Advances in Adaptive Data Analysis
Vol. 5, No. 1 (2013) 1350005 (19 pages)

DOI: 10.1142/S1793536913500052

World Scientific
www.worldscientific.com

ENSEMBLE EMPIRICAL MODE DECOMPOSITION WITH SUPERVISED CLUSTER ANALYSIS

CHIH-YU KUO

Research Center for Applied Sciences, Academia Sinica
Taipei, Taiwan 115, Republic of China

SHAO-KUAN WEI and PI-WEN TSAI*

Department of Mathematics, National Taiwan Normal University
Taipei, Taiwan 116, Republic of China
**pwtsai@math.ntnu.edu.tw*

Received 5 March 2013
Accepted 19 March 2013
Published 23 April 2013

Ensemble empirical mode decomposition (EEMD) is a noise-assisted data analysis method which decomposes a signal into a collection of intrinsic mode functions (IMFs). There nevertheless appears a multi-mode problem where signals with a similar timescale are decomposed into different IMF components. A possible solution to this problem is to recombine the multi-mode IMF components into a proper single mode but as of yet, no general rules have been proposed in the literature. This paper presents the incorporation of a statistical cluster analysis to assist in the diagnosis of multi-mode IMFs and to recombine them based on the classified clusters. As a result, signals are reorganized into a condensed set of clustered intrinsic mode functions (CIMFs). The method is applied to two sets of artificially synthesized signals and two sets of practical signals: wind turbine noise and earthquake motion. These applications demonstrate that, with the additional cluster analysis, the multi-mode problem can be largely eliminated in a statistically reliable manner, and *in situ* applications can be improved.

Keywords: Cluster analysis; clustered intrinsic mode function; cluster linkage; multi-mode.

1. Introduction

Ensemble mode decomposition (EMD) is an adaptive time-frequency data analysis method which decomposes time series or signals into a collection of intrinsic mode functions (IMFs) [Huang *et al.* (1998)]. This decomposition is based on the local characteristic timescale of the signal, which makes EMD applicable for analyzing nonlinear and nonstationary signals. It has been applied with great success in a broad range of applications, such as biological and medical sciences, geology, astronomy, engineering, and others, e.g. Huang *et al.* [1998], Echeverria *et al.* [2001], Yu *et al.* [2005], Wu *et al.* [2007]. Despite these successful applications, the major

drawback of EMD is found to be the mode mixing phenomena, wherein a single decomposed IMF may consist of signals of significantly different scales, and these mode mixed IMFs may contain mixed signal sources with mixed time scales, which hinder the interpretation of analysis.

To minimize the mode mixing problem, a noise-assisted data analysis method was proposed, the *ensemble empirical mode decomposition* (EEMD) [Wu and Huang (2009)]. The EEMD defines IMF components as the mean IMFs of ensemble trials of the original signals. Each ensemble trial is obtained by adding a set of white noise signals at a specified magnitude and, through this ensemble procedure, parasitic mode mixing of intermittent signals is largely removed; see for example applications in Chang *et al.* [2010], Yeh *et al.* [2010], Lei *et al.* [2011], Huang and Xu [2011] and Mhamdi *et al.* [2011]. As noted in Wu and Huang [2009] and Balocchi *et al.* [2003], however, EEMD usually accompanies multi-mode problems where signals with similar scales appear in different IMFs and causes an over-complete decomposition. When using EEMD to study the seismic signals of the Chi-Chi earthquake (Taiwan, 1999), coupled with the Tsaoling landslide motion, see e.g. Sec. 4.3, we find that the multi-mode renders the analysis somewhat inconclusive. To highlight the multi-mode phenomena, we use a synthesized signal with intermittent wave packets with the EEMD method in Sec. 2.

As addressed in Wu and Huang [2009], a possible solution to the multi-mode phenomena is to combine the multi-mode IMFs into a proper single mode, but to the authors' limited knowledge, no general guidelines have been established to date. In the present paper, we propose to incorporate statistical cluster analysis to diagnose the multi-mode IMFs and to group them according to the classified clusters. In the cluster analysis, correlation coefficients are used to measure the "closeness" among the IMFs and a hierarchical clustering technique, the dendrogram, is used to present the clues for a supervised decision for combing the IMFs. The result is that a condensed set of clustered IMFs (CIMFs), with minimized multi-modes, is formed. The details of the procedure will be described in Sec. 3.

Finally, we demonstrate with three application examples in Sec. 4 and recapitulate the findings in concluding remarks in Sec. 5 to show that robust results can be achieved with this additional supervised cluster analysis. These examples include one sophisticated artificial signal, wind turbine noise and an abbreviated analysis of the co-seismic ground motion of the Tsaoling landslide in the Chi-Chi earthquake.

2. EEMD and Multi-Mode

In this section, we briefly describe the EMD technique; its successor, EEMD; and the multi-mode phenomena by using a synthesized signal sequence. The EMD technique is a procedure through which a signal sequence $x(t)$ is decomposed into a set, or a hierarchy, of IMFs, $c_i(t)$. The decomposition reads:

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

where n is the total number of IMFs and $r_n(t)$ is the nth residue, called a trend. The procedure is a sifting process that begins by defining a transient mean of the signal and continues by extracting the offset signal from the mean to form IMFs. The transient mean is the average of a pair of bounding envelopes of the signal, wherein the bounding envelopes are spline curves of local extrema of the signal. A variety of spline or interpolation methods can be applied for obtaining the signal envelopes and this procedure is executed iteratively for the hierarchy of the n IMFs. To avoid a digression, we recommend that for further details, readers please refer to Huang *et al.* [1998].

When the sifting procedure stops, the set of n IMFs forms. These IMFs are sorted ascendantly according to the timescale of the IMFs. The IMFs have the following property: (i) their number of extrema and zero-crossings either are equal or differ at most by one, (ii) the mean values of their bounding envelopes are zero, and (iii) they are nearly orthogonal. The residue r_n, on the other hand, becomes a constant, a monotonic function, or a function that contains only one extremum from which no further IMFs can be extracted. In practice, the number of IMFs, n, is often determined by the length of the data, N, and Wu *et al.* [2007] suggest that n is about $\log_2 N$.

As mentioned in the introduction and existing literature, since EMD sometimes encounters a mode mixing problem and hence EEMD is proposed, [Huang *et al.* (1998)]. In EEMD, an additional ensemble average with the assistance of white noise is incorporated. The white noise contains a broad range of timescales, which helps separations of mode-mixed intermittent signal packets in EMD. This additional ensemble procedure is also an iterative process. In each iteration, an ensemble of the signal is prepared by adding the original signal with a white noise sequence at a prescribed root-mean-square (rms) amplitude. EMD is then applied on the ensemble to extract its IMFs. Finally, the EEMD is obtained by taking averages of the ensemble IMFs over a desired number of ensembles. To eliminate bias, a sufficiently large number is chosen for the number of ensembles, usually 30. Wu and Huang [2004] and Flandrin *et al.* [2004] point out that EEMD has the effect of a dyadic filter bank, and the filter bank is adaptive to the characteristics of the signal. With the ensemble averaging, the new IMFs of EEMD no longer satisfy the properties of the EMD IMFs. Especially, the relaxation from orthogonality enables the cluster analysis described in the next section.

