Some improvements of Hilbert-Huang transform for time-frequency analysis of gravitational waves

Ken-ichi Oohara
The Open University of Japan and Niigata University

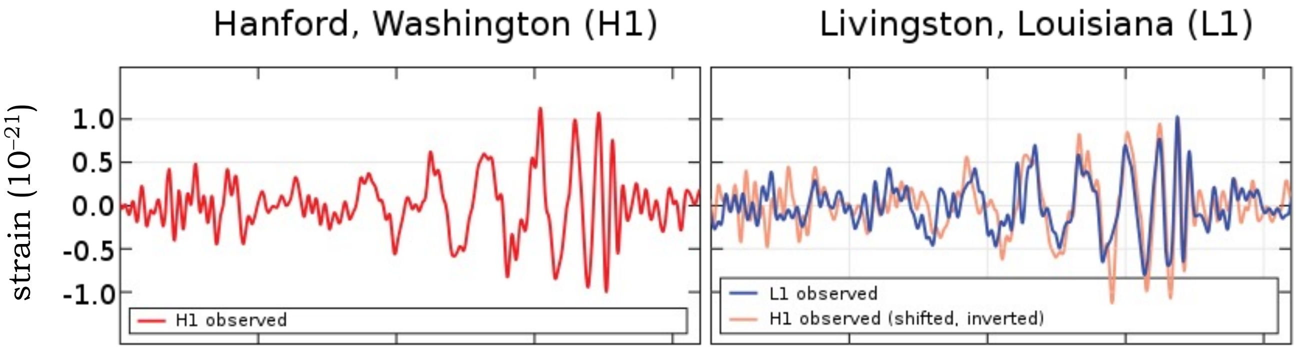
collaborating with

H. Takahashi, K. Sakai, Y. Hiranuma,
M. Kaneyama, M. Takeda and I. Yoda

TAUP2023@Vienna

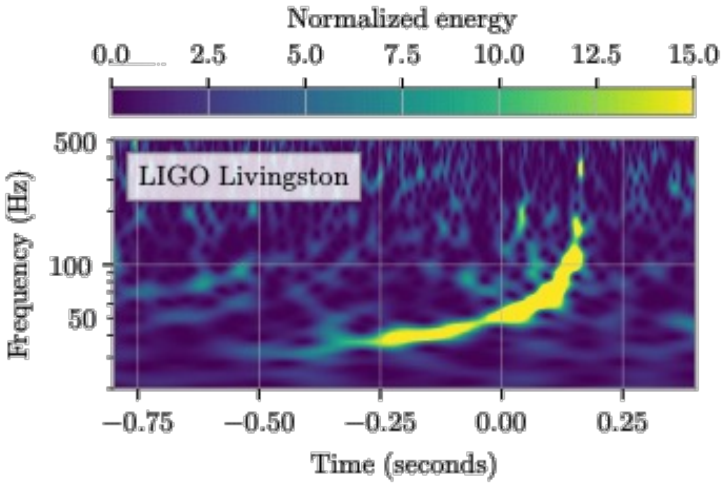
Introduction

- Many GW events from compact binary coalescence (CBC) comprised of black holes and neutron stars, while not from supernovae yet.



- Investigating temporal changes in the frequency and amplitude of GWs is important for studying the physics of GW sources.

- There are several different ways of T-F analysis
  - ➢ the short-time Fourier transform (STFT)
  - ➢ the wavelet analysis
  - ➢ Wigner distribution function (WDF) and its modifications
  - ➢ Non-harmonic analysis(NHA)
    - The resolutions in time and frequency are restricted by "**the uncertainty principle**" in methods based on Fourier transform.



spectrogram (time-frequency map)

- ➢ **Hilbert-Huang transform (HHT)** proposed by Norden E. Huang (1996)
  - HHT is not affected by the uncertainty principle of FT.
  - It is adaptive approach to time series analysis.
  - It consists of
    - ⋆ an empirical mode decomposition (EMD),
    - ⋆ the Hilbert spectral analysis (HSA).

Demodulation

- If a signal $h(t)$ can be divided into modulator $a(t)$ and carrier $c(t)$ or $\cos\phi(t)$:

$$h(t) = a(t)c(t) = a(t)\cos\phi(t)$$

  - $a(t)$: the time-varying amplitude or **the instantaneous amplitude (IA)**
  - $\phi(t)$: the instantaneous phase
  - **the instantaneous frequency (IF)** $f(t) = \frac{1}{2\pi}\frac{\mathrm{d}\phi(t)}{\mathrm{d}t}$

- The decomposition is **not unique**.

  ➡ For **a complex signal**, there is a reasonable way to define the IA and IF.

- $F(t)$: a complex signal $F(t) = h(t) + iv(t) = a(t)e^{i\phi(t)}$
  - $a(t) = |F(t)| = \sqrt{h(t)^2 + v(t)^2}$: IA
  - $\phi(t) = \mathrm{Arg}[F(t)] = \tan^{-1}\left[\frac{v(t)}{h(t)}\right]$; $f(t) = \frac{1}{2\pi}\frac{\mathrm{d}\phi(t)}{\mathrm{d}t}$: IF

Hilbert Spectral Analysis (HSA)

- How to find the complex signal $F(t)$ from a real signal $h(t)$.
- If $h(t)$ is the real part on the real axis of a holomorphic function $F(z)$ which approaches zero fast enough for $|z| \to \infty$, its imaginary part $v(t)$ is uniquely given by the Hilbert transform of $h(t)$.
- **Hilbert Transform**

$$v(t) = \mathcal{H}[h](t) \equiv \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{h(t')}{t-t'} dt' = h(t) * \left(\frac{1}{\pi t}\right)$$

P : the Cauchy's principal value,   * : the convolution

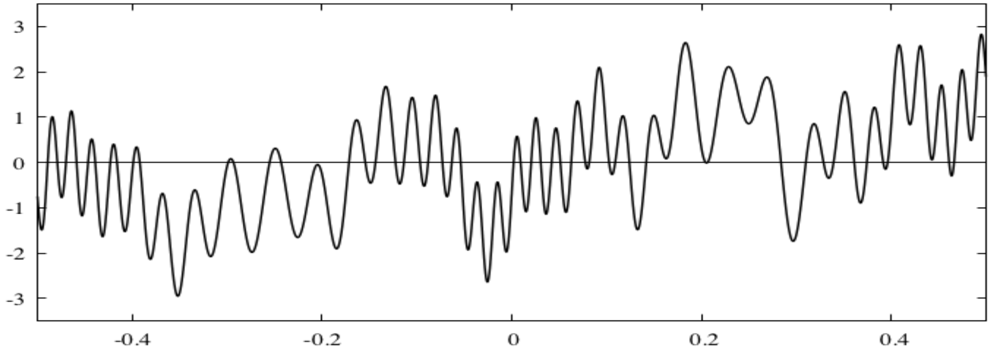
- The imaginary part $v(t)$ and therefore IF and IA are obtained as long as the integral converges.
- The IF is not always physically meaningful.
  - a cosine wave of of constant amplitude and frequency

$$h(t) = a\cos\omega t + b, \qquad \mathcal{H}[h](t) = a\sin\omega t$$

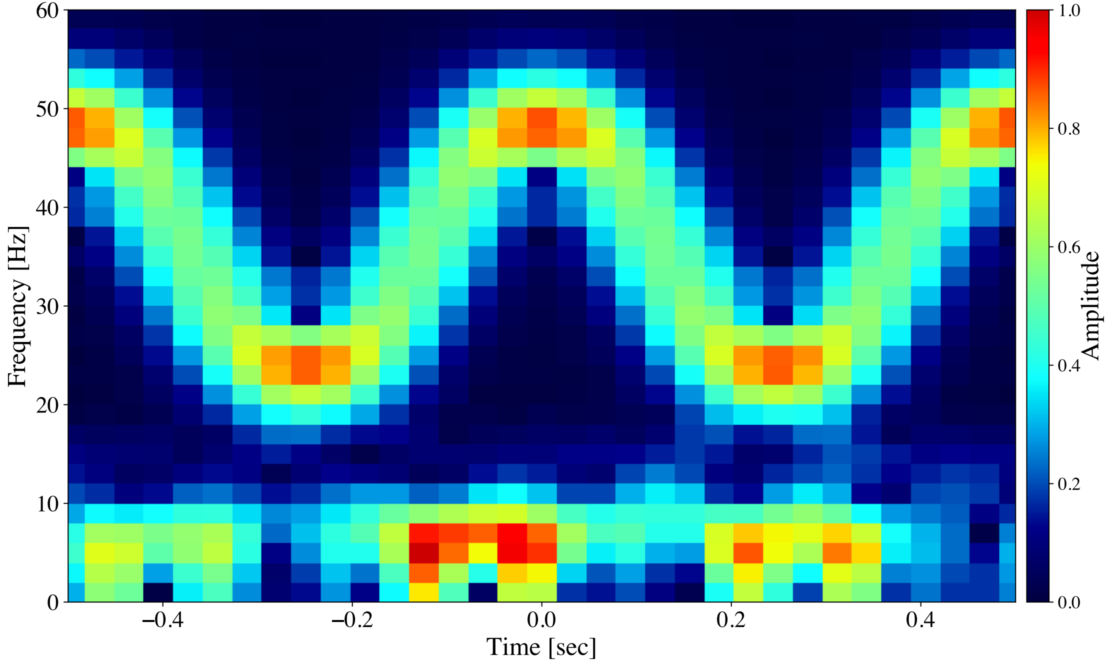
$$F(t) = h(t) + i\mathcal{H}[h](t) = ae^{i\omega t} + b = A(t)e^{i\phi(t)}$$

