New insights and best practices for the successful use of EMD, Iterative Filtering and derived algorithms

A. Cicone [1] A. Stallone [2] M. Materassi [3] H. Zhou [4]

[1]Università degli Studi dell'Insubria, Como, Italy

[2]Istituto Nazionale di Geofisica e Vulcanologia (INGV)

[3]Istituto dei Sistemi Complessi, CNR, Firenze, Italy

[4]Georgia Institute of Technology, Atlanta, USA

Contents

Contents

Motivation: The Empirical Mode Decomposition (EMD) and Iterative Filtering (IF) based algorithms could easily be misused if their known **limitations**, together with the **assumptions** they rely on, are not carefully considered.

We call attention to some of the pitfalls encountered when implementing these techniques. We have detected three critical factors that are often neglected:

1. Boundary effects
2. Presence of spikes/jumps in the original signal
3. Signals generated by processes characterized by a high degree of stochasticity

→ artefact-prone decomposition of the original signal

Contents

Problem: Decompose a nonlinear/non-stationary signal $s(x)$, $x \in \mathbb{R}$, into a few **simple components** that represent the features of the signal. We do not want to use any previous knowledge and assumptions on s.

Traditional methods, like Fourier or Wavelet Transform, require assumptions on the signal and do not work well with nonlinear and/or non-stationary signals.

EMD & IF methods

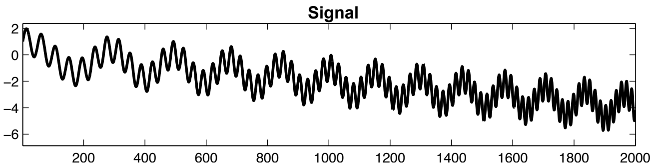
Advantages

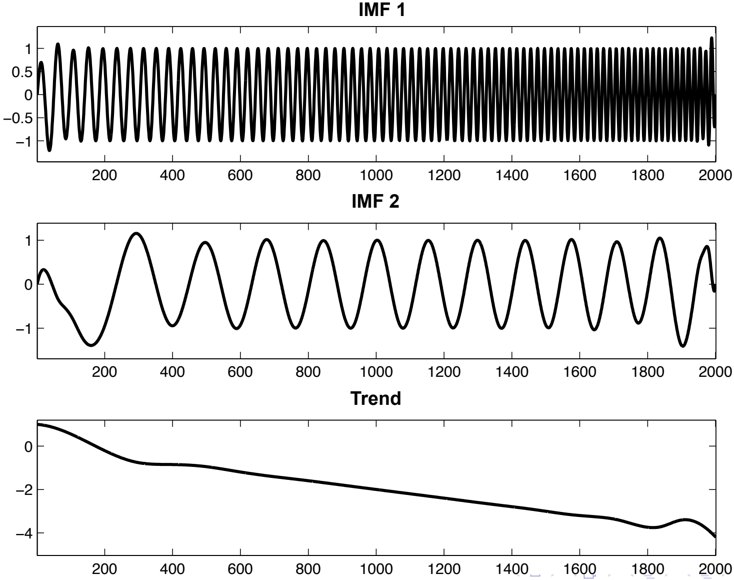
- Local and adaptive data-driven method
  $\longrightarrow$ can handle nonlinear and/or non-stationary signals
- Bypass partially the Heisenberg uncertainty principle
  $\longrightarrow$ accurate time-frequency representation of the signal

EMD & IF representation of a signal $x(t)$

$$x(t) = res(t) + \sum_{k=1}^{K} IMF_k(t)$$

Sifting process: given a signal $x(t)$, we capture its highest frequency oscillations by subtracting its moving average from the signal itself




IMF 1
1
0.5
0
−0.5
−1
200 400 600 800 1000 1200 1400 1600 1800 2000
IMF 2
1
0
−1
200 400 600 800 1000 1200 1400 1600 1800 2000
Trend
0
−2
−4
200 400 600 800 1000 1200 1400 1600 1800 2000

Contents

If not properly handled, **end effects** could arise:
Anomalously high amplitudes of the IMFs and artifact wave peaks towards the boundaries.

