The Korean Journal of Applied Statistics (2011) DOI: http://dx.doi.org/10.5351/KJAS.2011.24.6.995
24*(6), 995–1006*

A Hilbert-Huang Transform Approach Combined with PCA for Predicting a Time Series

Minjeong Park[1]

[1]Department of Statistics, Seoul National University

(Received August 2011; accepted October 2011)

Abstract

A time series can be decomposed into simple components with a multiscale method. Empirical mode decomposition(EMD) is a recently invented multiscale method in Huang *et al.* (1998). It is natural to apply a classical prediction method such a vector autoregressive(AR) model to the obtained simple components instead of the original time series; in addition, a prediction procedure combining a classical prediction model to EMD and Hilbert spectrum is proposed in Kim *et al.* (2008). In this paper, we suggest to adopt principal component analysis(PCA) to the prediction procedure that enables the efficient selection of input variables among obtained components by EMD. We discuss the utility of adopting PCA in the prediction procedure based on EMD and Hilbert spectrum and analyze the daily worm account data by the proposed PCA adopted prediction method.

Keywords: Multiscale method, empirical mode decomposition, Hilbert spectrum, Hilbert-Huang transform, principal component analysis, vector AR model.

1. Introduction

Prediction has always been an object of attention in statistics. In a classical time series analysis for predicting, a model for given data is first constructed in the time domain and then the future values are forecasted; however, most models are built under certain assumptions. As a typical example, we can consider the autoregressive(AR) model that is steadily used for analyzing a time series; however, the stationarity assumption is required. In order to alleviate such assumptions for data, we can take two different approaches. One is to construct general models having less strict assumptions, for example autoregressive integrated moving average(ARIMA) model and the other one is to decompose the data into simple components. The method that we deal with in this paper is related to simple components.

Decomposing a time series into several components can be performed by Fourier transform in the frequency domain or by wavelet transformation in the time-frequency domain. Recently, empirical mode decomposition(EMD) was invented in Huang *et al.* (1998), which enables to decompose a time series into several components in the time domain. Note that the decomposition process is performed

[1]Postdoc, Department of Statistics, Seoul National University, San 56-1 Shillim-Dong, Gwanak-Gu, Seoul 151-747, Korea. E-mail: mj.park.stat@gmail.com

in the time domain, hence we do not need an inverse transformation for reconstructing the original data. The component extracted by EMD is termed as intrinsic mode function(IMF). Through the Hilbert transform, the sum of IMFs can be represented as a generalized Fourier expansion (Huang *et al.*, 1998) that means that the data might be decomposed into IMFs according to the spectral characteristics of the data. Based on multiscale methods such as wavelet analysis or EMD, a time series can be decomposed into several components that may be simpler than the original data.

Once the data are decomposed into simple components, it can be more feasible to analyze and predict each component by classical models. In Oh *et al.* (2009), a time series of Korean stock price index(KOSPI) 200 was analyzed by a multi-resolution approach that especially discussed the applications of EMD and the Hilbert spectrum in Huang *et al.* (1998). The applications of EMD in various fields were also briefly reviewed in Oh *et al.* (2009). In addition, a specific prediction procedure based on EMD and the Hilbert spectrum was proposed in Kim *et al.* (2008) that analyzed a real time series of cyber attacks. They suggested the following steps for prediction: 1) Obtain IMFs of a time series through the EMD process. 2) Select some IMFs based on the cumulative energy for IMFs. 3) Apply a vector AR model for the selected IMFs and polynomial regression to the EMD residue. 4) Forecast each selected IMF and the EMD residue, respectively. 5) Predict the time series by adding up all forecasting results. The prediction result by the above procedure was compared with those by exponential smoothing and the Holt-Winters model in Kim *et al.* (2008).

In this paper, we are interested in the usage of EMD and the Hilbert spectrum among the multiscale methods for predicting a time series. Naturally, we consider the procedure in Kim *et al.* (2008) for predicting a time series. Instead selecting meaningful IMFs based on the cumulative energy in Step 2); however, we propose a new method to decide input variables for the vector AR model based on the principal component analysis(PCA) in order to predict the time series more efficiently. Throughout two examples, we discuss the utility of adopting PCA in the prediction procedure based on EMD and the Hilbert spectrum. By the newly proposed method adopting PCA, we also analyze a real data set, the time series of cyber attacks in Kim *et al.* (2008).

This paper is organized as follows. Section 2 briefly introduces EMD and the Hilbert spectrum, and explains the procedure for predicting cyber-attacks in Kim *et al.* (2008). Section 3 addresses the reason to adopt PCA to the prediction procedure in Kim *et al.* (2008) as well as the newly proposed procedure that adopts PCA. The real worm count data that represents cyber attacks are analyzed in Section 4. Finally, Section 5 contains some concluding remarks.

2. The HHT-Based Prediction Procedure

2.1. Empirical mode decomposition

The EMD process, which was invented to analyze nonstationary and nonlinear signals in Huang *et al.* (1998), decomposes a signal into locally zero symmetric components, IMFs. The IMF is defined to be locally zero-symmetric in the meaning as follows: 1) The number of extrema and the number of zero crossings in the whole data set must either be equal or differ by one. 2) The mean value of the upper and lower envelopes at any point is zero. The upper (lower) envelope is defined by the spline interpolation with knots at the local maxima (minima) in the signal.