IA: $A(t) = \sqrt{a^2 + b^2 + 2ab\cos\omega t}$ IF: $f(t) = \dfrac{1}{2\pi}\dfrac{d\phi(t)}{dt} = \dfrac{\omega}{2\pi}\dfrac{a(a+b\cos\omega t)}{a^2+b^2+2ab\cos\omega t}$

signal

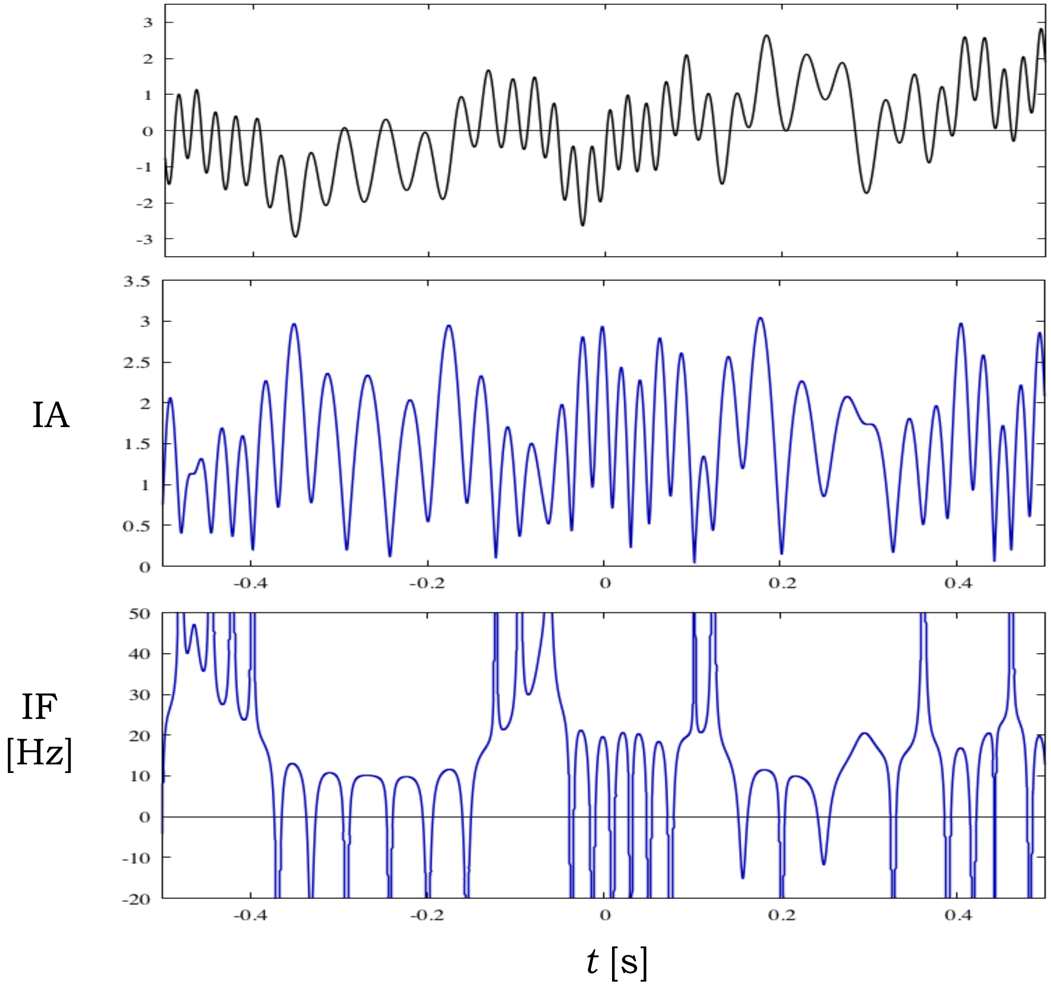


STFT-based Spectrogram
(Time-Frequency map)



signal

Hilbert Spectral Analysis



Decomposition of Signals

- To overcome this issue, we need to decompose signal $h(t)$ into some waves $c_k(t)$ called **intrinsic mode functions (IMFs)** and the non-wave part $r(t)$;

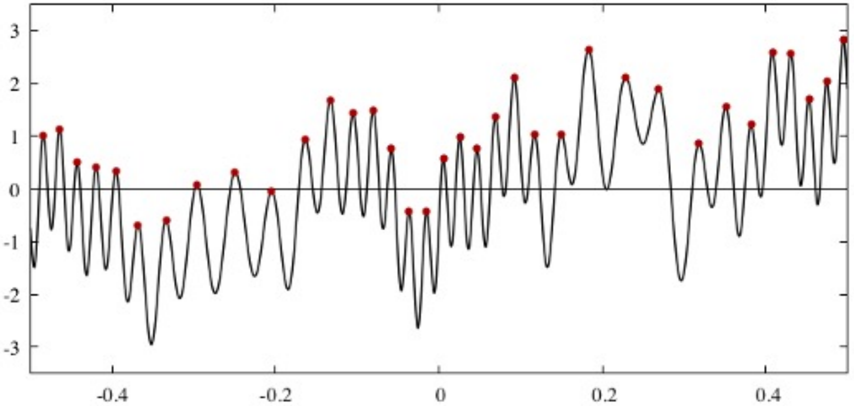
$$h(t) = \sum_k c_k(t) + r(t) = \sum a_k(t) \cos \phi_k(t) + r(t)$$

$c_k(t)$: IMF, $\quad r(t)$: the trend (non-wave part)

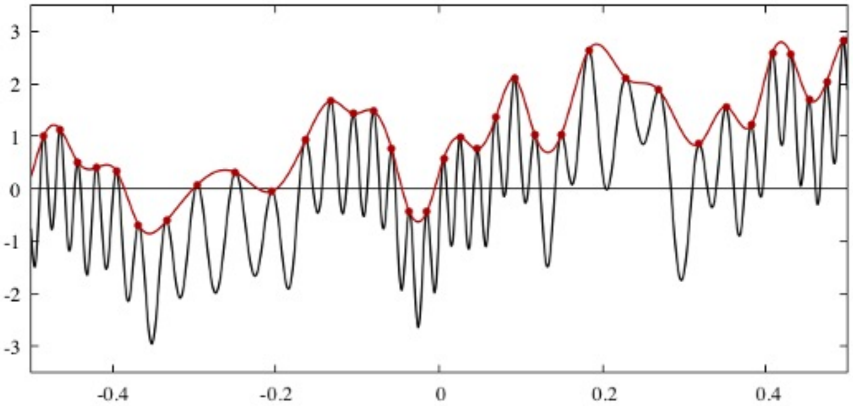
- **Each IMF must satisfy the following conditions** to obtain meaningful IF and IA using the HT:
  - **oscillating around zero**;
    in the whole data set, |# of extrema – # of zeros| = 0 or 1
  - **locally symmetric wrt zero**;
    the mean value of the upper and lower envelopes defined by the local maxima and minima = 0
- **Empirical Mode Decomposition (EMD)**:
  a sift procedure for decomposing a signal into IMFs.