Problem1: For the EMD–based methods, it is particularly difficult to estimate a priori these errors contributions
Problem2: There are many alternative versions of EMD algorithms, each with its own peculiar way of handling boundaries
Problem3: IF–based techniques algorithms are also prone to end effects, even though there is a complete mathematical foundation

Solution

Pre–extend the signal using an a priori chosen *optimal extension*, based on the specific features of the signal under study

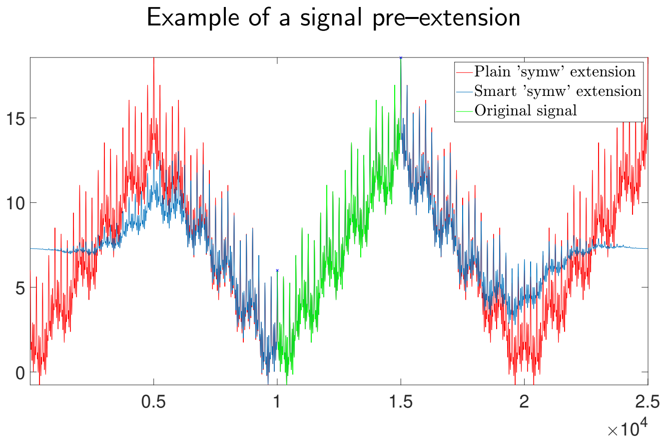
Signal Pre–extension Algorithm

1. Subtract from the signal s its mean value m
2. Extend $s - m$ outside the boundaries in the preferred or optimal way, producing an extended signal s_{ext}
3. Multiply s_{ext} by a characteristic function χ which has value one in the interval corresponding to the original signal s and goes smoothly to zero as we approach the new boundaries of s_{ext}
4. Add back the mean value m of the original signal

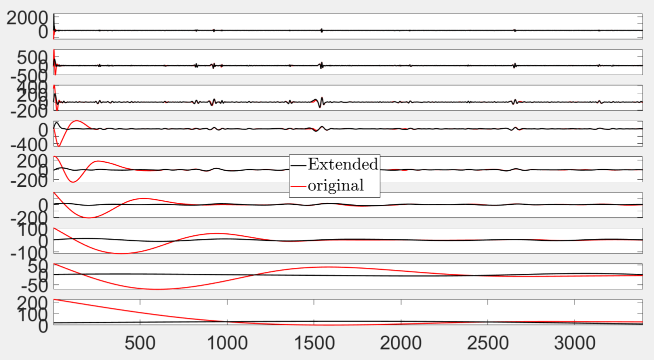
$$s_{\text{new}} = \chi \cdot s_{\text{ext}} + m$$

The produced signal s_{new} is periodical at the boundaries

Boundary conditions

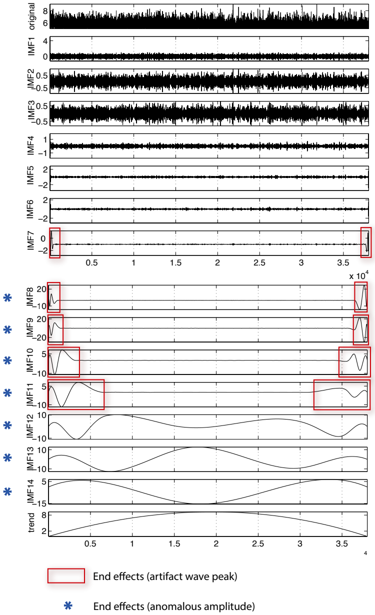


Artificial example: EMD decomposition of the signal without pre–extension (red) vs after pre–extension (black)