In order to extract an IMF from the given signal y, the EMD process composes the following steps for sifting: 1) Identify extrema in the signal. 2) Find the upper and lower envelopes by constructing a cubic spline interpolation using maxima and minima as knots, respectively. 3) Take the average of two envelopes and denote it as m_{11}. 4) Obtain the difference c_{11} between the signal y and m_{11}. That is, $c_{11} = y - m_{11}$. 5) Treating c_{11} as the signal y, repeat the steps 1)∼4) and obtain the next difference denoting as c_{12}. After k iterations, the difference can be represented as $c_{1k} = c_{1(k-1)} - m_{1k}$. In addition 6) stop the sifting repetition if c_{1k} satisfies the above two IMF conditions. Now we have the first IMF as $c_1 = c_{1k}$. From the residue $y - c_1$, we extract the second IMF and finally obtain the successive n IMFs when the component c_n is very small or the residue r_n is a monotonic function from which no more IMF can be extracted. The decomposition result can be denoted as $y = \sum_{j=1}^{n} c_j + r_n$. Since we extract a higher frequency component earlier, note that c_1 is the most frequently oscillating component.

Historically, the EMD process was suggested to apply Hilbert transform to a signal for finding the time-varying frequencies in it. The time-varying frequency named as instantaneous frequency(IF) can be mathematically defined by Hilbert transform (Boashash, 1992). Since Hilbert transform works correctly for zero-symmetric signals only, we need to decompose a signal into at least locally zero symmetric components, IMFs. For specific examples related to zero symmetric signals and Hilbert transform, refer to Huang *et al.* (1998). By EMD, we can properly apply Hilbert transform to each IMF and also find its IF. The orthogonality and completeness of IMFs as a basis system are well discussed in Huang *et al.* (1998). The stopping rule to prevent meaningless iteration of the above step 4) is also suggested. For other details of EMD, refer to Huang *et al.* (1998).

2.2. The Hilbert spectrum

Through the (frequency) spectrum obtained by Fourier transform, we can represent a signal in the frequency domain. If we expand a signal by Fourier transform, the representation would be $y(t) = \sum_{j=1}^{\infty} a_j e^{iw_j t}$ where a_j is the Fourier coefficient of the signal $y(t)$ at the frequency w_j. The powers based on a_j expressed versus frequency can be called the spectrum. That is, Fourier transform can support a global frequency-energy distribution for the signal. Note that we can only consider constant frequencies $\{w_j\}$ in the Fourier analysis.

On the other hand, if $x(t) = \cos\psi(t)$ is a sinusoid having a time varying frequency, then the IF of $x(t)$ is defined as

$$w(t) = \psi'(t) = \frac{1}{2\pi}\frac{d}{dt}\arg(z(t)), \tag{2.1}$$

where $z(t) = x(t) + iH[x(t)] = a(t)\exp(i\int w(t)dt)$ is a complex form or analytic signal of $x(t)$ (Boashash, 1992). The H denotes Hilbert transform and $a(t) = \sqrt{x(t)^2 + H[x(t)]^2}$ represents the amplitude of the analytic signal. After the signal $y(t)$ is decomposed into IMFs by EMD, the analytic signal can be represented as

$$z(t) = y(t) + iH[y(t)] = \sum_{j=1}^{n}\{c_j(t) + iH[c_j(t)]\} = \sum_{j=1}^{n} a_j(t)\exp\left(i\int w_j(t)dt\right). \tag{2.2}$$

Since Hilbert transform can be properly applied to IMFs, we can obtain the IFs, $\{w_j, j = 1, \ldots, n\}$, from $\int w_j(t)dt = \arctan(c_j(t)/H[c_j(t)])$. Now, similar to Fourier spectrum, we can have three dimensional time-frequency-energy distribution using $a_j(t)$ on $(t, w_j(t))$ for each IMF. This distribution for all IMFs is denoted as Hilbert (amplitude) spectrum $H(w,t)$, which is a local frequency-energy distribution. Since the residue usually represents the trend of given signal and occupies much of energy in the signal, we do not consider the residue for the efficient representation of IMFs in the Hilbert spectrum.

In Huang *et al.* (1998), the whole process of both EMD and the Hilbert spectrum is named Hilbert-Huang transform(HHT).

