This material may be protected by Copyright Law (Title 17 - U.S. Code)
Borrower: PMU

Lending String: *TET,NJS,IBS,IBS,KFH

Patron: Ierik Henrick

Journal Title: Proceedings. Mathematical, physical, and engineering sciences /

Volume: 454
Issue: 
Month/Year: 1998
Pages: 903-995

Article Author:

Article Title: 'Empirical mode decomposition and the Hilbert spectrum ...'

Imprint: London ; The Society, c1996-

ILL Number: 3830442
The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis

By Norden E. Huang\textsuperscript{1}, Zheng Shen\textsuperscript{2}, Steven R. Long\textsuperscript{3}, Manli C. Wu\textsuperscript{4}, Hsing H. Shih\textsuperscript{5}, Quanan Zheng\textsuperscript{6}, Nai-Chyuan Yen\textsuperscript{7}, Chi Chao Tung\textsuperscript{8} and Henry H. Liu\textsuperscript{9}

\textsuperscript{1}Laboratory for Hydrospheric Processes/Oceans and Ice Branch, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
\textsuperscript{2}Department of Earth and Planetary Sciences, The Johns Hopkins University, Baltimore, MD 21218, USA
\textsuperscript{3}Laboratory for Hydrospheric Processes/Observational Science Branch, NASA Wallops Flight Facility, Wallops Island, VA 23337, USA
\textsuperscript{4}Laboratory for Atmospheres, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
\textsuperscript{5}NOAA National Ocean Service, Silver Spring, MD 20910, USA
\textsuperscript{6}College of Marine Studies, University of Delaware, DE 19716, USA
\textsuperscript{7}Naval Research Laboratory, Washington, DC 20375-5000, USA
\textsuperscript{8}Department of Civil Engineering, North Carolina State University, Raleigh, NC 27695-7908, USA
\textsuperscript{9}Naval Surface Warfare Center, Carderock Division, Bethesda, MD 20084-5000, USA

Received 3 June 1996; accepted 4 November 1996

Contents

1. Introduction .......................... 904
2. Review of non-stationary data processing methods 907
   (a) The spectrogram 907
   (b) The wavelet analysis 907
   (c) The Wigner-Ville distribution 908
   (d) Evolutionary spectrum 909
   (c) The empirical orthogonal function expansion (EOF) 909
   (f) Other miscellaneous methods 911
3. Instantaneous frequency 915
4. Intrinsic mode functions 917
5. The empirical mode decomposition method: the sifting process 923
6. Completeness and orthogonality 928
7. The Hilbert spectrum 933
8. Validation and calibration of the Hilbert spectrum 948
9. Applications 949
   (a) Numerical results from classic nonlinear systems 949
   (b) Observational data from laboratory and field experiments 962


© 1998 The Royal Society

Printed in Great Britain
A new method for analysing nonlinear and non-stationary data has been developed. The key part of the method is the 'empirical mode decomposition' method with which any complicated data set can be decomposed into a finite and often small number of 'intrinsic mode functions' that admit well-behaved Hilbert transforms. This decomposition method is adaptive, and, therefore, highly efficient. Since the decomposition is based on the local characteristic time scale of the data, it is applicable to nonlinear and non-stationary processes. With the Hilbert transform, the 'intrinsic mode functions' yield instantaneous frequencies as functions of time that give sharp identifications of imbedded structures. The final presentation of the results is in an energy-frequency-time distribution, designated as the Hilbert spectrum. In this method, the main conceptual innovations are the introduction of 'intrinsic mode functions' based on local properties of the signal, which makes the instantaneous frequency meaningful; and the introduction of the instantaneous frequencies for complicated data sets, which eliminate the need for spurious harmonics to represent nonlinear and non-stationary signals. Examples from the numerical results of the classical nonlinear equation systems and data representing natural phenomena are given to demonstrate the power of this new method. Classical nonlinear system data are especially interesting, for they serve to illustrate the roles played by the nonlinear and non-stationary effects in the energy-frequency-time distribution.

**Keywords:** non-stationary time series; nonlinear differential equations; frequency-time spectrum; Hilbert spectral analysis; intrinsic time scale; empirical mode decomposition

---

1. Introduction

Data analysis is a necessary part in pure research and practical applications. Imperfect as some data might be, they represent the reality sensed by us; consequently, data analysis serves two purposes: to determine the parameters needed to construct the necessary model, and to confirm the model we constructed to represent the phenomenon. Unfortunately, the data, whether from physical measurements or numerical modelling, most likely will have one or more of the following problems: (a) the total data span is too short; (b) the data are non-stationary; and (c) the data represent nonlinear processes. Although each of the above problems can be real by itself, the first two are related, for a data section shorter than the longest time scale of a stationary process can appear to be non-stationary. Facing such data, we have limited options to use in the analysis.

Historically, Fourier spectral analysis has provided a general method for examining the global energy-frequency distributions. As a result, the term 'spectrum' has become almost synonymous with the Fourier transform of the data. Partially because of its prowess and partially because of its simplicity, Fourier analysis has dominated the data analysis efforts since soon after its introduction, and has been applied to all kinds of data. Although the Fourier transform is valid under extremely general conditions (see, for example, Titchmarsh 1948), there are some crucial restrictions of
the Fourier spectral analysis: the system must be linear; and the data must be strictly periodic or stationary; otherwise, the resulting spectrum will make little physical sense.

The stationarity requirement is not particular to the Fourier spectral analysis; it is a general one for most of the available data analysis methods. Therefore, it behooves us to review the definitions of stationarity here. According to the traditional definition, a time series, \( X(t) \), is stationary in the wide sense, if, for all \( t \),

\[
\begin{align*}
E(|X(t)|^2) &< \infty, \\
E(X(t)) &= m, \\
C(X(t_1), X(t_2)) &= C(X(t_1 + \tau), X(t_2 + \tau)) = C(t_1 - t_2),
\end{align*}
\]

in which \( E(\cdot) \) is the expected value defined as the ensemble average of the quantity, and \( C(\cdot) \) is the covariance function. Stationarity in the wide sense is also known as weak stationarity, covariance stationarity or second-order stationarity (see, for example, Brockwell & Davis 1991). A time series, \( X(t) \), is strictly stationary, if the joint distribution of

\[
[X(t_1), X(t_2), \ldots, X(t_n)] \quad \text{and} \quad [X(t_1 + \tau), X(t_2 + \tau), \ldots, X(t_n + \tau)]
\]

are the same for all \( t_i \) and \( \tau \). Thus, a strictly stationary process with finite second moments is also weakly stationary, but the inverse is not true. Both definitions are rigorous but idealized. Other less rigorous definitions for stationarity have also been used; for example, piecwise stationarity is for any random variable that is stationary within a limited time span, and asymptotically stationary is for any random variable that is stationary when \( \tau \) in equations (1.1) or (1.2) approaches infinity. In practice, we can only have data for finite time spans; therefore, even to check these definitions, we have to make approximations. Few of the data sets, from either natural phenomena or artificial sources, can satisfy these definitions. It may be argued that the difficulty of invoking stationarity as well as ergodicity is not on principle but on practicality: we just cannot have enough data to cover all possible points in the phase plane; therefore, most of the cases facing us are transient in nature. This is the reality; we are forced to face it.

Other than stationarity, Fourier spectral analysis also requires linearity. Although many natural phenomena can be approximated by linear systems, they also have the tendency to be nonlinear whenever their variations become finite in amplitude. Compounding these complications is the imperfection of our probes or numerical schemes; the interactions of the imperfect probes even with a perfect linear system can make the final data nonlinear. For the above reasons, the available data are usually of finite duration, non-stationary and from systems that are frequently nonlinear, either intrinsically or through interactions with the imperfect probes or numerical schemes. Under these conditions, Fourier spectral analysis is of limited use. For lack of alternatives, however, Fourier spectral analysis is still used to process such data. The uncritical use of Fourier spectral analysis and the inconsiderate adoption of the stationary and linear assumptions may give misleading results; some of those are described as follows.

First, the Fourier spectrum defines uniform harmonic components globally; therefore, it needs many additional harmonic components to simulate non-stationary data that are non-uniform globally. As a result, it spreads the energy over a wide frequency range. For example, using a delta function to represent a flash of light will give

a phase-locked wide white Fourier spectrum. Here, many Fourier components are added to simulate the non-stationary nature of the data in the time domain, but their existence diverts energy to a much wider frequency domain. Constrained by the energy conservation, these spurious harmonics and the wide frequency spectrum cannot faithfully represent the true energy density in the frequency space. More seriously, the Fourier representation also requires the existence of negative light intensity so that the components can cancel out one another to give the final delta function. Thus, the Fourier components might make mathematical sense, but do not really make physical sense at all. Although no physical process can be represented exactly by a delta function, some data such as the near-field strong earthquake records are of extremely short durations lasting only a few seconds to tens of seconds at most. Such records almost approach a delta function, and they always give artificially wide Fourier spectra.

Second, Fourier spectral analysis uses linear superposition of trigonometric functions; therefore, it needs additional harmonic components to simulate the deformed wave-profiles. Such deformations, as will be shown later, are the direct consequence of nonlinear effects. Whenever the form of the data deviates from a pure sine or cosine function, the Fourier spectrum will contain harmonics. As explained above, both non-stationarity and nonlinearity can induce spurious harmonic components that cause energy spreading. The consequence is the misleading energy-frequency distribution for nonlinear and non-stationary data.

