\pi}_2 = 0.337$ for States 0, 1 and 2.Figure 37displays the estimated diagonals of the corresponding covariance matrices. The noise spectrum is found fairly flat, whereas the first component has a high energy around [1.8 3.8] kHz and [9 11] kHz and the second component has its energy concentrated around [1.8 3.8] kHz. The spectrum of the LLR of the second component $\mathbf{X}^2(i)$reveals the BPFO of a bearing fault, i.e. $f_3 = 102.8$ Hz, while that of the first component $\mathbf{X}^1(i)$shows some smeared component in the low frequency, around $f_1 = 22.86$ and $f_2 = 124.4$Hz (see Fig. 38).

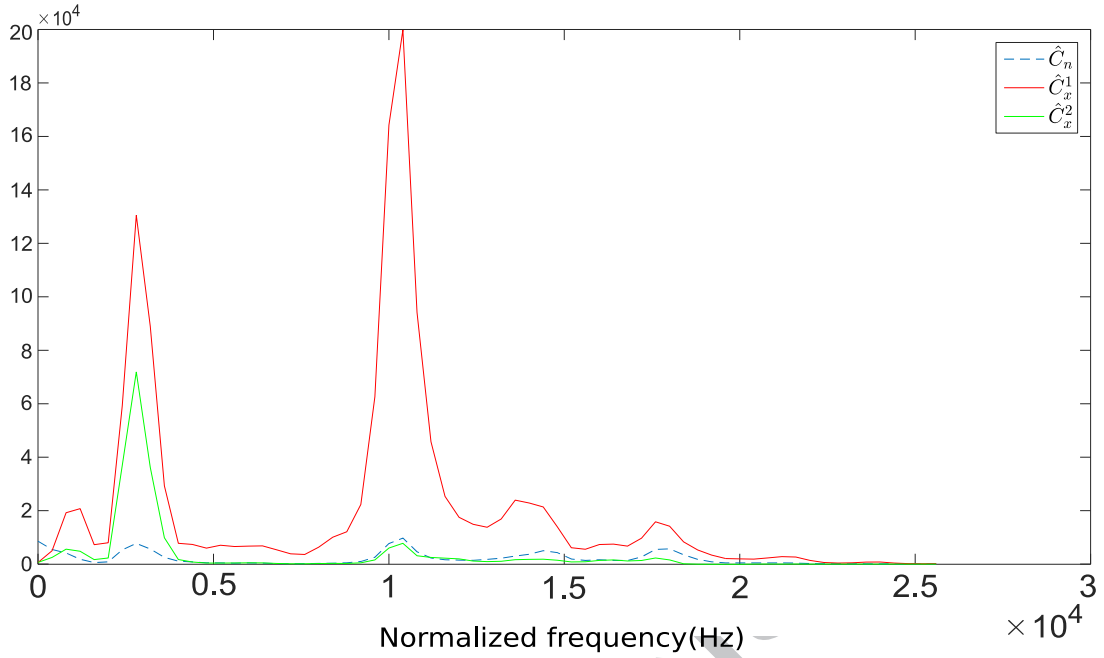


Figure 37. Diagonals of covariance matrices of the three components, $\mathbf{X}^1(i)$ (red solid line), $\mathbf{X}^2(i)$ (green solid line) and noise (blue dashed line).

