Detrending and denoising with empirical mode decompositions

Daniil LOTKOV
École Normale Supérieure Paris-Saclay

Abstract

This work is dedicated to the analysis of the paper "Detrending and denoising with empirical mode decompositions" by [Flandrin *et al.*, 2004a]. The report compiles several relevant papers on the topic and presents a deconvoluted explanation of important aspects. The algorithm is tested [1] against data of a variable fit, and the results show that it produces relatively good denoising/detrending even on non-standard data. The method for slight Hurst exponent estimation improvement is also proposed and tested yielding marginal gain in estimation quality.

1 Introduction

The problem of distinguishing signal from noise arises as an outcome of each and every real-world experiment. There exist many approaches for denoising the signal, which include the popular family of solving this problem in a spectral domain. The transition from the time to frequency domain could be performed in various ways, which include such well-studied operations as Fourier and Wavelet transformations. These spaces have been used as a representation for signal spectrum operations for a very long time. However, it is recognized that these methods have the limitations that come from the nature of the transformations themselves: Fourier transform expresses the signal as a composition of trigonometrical functions, and Wavelet transform utilizes wavelets for this. This apriori basis selection along with the basis functions properties imposes constraints on the final space that can result in a poor representation of signals of small support, and nonlinear or non-stationary signals. The Empirical Mode Decomposition (EMD) is a signal analysis method that allows us to adaptively choose the functions the signal is being decomposed into. This data-driven approach coupled with an algorithmic constraint of EMD basis functions being ordered by their frequency enables us to perform superior quality denoising and detrending operations by excluding certain components. The answer to the question of which components to exclude lies in the behavior of the EMD on the signal that has no actual information — noise. This report will discuss the EMD theoretical aspect, its algorithmic side, resulting in basis filtering criteria, and it will demonstrate examples of denoising done using this technique.

2 Empirical Mode Decomposition

The core assumption of the EMD algorithm is that a signal is comprised of multiple oscillatory components with different frequencies. In order to decompose the signal $x(t),\ t \in [0, T]$ into these components we run a *sifting* procedure multiple times. On each round of the sifting procedure, we obtain a new Intrinsic Mode Function (IMF) $d_i(t)$ that can be viewed as the basis component (the "frequency" of these components decreases with i). Each IMF when obtained is subtracted from the signal, and the procedure repeats on the residual. In the end the signal is represented as a linear sum of IMFs $d_n(t)$, and a final constant residual $m_N(t)$.

$$x(t) = \sum_{i=1}^{N} d_i(t) + m_N(t) \tag{1}$$

To get one IMF the sifting procedure is run multiple times until it meets the stopping criteria (e.g. std or energy threshold):

1. find all signal $x(t)$ extrema
2. build envelope around extrema
3. get mean of that envelope $m(t)$
4. extract the detail $d(t) = x(t) - m(t)$
5. iterate on the extracted detail signal $x(t) = d(t)$

The algorithm is guaranteed to converge since we decrease the number of extrema at every step; the maximum running time is bounded by $O(log_2 N)$. A typical number of basis functions is, therefore, also small, while these functions' support is identical to the initial signal's. This qualitatively distinguishes the given method from the other transformations, which in general have infinite support basis functions for signals limited in time. These properties coupled with a fact that IMFs still possess order in terms of their frequency give us an opportunity for a relatively easy frequency-space analysis and filtering.

[1] Google Colab access link: https://colab.research.google.com/drive/1Nsv7k2cvYNKL_9_ZjdrhQw2IE3Nmo8tb?usp=sharing

3 Detrending and denoising

In order to perform any kinds of operations on resulting basis functions we need to understand their properties. The main criteria upon which we decide whether the IMF represents noise or signal is energy. In the [Wu and Huang, 2004] authors using the Fourier transformation deduce the definition of the IMFs energy as:

$$E_i = \frac{1}{T}\sum_{t=1}^{T} d_i(t)^2 \tag{2}$$

The answer to the question of whether the given signal contains information or not can be given in a probabilistic manner. The criteria would build upon if the given IMF's energy exceeds some threshold, e.g. falls in the *confidence interval*. To find how can we obtain this confidence interval we need to look at the energy distribution for the IMFs that were produced by trying to encode noise. The authors of the mentioned paper obtain the IMFs energy statistics by running the Monte-Carlo simulation of EMD on uniform noise. The result can be seen in Fig. 1:

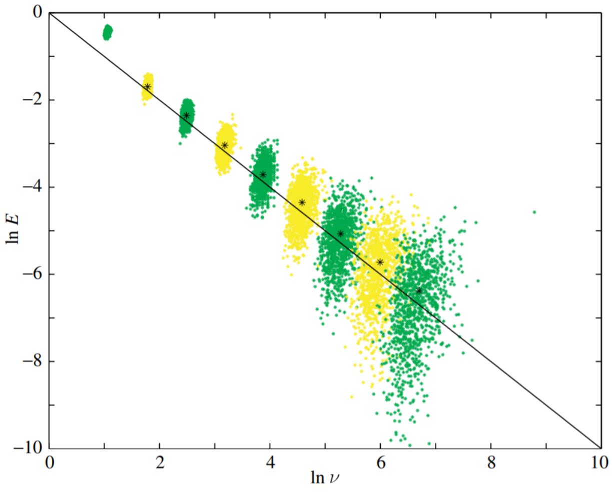

Figure 1: Relation between the average IMF frequency ν and its energy E (adapted from [Flandrin *et al.*, 2004c]).

From the Figure we can observe the law of the energy variance depending on its IMF's frequency, which means that we can estimate how likely a given IMF encodes only noise. The authors deduce the lines of confidence levels using the theoretical deduction. The resulting 99% confidence interval lines can be seen on the Fig. 2:

In contrast, the studied article [Flandrin *et al.*, 2004a] takes the empirical approach, and finds the same confidence interval line by analyzing the Monte-Carlo energy spread data. The big difference is that the given article studies a fractional Gaussian Noise (fGn), but not the uniform or regular Gaussian. The fGn is a broad-band noise with no dominant frequency band, which makes it a good property for the task of mode decomposition. Noise that the authors are using is also known as a Fractional Brownian motion [Decreusefond, Laurent and Üstünel, Ali Suleyman, 1998] which is zero-mean, has a property of long-range dependence and is controlled by the *Hurst exponent* $H \in [0, 1]$:

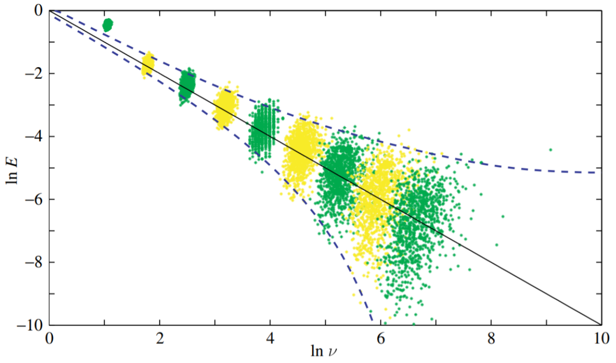

Figure 2: Energy E from frequency ν statistics and corresponding 99% confidence interval line (adapted from [Flandrin *et al.*, 2004c]).

- $H > \frac{1}{2}$ means that increments are positively correlated
- $H = \frac{1}{2}$ means that increments are independent
- $H < \frac{1}{2}$ means that increments are negatively correlated.

The results of the Monte-Carlo simulation for different values of H can be seen in Fig. 3

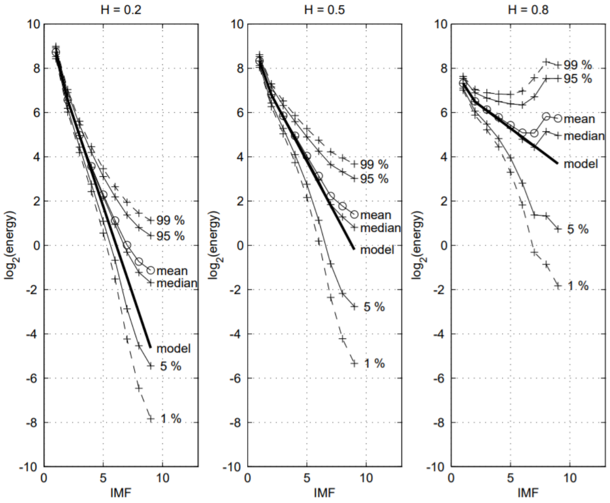

Figure 3: IMFs energy E statistics with the bold line corresponding to theoretical model given by the (3) (adapted from [Flandrin *et al.*, 2004a]).