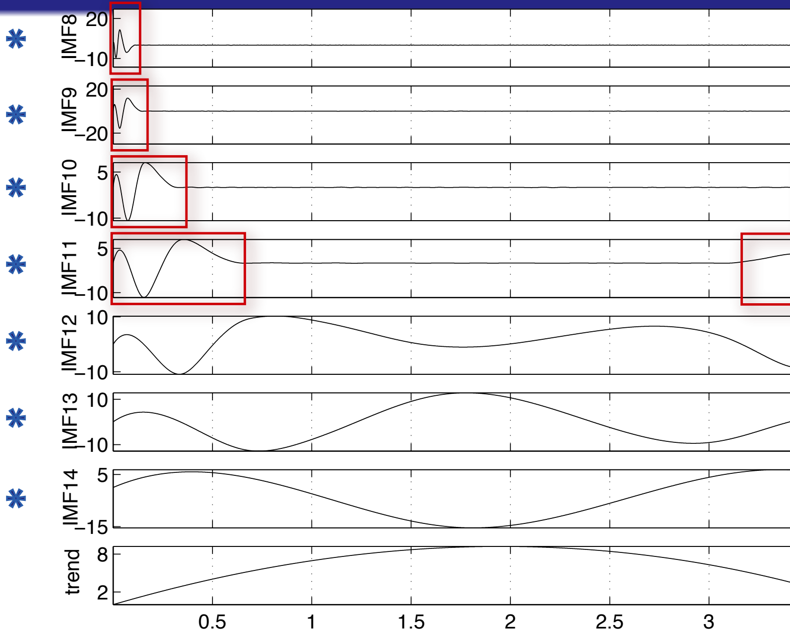


Real life example: from *Micro-scale, mid-scale, and macro-scale in global seismicity identified by empirical mode decomposition and their multifractal characteristics* by Sarlis, N. V. et al. *Scientific reports* 8, 9206 (2018)

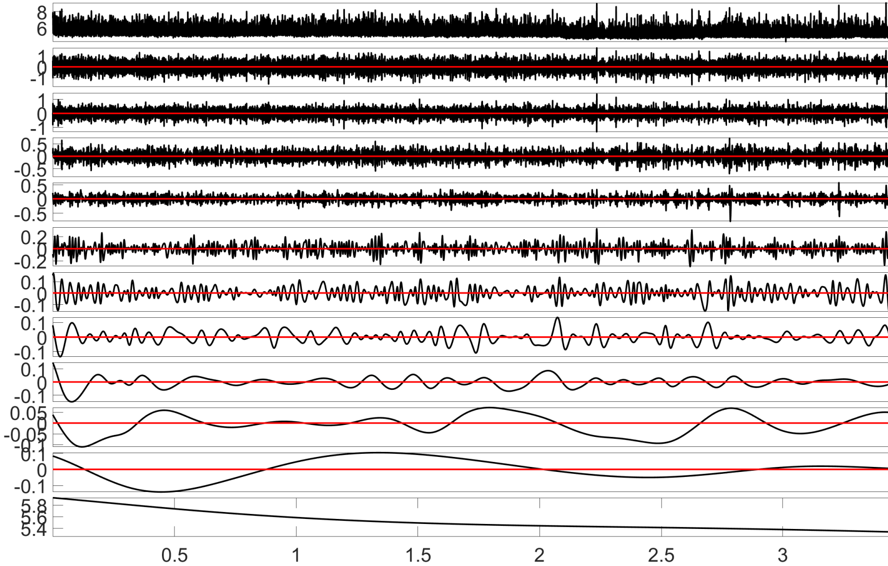
EMD decomposition of the magnitude time series of the global seismicity, for events of $M \geq 5.0$.



Boundary conditions



Boundary conditions



EMD decomposition after pre–extension

Contents

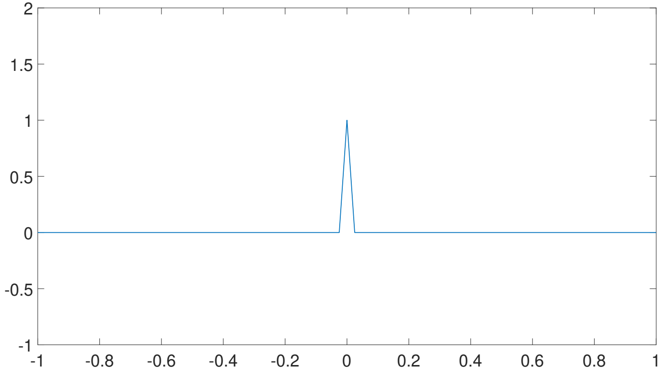
If spikes/jumps are present in the time series, they can substantially affect the signal decomposition

Problem: When the EMD and IF techniques are applied to an impulse, their decompositions introduce low frequencies at higher index IMFs $\rightarrow$ not meaningful from a physical point of view