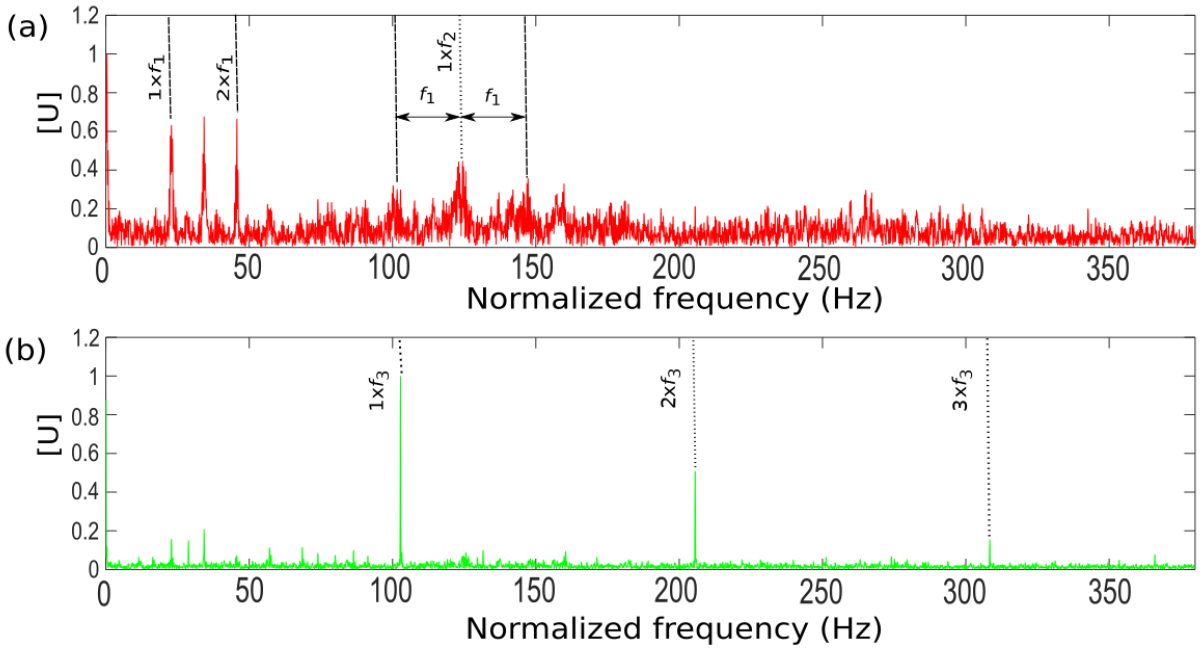


Figure 38. Spectra of the LLR of the two components: (a) $\mathbf{X}^1(i)$and (b) $\mathbf{X}^2(i)$, respectively (normalization to unit maximum value).

The reconstructed components are displayed in Fig. 39 and 40 in intervals$[1.9\ 2.3]\ s$ and $[3.45\ 3.85]\ s$. While the bearing fault is identified (f_3 = BPFO) in the second component $\mathbf{X}^2(i)$, the first component

$\mathbf{X}^1(i)$ corresponds to transients that occur in the second time interval only; since their frequency f_1 happens to be about twice the cage speed, it might indicate that the bearing is also misaligned.