2.3. The HHT-based prediction procedure

In Kim *et al.* (2008), the worm count data representing cyber attacks are analyzed and predicted based on HHT and a vector AR model. The procedure can be described as follows. First, given time series y is decomposed into n IMFs and a residue by the EMD process, that is, $y = \sum_{j=1}^{n} c_j + r_n$. For the worm count data, six IMFs and a residue are produced. Second, obtain the local-frequency-energy distribution or the Hilbert spectrum, $H(w,t)$, for all IMFs. By adding the power up to the k^{th} IMF from the first one, the cumulative energy E_k is defined as $E_k = \sum_{j=1}^{k} \sum_t a_j(t)^2/E$ where E denotes the total energy, $\sum_{j=1}^{n} \sum_t a_j(t)^2$. Third, since the lower frequency components may affect the main movement of the time series, several components starting from the n^{th} IMF are chosen based on the cumulative energy. Note that we select several lowest frequency components. Apply a vector AR model to the selected IMFs and predict each IMF. Three IMFs are used as input variables for the vector AR model in the worm count data analysis in Kim *et al.* (2008). Forth, apply a polynomial regression to the residue and predict it. Finally, obtain the prediction result of the given time series by joining all prediction results of the selected IMFs and the residue.

The specific result by the HHT approach for predicting a real data set are presented in Section 4 and compared with the HHT-PCA procedure suggested in the next section.

3. The Proposed Prediction Procedure

3.1. Necessity for adopting PCA

Consider a signal having several components as follows:

$$y = \sum_{i=1}^{p} a_i \cos 2\pi\theta_i t + \beta t, \tag{3.1}$$

where $\theta_1 > \cdots > \theta_p$. Under the assumption that the EMD process appropriately extracts IMFs, the i^{th} IMF may represent the i^{th} component $a_i \cos 2\pi\theta_i t$. For example, IMF1 can be an estimator of $a_1 \cos 2\pi\theta_1 t$ and the residue be an estimator of βt under the above assumption.

We generate two simple signals named Case A and Case B. Both have four components ($p = 4$) and are designed for the EMD process to work properly. Hence, we can obtain four IMFs and a residue in two examples as desired. With the same set of four frequencies for both cases, we put the amplitude differently as $(a_1, a_2, a_3, a_4)_A = (0.7, 2, \mathbf{4}, 1.5)$ for the Case A and $(a_1, a_2, a_3, a_4)_B = (0.7, \mathbf{4}, 2, 1.5)$ for the Case B. Two signals are shown in the first row in Figure 3.1.

Following the prediction procedure in Kim *et al.* (2008), the panels (a1) and (b1) show the cumulative energy for IMFs by a black line. The energy of each IMF is represented by a blue line. As expected in the amplitude design, the third component in the Case A has the largest part of energy whereas the second component does in the Case B.

The panels (a2)∼(a4) show the estimation (by blue line) and prediction (by red line) results using 2∼4 IMFs for the vector AR model, respectively. We used 90% of data for modeling and 10% for prediction. The panels (b2)∼(b4) are the same. In both cases, the IMFs representing slowly moving components are firstly considered. For predicting the residue, we employ a polynomial regression

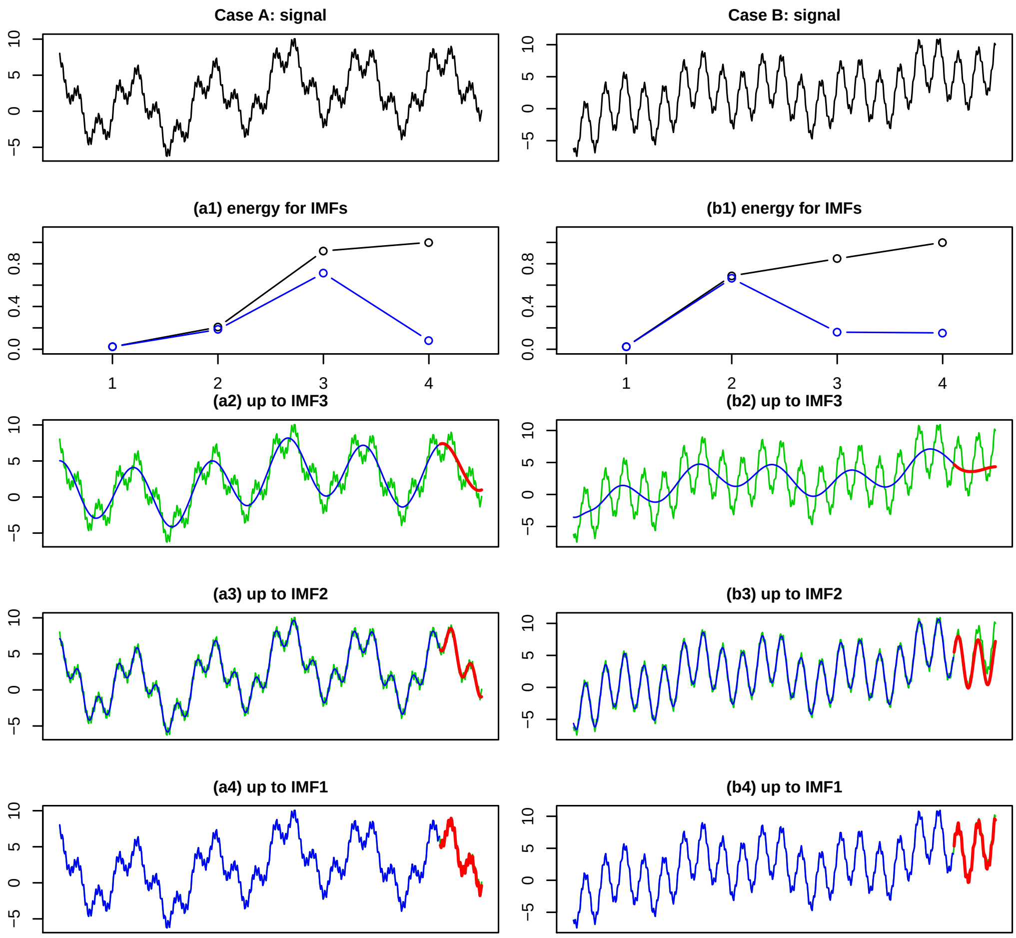

Figure 3.1. Toy examples of Case A and Case B: The estimation and prediction results based on the HHT approach

