Article ID: 1000-9116(2005)05-0602-09

Causes and solutions of overshoot and undershoot and end swing in Hilbert-Huang transform*

LONG Si-sheng (龙思胜) ZHANG Tie-bao (张铁宝) LONG Feng (龙 锋)

Earthquake Administration of Sichuan Province, Chengdu 610041, *China*

Abstract

There are overshoot and undershoot phenomenon and end swing phenomenon in the cubic spline fitting in Hilbert-Huang transform. The two problems influence data quality of the empirical mode decomposition seriously. The cubic spline fitting has been analysed, and the causes of producing the overshoot and undershoot phenomenon and the end swing phenomenon have been pointed out in this paper. Two new methods of cubic spline fitting and sine spline fitting and the new technique of handling the end points of the original data curve can completely remove the overshoot and undershoot phenomenon and the end swing phenomenon on the condition of unchanging original data, and have the advantages of the continuous fitting functions and its continuous one order derivative, the simple and convenient calculations, the small calculation amount and the easy work on it.

Key words: Hilbert-Huang transform; cubic spline fitting; sine spline fitting; overshoot and undershoot; end swings

CLC number: 315.01 **Document code**: A

Introduction

Hilbert-Huang transform (HHT) is a great break in processing nonlinear and non-stationary data and can be successfully used in many science domains. There are mainly two parts in this method. The first part is to decompose the original data into several intrinsic mode functions (IMF) with the empirical mode decomposition (EMD). IMF components are derived from the original data directly according to the local characteristics in the data under some rules, so that IMF are posteriori, adaptive, complete and almost orthogonal and can be used in the research of nonlinear and non-stationary data. Under the decomposition conditions, these IMFs are the intrinsic and imbedded oscillatory modes in the data and the wave trains approaching to harmonic wave with the intrawave amplitude and frequency modulations, for this reason, which have fine Hilbert transform property and the physical information of the data. The second part is to construct the energy-frequency-time distributions (Hilbert spectrum) with the Hilbert transform to IMFs, which can reveal the time and frequency position of each event for the detailed research of the process. (Huang *et al*, 1998, 1999, 2001).

To achieve empirical mode decomposition in HHT, the cubic spline fitting are used in con-

* **Received date**: 2004-07-05; **revised date**: 2005-06-27; **accepted date**: 2005-06-27.

Foundation item: The Foundation Research and Development Programs of China (2004CB418404).

E-mail of the first author: longsisheng@hotmail.com

structing upper and lower envelopes. This fitting is one of the most applied fitting methods. The cubic spline fitting is classified into three types: intrinsic boundary, natural boundary and periodical boundary (GUAN and CHEN, 2002), and there is also the viewpoint on five types of this methods (LI and QI, 1979). But there are the overshoot and undershoot and the end swing problems in the cubic spline fitting method used by Huang *et al*, which may influence the results of EMD seriously (Huang *et al*, 1998). None have researched the overshoot and undershoot problem up to now. Some researchers of China have discussed the end swing problem, but they have not explained the cause of the end swing phenomenon, and the solution methods pointed by them are not ideal for the end swing problem because these methods will change the original data before handling the original data (SHI and LUO, 2003; LUO and SHI, 2003; GAI and ZHANG, 2002).

The causation of the overshoot and undershoot phenomenon has been analysed and one new convenient method of cubic spline fitting with the advantages of a small amount of calculation has been given out which can clean up the overshoot and undershoot phenomenon in this paper. And the causation of the end swing phenomenon has been analysed and a new technique has been also given out which can remove the end swing without any change in the original data. The combination of the method and the technique in the empirical mode decomposition will greatly improve the data processing quality and establish a reliable basis for EMD.