Empirical Mode Decomposition (EMD)

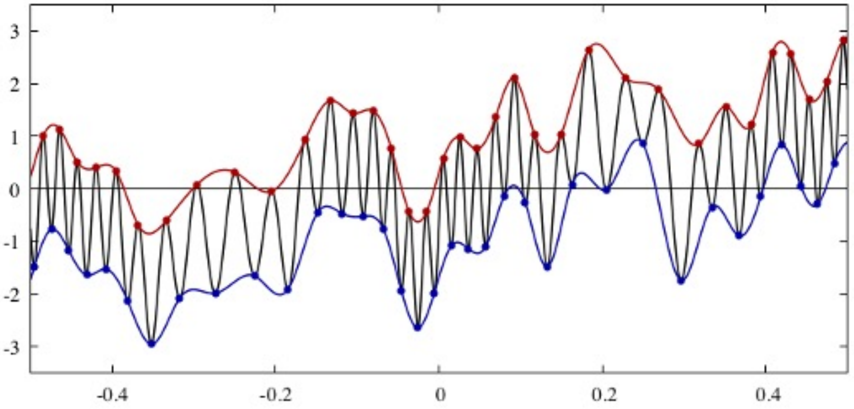
- Set $h_1(t) = h(t)$ (the original signal)
- for $i = 1$ to $i_{\max}$
  - ➢ $h_{i1}(t) = h_i(t)$
  - ➢ for $k = 1$ to $k_{\max}$
    1) Mark the local maxima and minima of $h_{ik}(t)$.
    2) Interpolate the maxima and minima by cubic splines
       $\longrightarrow$ the upper $U_{ik}(t)$ and lower $L_{ik}(t)$ envelopes.
    3) $m_{ik}(t) = (U_{ik}(t) + L_{ik}(t))/2$.
    4) $h_{i,k+1}(t) = h_{ik}(t) - m_{ik}(t)$.
  - ➢ Exit if a certain stoppage criterion is satisfied.
  - ➢ IMF i is obtained; $c_i(t) = h_{ik}(t)$.
  - ➢ Set $h_{i+1}(t) = h_i(t) - c_i(t)$.
- Set the final residual (the trend) $r(t) = h_{i\max+1}(t)$.



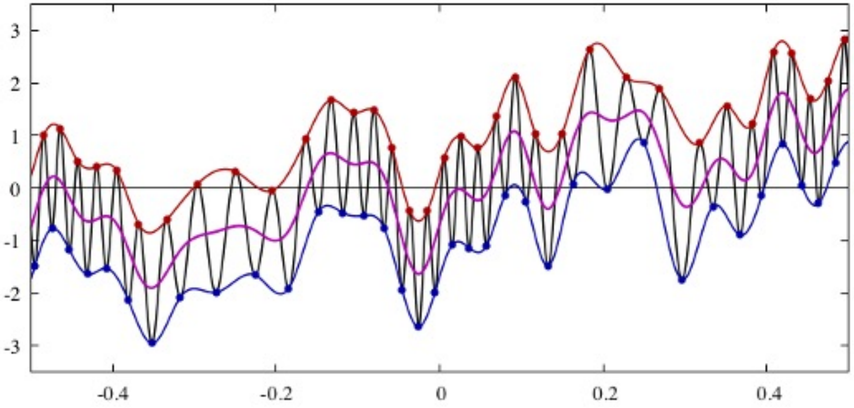
Mark the maxima of the original signal $h(t)$.



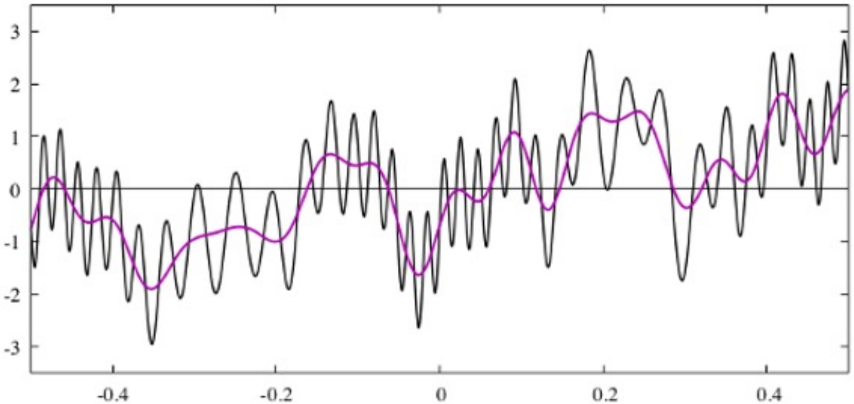
Interpolate the maxima to obtain the upper envelope $U_1(t)$ usually using **cubic spline**.



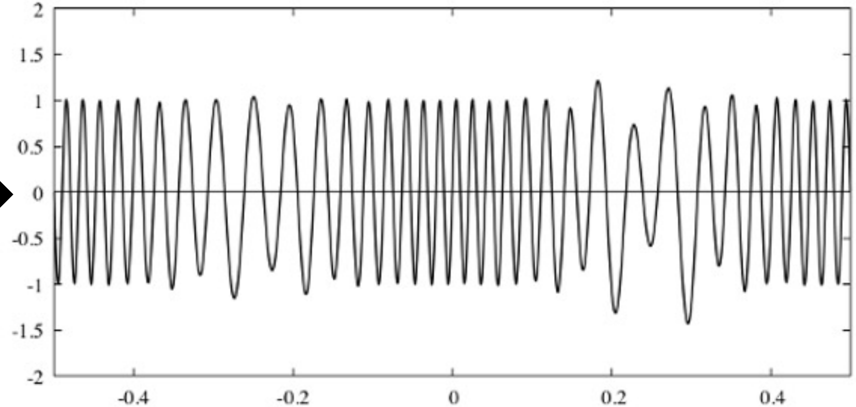
Follow the same procedure to obtain the lower envelope $L_1(t)$.



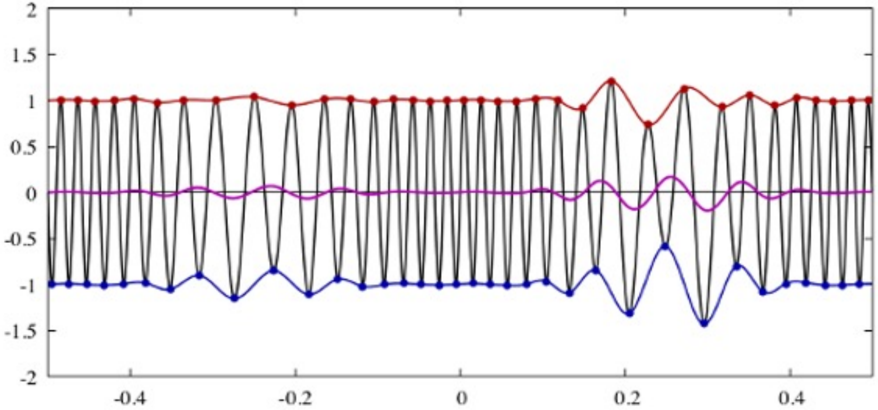
Calculate the local mean curve $m_1(t) = (U_1(t)+L_1(t))/2$.



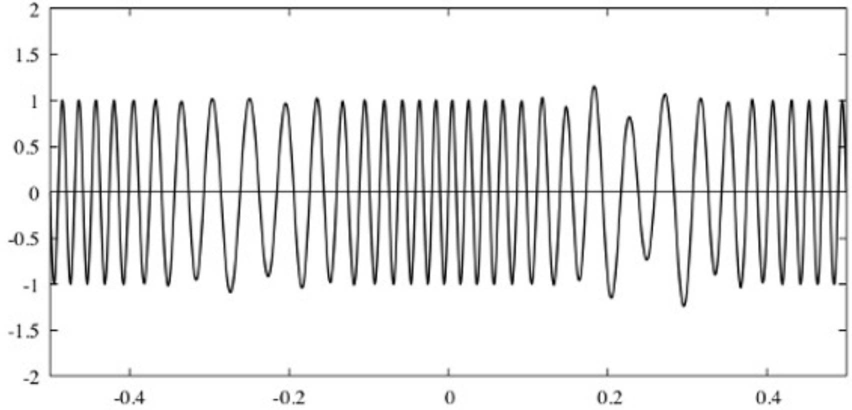
Subtract the mean $m_1(t)$ from the original signal $h(t)$



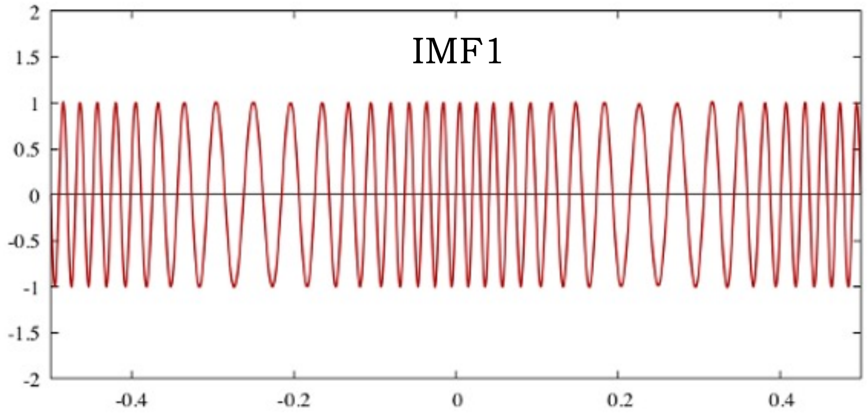
to obtain the residual $h_{12}(t) = h(t) - m_1(t)$.