and the prediction result for the residue is added to those from IMFs in all cases, (a2)∼(a4) and (b2)∼(b4).

In the Case A, where the third component is the most dominant, the prediction result (a2) that considers IMF4 and IMF3 for the vector AR process seems to be satisfactory to explain the main trend of the signal in the Case A. However, the prediction result (b2) does not because the most dominant component is estimated by IMF2 and IMF2 is not considered for the vector AR model in (b2). Only after IMF2 is considered for predicting the data, a satisfactory result is provided as shown in the panel (b3). The results (a4) and (b4) consider all four IMFs for prediction.

As demonstrated in these simple examples, we need to decide the number of IMFs being used for predicting a time series satisfactory, case by case. Sometimes, we may need all n or almost all $n-1$ IMFs to obtain a meaningful prediction result, which means inefficiency for dimension reduction. It would be better, therefore, to have a step for dimension reduction or proper feature extraction among all IMFs for the HHT-based prediction procedure.

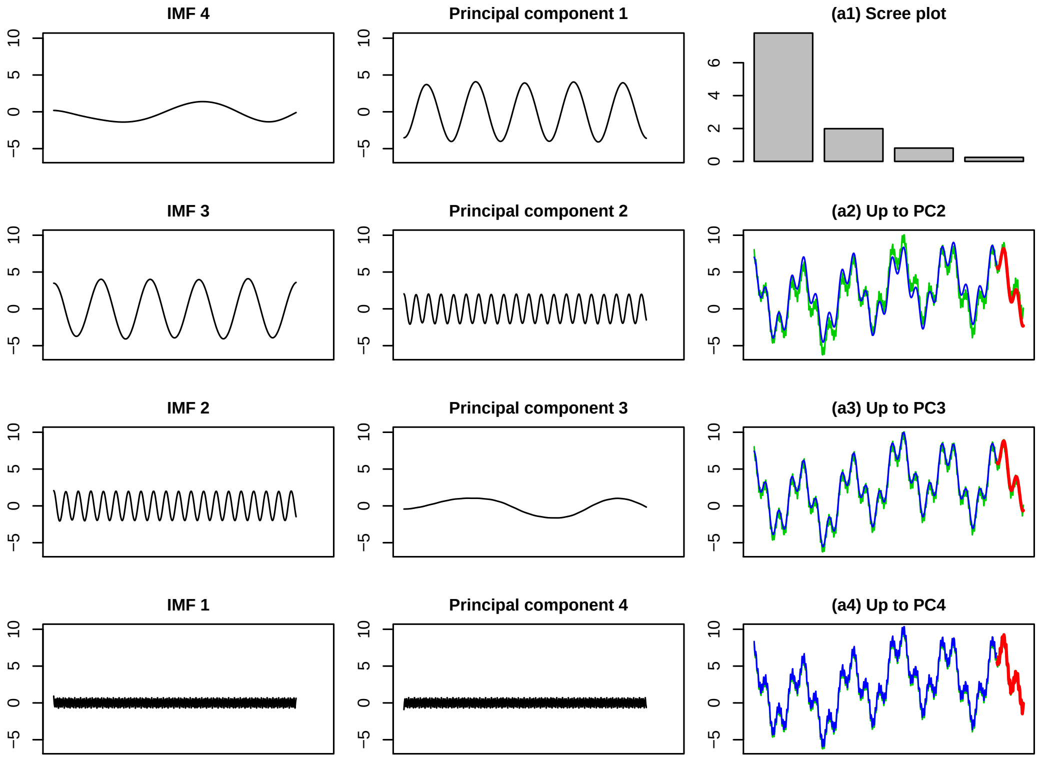

Figure 3.2. Case A: The estimation and prediction results based on the suggested HHT approach combined with PCA

PCA has been widely used for the purpose of the dimension reduction and feature extraction. We suggest to adopt PCA for the HHT prediction procedure and the detailed procedure is explained in the following section.

3.2. The proposed prediction procedure

The procedure of the proposed method is as follows: 1) Perform the EMD process to the given time series y. 2) Fulfill PCA to the n IMFs and obtain principal components(PCs). Based on the Scree plot, select some PCs. 3) Apply a vector AR model to the selected PCs and a polynomial regression model to the residue. Obtain the prediction result of each component. 4) Add all prediction results and predict y with the sum.

