Advances in Adaptive Data Analysis
Vol. 3, No. 4 (2011) 483–491

DOI: 10.1142/S179353691100091X

World Scientific
www.worldscientific.com

ASSESSING DISCONTINUOUS DATA USING ENSEMBLE EMPIRICAL MODE DECOMPOSITION

BRADLEY LEE BARNHART*
IIHR-Hydroscience & Engineering,
The University of Iowa, 130 W. Harrison St.,
Iowa City, Iowa 52242, USA
bradleybarnhart@gmail.com

HONDA KAHINDO WA NANDAGE
Department of Computer Science,
The University of Iowa, 702 18th Ave #4,
Coralville, Iowa 52241, USA
hondakahindo@hotmail.com

WILLIAM EICHINGER
IIHR-Hydroscience & Engineering,
The University of Iowa, 130 W. Harrison St.,
Iowa City, Iowa 52242, USA
william-eichinger@uiowa.edu

This investigation presents an improved ensemble empirical mode decomposition (EEMD) algorithm that can be applied to discontinuous data. The quality of the algorithm is assessed by creating artificial data gaps in continuous data, then comparing the extracted intrinsic mode functions (IMFs) from both data sets. The results show that errors increase as the gap length increases. In addition, errors in the high-frequency IMFs are less than the low-frequency IMFs. The majority of the errors in the high-frequency IMFs are due to end-effect errors associated with under-defined interpolation functions near the gap endpoints. A method that utilizes a mirroring technique is presented to reduce the errors in the discontinuous decomposition. The improved algorithm provides a more locally accurate decomposition of the data amidst data gaps. Overall, this simple but powerful algorithm expands EEMD's ability to locally extract periodic components from discontinuous data.

Keywords: Ensemble empirical mode decomposition; end-effect error; discontinuous data.

1. Introduction

There are a number of empirical mode decomposition (EMD) algorithms available today. These include commercial software called the Hilbert-Huang transform data-processing system (HHT-DPS), which was developed by Norden Huang at NASA

*Corresponding author.

and is available at http://techtransfer.gsfc.nasa.gov/HHT. There are also publicly available MATLAB codes by Patrick Flandrin (http://perso.ens-lyon.fr/patrick.flandrin/emd.html) and R code by Kim and Oh [2009], which extract intrinsic mode funcitons (IMFs) from a given input data series. The National Central University has an EMD package that is available at http://rcada.ncu.edu.tw/research1.htm and Imperial College offers a number of EMD tools and data that are available at http://www.commsp.ee.ic.ac.uk/∼mandic/research/emd.htm.

However, no investigation to our knowledge has utilized these algorithms to process discontinuous data. For example, the ensemble empirical mode decomposition (EEMD) algorithm introduced by Wu and Huang [2005] will not run when there exist gaps, or NaNs, in the input data.

Oftentimes instruments fail in the field, resulting in data gaps within the data. These gaps prevent the EMD algorithm from properly sifting through the data. Typically, scientists utilize interpolation values, such as the mean of surrounding data points, to fill data gaps. This may be useful for small data gaps; however, the data is assumed to remain constant during the time period, and so is insufficient for the larger gaps of fluctuating data. In the spirit of a local and adaptive decomposition tool, it is important to manipulate the data as little as possible, and merely describe the data that does exist. This investigation suggests an improvement to the EEMD algorithm that allows for gaps in the input data. The implications of applying EEMD to varying the sizes of discontinuous data will be discussed. In addition, we will suggest an error reduction technique that extracts IMFs that are more locally accurate.

2. Ensemble Empirical Mode Decomposition

This investigation utilizes the EEMD algorithm that was pioneered by Wu and Huang [2005]. The algorithm utilizes the original EMD sifting method, which is described fully in Huang *et al.* [1998]. An overview of the EEMD algorithm is as follows:

(1) Add finite amplitude noise to the original input signal.
(2) Decompose the signal into a finite set of IMFs using the original EMD sifting method.
(3) Repeat Steps 1 and 2 with different noise data sets of same noise standard deviation.
(4) Average the ensemble of extracted IMFs to average out the noise and obtain their mean IMFs.