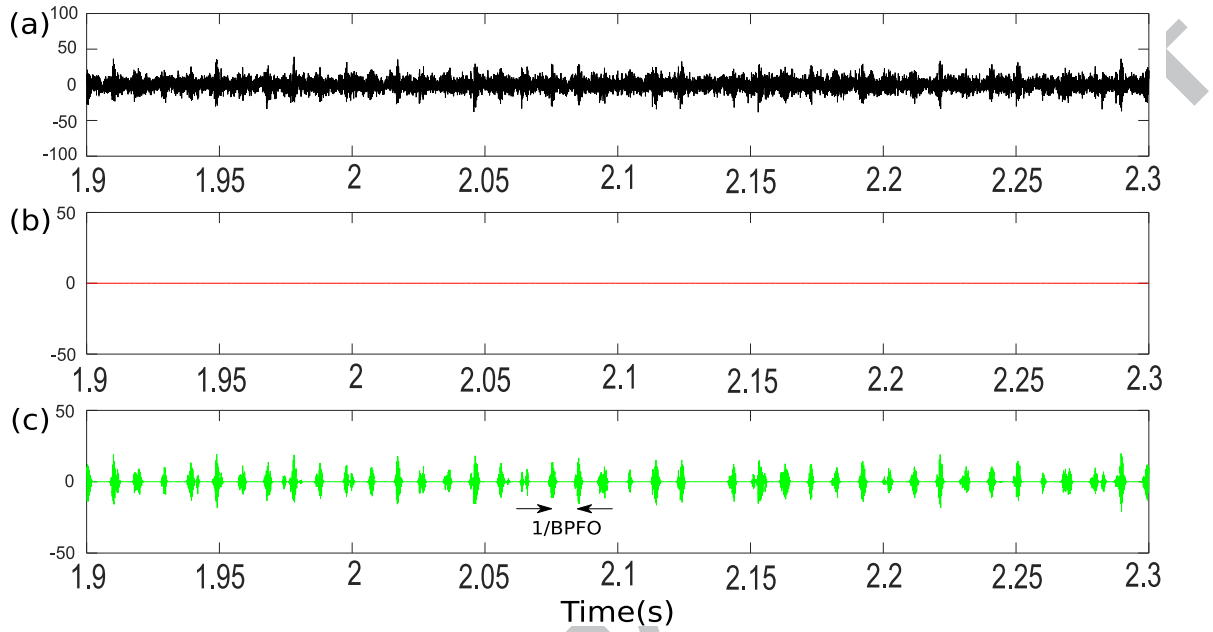


Figure 39. (a) Measured signal from 1.9 to 2.3s. Reconstructed components (b) $\mathbf{X}^1(i)$ and (c) $\mathbf{X}^2(i)$ from the HMM-based time-varying filter.

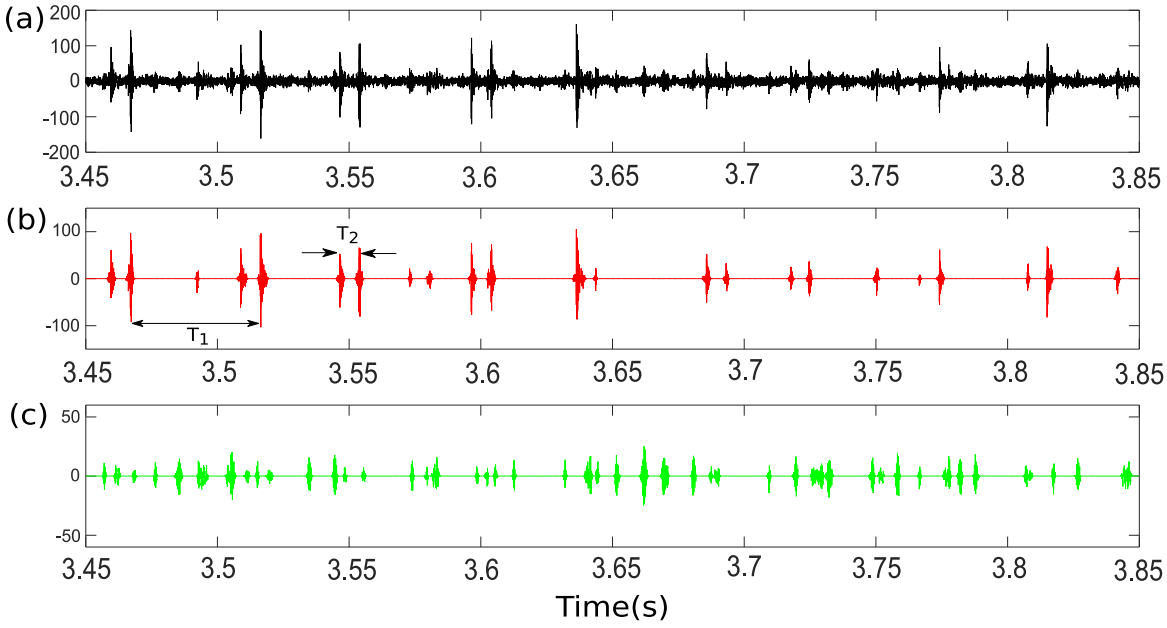


Figure 40. (a) Measured signal from 3.45 to 3.85s. Reconstructed components (b)$\mathbf{X}^1(i)$and (c) $\mathbf{X}^2(i)$from the HMM-based time-varying filter.

7. Conclusion

The HMM provides a versatile tool for characterizing the non-stationarity of a signal from its second-order statistics only, regardless of operating speed.