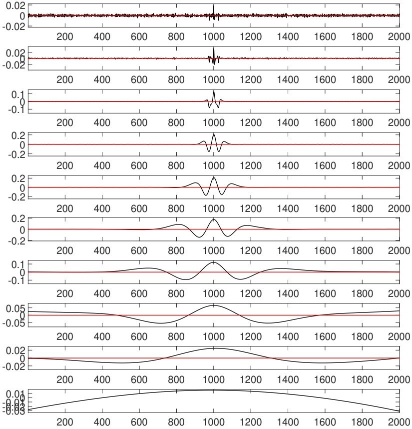
Solution

Always *pre-process* the signal, by removing spikes and splitting the signal in two subsets, before and after the jump

Artificial Example: Constant–amplitude signal with an impulsive spike

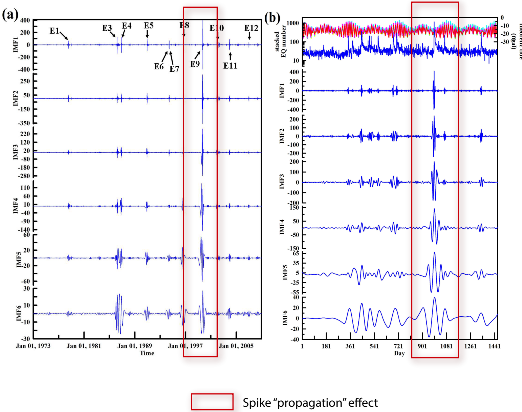


EMD decomposition of the impulsive spike

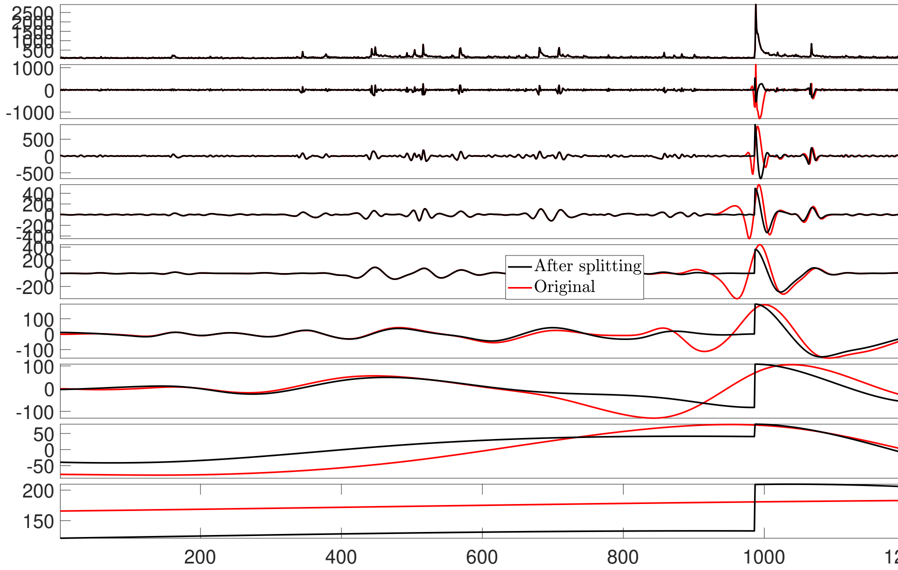


Real life example: from *Effect of tidal triggering on seismicity in taiwan revealed by the empirical mode decomposition method* by Chen, H. J. et al. *Nat. Hazards Earth Syst. Sci.* 12, 2193 (2012)

EMD decomposition of Taiwanese daily earthquakes number time series, for events of magnitude $M_L \geq 3.0$.



Spikes/jumps in the signal



EMD decomposition after jump preprocessing

Contents

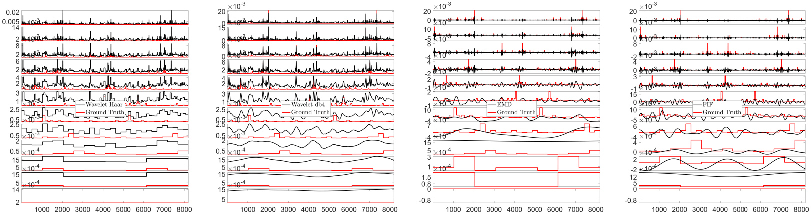
Problem: The EMD- and IF- based methods can be effectively applied to signals that can be described as oscillatory solutions of differential equations

Open question

Can these techniques successfully reproduce the features of signals containing a certain degree of stochasticity?

