Circuits, Systems, and Signal Processing (2023) 42:4669–4690
https://doi.org/10.1007/s00034-023-02319-0

Local Change Point Detection and Cleaning of EEMD Signals

Kentaro Hoffman[1] · Jonathan Lees[2] · Kai Zhang[3]

Received: 19 March 2022 / Revised: 3 February 2023 / Accepted: 4 February 2023 /
Published online: 20 March 2023

Abstract

The ensemble empirical mode decomposition (EEMD) has become a preferred technique to decompose nonlinear and non-stationary signals due to its ability to create time-varying basis functions. However, current EEMD signal cleaning techniques are unable to deal with situations where a signal only occurs for a portion of the entire recording length. By combining change point detection and statistical hypothesis testing, we demonstrate how to clean a signal to emphasize unique local changes within each basis function. This not only allows us to observe which frequency bands are undergoing a change, but also leads to improved recovery of the underlying information. Using this technique, we demonstrate improved signal cleaning performance for acoustic shockwave signal detection.

Keywords Change point detection · Signal cleaning · EEMD · Sparsity · LCDSC

✉ Kentaro Hoffman
khoffm34@jh.edu

Jonathan Lees
jonathan.lees@unc.edu

Kai Zhang
zhangk@email.unc.edu

[1] Whiting School of Engineering, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218-2681, USA

[2] Department of Earth, Marine, and Environmental Sciences, University of North Carolina at Chapel Hill, 313 Mitchell Hall, Chapel Hill, NC 27599, USA

[3] Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, 318 Hanes Hall, Chapel Hill, NC 27514, USA

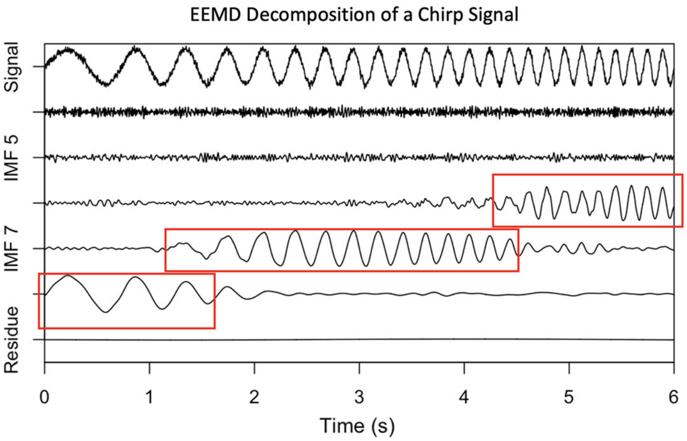

Fig. 1 A chirp with white noise decomposed by EEMD. The boxed-in areas identify when each basis function is picking up the sinusoidal signal. Notice how the increasing frequency of the sinusoid makes it such that no basis function picks up the signal for the entire duration

1 Introduction

The ensemble empirical mode decomposition (EEMD) method has become an important technique for the decomposition of nonlinear and non-stationary signals in fields including medicine [20, 22], hydrology [29], seismology [28], and mechanical engineering [6, 35]. A reason for its success has been EEMD's ability to create data-adaptive, rather than predefined, basis functions called intermediate mode functions (IMFs). These adaptive basis functions can be non-stationary and nonlinear, making them ideal for complex signals that are not as natural to express in Fourier or wavelet bases.

However, this data-adaptive nature of the EEMD's basis functions can make it hard to know *a priori* in which basis function a signal may end up. For instance, consider a chirp signal linearly increasing in frequency perturbed with white noise. When decomposed by EEMD, we can see in Fig. 1 that the signal glides between IMFs 8 and 6. Common EEMD signal cleaning techniques such as those used in [5, 8, 10, 15, 18, 19, 21, 32] first decompose the signal into its base IMF functions, but then treat the entire length of an IMF as either signal or noise. However, in this example, due to the increasing frequency of the chirp signal, no basis function is consistently signal or noise. To properly clean this signal, a more nuanced technique that is able to identify subsections of IMF as signal or noise is necessary. In this paper, we provide a novel example of an EEMD signal cleaning technique, local change detection, and signal cleaning (LCDSC), that is able to identify and clean subsections of EEMD signals. Moreover, we show how this technique can improve the identification of acoustic shock waves.