Figure 3.2 shows the estimation (by blue line) and prediction (by red line) results of the Case A by the proposed method in the third column. All green lines represent the given signal. The extracted PCs are in the second column. We can see that the PC1 is close to IMF3, which is designed to be the dominant component in the signal. In the Case A, each PC is similar to one of the IMFs. By adopting a PCA step and using PCs as input variables of a vector AR model, we can obtain the desired prediction result in (a2). Note that in (a2), where we used two PCs, at Figure 3.2, we had the similar prediction result to (a3), where we hired three IMFs, at Figure 3.1. This implies that we can more automatically select appropriate variables by adopting a PCA step.

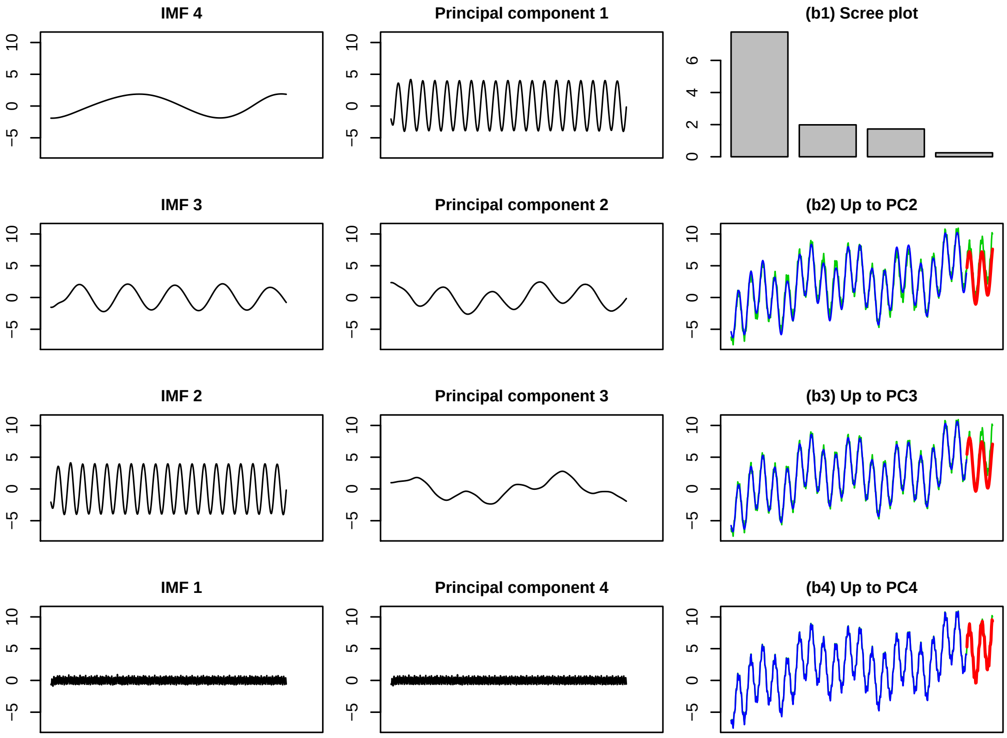

Figure 3.3. Case B: Estimation and prediction results based on HHT and PCA

Table 3.1. MSEs for the reconstruction results (by using 90 % of data)

	Case A			Case B		
Number of inputs	2	3	4	2	3	4
HHT	2.246	0.246	0.000	8.206	0.249	0.000
HHT & PCA	1.118	0.335	0.092	0.499	0.243	0.005

The results for the Case B are shown in the Figure 3.3. In this case, the PC1 indicates the dominant component IMF2, and the PC2 and PC3 are the mixtures of the IMF3 and IMF4. Since PCA can extract main features - PCs as linear combinations of IMFs based on the variance structure, we expect efficient prediction results using a few PCs. The result (b2) by two PCs in Figure 3.3 is similar to (b3) by three IMFs in Figure 3.1.

Table 3.1 shows the mean square error(MSE) when we use 90% of data for a vector AR modeling. HHT means that we apply the HHT approach with several IMFs whereas HHT & PCA denotes that we use the proposed prediction procedure by adopting a PCA step. Comparing to the MSE values, 2.246 for Case A and 8.206 for Case B, when we use 2 IMFs(IMF4 and IMF3) as input variables, the MSEs when we employ 2 PCs(PC1 and PC2) as input variables is relatively lower as 1.118 for Case A and 0.499 for Case B.

Consequently, adopting a PCA step to the prediction approach based HHT could perform a role of more automatic variable selection or proper feature extraction for a vector AR model in the HHT prediction approach.