Repeat the procedure to obtain $m_2(t) = (U_2(t)+L_2(t))/2$



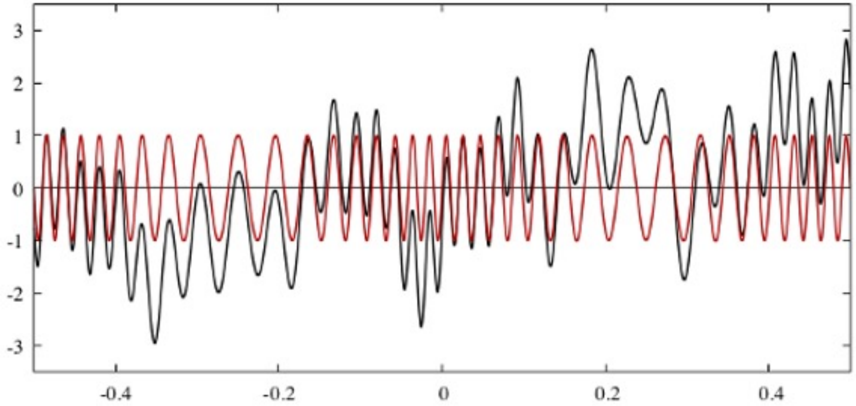
and subtract $m_2(t)$ from $h_{12}(t)$ to obtain $h_{13}(t)$.



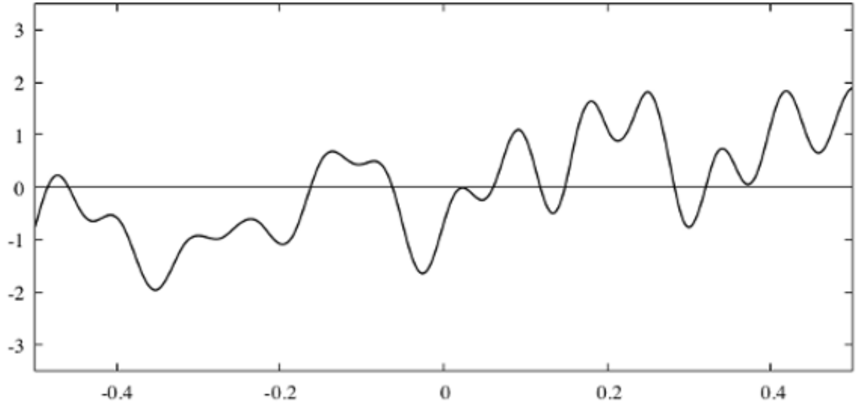
Repeat the procedure until a stoppage criterion is satisfied to adopt $h_{1k}(t)$ as $c_1(t)$ (IMF1).

stoppage criterion: $m_k(t)$ is sufficiently small

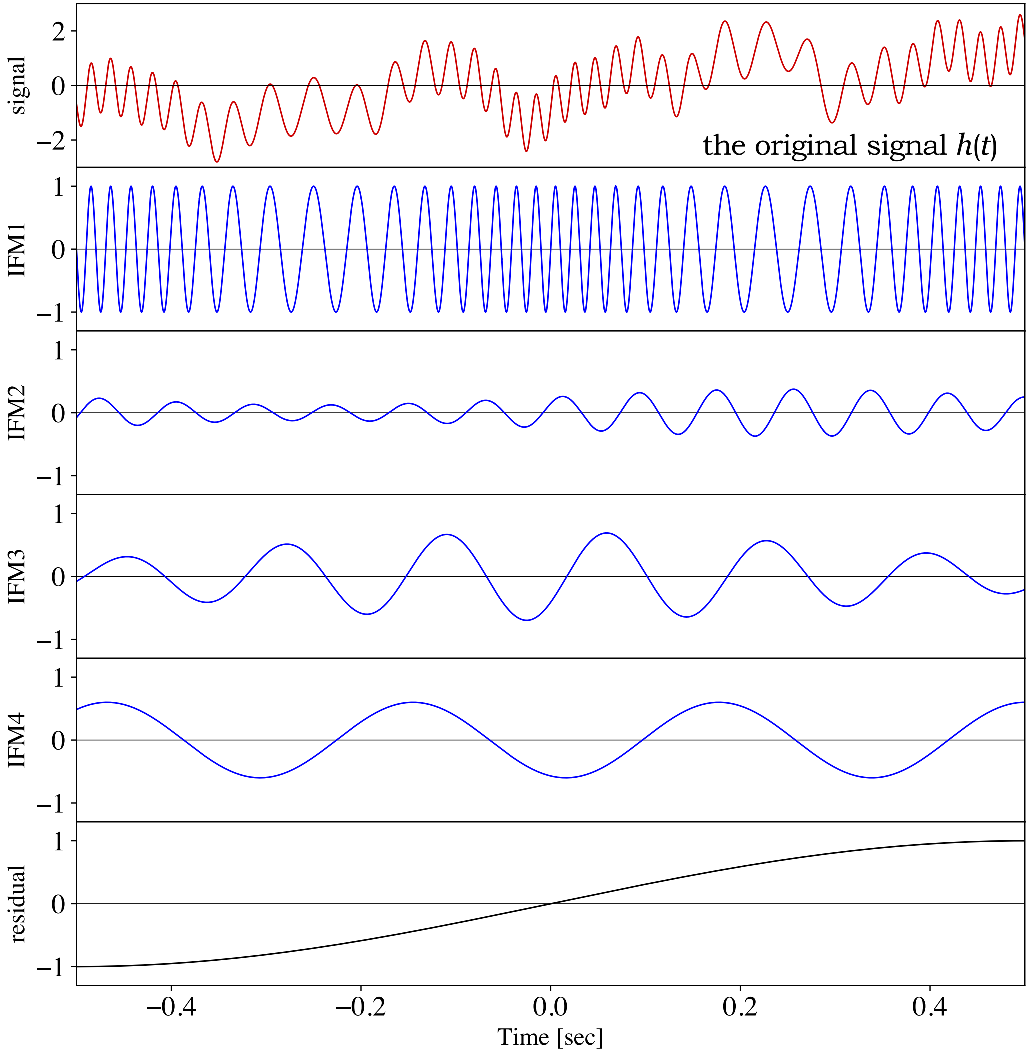
$$\|m_k(t)\| < \varepsilon \|h_{1k}(t)\|$$



Subtract IMF1 $c_1(t)$ from the original signal $h(t)$



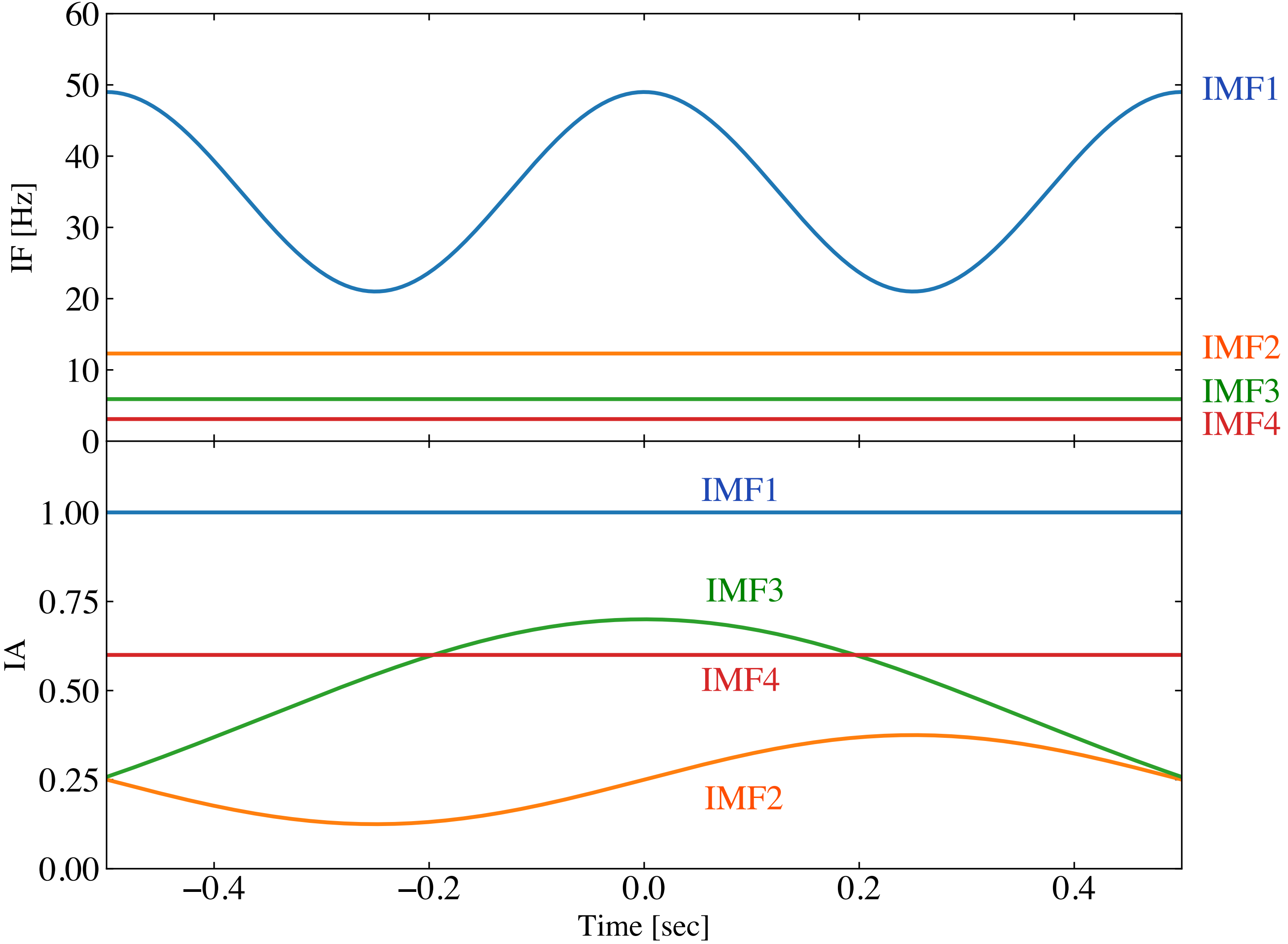
and apply the sifting process on the residual again to obtain IMF2, IMF3 ....