Under some circumstances, EEMD encounters multi-mode phenomena. These phenomena are seen as the multiple number of decomposed IMFs with similar timescales. To exaggerate them, we illustrate the EEMD analysis of a synthesized signal with intermittent wave packets. The synthesized signal is defined as

$$x(t) = \begin{cases} \sin(t) + 0.1\cos(10t), & t \in \left[\dfrac{(2i-1)}{2}\pi \pm \dfrac{\pi}{5}\right], \ i \in \mathbb{Z}, \\ \sin(t), & \text{otherwise.} \end{cases} \tag{2}$$

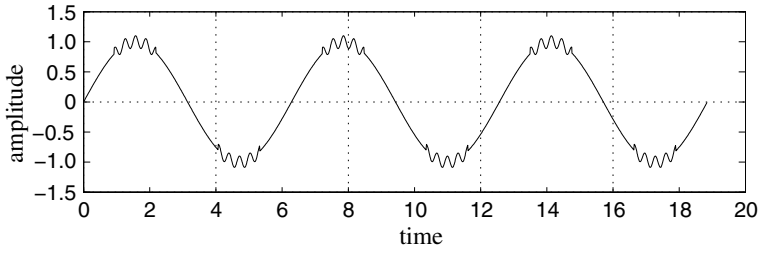


Fig. 1. Synthesized signal, (2).

which contains a fundamental harmonic sine wave at a unit amplitude and period 2π. The intermittent wave packets are generated by a cosine wave at amplitude 0.1 and period $\pi/5$, and they ride on the crests of the sine wave. The signal, discretized at a sampling frequency of 100 per unit time, is plotted in Fig. 1.

The signal is decomposed with EEMD by using the benchmark Matlab programs downloadable at the internet website http://rcada.ncu.edu.tw/research1.htm. In the calculation, the white noise is set at an amplitude of 10% of the rms value of the signal, and 30 ensembles are used. The decomposed IMFs are sketched in Fig. 2. With the benchmark program, the signal is decomposed into nine IMFs and one residue trend. They are sorted from the mode with the shortest timescale to the one with the longest. The first two modes present similar characteristics to the leading modes in many other EEMD applications that contain mostly the white noise. Their randomness can be confirmed by the significance test proposed in Wu and Huang [2004], as shown by the two modes lying on the margin of white noise in Fig. 3. In addition to the white noise, the second IMF has outstanding fluctuations at both starting and ending time instances of each high frequency intermittent wave packet. These two leading IMFs may collapse into one if the signal is discretized using a slower sampling rate.

The third and fourth IMFs, $\text{IMF}_{3,4}$, clearly include the intermittent signals. Though the intermittent signals have a monotonic frequency, they are decomposed into two IMFs. This phenomenon is referred as the multi-mode problem. The same phenomenon is also seen in the decomposition of the fundamental harmonic wave, which is in the fifth and sixth IMFs, $\text{IMF}_{5,6}$. From their amplitudes, the harmonic components in the original signal are almost equally distributed into the multi-mode IMF pairs. Because the EEMD has the effect of an adaptive dyadic filter bank, the multi-mode phenomena may be associated with the overlaps of the frequency response bands of adjacent filters in the filter bank. The last three IMFs and the residue trend are insignificant low frequency modes because of their small amplitudes and no more than two representative periods in the entire signal duration.

A few methods are proposed to overcome this multi-mode problem. The most notable alternative is to tune the level of the added white noise and the number of ensemble trials. With this approach, we report the IMFs of interest (IMF_3 to

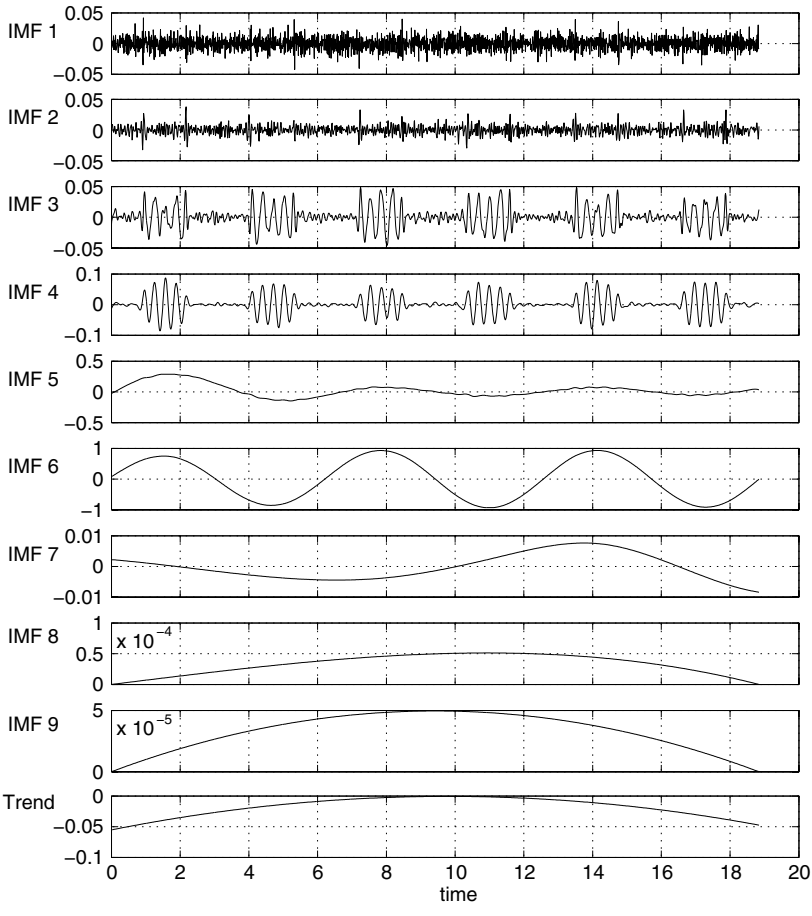


Fig. 2. IMFs of the synthesized signal, (2). The white noise of 10% of amplitude of the signal. Note that the vertical axes are not equally scaled.

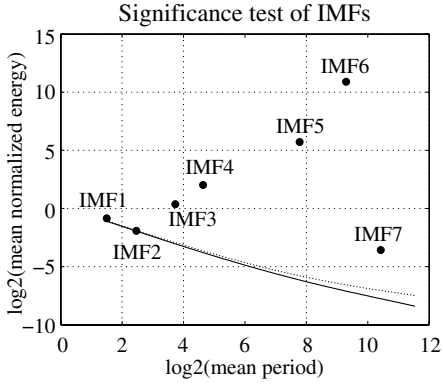


Fig. 3. Significance test of IMFs of the synthesized signal (2). The lower thick solid (upper dotted) line represents the upper bound of Gaussian noise at a 95% (99%) confidence level.