This paper has introduced a new stochastic model for representing the vibration signature of bearing faults in their incipient stage, when they have the typical form of a series of transients overlaid with stationary background noise. It is based on a HMM with intermittent switches between states characterized by Gaussian distributions with different covariance matrices. The main goal of the model is to provide a semi-automated diagnosis method that can handle at the same time the tasks of fault detection, fault identification and fault extraction. As compared to other semi-automated diagnostic solutions, the proposed one provides several advantages: first, the detection test is optimal in the Neyman-Pearson sense, second the spectrum of the LLR – a quantity equivalent to the SES– does not require pre-processing, and third the extraction of the fault signal is full-band and based on a time-varying filter superior to alternative stationary filters. The performance of the proposed method has been demonstrated on several vibration signals from test rigs (gear and bearing). Superior or equivalent results to the conventional methodology based on the fast kurtogram and envelope analysis and have been observed in all cases. Results are clearly superior in terms of reconstruction of the fault signal when compared to conventional band-pass filtering. The proposed model also deals with the case where there exists multiple components and with signals acquired under non-stationary operating conditions of the machine. The proposed methodology requires the setting of only one parameter, the frequency resolution; it has been verified to be quite robust in this respect provided that the frequency resolution is set greater than the potential fault frequencies. The performance is however based on the validity of the stochastic model and there might be cases which cannot be represented by the latter; in particular, a marked limitation, is that the proposed HMM applies only to incipient/localized faults with a transient nature. Nevertheless, the methodology introduced in the paper may open the way to the proposal of other types of stochastic models that can capture fault configurations not covered in the present work.

Acknowledgments

This work was supported financially by the Fundamental Research Funds for the Central Universities (Grant no. 2018RC009), the National Key Research and Development Program of China (Grant no. 2016YFB1200402), a scholarship from the China Scholarship Council (Grant no. 201304490040) and was performed within the framework of the Labex CeLyA of University of Lyon, operated by the French National Research Agency.

Reference

[1] R.B. Randall, J. Antoni, Rolling element bearing diagnostics-a tutorial, Mech. Syst. Signal Process. 25(2011) 485-520.

[2] W.A. Smith, R.B. Randall, Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study, Mech. Syst. Signal Process. 64(2015) 100-131.

[3] J. Antoni, F.Bonnardot, A.Raad, M.ElBadaoui, Cyclostationary modelling of rotating machine vibration signals, Mech. Syst. Signal Process. 18(2004) 1285-1314.

[4] B. Boashash, Time-Frequency Signal Analysis and Processing, second ed., Oxford, 2016.

[5] Z. Feng, M. Liang, F. Chu, Recent advances in time–frequency analysis methods for machinery fault diagnosis: a review with application examples,Mech. Syst. Signal Process. 38 (2013) 165-205.

[6] N. Harish Chandra, A.S. Sekhar, Fault detection in rotor bearing systems using time frequency techniques, Mech.

Syst. Signal Process. 72(2016) 105-133.

[7] A.Ghods, H. H. Lee, Probabilistic frequency-domain discrete wavelet transform for better detection of bearing faults in induction motors, Neurocomputing. 188(2016) 206-216.

[8] J.J. Shi, M. Liang, Intelligent bearing fault signature extraction via iterative oscillatory behavior based signal decomposition (IOBSD), Expert Syst. with Applications. 45(2016) 40-55.

[9] J. Yuan, Y. Wang, Y.Z. Peng, C.J. Wei, Weak fault detection and health degradation monitoring using customized standard multiwavelets, Mech. Syst. Signal Process. 94(2017) 384-399.

[10] Z.H. Lai, Y.Z. Leng, Weak-signal detection based on the stochastic resonance of bistableDuffing oscillator and its application in incipient fault diagnosis, Mech. Syst. Signal Process. 81(2016) 60-74.

[11] W.Guo, Z.M. Zhou, C. Chen, X. Li, Multi-frequency weak signal detection based on multi-segment cascaded stochastic resonance for rolling bearings, Microelectronics Reliability. 2017.

[12] B. Li, P.L. Zhang, Z.J. Wang, S.S.Mi, Y.T. Zhang, Gear fault detection using multi-scale morphological filters, Measurement. 44(2011) 2078-2089.

[13] Y.X. Wang, J.W. Xiang, R.Markert, M. Liang, Spectral kurtosis for fault detection, diagnosis and prognostics of rotating machines: A review with applications, Mech. Syst. Signal Process. 66(2016) 679-698.