In this paper, we will present a new data analysis method based on the empirical mode decomposition (EMD) method, which will generate a collection of intrinsic mode functions (IMF). The decomposition is based on the direct extraction of the energy associated with various intrinsic time scales, the most important parameters of the system. Expressed in IMFs, they have well-behaved Hilbert transforms, from which the instantaneous frequencies can be calculated. Thus, we can localize any event on the time as well as the frequency axis. The decomposition can also be viewed as an expansion of the data in terms of the IMFs. Then, these IMFs, based on and derived from the data, can serve as the basis of that expansion which can be linear or nonlinear as dictated by the data, and it is complete and almost orthogonal. Most important of all, it is adaptive. As will be shown later in more detail, locality and adaptivity are the necessary conditions for the basis for expanding nonlinear and non-stationary time series; orthogonality is not a necessary criterion for our basis selection for a nonlinear system. The principle of this basis construction is based on the physical time scales that characterize the oscillations of the phenomena. The local energy and the instantaneous frequency derived from the IMFs through the Hilbert transform can give us a full energy-frequency-time distribution of the data.

We have obtained good results and new insights by applying the combination of the EMD and Hilbert spectral analysis methods to various data: from the numerical results of the classical nonlinear equation systems to data representing natural phenomena. The classical nonlinear systems serve to illustrate the roles played by the nonlinear effects in the energy-frequency-time distribution. With the low degrees of freedom, they can train our eyes for more complicated cases. Some limitations of this method will also be discussed and the conclusions presented. Before introducing the new method, we will first review the present available data analysis methods for non-stationary processes.

2. Review of non-stationary data processing methods

We will first give a brief survey of the methods available for processing non-stationary data. Since most of the methods still depend on Fourier analysis, they are limited to linear systems only. Of the few methods available, the adoption of any method is almost strictly determined according to the special field in which the application is made. The available methods are reviewed as follows.

(a) The spectrogram

The spectrogram is the most basic method, which is nothing but a limited time window–width Fourier spectral analysis. By successively sliding the window along the time axis, one can get a time–frequency distribution. Since it relies on the traditional Fourier spectral analysis, one has to assume the data to be piecewise stationary. This assumption is not always justified in non-stationary data. Even if the data are piecewise stationary how can we guarantee that the window size adopted always coincides with the stationary time scales? What can we learn about the variations longer than the local stationary time scale? Will the collection of the locally stationary pieces constitute some longer period phenomena? Furthermore, there are also practical difficulties in applying the method: in order to localize an event in time, the window width must be narrow, but, on the other hand, the frequency resolution requires longer time series. These conflicting requirements render this method of limited usage. It is, however, extremely easy to implement with the fast Fourier transform; thus, it has attracted a wide following. Most applications of this method are for qualitative display of speech pattern analysis (see, for example, Oppenheim & Schafer 1989).

(b) The wavelet analysis

The wavelet approach is essentially an adjustable window Fourier spectral analysis with the following general definition:

\[ W(a,b; X, \psi) = |a|^{-1/2} \int_{-\infty}^{\infty} X(t) \psi^*(t, b/a) \, dt, \quad (2.1) \]

in which \( \psi^*(\cdot) \) is the basic wavelet function that satisfies certain very general conditions, \( a \) is the dilation factor and \( b \) is the translation of the origin. Although time and frequency do not appear explicitly in the transformed result, the variable \( 1/a \) gives the frequency scale and \( b \), the temporal location of an event. An intuitive physical explanation of equation (2.1) is very simple: \( W(a,b; X, \psi) \) is the ‘energy’ of \( X \) of scale \( a \) at \( t = b \).

Because of this basic form of \( at + b \) involved in the transformation, it is also known as affine wavelet analysis. For specific applications, the basic wavelet function, \( \psi^*(\cdot) \), can be modified according to special needs, but the form has to be given before the analysis. In most common applications, however, the Morlet wavelet is defined as Gaussian enveloped sine and cosine wave groups with 5.5 waves (see, for example, Chan 1995). Generally, \( \psi^*(\cdot) \) is not orthogonal for different \( a \) for continuous wavelets. Although one can make the wavelet orthogonal by selecting a discrete set of \( a \), this discrete wavelet analysis will miss physical signals having scale different from the selected discrete set of \( a \). Continuous or discrete, the wavelet analysis is basically a linear analysis. A very appealing feature of the wavelet analysis is that it provides a
uniform resolution for all the scales. Limited by the size of the basic wavelet function, the downside of the uniform resolution is uniformly poor resolution.

Although wavelet analysis has been available only in the last ten years or so, it has become extremely popular. Indeed, it is very useful in analysing data with gradual frequency changes. Since it has an analytic form for the result, it has attracted extensive attention of the applied mathematicians. Most of its applications have been in edge detection and image compression. Limited applications have also been made to the time-frequency distribution in time series (see, for example, Farge 1992; Long et al. 1993) and two-dimensional images (Spedding et al. 1993).

Versatile as the wavelet analysis is, the problem with the most commonly used Morlet wavelet is its leakage generated by the limited length of the basic wavelet function, which makes the quantitative definition of the energy–frequency–time distribution difficult. Sometimes, the interpretation of the wavelet can also be counter-intuitive. For example, to define a change occurring locally, one must look for the result in the high-frequency range, for the higher the frequency the more localized the basic wavelet will be. If a local event occurs only in the low-frequency range, one will still be forced to look for its effects in the high-frequency range. Such interpretation will be difficult if it is possible at all (see, for example, Huang et al. 1996). Another difficulty of the wavelet analysis is its non-adaptive nature. Once the basic wavelet is selected, one will have to use it to analyse all the data. Since the most commonly used Morlet wavelet is Fourier based, it also suffers the many shortcomings of Fourier spectral analysis: it can only give a physically meaningful interpretation to linear phenomena; it can resolve the interwave frequency modulation provided the frequency variation is gradual, but it cannot resolve the intrawave frequency modulation because the basic wavelet has a length of 5.5 waves. In spite of all these problems, wavelet analysis is still the best available non-stationary data analysis method so far; therefore, we will use it in this paper as a reference to establish the validity and the calibration of the Hilbert spectrum.

(c) The Wigner–Ville distribution

The Wigner–Ville distribution is sometimes also referred to as the Heisenberg wavelet. By definition, it is the Fourier transform of the central covariance function. For any time series, $X(t)$, we can define the central variance as

$$C_c(\tau, t) = X(t - \frac{1}{2}\tau)X^*(t + \frac{1}{2}\tau).$$

Then the Wigner–Ville distribution is

$$V(\omega, t) = \int_{-\infty}^{\infty} C_c(\tau, t)e^{-i\omega \tau} d\tau.$$  

This transform has been treated extensively by Clausen & Mecklenbräuker (1980a, b, c) and by Cohen (1995). It has been extremely popular with the electrical engineering community.

The difficulty with this method is the severe cross terms as indicated by the existence of negative power for some frequency ranges. Although this shortcoming can be eliminated by using the Kernel method (see, for example, Cohen 1995), the result is, then, basically that of a windowed Fourier analysis; therefore, it suffers all the limitations of the Fourier analysis. An extension of this method has been made by Yan (1994), who used the Wigner–Ville distribution to define wave packets that reduce...
a complicated data set to a finite number of simple components. This extension is very powerful and can be applied to a variety of problems. The applications to complicated data, however, require a great amount of judgement.

(d) Evolutionary spectrum

The evolutionary spectrum was first proposed by Priestley (1965). The basic idea is to extend the classic Fourier spectral analysis to a more generalized basis: from sine or cosine to a family of orthogonal functions \( \{ \phi(\omega, t) \} \) indexed by time, \( t \), and defined for all real \( \omega \), the frequency. Then, any real random variable, \( X(t) \), can be expressed as

\[
X(t) = \int_{-\infty}^{\infty} \phi(\omega, t) \, dA(\omega, t),
\]

in which \( dA(\omega, t) \), the Stieltjes function for the amplitude, is related to the spectrum as

\[
E(|dA(\omega, t)|^2) = d\mu(\omega, t) = S(\omega, t) \, d\omega,
\]

where \( \mu(\omega, t) \) is the spectrum, and \( S(\omega, t) \) is the spectral density at a specific time \( t \), also designated as the evolutionary spectrum. If for each fixed \( \omega \), \( \phi(\omega, t) \) has a Fourier transform

\[
\phi(\omega, t) = a(\omega, t)e^{i\Omega(\omega)t},
\]

then the function of \( a(\omega, t) \) is the envelope of \( \phi(\omega, t) \), and \( \Omega(\omega) \) is the frequency. If, further, we can treat \( \Omega(\omega) \) as a single valued function of \( \omega \), then

\[
\phi(\omega, t) = \alpha(\omega, t)e^{i\omega t}.
\]

Thus, the original data can be expanded in a family of amplitude modulated trigonometric functions.

The evolutionary spectral analysis is very popular in the earthquake community (see, for example, Liu 1970, 1971, 1973; Lin & Cai 1995). The difficulty of its application is to find a method to define the basis, \( \{ \phi(\omega, t) \} \). In principle, for this method to work, the basis has to be defined \textit{a posteriori}. So far, no systematic way has been offered; therefore, constructing an evolutionary spectrum from the given data is impossible. As a result, in the earthquake community, the applications of this method have changed the problem from data analysis to data simulation: an evolutionary spectrum will be assumed, then the signal will be reconstituted based on the assumed spectrum. Although there is some general resemblance to the simulated earthquake signal with the real data, it is not the data that generated the spectrum. Consequently, evolutionary spectrum analysis has never been very useful. As will be shown, the EMD can replace the evolutionary spectrum with a truly adaptive representation for the non-stationary processes.