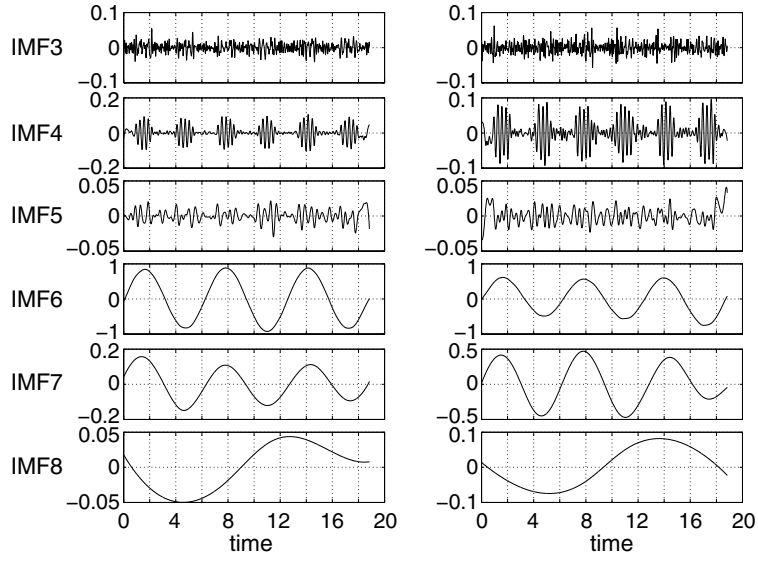


Fig. 4. IMF_3 to IMF_8 of the synthesized signal (2) with (a) 20% and (b) 30% level of white noise.

IMF_6) with a noise levels 20% and 30% in Fig. 4. Significant improvements can be found for the intermittent wave packets, which are concentrated into IMF_4 in both cases. The side effects are that two IMFs with random waves replace IMF_3 and IMF_5. Because the noise levels are not comparable to the fundamental wave and the signal duration is limited, the EEMD does not resolve the multi-modes IMF_6 and IMF_7.

The other alternative is to recombine IMFs into a single mode. With the knowledge of this synthesized signal, we can combine the two pairs ($IMF_{3,4}$ and $IMF_{5,6}$) into two modes for the fundamental and intermittent waves. However, this prior information does not usually exist in real applications. A conceptual approach on utilizing the orthogonal properties of IMFs was mentioned in Wu and Huang [2004], but guidelines have not been proposed at the time of writing. This concept inspires the present proposal of incorporation of cluster analysis, which will be described in the next section.

3. Incorporation of Cluster Analysis

Cluster analysis is a set of statistical methods that is used to identify and classify objects or variables into groups, called clusters. After the cluster analysis, objects in the same cluster are statistically similar to each other. There are two common clustering procedures: hierarchical and nonhierarchical procedures. In hierarchical clustering, a set of nested clusters is produced, and this set is often displayed with a

tree-like diagram, called a dendrogram. In dendrograms, the relations among clusters and subclusters can be inspected straightforwardly. In nonhierarchical clustering, on the other hand, objects are clustered by merging and splitting algorithms. The fundamentals of cluster analysis can be found in many standard statistics textbooks, such as Hair Jr. *et al.* [1992].

In the present paper, the hierarchical clustering method begins by treating each object as a singleton cluster, and then it successively merges clusters until all points have been merged into a single remaining cluster. This approach is called an agglomerative clustering algorithm, which clusters objects in a bottom-up manner. Two distance measures are needed in the clustering procedures: One is to measure the distance between any pair of objects (or variables), and the other is to measure the distance between clusters that contain single or multiple objects. We refer the former as the distance and the latter as the cluster distance or cluster linkage. These distance measures are the amalgamation rule that is applied to determine if objects or clusters are sufficiently "similar" to be linked together.

When applied with EEMD, the IMFs are the objects for clustering. We define the distance between IMF_i and IMF_j as $d_{i,j} = 1 - \text{cor}(c_i, c_j)$, where $\text{cor}(c_i, c_j)$ is the correlation coefficient between IMF time sequences c_i and c_j. This distance measure is chosen because the "closeness" of the oscillatory IMFs is calculated based on the in-phase components: i.e. the effect of closeness of timescales in the IMFs is included. If two IMFs contain higher portions of in-phase fluctuations, the distance between them is smaller. In applications, the EEMD extraction process is performed by repetitively removing transient means from the signals so that the process does not produce negatively correlated *neighboring* multi-mode IMFs under usual circumstances, such that the distance between the multi-mode IMFs is between 0 and 1. This type of cluster linkage is also used in a variety of subject fields, such as Bien and Tibshirani [2011].

The cluster distance, on the other hand, has various definitions according to different linkage criteria. Let us denote the cluster distance as $D_{\alpha,\beta}$ for a cluster pair α and β, and briefly describe two common linkage criteria used in the present paper. For a cluster pair that may contain multiple IMFs in each cluster, their distance can be defined to be the closest distance or to be the farthest distance between the elementary IMFs that join the cluster pair. The first linkage criterion is called the single linkage, and the second is the complete linkage. These linkage criteria can be expressed symbolically as $D_{\alpha,\beta} = \min\{d_{i,j} : i \in \alpha, j \in \beta\}$ and $D_{\alpha,\beta} = \max\{d_{i,j} : i \in \alpha, j \in \beta\}$. There are many other linkage criteria in the literature, but we shall not elaborate further for the sake of simplicity.

The agglomerative hierarchical clustering procedure can be executed in the following algorithmic steps:

(i) Treat each IMF as an individual cluster and assign an initial cluster as cluster α for iteration (excluding the insignificant residue trend),

(ii) Compute the cluster distance $D_{\alpha,\beta}$ by exhausting cluster β in the rest cluster population and find the cluster pair (α, β) in which $D_{\alpha,\beta}$ is minimum,
(iii) Amalgamate the clusters α, β, into a new cluster and re-index the new cluster as α,
(iv) Repeat steps (ii) and (iii) until all of the clusters are in the hierarchy of clusters, and
(v) Inspect the resultant representative dendrogram of the cluster hierarchy and draw a clustering level to regroup the member IMFs (supervised clustering).

Note that when the dendrogram representation of the cluster hierarchy is obtained at the end of the algorithm, a subjective inspection for the multi-mode IMFs is performed. Terminologically, it is a supervised clustering. We do not particularly emphasize the performance efficiency in the cluster algorithm because the number of IMFs is moderate.

The full procedure is illustrated using signal (2) and the IMFs in Fig. 2. The complete linkage strategy is applied for this example, and Fig. 5 presents the dendrogram of the cluster analysis. In dendrograms, attention is paid to the positions of the joints, called nodes, connecting the IMFs and clusters. In the figure, the joints are the horizontal line segments, and their vertical positions indicate the cluster distance between the connected sibling clusters. The vertical scale is plotted for the reference of the supervised inspection.