[14] J. Antoni, Cyclic spectral analysis in practice. Mech. Syst. Signal Process.21(2007) 597-630.

[15] J. Antoni, G. Xin, N.Hamzaoui, Fast computation of the spectral correlation, Mech. Syst. Signal Process. 92(2017) 248-277.

[16] Z.H. Du, X.F. Chen, H. Zhang, B.Y. Yang, Z.Zhai, R.Q. Yan, Weighted low-rank sparse model via nuclear norm minimization for bearing fault detection, J. Sound Vib.400(2017) 270-287.

[17]P. D. Mcfadden, J. D. Smith, Vibration monitoring of rolling element bearings by the high-frequency resonance technique—a review, Tribology international.(17)1984 3-10.

[18] R. B. Randall,J. Antoni, S. Chobsaard, The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults and other cyclostationary machine signals. Mech. Syst. Signal Process.15(2001) 945-962.

[19] J. Antoni, Cyclic spectral analysis of rolling-element bearing signals: facts and fictions. J. Sound Vib. 304(2007) 497-529.

[20] J. Antoni, R. B. Randall, Differential diagnosis of gear and bearing faults. J. Vib. Acoust.124(2002) 165-171.

[21] D. Abboud, J. Antoni, S. Sieg-Zieba, M. Eltabach, Envelope analysis of rotating machine vibrations in variable speed conditions: A comprehensive treatment, Mech. Syst. Signal Process. 84(2017) 200-226.

[22] A.B. Ming, W. Zhang, Z.Y. Qin, F.L. Chu, Envelope calculation of the multi-component signal and its application to the deterministic component cancellation in bearing fault diagnosis, Mech. Syst. Signal Process. 50(2015) 1-31.

[23] B. Picinbono, On instantaneous amplitude and phase of signals, IEEE Transactions on Signal Process. 45(1997) 552-560.

[24] J. Antoni, The spectral kurtosis: a useful tool for characterising non-stationary signals. Mech. Syst. Signal Process. 20(2006) 282-307.

[25] J. Antoni,R. B. Randall, The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines. Mech. Syst. Signal Process. 20(2006) 308-331.

[26] P. BORGHESANI,P. PENNACCHI, S. CHATTERTON, The relationship between kurtosis-and envelope-based indexes for the diagnostic of rolling element bearings. Mech. Syst. Signal Process. 43(2014) 25-43.

[27] J. Antoni, The infogram: Entropic evidence of the signature of repetitive transients, Mech. Syst. Signal Process.74(2016) 73-94.

[28] D. Wang, An extension of the infograms to novel Bayesian inference for bearing fault feature identification, Mech. Syst. Signal Process. 80(2016) 19-30.

[29] C. Li, D. Cabrera, J. Valente de Oliveira, R.V. Sanchez, M.Cerrada, G.Zurita, Extracting repetitive transients for rotating machinery diagnosis using multiscale clustered grey infogram, Mech. Syst. Signal Process. 76(2016) 157-173.

[30] J. Wang, Q.B. He, F.R. Kong, Multiscale envelope manifold for enhanced fault diagnosis of rotating machines, Mech. Syst. Signal Process. 52(2015) 376–392.

[31] D.J. Yu, J.S. Cheng, Y. Yang, Application of EMD method and Hilbert spectrum to the fault diagnosis of roller bearings, Mech. Syst. Signal Process. 19(2005) 259-270.

[32] H. Endo, R. Randall, Enhancement of autoregressive model based gear tooth fault detection technique by the use of minimum entropy deconvolution filter, Mech. Syst. Signal Process. 21(2007) 906–919.

[33] J. Obuchowski, R. Zimroz, A. Wylomanska, Blind equalization using combined skewness-kurtosis criterion for gearbox vibration enhancement, Measurement: Journal of the International Measurement Confederation 88 (2016) 34–44.

[34] G. L. McDonald, Q. Zhao, Multipoint Optimal Minimum Entropy Deconvolution and Convolution Fix: Application to vibration fault detection, Mech. Syst. Signal Process. 82 (2017) 461–477.