(e) The empirical orthogonal function expansion (EOF)

The empirical orthogonal function expansion (EOF) is also known as the principal component analysis, or singular value decomposition method. The essence of EOF is briefly summarized as follows: for any real \( z(x, t) \), the EOF will reduce it to

\[
z(x, t) = \sum_{k=1}^{n} a_k(t) f_k(x),
\]

in which

\[ f_j \cdot f_k = \delta_{jk}. \]  

(2.9)

The orthonormal basis, \( \{ f_k \} \), is the collection of the empirical eigenfunctions defined by

\[ C \cdot f_k = \lambda_k f_k, \]  

(2.10)

where \( C \) is the sum of the inner products of the variable.

EOF represents a radical departure from all the above methods, for the expansion basis is derived from the data, therefore, it is \textit{a posteriori}, and highly efficient. The critical flaw of EOF is that it only gives a distribution of the variance in the modes defined by \( \{ f_k \} \), but this distribution by itself does not suggest scales or frequency content of the signal. Although it is tempting to interpret each mode as independent variations, this interpretation should be viewed with great care, for the EOF decomposition is not unique. A single component out of a non-unique decomposition, even if the basis is orthogonal, does not usually contain physical meaning. Recently, Vautard & Ghil (1989) proposed the singular spectral analysis method, which is the Fourier transform of the EOF. Here again, we have to be sure that each EOF component is stationary, otherwise the Fourier spectral analysis will make little sense on the EOF components. Unfortunately, there is no guarantee that EOF components from a nonlinear and non-stationary data set will be linear and stationary. Consequently, singular spectral analysis is not a real improvement. Because of its adaptive nature, however, the EOF method has been very popular, especially in the oceanography and meteorology communities (see, for example, Simpson 1991).

\section*{(f) Other miscellaneous methods}

Other than the above methods, there are also some miscellaneous methods such as least square estimation of the trend, smoothing by moving averaging, and differencing to generate stationary data. Methods like these, though useful, are too specialized to be of general use. They will not be discussed any further here. Additional details can be found in many standard data processing books (see, for example, Brockwell & Davis 1991).

All the above methods are designed to modify the global representation of the Fourier analysis, but they all failed in one way or the other. Having reviewed the methods, we can summarize the necessary conditions for the basis to represent a nonlinear and non-stationary time series: (a) complete; (b) orthogonal; (c) local; and (d) adaptive.

The first condition guarantees the degree of precision of the expansion; the second condition guarantees positivity of energy and avoids leakage. They are the standard requirements for all the linear expansion methods. For nonlinear expansions, the orthogonality condition needs to be modified. The details will be discussed later. But even these basic conditions are not satisfied by some of the above mentioned methods. The additional conditions are particular to the nonlinear and non-stationary data. The requirement for locality is the most crucial for non-stationarity, for in such data there is no time scale; therefore, all events have to be identified by the time of their occurrences. Consequently, we require both the amplitude (or energy) and the frequency to be functions of time. The requirement for adaptivity is also crucial for both nonlinear and non-stationary data, for only by adapting to the local variations of the data can the decomposition fully account for the underlying physics.
of the processes and not just to fulfill the mathematical requirements for fitting the data. This is especially important for the nonlinear phenomena, for a manifestation of nonlinearity is the 'harmonic distortion' in the Fourier analysis. The degree of distortion depends on the severity of nonlinearity; therefore, one cannot expect a predetermined basis to fit all the phenomena. An easy way to generate the necessary adaptive basis is to derive the basis from the data.

In this paper, we will introduce a general method which requires two steps in analysing the data as follows. The first step is to preprocess the data by the empirical mode decomposition method, with which the data are decomposed into a number of intrinsic mode function components. Thus, we will expand the data in a basis derived from the data. The second step is to apply the Hilbert transform to the decomposed IMFs and construct the energy–frequency–time distribution, designated as the Hilbert spectrum, from which the time localities of events will be preserved. In other words, we need the instantaneous frequency and energy rather than the global frequency and energy defined by the Fourier spectral analysis. Therefore, before going any further, we have to clarify the definition of the instantaneous frequency.

3. Instantaneous frequency

The notion of the instantaneous energy or the instantaneous envelope of the signal is well accepted; the notion of the instantaneous frequency, on the other hand, has been highly controversial. Existing opinions range from editing it out of existence (Shackel 1953) to accepting it but only for special 'monocomponent' signals (Boashash 1992; Cohen 1995).

There are two basic difficulties with accepting the idea of an instantaneous frequency as follows. The first one arises from the deeply entrenched influence of the Fourier spectral analysis. In the traditional Fourier analysis, the frequency is defined for the sine or cosine function spanning the whole data length with constant amplitude. As an extension of this definition, the instantaneous frequencies also have to relate to either a sine or a cosine function. Thus, we need at least one full oscillation of a sine or a cosine wave to define the local frequency value. According to this logic, nothing shorter than a full wave will do. Such a definition would not make sense for non-stationary data for which the frequency has to change values from time to time. The second difficulty arises from the non-unique way in defining the instantaneous frequency. Nevertheless, this difficulty is no longer serious since the introduction of the means to make the data analytical through the Hilbert transform. Difficulties, however, still exist as 'paradoxes' discussed by Cohen (1995). For an arbitrary time series, \( X(t) \), we can always have its Hilbert Transform, \( Y(t) \), as

\[
Y(t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{X(t')}{t - t'} \, dt',
\]

where \( P \) indicates the Cauchy principal value. This transform exists for all functions of class \( L^p \) (see, for example, Titchmarsh 1948). With this definition, \( X(t) \) and \( Y(t) \) form the complex conjugate pair, so we can have an analytic signal, \( Z(t) \), as

\[
Z(t) = X(t) + iY(t) = a(t)e^{i\theta(t)},
\]

in which

\[
a(t) = [X^2(t) + Y^2(t)]^{1/2}, \quad \theta(t) = \arctan \left( \frac{Y(t)}{X(t)} \right).
\]
Theoretically, there are infinitely many ways of defining the imaginary part, but the Hilbert transform provides a unique way of defining the imaginary part so that the result is an analytic function. A brief tutorial on the Hilbert transform with the emphasis on its physical interpretation can be found in Bendat & Piersol (1986). Essentially equation (3.1) defines the Hilbert transform as the convolution of \( X(t) \) with \( 1/t \); therefore, it emphasizes the local properties of \( X(t) \). In equation (3.2), the polar coordinate expression further clarifies the local nature of this representation; it is the best local fit of an amplitude and phase varying trigonometric function to \( X(t) \). Even with the Hilbert transform, there is still considerable controversy in defining the instantaneous frequency as

\[
\omega = \frac{d\theta(t)}{dt}.
\] (3.4)

This leads Cohen (1995) to introduce the term, 'monocomponent function'. In principle, some limitations on the data are necessary, for the instantaneous frequency given in equation (3.4) is a single value function of time. At any given time, there is only one frequency value; therefore, it can only represent one component, hence 'monocomponent'. Unfortunately, no clear definition of the 'monocomponent' signal was given to judge whether a function is or is not 'monocomponent'. For lack of a precise definition, 'narrow band' was adopted as a limitation on the data for the instantaneous frequency to make sense (Schwartz et al. 1966).

There are two definitions for bandwidth. The first one is used in the study of the probability properties of the signals and waves, where the processes are assumed to be stationary and Gaussian. Then, the bandwidth can be defined in terms of spectral moments as follows. The expected number of zero crossings per unit time is given by

\[
N_0 = \frac{1}{\pi} \left( \frac{m_2}{m_0} \right)^{1/2},
\] (3.5)

while the expected number of extrema per unit time is given by

\[
N_1 = \frac{1}{\pi} \left( \frac{m_4}{m_2} \right)^{1/2},
\] (3.6)

in which \( m_i \) is the \( i \)th moment of the spectrum. Therefore, the parameter, \( \nu \), defined as

\[
N_1^2 - N_0^2 = \frac{1}{\pi^2} \frac{m_4 m_0 - m_2^2}{m_0 m_2} = \frac{1}{\pi^2} \nu^2,
\] (3.7)

offers a standard bandwidth measure (see, for example, Rice 1944a, b, 1945a, b; Longuet-Higgins 1957). For a narrow band signal \( \nu = 0 \), the expected numbers of extrema and zero crossings have to equal.