The sifting is completed when the residual has at most one extremum or is smaller than the predetermined value.

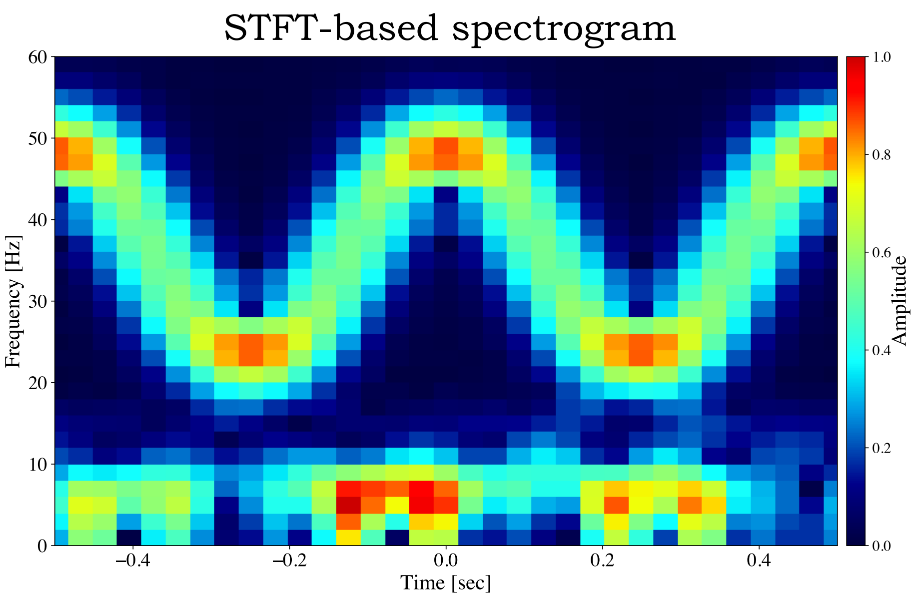
Finally, the original signal is decomposed in terms of IMFs.

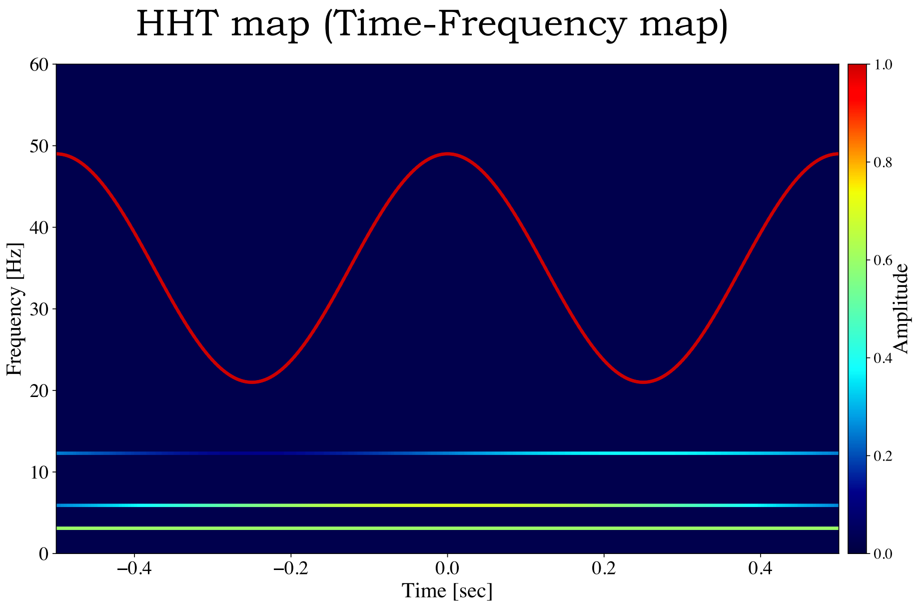
$$h(t) = \sum_{k=1}^{n} c_k(t) + r(t)$$



IF [Hz]
IA
Time [sec]
IMF1
IMF2
IMF3
IMF4
IMF1
IMF3
IMF4
IMF2
0
10
20
30
40
50
60
0.00
0.25
0.50
0.75
1.00
−0.4
−0.2
0.0
0.2
0.4

Spectrogram vs HHT map





- Both of a spectrogram and an HHT map are graph with two geometric dimension.
- A spectrogram, in principle, fills the entire area of 2D graph.
- An HHT map consists of several curves representing IFs $f_k(t)$ with color gradients representing IAs $a_k(t)$ of IMFs.

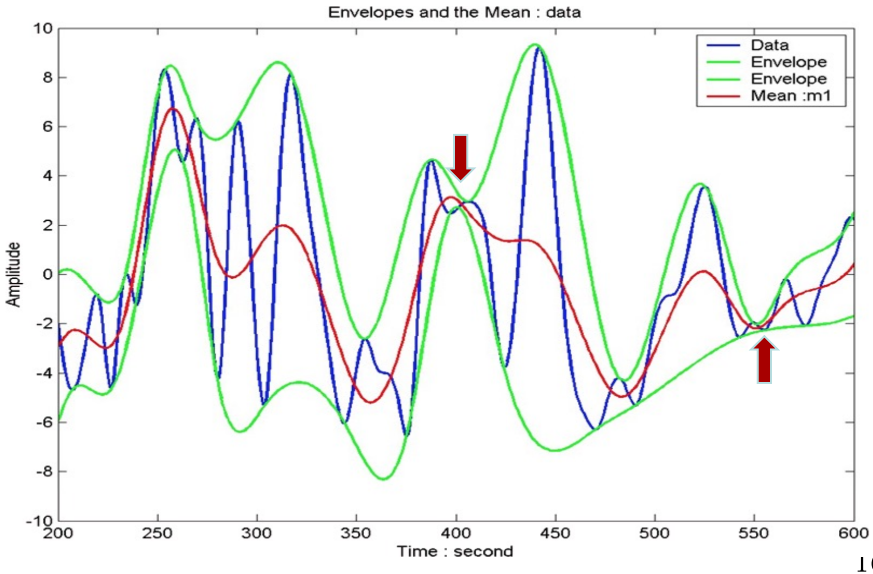
Problems with EMD

- The original EMD is sensitive to noise.
- In the original form of the EMD, mode mixing and/or splitting frequently appears.
  - ➢ **mode mixing / mode splitting**
    - A single IMF consists of signals of widely disparate scale.
    - Signals of a similar scale reside in different IMF components.

➡ serious aliasing in the time-frequency distribution

➡ not physically meaningful IMF

- Mode mixing often occurs if envelopes are close together at height away from zero.



Ensemble EMD

- **Solution: Ensemble EMD (EEMD)**
  - Proposed by Huang et al. (2009)
  - Inspired by the study of white noise using EMD

- **Algorithm:**
  1) Add white noise to the original data to form a "trial", $h_i(t) = h(t) + n_i(t)$.
  2) Perform EMD on each $h_i(t)$ with different $n_i(t)$.
  3) For each IMF, take ensemble mean among the trials (i = 1, 2, …) to obtain the final answer.

