Introduction to EMD (Empirical Mode Decomposition) with application to scientific data

Dongho Kim

May 27, 2006
1. Overview
2. Introduction to EMD
3. Applications to Statistics
4. Future Researches
Overview

1. A signal with different scales of frequencies
   - For example, sun-spot data is fluctuated over about 11 year and 85 year.
   - Economic data is supposed to be compound of seasonal component, cyclic component and long-term trend.

2. Fourier series analysis, Wavelet, Empirical Mode Decomposition (Huang 1998)

3. Empirical Mode Decomposition
   - Fast oscillations superimposed to slow oscillations
   - Data-driven, locally adaptive, multiscale, robust to non-stationary
   - Many applications to speech analysis (Liu 2005), biological data (Huang 1998, 2002), earthquake (Zhang et. al 2003), climate (Coughlin and Tung 2004)
What is Frequency?

1) Oscillating and periodic patterns are repeated
2) Local mean is zero and the signal is symmetric to its local mean
3) One cycle of oscillation: sinusoidal function starting at 0 and ending at 0 with passing through zero between two zero crossings. Or starting at local maximum and terminating at consecutive local maximum with passing through two zeros and local minimum.
Huang et. al (1998) defined an oscillating wave as Implicit Mode Function (IMF) if it satisfies two conditions

1) the number of extrema and the number of zero crossing differs only by one and

2) the local average is zero. The condition that the local average is zero implies that envelop mean of upper envelop and lower envelop is zero.
A signal observed in real world consist of low and high frequencies.

Suppose we observe signal $x(t)$ which is of the form

$$x(t) = 0.5t + \sin(\pi t) + \sin(2\pi t) + \sin(6\pi t) + \epsilon,$$

where $\epsilon$ is the noise.

For illustrative purpose, ignore noise this time. The signal consist of 4 components from the highest frequency $\sin(6\pi t)$ to the lowest frequency $0.5t$. 
1-1. First identify the local extrema.
1-2. Consider the two functions interpolated by local maximum and local minimum (upper and lower envelop).
1-3. Their average, envelop mean will yields the lower frequency component than the original signal.
1-4. By subtracting envelop mean, from the original signal $x(t)$, the highly oscillated pattern $h$ is separated.
One iteration of above procedure does not guarantee that the resulting residue signal $h$ is IMF. The same procedure is applied to the residue signal $h$ until properties of IMF is satisfied.

This iterative algorithm is called sifting.

Sifting makes the remaining signal more symmetric, local mean toward to zero, all the maximum positive and all the minimum negative.

In other word, sifting makes envelop mean to bisect the signal evenly so that there are no overshoots nor undershoots.

It is known that cubic spline is optimal for interpolating.
Introduction to EMD

Sifting – Extracting Implicit Mode Function (IMF)
Introduction to EMD
Sifting – Extracting Implicit Mode Function (IMF)

- Note that as the name sifting implies, the lower frequency component is repeatedly removed from the highest frequency.
- The first IMF $imf_1$ produced by sifting is the highest frequency by its construction.
- Residue signal $r$ less oscillated than the original signal. Remaining signal $r = x - imf_1$ still may be compound of several frequencies.
- The same procedure is applied on the residue signal $r$ to obtain the next IMF.
- By the construction, the number of extrema will eventually decreased as the procedure continues so that a signal is sequently decomposed into the highest frequency component $imf_1$ to the lowest frequency component $imf_n$, for some finite $n$ and residue $r$.
- Finally we have $n$ empirical mode and residue as

$$x(t) = \sum_{i=1}^{n} imf_i(t) + r(t).$$
Introduction to EMD
Sifting – Extracting Implicit Mode Function (IMF)

First two IMF’s

\[ \text{Signal} = 1\text{-st IMF} + 1\text{-st residue} \]

\[ 1\text{-st residue} = 2\text{-nd IMF} + 2\text{-nd residue} \]
Introduction to EMD
Sifting – Extracting Implicit Mode Function (IMF)

Decomposition Result

1-st IMF

2-nd IMF

3-rd IMF

residue
Introduction to EMD
EMD Algorithm

a. Take input signal $r_{k-1}$ to decompose. $r_0$ is the original signal $x$.
   1. Identify the local extrema of the signal $r_{k-1}$.
   2. Construct upper envelop $emax_k$ and lower envelop $emin_k$ interpolating maximum and minimum, respectively.
   3. Approximate local average by envelop mean $em_k$ taking average of two envelops $emax_k$ and $emin_k$. That is $em_k = (emax_k + emin_k)/2$.
   4. Compute candidate implicit mode $h_{k1} = r_{k-1} - em_k$.
   5. If $h_{ki}$ is IMF, decompose the signal $r_{k-1}$ as IMF $imf_k = h_{ik}$ and the residue signal $r_k = r_{k-1} - imf_k$. Otherwise repeat the step 1 through 5.