2 Local Change Point Detection and Signal Cleaning

2.1 EMD

The empirical mode decomposition (EMD) was invented in 1998 as a novel technique for analyzing nonlinear and non-stationary time series data [12]. Using iteratively computed, adaptive filters, the EMD performs an additive decomposition of a signal $X(t)$ into:

$$X(t) = \sum_{j=1}^{n} \mathrm{IMF}_j(t) + r(t). \tag{1}$$

Here, $\mathrm{IMF}_j(t)$ is a approximately narrow-banded the jth basis function, referred to as an intermediate mode function (IMF), and $r(t)$ is the residual. As the EMD is a numerical algorithm, there exists a variety of stopping criteria to indicate when the algorithm has converged. One of the most common stopping criteria, S-stoppage, [13] results in the remainder term $r(t)$ becoming a monotonic or a constant function. In either case, the resulting $r(t)$ can easily be subtracted from the original signal $X(t)$ to create a decomposition with no residual term. Thus, for the purposes of this paper, we will assume that either $X(t)$ has an $r(t)$ of 0, or $X(t)$ has had its remainder subtracted out resulting in

$$X(t) = \sum_{j=1}^{n} \mathrm{IMF}_j(t). \tag{2}$$

Using the Hilbert transform, each IMF_j's instantaneous amplitude $a_j(t)$ and instantaneous frequency $w_j(t)$ time series can be extracted as the sum of a time-varying amplitude function $a_j(t)$ multiplied by an equally time-varying frequency function $\mathrm{e}^{i w_j(t)}$,

$$X(t) = \sum_{j=1}^{n} \mathrm{IMF}_j(t) = \mathcal{R}\left[\sum_{j=1}^{n} a_j(t)\mathrm{e}^{i w_j(t)}\right]. \tag{3}$$

As $a_j(t)$ and $w_j(t)$ are functions of time, this decomposition allows for the analysis of time-varying amplitude and frequency signals. This contrasts with the Fourier decomposition in which the amplitude a_j and instantaneous frequency w_j are no longer functions of time, but constants.

IMFs also come with several crucial properties. By definition, an IMF is a nonlinear oscillatory function that satisfies the requirements [2]:

1. For each IMF, the number of local extrema and zero crossings must differ by at most one.
2. Let $g_{j,\max}(t)$ and $g_{j,\min}(t)$ be smooth functions connecting the local maxima and minima of the jth IMF (These functions are commonly referred to as the upper

and lower envelope of $X(t)$). At any time point t, the mean of the upper envelope of the jth IMF, $g_{j,\max}(t)$, and the lower envelope, $g_{j,\min}(t)$, is zero:

$$g_{j,\max}(t) + g_{j,\min}(t) = 0. \tag{4}$$

Moreover, [14] illustrated that IMFs are close to mutually orthogonal; this allows us, with a high degree of accuracy, to decompose the total energy of the signal, $[\sum_{t=1}^{T} X(t)^2]$ into the sum of the energy of the individual IMFs:

$$\sum_{t=1}^{T} X(t)^2 = \sum_{t=1}^{T} \sum_{j=1}^{n_{\text{IMF}}} \text{IMF}_j(t)^2 \tag{5}$$

2.2 EEMD

While the EMD demonstrated its effectiveness as a signal decomposition tool, it was noted in [11] that signals which exhibit intermittency can lead to a phenomena dubbed “mode-mixing” where an IMF fails to separate the high-frequency intermittency from more continuous behavior. This behavior ruins the desirable narrow-bandedness of IMFs potentially complicating the analysis [11]. To account for this, [33] developed a variant of EMD known of the “Ensemble Empirical Mode Decomposition” (EEMD). In the EEMD, one generates many signals $\tilde{X}(t)$ which are perturbed with white noise $w(t)$, $\tilde{X}(t) = X(t) + w(t)$. $\tilde{X}(t)$ is then decomposed using EMD into its constituent IMFs. This procedure is then repeated over multiple replicates, and the IMFs from each replicate are average together. These ensembled IMFs demonstrate greater robustness to mode-mixing, which improves the decomposition and delineation of signal.

2.3 Additive Noise Model

In performing local change point detection, we will operate under the assumption that there exists an observed signal $X(t)$ that consists of mean zero Gaussian noise $R(t)$ occurring throughout the entire duration and an underlying true signal $S(t)$ which is only observable during the interval A. If we assume an additive decomposition, this gives the setup

$$X(t) = S(t)I_A(t) + R(t), \tag{6}$$

where $I_A(t)$ is an indicator function that returns 1 if $t \in A$ and 0 otherwise. The additional assumption of statistical independence between $R(t)$ and both $S(t)$ and the set A completes the additive local noise model.