The second definition is a more general one; it is again based on the moments of the spectrum, but in a different way. Let us take a complex valued function in polar coordinates as

\[
z(t) = a(t)e^{i\theta(t)},
\] (3.8)

with both \( a(t) \) and \( \theta(t) \) being functions of time. If this function has a spectrum, \( S(\omega) \), then the mean frequency is given by

\[
\langle \omega \rangle = \int \omega |S(\omega)|^2 d\omega,
\] (3.9)
which can be expressed in another way as
\[
\langle \omega \rangle = \int z^*(t) \frac{1}{i} \frac{d}{dt} z(t) \, dt
\]
\[
= \int \left( \dot{\theta}(t) - i \frac{\dot{\theta}(t)}{\omega(t)} \right) \dot{\omega}^2(t) \, dt
\]
\[
= \int \dot{\theta}(t) \dot{\omega}^2(t) \, dt. \tag{3.10}
\]

Based on this expression, Cohen (1995) suggested that \( \dot{\theta} \) be treated as the instantaneous frequency. With these notations, the bandwidth can be defined as
\[
\nu^2 = \frac{(\omega - \langle \omega \rangle)^2}{\langle \omega \rangle^2} = \frac{1}{\langle \omega \rangle^2} \int (\omega - \langle \omega \rangle)^2 |S(\omega)|^2 \, d\omega
\]
\[
= \frac{1}{\langle \omega \rangle^2} \int z^*(t) \left( \frac{1}{i} \frac{d}{dt} - \langle \omega \rangle \right)^2 z(t) \, dt
\]
\[
= \frac{1}{\langle \omega \rangle^2} \left[ \int \dot{a}^2(t) \, dt + \int (\dot{\theta}(t) - \langle \omega \rangle)^2 \dot{\omega}^2(t) \, dt \right]. \tag{3.11}
\]

For a narrow band signal, this value has to be small, then both \( a \) and \( \theta \) have to be gradually varying functions. Unfortunately, both equations (3.7) and (3.11) defined the bandwidth in the global sense: they are both overly restrictive and lack precision at the same time. Consequently, the bandwidth limitation of the Hilbert transform to give a meaningful instantaneous frequency has never been firmly established. For example, Melville (1983) had faithfully filtered the data within the bandwidth requirement, but he still obtained many non-physical negative frequency values. It should be mentioned here that using filtering to obtain a narrow band signal is unsatisfactory for another reason: the filtered data have already been contaminated by the spurious harmonics caused by the nonlinearity and non-stationarity as discussed in the introduction.

In order to obtain meaningful instantaneous frequency, restrictive conditions have to be imposed on the data as discussed by Gabor (1946), Bedrosian (1963) and, more recently, Boashash (1992): for any function to have a meaningful instantaneous frequency, the real part of its Fourier transform has to have only positive frequency. This restriction can be proven mathematically as shown in Titchmarsh (1948) but it is still global. For data analysis, we have to translate this requirement into physically implementable steps to develop a simple method for applications. For this purpose, we have to modify the restriction condition from a global one to a local one, and the basis has to satisfy the necessary conditions listed in the last section.

Let us consider some simple examples to illustrate these restrictions physically, by examining the function,
\[
x(t) = \sin t. \tag{3.12}
\]
Its Hilbert transform is simply \( \cos t \). The phase plot of \( x-y \) is a simple circle of unit radius as in figure 1a. The phase function is a straight line as shown in figure 1b and the instantaneous frequency, shown in figure 1c, is a constant as expected. If we move the mean off by an amount \( a \), say, then,
\[
x(t) = a + \sin t. \tag{3.13}
\]
any spatial and temporal localisation of the frequency information associated with the function and its instantaneous frequency. This has led to the proposal of a new approach to the localisation of the frequency of a function, which is based on the existence of a localised frequency function.

![Figure 1](image.png)

The physical interpretation of instantaneous frequency is shown in figure 1. The phase plane for the model function $x(t) = \alpha \sin t$ is given in figure 1a. The phase plane for the model function $x(t) = \alpha \sin t$ is shown in figure 1b. The instantaneous frequency computed according to equation (3.4) is given in figure 1c.

The phase plot of $x-y$ is still a simple circle independent of the value of $\alpha$, but the centre of the circle will be displaced by the amount of $\alpha$ as illustrated in figure 1a. If $\alpha < 1$, the centre is still within the circle. Under this condition, the function has already violated a restriction, for its Fourier spectrum has a DC term; nevertheless, the mean zero-crossing frequency is still the same as in the case for $\alpha = 0$, but the phase function and the instantaneous frequency will be very different as shown in figures 1b, c. If $\alpha > 1$, the centre is outside the circle; thus, the function no longer satisfies the required conditions. Then both the phase function and the instantaneous frequency will assume negative values as shown in figures 1b, c, which are meaningless. These simple examples illustrate physically that, for a simple signal such as a sine function, the instantaneous frequency can be defined only if we restrict the function to be symmetric locally with respect to the zero mean level.

For general data, any riding waves would be equivalent to the case of $\alpha > 1$ locally;
any asymmetric wave form will be equivalent to the case of $\alpha < 1$, but non-zero, locally. In order to have a meaningful instantaneous frequency, this local restriction should be used in lieu of the global requirements given previously. Furthermore, this local restriction also suggests a method to decompose the data into components for which the instantaneous frequency can be defined. The examples presented above, however, actually lead us to the definition of a class of functions, based on its local properties, designated as intrinsic mode function (IMF) for which the instantaneous frequency can be defined everywhere. The limitation of interest here is not on the existence of the Hilbert transform which is general and global, but on the existence of a meaningful instantaneous frequency which is restrictive and local.

4. Intrinsic mode functions

The simple examples given above provide more physical interpretation of the restrictive conditions; they also suggest a practical way to decompose the data so that the components all satisfy the conditions imposed on them. Physically, the necessary conditions for us to define a meaningful instantaneous frequency are that the functions are symmetric with respect to the local zero mean, and have the same numbers of zero crossings and extrema. Based on these observations, we propose a class of functions designated as intrinsic mode functions here with the following formal definition.

An intrinsic mode function (IMF) is a function that satisfies two conditions: (1) in the whole data set, the number of extrema and the number of zero crossings must either equal or differ at most by one; and (2) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

The first condition is obvious; it is similar to the traditional narrow band requirements for a stationary Gaussian process. The second condition is a new idea; it modifies the classical global requirement to a local one; it is necessary so that the instantaneous frequency will not have the unwanted fluctuations induced by asymmetric wave forms. Ideally, the requirement should be ‘the local mean of the data being zero’. For non-stationary data, the ‘local mean’ involves a ‘local time scale’ to compute the mean, which is impossible to define. As a surrogate, we use the local mean of the envelopes defined by the local maxima and the local minima to force the local symmetry instead. This is a necessary approximation to avoid the definition of a local averaging time scale. Although it will introduce an alias in the instantaneous frequency for nonlinearly deformed waves, the effects of nonlinearity are much weaker in comparison with non-stationarity as we will discuss later. With the physical approach and the approximation adopted here, the method does not always guarantee a perfect instantaneous frequency under all conditions. Nevertheless, we will show that, even under the worst conditions, the instantaneous frequency so defined is still consistent with the physics of the system studied.

The name ‘intrinsic mode function’ is adopted because it represents the oscillation mode imbedded in the data. With this definition, the IMF in each cycle, defined by the zero crossings, involves only one mode of oscillation, no complex riding waves are allowed. With this definition, an IMF is not restricted to a narrow band signal, and it can be both amplitude and frequency modulated. In fact, it can be non-stationary. As discussed above, purely frequency or amplitude modulated functions can be IMFs
even though they have finite bandwidth according to the traditional definition. A typical IMF is shown in figure 2.

Having defined IMF, we will show that the definition given in equation (3.4) gives the best instantaneous frequency. An IMF after the Hilbert transform can be expressed as in equation (3.2). If we perform a Fourier transform on $Z(t)$, we have

$$W(\omega) = \int_{-\infty}^{\infty} a(t) e^{i\theta(t)} e^{-i\omega t} dt = \int_{-\infty}^{\infty} g(t) e^{i(\theta(t) - \omega t)} dt. \tag{4.1}$$

Then by the stationary phase method (see, for example, Copson 1967), the maximum contribution to $W(\omega)$ is given by the frequency satisfying the condition

$$\frac{d}{dt}(\theta(t) - \omega t) = 0; \tag{4.2}$$

therefore, equation (3.4) follows. Although mathematically, the application of the stationary phase method requires a large parameter for the exponential function, the adoption here can be justified if the frequency, $\omega$, is high compared with the inversely local time scale of the amplitude variation. Therefore, this definition fits the best for gradually changing amplitude. Even with this condition, this is still a much better definition for instantaneous frequency than the zero-crossing frequency; it is also better than the integral definition suggested by Cohen (1995) as given in equation (3.10). Furthermore, it agrees with the definition of frequency for the classic wave theory (see, for example, Whitham 1975).

As given in equation (4.2) and the simple analogy given in equations (3.2)–(3.4), the frequency defined through the stationary phase approximation agrees also with the best fit sinusoidal function locally; therefore, we do not need a whole oscillatory period to define a frequency value. We can define it for every point with the value changing from point to point. In this sense, even a monotonic function can be treated as part of an oscillatory function and have instantaneous frequency assigned according to equation (3.4). Any frequency variation is designated as frequency modulation. There are actually two types of frequency modulations: the interwave and the intrawave modulations. The first type is familiar to us; the frequency of the oscillation is gradually changing with the waves in a dispersive system. Technically, in the dispersive waves, the frequency is also changing within one wave, but that

was not emphasized either for convenience, or for lack of a more precise frequency definition. The second type is less familiar, but it is also a common phenomenon: if the frequency changes from time to time within a wave its profile can no longer be a simple sine or cosine function. Therefore, any wave-profile deformation from the simple sinusoidal form implies the intrawave frequency modulation. In the past such phenomena were treated as harmonic distortions. We will show in detail later that most such deformations are better viewed as intrawave frequency modulation, for the intrawave frequency modulation is more physical.

In order to use this unique definition of instantaneous frequency, we have to reduce an arbitrary data set into IMF components from which an instantaneous frequency value can be assigned to each IMF component. Consequently, for complicated data, we can have more than one instantaneous frequency at a time locally. We will introduce the empirical mode decomposition method to reduce the data into the needed IMFs.