A complete description of EEMD can be found in the paper by Wu and Huang [2005]. The standard deviation of random noise, which is added to the original data before decomposition, can be specified by the user. For this investigation, we used a standard deviation of 0.2.

In order to accommodate for discontinuous data, the MATLAB version of the EEMD algorithm by Wu and Huang [2005] was modified. In order to accomplish this, the sifting process, which fits spline functions to the local maxima and the minima of the signal, must be performed on each individual continuous data segment. When the algorithm encounters a data gap, the splines must be halted and a new set of piece-wise spline functions begin on the next set of continuous data.

Figure 1 shows original data and its decomposed IMFs, as well as the decomposed set with artificial data gaps created.

The improvement is simple but powerful. The majority of the discontinuous IMFs seem to replicate the original IMFs that are away from the gap section. For instance, look at approximately Arbitrary Time 90 for IMF2 in Fig. 1. The local oscillations are still captured even with the gap present. The two decompositions are not identical, however. This investigation will now assess the errors associated with decomposing discontinuous data in a quantifiable manner. In addition, suggestions will be made for how to reduce such errors.

3. Errors Due to Data Gaps

In order to assess the abilities of EEMD applied to discontinuous data, we present an error analysis comparing an original IMF with a discontinuous IMF. Equation (1) shows the root mean square error equation.

$$Err_{\mathrm{IMF}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(IMF_{i,\,\mathrm{nogap}} - IMF_{i,\mathrm{gap}})^2}. \tag{1}$$

The total error for a single decomposition, then, is the sum of the rms errors from all IMFs. It is interesting to compare the rms error with the size of gaps within the data. Figure 2 shows the error calculated for a number of data gap sizes. That is, a data gap was artificially created in the exact middle of the data. The gap length was created as some percentage of the original data length. Then, the IMFs were decomposed and compared with the IMFs from the original data.

The errors, as calculated by Equation (1), are shown in Fig. 2 for the first six IMFs extracted using the new discontinuous EEMD algorithm. The errors were also normalized by dividing the error by the standard deviation of each IMF. Notice that the errors increase with the increased gap size, as expected. The increase is fairly consistent, especially up until the gap size is greater than half the total length of the data.

In addition, the highest-frequency IMFs have the smallest differences with the original data. As the IMF number increases, which represents the lower-frequency IMFs, the errors increase more quickly with the increasing data gap size. This shows that the algorithm works better with high-frequency IMFs.

This investigation has also found that the majority of the errors in the high-frequency IMFs occur near the gaps. Figure 3 shows the errors plotted against time