[35] R. Boustany, J.Antoni, Blind extraction of a cyclostationary signal using reduced-rank cyclic regression—A unifying approach, Mech. Syst. Signal Process. 22(2008) 520-541.

[36] G. Cai, X. Chen, Z. He, Sparsity-enabled signal decomposition using tunable Q-factor wavelet transform for fault feature extraction of gearbox, Mech. Syst. Signal Process. 41 (2013) 34–53.

[37] W. He, Y. Ding, Y. Zi, I.W. Selesnick,Sparsity-based algorithm for detecting faults in rotating machines, Mech. Syst. Signal Process. 72 (2016) 46–64.

[38] H. Zhang, X. Chen, Z. Du, R. Yan, Kurtosis based weighted sparse model with convex optimization technique for bearing fault diagnosis, Mech. Syst. Signal Process. 80(2016) 349–376.

[39] W. He, Y. Ding, Y. Zi, I.W. Selesnick, Repetitive transients extraction algorithm for detecting bearing faults, ,Mech. Syst. Signal Process. 84(2017) 227–244.

[40] X. Ding, Q. He, Time–frequency manifold sparse reconstruction: A novel method for bearing fault feature extraction, Mech. Syst. Signal Process. 80(2016) 392-413.

[41] A.Srividya, A. K. Verma, B. Sreejith, Automated diagnosis of rolling element bearing defects using time-domain features and neural networks, International Journal of Mining, Reclamation and Environment.23(2009).

[42] M. Cococcioni, B. Lazzerini,S. L. Volpi, Automatic Diagnosis of Defects of Rolling Element Bearings Based on Computational Intelligence Techniques. 2009 970-975.

[43] C. Castejón, O. Lara, J.C. García-Prada, Automated diagnosis of rolling bearings using MRA and neural networks, Mech. Syst. Signal Process. 24(2010) 289-299.

[44] K. C. Gryllias, C.Yiakopoulos, I. Antoniadis, Automated diagnostic approaches for deffective rolling element bearing using minimal training pattern classification methods, Engineering Asset Lifecycle Management: Proceedings of the 4th World Congress on Engineering Asset Management. 2009 862—876.

[45] Y. Lei, An intelligent fault diagnosis method using unsupervised feature learningtowards mechanical big data,IEEE Transactions on Industrial Electronics.63(2015) 3137-3147.

[46] S.A. Khan, J.M. Kim, Hindawi, Automated Bearing Fault Diagnosis Using 2D Analysis of Vibration Acceleration Signals under Variable Speed Conditions, Shock and Vibration. 2016.

[47] R. DiBiano, S.Mukhopadhyay, Automated diagnostics for manufacturing machinery based on well-regularized deep neural networks, Integration, the VLSI Journal. 58(2017) 303-310.

[48] H. Yang, J. Mathew, L. Ma, Basis pursuit-based intelligent diagnosis of bearing faults, Journal of Quality in Maintenance Engineering. 13(2007) 152-162.

[49] J. Rafiee, M.A. Rafiee, P.W. Tse, Application of mother wavelet functions for automatic gear and bearing fault diagnosis, Expert Systems with Applications. 37(2010) 4568–4579.

[50] W. Y. Chen, J. X. Xu,S. K. Panda, A study on automatic machine condition monitoring and fault diagnosis for bearing and unbalanced rotor faults, IEEE International Symposium on Industrial Electronics, 2011 2105-2110.

[51] Z. Yang, Hindawi, Automatic Condition Monitoring of Industrial Rolling-Element Bearings Using Motor's Vibration and Current Analysis, Shock and Vibration. 2015.

[52] T. Gerber, N. Martin, C. Mailhes, Time-Frequency Tracking of Spectral Structures Estimated by a Data-Driven Method, IEEE Transactions on Industrial Electronics, 52(2015) 1-11.

[53] M. Firla, Z.Y. Li, N. Martin, C. Pachaud, T. Barszcz, Automatic characteristic frequency association and all-sideband demodulation for the detection of a bearing fault, Mech. Syst. Signal Process. 80 (2016) 335–348.