The theoretical energy (which we will need as a reference point for our relative confidence interval) is deduced in [Flandrin *et al.*, 2004b], and equals to:

$$\hat{E}_H[i] = C_H \rho_H^{-2(1-H)i},\ i \geq 2 \tag{3}$$

where $C_H = \hat{E}_H[1]/\beta_H$, β_H is obtained from the experimental data given in the Fig. (3.6) of [Flandrin *et al.*, 2004b] and available in Table 1, ρ_H is a dependence of IMF frequency decline with IMF index from H, which is approximately 2 for all $H \in [0, 1]$.

H	β_H	$a_H(99\%)$	$b_H(99\%)$
0.2	0.487	0.452	-1.951
0.5	0.719	0.460	-1.919
0.8	1.025	0.495	-1.833

Table 1: Confidence Interval Parameters for the Linear Model (from [Flandrin *et al.*, 2004b])

Now let us compute the relative confidence interval for a given H from the data of the Fig. 3. To that end, we select a 99% percentile curve $T_H[i]$, and an energy estimate $\bar{E}_H[i]$ that in the case of our data is best of all picked as a median. By parameterizing this value as $log_2(log_2(T_H[i]/\bar{E}_H[i]))$, we can plot the dependency on the Fig. 4:

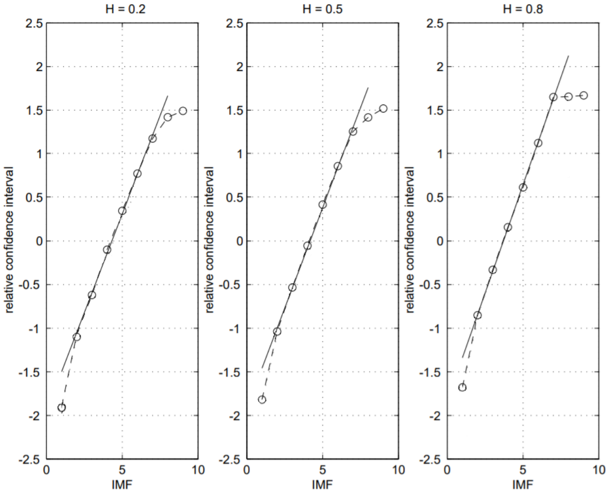

Figure 4: IMF energy relative confidence intervals computed as a log log of 99% percentile divided my a median estimate (dashed), best linear fit (solid) (adapted from [Flandrin *et al.*, 2004a]).

The graph indicates that we are able to express the relative confidence interval by using a linear function that can be expressed as:

$$log_2(log_2(T_H[i]/\bar{E}_H[i])) = a_H i + b_H \tag{4}$$

with the resulting a_H and b_H values given in the Table 1.

Now for us to decide which IMFs to keep we can go through this steps:

1. compute all IMF's energies $E_H[i]$ using (2)
2. get the model energy estimates of $\hat{E}_H[i]$ using $E_H[1]$ and (3)
3. get confidence interval $T_H[i]$ for all i using $(a_H i + b_H)\ \hat{E}_H[i]$
4. if $E_H[i]$ exceeds $T_H[i]$, it can be considered signal component, otherwise — noise.

By doing a linear sum of the IMFs that are selected by this procedure one can obtain denoised or detrended versions of the source signal.

4 Experiments

4.1 Simulated

To verify the correctness of the implemented procedure the experiment is conducted on synthetic data that is produced by mixing the clear signal with a fGn noise of $H = 0.2$ as seen in Fig. 5.

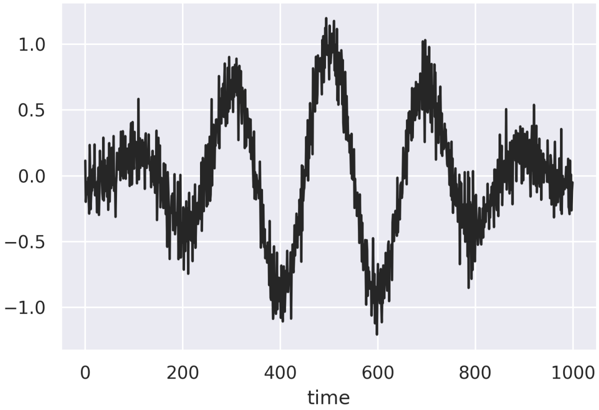

Figure 5: Synthetic signal with noise.