It is clearly seen from the small distances that there are three multi-mode pairs of IMFs: (IMF_3, IMF_4), (IMF_5, IMF_6) and (IMF_8, IMF_9). The small distance between the clustered IMFs implies that the correlation coefficients in the pairs are higher than 0.5. A gap region is found between 0.3 and 0.8, and a straightforward selection of the clustering level is to choose a level in this gap. By doing so, we can cluster the IMFs into six clustered IMFs, illustrated as the six light-grey shaded blocks in Fig. 5.

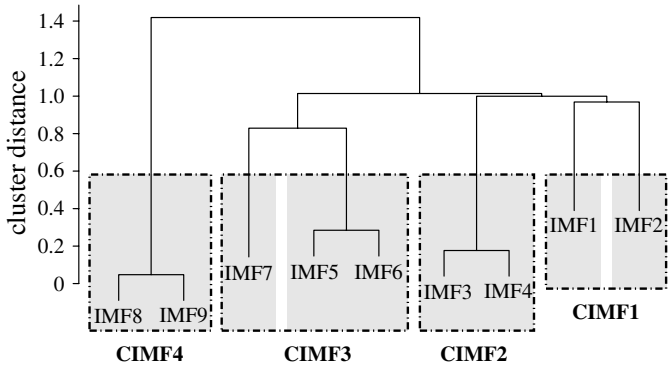


Fig. 5. Dendrogram with complete linkage of the synthesized signal (2). The vertical axis is the scale for the cluster distance. The horizontal appearance order of IMFs is irrelevant in the cluster analysis.

The supervised approach allows us for a slightly more aggressive clustering strategy to further reduce the complexity of the IMF set. This aggressive strategy involves increasing the clustering level close to distance 1. This means clustering noncorrelated IMFs, so special cautions are needed. The cautions involve inspecting the inter-relations between the IMFs and clusters and the IMF properties: e.g. timescales and amplitudes. From the dendrogram, the next clustering is to combine IMF_7 into (IMF_5, IMF_6), and it is acceptable because IMF_7 has a small amplitude at a neighboring timescale (Fig. 2) and a relatively insignificant signal (Fig. 3) to the merged pair. Finally, even though the distance is almost 1, IMF_1 and IMF_2 actually arise from the white noise of the ensemble average process, and they can be safely discarded or, as done here, regrouped for illustration.

After the supervised clustering, the resultant clustering strategy is shown by the clusters outlined with the bold dash-dot lines in Fig. 5. There are four clusters. The IMFs in each cluster are summed up and a set of clustered IMFs, now called CIMFs, is formed. These CIMFs are plotted in Fig. 6. The intermittent signals are unambiguously clustered into CIMF_2, while the main harmonic signal is clustered into CIMF_3. CIMF_1 retains the characteristics of white noise, and CIMF_4 is the negligible mode of long timescale caused by the end effects of the original signal. Although there is a subjective supervised procedure, the cluster analysis provides a deterministic clue to iteratively combine the multi-mode IMFs.

The aforementioned complete analysis is repetitively elaborated to verify the algorithmic robustness for a range of signal sampling, ensembling parameters and the effect of cluster analysis linkage types. The parameter adjustments include varying the noise level and number of ensemble trials of EEMD, resampling the original signal to different sampling rates, etc. It is found that their effects are at most relocations of the joint positions (levels) in the dendrograms or, sometimes,

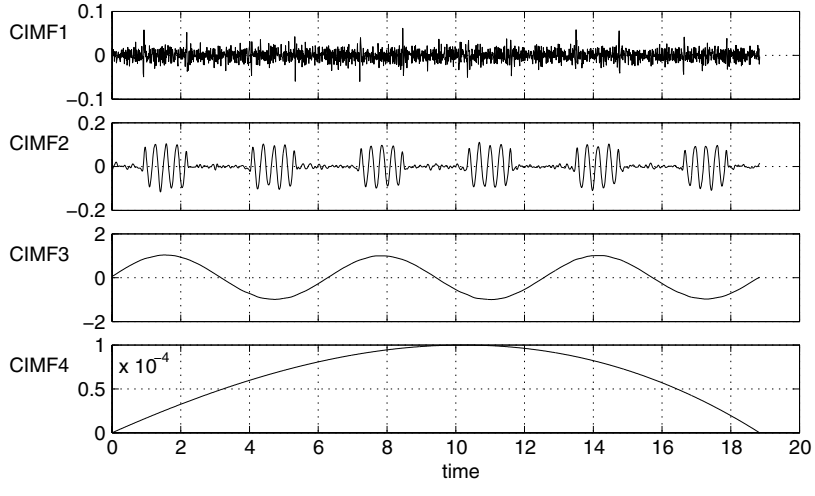


Fig. 6. CIMFs of the synthesized signal (2).

the hierarchical relation among the insignificant high IMF modes. In conclusion, the cluster analysis is robust for resolving the multi-mode EEMD phenomena.

4. Applications

The method is applied to three sets of different data. The first is a synthesized signal and the others are applications to two sets of practical data: wind turbine noise and an earthquake signal.

4.1. *Synthesized signal with three components*

The synthesized signal in this example is composed of two harmonic sine waves and a set of high frequency intermittent wave packets. The signal reads

$$x(t) = \begin{cases} \sin(t) + 0.1\sin(10t) + 0.1\cos(40t), & t \in \left[\dfrac{(2i-1)}{2}\pi \pm \dfrac{\pi}{5}\right],\ i \in \mathbb{Z}, \\ \sin(t) + 0.1\sin(10t), & \text{otherwise} \end{cases} \tag{3}$$

and is plotted in Fig. 7. The signal is decomposed by EEMD with a white noise level at 10% of the rms values of the signal for 30 ensemble trials. Figure 8 shows the IMFs and the intermittent wave packets are found in the multi-modes IMF_2 and IMF_3. Similarly, multi-mode pairs are also found in (IMF_4, IMF_5) and (IMF_6, IMF_7) pairs.

The cluster analysis yields a dendrogram sketched in Fig. 9, and following the same supervised procedures as described in Sec. 3, we finalize the five clusters indicated in the figure. The CIMFs are plotted in Fig. 10. It is clear that the three main components of the signal are cleanly clustered in CIMF modes 2 to 4. As usual, $CIMF_1$ is a random noise of the ensemble average and $CIMF_5$ is the long timescale trend arising from the end effect.

4.2. *Wind turbine signal*

Wind power is becoming an important sustainable source of energy. Because of the interactions between rotating blades and wind flows, wind turbines produce noise

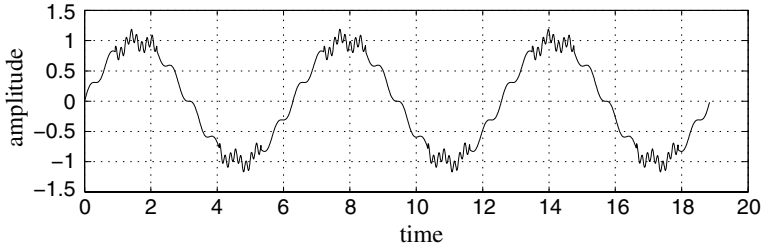


Fig. 7. Synthesized signal, (3).

