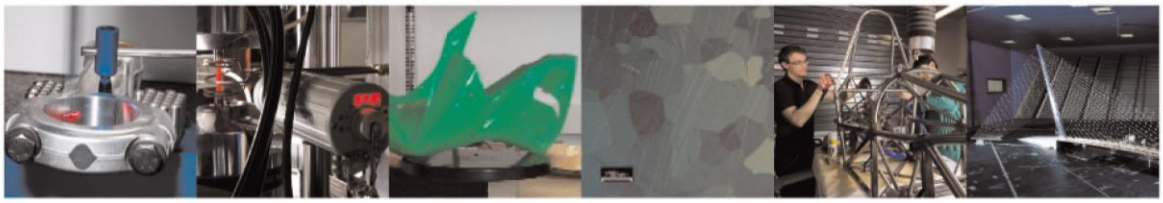

A data-driven method to enhance vibration signal decomposition for rolling bearing fault analysis

Grasso, M; Chatterton, S.; Pennacchi, P.; Colosimo, B.M.

This is a post-peer-review, pre-copyedit version of an article published in MECHANICAL SYSTEMS AND SIGNAL PROCESSING. The final authenticated version is available online at: http://dx.doi.org/10.1016/j.ymssp.2016.02.067

DIPARTIMENTO DI MECCANICA ■ POLITECNICO DI MILANO
via G. La Masa, 1 ■ 20156 Milano ■ EMAIL (PEC): pecmecc@cert.polimi.it
http://www.mecc.polimi.it
Rev. 0

A Data-Driven Method to Enhance Vibration Signal Decomposition for Rolling Bearing Fault Analysis

M. Grasso[1a], S. Chatterton[a], P. Pennacchi[a], B. M. Colosimo[a]

[a]*Dipartimento di Meccanica, Politecnico di Milano, Via La Masa 1, 20156 Milan, Italy*

[1] Corresponding author.

Tel.: (+39) 0523-623190

Fax: (+39) 0523-645268

e-mail: marcoluigi.grasso@polimi.it; steven.chatterton@polimi.it; paolo.pennacchi@polimi.it; biancamaria.colosimo@polimi.it;

Abstract— Health condition analysis and diagnostics of rotating machinery requires the capability of properly characterizing the information content of sensor signals in order to detect and identify possible fault features. Time-frequency analysis plays a fundamental role, as it allows determining both the existence and the causes of a fault. The separation of components belonging to different time-frequency scales, either associated to healthy or faulty conditions, represents a challenge that motivates the development of effective methodologies for multi-scale signal decomposition. In this framework, the Empirical Mode Decomposition (EMD) is a flexible tool, thanks to its data-driven and adaptive nature. However, the EMD usually yields an over-decomposition of the original signals into a large number of intrinsic mode functions (IMFs). The selection of most relevant IMFs is a challenging task, and the reference literature lacks automated methods to achieve a *synthetic* decomposition into few physically meaningful modes by avoiding the generation of spurious or meaningless modes. The paper proposes a novel automated approach aimed at generating a decomposition into a minimal number of relevant modes, called Combined Mode Functions (CMFs), each consisting in a sum of adjacent IMFs that share similar properties. The final number of CMFs is selected in a fully data driven way, leading to an enhanced characterization of the signal content without any information loss. A novel criterion to assess the dissimilarity between adjacent CMFs is proposed, based on probability density functions of frequency spectra. The method is suitable to analyze vibration signals that may be periodically acquired within the operating life of rotating machineries. A rolling element bearing fault analysis based on experimental data is presented to demonstrate the performances of the method and the provided benefits.

Keywords: Empirical Mode Decomposition, Combined Mode Functions, Vibration, Bearing, Fault Detection

Nomenclature

BSF	Ball Spin Frequency
$c_i(t)$	ith IMF extracted from the signal $Y(t)$, $i = 1, \dots, n$,
$c_{s_k}(t)$	kth sequential CMF extracted from the signal $Y(t)$, $k = 1, \dots, n$
$c^*_{s_k}(t)$	kth final CMF extracted from the signal $Y(t)$, $k = 1, \dots, K$
CMF	Combined Mode Function
$D_{k,k+1}$	distance (dissimilarity) between the kth and the (k+1)th IMFs from $Y(t)$
EMD	Empirical Mode Decomposition
$EEMD$	Ensemble Empirical Mode Decomposition
n	Number of IMFs extracted from the signal $Y(t)$
$f(x)$	probability density function of random process $x(t)$
$\hat{f}(x)$	kernel estimator of the probability density function of random process $x(t)$
F_s	sampling frequency
FFT	Fast Fourier Transform
FTF	fundamental train frequency
h	bandwidth of the kernel function
$\hat{h}$	optimal bandwidth of the kernel function
$h_u(t)$	difference between the signal $Y(t)$ and $m_u(t)$, at uth step of the sifting algorithm
K^*	number of iteratively generated CMFs
K	final number of CMFs extracted from the signal $Y(t)$
$Ker(x)$	kernel function
IMF	Intrinsic Mode Function
M	number of peaks in the $D_{k,k+1}$ function
$m_u(t)$	mean of envelopes at the uth step of the sifting algorithm
$\boldsymbol{p}$	vector of "locations" k corresponding to peaks in the $D_{k,k+1}$ function
$\tilde{\boldsymbol{p}}$	vector $\boldsymbol{p}$ with elements sorted in descending peak amplitude order
PDF	probability density function
q_k	number of IMFs included into the kth CMF, $k = 1, \dots, K$
r_i	normalized sample correlation coefficient between $Y(t)$ and the ith IMF
$r_n(t)$	residue of the EMD for the signal $Y(t)$
rms	root mean square
$SSB(K^*)$	sum-of-squares between the K^* CMFs from $Y(t)$

$SSW(K^*)$	sum-of-squares within the K^* CMFs from $Y(t)$
T	time window length
tol	tolerance threshold
UCV	Unbiased Cross Validation
$UCV_k(\hat{h})$	unbiased cross-validation statistic for kth CMF from $Y(t)$, with bandwidth $\hat{h}$
$w_k(\omega)$	weight function in the PDF for the kth CMF from $Y(t)$
$x_k(\omega)$	amplitude of frequency spectrum of the kth CMF from $Y(t)$
$Y(t)$	vibration signal
λ	threshold used in the index-based approach for IMF selection
$\rho(\cdot,\cdot)$	sample cross-correlation coefficient
ω	frequency location

1 Introduction

Sensor signals involved in health condition analysis of rotating machinery usually exhibit a multi-scale information content, due to the superimposition of features on different time-frequency scales, either stationary or non-stationary. Typical rolling bearing faults are caused by localized defects that generate impact vibrations. Thus, time-frequency analysis is a powerful approach to characterize both the time of impacts and the corresponding frequency ranges. Empirical Mode Decomposition (EMD) gained increasing influence in the technical literature to this aim. This kind of analysis relies on a decomposition of vibration signals into their embedded modes. Signal decomposition is a critical step that strongly influences the capability of isolating fault features and determining the health condition of the system. The achievement of a good characterization of the multi-scale content of the signal is of great importance to detect and diagnose faulty conditions in order to reduce plant downtime and to rapidly react to performance worsening caused by degraded states of the machine components. Although several multi-resolution techniques may be applied to this aim [1 – 2], the achievement of a synthetic decomposition into a minimal number of physically meaningful and interpretable oscillation modes still represents an open issue. In some cases, a number of decomposition levels is imposed *a-priori* and the signal is reconstructed by applying level-dependent thresholding techniques [3]. In many practical applications, the signal is first decomposed into a number of scales (usually larger than the one required to describe the relevant content), and a subset of modes of interest is then selected [4 – 6]. However, this latter approach yields a potential information loss. Furthermore, mode

selection may be a troublesome task in practice, being usually based on the human expert's knowledge and difficult to apply in an automatic way. Among time-frequency analysis techniques, the EMD proposed by Huang *et al*. [7] has several attractive properties that make it suitable to fault detection and diagnosis problems. The EMD is a nonparametric, data-driven and adaptive method that allows decomposing any signal into a number of Intrinsic Mode Functions (IMFs), without any prior basis selection. Due to its data-driven nature, the number of IMFs may vary over time, when the EMD is applied to periodically acquired signals. As an example, a higher frequency content in vibration signals caused by a defective bearing may lead to a larger number of IMFs than the ones extracted from the signal under healthy conditions, due to increased frequency ranges and energy levels [8]. In addition, the sifting algorithm is known to be affected by the so-called "*mode mixing*" problem [9], which may cause either a splitting of one intrinsic mode into two (or more) adjacent IMFs, or a merging of different scales into a single IMF. The EMD usually yields an over-decomposition of the signal, which can be inflated by the *mode mixing* effect and/or by the specific choice of the stopping criterion, leading to the presence of IMFs with no physical meaning. The introduction of the Ensemble Empirical Mode Decomposition (EEMD) [9] helped to mitigate mode mixing effects, but the EEMD is not able to avoid the over-decomposition imposed by the sifting algorithm. As a matter of fact, the literature devoted to the EMD and other multi-scale analysis methods lacks automated approaches for the achievement of a synthetic decomposition into a minimal number of relevant and interpretable modes. The combination of adjacent IMFs into the so-called Combined Mode Functions (CMFs) was proposed to synthetize the signal decomposition and to cope with split modes [6]. Such a combination can be interpreted as a new adaptive filter bank, which has the benefit of increasing the EMD accuracy [6]. Nevertheless, the literature lacks methods to automatically determine which IMFs should be summed together, because the proposed procedures rely on a visual inspection of the IMFs and the choice of problem-dependent criteria [5, 10 – 12].

The development of automated and data-driven tools to enhance the signal decomposition is expected to provide multiple benefits: (i) it may speed up the fault detection and diagnosis by improving the expert's decisional process, (ii) it may simplify (or even avoid) the selection of single modes of interest, and (ii) it allows implementing in-process monitoring functionalities.

This paper proposes a novel and automated approach to enhance signal decomposition via EMD, which is suitable for vibration signal analysis. The methodology works by automatically converting the original IMFs into a minimal number of CMFs. It consists in combining together adjacent IMFs such that the final CMF decomposition allows capturing distinct signal features via a parsimony-oriented procedure. Each eventually generated CMF consists in adjacent IMFs with similar spectral

properties described in terms of the probability density function of their frequency spectrum. The optimal number of CMFs is selected by minimizing the dissimilarity between IMFs included into the same CMF.

This paper represents a follow-on of a previous study authored by Grasso *et al*. [13], which proposed an EMD-based approach for in-process monitoring of multi-scale signals. Those authors highlighted the need for an effective and automated approach to achieve a synthetic and suitable separation of the CMFs, by showing that monitoring performances may be considerably influenced by the final CMF decomposition [13].

The performances of the method are discussed by means of experimental data acquired in a rolling element bearing diagnostics application. Vibration signals under both healthy and faulty conditions are processed in order to demonstrate the benefits of the proposed methodology and possible critical issues that deserve future research. A comparison with benchmark EMD-based analysis is presented, to further highlight the potential of the proposed approach.

Section 2 introduces a real case study devoted to the fault analysis of an evolving faulty state of a rolling element bearing; Section 3 briefly reviews the theoretical background of EMD and CMF methodologies; Section 4 presents the proposed approach; Section 5 demonstrates the application of the proposed methodology in the rolling element bearing fault analysis application; Section 6 eventually concludes the paper.

2 A real case study

The experimental data employed in this paper are relative to the condition monitoring of a NU1040M1 cylindrical roller bearing installed on the driven end of a stubby shaft in a test-rig employed for endurance testing (bearing A in Fig 1). The non-driven end is equipped by a spherical roller bearing, while a variable direction in the vertical plane and magnitude load (0-67.32 kN) is applied on the central part of the stubby shaft. The housings of the two bearings at the end of the stubby shaft can be independently moved in vertical.

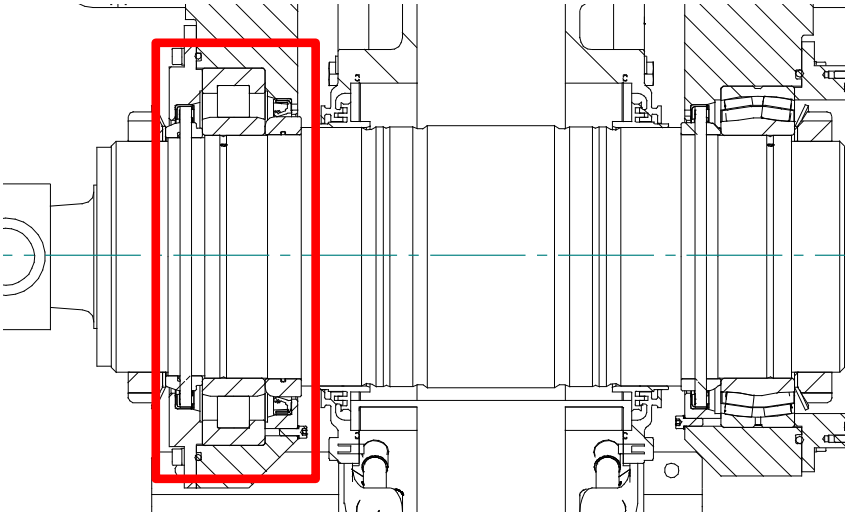

Fig. 1 – Bearing A on the stubby shaft of the test-rig

A long run-in has been performed (about 5 millions of bearing cycles from the beginning of the experimental activity), with the two seats aligned and load magnitudes within the bearing design and equal to 13.323 kN, at constant shaft rotational speed of 970 rpm. Then, an angular misalignment of the shaft with respect to the bearing housing of about 0.5° has been imposed. In this operating condition, a local overload of bearing A occurred, boosting the intentional premature failure of the bearing. Bearing A was monitored by a dual-probe (accelerometer and temperature sensor) SKF – CMSS 786T-IS and the data are acquired by using a NI PXI-1042Q chassis with a NI PXI-4472 board with sampling rate of 20 kS/s.