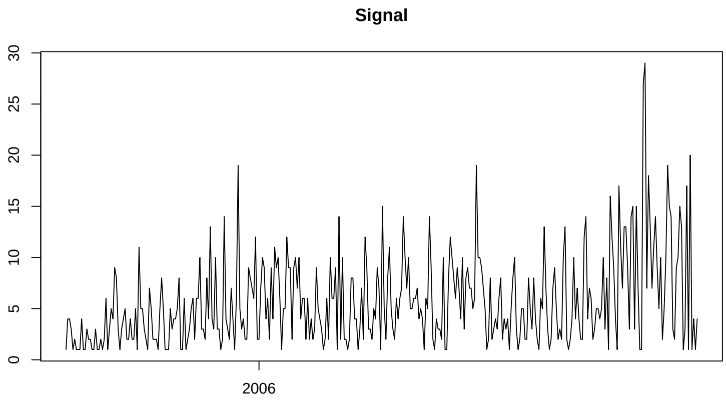

Figure 4.1. The worm count data from August 1, 2005 to October 9, 2006

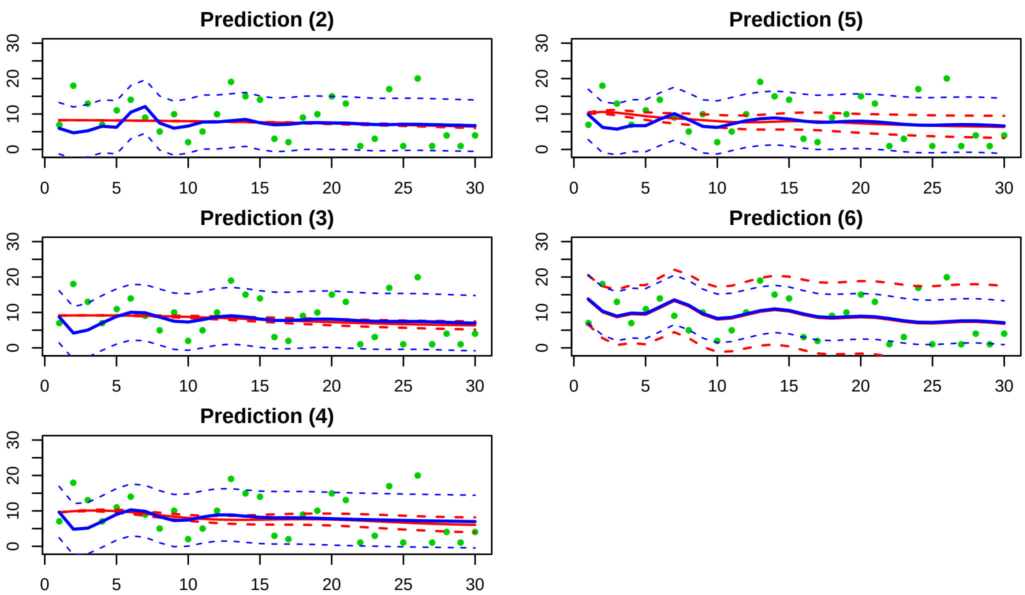

Figure 4.2. The worm count data analysis: Comparison the prediction results based on HHT *vs.* HHT and PCA

4. Worm Count Data Analysis

As a signal representing cyber-attacks, the worm count data are shown in Figure 4.1. The data were observed for 435 days from August 1, 2005 to October 9, 2006 and the original source of the data is on the website of Ahn Lab, http://www.ahnlab.com. As explained in Kim *et al.* (2008), it seems to be difficult to find a specific pattern in the signal. However, we would like to predict the last 30 days of the time series based on the proposed HHT-PCA prediction procedure.

The close looks on the prediction results by the HHT *vs.* HHT-PCA approaches, respectively, are shown in Figure 4.2. The prediction results by the HHT approach in Kim *et al.* (2008) by using 2∼6 IMFs as input variables are shown by red lines. The dotted lines indicate 95% confidence intervals. The blue lines show the prediction results based on the proposed HHT-PCA prediction procedure. When we used all six IMFs or PCs, the results are similar as shown in the panel of Prediction (6). However, we can see that the usage of two PCs in the panel of Prediction (2) can follow the prediction result by using all IMFs. Since the proposed method can give priority to small scale movement in the signal when it dominates the signal, it seems to obtain a more desirable result with less number of input variables.

The specific process is as follows. The IMFs and the residue by EMD are shown in the left column in Figure 4.3. The First IMF has the most energy in the signal as shown in the first panel on the right column. By following the HHT-based prediction procedure, the prediction result for the last 30 days are represented by red lines in (2)∼(6). The panel (2) considers IMF6 and IMF5, (3) does IMF6∼IMF4, . . . , and finally (6) contains all IMFs for the vector AR modeling. The MSEs between the signal (represented by a green line) and the reconstruction result by several IMFs (represented by a blue line in each panel) are displayed in the last panel. We also observe that the first IMF dominates the MSE plot.

Based on the proposed HHT and PCA prediction approach, the reconstruction and prediction results are shown in Figure 4.4. The PCs automatically selected by the variance structure are shown in the left column. The first PC reflects the first IMF and other PCs seems to be linear combinations of IMFs. The first PC dominates the signal as revealed in the Scree plot. The reconstruction results are described by blue lines and the prediction results are by red lines. By using only two PCs, we can obtain the similar pattern to (6) by all IMFs in Figure 4.3. The panel showing MSE versus the number of PCs supports the efficiency of the feature selection by adopting the PCA step.