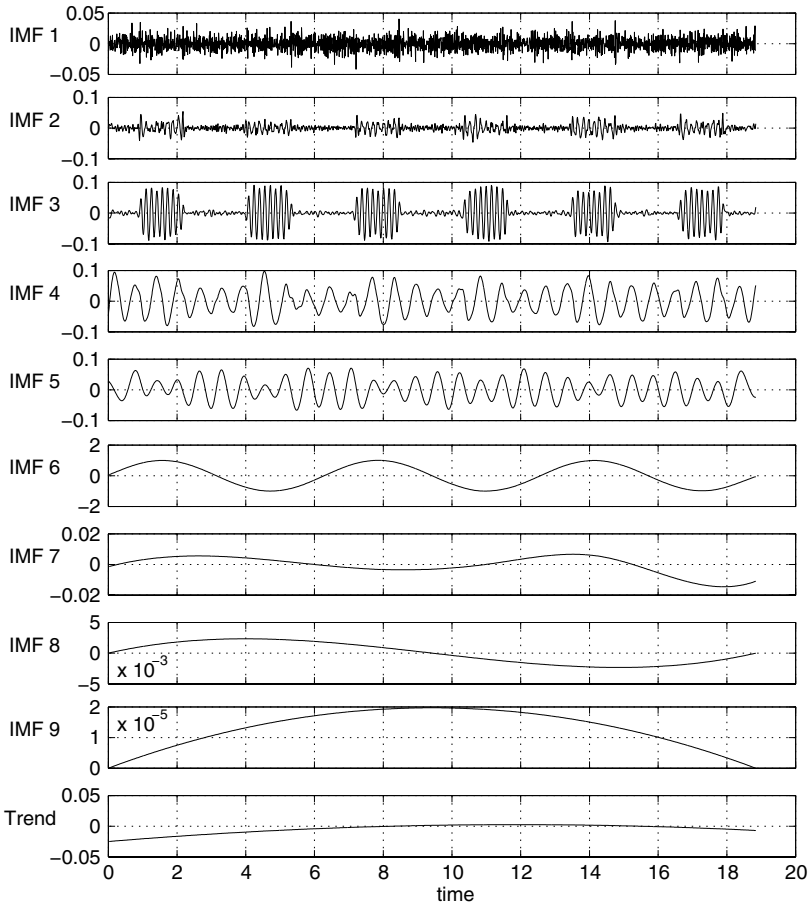


Fig. 8. IMFs of the synthesized signal (3).

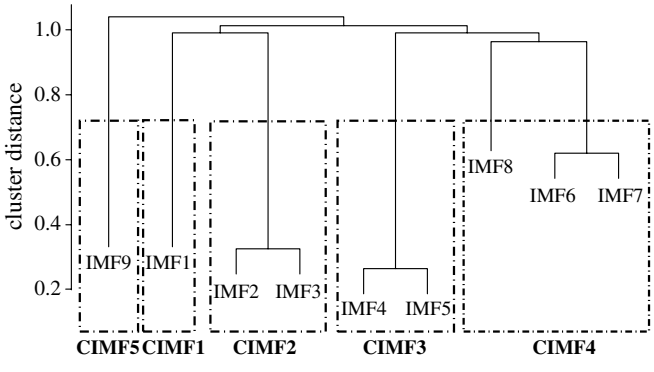


Fig. 9. Dendrogram with complete linkage of the synthesized signal (3).

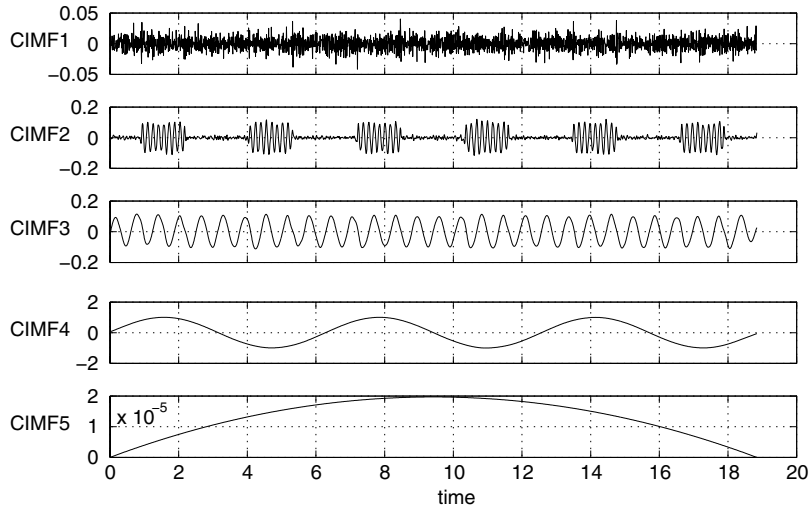


Fig. 10. CIMFs of synthesized signal (3).

radiation. The numbers of articles and reports on the effects of wind turbine noise on nearby living communities are growing fast [Colby *et al.* (2009); Pierpont (2009)]. Wind turbine noise is mainly composed of broadband low frequency noise [Wagner *et al.* (1996)]. The frequency range starts from an inaudible few Hertz (infrasound) to a few hundred Hertz. It is argued that low frequency noise may interact with the human body through mechanical resonances and cause undesirable physiological and psychological effects.

In addition to the broadband low frequency spectrum, parts of wind turbine noise have intermittent characteristics because a portion of the noise comes from the tip-flow interactions modulated with blade passing frequency. This type of signal has multiple timescales. The tip-flow noise has relatively high frequencies, while blade passing is at the low frequency end. Under practical operational conditions, these noises are also affected by the transient wind field, such as unsteady wind turbulence and gusts. In this example, we demonstrate that the present EEMD analysis can provide a valuable technique to evaluate these nonstationary effects on noise radiation.

A pilot measurement was performed at Chunan town, Miaoli County, Taiwan, on the 17 November 2011. The wind was blowing north or north-east and the wind ground speed was about $7 \sim 8$ m/s. The wind turbine was a typical horizontal axial type (HAWT) with three blades. The tower height was 67 m, the blade length was 35 m and the power was rated at 2 MW. The rotational speed was slightly lower than 20 rpm at the time of measurement, which yielded a tip velocity of the blade of about 70 m/s. A 1/2" free field microphone, Brüel and Kjær 4190, was used. The microphone was positioned 67 m north (upwind) of the wind turbine. For demonstration of the EEMD capabilities, this position was chosen without regard

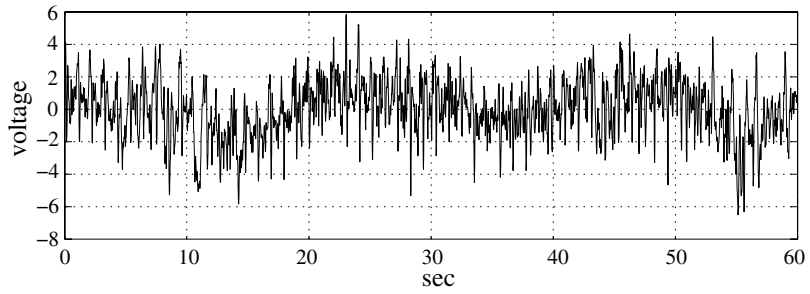


Fig. 11. The wind turbine noise. The sampling rate is 1,087 Hz.

to the worst direction of noise pollution, or the sound pressure calibration from the voltage reading of the digital data acquisition.