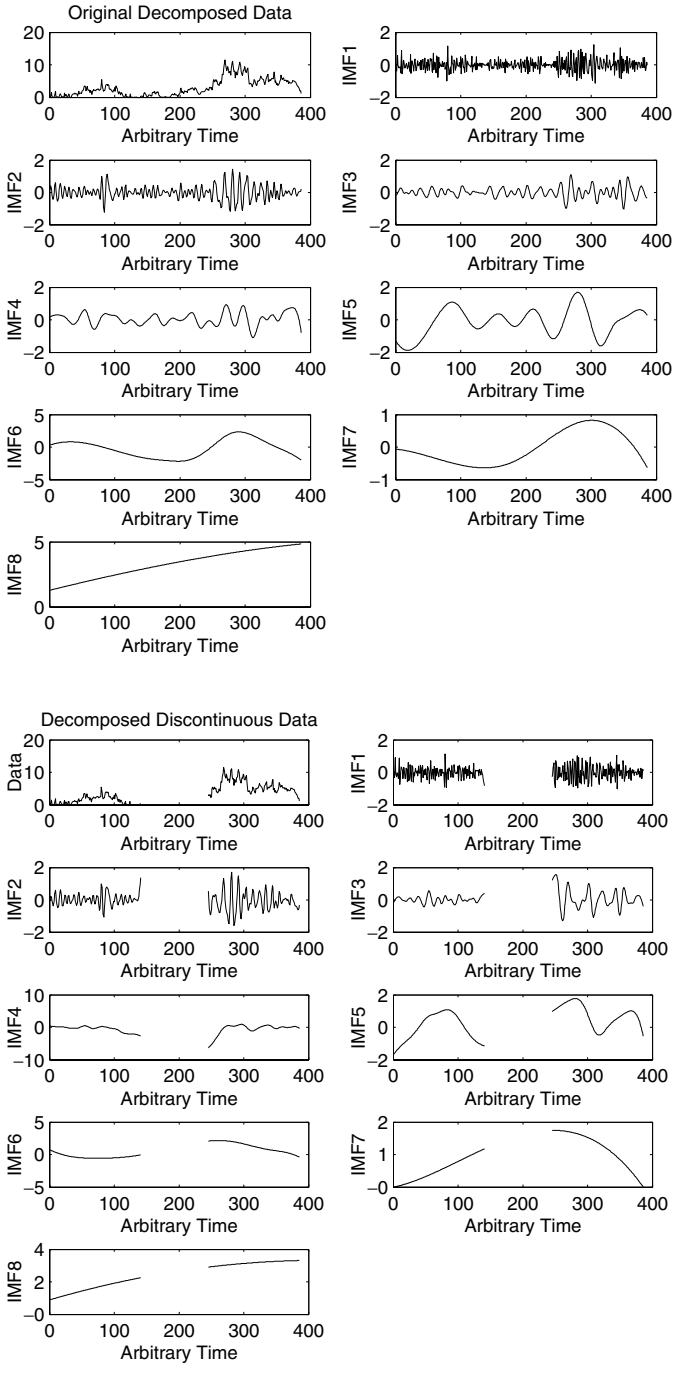

Fig. 1. Original data decomposed into its IMFs as well as the data and IMFs decomposed from the discontinuous data.

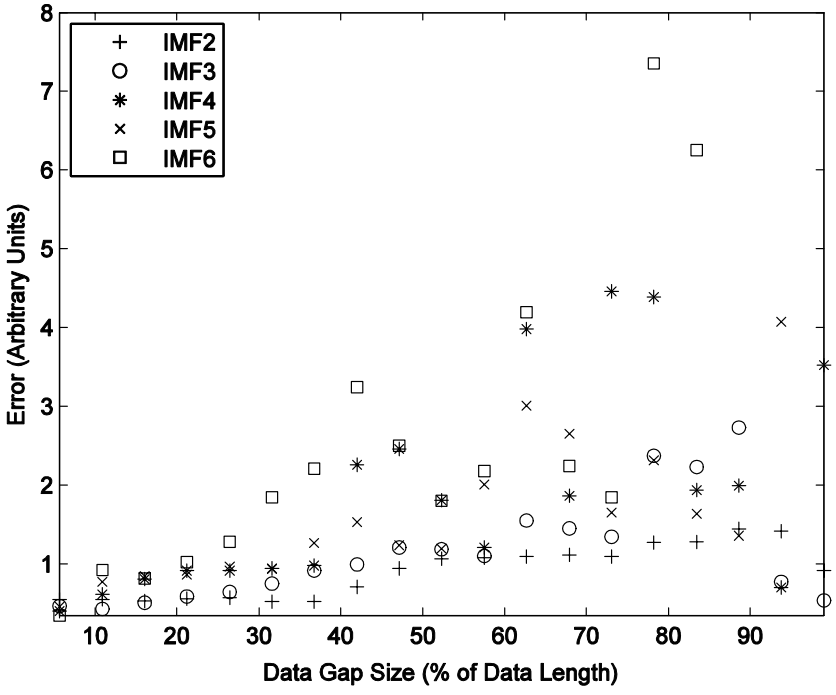

Fig. 2. Error as defined by the summed differences between the discontinuous and continuous extracted IMFs, plotted against data gap size.

for all IMFs with a gap size of 80 points, which is approximately 20 % of the total data.

Notice that the high-frequency IMFs have the largest errors near the endpoints of the gaps. The low-frequency IMFs also have errors, but they are not specifically located near the gap endpoints.