5. The empirical mode decomposition method: the sifting process

Knowing the well-behaved Hilbert transforms of the IMF components is only the starting point. Unfortunately, most of the data are not IMFs. At any given time, the data may involve more than one oscillatory mode, that is why the simple Hilbert transform cannot provide the full description of the frequency content for the general data as reported by Long et al. (1995). We have to decompose the data into IMF components. Here, we will introduce a new method to deal with both non-stationary and nonlinear data by decomposing the signal first, and discuss the physical meaning of this decomposition later. Contrary to almost all the previous methods, this new method is intuitive, direct, a posteriori and adaptive, with the basis of the decomposition based on, and derived from, the data.

The decomposition is based on the assumptions: (1) the signal has at least two extrema—one maximum and one minimum; (2) the characteristic time scale is defined by the time lapse between the extrema; and (3) if the data were totally devoid of extrema but contained only inflection points, then it can be differentiated once or more times to reveal the extrema. Final results can be obtained by integration(s) of the components.

The essence of the method is to identify the intrinsic oscillatory modes by their characteristic time scales in the data empirically, and then decompose the data accordingly. According to Drazin (1992), the first step of data analysis is to examine the data by eye. From this examination, one can immediately identify the different scales directly in two ways: by the time lapse between the successive alternations of local maxima and minima; and by the time lapse between the successive zero crossings. The interlaced local extrema and zero crossings give us the complicated data: one undulation is riding on top of another, and they, in turn, are riding on still other undulations, and so on. Each of these undulations defines a characteristic scale of the data; it is intrinsic to the process. We have decided to adopt the time lapse between successive extrema as the definition of the time scale for the intrinsic oscillatory mode, because it not only gives a much finer resolution of the oscillatory modes, but also can be applied to data with non-zero mean, either all positive or all negative values, without zero crossings. A systematic way to extract them, designated as the sifting process, is described as follows.

Figure 3. Illustration of the sifting processes: (a) the original data; (b) the data in thin solid line, with the upper and lower envelopes in dot-dashed lines and the mean in thick solid line; (c) the difference between the data and \( m_1 \). This is still not an IMF, for there are negative local maxima and positive minima suggesting riding waves.

By virtue of the IMF definition, the decomposition method can simply use the envelopes defined by the local maxima and minima separately. Once the extrema are identified, all the local maxima are connected by a cubic spline line as the upper envelope. Repeat the procedure for the local minima to produce the lower envelope. The upper and lower envelopes should cover all the data between them. Their mean is designated as \( m_1 \), and the difference between the data and \( m_1 \) is the first component, \( h_1 \), i.e.

\[
X(t) - m_1 = h_1. \tag{5.1}
\]

The procedure is illustrated in figures 3a–c (figure 3a gives the data; figure 3b gives

the data in the thin solid line, the upper and the lower envelopes in the dot-dashed lines, and their mean in the thick solid line, which bisects the data very well; and figure 3c gives the difference between the data and the local mean as in equation (5.1)).

Ideally, \( h_1 \) should be an IMF, for the construction of \( h_1 \) described above seems to have been made to satisfy all the requirements of IMF. In reality, however, overshoots and undershoots are common, which can also generate new extrema, and shift or exaggerate the existing ones. The imperfection of the overshoots and undershoots can be found at the 4.5 and 4.7 s points in figure 3b. Their effects, however, are not direct, for it is the mean, not the envelopes, that will enter the sifting process. Nevertheless, the problem is real. Even if the fitting is perfect, a gentle hump on a slope can be amplified to become a local extremum in changing the local zero from a rectangular to a curvilinear coordinate system. An example can be found for the hump between the 4.5 and 4.6 s range in the data in figure 3a. After the first round of sifting, the hump becomes a local maximum at the same time location as in figure 3c. New extrema generated in this way actually recover the proper modes lost in the initial examination. In fact, the sifting process can recover low-amplitude riding waves with repeated siftings.

Still another complication is that the envelope mean may be different from the true local mean for nonlinear data; consequently, some asymmetric wave forms can still exist no matter how many times the data are sifted. We have to accept this approximation as discussed before.
Other than these theoretical difficulties, on the practical side, serious problems of the spline fitting can occur near the ends, where the cubic spline fitting can have large swings. Left by themselves, the end swings can eventually propagate inward and corrupt the whole data span especially in the low-frequency components. We have devised a numerical method to eliminate the end effects; details will be given later. At any rate, improving the spline fitting is absolutely necessary. Even with these problems, the sifting process can still extract the essential scales from the data.

The sifting process serves two purposes: to eliminate riding waves; and to make the wave-profiles more symmetric. Toward this end, the sifting process has to be repeated more times. In the second sifting process, \( h_1 \) is treated as the data, then

\[ h_1 - m_{11} = h_{11}. \]  

(5.2)

Figure 4a shows the much improved result after the second sifting, but there are still local maxima below the zero line. After another sifting, the result is given in figure 4b. Now all the local maxima are positive, and all the local minima are negative, but many waves are still asymmetric. We can repeat this sifting procedure \( k \) times, until \( h_{1k} \) is an IMF, that is

\[ h_{1(k-1)} - m_{1k} = h_{1k}, \]  

(5.3)

the result is shown in figure 2 after nine siftings. Then, it is designated as

\[ c_1 = h_{1k}, \]  

(5.4)

the first IMF component from the data.

As described above, the process is indeed like sifting: to separate the finest local mode from the data first based only on the characteristic time scale. The sifting process, however, has two effects: (a) to eliminate riding waves; and (b) to smooth uneven amplitudes.

While the first condition is absolutely necessary for the instantaneous frequency to be meaningful, the second condition is also necessary in case the neighboring wave amplitudes have too large a disparity. Unfortunately, the second effect, when carried to the extreme, could obliterate the physically meaningful amplitude fluctuations. Therefore, the sifting process should be applied with care, for carrying the process to an extreme could make the resulting IMF a pure frequency modulated signal of constant amplitude. To guarantee that the IMF components retain enough physical sense of both amplitude and frequency modulations, we have to determine a criterion for the sifting process to stop. This can be accomplished by limiting the size of the standard deviation, SD, computed from the two consecutive sifting results as

\[ \text{SD} = \sum_{i=0}^{T} \left[ \frac{|(h_{1(k-1)}(t) - h_{1k}(t))|^2}{h_{1(k-1)}^2(t)} \right]. \]  

(5.5)

A typical value for SD can be set between 0.2 and 0.3. As a comparison, the two Fourier spectra, computed by shifting only five out of 1024 points from the same data, can have an equivalent SD of 0.2–0.3 calculated point-by-point. Therefore, a SD value of 0.2–0.3 for the sifting procedure is a very rigorous limitation for the difference between siftings.

Overall, \( c_1 \) should contain the finest scale or the shortest period component of the signal. We can separate \( c_1 \) from the rest of the data by

\[ X(t) - c_1 = r_1. \]  

(5.6)
Since the residue, $r_1$, still contains information of longer period components, it is treated as the new data and subjected to the same sifting process as described above. This procedure can be repeated on all the subsequent $r_j$s, and the result is

$$r_1 - c_2 = r_2, \ldots, r_{n-1} - c_n = r_n.$$  \hspace{1cm} (5.7)

The sifting process can be stopped by any of the following predetermined criteria: either when the component, $c_n$, or the residue, $r_n$, becomes so small that it is less than the predetermined value of substantial consequence, or when the residue, $r_n$, becomes a monotonic function from which no more IMF can be extracted. Even for data with zero mean, the final residue can still be different from zero; for data with a trend, then the final residue should be that trend. By summing up equations (5.6) and (5.7), we finally obtain

$$X(t) = \sum_{i=1}^{n} c_i + r_n.$$  \hspace{1cm} (5.8)

Thus, we achieved a decomposition of the data into $n$-empirical modes, and a residue, $r_n$, which can be either the mean trend or a constant. As discussed here, to apply the EMD method, a mean or zero reference is not required; EMD only needs the locations of the local extrema. The zero references for each component will be generated by the sifting process. Without the need of the zero reference, EMD eliminates the troublesome step of removing the mean values for the large DC term in data with non-zero mean, an unexpected benefit.

To illustrate the sifting process, we will use a set of wind data collected in a laboratory wind-wave tunnel (Huang & Long 1980) with a high-frequency response Pitot tube located 10 cm above the mean water level. The wind speed was recorded under the condition of the initial onset of water waves from a calm surface. Calibrated wind data are given in figure 5. Clearly, the data are quite complicated with many

Figure 6. The resulting empirical mode decomposition components from the wind data: (a) the original data and the components $c_1$-$c_4$; (b) the components $c_5$-$c_9$. Notice the last component, $c_9$, is not an IMF; it is the trend.

local extrema but no zero crossings, for the time series represents all positive numbers. Although the mean can be treated as a zero reference, defining it is hard, for the whole process is transient. This example illustrates the advantage of adopting the successive extrema for defining the time scale; it also illustrates the difficulties of dealing with non-stationary data: even a meaningful mean is impossible to define, but for EMD this difficulty is eliminated. Figures 6a, b summarize all the IMF obtained from this repeated sifting processes. We have a total of nine components. Comparing this with the traditional Fourier expansion, one can immediately see the efficiency of the EMD: the expansion of a turbulence data set with only nine terms. From the result, one can see a general separation of the data into locally non-overlapping time scale components. In some components, such as c_3 and c_5, the signals are intermittent, then the neighbouring components might contain oscillations of the same scale, but signals of the same time scale would never occur at the same locations in two different IMF components.