A 60 s noise signal sampled at 1,087 Hz was applied for EEMD analysis. The original signal is plotted in Fig. 11, in which the multiple time scales of the noise and the blade passing pulses are clearly seen. This signal is decomposed into 14 IMFs with a 10% white noise level and 30 ensemble trials. The dendrogram of the IMFs with the complete linkage criterion is plotted in Fig. 12. In the following discussion, a slightly aggressive clustering strategy, as shown in the figure, is adopted for a smaller set of CIMFs for convenience. The clustering level is set slightly less than 1, about the joining level of $IMF_{5,6}$ and IMF_7.

Figure 13 depicts the CIMFs. The significance test indicates that $CIMF_1$ is a negligible random noise mode. $CIMF_2$ has a maximum frequency content near 20 Hz, although it is also a broadband mode. If this noise is excited by turbine blades cutting through turbulent eddies, the eddy size is estimated to be on the order of a few meters based on the blade tip velocity. The agreement of this size with the blade chord length implies that the speculation is reasonable. $CIMF_3$ and $CIMF_4$ obviously contain the blade passing noise, roughly at 60 rpm (three blades at

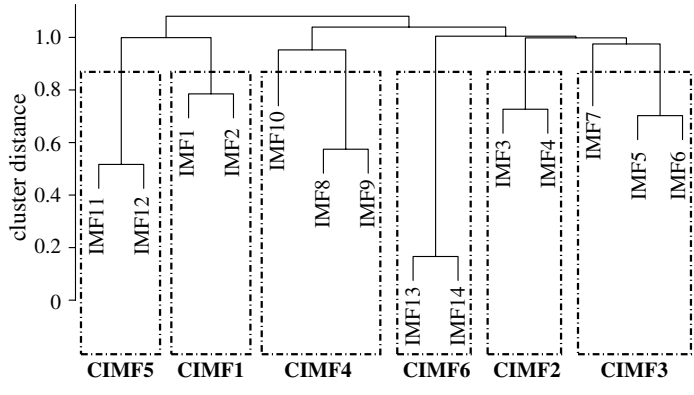


Fig. 12. Dendrogram with complete linkage of the wind turbine noise.

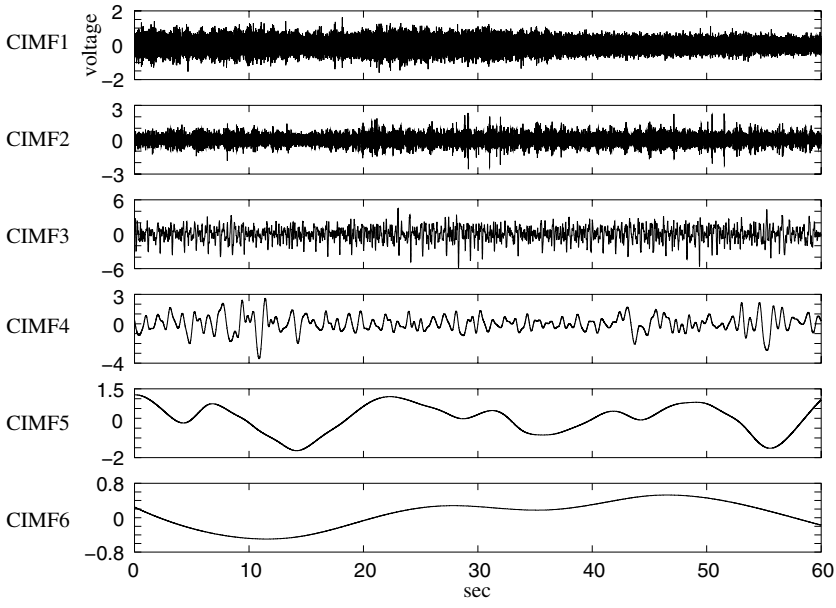


Fig. 13. CIMFs of the wind turbine noise.

20 rpm). A close inspection on the wave forms reveals more details, a brief overview of which will be provided in the next paragraph. $CIMF_5$ has a long fluctuation period of about 10 s, despite some distortion. Assuming the wind speed is 10 m/s, fluctuations at this period have length scales about the order of one hundred meters. Such length scales correspond to a flow eddy size about the diameter of the windex turbine. $CIMF_6$ has an extremely long period that is comparable to the entire signal duration; hence, it may contain signal-end effects, and no conclusion on its sources can be drawn.

From their amplitudes, $CIMF_3$ and $CIMF_4$ have the largest portion of spectral energy and they show fluctuations and signal spikes at the blade passing frequency. For closer inspection, both of the CIMFs between 0 and 6 s are plotted in Fig. 14, with an equal vertical scale. $CIMF_2$ is also superposed in the figure for comparison. Cross-examining the cluster distance between $CIMF_3$ and $CIMF_4$, we find that they are negatively correlated, and from the figure, these two modes present two distinctive characteristics: $CIMF_3$ has discrete sharp peaks between each blade passing and has maximum peak spikes at instances slightly retarded after the maxima of $CIMF_4$. The sound sources of these peaks are clearly the blade tip vortex sound which makes the "swish" noise that observers hear near wind turbines. Because the sound generating mechanism is localized at the blade tip, the sound becomes prodominant only when the blade passes a particular position relative to the observer. $CIMF_4$, on the other hand, is a smooth fluctuation that produces an infra-sound at about 1 Hz. In each oscillation period, it has a steeper ascending rate than the descending

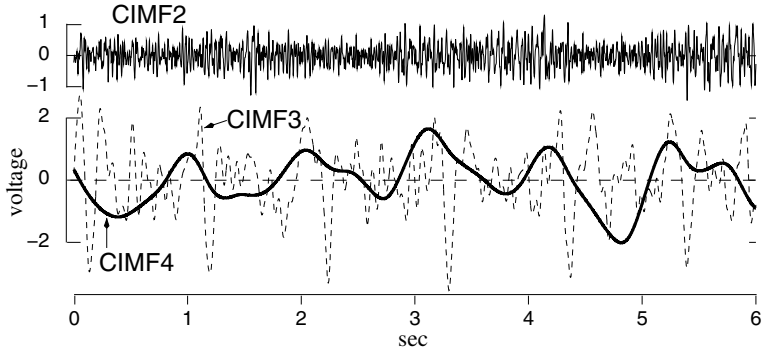


Fig. 14. Close view of $CIMF_2$, $CIMF_3$ and $CIMF_4$.

rate, which signifies the Doppler frequency shift of the blade rotation. $CIMF_2$ shows no correlation to the other two modes. Further quantification of the wind turbine noise is beyond the scope of the present paper and will be reported in a designated article.