2.4 Change Point Detection of the IMFs

Under the additive local noise model, the goal of signal cleaning is to recover the true signal $S(t)$ by first estimating the interval A, or when the true signal is occurring, and then performing a signal cleaning on $X(t)$ for $t \in A$ to recover $S(t)$. To identify when changes are occurring in $X(t)$, we first decompose $X(t)$ into its constituent IMFs and then perform a change point detection procedure on each IMF. Here, IMF is the ensembled IMF from the runs of the EEMD. From a statistical perspective, identifying change points entails finding the set of time points $\{\tau_1^{(i)}, \ldots, \tau_{n_j}^{(i)}\}$ such that:

$$\begin{aligned} f(\mathrm{IMF}_j(t_1)) \neq f(\mathrm{IMF}_j(t_2)),& \\ \forall t_1 \in [\tau_k^{(i)}, \tau_{k+1}^{(i)}],& \\ \forall t_2 \in (\tau_{k+1}^{(i)}, \tau_{k+2}^{(i)}],& \\ \forall k \in [1, \ldots, k-2].& \end{aligned} \tag{7}$$

Here, $f(\mathrm{IMF}_j(t))$ represents the distribution of the ensembled IMF_j at time t. However, as the distribution of each IMF is generally unknowable *a priori* outside of well-known distributions such as white noise [30], it can be difficult to create a change point detection algorithm that is able to rapidly identify when a change is occurring. To make this more tractable, we utilize several of the properties of IMFs and the additive local noise model to construct a more feasible change point detection problem.

According to our additive local noise model

$$X(t)^2 = \begin{cases} (S(t) + R(t))^2 & \text{If } t \in A \\ R(t)^2 & \text{If } t \notin A. \end{cases} \tag{8}$$

Combining this with the statistical independence between $R(t)$, $S(t)$ and A we assumed in the additive local noise model, this implies that

$$E\big[X(t)^2\big] = \begin{cases} E\big[S(t)^2\big] + E\big[R(t)^2\big] & \text{If } t \in A \\ E\big[R(t)^2\big] & \text{If } t \notin A. \end{cases} \tag{9}$$

Thus, when we are in interval A, there is an increase in expected power in $X(t)$ (power being $X(t)^2$). Furthermore, by the orthogonality of the IMFs, this directly implies that an increase in power in $X(t)$ must lead to a corresponding increase in at least one of the constituent IMFs. Formally, if $t \in A$, then there exists a subset of IMFs $\eta \subset \{1, \ldots, n\}$ such that for $j \in \eta$, $\mathrm{IMF}_j(t)$ displays an increase in power during $t \in A$.

$$\forall t \in A, t^* \notin A, \exists j \in \eta \neq \emptyset : E\left(\mathrm{IMF}_j(t)^2\right) \geq E\left(\mathrm{IMF}_j(t^*)^2\right) \tag{10}$$

Additionally, as each IMF has a mean of zero with respect to its envelope, an increase in power in an IMF implies an increase in the variance in that IMF

$$E\left[\mathrm{IMF}_j(t)^2\right] = E\left[(\mathrm{IMF}_j(t) - E(\mathrm{IMF}_j(t))^2\right] = Var\Big(\mathrm{IMF}_j(t)\Big). \tag{11}$$

Using Eq. 3, we can write the variance of an IMF as:

$$\mathrm{Var}\Big(\mathrm{IMF}_j(t)\Big) = Var(\mathcal{R}[a_j(t)\mathrm{e}^{iw_j(t)}]) \tag{12}$$

If the variance of an IMF increases, this could be because $a_j(t)$ has changed or because $\mathrm{e}^{iw_j(t)}$ has changed. However, in simulation, we observe that changes in the variance of an IMF are better expressed in the amplitude term, $a_j(t)$ rather than the frequency term, $\mathrm{e}^{iw_j(t)}$. Thus, to identify a local signal, we will look for IMFs which are exhibiting a change in the variance of their amplitudes.

2.5 Change Point Detection

To identify when the amplitude of an IMF is experiencing an increase in variance, we employ techniques from statistical change point detection. Many change point detection problems can be framed in the form of minimizing an objective function of the form:

$$\min_{m} \min_{\tau_1,\ldots\tau_{m-1}} \sum_{i=1}^{m-1} L\big(X_{\tau_{i-1}}, X_{\tau_i}, X_{\tau_{i+1}-1}\big) + \beta D(m), \tag{13}$$

where τ_0 is 1 and τ_m is the length of the signal, m is the number of change points, τ_i is the location of the ith change point, β is a constant, L is a function that decreases when τ is a true change point, and $D(m)$ is a penalization function that increases with the number of change points selected. By balancing L and $D(m)$, the objective seeks to select the correct number and locations of changes in variance.