Finally, let us examine the physical meaning of each IMF component. The components of the EMD are usually physical, for the characteristic scales are physical. Nevertheless, this is not strictly true, for there are cases when a certain scale of a phenomenon is intermittent. Then, the decomposed component could contain two scales in one IMF component. Therefore, the physical meaning of the decomposition comes only in the totality of the decomposed components in the Hilbert spectrum. Even with the entire set of decomposed components, sound physical interpretation is still not guaranteed for other decompositions such as Fourier expansion. Further discussions will be given later in detail. Having established the decomposition, we should check the completeness and orthogonality of this decomposition. Because this decomposition is a posteriori, the check should also be a posteriori.

6. Completeness and orthogonality

By virtue of the decomposition, completeness is given, for equation (5.8) is an identity. As a check of the completeness for the wind data numerically, we can reconstruct the data from the IMF components starting from the longest to the shortest periods in the sequence from figure 7a. Figure 7c gives the data and the longest period component, c_9, which is the residue trend, not an IMF. By itself, the fitting of the trend is quite impressive, and it is very physical: the gradual decrease of the mean wind speed indicates the lack of drag from the calm surface initially and the increase of drag after the generation of wind waves. As the mean wind speed decreases, the amplitude of the fluctuation increases, another indication of wind-wave interactions. If we add the next longest period component, c_8, the trend of the sum, c_9 + c_8, takes a remarkable turn, and the fitting to the data looks greatly improved as shown in figure 7b. By successively adding more components with increasing frequency, we have the results in the series of figures 7c–i. The gradual change from the monotonic trend to the final reconstruction is illustrative by itself. By the time we reach the sum of IMF components up to c_3 in figure 7g, we essentially have recovered all the energy containing eddies already. The components with the highest frequencies add little more energy, but they make the data look more complicated. In fact, the highest frequency component is probably not physical, for the digitizing rate of the Pitot tube is too slow to capture the high-frequency variations. As a result, the data are jagged artificially by the digitalizing steps at this frequency.

Figure 7. Numerical proof of the completeness of the EMD through reconstruction of the original data from the IMF components. (a) Data (in the dotted line) and the \( c_0 \) component (in the solid line). \( c_0 \) serves as a running mean, but it is not obtained from either mean or filtering. (b) Data (in the dotted line) and the sum of \( c_0 - c_k \) components (in the solid line). (c) Data (in the dotted line) and the sum of \( c_0 - c_7 \) components (in the solid line). (d) Data (in the dotted line) and the sum of \( c_0 - c_6 \) components (in the solid line).

The difference between the reconstructed data from the sum of all the IMFs and the original data is shown in figure 7j, in which the maximum amplitude is less than \( 5 \times 10^{-15} \), the roundoff error from the precision of the computer. Thus, the completeness is established both theoretically by equation (5.8), and numerically by figure 7j.

The orthogonality is satisfied in all practical sense, but it is not guaranteed theoretically. Let us discuss the practical aspect first. By virtue of the decomposition, the elements should all be locally orthogonal to each other, for each element is obtained from the difference between the signal and its local mean through the maximal and minimal envelopes; therefore,

\[
(x(t) - \bar{x}(t)) \cdot \bar{x}(t) = 0. \tag{6.1}
\]

Nevertheless, equation (6.1) is not strictly true, because the mean is computed via the envelopes, hence it is not the true mean. Furthermore, each successive IMF component is only part of the signal constituting \( x(t) \). Because of these approximations, leakage is unavoidable. Any leakage, however, should be small.

The orthogonality of the EMD components should also be checked \textit{a posteriori} numerically as follows: let us first write equation (5.8) as

\[
X(t) = \sum_{j=1}^{n+1} C_j(t), \tag{6.2}
\]

in which we have included \( r_n \) as an additional element. To check the orthogonality of the IMFs from EMD, we define an index based on the most intuitive way: first,
Figure 7. Cont. (e) Data (in the dotted line) and the sum of $c_2$—$c_5$ components (in the solid line). (f) Data (in the dotted line) and the sum of $c_0$—$c_4$ components (in the solid line). (g) Data (in the dotted line) and the sum of $c_5$—$c_3$ components (in the solid line). By now, we seem to have recovered all the energy containing eddies. (h) Data (in the dotted line) and the sum of $c_0$—$c_2$ components (in the solid line). (i) Data (in the dotted line) and the sum of $c_3$—$c_1$ components (in the solid line). This is the final reconstruction of the data from the IMFs. It appears no different from the original data. (j) The difference between the original data and the reconstructed one; the difference is the limit of the computational precision of the personal computer (PC) used.

Let us form the square of the signal as

$$X^2(t) = \sum_{j=1}^{n+1} C_j^2(t) + 2 \sum_{j=1}^{n+1} \sum_{k=1}^{n+1} C_j(t)C_k(t).$$  \hspace{1cm} (6.3)$$

If the decomposition is orthogonal, then the cross terms given in the second part on the right-hand side should be zero. With this expression, an overall index of orthogonality, \(\text{IO}\), is defined as

$$\text{IO} = \sum_{t=0}^{T} \left( \sum_{j=1}^{n+1} \sum_{k=1}^{n+1} C_j(t)C_k(t)/X^2(t) \right).$$  \hspace{1cm} (6.4)$$

For the wind data given above, the \(\text{IO}\) value is only 0.0067. Orthogonality can also

Figure 8. The Hilbert spectrum for the wind data with 200 frequency cells. The wind energy appears in skeleton lines representing each IMF.

Figure 9. The Morlet wavelet spectrum for the wind data with the same number of frequency cells. Wind energy appears in smoothed contours with a rich energy distribution in the high harmonics.

be defined for any two components, \( C_f \) and \( C_g \). The measure of orthogonality will, then, be

\[
\text{IO}_{fg} = \sum \frac{C_f C_g}{C_f^2 + C_g^2}.
\] (6.5)

It should be noted that, although the definition of orthogonality seems to be global, the real meaning here applies only locally. For some special data, the neighbouring components could certainly have sections of data carrying the same frequency at different time durations. But locally, any two components should be orthogonal for all practical purposes. The amount of leakage usually depends on the length of data as well as the decomposition results. In reality, because of the finite data length, even pure sinusoidal components with different frequencies are not exactly orthogonal. That is why the continuous wavelet in the most commonly used Morlet form suffers severe leakage. For the EMD method, we found the leakage to be typically less than 1%. For extremely short data, the leakage could be as high as 5%, which is comparable to that from a set of pure sinusoidal waves of the same data length.

Theoretically, the EMD method guarantees orthogonality only on the strength of equation (6.1). Orthogonality actually depends on the decomposition method. Let us consider two Stokian waves each having many harmonics. Separately, they are IMFs, of course, and each represents a free travelling wave. If the frequency of one Stokian wave coincides with the frequency of a harmonic of the other, then the two waves are no longer orthogonal in the Fourier sense. EMD, however, can still separate the two Stokian wave trains. Such a decomposition is physical, but the separated IMF components are not orthogonal. Therefore, orthogonality is a requirement only
for linear decomposition systems; it would not make physical sense for a nonlinear decomposition as in EMD. Fortunately, in most cases encountered, the leakage is small.

7. The Hilbert spectrum

Having obtained the intrinsic mode function components, we will have no difficulties in applying the Hilbert transform to each component, and computing the instantaneous frequency according to equation (3.4). After performing the Hilbert transform on each IMF component, we can express the data in the following form:

\[ X(t) = \sum_{j=1}^{r} a_j(t) \exp \left( i \int \omega_j(t) \, dt \right). \]  

(7.1)

Here we have left out the residue, \( r_n \), on purpose, for it is either a monotonic function, or a constant. Although the Hilbert transform can treat the monotonic trend as part of a longer oscillation, the energy involved in the residual trend could be overpowering. In consideration of the uncertainty of the longer trend, and in the interest of information contained in the other low-energy and higher-frequency components, the final non-IMF component should be left out. It, however, could be included, if physical considerations justify its inclusion.

Equation (7.1) gives both the amplitude and the frequency of each component as functions of time. The same data if expanded in Fourier representation would be

\[ X(t) = \sum_{j=1}^{\infty} a_j e^{i\omega_j t}, \]  

(7.2)

with both \( a_j \) and \( \omega_j \) constants. The contrast between equations (7.1) and (7.2) is clear: the IMF represents a generalized Fourier expansion. The variable amplitude and the instantaneous frequency have not only greatly improved the efficiency of the expansion, but also enabled the expansion to accommodate non-stationary data. With IMF expansion, the amplitude and the frequency modulations are also clearly separated. Thus, we have broken through the restriction of the constant amplitude and fixed-frequency Fourier expansion, and arrived at a variable amplitude and frequency representation. This expression is numerical. If a function is more desired, an empirical polynomial expression can be easily derived from the IMFs.

Equation (7.1) also enables us to represent the amplitude and the instantaneous frequency as functions of time in a three-dimensional plot, in which the amplitude can be contoured on the frequency–time plane. This frequency–time distribution of the amplitude is designated as the Hilbert amplitude spectrum, \( H(\omega, t) \), or simply Hilbert spectrum. If amplitude squared is more desirable commonly to represent energy density, then the squared values of amplitude can be substituted to produce the Hilbert energy spectrum just as well.