Completeness of decomposition

- The original EMD preserves completeness;
  the sum of the IMFs and residual recovers the original signal.

$$h(t) = \sum_k c_k(t) + r(t)$$

- In the EEMD, the residual noise destroys the completeness,

$$h(t) + n_i(t) = \sum_k c_{i,k}(t) + r_i(t)$$

$$c_k(t) = \frac{1}{N} \sum_{i=1}^{N} c_{i,k} = \langle c_{i,k}(t) \rangle; \quad r(t) = \langle r_i(t) \rangle$$

$$\Rightarrow h(t) + \langle n_i(t) \rangle = \sum_k c_k(t) + r(t)$$

  while the residue noise could be reduced with large enough ensemble.

Complementary EEMD (CEEMD)

To resolve the problem, J.-R. Yeh and J.-S. Shieh (2010) proposed Complementary EEMD (CEEMD).

- If trials with noise $-n_i(t) \equiv n_{-i}(t)$ are added to the ensemble,

$$h(t) + n_i(t) = \sum_k c_{i,k}(t) + r_i(t)$$

$$h(t) - n_i(t) = h(t) + n_{-i}(t) = \sum_k c_{-i,k}(t) + r_{-i}(t)$$

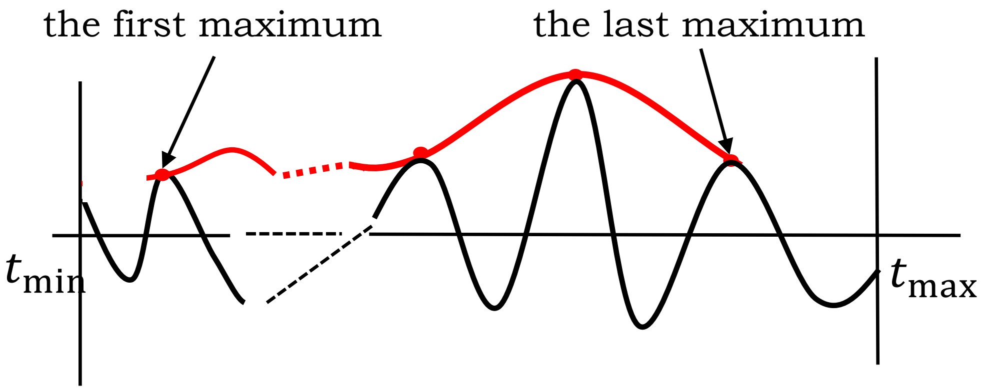
$$c_k(t) = \frac{1}{2N} \sum_{i=1}^{N} (c_{i,k} + c_{-i,k}) = \langle c_{i,k}(t) \rangle$$

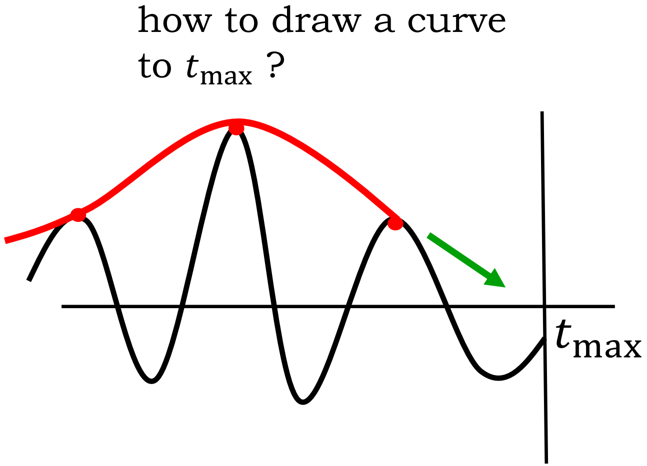
  then the completeness will be recovered.

$$h(t) = \sum_k c_k(t) + r(t)$$

near both ends of data

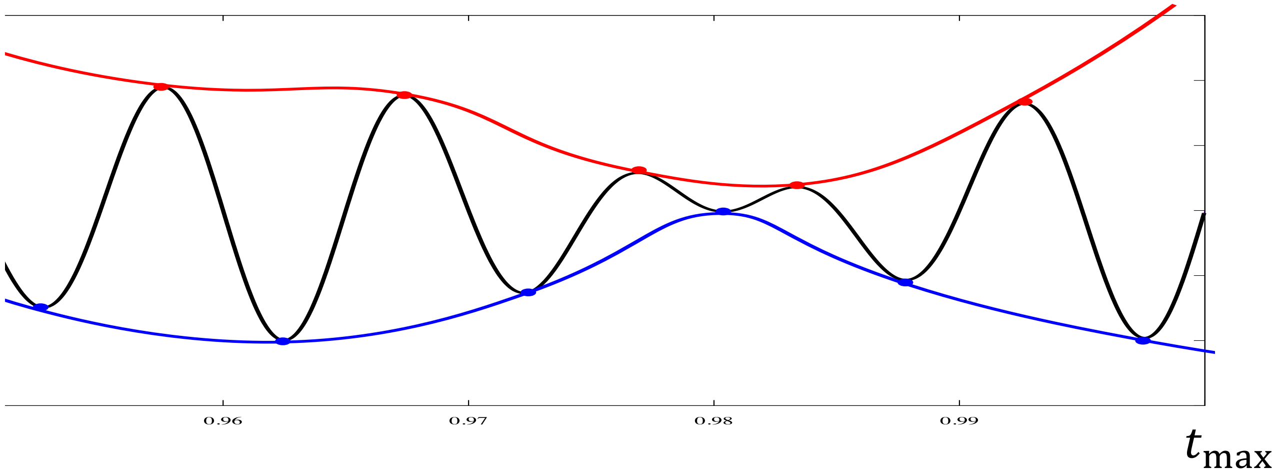
- A cubic spline (cspline) is usually used to interpolate the local extrema to find the upper and lower envelopes.
- Extrapolation is needed between the first (last) extremum and the end of data.





near both ends of data

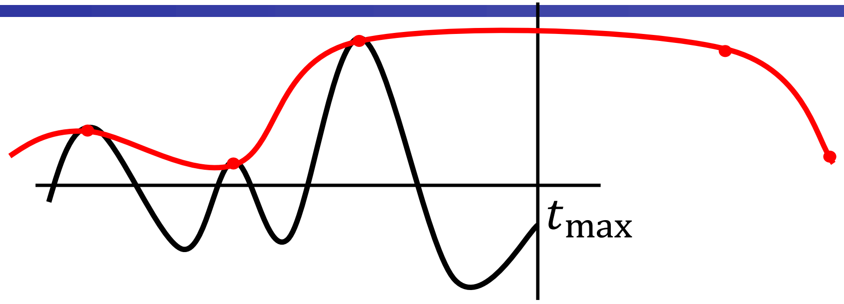
- Extrapolation may give an extremely large value at the edge.



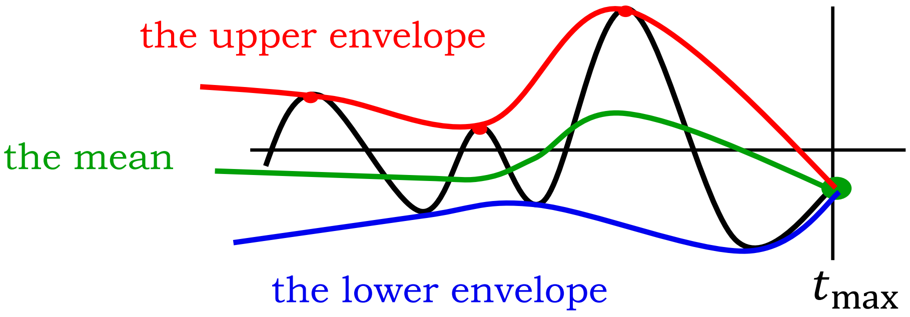
- It gives an erroneous value of the mean of upper and lower envelope due to loss of digits.
- To suppress this issue, we add an extra point on the outside of the edge in calculating the coefficients of cspline. It works well in most cases.