5. Concluding Remarks

In this paper, we focus on the prediction procedure based on the HHT approach, and propose a new prediction method that is a novel combination of HHT and PCA. By adopting PCA, the proposed prediction procedure can consider the dominant component first even though the dominant one is not the most slowly oscillatory. We can also obtain a new feature as an input variable for prediction modeling that is a linear combination of some IMFs.

Before closing this paper, we would like to mention an accompanying benefit of the adoption PCA and to pay attention in the use of the decomposition approach. First, the EMD process is not perfect and can product inappropriate IMFs so that the direct usage of IMFs for a vector AR model could be risky for the prediction. For example, EMD can divide four components into seven meaninglessly distorted ones. In this case, to select meaningless two lowest frequency components (least oscillating IMFs) might produce a nonsensical prediction result. On the other hand, since PCA can produce a linear combination of IMFs which could become a meaningful component, we may expect the alleviation of the negative effect of EMD to the prediction.

Second, as stated in the introduction, we choose to decompose given signal into simple components for appropriate prediction. As a prediction method, we use a vector AR model that is widespread but assumes the stationarity of input variables. Therefore, we need to pay attention not to apply the proposed method to apparently nonstationary IMFs, even though EMD can support nonlinear and nonstationary signals and IMFs can have time-varying frequencies. For example, if a dominant component is a linear chirp signal whose frequency is rapidly linearly-increasing or decreasing, then the prediction result cannot be appropriate. For the components having rapidly time-varying frequencies, we can apply an IF estimation method but this is beyond the scope of this paper.

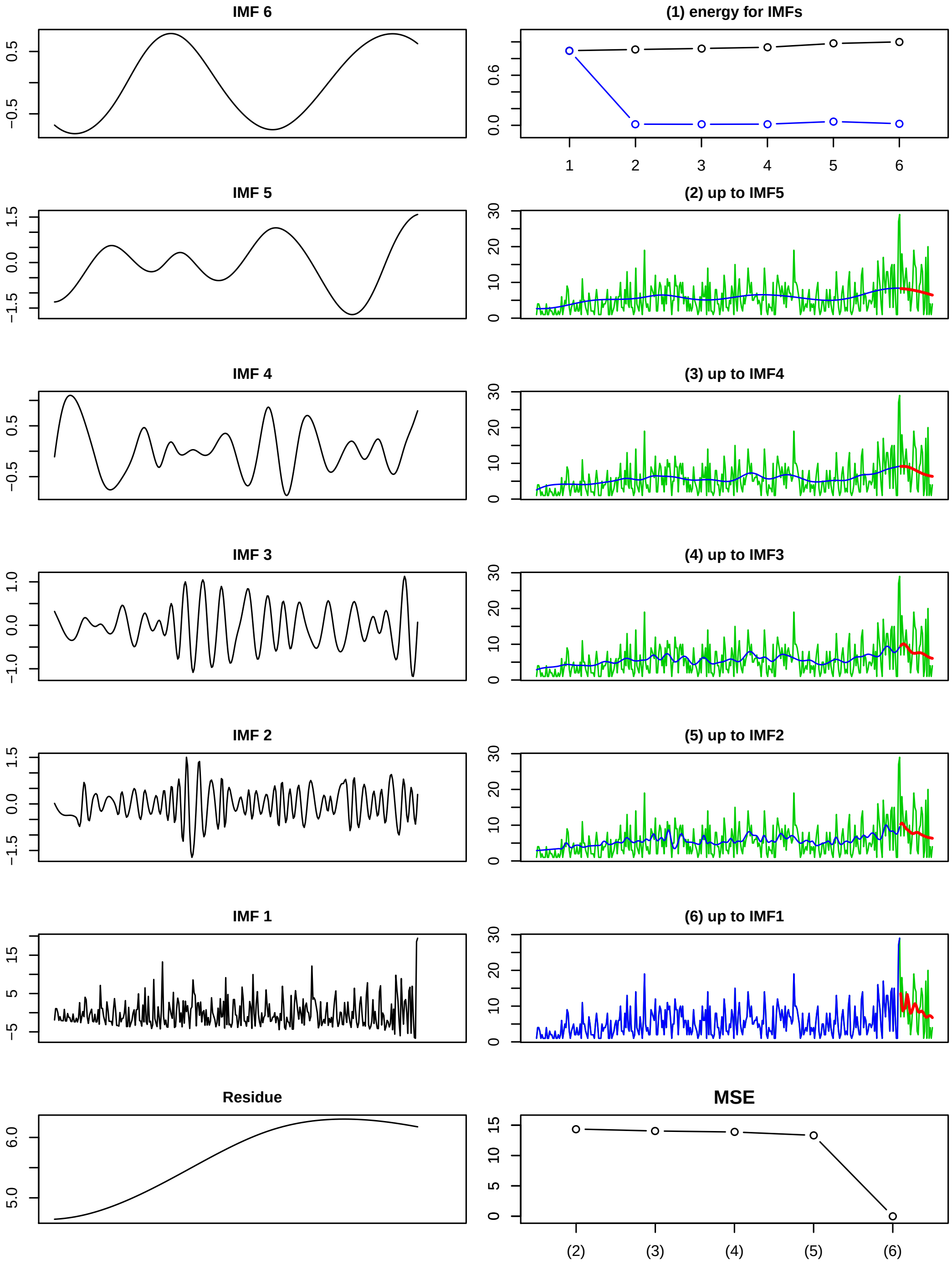

Figure 4.3. The worm count data analysis: Estimation and prediction results based on the HHT approach