Various forms of Hilbert spectra presentations can be made: the colour coded maps and the contour maps all with or without smoothing. The Hilbert spectrum in the colour map format for the wind data is given in figure 8. This spectrum gives a very different appearance in comparison with the corresponding wavelet spectrum shown in figure 9. While the Hilbert spectrum appears only in the skeleton form with emphasis on the frequency variations of each IMF, the Wavelet analysis result...
gives a smoothed energy contour map with a rich distribution of higher harmonics. If a more continuous form is preferred, two smoothing methods can be applied. The first one is to use a weighted spatial filter with which to average over a range of cells. However, such a smoothing degrades both frequency and time resolutions. If we applied this approach with large enough spatial averaging, we would get a result similar to what the wavelet analysis would give. Even here, we still will not be encumbered by the many spurious harmonics as in the Fourier-based wavelet analysis. For example, a 15 × 15 weighted Gaussian filter will give the smoothed spectrum as in figure 10, which gives a better resemblance to the wavelet result, though not necessarily a better physical interpretation than the original, for the information content has been degraded. In the smoothed form, however, the energy density and its trends of evolution as functions of frequency and time are easier to identify. In general, if more quantitative results are desired, the original skeleton presentation is better; if more qualitative results are desired, the smoothed presentation is better. As a guide, the first look of the data is better in the smoothed format.

The alternative to the spatial smoothing is to select a lower frequency resolution and leave the time axis undisturbed. The advantage of this approach is the preservation of events’ locations, but gives a more continuous frequency variation. Furthermore, with low-frequency resolution, we also save some computation time. The optimal frequency resolution in the Hilbert spectrum can be computed as follows.

Let the total data length be T, and the digitizing rate be Δt. Then the lowest frequency one can extract from the data is 1/T Hz, which is also the limit of frequency resolution for the data. The highest frequency one can extract from the data is 1/(nΔt) Hz, in which n represents the minimum number of Δt needed to define the frequency accurately. Since the Hilbert transform defines instantaneous frequency by differentiation, we do need more data points to define an oscillation. The absolute minimum number of data points is five for a whole sine wave. We do not need a whole sine wave to define its frequency, but we do need many points within any part of the wave to get a stable derivative. Therefore, the maximum number of the frequency cells, N, of the Hilbert spectrum should be

\[ N = \frac{(1/n\Delta t)}{(1/T)} = \frac{T}{n\Delta t}. \]

(7.3)

In order to make the derivative stable, we have averaged over three adjacent cell values for the final presentation. The wind data, digitized at a rate of 0.01 s, has a total length of 30 s. Therefore, the highest frequency we can extract is 25 Hz. The total cell size could be 600 but they have been averaged to 200 as used in figure 8.

With the Hilbert spectrum defined, we can also define the marginal spectrum, h(ω), as

\[ h(\omega) = \int_0^T H(\omega, t) \, dt. \]

(7.4)

In figure 11 the solid line gives the corresponding marginal spectrum of the Hilbert spectrum given in figure 8. The marginal spectrum offers a measure of total amplitude (or energy) contribution from each frequency value. It represents the cumulated amplitude over the entire data span in a probabilistic sense. As pointed out by Huang et al. (1996), the frequency in either H(ω, t) or h(ω) has a totally different meaning from the Fourier spectral analysis. In the Fourier representation, the existence of energy at a frequency, ω, means a component of a sine or a cosine wave persisted
through the time span of the data. Here, the existence of energy at the frequency, \( \omega \), means only that, in the whole time span of the data, there is a higher likelihood for such a wave to have appeared locally. In fact, the Hilbert spectrum is a weighted non-normalized joint amplitude–frequency–time distribution. The weight assigned to each time–frequency cell is the local amplitude. Consequently, the frequency in the marginal spectrum indicates only the likelihood that an oscillation with such a frequency exists. The exact occurrence time of that oscillation is given in the full Hilbert spectrum. The corresponding Fourier spectrum of the wind data is also given in figure 11 in a dotted line. There is little similarity between the Fourier spectrum and the marginal spectrum. While the Fourier spectrum is dominated by the DC term because of the non-zero mean wind speed, the marginal spectrum gives a nearly continuous distribution of energy. The Fourier spectrum is meaningless physically, for the data is not stationary.

In addition to the marginal spectrum, we can also define the instantaneous energy density level, IE, as

\[
IE(t) = \int_\omega H^2(\omega, t) \, d\omega. \tag{7.5}
\]

Obviously, this IE also depends on time; it can be used to check the energy fluctuation.

With the energy–frequency–time distribution given, we can discuss the definition of stationarity quantitatively now. As we stated in the introduction, the classic defi-
The variation of the degree of stationarity and the degree of statistical stationarity are depicted in the figure. The DSS, time averages of 10, 50, 100 and 300 have been used. The DSS indeed decreases with the length of averaging especially in the high-frequency range.

Definitions of stationarity given in equations (1.1) and (1.2) are dichotomous: a process is either stationary or non-stationary, the description is only qualitative. Such definitions are both stringent and useless at the same time: few data sets can satisfy the rigour of these definitions; consequently, no one even bothers using them for checking the stationarity. As a result, data as non-stationary as earthquake and seismological signals are routinely treated as stationary (see, for example, Hu et al. 1996). Sometimes, for some obvious non-stationary data, two less stringent definitions are invoked: the piecewise stationary; and the asymptotically stationary. These definitions are still dichotomous. To quantify the statistical processes further, an index is needed to give a quantitative measure of how far the process deviates from stationarity; a prerequisite for such a definition is a method to present the data in the frequency time space. Now, having established the Hilbert spectrum, we can introduce the index of stationarity as follows.

The first step to define the degree of stationarity, \( DS(\omega) \), is to find the mean marginal spectrum, \( n(\omega) \), as

\[
n(\omega) = \frac{1}{T} b(\omega).
\]

Then, the degree of stationarity is defined as

\[
DS(\omega) = \frac{1}{T} \int_0^T \left( 1 - \frac{H(\omega, t)}{n(\omega)} \right)^2 dt,
\]

in which the integrand is similar to the intermittency defined in the wavelet analysis. We decided to adopt the integration representation because it gives a quantitative

Figure 13. The intermittence and the DSS. (a) A linear plot of the DSS with 100 time step average. Energy at 7 and 17 Hz are highly energetic and intermittent. (b) A section of wind data demonstrating the presence of energetic and intermittent oscillations at 7 and 17 Hz.

measure of the whole data set. Obviously, for a stationary process, the Hilbert spectrum cannot be a function of time. Then, the Hilbert spectrum will consist of only horizontal contour lines; $DS(\omega)$ will then be identically zero. Only under this condition will the marginal spectrum be identical to the Fourier spectrum, then the Fourier spectrum will also make physical sense. If the Hilbert spectrum depends on time, the index will not be zero, then the Fourier spectrum will cease to make physical sense. The higher the index value, the more non-stationary is the process. The $DS$ for the wind data is shown in figure 12. As the index shows, the data are highly non-stationary especially for the high-frequency components.

Equation (7.7) defines the stationarity as a function of frequency. This is necessary, for certain frequency components can be non-stationary while other components remain stationary. An example is the sporadic riding wind waves on otherwise uniform swell: the low-frequency swell component is stationary, while the high-frequency wind waves are intermittent, and hence non-stationary.

The degree of stationarity can also be a function of time implicitly, for the definition depends on the time length of integration in equations (7.6) and (7.7). Therefore, a process can be piecewise stationary. On the other hand, for a singular outburst in an otherwise stationary signal, the process can be regarded as almost stationary if we take a long time integral, but non-stationary if we look at the immediate neighbourhood of the outburst. Stationarity can be a complicated property of the process: for any data shorter than a typical long wave period, the process may look transient, yet as the data length gets longer, the process can have many longer wave periods and becomes stationary. On the other hand the data can be locally stationary while

in a long time sense non-stationary. An index is therefore not only useful but also necessary to quantify the process and give a measure of the stationarity.

The degree of stationarity defined in equation (7.7) can be modified slightly to include the statistically stationary signals, for which the degree of stationarity, $DSS(\omega, \Delta T)$, is defined as

$$DSS(\omega, \Delta T) = \frac{1}{T} \int_0^T \left( 1 - \frac{H(\omega, t)}{n(\omega)} \right)^2 \, dt,$$

(7.8)

where the overline indicates averaging over a definite but shorter time span, $\Delta T$, than the overall time duration of the data, $T$. For periodic motions, the $\Delta T$ can be the period. Such a time scale, however, is hard to define precisely for high-dimensional non-stationary dynamic systems. Even with this difficulty, the definition for $DSS$ could be more useful in characterizing random variables from natural phenomena.

Obviously, $DSS$ will depend on both frequency and the averaging time span. For the wind data taken as an example, the $DSS$ with $\Delta T$ as a parameter is given in figure 12 with $\Delta T = 10, 50, 100$ and 300 time steps, respectively. The results show that while the high-frequency components are non-stationary, they can still be statistically stationary. Two frequency bands at 7 and 17 Hz are highly non-stationary as the $DSS$ averaged at 100 time steps shows in figure 13. These components are intermittent as can be seen in the IMF components and the marginal spectrum. A section of the original wind data is also plotted in figure 13b to demonstrate that there are indeed prominent 7 and 17 Hz time scale oscillations.