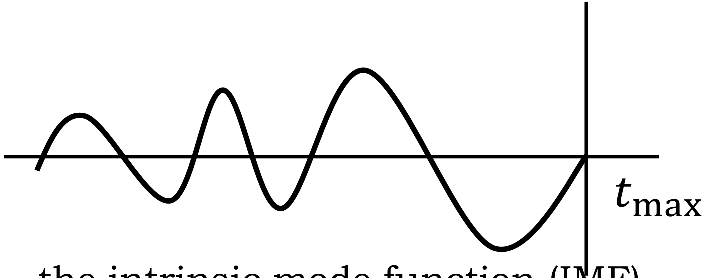
near both ends of data

1. to add knots assuming reflection symmetry wrt $t_{\min}$ and $t_{\max}$



2. to add an extra knot on the signal at $t_{\min}$ and $t_{\max}$ both for the upper and lower envelopes



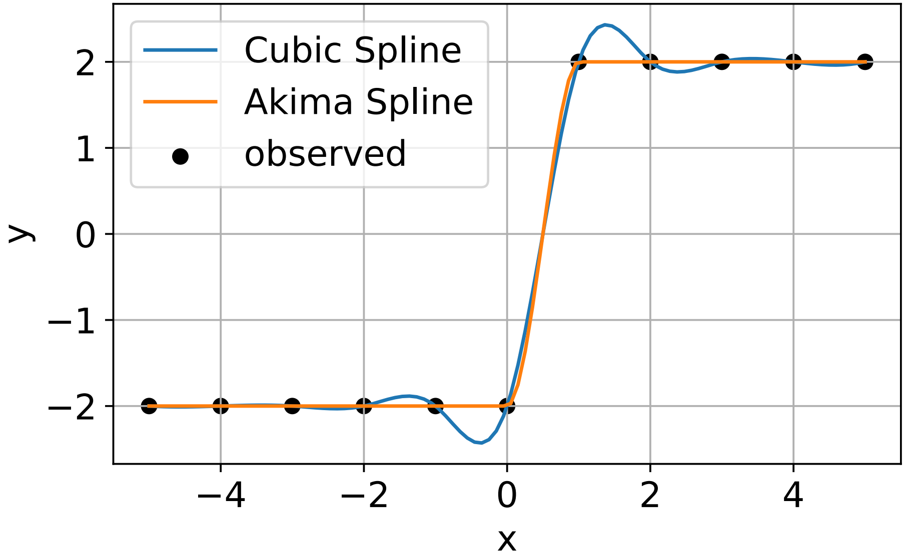


the intrinsic mode function (IMF)

- Overshoot in cspline sometimes occurs with both cases.

Akima spline

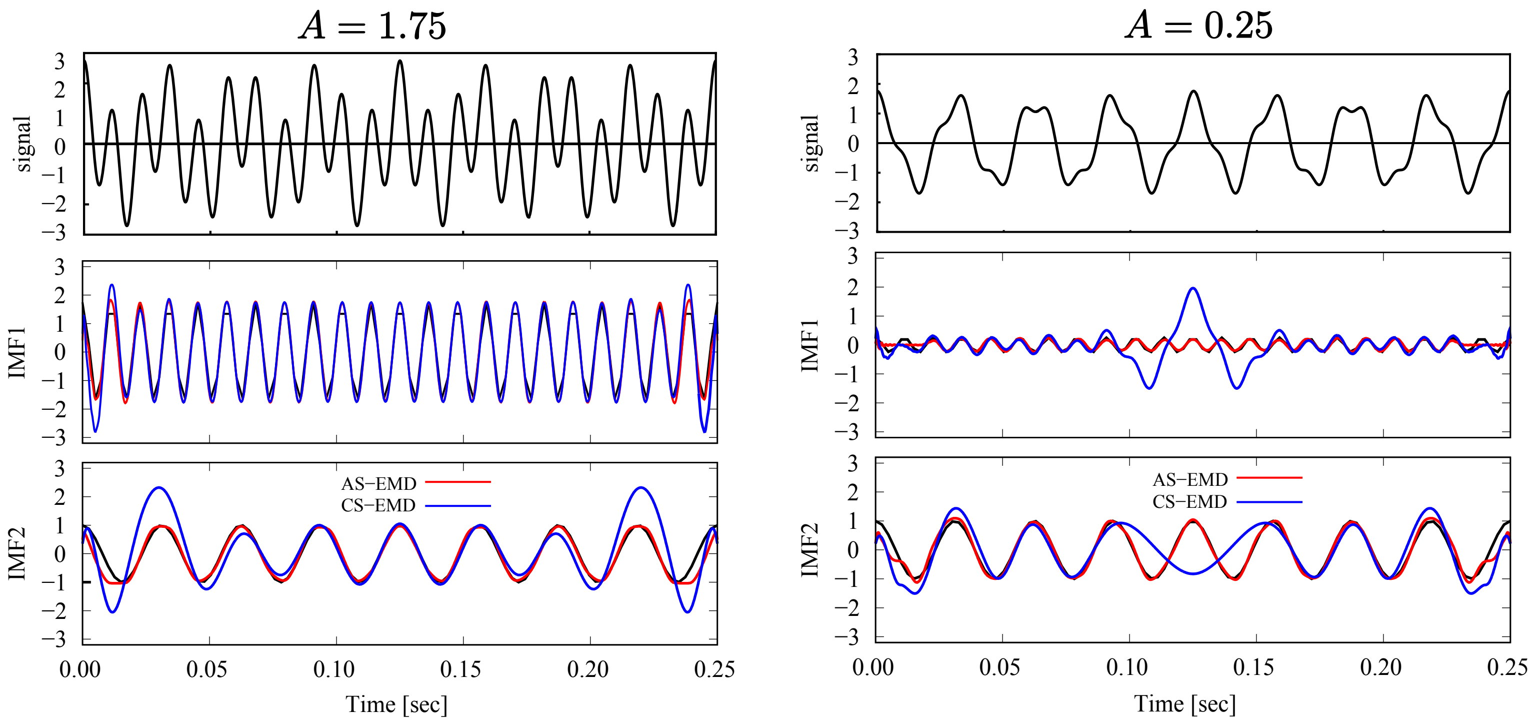
- Cubic spline often causes pseudo-oscillations and overshoots where there are jumps or the second derivative changes markedly.
- Akima spline generates natural interpolated curves like handwriting, by abandoning the continuity of the second derivative.



Akima spline vs cubic spline

- Akima spline tends to suppress overshoots near the boundary of time and mode mixing. [I. Yoda et al., Prog. Theor. Exp. Phys. in press (2023)]

$$s(t) = \cos(2\pi f_1 t) + A\cos(2\pi f_2 t) \quad \left[f_1 = 32\ \mathrm{Hz}, \quad f_2 = 88\ \mathrm{Hz}\right]$$



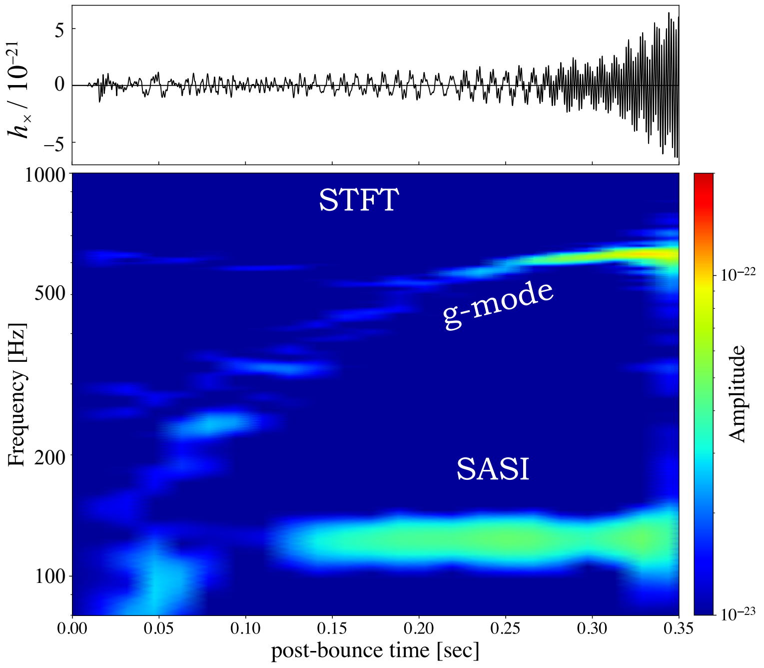
Parameters to be Predetermined

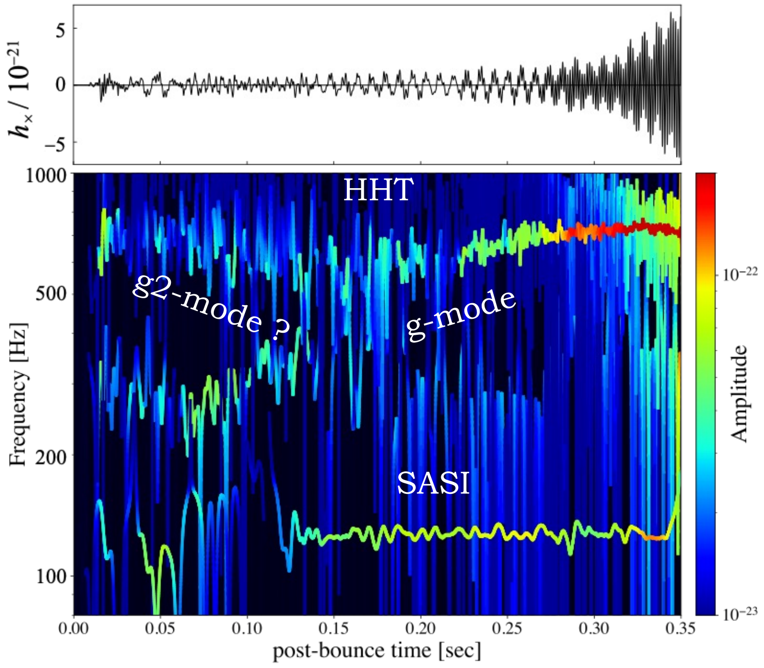
- To perform the EMD or the EEMD (CEEMD), we must predetermine
  - ➢ stoppage criterion:
    - the value of ε
  - ➢ the magnitude of noise σ_E to be added for EEMD
    - The parameter σ_E is given as a value relative to the rms of input data.
  - ➢ the size of the ensemble for EEMD, which depends on σ_E.
  - ➢ how to calculate envelopes near the edge of data