b. If $r_k$ has implicit oscillation mode, set $r_k$ as input signal and repeat step a.
- Solar irradiance variation is supposed to be substantial effect on both high and low frequency variations in climate.
- Observed sunspot series: Satellites measure solar irradiance directly in the recent decades. Represent 11 year solar mode relatively well.
- Proxy-based reconstructions: For the lower frequent component, proxy-based reconstructions (Lean et. al, 1995) are widely used.
- Is this proxy-based reconstruction proper one?
IMF 2 and IMF 6 of Lean’s reconstruction represent two solar mode well.
IMF 2 matches the sunspot data (11 years solar mode) with respect to frequency and timing.
IMF 6 corresponds to lower frequent component known as 85 years solar mode.
What happen if we apply EMD to noisy signal?

Example noisy signal

\[ x(t) = f(t) + \epsilon(t), \quad t = 1, 2, \ldots, n, \]

where \(\{\epsilon(t)\}\) is error with mean 0 and variance \(\sigma^2\) and \(f\) is

\[
    f(t) = \begin{cases} 
    \sin(\pi t) + \sin(2\pi t) + \sin(6\pi t) & 0 < t \leq 3, \\
    \sin(\pi t) + \sin(6\pi t) & 3 < t \leq 6, \\
    \sin(\pi t) + \sin(6\pi t) + \sin(12\pi t) & 6 < t \leq 9. 
    \end{cases}
\]
EMD result for noisy signal $x(t)$: $imf_1$ and $imf_2$ are noise?

Residue Signal $r(t) = x(t) - imf_1(t) - imf_2(t)$
How to remove the noise efficiently? One of solution is thresholding.

\[ d_i(t) = \text{imf}_i(t) \times I(|\text{imf}_i(t)| > \lambda) \]
If $\lambda_1 = 0.8107$ and $\lambda_2 = 0.3878$, $\hat{f}(t) = d_1(t) + d_2(t) + r(t)$.

Thresholded IMF, $d_1$

Thresholded IMF, $d_2$

Denoised Signal $\hat{f} = d_1 + d_2 + r$
Cross-validation for Threshold.

1. Calculate the prediction error $PE$

$$PE(\lambda_1, \ldots, \lambda_s) = \frac{1}{n} \sum_{t=1}^{n} \{x(t) - \hat{f}_{\lambda_1,\ldots,\lambda_s}(t)\}^2,$$

for given a threshold $\lambda_i$ for the $i$th IMF $i = 1, \ldots, s$.

1-1. Apply the EMD for the data sets expelling $t$th data point, and then obtain IMF $imf_i, i = 1, \ldots, s$ and residue signal $r$.

1-2. Given a threshold $\lambda_i$ for the $i$th IMF $i = 1, \ldots, s$, construct denoised IMF’s $d_i, i = 1, \ldots, s$ and $\hat{f} = d_1 + \ldots + d_s + r$.

1-3. Calculate the predicted estimate $\hat{f}_{\lambda_1,\ldots,\lambda_s}(t)$ by interpolating $\hat{f}$’s nearest to $t$.

1-4. Repeat step 1-1 through 1-3 for all $t$ and calculate $PE$.

2. By grid search algorithm, find $\hat{\lambda}_1, \ldots, \hat{\lambda}_s$ to minimize prediction error $PE$. 
Applications
Smoothing

Comparison with wavelet methods: \(MSE = 43.43, 234.08, 61.29, 63.46.\)
1. Comparison with three wavelet method
   1. \(wc\): Nason’s Two-fold Cross-validation
   2. \(ws\): SureThresh of Donoho and Johnstone (1995)
   3. \(wu\): Universal threshold of Donoho and Johnstone (1994)
   4. \(emd.cv\): Proposed method. Threshold \(imf_1\) and \(imf_2\) based on 2-fold CV.

2. Test functions

   \[
   f_1(t) = \begin{cases} 
   \sin(\pi t) + \sin(2\pi t) + \sin(6\pi t) & 0 < t \leq 3, \\
   \sin(\pi t) + \sin(6\pi t) & 3 < t \leq 6, \\
   \sin(\pi t) + \sin(6\pi t) + \sin(12\pi t) & 6 < t \leq 9, 
   \end{cases}
   \]

   \[
   f_2(t) = \exp(-0.01t) \cos(\pi t/10), \quad 0 < t \leq 500.
   \]

3. 1,024 observations by noise \(\sim N(0, \sigma_{snr=3})\), \(snr=\|f\|/\sigma = 3\)

4. MSE based on 100 replications.
Test Functions

\begin{align*}
  f_1 & \\
  f_2 &
\end{align*}
Boxplot of MSE: The proposed procedure outperformed.
Future works
Future Researches

- Frequency analysis
- Forecasting
- 2-dimensional signal: Image
References