8. Validation and calibration of the Hilbert spectrum

Through the empirical mode decomposition and the associated Hilbert spectral analysis we obtained the probabilistic Hilbert spectrum representation of the non-stationary data. Now, we will validate the approach and the results, and calibrate its fidelity compared with the best existing method, the wavelet analysis. Before proceeding further, let us first examine the physical meaning of EMD.

Historically, there are two important methods to expand a function in series: the Taylor and the Fourier expansions, both powerful but based on totally different approaches. The Taylor series for a function, $f(t)$, is expanded near a point, $t_0$, as

$$f(t+t_0) = f(t_0) + f'(t_0) (t-t_0) + f''(t_0) \frac{1}{2} (t-t_0)^2 + \cdots + f^{(n)}(t_0) \frac{1}{n!} (t-t_0)^n + \cdots,$$

(8.1)

in which the primes and $(n)$ indicate the order of differentiation. The Fourier expansion for the same function is expanded globally as,

$$f(t) = \sum_{j=0}^{n} a_j e^{i\omega_j t},$$

(8.2)

in which the coefficients, $a_j$, are given by

$$a_j = \frac{1}{2\pi} \int_t f(t) e^{-i\omega_j t} \, dt.$$  

(8.3)

From equations (8.1) and (8.2), the different mathematical principles in deriving the expansions are quite clear. While the Taylor expansion is based on the properties of
the function at a point through differentiation, the Fourier expansion is based on the properties of the function over the whole time span through integration. Although both expansions are valid within the range of convergence, there are different requirements on each function: one of the necessary conditions for the Taylor expansion to exist is that the function has to be analytic to the nth order. Such a restriction is not necessary for the Fourier expansion, which can be applied to functions with denumerable finite jumps. The Fourier expansion introduces a powerful idea: to expand a function with any orthogonal and complete basis. Therefore, there are infinitely many ways to decompose a signal. For example, a signal can be decomposed into simple harmonic components with the Fourier transform. By the same token, the same signal can also be decomposed into Hermite polynomials, or some other special functions such as Legendre or Laguerre polynomials. In fact any collection of orthogonal functions can serve as the basis for a decomposition for a linear system. Once we have a complete and orthogonal basis, the expansion is mathematically correct even for a transient nonlinear system, as testified by the results produced by the classical perturbation analysis for nonlinear differential equations. Whether this mathematically correct expansion also makes physical sense is entirely a separate, but crucial, problem. In general, we believe the answer provided by any linear expansion for a nonlinear system, such as the answer obtained by the perturbation method, should not make physical sense. This conclusion can be argued from two different direc-

Nonlinear and non-stationary time series analysis

Figure 15. Validation of the intrawave frequency modulation with Stokian waves. (a) The profile of a second-order Stokes wave in deep water with sharp crests and rounded-off troughs in comparison with the pure cosine waves. (b) The IMF generated by the Stokes wave, there is only one component; the constant off-set is not shown.

...tions: first, a nonlinear system does not admit an answer by superposition; second, the perturbation method is forcing a linear approximation on a nonlinear system. Both of the above arguments are true; therefore, it is fortuitous to have the sum of the linear system to approximate the full nonlinear system. It is entirely different to also ask each individual component in the linear system to have physical meaning related to the full nonlinear system.

EMD expansion is at the same time both different from and similar to the above expansions. It is based on both integration, as in the Hilbert transform given in equation (3.1), and differentiation, as in the instantaneous frequency given in equation (3.4). The integration in the Hilbert transform is not exactly global, for it is the convolution of the function with $1/t$, which makes the result extremely local. Essentially, the Hilbert transform gives the best fit with a sinusoidal function to the data weighted by $1/t$.

Let us compare how local the Hilbert transform can be with the result from wavelet analysis by considering an isolated sine wave given in figure 14a. With wavelet analysis we get the spectrum in figure 14c, in which the event is well defined on the time axis by the high-frequency components, even though the event is a low-frequency wave. In the result, neither the energy density nor the frequency is well localized; they give a counter-intuitive interpretation of the wavelet spectrum: to look for definition of a low-frequency event in the high-frequency range. When the same data are treated by the Hilbert spectral analysis, we have the result in figure 14b, in which the

energy is well localized in both frequency and time domains. This simple example illustrates the unique property of the Hilbert spectrum in elimination of the spurious harmonic components to represent the non-stationary data.

Next, let us examine the validity and the implication of the instantaneous frequency for nonlinear data through the idealized Stokes wave in deep water, which is a classic example of using harmonic components to represent nonlinear wave form distortion (see, for example, Lamb 1932). It is also the first successful application of the perturbation method to solve a nonlinear analytic equation system for a natural phenomenon. To second order, the profile is given by

$$X(t) = \frac{1}{2}a^2k + a \cos \omega t + \frac{1}{2}a^2k \cos 2\omega t + \cdots,$$

in which $a$ is the amplitude and $k$ is the wave number. For $a = 1$, $ak = 0.2$, and $\omega = \frac{1}{12}$ Hz, the wave-profile is shown in figure 15a. Because of the harmonic distortion, the wave form shows sharpened crests and rounded-off troughs, which make the crests and troughs asymmetric with respect to the mean surface. Processed with EMD, these data yield only one IMF component, shown also in figure 15b, with a constant offset component not shown. The Stokes wave is deformed due to the harmonic distortion; it is asymmetric with respect to the mean, while the IMF is symmetric. Although this wave has only one characteristic scale, the wavelet analysis result shown in figure 16 gives two bands of energy. The IMF data can be processed immediately to give the Hilbert spectrum as shown in figure 17, which has only one frequency band centred around 0.03 Hz, the fundamental frequency of the wave train.

The Hilbert spectrum from the IMF with forced symmetry of the profile in which the intrawave frequency modulation interpretation of the Stokes wave is clear. In the Hilbert spectrum the frequency variation is bounded in a narrow range around 0.03 Hz with no harmonics.

but there is an intrawave frequency modulation with a magnitude range of 0.02–0.04 Hz. This intrawave frequency modulation has been totally ignored in the past, for the traditional definition of frequency is based on the reciprocal of periodicity. Consequently, the notion of intrawave frequency modulation cannot be represented or discussed. Is the intrawave frequency modulation meaningful? What is its physical significance? To answer these questions, let us consider a simple model wave given by

\[ X(t) = \cos(\omega t + \epsilon \sin \omega t). \]  

Figure 18a gives the profile of this wave together with the cosine wave given by the dotted line. This wave shows the harmonic distortion similar to the Stokes wave given in figure 15c: sharpened crests and rounded-off troughs. In the past, this kind of wave form has been treated as a typical case of harmonic distortion. Indeed, the second-order approximation of equation (8.5) is

\begin{align*}
X(t) &= \cos(\omega t + \epsilon \sin \omega t) \\
&= \cos \omega t \cos(\epsilon \sin \omega t) - \sin \omega t \sin(\epsilon \sin \omega t) \\
&\approx \cos \omega t - \epsilon \sin^2 \omega t \approx (1 - \frac{1}{2} \epsilon) \cos \omega t + \frac{1}{2} \epsilon \cos 2\omega t.
\end{align*}

This expression is very similar to the one given in equation (8.4).

Equation (8.6), however, is only an approximation to the full expression of equa-
Figure 18. Model wave with intrawave frequency modulation to simulate the Stokes wave. (a) Profile of the model wave showing the same characteristics as the Stokes wave: sharp crest and rounded-off trough. A pure cosine wave is shown by a dotted line for reference. (b) Frequency of the model wave computed from Hilbert transform (solid line) and equation (8.8) (dotted line). The agreement is excellent for this case.

Equation (8.5), which satisfies the following highly nonlinear differential equation:

\[
\frac{d^2 x}{dt^2} + (\omega + \epsilon \omega \cos \omega t)^2 x - \omega^2 \sin \omega t (1 - x^2)^{1/2} = 0. \tag{8.7}
\]

Although this equation bears no relationship to the system generating the Stokes wave, the example illustrates that harmonic components are the results of using a linear system to simulate the nonlinear one. Although the final solution form is similar to the real one, the harmonic components have no physical meaning other than providing better mathematical approximation.

As given in equation (8.5), the function is clearly a frequency modulated wave with the modulating frequency equal to that of the main oscillating frequency. The frequency given by the classic wave theory for this model wave is

\[
\frac{d \theta}{dt} = \Omega = \omega (1 + \epsilon \cos \omega t). \tag{8.8}
\]

With \( \omega = 1 \) Hz and \( \epsilon = 0.5 \), the frequency values are plotted in figure 18, with the solid line representing the values given by the Hilbert transform and equation (3.4), and the dotted line representing the values given in equation (8.8). Both methods give similar variation of the frequency with only minor deviation at the extreme values.

The agreement between the two methods confirms that a deformed sinusoidal wave

\[\dagger\] Note added in proof: It can also be shown that a general form of equation (8.5), \( X(t) = \cos(\omega t + \phi(t)) \), satisfies a linear differential equation with variable coefficients.

Figure 19. Comparison of the wavelet spectrum with the Hilbert spectrum for the model intrawave frequency modulated wave. (a) The Morlet wavelet spectrum of the model wave showing a rich complement of harmonics to simulate the wave-profile deformation due to the intrawave frequency modulation; the main energy is located around 0.03 Hz, with the harmonic at 0.06 Hz as expected. (b) The Hilbert spectrum for the same data showing the intrawave frequency modulation as the model wave simulated. These results are almost identical to those shown in figures 16 and 17.