A common problem within the old EMD sifting algorithm has been one dealing with so-called *end-effect errors* [Huang *et al.* (1998); Qingjie *et al.* (2010)]. These errors, which have been well studied, have traditionally existed at the endpoints of the data [Huang *et al.* (1998); Qingjie *et al.* (2010)]. This is because the first or second derivatives that are required for spline fitting are unavailable. The endpoints, then, have large fluctuations that do not represent the real signal. For our particular algorithm that is dealing with discontinuous data, these end-effect errors occur not only at the beginning and the end of the input data, but also at the start and the end of every data gap. While the differences in the low-frequency IMFs are primarily due to the size of the data gap, the high-frequency IMFs have differences primarily due to gap end-effect errors. Therefore, in order to reduce the errors from the high-frequency IMFs, we can use traditional end-effect mitigation tools as described in the subsequent section.

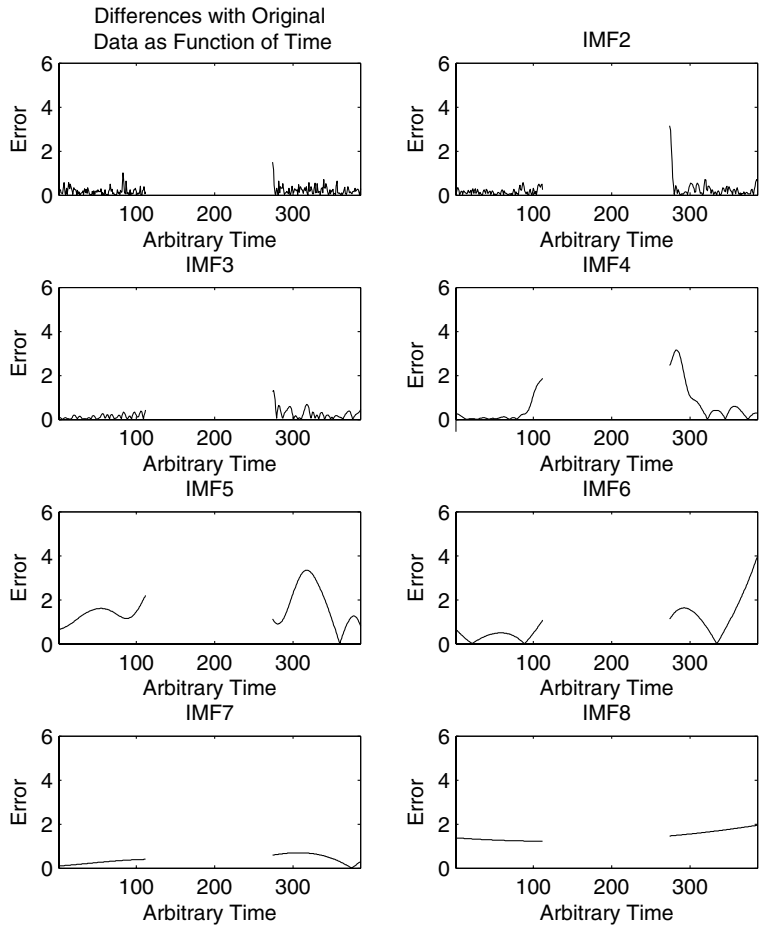

Fig. 3. Errors plotted as a function of time. For the high-frequency IMFs, the errors occur largely near the gap endpoints.

4. Error Reduction Methods

There are a number of investigations that have dealt with end-effect errors [Huang *et al.* (1998); Qingjie *et al.* (2010); Zhao and Huang (2001)]. Therefore, it may be possible to utilize these end-effect mitigation tools in order to decrease the errors in the high-frequency IMFs due to data gaps.

Qingjie *et al.* [2010] suggest that IMFs should use a mirror extension technique to lower the errors due to end effects. This would restrict the spline from varying extravagantly at the ends of the gaps.