Fig. 2 - Rollers and cage of the damaged bearing after dismounting

Table 1 – Signals acquired in the real case study and corresponding epochs of machine bearing life

Data-sets	Epoch (millions of cycles)
$Y_1(t)$	2.03
$Y_2(t)$	27.5
$Y_3(t)$	34.4
$Y_4(t)$	36.2
$Y_5(t)$	38.4
$Y_6(t)$	Faulty bearing replaced by new one

Failure of bearing A occurred after a little less than 39 millions of bearing cycles from the beginning of the experimental activity, when high level of vibrations had been detected by the sensor. Visual inspection revealed the heavy deformed brass cage shown in Fig. 2 and the presence of metal debris mixed with oil residue. Flaking was also present on the roller surface.

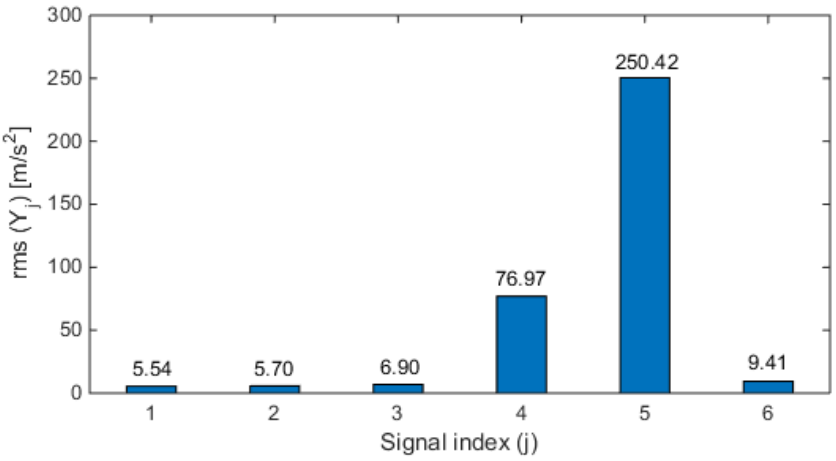

Fig. 3 – rms of bearing vibration signals $Y_1(t), \dots, Y_6(t)$

Whilst the vibration data of bearing A are available for all the test campaign, the data used in this paper refer to five different epochs of the machine life (see Table 1), from a healthy bearing condition to a severely faulty state before the substitution of the component. One additional data set was acquired after the installation of a new bearing, following the maintenance intervention. The row signals of the data sets, denoted by $Y_j(t)$, where $j = 1, \dots, 6$, correspond to 6 successive acquisitions within a time window of length $T = 5\ s$.

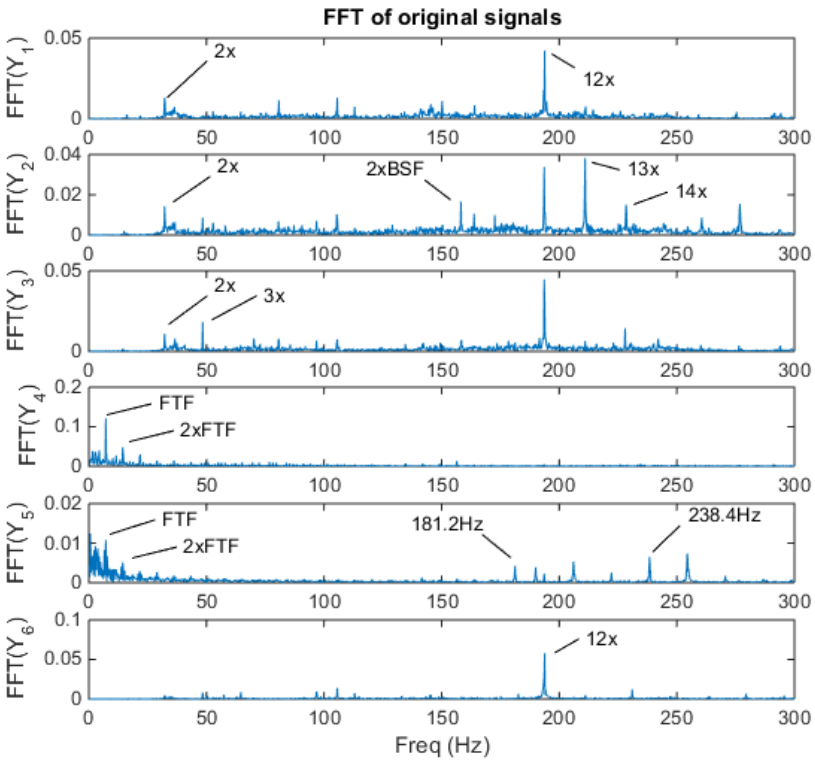

Fig. 4 – Frequency spectra of bearing vibrations signals $Y_1(t), \dots, Y_6(t)$

Fig. 3 shows the growth of the signal rms during the epochs described in Table 1. The rms remains about constant for signals $Y_1(t)$, $Y_2(t)$ and $Y_3(t)$, and it starts growing for signal $Y_4(t)$. The frequency spectra and the envelope spectra of the signals are shown in Fig. 4 and Fig. 5, respectively. Fig. 4 shows that the frequency content of the signal strongly changes passing from $Y_3(t)$ to $Y_4(t)$. In $Y_4(t)$ the energy of the signal shifts towards the low frequency range, being dominated by the Fundamental Train Frequency (FTF) at about $7.28\,Hz$ and its multiples. Fig. 5 shows the modulating effect imposed by the Ball Spin Frequency (BSF) at about $78.42\,Hz$ in $Y_4(t)$, and it also shows that such a modulation is already evident in $Y_3(t)$, which suggests that defects were present before the growth of the signal rms. Generally speaking, the rms provides no information about the time-frequency content of the signal, which means that it is not suitable for diagnostic purposes, but it is also poorly reliable for the early detection of a defect onset. In $Y_5(t)$ the modulation of the fundamental rotation frequency $(1x)$ predominates the envelope spectrum. When signal $Y_5(t)$ was acquired, the rolling element bearing was already in a severely faulty state, which justified its replacement.

The peaks at $f = 193.8\,Hz$ (visible in signals $Y_1(t)$, $Y_2(t)$, $Y_3(t)$ and $Y_6(t)$), $f = 211\,Hz$ (mainly visible in signal $Y_3(t)$), and $f = 228.4\,Hz$ (mainly visible in signal $Y_3(t)$) correspond to $12x$, $13x$ and $14x$ components, respectively. They are related to the inverter of the electrical motor driving the test rig.

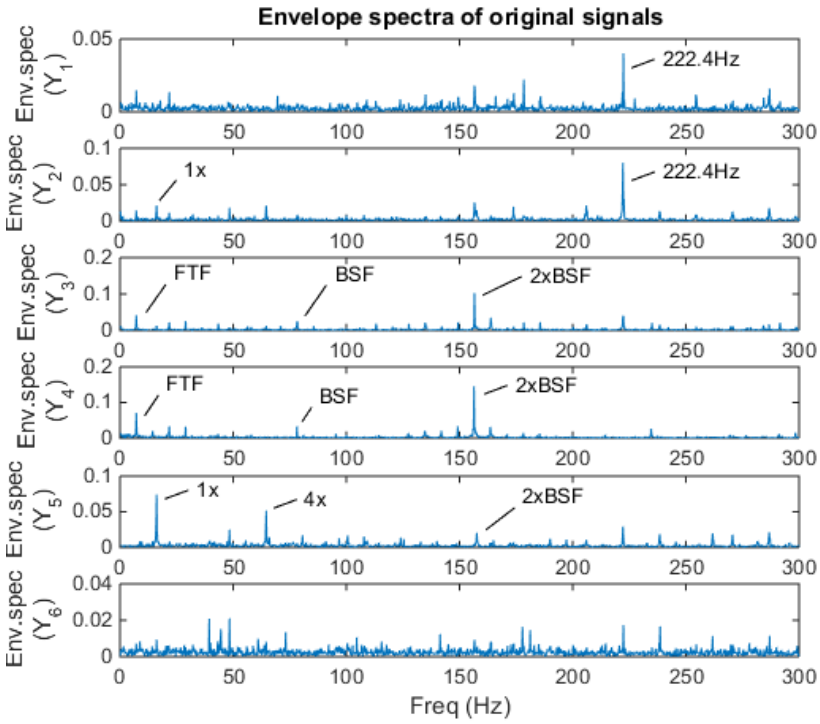

Fig. 5 – Envelope spectra of bearing vibrations signals $Y_1(t), \dots, Y_6(t)$

3 Theoretical background

Typical rolling element bearing defects modify the time-frequency content of vibration signals by affecting specific embedded oscillation modes, which motivates the use of multi-scale decomposition methods. Wavelet analysis represents one of the most employed signal processing techniques in a wide range of applications. Nevertheless, it is known to have different deficiencies [14 – 15] that reduce its effectiveness for rolling bearing fault detection, e.g., due to the incapacity of achieving fine resolutions in both time and frequency (i.e. scale) domains, simultaneously. In addition, the wavelet-based decomposition relies on the prior definition of a basis function (the *mother* wavelet) and other problem-dependent parameters to design a machine health monitoring and diagnosis tool. On the contrary, the EMD was proposed as a nonparametric alternative to the time-frequency methods [7]. It is a data-driven and adaptive technique that has a notable potential for automated inspections and in-process monitoring applications. In addition, the Hilbert-Huang spectrum, based on the EMD, allows one to overcome the simultaneous time and frequency resolution problem, providing a more effective alternative to wavelet analysis for rolling bearing diagnostics [15]. The use of EMD (or its variants) for bearing fault detection and diagnosis was investigated by several authors, including [6,

8, 16 – 18]. A review of the literature devoted to the application of EMD in rotating machinery diagnosis was presented by Lei *et al*., [19].

The EMD methodology exploits the so-called "*sifting*" algorithm [7] to decompose a signal $Y(t)$, into a number n of IMFs, which work as basis functions, and a residual term as follows:

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

where $c_i(t)$ is the i^{th} IMF and $r_n(t)$ is the residue obtained after extracting n IMFs.

A brief review of the sifting algorithm is presented in Appendix A. By way of example, Fig. 6 shows the EMD of signal $Y_1(t)$, i.e., the first bearing vibration signal considered, which corresponds to healthy conditions in the case study introduced in Section 1, whereas Fig. 7 shows the EMD of signal $Y_4(t)$, i.e., the one corresponding to a faulty condition and an increase of the vibration rms.

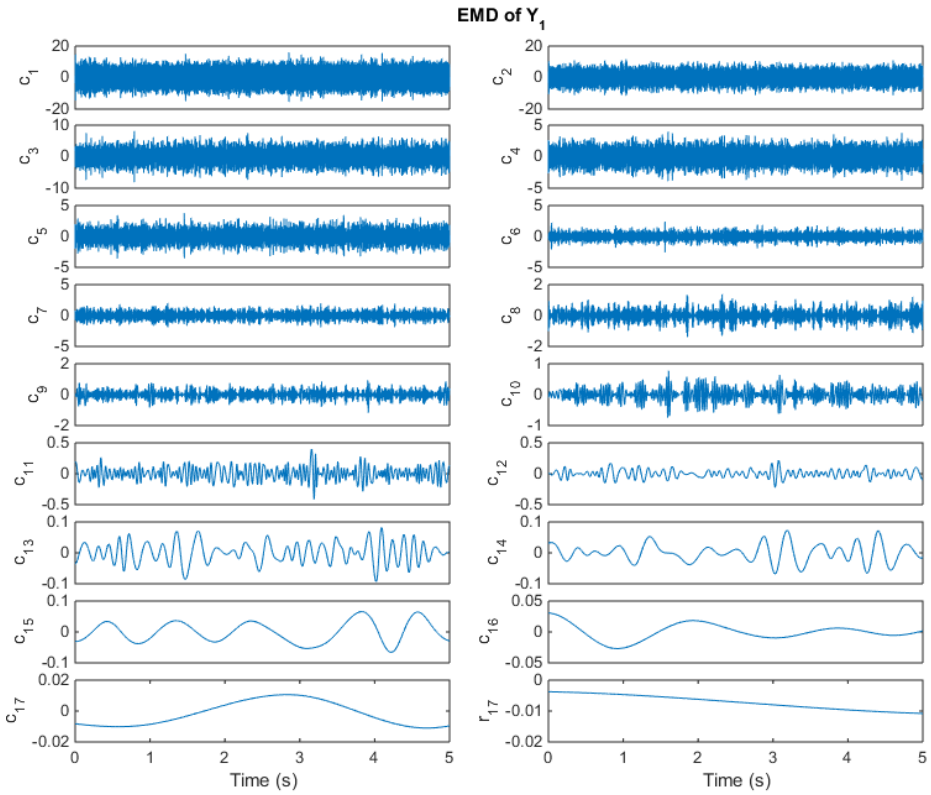

Fig. 6 – Empirical mode decomposition of bearing vibration signal $Y_1(t)$ – healthy conditions

Notice that the number of IMFs extracted from the two signals, denoted by n_1 and n_4, are different. In particular, the higher frequency content caused by the presence of defects yields, in this case, a larger number of IMF, $n_4 = 21$, than the one under healthy conditions, $n_1 = 17$. However, the number of IMFs is influenced by many additional factors, including the possible splitting or merging

of embedded modes caused by the mode mixing effect, which is influenced by the intermittency and noise properties of the signal itself. Fig. 7 clearly shows the defect-induced impacts on the vibration signal on different scales, which are absent in the EMD shown in Fig. 6. Nevertheless, the EMD results in an over-decomposition of the signals. The decomposition of both $Y_1(t)$ and $Y_4(t)$ generated some low-amplitude and meaningless-frequency components in the higher-order range of IMFs, i.e., the order that captures lower frequency regions. The presence of spurious modes further complicates the analysis and reduces the accuracy of fault feature extraction.