IMFs we got after applying EMD are pictured at the Fig. 6:

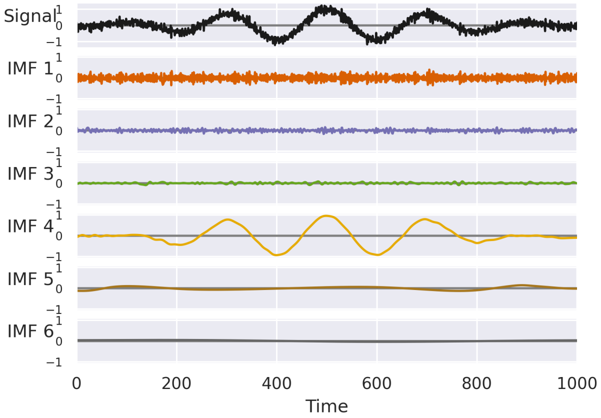

Figure 6: IMFs for the synthetic signal.

We can observe that, as expected, the first IMFs are corresponding to the noise part of the signal. IMF number 4 contains the majority of the signal, but it can be seen that the $5th$ component carries some oscillations that contribute to the beginning and to the end of the signal.

Now let us apply the algorithm described in the previous section and compute the criteria for IMF filtering. The resulting curves are presented at the Fig. 7:

We can see that the graph is in accordance with our observations of IMFs. Indeed, $4th$ and $5th$ components are considered to contribute to the signal, however, the $6th$ IMF is also included despite having no distinct signal features. The

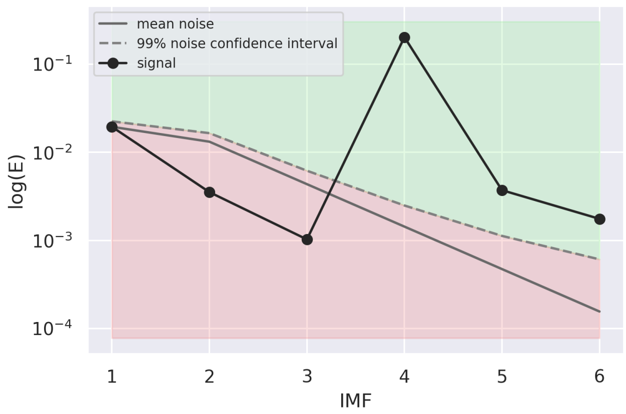

Figure 7: Signal IMF's energies and their estimated statistical characteristics. IMF components with $\hat{E}_H[i] > T_H[i]$ (green zone) are considered signal-carrying , while others (red zone) — noise-carrying.

reason for that is that the last component carries the small oscillations of low frequency that were not included in the previous two components, but contribute to the signal.

To get the denoised signal, we remove the components that are not included in the confidence interval and do a linear sum of the rest. You can see the result on the Fig. 8:

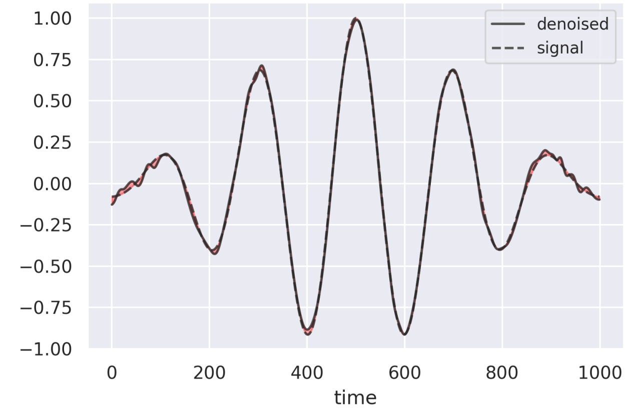

Figure 8: Denoised and pure signal. The difference is highlighted in red.