1 Cubic spline fitting method

1.1 Mathematical meaning of the overshoot and undershoot

No strict definition of the overshoot and undershoot has been given in Huang's papers. According to Huang's meaning about the overshoot and undershoot and our realization in the data processing practice, we think the overshoot and undershoot phenomenon is that some of the fitted numbers are bigger than the bigger one or smaller than the smaller one of the couple of the two adjacent data points in the closed subinterval which is closed by the two data points, *i.e.* some of the fitted numbers go up or down beyond the range of the two adjacent data points in the vertical directions. In order to remove the overshoot and undershoot phenomenon from the fitted curve, it is the requirement that all fitted numbers are not bigger than the bigger one or smaller than the smaller one of the two adjacent data points in the every closed subinterval which is closed by the two data points.

1.2 The requirement of EMD for fitting function

According to Huang′s meaning about EMD and our realization in the data processing practice, it is the aim of EMD to decompose the data curve into several curves of the different components with the narrower frequency subintervals which are not overlap each other. Each of these components is ″purer″ enough to be handling easily and to include the partial physical meaning in the original data set. And the data set can be directly obtained from all the components by adding all the components each other.

The major local extreme value points are the ones that represent the major characters of the data curve. The important tools for EMD are the upper and lower envelopes and their average value curve, which are fitted to these major local extreme points. Thus, the disturbance of the overshoot and undershoot and the end swing must be removed from the fitted curves if we do not want to reduce the handled data quality after fitting. On the other way, the requirement of the degree of smoothing of the fitted curves can be reduced lightly. Therefore it is the requirement of the fitting function that the fitting function and its first-order derived function should be continuous at every nodal point, the first-order derived function should equal to zero at every nodal point, nothing required for the second order derived function, and there is not the overshoot and undershoot phenomenon in each subinterval.

1.3 The limitation of the existing cubic spline fitting method

The existing cubic spline fitting method has been developing for shipping, automobile and aircraft industry, and it has different emphasis in different fields. It is its requirements that the fitted curve has excellent smoothing degree in the necessary interval, the overshoot and undershoot are not been limited, *i.e.* that the fitting function, its first and second order derived functions are continuous at every inner nodal point, and three requirements at the first and the last data points are the fixed, natural and periodic boundary conditions respectively (GUAN and CHEN, 2002). Because the fitting functions, its first and second derived functions on the left and right sides of every inner nodal point are respectively equal to each other at the inner nodal point (continuity condition), it is almost impossible that the first order derived function equals to zero at the inner nodal points (except the data set is given specially). The overshoot and undershoot phenomenon must adhere the fitted curves. But it is the necessary condition for the local extreme points of the fitted curves that the first order derived function equals to zero. Only the local extreme value points can confine the lengthening ranges of the curves at up or down directions in the neighborhood of these points. The directions of the overshoot and undershoot may be upward or downward, the amplitudes may be big or small, and the directions and amplitudes may be changed according to the variable combinations of the numerical values of the adjacent data points in the data set, and it is difficult to accurately estimate the influence degree of the handled data by the overshoot and undershoot (Figures 2, 3, 4). Therefore it is necessary to put forward new fitting conditions.

The function building and resolving of the existing cubic spline fitting method must resolve the simultaneous equations with matrices; its calculation is quite complicated. The larger the amounts of the data sets are, the higher the orders of the matrices are, and the larger the amounts of calculating and compiling are.

1.4 The cubic spline fitting method without the overshoot and undershoot

We attempted to find a simpler cubic spline fitting method without the overshoot and undershoot. There are only two extreme value points in a cubic spline fitting curve; the curve segments between the two extreme value points are monotone increasing (the coefficient of the three time term a<0) or monotone decreasing (the coefficient of the three time term a>0); and the two extreme value points are also the greatest and smallest value points of the curve segments; the curve segment is limited between the two extreme value points at the vertical and horizontal directions without the overshoot and undershoot phenomenon (Figure 1). If the fitting function had been built only with this segments of the cubic spline curve, we can set up a cubic spline fitting method without the overshoot and undershoot.