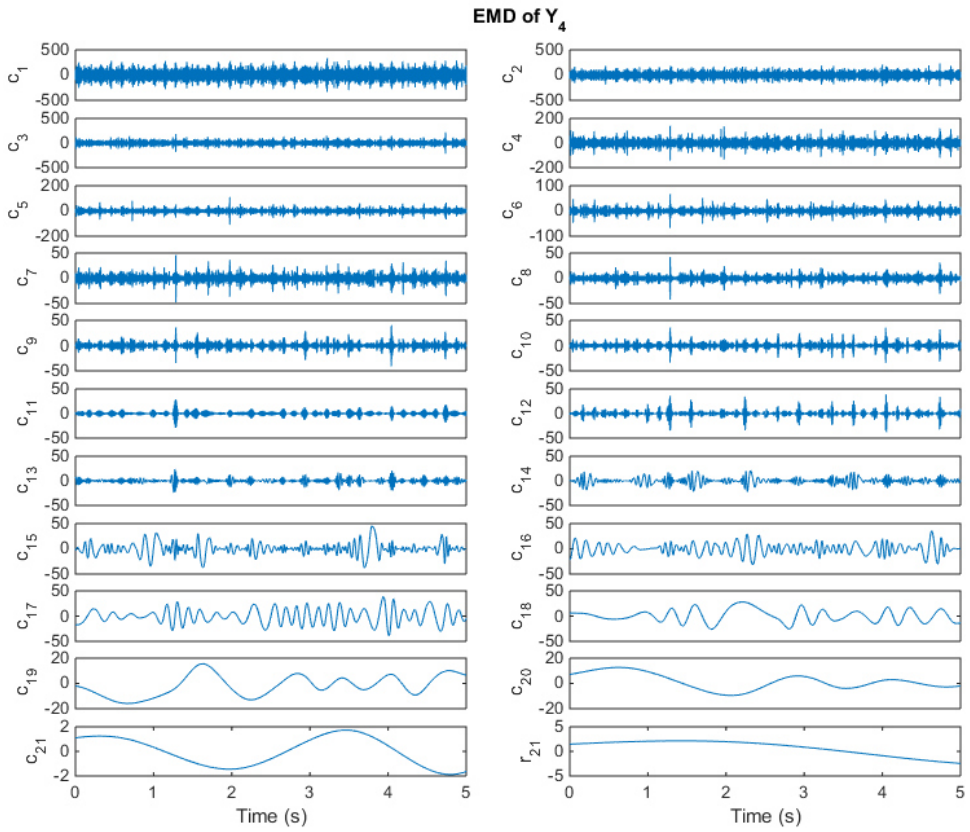

Fig. 7 – Empirical mode decomposition of bearing vibration signal $Y_4(t)$ – faulty conditions

The decomposition can be enhanced by avoiding the mode mixing effect. To this aim, Wu and Huang [9] proposed the EEMD approach, which consists in defining the "true" IMFs as the mean of an ensemble of trials, each one involving a sum of a white noise of finite amplitude to the original signal. The main limitation of EEMD is its high computational cost, because it requires the computation of a sufficient number of ensemble trials. Although some more efficient variants of the EEMD were proposed [20], the computational cost is still considerably higher than the one of the basic EMD. In addition, the usage of the EEMD methodology is not sufficient to avoid the over-decomposition of the signal, and post-processing analysis (often visual inspection) are usually applied to select and isolate relevant IMFs.

A more interesting and effective approach, to enhance the diagnostic relevance of the decomposition and to achieve a more synthetic representation of actually significant embedded modes, is the CMF approach [6]. This method consists in summing up adjacent IMFs $c_i(t), c_{i+1}(t), \dots, c_{i+q_k-1}(t)$ to obtain a new CMF, $c_{s_k}(t)$, as follows:

$$c_{s_k}(t) = c_i(t) + c_{i+1}(t) + \cdots + c_{i+q_k-1}(t), \quad k = 1, \dots, K \tag{2}$$

where q_k is the number of IMFs combined into the k^{th} CMF, $k = 1, \dots, K$, being $1 \leq q_k \leq n$ and $K \leq n$. The combination of adjacent IMFs allows one to cope with the over-decomposition produced by the sifting algorithm and the splitting of intrinsic modes into multiple IMFs. Gao *et al*. [6] showed that such a combination of subsets of IMFs can be interpreted as a new adaptive filter bank based on the intrinsic time scales of the signal, which is expected to increase the EMD accuracy. The aim is to convert the starting decomposition described by n IMFs, $c_i(t), c_2(t), \dots, c_n(t)$, into a more synthetic decomposition described by $K < n$ CMFs, $c_{s_1}(t), c_{s_2}(t), \dots, c_{s_K}(t)$. The transformation implies not only an information synthesis, but also the capability of separating IMFs characterized by different properties and grouping together IMFs that shares similar patterns, leading to a better interpretation of underlying phenomena by means of a clustering-like procedure.

Grasso *et al*. [13] showed that the CMF methodology may be exploited to design signal monitoring techniques and, at the same time, to achieve a better characterization of fault effects on different scales. Nevertheless, the selection of the number K of final CMFs, together with the determination of which IMFs should be summed up together in each CMF still represent two open issues. A criterion based on local frequency changes captured by the instantaneous frequencies of IMF was proposed by Gao *et al*. [6]. This criterion is suitable to identify adjacent IMFs that share similar instantaneous frequency pattern, but it is not suitable for automatic implementation, as it relies on visual inspection of the IMFs. The selection of subsets of IMFs has attracted the attention of different authors, either for signal de-noising, de-trending or band-pass filtering. The mainstream methods are based on the computation of synthetic indexes [4 – 5, 10 – 12], but they share different limitations: (i) the selection of the most suitable index is a problem-dependent issue, (ii) there is a lack of automated ways to define thresholds associated to those indexes, apart from few heuristic solutions, and (iii) most of those methods are typically inapplicable when multiple groups of IMFs are of practical interest. These limitations, coupled with the potential benefits of enhancing the multi-scale decomposition via a CMF-based procedure, represent the motivation for the present study.

4 The proposed methodology

Let $Y(t)$ be a signal acquired within a time window $t \in [T_{start}, T_{stop}]$ with same sampling frequency, F_s, and decomposed into n IMFs.

The proposed approach is aimed at automatically reducing the n IMFs into a number $K < n$ of CMFs that are expected to better represent the multi-scale content of the signal, via a parsimony-driven procedure. The proposed approach for CMF computation involves four consecutive steps: (i) preliminary computation of sequential CMFs, denoted by $c_{s_k}(t),\ k = 1, \dots, n$, (ii) computation of a dissimilarity index to determine a possible separation of IMFs into fewer CMFs, denoted by $c^*_{s_k}(t)$, $k = 1, \dots, K^*$, (iii) iterative decomposition into different numbers K^* of CMFs, and (iv) determination of the optimal number $K < n$ of final CMFs, $c^*_{s_k}(t)$.

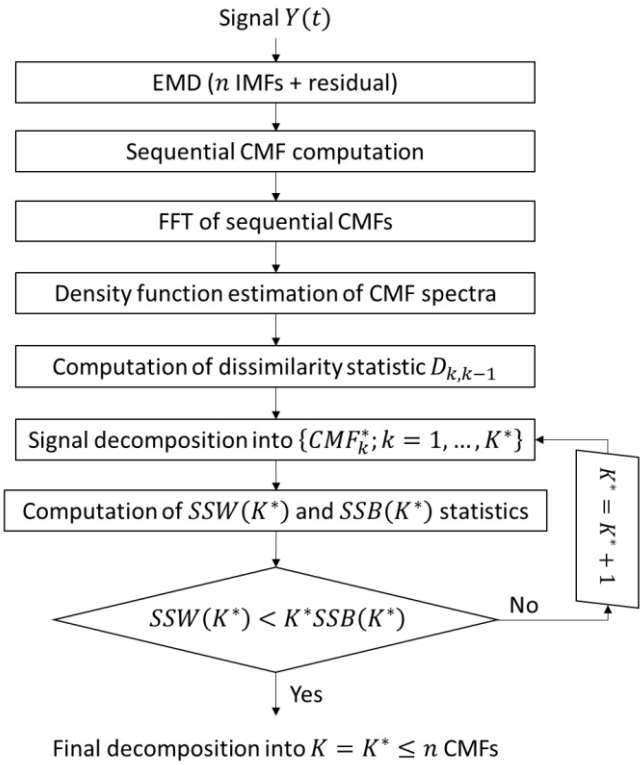

Fig. 8 – Conceptual scheme of the proposed approach

Notice that two CMF decompositions are successively generated. The first one, called "sequential CMF" decomposition and denoted by $\{c_{s_k}(t), k = 1, \dots, n\}$, implies no dimensionality reduction with respect to the EMD. The second one, called "final CMF" decomposition and denoted by $\{c^*_{s_k}(t), k = 1, \dots, K\}$, is the final result of an iterative procedure aimed at finding the best compromise between information synthesis and separation of relevant modes.

A conceptual scheme of the proposed approach is depicted in Fig. 8.

Once the EMD has been applied to the input signal $Y(t)$, the first step consists in an iterative computation of sequential CMFs $\{c_{s_k}(t) = \sum_{i=1}^{k} c_i(t); k = 1, \dots, n\}$, such that the k^{th} CMF is the sum of all the IMFs, c_i, from the first one to the k^{th} one, i.e., each CMF is simply generated by adding one IMF to all the previous ones. The resulting decomposition into n CMFs is then used to determine if the addition of a higher order IMF to the lower order ones yields a significant modification of the spectral properties. Assume that by adding the i^{th} IMF, $c_i(t)$, to a CMF that consists in the sum of all the previous IMFs, $c_{s_{i-1}}(t) = c_1(t) + \cdots + c_{i-1}(t)$, no significant change of the frequency spectrum is observed. This means that the i^{th} IMF does not provide any novel information with respect to $c_{s_{i-1}}(t)$, and hence it may be added to the previous IMFs. On the contrary, if the addition of the i^{th} IMF yields a significant change of the spectral properties, an embedded scale change is detected, and an enhanced signal decomposition may be achieved by separating $c_i(t)$ from $c_{s_{i-1}}(t)$.

It is clear that the CMF separation mechanism requires a criterion to decide when a scale change occurs, i.e., when the information captured by one IMF is significantly different from the one captured by the sum of previous IMFs. In our proposed approach, such a decision exploits a dissimilarity measure that is based on the probability density functions (PDFs) of CMF frequency spectra. The higher is the dissimilarity between the PDFs of two frequency spectra, the higher is the benefit of separating the corresponding modes to achieve a meaningful final decomposition. The dissimilarity statistic between $c_k(t)$ and $c_{s_{k-1}}(t)$ will be denoted by $D_{k,k-1}$. The rationale for the choice of a PDF-based criterion and the details of the proposed procedure are discussed in sub-section 4.1.

Every local peak in the dissimilarity function $D_{k,k-1}$ represents a potential scale change, and hence a potential separation into distinct CMFs $c^*_{s_k}(t)$. Generally speaking, the number of potential CMFs is upper-bounded by $M + 1$, where M is the number of local maxima of the dissimilarity function, $D_{i,i-1}$. Thus, multiple potential decompositions $\{CMF^*_k; k = 1, \dots, K^*\}$, where $K^* = 1, \dots, M + 1$ are possible. Because of this, the last step of the proposed method consists in an automated way to decide the optimal number, K, of final CMFs. The underlying idea consists in computing two sum-of-squares statistics that describes the variability within and between the CMFs (respectively called "sum-of-squares within CMFs", denoted by $SSW(K^*)$ and "sum-of-squares between CMFs", denoted by $SSB(K^*)$) and to select the minimum number K that corresponds to the best compromise between them. The methodology is explained in sub-section 4.3. The final result is a decomposition into $K \leq n$ CMFs, $c^*_{s_k}(t)$, which are expected to synthetically capture distinct embedded modes.

4.1. Dissimilarity computation between consecutive CMFs

The sequential CMF decomposition represents a suitable source of information to decide whether each IMF, $c_i(t)$, introduces a relevant scale change with respect to the sum of previous ones, $c_{s_{i-1}}(t)$, or it may be included into the former CMF without distorting its spectral properties.

The methodology is illustrated by means of an example. Let $Y(t)$ be a signal composed by a white noise terms and a superimposition of two frequency components at $f_1 = 15\ Hz$ and $f_2 = 75\ Hz$, respectively. The signal is defined as follows:

$$Y(t) = Y_n(t) + Y_{S1}(t) + Y_{S2}(t)$$
$$Y_n(t) \sim N(0, \sigma_n^2);\ Y_{S1}(t) = A_1 \sin(2\pi f_1 t);\ Y_{S2}(t) = A_2 \sin(2\pi f_2 t) \quad (3)$$

where $\sigma_n^2 = 11$, $A_1 = 100$ and $A_2 = 50$.

The signal is generated over a time window $t \in [0, T]$ of length $T = 1$ s and sampled at $F_s = 1\ kHz$. Fig. 9 shows a time plot of the signal, whereas its EMD is shown in Fig. 10.

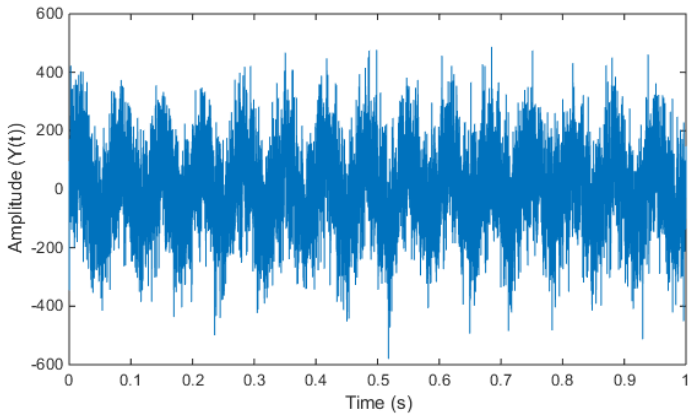

Fig. 9 – Synthetic signal consisting in a superimposition of two oscillation modes and a noise term

The frequency component at $f_2 = 75\ Hz$ is mainly captured by IMFs $c_6(t)$ and $c_7(t)$, whereas the IMFs $c_8(t)$ and $c_9(t)$ capture the component at $f_1 = 15\ Hz$. Notice that a mode mixing effect occurred, causing a partial splitting of the low frequency components into $c_8(t)$ and $c_9(t)$. This is an undesired effect, which has a detrimental impact on the interpretation of the IMF information content. The noise terms is split into the first five IMFs, $c_1(t), \ldots, c_5(t)$, whereas the two last IMFs in the low frequency region, $c_{10}(t)$ and $c_{11}(t)$, represent meaningless modes.