$\rightarrow$ We consider the multiscale statistical analysis of a synthetic signal.
We generate it by the **p-model**: branching model which mimics the irregular and intermittent distribution of energy in turbulent media

We model a signal $p(x)$ obtained from the summation of $n = 12$ generations $u_{k=0,\dots,12}$, each representing a different realization of a p–model $p(x) = \sum_{k=0}^{12} u_k(x)$



We decompose this signal via DWT, using both "Haar" and "Daubachies 4" (db4) bases, EMD and IF algorithms (from left to right, respectively)

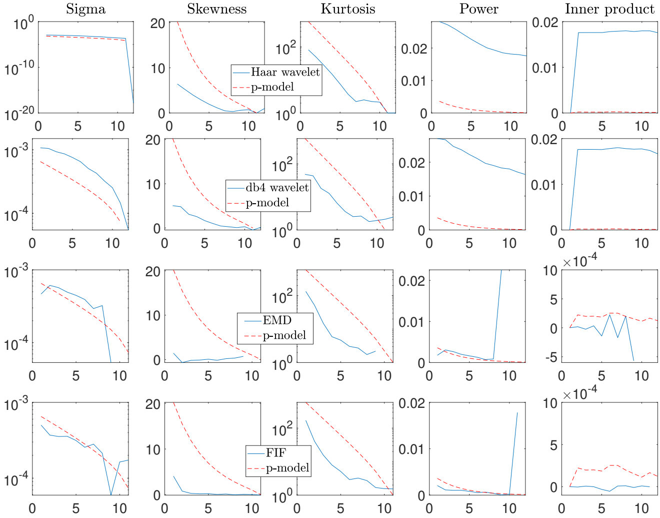
Problem

It is evident that **none of the aforementioned techniques is able to extract components which resemble the corresponding ground truth** u_k generations
These techniques will always produce a decomposition of the signal, no matter what the process behind it is

We now test the ability of all these techniques in reconstructing *multiscale statistical features* of the given signal
In particular, we consider:

- Standard deviation $\sigma\left(p_h(x)\right)$
- Skewness $S\left(p_h(x)\right)$
- Excess of kurtosis $K\left(p_h(x)\right)$
- Total energy of the h-th scale $\mathcal{E}\left(p_h(x)\right)$
- Inner product between two nearby scales $\mathcal{C}\left(p_h(x)\right)$

Decomposition of highly-stochastic signals



Discussion

- EMD and IF decompositions prove to be more accurate in replicating the standard deviation and power of the different components
- The skewness values, as expected, are close to zero. Both EMD and IF enforce symmetry with respect to the horizontal axis
- The inner products tend to zero with EMD and IF methods, since they produce IMF components which are locally almost orthogonal each other
- Regarding the kurtosis, all the techniques are all able to properly reproduce its trend as a function of the scale

$\rightarrow$ *All* the multiscale analysis tools compared in this example are able to detect the intermittency of the signal under analysis

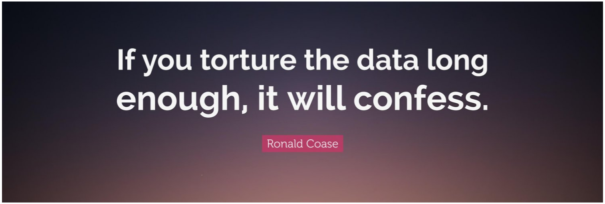
Contents

Take Home Messages

- The impact of boundary errors can always be reduced by properly **pre–extending the signal** under study. For this reason, we propose a new approach for the pre–extension of a given signal
- When dealing with a signal containing spikes, the optimal solution would be to study its decomposition **before and after removing the spikes** from the signal itself
- When a jump is present in the data set, a good practice would be to split it into *before* and *after* the jump, and to **analyse the two portions separately**

Take Home Messages - cont'ed

- **Although the derived decomposition is always correct from a mathematical stand point, it may be the case that there is not a corresponding evident physical meaning of each IMF**, as for the case of stochastic signals
- Nevertheless, the EMD– and IF–based methods proved to have good performance from a **multiscale statistical analysis** prospective
- It remains, an open problem to understand **up to which degree of stochasticity these techniques are able to reproduce with a good accuracy the single components** contained in a given signal


If you torture the data long enough, it will confess.
Ronald Coase