The denoised signal is very close to the pure version, which validates the correctness of the hypothesis and implemented algorithm. The difference in the middle part of the signal is negligible, however, on the edges, we can observe oscillations that are not present in the pure signal. The latter phenomenon can be explained by the contribution of two factors. First, EMD itself is known to have artifacts at the end of the signal due to the cubic interpolation (which should not play a role for this particular signal), second, noise to signal ratio falls closer to edges, making it harder to decompose the signal correctly.

As for the detrending part, the results are presented in Fig. 9

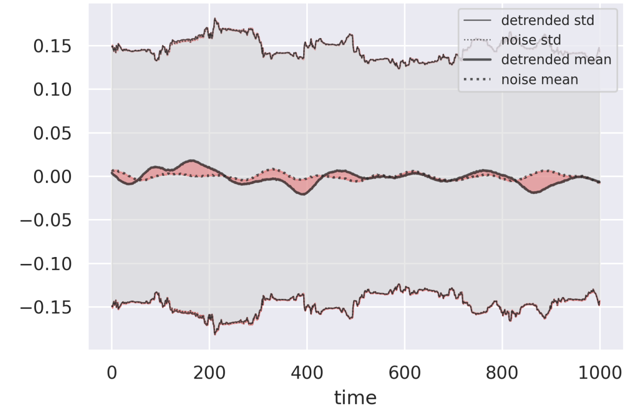

Figure 9: Detrended and denoised statistics. The difference is highlighted in red.

We can see that detrending removed all the signal components leaving only noise as a residue. The difference of a known noise std and detrended signal std is almost non-existent, while the discrepancy in mean has a maximum absolute value of only 0.2 for the signal of amplitude 1.

4.2 Fiber laser ACF

For the sake of seeing the limitations of this approach let's test it on unfit data. The first signal would be the Auto Correlation Function (ACF) of a fiber laser impulse [Kokhanovskiy *et al.*, 2019]. The issue we have had when working with these impulses was that their background noise interfered with some of the transformations that were necessary for the research. The example of the signal is demonstrated at the Fig. 10.

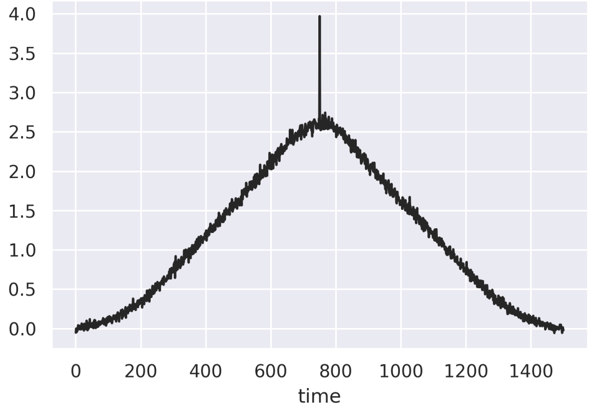

Figure 10: Fiber laser impulse ACF.

The problem with that signal is the fact that we don't know if the noise was following the fGn distribution, and we also don't know the H of that noise. Following the assumption of noise being isotropic, which is the default property of most real-world noises, we pick the $H = 0.5$, and since fGn with $H = 0.5$ is effectively a Gaussian distribution law, it fits the nature of the task. Figures 11, 12, and 13 display the EMD, energy statistics, and denoised signal respectively.

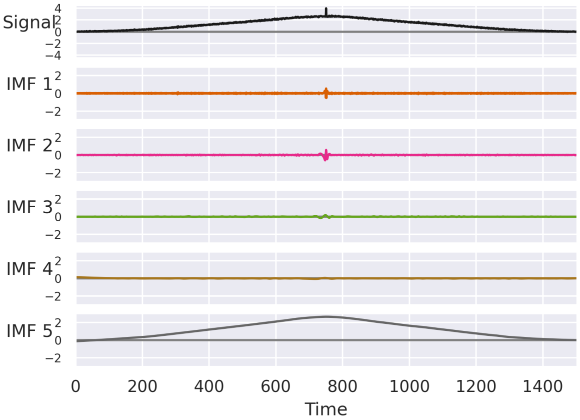

Figure 11: Fiber laser impulse ACF EMD.