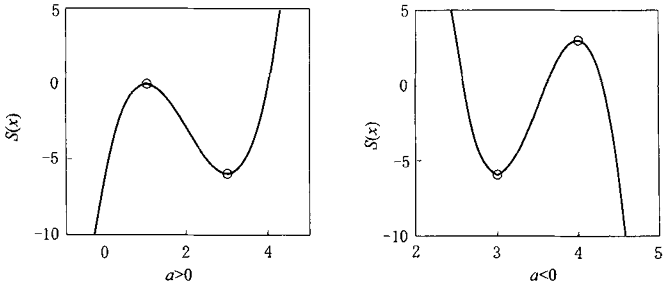

Figure 1 Cubic spline fitting function curve and its extreme value points

Suppose a segmentation $p=x_0<x_1<\cdots<x_n=q$ is given in a range $[p, q]$, it is the necessary conditions of the cubic spline fitting function that the left and right limits of the fitting function at the inner nodal points are equal to each other and to the numbers of the inner nodal point given beforehand, the left and right limits of the first order derivatives are equal to each other and to zero at all inner nodal points; and the values of the fitting functions at the first and last points of the data set are equal to the numbers of the first and last data points given beforehand and the values of the first order derivatives are equal to zero at the first and last data points. That is

$$\begin{cases} S_i(x_{i+1}-0)=f(x_{i+1}) \\ S_i(x_i+0)=f(x_i) \\ S_i'(x_{i+1}-0)=0 \\ S_i'(x_i+0)=0 \end{cases} \tag{1}$$

and $i=1, 2, \cdots, n$.

There are n subintervals and n cubic spline fitting equation groups in the curve interval with $(n+1)$ nodal points (include the first and last data points). These nodal points bring forward $4n$ conditions. Every group of 4 conditions in every subinterval can just resolve the 4 unknown coefficients of cubic spline fitting function $S_i(x)$ in this subinterval. The equality of the left and right extreme values of the fitting functions and its first order derivatives at the inner nodal points ensure the continuity of the fitting functions and its first order derivatives. In every subinterval, $S_i(x)$ is a monotone function, its two extreme points are not only its two extreme value points, but also the largest and smallest points of $S_i(x)$, so that all values of $S_i(x)$ are confined to the range between $f(x_{i-1})$ and $f(x_i)$ and the overshoot and under shoot will not appear at all. On the other hand, it is not necessary to resolve the simultaneous equations with the all unknown coefficients of n subintervals, and it is necessary to resolve the one variant and three time equation of every subinterval simply and conveniently.

According to the unique existence theorem of the cubic spline fitting (interpolating) function $S(x)$, the supposition below is suitable (GUAN and CHEN, 2002, LI and QI, 1979):

$$\begin{cases} S(x)=\sum_{i=1}^{n} S_i(x) \\ S_i(x)=a_i x^3+b_i x^2+c_i x+d_i \end{cases} \tag{2}$$

in which $x_i \le x \le x_{i+1}$, $i=1, 2, \cdots, n$.

We can obtain the equation group (3) from the conditions (1).

$$\begin{cases} S_i(x_i)=a_i x_i^3+b_i x_i^2+c_i x_i+d_i=f(x_i) \\ S_i(x_{i+1})=a_i x_{i+1}^3+b_i x_{i+1}^2+c_i x_{i+1}+d_i=f(x_{i+1}) \\ S_i'(x_i)=3a_i x_i^2+2b_i x_i+c_i=0 \\ S_i'(x_{i+1})=3a_i x_{i+1}^2+2b_i x_{i+1}+c_i=0 \end{cases} \tag{3}$$

The equation group (4) can be get from the equation group (3):

$$\begin{cases} a_i = -2L_i \\ b_i = 3(x_i + x_{i+1})L_i \\ c_i = -6x_{i+1}x_iL_i \\ d_i = f(x_i) + x_i^2(3x_{i+1} - x_i)L_i \end{cases} \tag{4}$$

and $L_i = \dfrac{f(x_{i+1}) - f(x_i)}{(x_{i+1} - x_i)^3}$ $(i=1, 2, \cdots, n)$.