4.3. *A seismic signal from station CHY080 at Chi-Chi*

On the 21th September 1999, the Chi-Chi earthquake in Taiwan triggered a major landslide in the Tsaoling area [Hung (2000); Chang and Taboada (2009)]. The detached landslide mass had a volume about $125 \times 10^6\,m^3$. A strong seismic station, code-named CHY080, recorded the ground acceleration during the earthquake and the landslide. CHY080 is located about 200 m from the north-east boundary of the scar area and its projection onto the NE-SW profile is shown in Fig. 15. The NE–SW (north-east to south-west) profile, the solid thick black line in the inset figure, is in parallel to the slide direction.

The acceleration have three vectoral components: the EW, NS, and vertical components. In this application, we are particularly interested in a burst of high

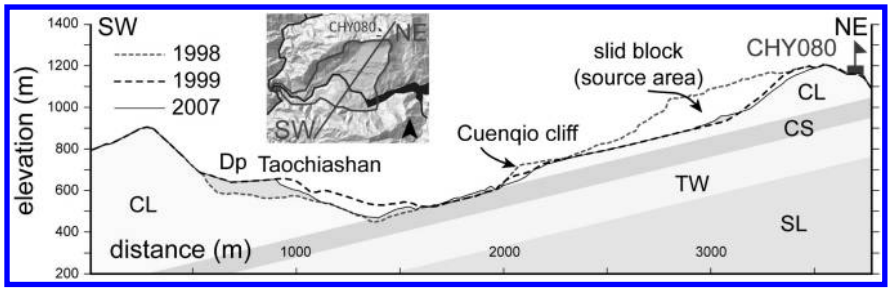
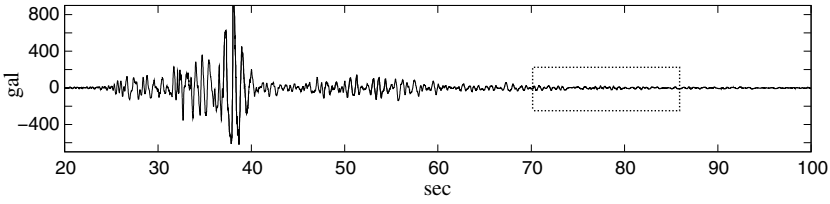
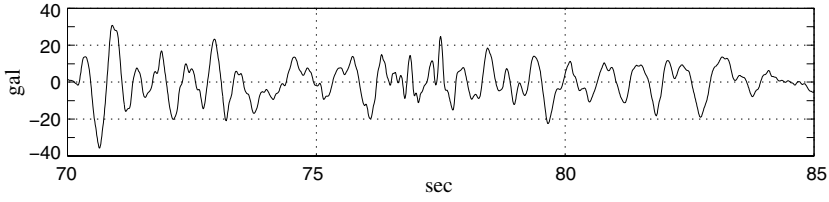


Fig. 15. Geological profile of the Tsaoling area. The inset figure depicts the surrounding area of the landslide site. The NE-SW profile is defined through the gravity center of the slid mass and is parallel to the slide direction. Geological formation: SL = Shihliufeng shale; TW = Tawo sandstone; CS = Chinshui shale; CL = Cholan formation; Dp = debris deposit.



(a)



(b)

Fig. 16. EW ground acceleration component of the CHY080 seismic record. (a) Full record, p-wave arrives at 20 s and s-wave arrives at 25.2 s. (b) Excerpted signal section between 70 and 85 s, the highlighted signal section in (a). The signal is conditioned for instrumental errors and low-pass filtered at 20 Hz.

frequency which was visually identified around 76 s from the start of the record and, for brevity, only the EW component is addressed here. The sampling rate of the seismic station is 200 Hz. For EEMD analysis, the signal is conditioned by removing the instrumental offset and baseline [Boore (2001); Wu and Chen (2011)], and then applied with a 20 Hz low pass FIR filter. The full and excerpted EW signals (conditioned) are plotted in Fig. 16. With the usual 10% noise level and 30 ensemble trials, the EEMD of the excerpted signal yields 9 IMFs.

After a complete linkage cluster analysis, the dendrogram of the IMFs is sketched in Fig. 17. Based on the supervised procedure described in Sec. 3 and the timescales of IMFs, we set the clustering level between 0.7 and 0.9, which leads to five CIMFs. These CIMFs are plotted in Fig. 18. It is confirmed by the significance test that CIMF_1 is a negligible ensemble averaged random noise. CIMF_3 and CIMF_4 contain more than 81% of the total spectral energy of the acceleration signal and are the main earthquake components. Awaiting further investigation is whether any mechanisms divide the main earthquake signal into two CIMFs with distinctive time scales. CIMF_5 absorbs the end effects of the sectioned signal and is insignificant to the ground motion.

The striking finding in the clustered IMFs is the localized wave packets in CIMF_1. In addition, similar localized wave packets are also found in the other two acceleration components. The timescale of these wave packets is much shorter than that of the ambient ground motion. From the high frequency contents, these wave packets are likely caused by the coseismic near-field landslide motion. Using numerical simulations, Kuo *et al.* [2009] found that the emerging instance of the

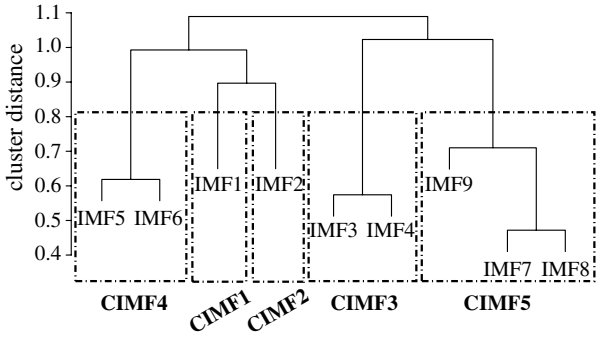


Fig. 17. Dendrogram of EEMD cluster analysis of the CHY080 seismic signal (EW component).