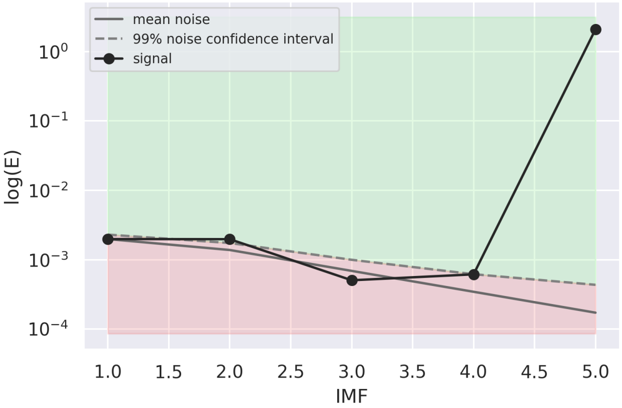

Figure 12: Fiber laser impulse ACF IMF's energies and their estimated statistical characteristics.

We can see that the result of the denoising while being good on most of the signal, negatively impacted the autocorrelation peak, changing its structure from δ-function-like, and reducing its peak to max envelope height ratio, which is an important characteristic used in the task. Despite these flaws, the decision algorithm based on the confidence interval has actually included not only the signal main shape but the second IMF also, which could be easily interpreted as noise. Coupling it with a fact that we have only guessed the H of the process, it gives a satisfactory result, which could be improved by further signal composition that includes some task apriori knowledge.

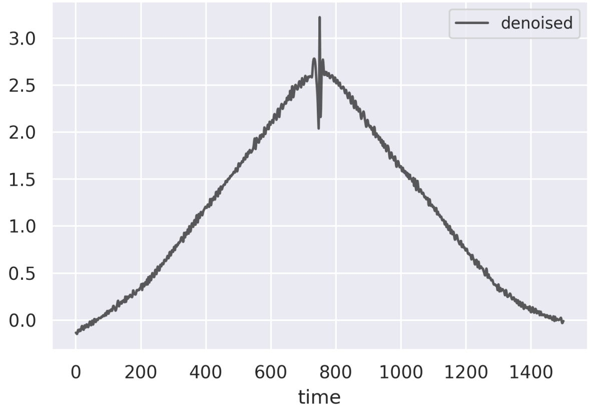

Figure 13: Denoised fiber laser impulse ACF.

4.3 Temperature data

As an example of data that is complete does not lie under the constraints of these methods we can apply the algorithm to the weather dataset [Beniaguev, 2017]. We will take the temperature in Portland city as a source of the signal. You can see the initial data, and the denoised signal on the Fig. 14 and Fig. 15.

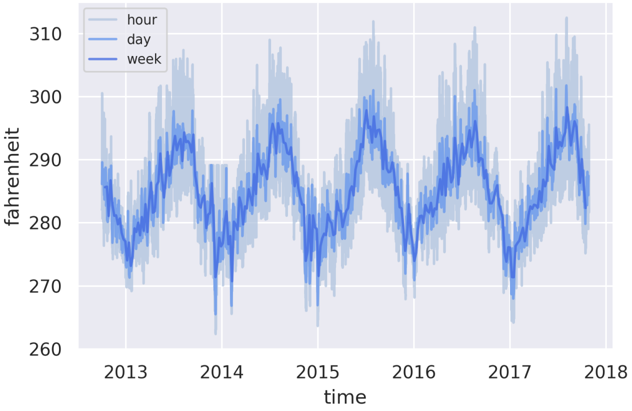

Figure 14: Temperature in the Portland city resampled by different frequency.

It selects only the IMF that corresponds to the yearly temperature oscillations. The $Hurst$ exponent for this process was chosen to display a positive correlation at a level of 0.8 based on a hypothesis that naturally occurring processes often have a positive Hurst exponent [Hurst, 1951].