Five times of plus and reduce operation, 13 times of multiply and divide operation are necessary to get 4 coefficients in every subinterval; $5n$ times of plus and reduce operation, $13n$ times of multiply and divide operation are necessary to get $4n$ coefficients in all fitted curve.

If $f(x_{i+1})-f(x_i)=0$, $S_i(x)$ is a horizontal line segment; if $f(x_{i+1})-f(x_i)>0$ (*i.e.* $a<0$), $S_i(x)$ is a monotone increasing curve segment; if $f(x_{i+1})-f(x_i)<0$ (*i.e.* $a>0$), $S_i(x)$ is a monotone decreasing curve segment. The fitted curve segment in every subinterval is a sect of a one variant and three time curve.

In addition, if the equation groups (3) and (4) are exchanged for the group (5) below, the fitted curve segment in every subinterval will be a sect of a sinusoidal curve (the total fitted curve in all interval will be a more approximated sinusoidal curve). And there are these advantages in fitting method that the fitting function and its first order derivatives are continuous; no the overshoot and undershoot appear in the fitted curves; the calculating is simple and convenient; the amount of calculating is quite small.

$$\begin{cases} h_i = \dfrac{f(x_{i+1}) - f(x_i)}{2} \\ S_i(x) = f(x_i) + h_i\left[1 + \sin\left(\dfrac{k-1}{k_2-1} - \dfrac{1}{2}\right)\pi\right] \end{cases} \tag{5}$$

and k_2 is the difference between the position numbers of the latter and the former extreme value points of every subinterval in the entire fitted interval, $k=1, 2, \cdots, k_2$.

1.5 Calculated example

12 couples of data given, calculated results are listed in Tabal 1, and the fitted curve is in Figure 2. From Figure 2, 5 characteristics can be evidently seen, *i.e.* ① the first and last data points and the inner nodal points are the data points; ② the first and last data points and the inner nodal points are also the zero points of the first order derivative, and may be the local extreme value points; ③ The curve segment in every subinterval is monotone; ④ every curve segment is limited between its two extreme points and there is not the overshoot and undershoot in every subinterval; ⑤ the method do not increase any new extreme value points in all fitted curve.

The envelope curves formed with the fitted curve can satisfy the necessity of EMD because of the five characteristics. The estimation of error bound is not given in this paper because of no analytic expressions of the original data sets. It is very easy to prove the monotonicity and limitation of the fitted function in every subinterval, so the demonstration is omitted.

The comparison between two fitting methods is diagramed in Figure 2 (with the same data set in Table1). It is very obvious that the new method do not make any overshoot and undershoot in the fitted curve at all, while the existent method brings some serious phenomena of the overshoot and undershoot to the fitted curve. The overshoot and undershoot in different subintervals

Table 1 Data, calculated interspace and result numbers of the example

i	x_i	f_i	$f_{i+1}-f_i$	$(x_{i+1}-x_i)^3$	L_i	$x_{i+1}+x_i$	$x_i^2\ (3x_{i+1}-x_i)$	a_i	b_i	c_i	d_i
1	–3	0	–2	8	–0.25	–4	0	0.50	3.00	4.5	0.00
2	–1	–2	2	8	0.25	0	4	–0.50	0.00	1.5	–1.00
3	1	0	–6	8	–0.75	4	8	1.50	–9.00	13.5	–6.00
4	3	–6	9	1	9.00	7	81	–18.0	189	–648	723
5	4	3	–2	1	–2.00	9	176	4.0	–54.0	240	–349
6	5	1	–3	27	–0.11	13	475	0.22	–4.33	26.67	–51.8
7	8	–2	3	8	0.38	18	1 408	–0.75	20.3	–180	526
8	10	1	2	1	2.00	21	2 300	–4.00	126	–1 320	4 601
9	11	3	0	64	0.00	26	4 114	0.00	0.00	0.00	3.00
10	15	3	2	1	2.00	31	7 425	–4.00	186	–2 880	14 853
11	16	5	–3	64	–0.05	36	11 264	0.09	–5.06	90.0	–523
12	20	2									