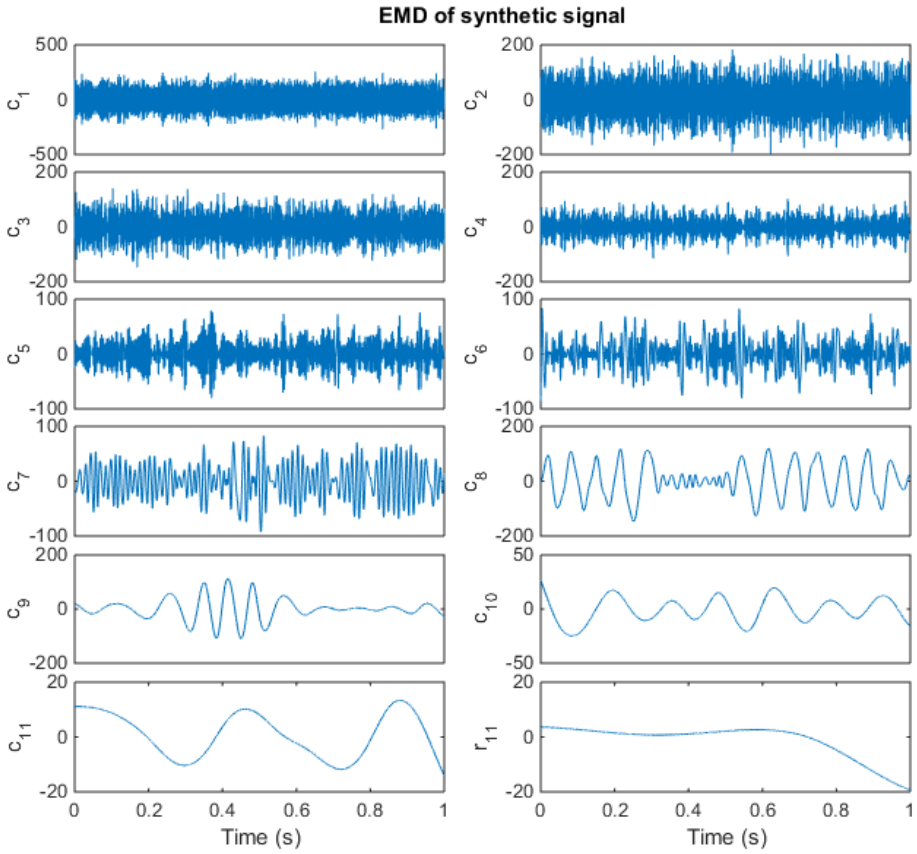

Fig. 10 – Empirical mode decomposition of the synthetic signal

Fig. 11 shows the frequency spectra via Fast Fourier Transform (FFT) of the sequential CMF decomposition, $\{c_{s_k}(t), k = 1, \dots, n = 11\}$, for the synthetic signal $Y(t)$. Fig. 11 shows that the sum of the i^{th} IMFs to the previous ones does not modify the spectrum of the CMFs unless $i > 5$. As a matter of fact, the first five CMFs, $c_1(t), \dots, c_5(t)$, result from a splitting of the noise term, and hence they can be summed up into a single CMF without losing relevant information about the signal pattern. When the IMF $c_6(t)$ is summed to the previous ones, instead, a peak at $f_2 = 75\ Hz$ is clearly observed. The further addition of the IMF $c_7(t)$ has the only consequence of inflating the energy associated to $f_2 = 75\ Hz$, without introducing additional information. When the IMF $c_8(t)$ is added to the previous ones, the contribution of the frequency component at $f_1 = 15\ Hz$ becomes visible, too. By summing up the following IMFs, no relevant modification of the spectral content is achieved, apart from a slight increase of the energy at the low frequency component compared with the high frequency one.

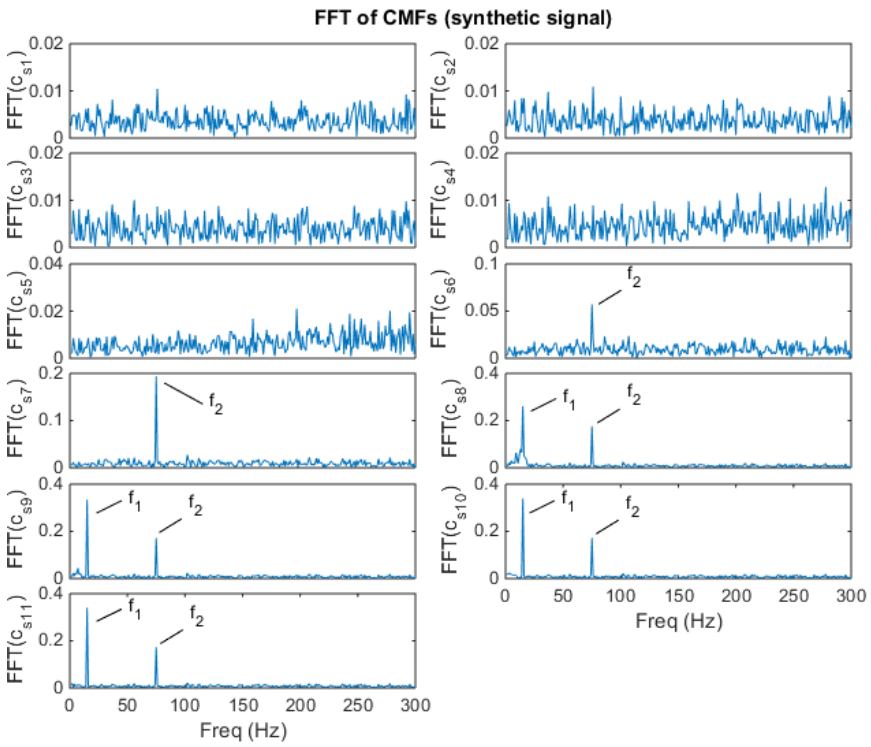

Fig. 11 – Frequency spectra of the sequential CMFs generated from the synthetic signal

Therefore, the spectral analysis of the sequential CMFs allows determining the relevant scale changes that occur in the original IMF decomposition. One critical issue consists in assessing the dissimilarity between the frequency spectra for two consecutive CMFs, $c_{s_k}(t)$ and $c_{s_{k-1}}(t)$. To this aim, we propose a procedure based on the probability density function (PDF) of the CMFs, which involves the following steps: (i) estimation of the PDF for each sequential CMF, $c_{s_k}(t)$, for $k = 1, \dots, n$; (ii) computation of a dissimilarity measure between each pair of PDFs as a function of the index k, and (iii) identification of local maxima of the dissimilarity function.

The PDF estimation allows transforming the spectrum, denoted by $x_k(\omega),\ k = 1, \dots, n$, where ω represents the frequency location, into a function, $f_k(x)$, that is believed to improve the dissimilarity computation. However, in order to properly recognize differences between spectra that involve mainly local changes (like the ones shown in Fig. 11), we advocate the use of a weighted PDF estimate. The weight function $w_k(\omega)$ associates a weight to each frequency location, ω. By defining $w_k(\omega) = x_k^2(\omega)$, local peaks with high amplitude will have a larger weight than other frequency bands characterized by a low amplitude, and hence a more effective detection of scale changes will be achieved.

Since the form of the distribution of sequential CMFs is unknown, a nonparametric density estimation is required [21], which exploits the kernel fitting technique. Two relevant issues consist in the choice

of the kernel function, $Ker(x)$, and the selection of an optimal kernel bandwidth, h. The latter issue is the most critical one, and several methods have been proposed thus far. One simple approach is to use rule-of-thumb estimates, which are known to approximate the optimal choice in the presence of normal data [22]. However, when strong departures from normality are observed, other methods should be preferred, which are aimed at estimating the bandwidth, h, in a data-driven way. The most effective methods can be classified into two major categories, i.e., plug-in methods [23] and cross-validation methods [24]. Among these, the method based on unbiased cross-validation (UCV) is probably the most popular and studied one [25 – 26]. Although different studies compared the above mentioned algorithms [27 – 29], there is no general rule to prefer one approach over the other, with the only exception of rule of thumb methods that should be used only if the expected shape of the function is close to a Gaussian. The UCV-based approach is used as a baseline in this study as it was demonstrated to yield good results in a wide range of applications [24]. Future studies may be aimed at investigating an implementation based on different KDE techniques. Appendix B briefly revises the UCV-based method. The kernel function choice has a reduced impact on KDE performances. Because of this, the most common choice, i.e., the Gaussian function, is used in this study.

The symbol $\hat{f}(x)$ is used to denote the weighted kernel estimator of the (unknown) density function. By way of example, the weighted PDF estimates, $\hat{f}_k(x)$, for $k = 1, \dots, n = 11$, of the frequency spectra of sequential CMFs shown in Fig. 11 are depicted in Fig. 12 (left panel).

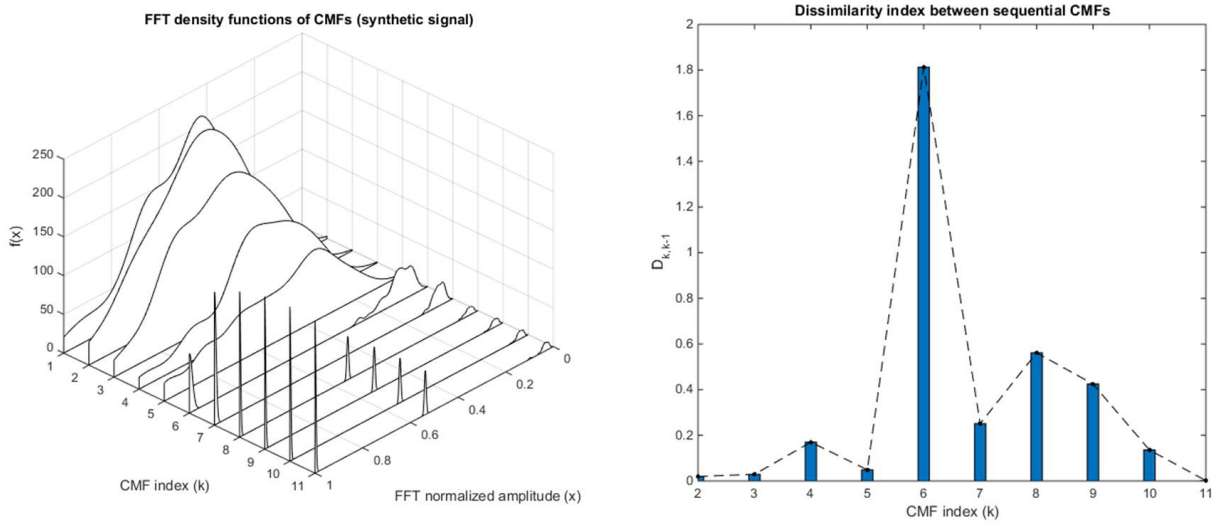

Fig. 12 – Probability density functions of CMF frequency spectra (left panel) and corresponding dissimilarity index function (right panel) for the synthetic signal

For sake of clarity, Fig. 12 (left panel) shows the amplitudes of the CMF frequency spectra after normalization, i.e., $x_k(\omega) \in [0,1]$ for every sequential CMF. The shape modification of the estimated

PDFs occurred at $k = 6$ is clearly visible in Fig. 12 (left panel), which corresponds to the high frequency oscillation at $f_2 = 75\ Hz$. Another shape modification occurs at $k = 8$, when the CMF includes both the low-frequency and high-frequency modes. The next step consists in quantifying the dissimilarity between these PDFs.

Let $\hat{f}_k(x)$ and $\hat{f}_{k-1}(x)$ be the estimated density functions of the frequency spectra of sequential CMFs $c_{s_k}(t)$ and $c_{s_{k-1}}(t)$ generated from the signal $Y(t)$. Then, their dissimilarity can be expressed in terms of their cross-correlation coefficient as follows:

$$D_{k,k-1} = 1 - \rho\left(\hat{f}_k(x), \hat{f}_{k-1}(x)\right), \qquad k = 1, \dots, n \tag{4}$$

where $\rho(\cdot,\cdot)$ is the sample cross-correlation coefficients, such that $\rho(\cdot,\cdot) \in [-1,1]$. In this study, the Spearman's correlation coefficient is used, because, differently from the Pearson's coefficient, it is able to capture cross-correlations that are not limited to simple linear functions. Such a property is expected to enhance the capability of quantifying not only global shape dissimilarities but also local ones. The dissimilarity function for the synthetic signal example is shown in Fig. 12 (right panel). It exhibits two major peaks at $k = 6$ and $k = 8$, which correspond to the actual scale changes of interest. The following steps of the proposed approach are described in the next two sub-sections, respectively devoted to the iterative decomposition into few potentially relevant CMFs and to the final selection of the optimal number of final CMFs.

4.2. Decomposition into a reduced number of CMFs

A local maximum in the dissimilarity function, $D_{k,k-1}$, represents a potential scale change in the sequential CMF decomposition, such that the inclusion of the k^{th} IMF, $c_{i=k}(t)$, into $c_{s_{k-1}}(t)$ yields a large dissimilarity between $c_{s_{k-1}}(t)$ and $c_{s_k}(t)$, but the further inclusion of the $(k+1)^{th}$ IMF, $c_{i=k+1}(t)$, into $c_{s_k}(t)$ yields a smaller dissimilarity between $c_{s_k}(t)$ and $c_{s_{k+1}}(t)$. This means that the k^{th} IMF, $c_{i=k}(t)$, captures some dissimilar information with respect to the sum of previous IMFs, but similar information with respect to the next IMF. Thus, one possible criterion to decide whether each IMF should be added to the lower order ones or separated from them consists in dividing the final CMFs at indexes k corresponding to local maxima in the $D_{k,k-1}$ function. Local peak detection can be easily performed by finding sign changes in the successive differences of the $D_{k,k-1}$ function. Nevertheless, only the largest peaks are expected to bring some relevant information about actual scale changes in embedded modes. Two operations are required: first, an iterative approach to generate potential CMF decompositions, driven by the analysis of the dissimilarity function, $D_{k,k-1}$, must be applied (briefly explained in this sub-section); second, an optimality criterion should be

applied to select the final CMF decomposition that better represents the multi-scale content of the original signal (explained in sub-section 4.3).