The mirroring technique used in this investigation is now reviewed. When a data gap is encountered, it is split in two sections. The first section is filled by the mirror image of the data directly before the gap. The second section is filled by the mirror image of the data directly after the gap. The amount of mirroring

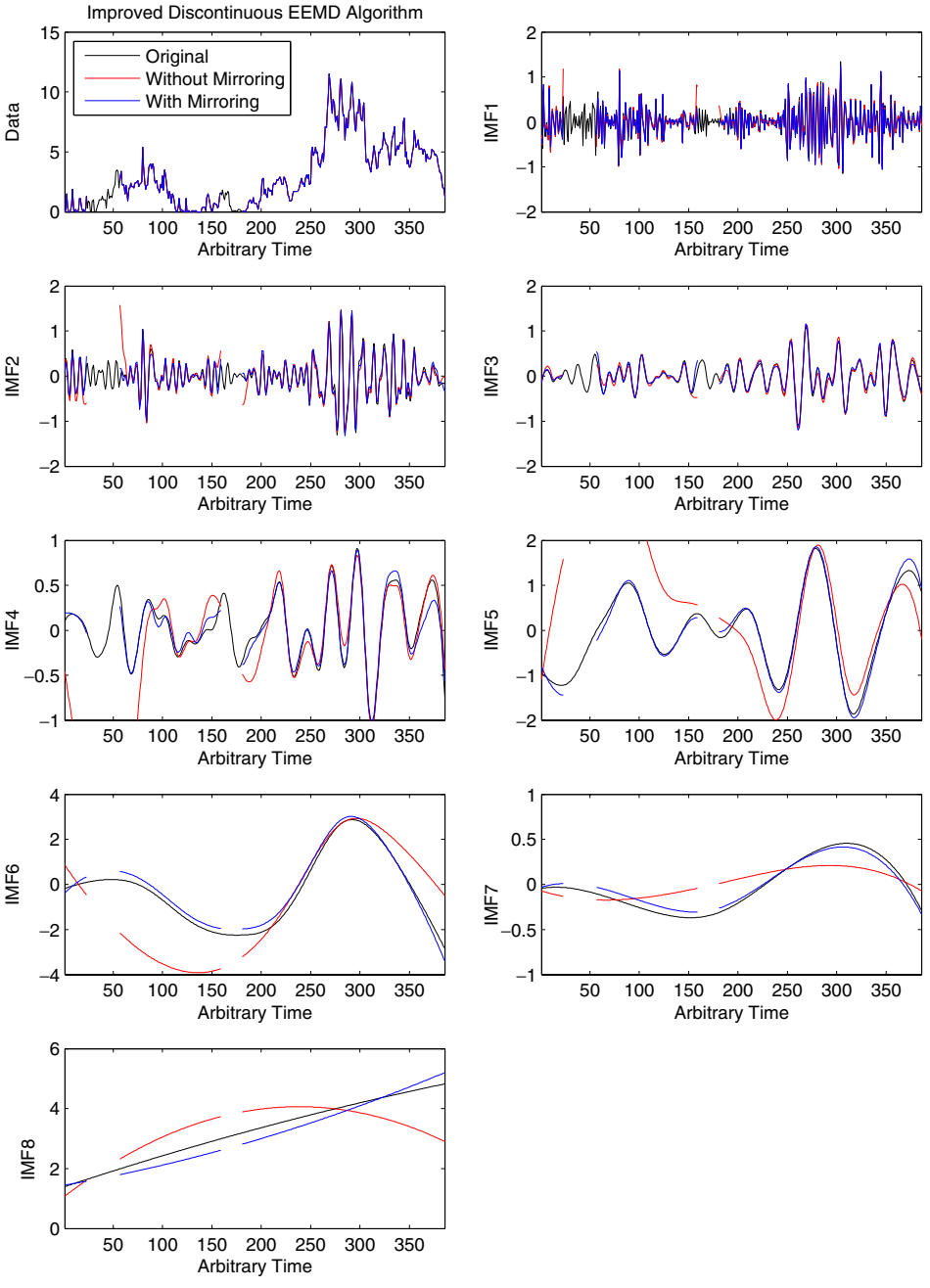

Fig. 4. Comparison of three different decompositions of data. The original signals (black) contain no gaps, the red signals are the decomposed IMFs from the discontinuous EEMD algorithm, and the blue signals are the discontinuous EEMD algorithm used after a mirroring technique was performed.

needed is dependent on the gap size. The result is a continuous data set. The traditional EEMD algorithm is then used to decompose the data into its IMFs. Once decomposed, the data gaps are recreated by removing the mirrored data.