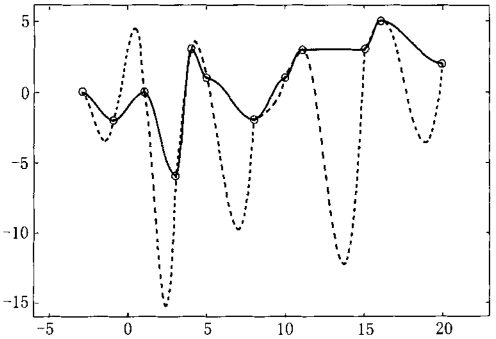

Figure 2 Comparison of the data points and the fitted curve

The solid line and dashed line are the fitted curves by the new and existing methods separately, the former do not has and the latter has the overshoot and undershoot

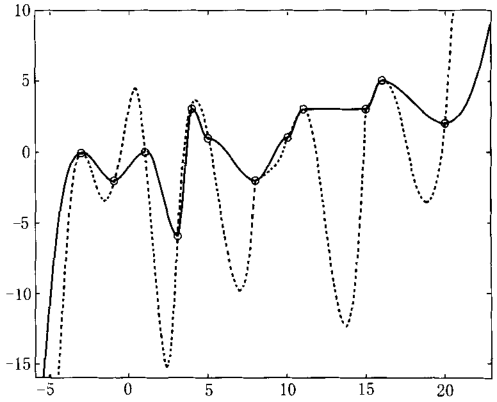

Figure 3 End swing phenomenon of the fitted curve

The solid line and dashed line are the fitted curves by the new and existing methods separately, two methods bring the end swings to the fitted curves

are almost different each from others; the directions and amplitudes are protean according to variations of the ratios of the differences of the vertical and horizontal coordinates of these contiguous data points; and the changes are very sensitive to every tittle of the variety of the original data. So it is difficult to estimate how much of the influence by the overshoot and undershoot in the handled data.

2 The end swing problem of the fitted curve

2.1 The end swing phenomenon

In Hilbert-Huang transform, the upper and lower envelope curves are constructed with the maximal value points and the minimal value points separately by the cubic spline fitting method. When the first segment from the first data point to the first maximal or minimal value point is been handling, there is no suitable fitting function to be utilized as a necessary tool. Only the segment of every fitting curve corresponding to the suitable subinterval accords with the fitting conditions, this fitted curve segment will continue to stretch beyond the suitable subinterval and to form two end swings at the left and right directions outside the suitable subinterval (Figure 1 and

Figure 3).If the upper or lower envelope curves never-ending stretch forward the beginning part and the last part of the data curve to obtain the complete fitted curve, the end swings must appear at two end-parts of the fitted curve. Both the new and the existing cubic spline fitting methods will form the end swing phenomenon in the fitted curves that will seriously depress the handled data quality as it has been exhibited in Figure 2 and Figure 3.

2.2 The appearances of the end swing phenomenon

The appearances of the end swing phenomenon have many forms (Figure 3 and Figure 4): the direction change of the beginning part of the fitted curve is determined by the difference σ between the vertical coordinates of the first and the second extreme value points: if $\sigma<0$, the swing is up (Figure 4a and 4b); if $\sigma>0$, the swing is down (Figure 4c); if $\sigma=0$, the swing is a horizontal line (Figure 4d); and the changing speed of the swing up or down is relative to the absolute values of the ratio of the differences of the vertical and the horizontal coordinates of the first and second extreme value points (Figure 4a and 4b). The swing phenomenon at the last part of the data curve is similar to the one at the beginning part. The appearances of the end swings depend on the local and specific data numbers at the beginning and last parts of the data set. Therefore the pictures produced by the end swings of different data sets are almost different, as the pictures of the overshoot and undershoot of the fitted curves depend on the specific structure of the data sets.