[54] C. Wang, M. Gan, C. Zhu, Intelligent fault diagnosis of rolling element bearings using sparse wavelet energy based on overcomplete DWT and basis pursuit, Journal of Intelligent Manufacturing.28(2017) 1377-1391.

[55] N. Sawalhi, R.B. Randall, Semi-automated bearing diagnostics-three case studies, NON DESTRUCTIVE TESTING AUSTRALIA. 45(2008) 59.

[56] P. Borghesani, P. Pennacchi, R.B. Randall, N. Sawalhi, R. Ricci, Application of cepstrum pre-whitening for the diagnosis of bearing faults under variable speed conditions, Mech. Syst. Signal Process. 36 (2013) 370-384.

[57]C.Peeters, P. Guillaume, J.Helsen, A comparison of cepstral editing methods as signal pre-processing techniques for vibration-based bearing fault detection, Mech. Syst. Signal Process. 91(2017) 354-381.

[58] L. Barbini, A.P. Ompusunggu, A.J. Hillis, J.L. du Bois, A. Bartic, Phase editing as a signal pre-processing step for automated bearing fault detection, Mech. Syst. Signal Process. 91(2017) 407-421.

[59] C. Peeters, P. Guillaume, J. Helsen, VIBRATION DATA PRE-PROCESSING TECHNIQUES FOR ROLLING ELEMENT BEARING FAULT DETECTION, 23 rd International Congress on Sound & Vibration.

[60] N. Sawalhi, R.B. Randall, H. Endo, The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis, Mech. Syst. Signal Process. 21(2007) 2616-2633.

[61] V.Girondin, K.M.Pekpe, H. Morel, J.P.Cassar, Bearings fault detection in helicopters using frequency readjustment and cyclostationary analysis, Mech. Syst. Signal Process. 38(2013) 499-514.

[62] P.Borghesani, MdRifatShahriar, Cyclostationary analysis with logarithmic variance stabilisation, Mech. Syst. Signal Process. 70(2016) 51-72.

[63] P. Borghesani, J. Antoni, CS2 analysis in presence of non-Gaussian background noise – Effect on traditional estimators and resilience of log-envelope indicators, Mech. Syst. Signal Process. 90(2017) 378-398.

[64] D.R. Brillinger, Time Series: Data Analysis and Theory, SIAM: Society for Industrial and Applied Mathematics.

[65] Z. Li, Z. Wu, Y. He, Hidden Markov model-based fault diagnostics method in speed-up and speed-down process for rotating machinery, Mech. Syst. Signal Process. 19(2005) 329-339.

[66]D.A. Tobon-Mejia, K. Medjaher, N. Zerhouni,A mixture of gaussians hidden markov model for failure diagnostic and prognostic, IEEE International Conference on Automation Science and Engineering. 2010 338-343.

[67] Q. Miao, V. Makis, Condition monitoring and classification of rotating machinery using wavelets and hidden

Markov models, Mech. Syst. Signal Process. 21(2007) 840-855.

[68] H. Ocak, K.A. Loparo, HMM-based fault detection and diagnosis scheme for rolling element bearings. J. Vib. Acoust.127(2005) 299-306.

[69] T. Heyns, P.S. Heyns,J.P. De Villiers, Combining synchronous averaging with a Gaussian mixture model novelty detection scheme for vibration-based condition monitoring of a gearbox. Mech. Syst. Signal Process. 32(2012) 200-215.

[70] H. Ocak,K.A. Loparo, A new bearing fault detection and diagnosis scheme based on hidden Markov modeling of vibration signals, Acoustics Speech and Signal Process. 2001 3141-3144.

[71] D.A. Tobon-Mejia, K. Medjaher, N. Zerhouni, A data-driven failure prognostics method based on mixture of Gaussians hidden Markov models, IEEE Transactions on reliability.61(2012) 491-503.

[72] J. Yu, Health condition monitoring of machines based on hidden markovmodel and contribution analysis, IEEE Transactions on Instrumentation and Measurement.61(2012) 2200-2211.

[73]Schmidt, S., Heyns, P. S., De Villiers, J. P., A novelty detection diagnostic methodology for gearboxes operating under fluctuating operating conditions using probabilistic techniques. Mechanical Systems and Signal Processing, 100(2018), 152-166.