Analysis of Gravitational Waves in a Core-collapse supernova

M. Takeda, Y. Hiranuma, et al., Phys. Rev. D104, 084063 (2021)

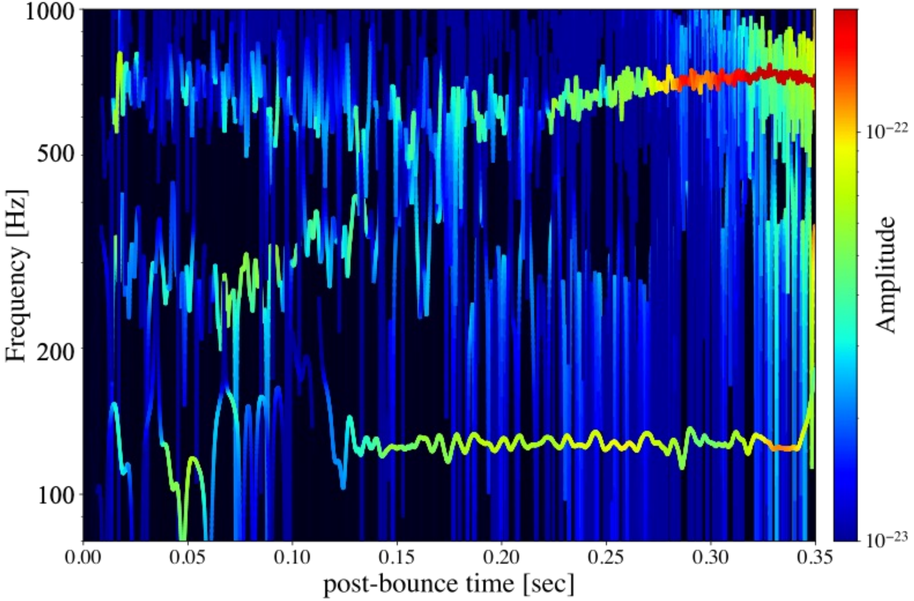
- Numerical simulation of core collapse supernova (Kuroda, et al. 2016) suggests that strong GWs are emitted from g-mode oscillation in the core and standing accretion shock instability (SASI).
- The SASI mode as well as g-mode can be clearly detected using the HHT.

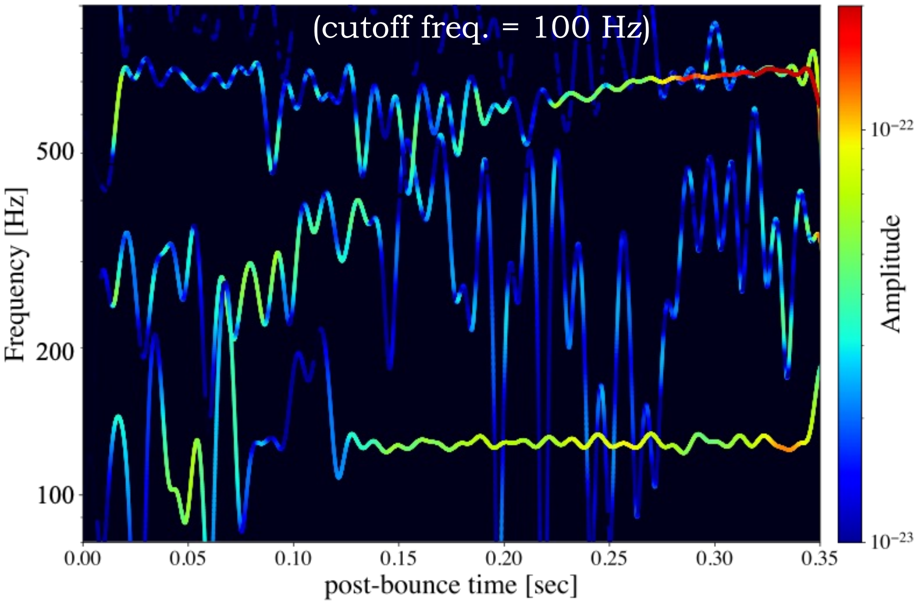




Analysis of Gravitational Waves in a Core-collapse supernova

- Short-time oscillations in the frequency $f(t)$ are considered to be due to numerical errors.
  - ➢ The frequency is obtained as the first derivative of the phase.
- Assuming it is true,
  the short-time oscillations can be eliminated by applying a lowpass filter to $f(t)$,
  since the HHT gives $f(t)$ (the instantaneous frequency) as a function of time.

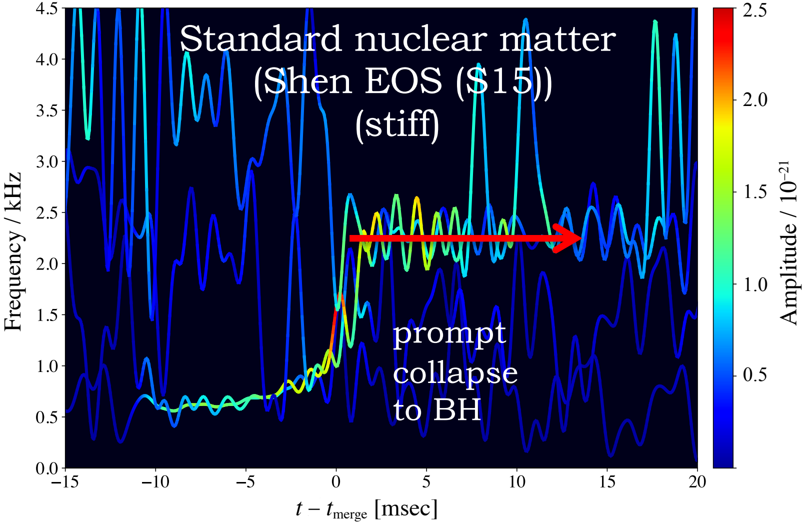


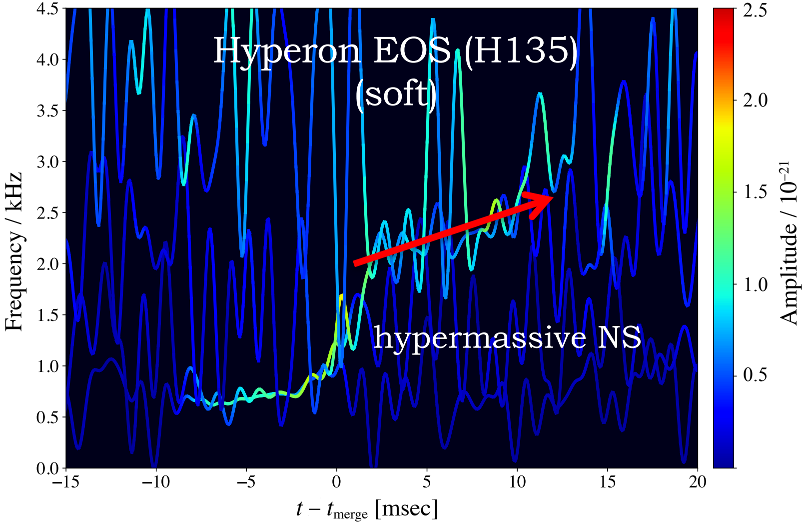


GWs from binary neutron star merger

M. Kaneyama, et al., Phys. Rev. D93, 123010 (2016); I. Yoda et al., Prog. Theor. Exp. Phys. in press (2023)

- Numerical relativity revealed that GWs in the post-merger phase depend greatly on EOS of the NS matter (Sekiguchi et al, 2011).
- The structure of the signal clearly appear in HHT map (T-F map).





- Using the original EEMD, we can distinguish the EOS from the evolution of IF in the post-merger phase up to ∼ 20 Mpc with Advanced-LIGO.
- Using the Akima spline, it is extended to ∼ 45 Mpc.

Summary and future work

- We apply the HHT to analysis of GWs with some improvement of EEMD, including substituting Akima spline interpolation for the cubic spline and careful treatment near both ends of time series data, etc.
- It is easy to parallelize the code using MPI to reduce computation time.
- Various kinds of future improvement have been proposed.
  We will investigate whether they are useful for analysis of GWs with real observed data.