It is necessary to notice that the different fitting methods will make out the different pictures of the end swing. The variety regularity of the end swing is very obvious with the new fitting method without the overshoot and undershoot, as exhibited in Figure 4. There are more complex cases with the fitting methods bringing the overshoot and undershoot.

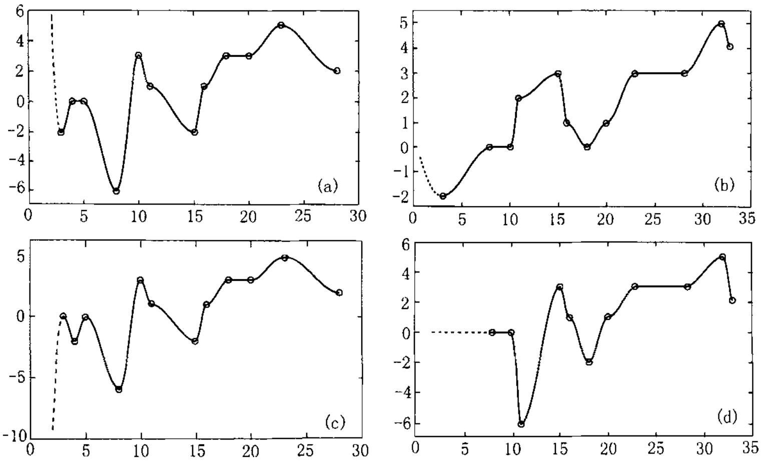

Figure 4 Changes of the end swings

The small circles are the nodal points, the solid curves are the fitted curves, the left dashed lines are the end swing lines. The direction of the end swing (up or down) depends on the difference of the vertical coordinates of the first and the second nodal points, the change speed of the end swing moving up or down correlates with the ratio of the differences of the vertical and horizontal coordinates of the first and the second nodal points

2.3 The resolution of the end swing problem

According to Figure 3 and Figure 4, it is the common characteristic of the end swing there is no a tip nodal point to control the end swing at the beginning tip of the data curve. Some methods have been brought forward to eliminate the end swing (GAI and ZHANG, 2002; SHI and LUO, 2003). Although the kind of methods has various forms, it is the pivotal part to supply the tip nodal point at the beginning tip of the data curve. To reduce the man-made interference in calculate process and improve quality of the handled result as possible, we recognize it is the precondition not to change the original data for eliminating the end swing phenomenon. Therefore the best technique is to make the first point of the data set as the first nodal point at the beginning tip, *i.e.* the zeroth maximal and the zeroth minimal points at the same time; and the last point of the data set is made as the zeroth maximal and the zeroth minimal points in reversed order at the same time. In other words, it is to separately increase a fitted curve segment and a fitting function in the beginning and the last subintervals of the data set at the same time. So that the end swing phenomenon will be eliminated from the fitted curves entirely and the upper and lower envelopes will connect with each other forming a closed curve loop. The technique will not bring any baneful influence on the upper and lower envelopes and their average value curve at all. The difference curve between the data curve and the average value curve fluctuates round the horizontal coordinate axis up and down, both the first and the last points of the difference curve are equal to zero at the horizontal coordinate axis. The numbers of the data points of the data curve, the envelope curves, the average value curve and the difference curve are all same, so that it is convenient to calculate these curves.

The effect of the technique is illustrated with two examples below. The data set in Figure 5 is one part of a earthquake record, the data set in Figure 6 is a composite imitation signal of a 10 Hz sine signal and a linear frequency modulation signal sin[2π*t*(10+10*t*)]. It is obvious that the effects are very good in two examples.