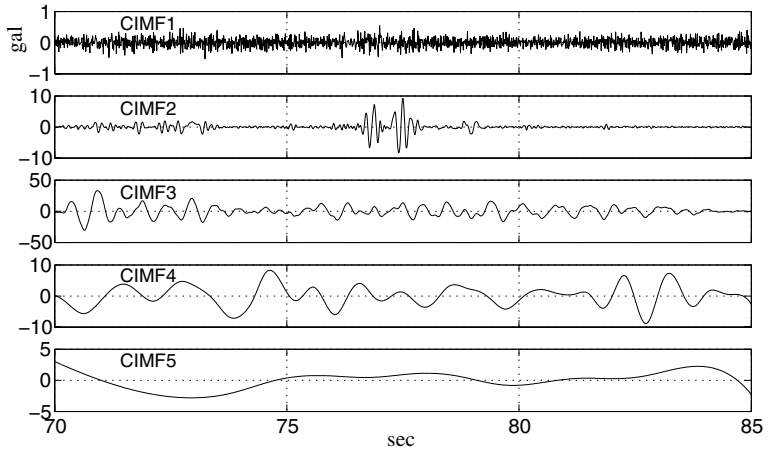


Fig. 18. CIMFs of the CHY080 seismic signal (EW component). The vertical axes are not equally scaled.

wave packets coincides with the instance of the landslide flow impacting the deposit river valley. It is further verified in Chang *et al.* [2012] that the magnitude of the wave packets is found to be consistent with that of the corresponding energy release of the landslide impact. For illustration purposes of the modified EEMD, we do not proceed any further but conclude here with the encouraging finding of the impact signal. In summary, although it is a preliminary application and there is in a great lack of details, the single linkage criterion was actually used in the last citation [Chang *et al.* (2012)] to which the interested readers are referred for extra verification of the robustness of the cluster analysis.

5. Concluding Remarks

In this paper, we present the incorporation of supervised cluster analysis with EEMD. The additional cluster analysis provides a deterministic method to diagnose

the multi-mode phenomena of EEMD and a formal approach to cluster the multi-mode IMFs. An agglomerative hierarchical clustering procedure is applied and the cluster distance (linkage) is defined on the correlation coefficient between the IMFs pairs. Dendrograms of IMF hierarchies produced after the analysis provide the visual clues for the supervised inspection and clustering. The complete procedure is demonstrated with details using two synthesized and two practical signals. These applications conclude that the multi-mode problem can be largely eliminated in a statistically reliable manner and *in situ* applications can be improved.

Acknowledgments

This work is supported in parts by National Science Council, Taiwan, under grants NSC-100-2625-M-001-002-MY3 (CYK) and NSC-99-2118-M-003-001-MY2 (SKW, PWT). The help of Dr. Gwo-Shyang Hwang and Mr. Yun-Peng Wu, National Taiwan University, in data acquisition of the wind turbine noise is credited.

References

Balocchi, R., Menicucci, D. and Varanini, M. (2003). Empirical mode decomposition to approach the problem of detecting sources from a reduced number of mixtures, in *Proc. Int. Conf. IEEE EMBS*, Cancun, Mexico, pp. 2443–2446.

Bien, J. and Tibshirani, R. (2011). Hierarchical clustering with prototypes via minimax linkage. *J. Am. Stat. Assoc.*, **106**: 1075–1084.

Boore, D. M. (2001). Effect of baseline corrections on displacements and response spectra for several recordings of the 1999 Chi-Chi, Taiwan, earthquake. *Bull. Seis. Soc. Am.*, **91**: 1199–1211.

Chang, K. J. and Taboada, A. (2009). Discrete element simulation of the Jiufengershan rock-and-soil avalanche triggered by the 1999 Chi-Chi earthquake, Taiwan. *J. Geophys. Res.*, **114**: F03003.

Chang, K. J., Wei, S. K., Chen, R. F., Chan, Y. C., Tsai, P. W. and Kuo, C. Y. (2012). Empirical modal decomposition of near field seismic signals of Tsaoling landslide, in *Earthquake-Induced Landslides*, eds. K. Ugai, H. Yagi and A. Wakai, Int. Symp. Earthquake-induced landslides, Springer, Kiryu, pp. 421–430.

Chang, Y.-M., Wu, Z., Chang, J. and Huang, N. E. (2010). Model validation based on ensemble empirical mode decomposition. *Adv. Adapt. Data Anal.*, **2**: 415–428.

Colby, W. D., Dobie, R., Leventhall, G., Lipscomb, D. M., McCunney, R. J., Seilo, M. T. and Søndergaard, B. (2009). *Wind Turbine Sound and Health Effects*: *An Expert Panel Review*, Prepared for American Wind Energy Association and Canadian Wind Energy Association.

Echeverria, J., Crowe, J., Woolfson, M. and Hayes-Gill, B. (2001). Application of empirical mode decomposition to heart rate variability analysis. *Med. Biol. Eng. Comput.*, **39**: 471–479.

Flandrin, P., Rilling, G. and Goncalves, P. (2004). Empirical mode decomposition as a filter bank. *IEEE Signal Process. Lett.*, **11**: 112–114.

Hair, Jr. J. F., Anderson, R. E., Tatham, R. L. and Black, W. C. (1992). *Mutivariate Data Analysis with Reading*. Macmillan Publishing Company.

Huang, D. and Xu, Y. (2011). A new application of ensemble emd ameliorating the error from insufficient sampling rate. *Adv. Adapt. Data Anal.*, **3**: 493–508.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. L., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C. and Liu, H. H. (1998). The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proc. R. Soc. Lond. A*, **454**: 903–995.

Hung, J. J. (2000). Chi-Chi earthquake induced landslides in Taiwan. *Earthquake Eng. Eng. Seismol.*, **2**: 25–33.

Kuo, C. Y., Tai, Y. C., Bouchut, F., Mangeney, A., Pelanti, M., Chen, R. F. and Chang, K. J. (2009). Simulation of Tsaoling landslide, Taiwan, based on Saint Venant equations over general topography. *Eng. Geol.*, **104**: 181–189.

Lei, Y., He, Z. and Zi, Y. (2011). Application of the eemd method to rotor fault diagnosis of rotating machinery. *Mech. Syst. Signal Process.*, **23**: 1327–1338.

Mhamdi, F., Poggi, J.-M. and Jadane, M. (2011). Trend extraction for seasonal time series using ensemble empirical mode decomposition. *Adv. Adapt. Data Anal.*, **3**: 363–383.

Pierpont, N. (2009). *Wind Turbine Syndrome*: *A Report on a Natural Experiment*. K-Selected Books.

Wagner, S., Bareiss, R. and Guidati, G. (1996). *Wind Turbine Noise*. Springer, Berlin.

Wu, J. H. and Chen, C. H. (2011). Application of dda to simulate characteristics of the Tsaoling landslide. *Comp. Geotech.*, **38**: 741–750.

Wu, Z. and Huang, N. E. (2004). A study of the characteristics of white noise using the empirical mode decomposition method. *Proc. R. Soc. Lond. A*, **460**: 1597–1611.

Wu, Z. and Huang, N. E. (2009). Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Adv. Adapt. Data Anal.*, **1**: 1–41.

Wu, Z., Huang, N. E., Long, S. R. and Peng, C.-K. (2007). On the trend, detrending, and variability of nonlinear and nonstationary time series. *PNAS*, **104**: 14889–14894.

Yeh, J.-R., Shieh, J.-S. and Huang, N. E. (2010). Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method. *Adv. Adapt. Data Anal.*, **2**: 135–156.

Yu, D., Cheng, J. and Yang, Y. (2005). Application of emd method and Hilbert spectrum to the fault diagnosis of roller bearings. *Mech. Syst. Signal Process.*, **19**: 259–270.