Let $\boldsymbol{p} = [p_1, \dots, p_M]$ be the vector of "locations" k corresponding to these local peaks, where M is the number of peaks and let $\tilde{\boldsymbol{p}}$ be the vector with elements sorted in descending peak amplitude order. The goal is to pass from the n sequential CMFs, $c_{s_k}(t)$, to a smaller number of final CMFs, $c^*_{s_k}(t)$, by iteratively imposing a further separation between CMFs corresponding to each element of the sorted location vector $\tilde{\boldsymbol{p}}$. The algorithm consists in the following steps:

1 Initialize the counter $c = 1$;

2 Generate a new CMF decomposition by dividing the IMFs into $K^* = c + 1$ CMFs, $c^*_{s_k}(t)$, such that the c^{th} element of the sorted location vector $\tilde{\boldsymbol{p}}$, i.e. $\tilde{p}_c$, represents the argument of the first IMF of one of K^* CMFs;

3 Set $c = c + 1$: if $c > M$ the procedure is over, otherwise go to step 2 and generate the c^{th} CMF decomposition.

The result consists in M possible CMF decompositions $\{c^*_{s_k}(t); k = 1, \dots, K^*\}$, each comprised of $K^* \in [2, M + 1]$ CMFs, $c^*_{s_k}(t)$.

4.3. Selection of the final number of CMFs

The problem of selecting the optimum number of CMFs is analogous to the problem of finding the best number of clusters in an unsupervised classification application, which is also known as "cluster validity" in the statistical learning literature [30 – 31]. Analogously to clustering problems, our aim consists in determining a CMF decomposition such that the inclusion of dissimilar modes into the same CMF is possibly avoided by contemporary keeping the number of final CMFs as small as possible. A category of cluster validity criteria relies on the measure of the within and the between group variance components: they include the *Ball & Hall* index [31], the *Calinski & Harabasz* index [32], the *Hartigan* index [33], and many others [34 – 35]. A simple and effective method can be designed by considering the variability within and between the CMFs as a measure of the necessity of applying a further decomposition. On the one hand, if the variability within the CMFs decreases by passing from K^* to $K^* + 1$ final CMFs, then at least two dissimilar modes have been separated into distinct CMFs, and hence the decomposition that comprises $K^* + 1$ CMFs should be preferred. However, if the variability within the CMFs is still larger than the variability between the CMFs, a further decomposition is required. In analogy with the cluster validity indexes proposed by other authors [31 – 35], these two variabilities can be estimated via the sum-of-squares within and between

the K^* CMFs, respectively. These indexes are denoted by $SSW(K^*)$ and $SSB(K^*)$ and computed as follows:

$$SSW(K^*) = \sum_{k=1}^{K^*} \sum_{i \in c^*_{s_k}(t)} 1 - \rho\left(\hat{f}_{k,i}(x) - \bar{\hat{f}}_k(x)\right)$$
$$SSB(K^*) = \sum_{k=1}^{K^*} n(k)\left[1 - \rho\left(\bar{\hat{f}}_k(x) - \bar{\hat{f}}(x)\right)\right] \quad (5)$$

where $\bar{\hat{f}}_k(x)$ is the average density function of frequency spectra within the kth CMF, $c^*_{s_k}(t)$, and $\bar{\hat{f}}(x)$ is the average density function of the original decomposition. The proposed criterion to determine the best number of CMFs is based on the following inequality:

$$SSW(K^*) < K^* SSB(K^*) \quad (6)$$

When $SSW(K^*) > K^* SSB(K^*)$, the variability within the CMFs (in terms of distances between density functions) is larger than the variability between the CMFs: this means that a further decomposition into $K + 1$ CMFs allows separating distinct modes that are currently included into the same CMF. When $SSW(K^*) < K^* SSB(K^*)$ the desired condition is achieved, i.e., the IMFs in each CMF are close to each other, and different intrinsic modes are separated into distinct CMFs. From a parsimony viewpoint, the multiplication by K^* is used to penalize the selection of larger number of CMFs. Grasso *et al.* [36] showed that this criterion is effective in practice and it may outperform other benchmark validity criteria when coupled with PDF-based dissimilarity functions. Nevertheless, future studies may investigate possible ways to tune the selection criterion for general-purpose use.

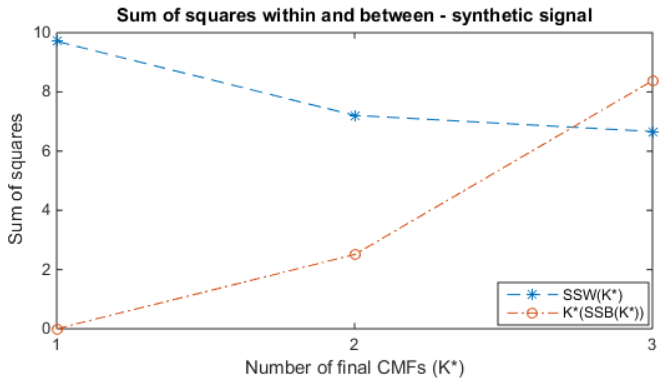

Fig. 13 – $SSW(K^*)$ and $K^* SSB(K^*)$ values for different numbers K^* of CMFs (synthetic signal)

Fig. 13 shows the values of $SSW(K^*)$ and $K^* SSB(K^*)$ for $K^* = 1,2,3$ in the synthetic signal example. In this case, the proposed criterion leads to the choice of $K = 3$ final CMFs, which are depicted in Fig. 14. Fig. 14 shows that the signal decomposition has been synthetized from $n = 11$ starting IMFs

to only three final CMFs, which actually capture the three embedded components, i.e., the noise term in $c^*_{s_1}(t)$, the high frequency oscillation in $c^*_{s_2}(t)$ and the low frequency oscillation in $c^*_{s_3}(t)$. The result is a more synthetic and easier to interpret representation of the multi-scale content of the signal than the one provided by the basic EMD methodology.

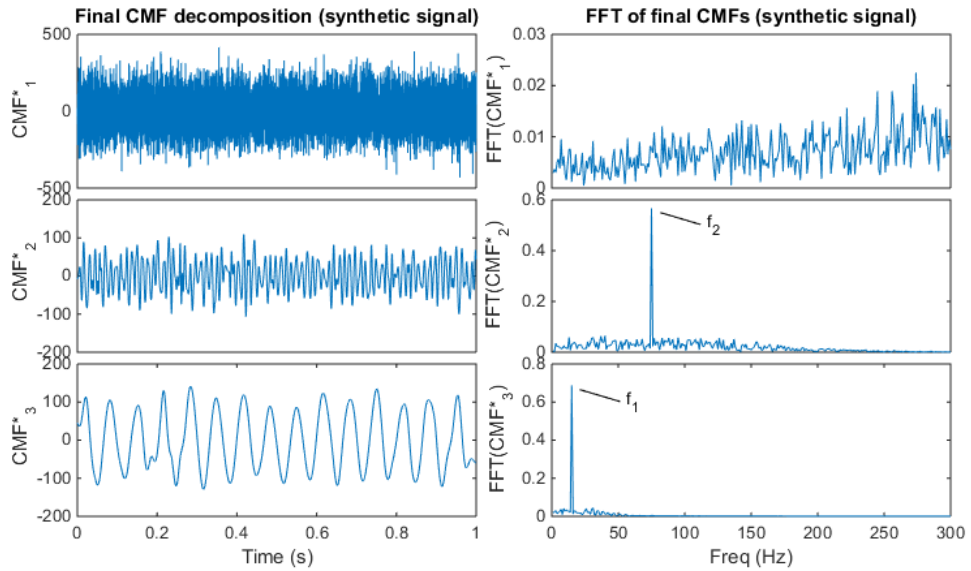

Fig. 14 – Final CMF decomposition of the synthetic signal (left panels) and corresponding frequency spectra for different CMFs (right panels)

5 Analysis of rolling element bearing conditions

5.1. Performance analysis

The effectiveness of the proposed approach is proven by means of the application to the vibration signals acquired in the case study introduced in Section 1, for rolling element bearing fault analysis. In the following, subscripts $1, \dots, 6$ are included into the notation to identify distinct signals corresponding to different epochs. Fig. 15 and Fig. 16 show, respectively, the frequency spectra of the sequential CMFs generated from the health bearing signal $Y_1(t)$ and the faulty bearing signal $Y_4(t)$. Relevant frequency components were labeled as multiple of defect-related and known frequencies, whereas sidebands and other components are indicated in Hz. Fig. 15 shows that a change in the spectral pattern occurs at CMF $c_{S_{7,1}}(t)$, where the peak at $f = 193.8\,Hz$ (i.e., the $12x$ component) starts predominating the frequency spectrum. The addition of following IMFs in sequential CMFs does not significantly change the spectra, apart from making more evident the

contribution of multiples of the fundamental rotating frequency and other components that are not related with known defects.

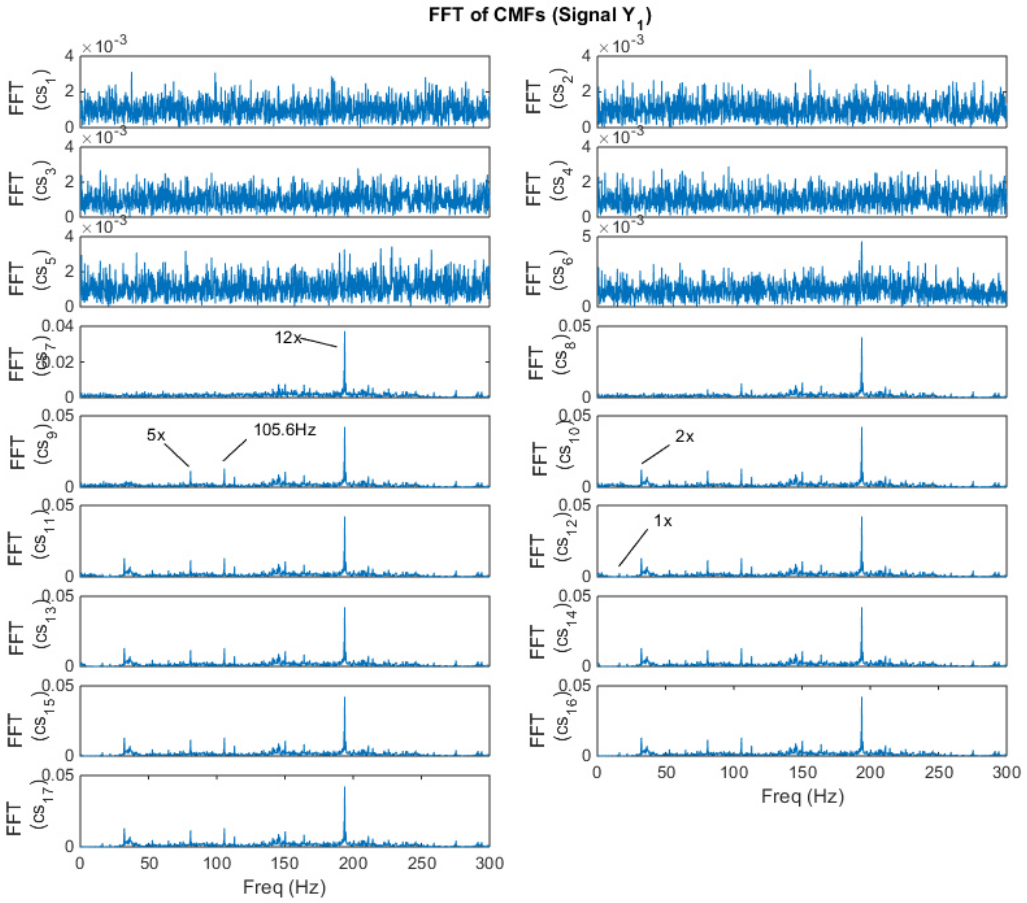

Fig. 15 – Frequency spectra of the sequential CMFs generated from the bearing vibration signal $Y_1(t)$

Fig. 16 shows that, for signal $Y_4(t)$, the noise frequency spectrum characterizes the first 9 sequential CMFs, $c_{s_{1,4}}(t), \ldots, c_{s_{9,4}}(t)$. A frequency spectrum change occurs at CMF $c_{s_{10,4}}(t)$, where peaks in the high frequency range become evident together with sidebands caused by a modulation effect imposed by bearing defects. In correspondence of CMF $c_{s_{15,4}}(t)$ a further frequency spectrum change seems to occur, which shifts the signal energy towards the very low frequency region, with predominant peaks at the FTF frequency and its multiples. Although the frequency spectrum is known to be not sufficient to properly characterize the bearing defects, it provides a suitable source of information to detect possible scale changes between sequential CMFs. As a matter of fact, the frequency spectrum is here used not for direct diagnostic purposes, but as an intermediate step to enhance the signal decomposition into its embedded modes. The analysis of envelope spectra of final CMFs for diagnosis purposes is recommended.