To test the effectiveness of this technique with discontinuous data, two gaps were created in the continuous data. Three different algorithm iterations were performed. The first one was a traditional EEMD algorithm that is decomposing the original data without gaps. The second was the discontinuous EEMD algorithm as applied to the data set without mirroring. The third was the discontinuous EEMD algorithm as applied to the data that had undergone the mirroring technique.

Figure 4 shows the three decompositions compared. This verifies, at least visually, the effectiveness in the mirroring technique to reduce end-effect errors near the endpoints of the gaps. In addition, the low-frequency IMFs more closely match the original decomposition.

Figure 5 shows the rms error from each IMF as compared between the discontinuous EEMD algorithms with and without mirroring applied.

The first IMF is the actual data and can be ignored. For all the other IMFs, the mirroring technique greatly reduces the error in the decompositions. Therefore, the processing of discontinuous data is greatly improved by using the mirroring technique. The mirroring technique used is not the only technique that could be used to fill data gaps. It is possible to utilize neural networks or other predictive methodologies in order to fill the data gaps based upon the previous and subsequent data. The primary goal of gap-filling techniques, though, is to constrain the

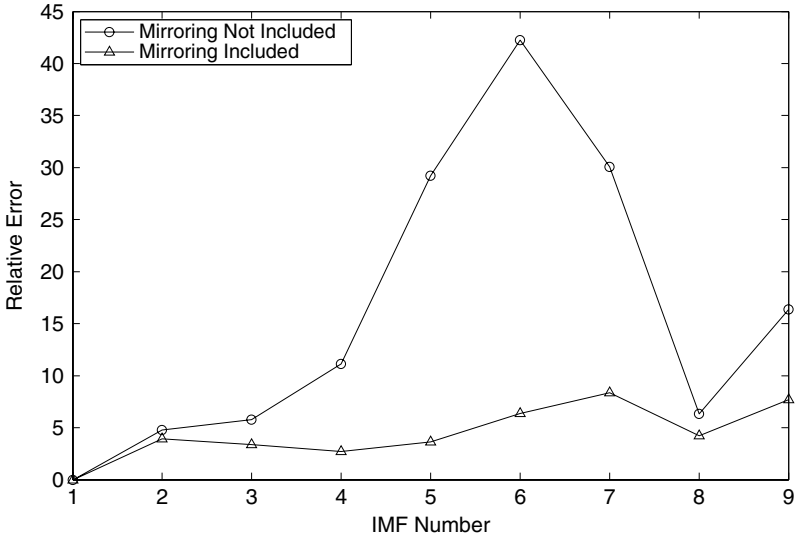

Fig. 5. Comparison of relative error associated with including or not including the mirror technique when using the discontinuous EEMD decomposition.

splines and is not to conjecture what data may actually be within the gaps. Therefore, further research should be pursued, which would compare different gap-filling and spline-constraining techniques to determine if there are superior methods for analyzing discontinuous data with this new EEMD algorithm.

5. Discussion

Overall, this investigation presents a new version of the EEMD algorithm that is now applicable to discontinuous data. For short gap durations, the errors are small and the decomposition is locally representative. A mirroring technique was utilized to improve the discontinuous decomposition. This makes for a more local and adaptive decomposition of data that may contain one or more gaps. Further research should be pursued, which utilizes neural networks or prediction models to fill gaps and improve on the reduction of errors.

References

Huang, N. E., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N., Tung, C. and Liu, H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London A*, **454**: 903–995.

Kim, D. and Oh, H.-S. (2009). EMD: a package for empirical mode decomposition and Hilbert spectrum. *The R J.*, **1**: ISSN 2073–4859.

Qingjie, Z., Huayong, Z. and Lincheng S. (2010). A new method for mitigation of end effect in empirical mode decomposition. *2nd International Asia Conference on Informatics in Control, Automation, and Robotics,* 978-1-4244-5194-4/10.

Wu, Z. and Huang, N. E. (2005). Ensemble empirical mode decomposition: a noise assisted data analysis method. *COLA Tech Rep. 193.* Center for Ocean Land Atmos Studies. ftp://grads.iges.org/pub/ctr/ctr_193.pdf.

Zhao, J. and Huang, D. (2001). Mirror extending and circular spline function for empirical mode decomposition method. *J. Zhejiang Univ.* (Science), **2**: 247–252.