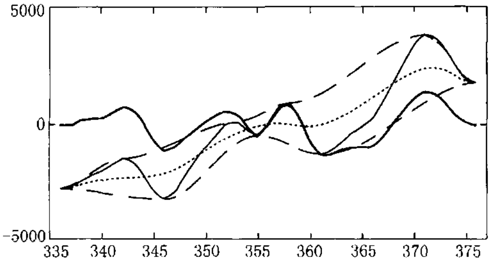

Figure 5 The fitting technique without the end swing (earthquake data)

Thin solid line is the data line, dashed line is the upper and lower envelope lines, dotted line is the average value line of the envelopes; thick solid line is the difference line between the data line and the average value line

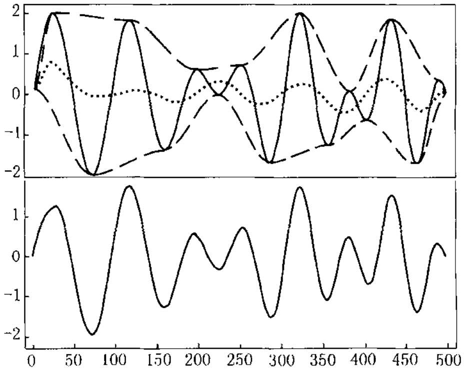

Figure 6 The fitting technique without the end swing (imitation signal)

Upper: solid line is the data line, dashed line is the upper and lower envelope lines, dotted line is the average value line of the envelopes; Lower: solid line is the difference line between the data line and the average value line

Although the advantages above are obvious, the technique has the shortages of compelling the first and the last data points to be the extreme value points and possibly changing the convergent tendency of the beginning and the last part of the fitted curve. The technique can not eliminate the end effect in processing.

3 Conclusions

The Hilbert-Huang transform is a important development in the domain of digital signal recently. The new method of cubic spline fitting in this paper is named "the zero point method of first order derivate". It is the advantages of the new method that the fitting function and its first order derivate function are continuous, that the overshoot and undershoot have been eliminated from the fitted curves, that the calculation is simple and convenient; that the amount of calculate is small. The handling technique eliminating the end swing from the fitted curves is called "the designative method of extreme endpoints", and it has the advantages that the technique do not change the original data; that the original data curve, the envelope curves, the average value curve and the difference curve have a same number of the data points; and that the calculation is simple and convenient. The combination of the methods and the technique can provide a reliable technical tool to decompose the original data into some of intrinsic mode functions (IMF) successfully in Hilbert-Huang transform.

References

GAI Qiang and ZHANG Hai-yong. 2002. Compare and study about several decomposition method of interval waves [J]. *System Engineering and Electronic Technology*, **2**: 57~59 (in Chinese).

GUAN Zhi and CHEN Jing-liang. 2002. *Numerical Calculation Methods* [M]. Beijing Tsinghua University Press, 126~134 (in Chinese).

Huang N E, Shen Z, Long S R *et al*. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series [J]. *Proc R Soc London (Series A)* **454**: 903~995.

Huang N E, Shen Z, Long S R. 1999. A new view of nonlinear water waves: The Hilbert spectrum [J]. *Annu Rev Fluid Mech*, **31**: 417~457.

LI Yue-sheng and QI Dong-xu. 1979. *Spline Function Method* [M]. Beijing: Science Press, 89~121 (in Chinese).

LUO Qi-feng and SHI Chun-xiang. 2003. Hilbert-Huang transform theory and the problem in calculation [J]. *Journal of Tongji University*, **31**(6): 637~640 (in Chinese).

Huang N E, Chern C C, Huang K *et al*. 2001. A new spectral representation of earthquake data: Hilbert spectral analysis of station TCU129, Chi-Chi, Taiwan, 21 September 1999 [J]. *Bulletin of the Seismological Society of America*, **91**(5): 1 310~1 338.

SHI Chun-xiang and LUO Qi-feng. 2003. Hilbert-Huang transform and wavelet analysis of time history signal [J]. *Acta Seismologica Sinica*, **16**(4): 422~439.