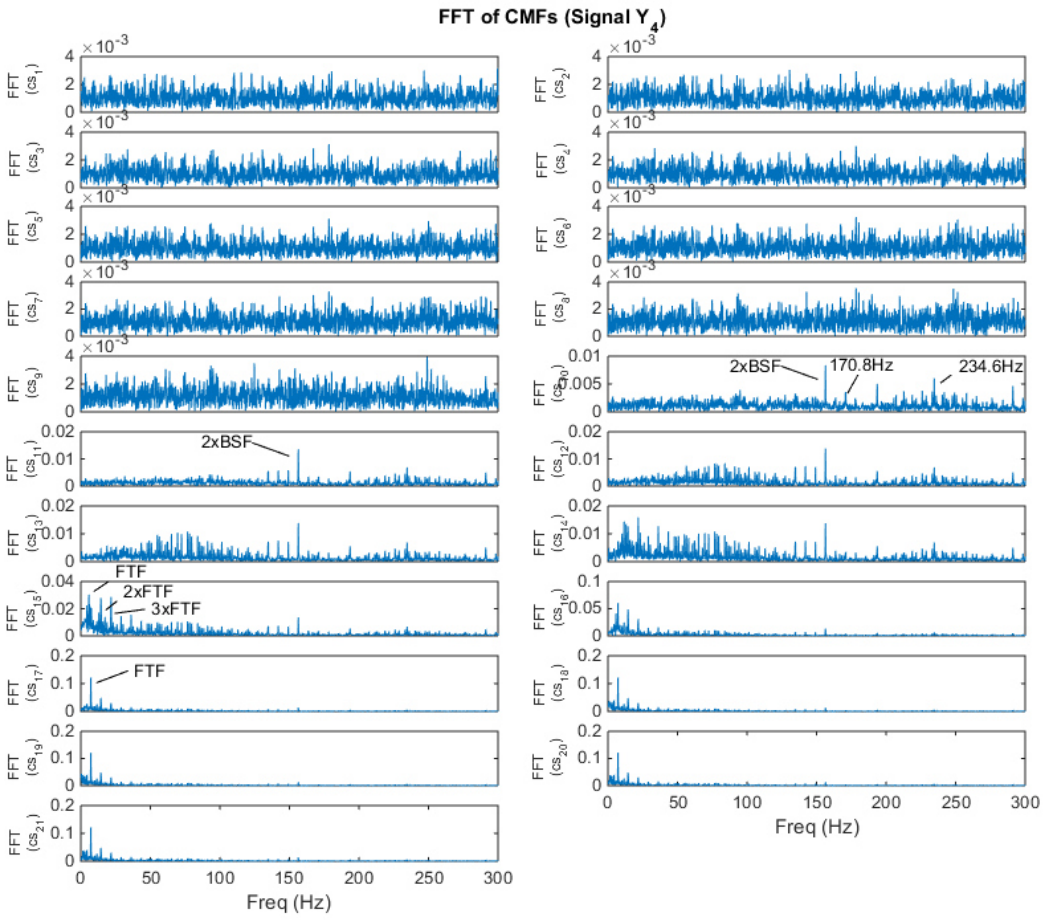

Fig. 16 – Frequency spectra of the sequential CMFs generated from the bearing vibration signal $Y_4(t)$

With regard to signal $Y_1(t)$, the probability density functions of the sequential CMF frequency spectra are shown in Fig. 17 (left panel), whereas Fig. 17 (right panel) shows the corresponding dissimilarity function. The shape of the density functions emphasizes the change of the vibration spectrum occurred at CMF $c_{s_{7,1}}(t)$ and it clearly shows that for $c_{s_{1,1}}(t), \dots, c_{s_{6,1}}(t)$ and for $c_{s_{7,1}}(t), \dots, c_{s_{17,1}}(t)$ the density shape remains about constant. The dissimilarity index, $D_{k,k-1,1}$, clearly exhibits a major peak at $k = 7$, a minor peak at $k = 3$ and two additional very low amplitude peaks at $k > 10$. In this case, the criterion based on the $SSW_1(K^*)$ and $SSB_1(K^*)$ statistics yields a final decomposition consisting in $K_1 = 2$ CMFs, such that $c^*_{s_{1,1}}(t) = \sum_{i=1}^{6} c_{s_{i,1}}(t)$ and $c^*_{s_{2,1}}(t) = \sum_{i=7}^{17} c_{s_{i,1}}(t)$. The final CMF decomposition for the healthy bearing signal, $Y_1(t)$, is shown in Fig. 18 (left panels), where both the CMF frequency spectra (middle panels) and envelope spectra (right panels) are depicted.

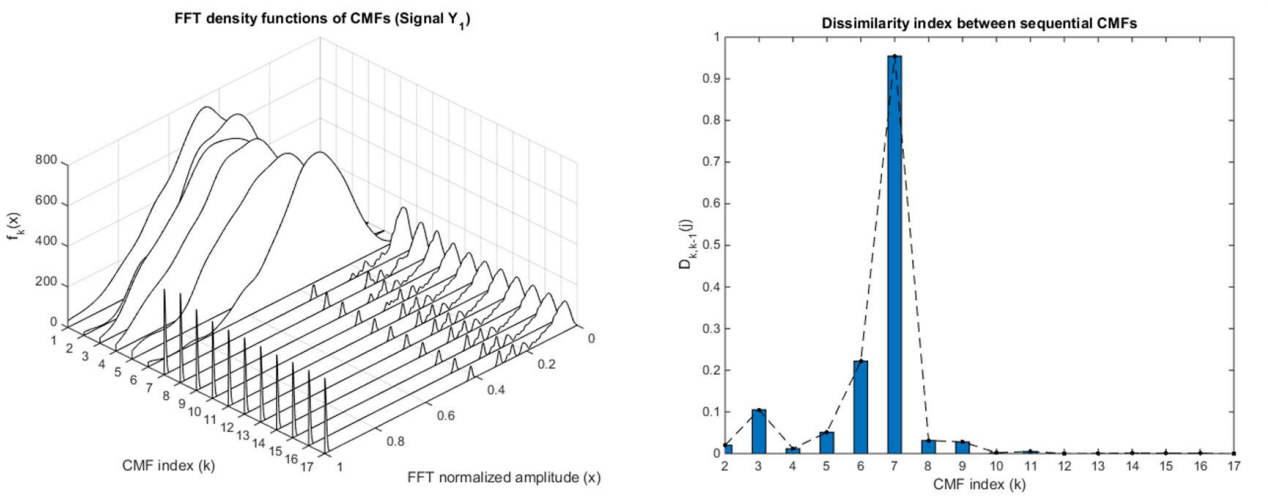

Fig. 17 – Probability density functions of CMF frequency spectra (left panel) and corresponding dissimilarity index function (right panel) for the bearing vibration signal $Y_1(t)$

Fig. 18 shows that the entire information content of the signal has been synthetized from the starting 17 IMFs to only two CMFs, $c^*_{s_{1,1}}(t)$ and $c^*_{s_{2,1}}(t)$. The first CMF, $c^*_{s_{1,1}}(t)$, seems to mainly capture the noise term, whereas the second CMF, $c^*_{s_{2,1}}(t)$, captures frequency components associated to the healthy state of the bearing.

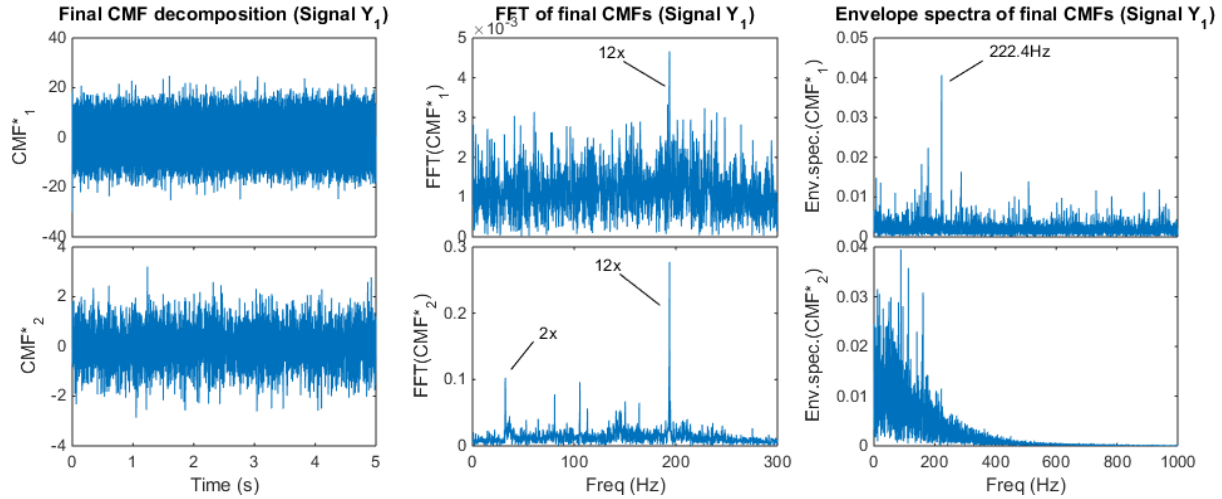

Fig. 18 – Final CMF decomposition for the bearing vibration signal $Y_1(t)$ (left panels) and corresponding frequency spectra (middle panels) and envelope spectra (right panels)

With regard to signal $Y_4(t)$, the probability density functions of the sequential CMF frequency spectra are shown in Fig. 19 (left panel), whereas Fig. 19 (right panel) shows the corresponding dissimilarity function. Also in this case, the shape of the density functions emphasizes the two changes of the vibration spectrum occurred at CMF $c_{s_{10,4}}(t)$ and at CMF $c_{s_{15,4}}(t)$, respectively. The variability in the range $c_{s_{10,4}}(t), \ldots, c_{s_{14,4}}(t)$ seems to be larger than in the upper and lower order ranges. This may

suggest that the IMFs belonging to that range are capturing some intermediate modes where the splitting of embedded oscillation scales is larger than in the lower and higher frequency regions. The dissimilarity index, $D_{k,k-1,1}$, clearly exhibits two major peaks at $k = 15$ and $k = 10$, together with some very low amplitude peaks at few other values of k. In this case, the criterion based on the $SSW_4(K^*)$ and $SSB_4(K^*)$ statistics leads to a final decomposition consisting in $K_4 = 3$ CMFs, such that $c^*_{s_{1,1}}(t) = \sum_{i=1}^{9} c_{s_{i,1}}(t)$, $c^*_{s_{1,1}}(t) = \sum_{i=10}^{14} c_{s_{i,1}}(t)$ and $c^*_{s_{2,1}}(t) = \sum_{i=15}^{21} c_{s_{i,1}}(t)$. The final CMF decomposition for the faulty bearing signal, $Y_4(t)$, is shown in Fig. 20 (left panels), where both the CMF frequency spectra (middle panels) and envelope spectra (right panels) are depicted.

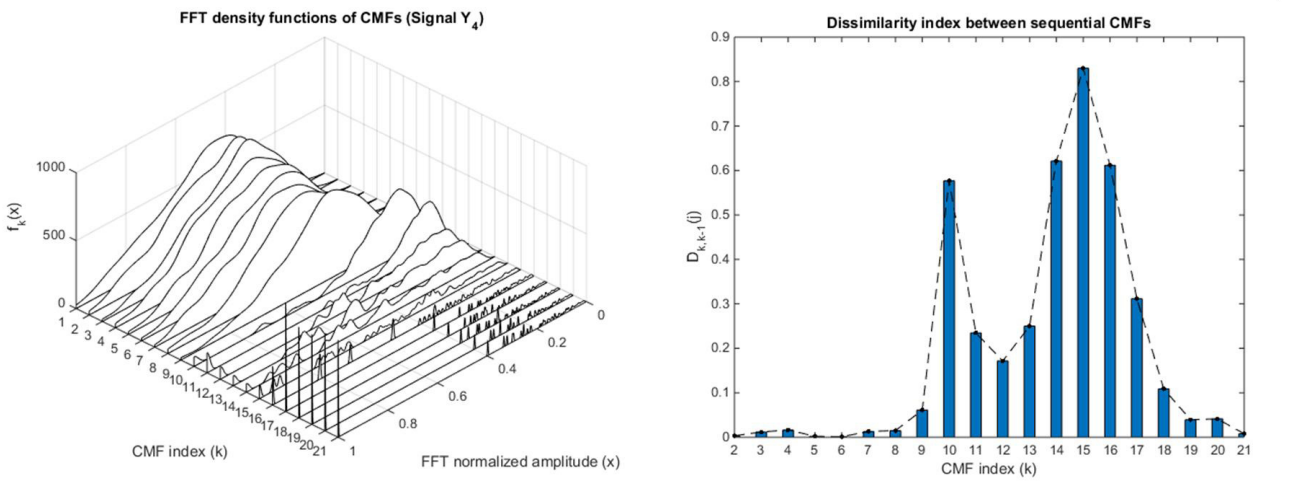

Fig. 19 – Probability density functions of CMF frequency spectra (left panel) and corresponding dissimilarity index function (right panel) for the bearing vibration signal $Y_4(t)$

Fig. 20 shows that the entire information content of the signal has been synthetized from the starting 21 IMFs to only three CMFs, $c^*_{s_{1,4}}(t)$, $c^*_{s_{1,4}}(t)$ and $c^*_{s_{3,4}}(t)$. Such a transformation has preserved the original information content, as neither filtering nor IMF selection were applied, but it seems to provide a clearer representation of the multi-scale pattern related with the presence of bearing defects. The first CMF, $c^*_{s_{1,4}}(t)$, captures most of the signal noise modulated by the BSF as shown by the envelope spectrum. The second CMF, $c^*_{s_{2,4}}(t)$, captures an intermediate spectral range, where the modulation at the FTF predominates the envelope spectrum, but the BSF-related component is still present. The third CMF, $c^*_{s_{3,4}}(t)$, captures the oscillation with frequency equal to the FTF.