5 On Hurst exponent estimation

As it is clear from the experiments, the EMD detrending and denoising process are accompanied by the problem of noise characteristics estimation. The bright side of this is that the given technique has been proven to work both with uniform and Gaussian noise in [Wu and Huang, 2004], which do not

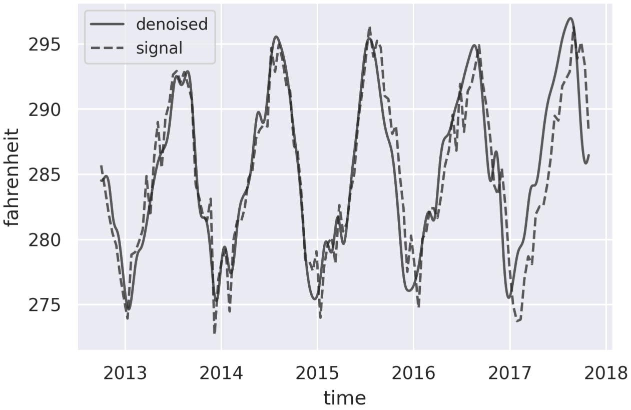

Figure 15: Denoised temperature signal (solid), 2 week mean resampling (dashed).

require H. However, Hurst exponent estimation, while not being in the scope of this work, has a solution.

The first way of estimating H is presented in [Flandrin *et al.*, 2004c], and based on the slope of the IMFs energy variance. They show that the variance slope values behave like $2(H-1)$ for $H > 1/2$, and have a nonlinear but prominent trend for $H < 1/2$. The real signal estimation procedure is not described, but it should involve calculating IMF's energies statistics by taking sub-parts of the signal.

The other way of solving this issue is estimating the Hurst exponent without the EMD by using the Hurst equation and a Rescaled Range technique. The approach combining these are implemented in the *hurst* python package [Mottl, 2021].

However, despite the fact that H estimation is a known technique, if we apply it to the signal directly, it can yield strongly biased predictions. It is clear that it is better to be done on a noise itself. The technique I propose is computing the EMD, and building a confidence interval for an arbitrary value of $H = 0.5$. The IMF components that are more likely to contain signals have the highest distance to that confidence interval. Removing the IMFs in this particular order will lead the resulting signal to be mostly comprised of noise. It has been tested against the approach of taking a full signal, and it has slightly increased the H exponent estimation quality.

6 Conclusion

In the scope of this work, the article ”Detrending and denoising with empirical mode decompositions technique” by [Flandrin *et al.*, 2004a] was studied, given algorithm was implemented and tested against both model-compliant synthetic data and non-fit real-world signals. The work went beyond a particular paper summarizing and compiling different aspects of related articles. The method was shown to be viable even in the case of non-standard data. The slightly improved Hurst exponent estimation method was proposed and tested showing marginal improvement in H estimation quality.

References

[Beniaguev, 2017] David Beniaguev. Historical hourly weather data 2012-2017, 12 2017.

[Decreusefond, Laurent and Üstünel, Ali Suleyman, 1998] Decreusefond, Laurent and Üstünel, Ali Suleyman. Fractional brownian motion: theory and applications. *ESAIM: Proc.*, 5:75–86, 1998.

[Flandrin *et al.*, 2004a] P. Flandrin, P. Gonçalvès, and G. Rilling. Detrending and denoising with empirical mode decompositions. In *2004 12th European Signal Processing Conference*, pages 1581–1584, 2004.

[Flandrin *et al.*, 2004b] P. Flandrin, G. Rilling, and P. Goncalves. Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, 11(2):112–114, 2004.

[Flandrin *et al.*, 2004c] Patrick Flandrin, Paulo Gonçalvès, and Gabriel Rilling. *EMD EQUIVALENT FILTER BANKS, FROM INTERPRETATION TO APPLICATIONS*, pages 57–74. 2004.

[Hurst, 1951] H. E. Hurst. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116(1):770–799, 1951.

[Kokhanovskiy *et al.*, 2019] Alexey Kokhanovskiy, Anastasia Bednyakova, Evgeny Kuprikov, Aleksey Ivanenko, Mikhail Dyatlov, Daniil Lotkov, Sergey Kobtsev, and Sergey Turitsyn. Machine learning-based pulse characterization in figure-eight mode-locked lasers. *Opt. Lett.*, 44(13):3410–3413, Jul 2019.

[Mottl, 2021] Dmitry Mottl. Hurst exponent evaluation and r/s-analysis, 2021.

[Wu and Huang, 2004] Zhaohua Wu and Norden E. Huang. A study of the characteristics of white noise using the empirical mode decomposition method. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 460(2046):1597–1611, 2004.