[74] Zhou, H., Chen, J., Dong, G., Wang, R., Detection and diagnosis of bearing faults using shift-invariant dictionary learning and hidden Markov model. Mechanical systems and signal processing, 72(2016), 65-79.

[75] T. Boutros, M. Liang, Detection and diagnosis of bearing and cutting tool faults using hidden Markov models, Mech. Syst. Signal Process. 25(2011) 2102-2124.

[76] J. Antoni,Cyclostationarity by examples. Mech. Syst. Signal Process. 23(2009)987-1036.

[77] J.A. Bilmes, A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, International Computer Science Institute.4(1998) 126.

[78] J. Antoni, Fast computation of the kurtogram for the detection of transient faults. Mech. Syst. Signal Process. 21(2007) 108-124.

[79] D. Abboud, J. Antoni,M. Eltabach, Speed-spectral whitening for enhancing envelope analysis in speed varying conditions. VISHNO 2014.

[80] D. Abboud, The spectral analysis of cyclo-non-stationary signals, Mech. Syst. Signal Process. (75)2016280-300.

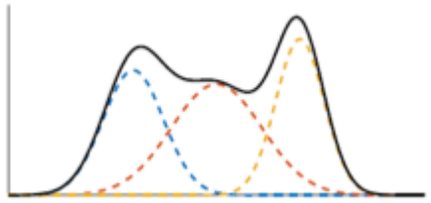
Highlights

- A mixture of Gaussians model is introduced for rolling element bearing vibrations.
- The model allows semi-automated diagnosis of bearing faults without need for pre-processing.
- It is optimal in the Neyman-Pearson sense for detecting repetitive transients.
- It allows full-band reconstruction of transients in the time domain.
- It applies under general assumptions, including time-varying operating conditions.

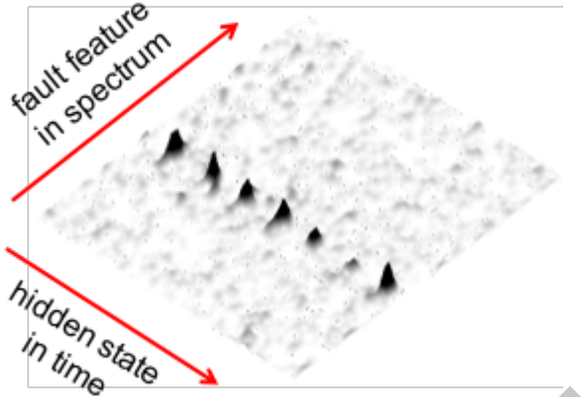
Model

Problem

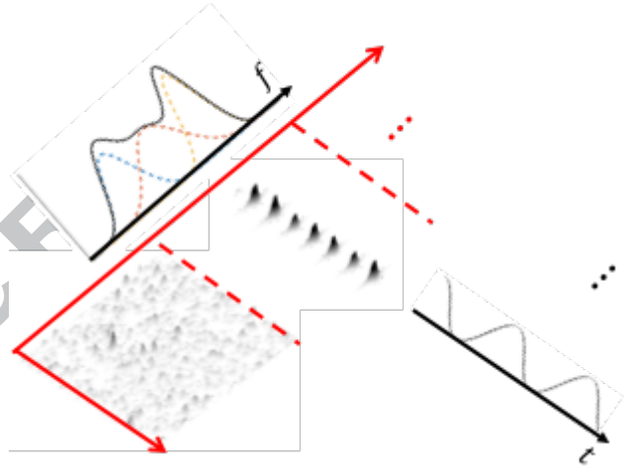
Solution



$$f(x) = \sum_{k=1}^{K} w_k f_k(x|\theta_k)$$



?



$$\boldsymbol{Y}(i) = \xi(i) \times \boldsymbol{X}(i) + \boldsymbol{N}(i)$$

$$\begin{cases} State\ 1: \boldsymbol{X}(i) + \boldsymbol{N}(i) \\ State\ 2: \boldsymbol{N}(i) \end{cases}$$

Hidden Variables: $\{\xi(i)\}_{i=1}^{N} \in \{0,1\}^{N}$