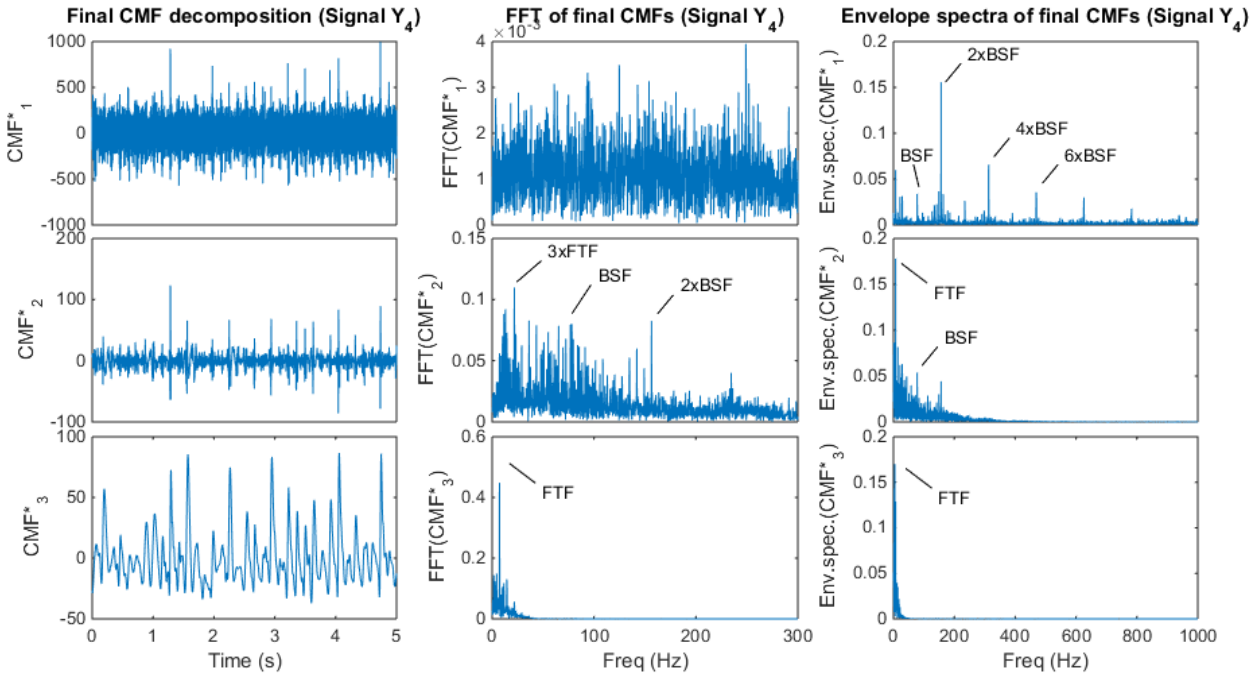

Fig. 20 – Final CMF decomposition for the bearing vibration signal $Y_4(t)$ (left panels) and corresponding frequency spectra (middle panels) and envelope spectra (right panels)

Fig. 21 shows the pattern of dissimilarity indexes $D_{k,k-1}$ computed for the five signals acquired from an healthy bearing state to a severely faulty state, i.e., $Y_1(t), \ldots, Y_6(t)$, together with the sixth signal, acquired after bearing substitution. It shows that the dissimilarity between sequential CMFs remains quite stable for signals $Y_1(t)$, $Y_2(t)$ and $Y_3(t)$, with a peak corresponding to IMF $c_7(t)$. The proposed approach yields two CMFs for all these signals, but the analysis of the CMF decomposition of signal $Y_3(t)$ highlights that the defect is already present, being evident the contribution of both BSF and FTF in the envelope spectrum of the first CMF, $c^*_{s_{1,3}}(t)$ (see Fig. 22).

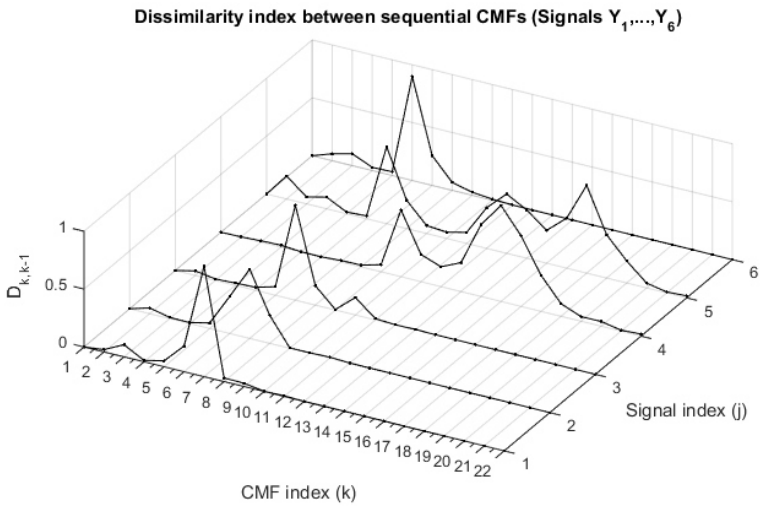

Fig. 21 – Dissimilarity index functions for bearing vibration signals $Y_1(t), \ldots, Y_6(t)$

Starting from the signal $Y_4(t)$ the number of extracted IMFs grows and the pattern of the dissimilarity index $D_{k,k-1}$ changes, including two major peaks at $Y_4(t)$ and three major peaks at $Y_5(t)$. Thus, the evolution over time of the CMF decomposition follows the evolution of the defect severity, and hence it is expected to provide a useful diagnostic tool. The three peaks at $Y_5(t)$ yield a decomposition into four CMFs, where the noise is separated from defect-related components, leading to a further highlight of the faulty state. After replacing the faulty bearing with a healthy one (signal $Y_6(t)$), the initial $D_{k,k-1}$ pattern characterized by a single peak is restored.

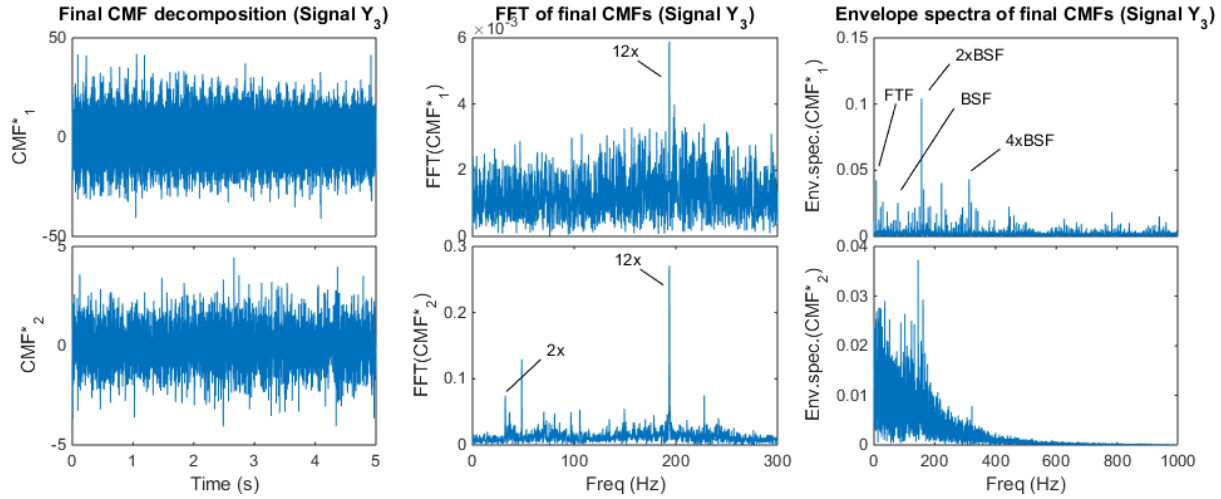

Fig. 22 – Final CMF decomposition for the bearing vibration signal $Y_3(t)$ (left panels) and corresponding frequency spectra (middle panels) and envelope spectra (right panels)

5.2. Comparison against index-based IMF selection

As stated in Section 2, different synthetic indexes were proposed in the literature devoted to the EMD methodology for the selection of specific IMFs. Apart from denoising or detrending applications, index-based selection of IMFs is often used to enhance the characterization of the signal and to increase the accuracy of the Hilbert-Huang transform by retaining only relevant modes. Because of this, a natural competitor for our proposed approach consists in selecting specific IMFs via index-based criteria instead of computing a more synthetic decomposition in terms of CMFs. Many synthetic indexes have been proposed thus far [4 – 5, 11 – 12] including the energy, the correlation between the IMF and the original signal, the peak frequency, etc. Among them, the cross-correlation coefficient between each IMF and the original signal has been used by different authors [4, 15] and it can be considered a benchmark in this frame. Moreover, Peng *et al.* [15] proposed a method to automatically select the relevant IMFs that works as follows: let r_i be the normalized sample

correlation coefficient between the signal $Y(t)$ and the i^{th} IMF, $c_i(t)$; then, if $r_i \geq \lambda$ the i^{th} IMF is retained, otherwise it is added to the residue, where $\lambda = \left(\max_i r_i\right)/10$.

By applying this index-based selection approach to the IMFs generated from signals $Y_1(t)$ and $Y_4(t)$, respectively, the following subsets of relevant IMFs are obtained: $\{c_{1,1}(t), \dots, c_{5,1}(t); c_{7,1}(t)\}$ and $\{c_{1,4}(t), \dots, c_{6,4}(t); c_{15,4}(t), \dots, c_{18,4}(t)\}$. This means, that the information content of the signals is described in terms of 6 and 10 modes, whereas in terms of CMFs only two and three modes were extracted.

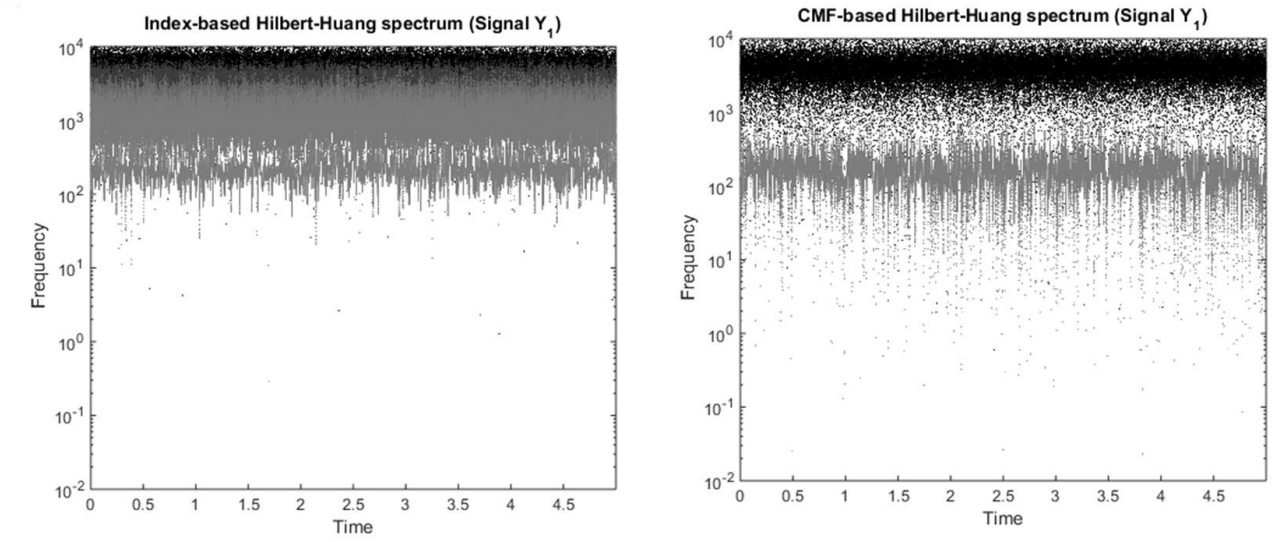

Fig. 23 – Hilbert-Huang spectrum based on IMFs selected via cross-correlation-based criterion (left panel) and Hilbert-Huang spectrum based on our proposed approach (right panel), for bearing vibration signal $Y_1(t)$

It is possible to compute the Hilbert-Huang spectrum on either the IMFs selected by using the correlation-based criterion or the CMFs generated via our proposed approach. The two resulting spectra will be denoted hereafter as "index-based" spectrum and "CMF-based" spectrum for sake of clarity. Fig. 23 shows the index-based (left panel) and CMF-based (right panel) Hilbert-Huang spectrum for the healthy bearing signal, $Y_1(t)$. In both cases, the pattern shows the lack of suspect defects, and the high frequency range dominates the time-frequency spectrum. Fig. 24 shows the index-based (left panel) and CMF-based (right panel) Hilbert-Huang spectrum for the bearing signal in the presence of a faulty state, $Y_4(t)$. In this case, the CMF-based spectrum seems to better depict the actual condition of the rolling element bearing. The second CMF, $c^*_{s_{2,4}}(t)$ produces a fluctuating pattern that oscillates about the BSF level, whereas the third CMF, $c^*_{s_{3,4}}(t)$, produces an oscillating pattern about the FTF level. The first CMF, $c^*_{s_{1,4}}(t)$, captures the high frequency content of the signal, the one with the highest energy, and this is visible also in the time-frequency spectrum. Analogously

to the CMF-based one, the index-based Hilbert-Huang spectrum allows to detect a changed condition of the bearing health state by comparing the spectra in Fig. 24 and those in Fig. 23. Nevertheless, IMF selection seems to reduce the capability of detecting some embedded phenomena. In particular, the presence of a modulating effect at both the FTF and the BSF is more difficult to appreciate, and the presence of different IMFs in the low frequency range has a detrimental effect on the fault analysis and the determination of its cause. Because of this, the enhanced computation of CMFs is expected to reduce the dimensionality of the problem and improve the interpretation of the system health conditions with respect to other IMF selection methods commonly used in practice.

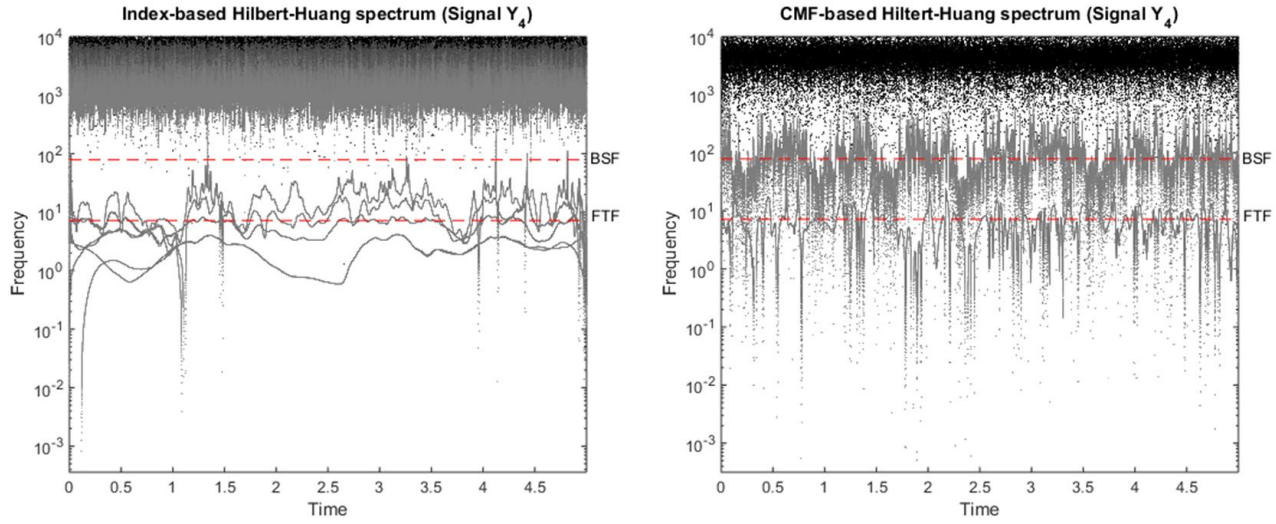

Fig. 24 – Hilbert-Huang spectrum based on IMFs selected via cross-correlation-based criterion (left panel) and Hilbert-Huang spectrum based on our proposed approach (right panel), for bearing vibration signal $Y_4(t)$

6 Conclusions

In rotating machinery fault detection and diagnosis applications, the capability of isolating fault features and determining the health condition of the system often requires a clear decomposition of sensor signals in their embedded modes. The data-driven and adaptive properties of the EMD methodologies attracted the attention of many researchers in this field, but the current literature lacks automated methods for the extraction of relevant modes. As a matter of fact, the EMD yields an-over decomposition of the signal, with spurious IMFs, especially in the low frequency range, and the possible occurrence of the mode mixing effect. The selection of specific IMFs of interest is often a challenging task, whose result is influenced by problem-dependent criteria.