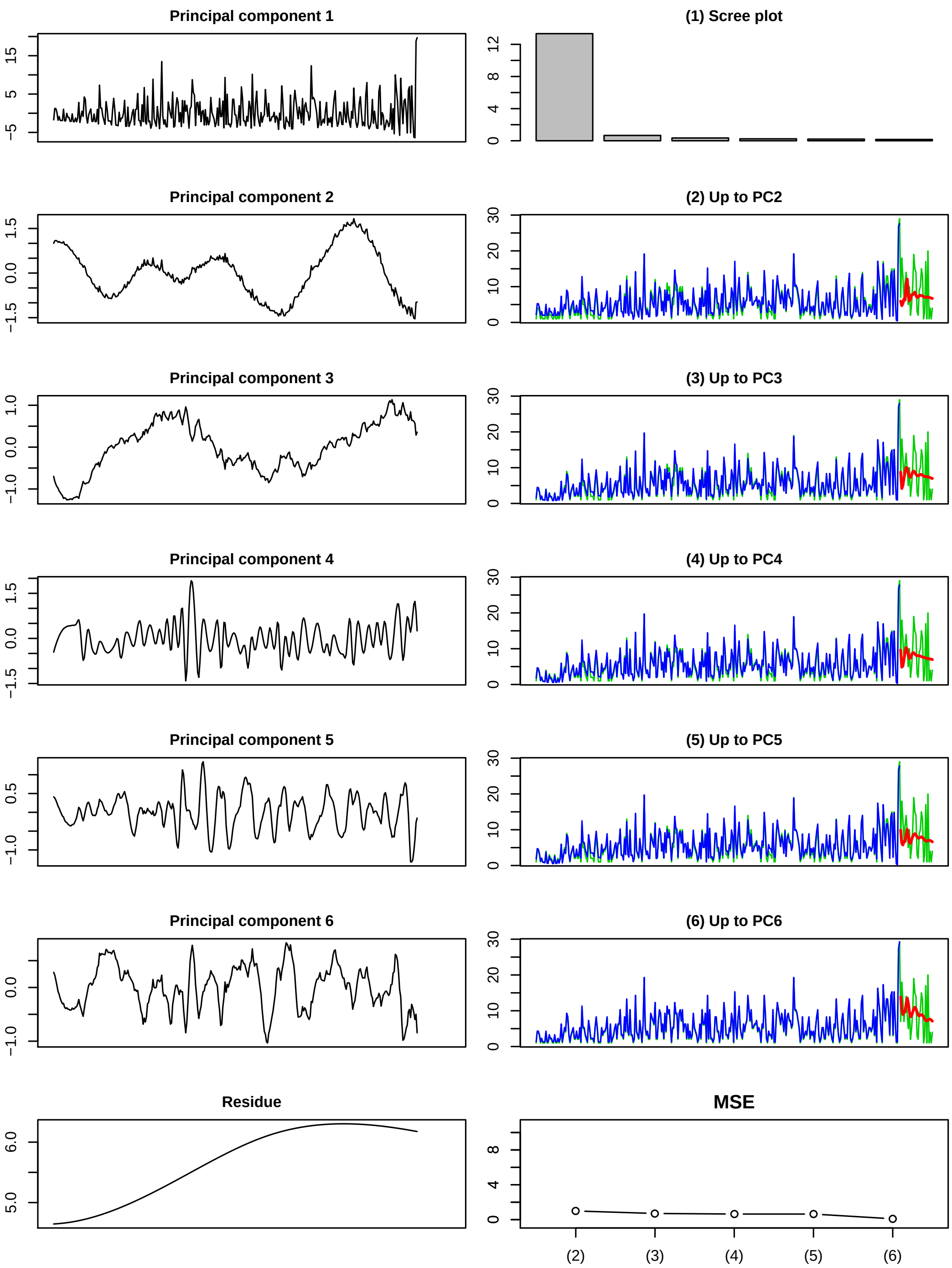

Figure 4.4. The worm count data analysis: Estimation and prediction results based on the proposed HHT and PCA approach

References

Boashash, B. (1992). Estimating and interpreting the instantaneous frequency of a signal - part I: Fundamentals, *Proceedings of the IEEE*, **80**, 519–538.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C. and Liu, H. H. (1998). The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis, *Proceeding of the Royal Society London A*, **454**, 903–995.

Kim, D., Paek, S.-H. and Oh, H.-S. (2008). A Hilbert-Huang transform approach for predicting cyber-attacks, *Journal of the Korean Statistical Society*, **37**, 277–283.

Oh, H.-S., Suh, J. H. and Kim, D. (2009). A multi-resolution approach to non-stationary financial time series using the Hilbert-Huang transform, *The Korean Journal of Applied Statistics*, **22**, 499–513.