This study presented an automated approach for the enhancement of vibration signal decomposition via EMD. The goal is to achieve a synthetic characterization of relevant modes by defining the optimal number of CMFs, where the term "optimal" refers to the best compromise between the number of final CMFs and the capability of capturing salient features on distinct scales.

The proposed approach works in a fully data-driven way by evaluating the role played by each IMF in determining the spectral property of the signal. The main idea of the approach is to compute the empirical probability density function of the CMFs frequency spectra and compute a dissimilarity index between density functions of adjacent IMFs to cluster them. In particular, the minimal number of final CMFs is eventually determined by applying a criterion that inherits the cluster validity principle used in unsupervised classification.

The proposed approach was illustrated by means of a simulation example consisting in a synthetic signal. The application of the method to a real case study involving rolling element bearing fault analysis showed that the method is suitable to reduce the number of relevant modes from many IMFs to few CMFs and, simultaneously, to enhance the interpretation and characterization of multi-scale phenomena of interest. The comparison with an index-based approach for IMF selection, with is believed to be representative of the common practice in EMD-based diagnostic problems, showed that the investigation of the nature and possible causes of bearing defects can be improved by using a CMF-based approach.

Future research developments will be aimed at evaluating the possible extension and generalization of the proposed approach to other application fields and different kinds of sensor signals. A direction of future research will be also devoted to couple the proposed procedure with alarm criterion usually applied in statistical process monitoring literature. In addition, the case study here discussed was characterized by a constant rotation speed. Future studies may be aimed at testing the robustness of the proposed approach in the presence of rotation speed fluctuations.

References

[1] R. B. Randall, J. Antoni, Rolling element bearing diagnostics—a tutorial, Mechanical Systems and Signal Processing, 25(2) (2011) 485-520

[2] Z. Feng, M. Liang, F. Chu, Recent Advances in Time-Frequency Analysis Methods for Machinery Fault Diagnosis: A Review with Application Examples, Mechanical Systems and Signal Processing, 38 (2013) 165-205

[3] F. Al-Badour, M. Sunar, L. Cheded, Vibration analysis of rotating machinery using time–frequency analysis and wavelet techniques, Mechanical Systems and Signal Processing, 25(6) (2011) 2083-2101

[4] A. Ayenu-Prah, N. Attoh-Okine, A Criterion for Selecting Relevant Intrinsic Mode Functions in Empirical Mode Decomposition, Advances in Adaptive Data Analysis, 2:1 (2010) 1-24

[5] P. Flandrin, P. Goncalves, G. Rilling, Detrending and Denoising with Empirical Mode Decomposition, XII European Signal Processing Conference (EUSIPCO), Vienna, Austria, September 6-10 (2004)

[6] Q. Gao, C. Duan, H. Fan, Q. Meng, Rotating Machine Fault Diagnosis using Empirical Mode Decomposition, Mechanical Systems and Signal Processing, 22(5) (2008) 1072-1081

[7] N.E. Huang, Z. Shen, S.R. Long, M.L. Wu, H.H. Shih, et al., The Empirical Mode Decomposition and Hilbert Spectrum for Nonlinear And Nonstationary Time Series Analysis, Proc. R. Soc. London Ser. A, 454 (1998) 903–95

[8] V. K. Rai, A. R. Mohanty, Bearing fault diagnosis using FFT of intrinsic mode functions in Hilbert–Huang transform, Mechanical Systems and Signal Processing, 21(6) (2007) 2607-2615

[9] Z. Wu, N.E. Huang, Ensemble Empirical Mode Decomposition: a Noise Assisted Data Analysis Method, Advances in Adaptive Data Analysis, 1:1 (2009) 1-41

[10] H. Liang, Q-H. Lin, J.D.Z. Chen, Application of the Empirical Mode Decomposition to the Analysis of the Esophageal Manometric Data in Gastroesophageal Reflux Disease, IEEE Transactions on Biomedical Engineering, 52:10 (2005) 1692-1701

[11] N. Bu, N. Ueno, O. Fukuda, Monitoring of Respiration and Heartbeat during Sleep using a Flexible Piezoelectric Film Sensor and Empirical Mode Decomposition, Proceedings of the 29th Annual International Conference of the IEEE EMBS, 2007, Lyon (France) (2007)

[12] R. Ricci, P. Pennacchi, Diagnostics of Gear Faults Based on EMD and Automatic Selection of Intrinsic Mode Functions, Mechanical Systems and Signal Processing, 25:3 (2011) 821-838

[13] M. Grasso, P. Pennacchi, B.M. Colosimo, Empirical mode decomposition of pressure signal for health condition monitoring in waterjet cutting, International Journal of Advanced Manufacturing Technology, 72:1-4 (2014) 347-364

[14] Z. Peng, F. Chu, Y. He, Vibration signal analysis and feature extraction based on reassigned wavelet scalogram. Journal of Sound and Vibration, 253(5) (2002) 1087-1100.

[15] Z. K. Peng, W. T. Peter, F. L. Chu, A comparison study of improved Hilbert–Huang transform and wavelet transform: application to fault diagnosis for rolling bearing, Mechanical systems and signal processing, 19(5) (2005) 974-988

[16] W. Guo, P.W. Tse, A novel signal compression method based on optimal ensemble empirical mode decomposition for bearing vibration signals, Journal of sound and vibration, 332(2) (2013) 423-441

[17] R. Yan, R. X. Gao, Rotary machine health diagnosis based on empirical mode decomposition, Journal of Vibration and Acoustics, 130(2) (2008) 2007

[18] C. Junsheng, Y. Dejie, Y. Yu, The application of energy operator demodulation approach based on EMD in machinery fault diagnosis, Mechanical Systems and Signal Processing, 21(2) (2007) 668-677

[19] Y. Lei, J. Lin, Z. He, M. J. Zuo, A review on empirical mode decomposition in fault diagnosis of rotating machinery, Mechanical Systems and Signal Processing, 35(1) (2013) 108-126

[20] J. Zhang, R. Yan, R.X. Gao, Z. Feng, Performance Enhancement of Ensemble Empirical Mode Decomposition, Mechanical Systems and Signal Processing, 24 (2010) 2104-2123

[21] A.J. Izenman Review papers: recent developments in nonparametric density estimation, Journal of the American Statistical Association, 86(413) (1991) 205-224

[22] A.W. Bowman, A. Azzalini, Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations, Oxford University Press, Oxford (1997)

[23] S.J. Sheather, Density Estimation, Statistical Science, 9(4), (2004) 588-597

[24] A.Z. Zambon, R. Dias, A review of Kernel Density Estimation with Applications to Econometrics, arXiv:1212.2812v1 [stat.ME], (2012) 1-35

[25] A.W. Bowman, An alternative method of cross-validation for the smoothing of kernel density estimates, Biometrika, 71 (1984) 353–360.

[26] W. Hardle, J.S. Marron, M.P. Wand, Bandwidth choice for density derivatives, Journal of the Royal Statistical Society, Series B (Methodological) (1990) 223-232

[27] M.C. Jones, R.F. Kappenman, On a class of kernel density estimate bandwidth selectors. Scandinavian Journal of Statistics, (1992) 337-349.

[28] M.C. Jones, J.S. Marron, S.J. Sheather, A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association, 91(433), (1996) 401-407.

[29] A.C. Guidoum, kedd: Kernel estimator and bandwidth selection for density and its derivatives. R package version 1.0.3, http://CRAN.R-project.org/package=kedd (2015)

[30] L. Vendramin, R.J. Campello, E.R. Hruschka, Relative clustering validity criteria: A comparative overview, Statistical Analysis and Data Mining: The ASA Data Science Journal, 3(4) (2010) 209-235

[31] G.H. Ball, D.J. Hall, ISODATA, A novel method of data analysis and pattern classification, Tech. Rep. NTIS No. AD 699616, Standford Research Institute, Menlo Park (1965)

[32] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Communication in statistics, 3 (1974) 1–27

[33] J.A. Hartigan, Clustering algorithms, Wiley Ed., New York (1975)

[34] Q. Zhao, M. Xu, P. Fränti, Sum-of-squares based cluster validity index and significance analysis. Adaptive and Natural Computing Algorithms, Lecture Notes in Computer Science, 5495 (2009) 313-322

[35] L. Xu, Bayesian Ying-Yang machine, clustering and number of clusters, Pattern Recognition Letters, 18 (1997) 1167–1178

[36] M. Grasso, B.M. Colosimo, An Automated Approach to Enhance Multi-Scale Signal Monitoring of Manufacturing Processes, Journal of Manufacturing Science and Engineering, 138(5), (2016) 051003 – 051003-16

[37] G. Rilling, P. Flandrin, P. Goncalves, On empirical mode decomposition and its algorithms, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03, Grado (Italy) (2003)

[38] S.J. Loutridis, Damage detection in gear systems using empirical model decomposition, Engineering Structures, 26(12) (2004) 1833–1841.

[39] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S. Springer, New York (2002)

Appendix A: the sifting algorithm

Let $Y(t)$ be a generic signal acquired at sampling frequency F_s. Then, the algorithm to extract the IMFs that capture intrinsic oscillation modes is called "sifting" algorithm, and it works as follows [7]:

1. All the local minima and maxima of the signal $Y(t)$ are identified and they are interpolated respectively by an upper and a lower envelope expressed on a cubic spline basis;
2. the mean of the two envelopes is calculated and designated as $m_1(t)$; then, the difference between the signal $Y(t)$ and $m_1(t)$ is calculated and designated as $h_1(t)$:

$$h_1(t) = Y(t) - m_1(t) \tag{A.1}$$

If $h_1(t)$ satisfies the following conditions:

a) in the entire dataset, the number of extremes and the number of zero crossings must be either equal or different at most by one;
b) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero;

then, $h_1(t)$ is taken as the first IMF of the signal and designated as $c_1(t)$. If $h_1(t)$ is not an IMF, $h_1(t)$ replaces the original signal and the above steps are repeated until an IMF is obtained.

3. The first IMF, $c_1(t)$, is separated from the signal $Y(t)$ by:

$$r_1(t) = Y(t) - c_1(t) \tag{A.2}$$

The residue, $r_1(t)$, is treated as the original signal and the above steps are repeated, leading to the extraction of the following IMFs $c_2(t), \dots, c_n(t)$ such that:

$$\begin{aligned} r_1(t) - c_2(t) &= r_2(t) \\ r_{n-1}(t) - c_n(t) &= r_n(t) \end{aligned} \tag{A.3}$$

At the end of the process, the signal is decomposed into n IMFs and a residue $r_n(t)$:

$$Y(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{A.4}$$

The residue is a signal such that no further decomposition is possible. In this study, the Amplitude Ratio criterion proposed by Rilling *et al.* [37] is used. The EMD algorithm usually converges rapidly in few iterative passes, producing a nearly orthogonal adaptive basis, as discussed in [7] and [38].

Appendix B: On kernel density estimation

The kernel density estimation is a nonparametric approach to estimate the probability density $f(x)$ of a random variable x. The basic idea is to estimate the density function at a point x_i using the neighboring observations, such that the influence of x_i on the estimate at any x vanishes asymptotically. The methodology is widely used in practice, and there is an extensive literature in this field [22, 24, 39]. Two relevant issues consist in the choice of the kernel function, $Ker(x)$, and the selection of an optimal kernel bandwidth, h. The latter issue is the most critical one, and several methods have been proposed thus far. One simple approach is to use rule-of-thumb estimates, which are known to approximate the optimal choice in the presence of normal data [22]. However, when strong departures from normality are observed, other methods should be preferred, which are aimed at estimating the bandwidth, h, in a data-driven way.

The kernel estimator of (x) from a random sample $X_1, \dots, X_p$, denoted by $\hat{f}(x)$, is given in [26]:

$$\hat{f}(x) = p^{-1} \sum_{l=1}^{p} h_k^{-1} Ker((x - X_l)/h_k), \qquad k = 1, \dots, n \tag{B.1}$$

where p is the number of frequency locations and h_k is the kernel bandwidth of the k^{th} CMF extracted from the signal profile $Y(t)$. Notice that, since the probability distribution may considerably change

from CMF to CMF, different optimal choices of h_k can be made, for $k = 1, \dots, n$. The essential idea of UCV is to use the bandwidth, $h_k = \hat{h}$, that minimizes the function:

$$UCV_k(\hat{h}) = \int \hat{f}_{\hat{h}}(x)^2 dx - 2p^{-1} \sum_l \hat{f}_{\hat{h},l}(X_l), \quad k = 1, \dots, n \tag{B.2}$$

where $\hat{f}_{\hat{h}}$ denotes the kernel estimator based on the choice $h_k = \hat{h}$, and $\hat{f}_{\hat{h},l}$ denotes the leave-one-out kernel estimator, defined as follows:

$$\hat{f}_{\hat{h},l}(x) = p^{-1} \sum_{\substack{u=1 \\ u \neq l}}^{p} h_k^{-1} Ker\big((x - X_l)/\hat{h}\big), \qquad k = 1, \dots, n \tag{B.3}$$

Thus, the optimal choice of h_k is defined by:

$$h_k = \hat{h} = arg \min_{h>0} UCV_k(h), \quad k = 1, \dots, n \tag{B.4}$$