For our particular type of local signal, since the background noise is Gaussian, we focus our attention to L and $D(m)$ that is well suited to noticing changes in Gaussian signals. One such L is the likelihood ratio test for changes in variance of Gaussians [16].

$$L\big(X_{\tau_{i-1}}, X_{\tau_i}, X_{\tau_{i+1}-1}\big) = \frac{C_{\tau_i}}{C_{\tau_{i+1}-1}} - \frac{\tau_i - \tau_{i-1}}{\tau_{i+1} - 1 - \tau_{i-1}} \tag{14}$$

Here C_{τ_i} is the cumulative normalized second moment, $\sum_{k=\tau_{i-1}+1}^{\tau_i}(X(k) - \overline{X_{\tau_i}})^2$ and $\overline{X_{\tau_i}}$ is the cumulative mean, $1/(\tau_i - \tau_{i-1} + 1)\sum_{k=\tau_{i-1}+1}^{\tau_i} X(k)$. This L has the ability not only to consistently select the correct location, but correct number of change points under an asymptotic scheme but also has strong performance in the finite sample case [16]. As for $\beta D(m)$, this is a penalization term that combines some function of the number of change points, $D(m)$, with a constant, β to ensure that the correct number of change points are selected [26]. While there exist many popular penalty terms such as Akaike's information criterion (βm) [1] and Bayesian information criterion ($m \log(n)$) [25] (n is the total signal length), many still lack theoretical justifications in the context of change point detection. One exception is the modified Bayesian information criterion (mBIC)

Table 1 List of EMD/EEMD signal cleaning techniques

Cleaning method	Description
LCDSC	Our method
k-Highest	Removal all but the k-highest IMFs [32]
l-Lowest	Removal all but the k-lowest IMFs [10]
k-Highest and l-Lowest	Combination of k-Highest and l-Lowest [5]
Power set cleaning	Perform a best subset selection over all possible subsets
WHT	Wavelet hard thresholding each IMF [17]
WIT	Wavelet interval thresholding each IMF [17]
No cleaning	No signal cleaning

$(-\frac{1}{2}(3\,m + \log(n) + \sum_{i=1}^{m+1} \log(\tau_i - \tau_{i-1}))$ [34]. The modified Bayesian Information Criterion frames change point detection as a model selection problem, where we are choosing between Gaussian processes that have differing number and size of change points even allowing for the detection of changes in means and variances. Under this perspective, the number and location of change points returned by mBIC accord with the model that yields the largest Bayes factor. For these principled properties, in our discussions below, we will be employing the mBIC as our penalization term.

2.6 IMF Cleaning

Once the significant segments are identified, we must determine how a signal is cleaned. Similar to how high-pass, low-pass, and band-pass filters clean signals by removing basis functions that are beyond a given threshold in the Fourier domain, in EEMD signal cleaning, there exist methods which remove IMFs which seem to contain background noise (see Table 1 for examples). In this spirit, we will set a signal segment to 0 if it is not identified as containing a significant spike in amplitude compared to surrounding segments. More specifically, we would like to test:

$$\begin{aligned} H_0 &: \sigma^2_{\text{during}} \leq \gamma * \max\left(\sigma^2_{\text{before}}, \sigma^2_{\text{after}}\right) \\ H_1 &: \sigma^2_{\text{during}} > \gamma * \max\left(\sigma^2_{\text{before}}, \sigma^2_{\text{after}}\right), \end{aligned} \tag{15}$$

where σ^2_{before} is the variance of the previous interval, σ^2_{during} is the variance of the current interval, σ^2_{after} is variance of following interval, and γ is assumed to be greater than or equal to 1.

By rearranging the alternate hypothesis, $\gamma > \frac{\sigma_{\text{during}}}{\max(\sigma_{\text{before}}, \sigma_{\text{after}})}$, we can see that γ serves as a measure of how much the ratio of variances much increase to be considered significant. Setting $\gamma = 1$ tests if there has been any statistically significant increase in variance. The test statistic for (15) is the F-statistic for change in variance

$$F_{\text{before/during}} = \frac{\gamma * \max\left(S^2_{\text{before}}, S^2_{\text{after}}\right)}{S^2_{\text{during}}},$$

where S_{before} is the sample variance used to estimate σ_{before}. $F_{\text{before/during}}$ is compared against the F distribution with degrees of freedoms, $df_1 = n_{\text{during}}$, $df_2 = \max(n_{\text{before}}, n_{\text{after}})$ (where n_{during} is the length of the during interval) to determine the p value and thus significance.

As this process involves performing a hypothesis test at every potential change point, across every IMF, this can quickly lead to a large number of tests being performed for the same goal: identifying a significant segment. This large number of tests can lead to the multiplicity issue where one or more spurious false positives may occur. To perform these tests so they collectively have an α $(1 > \alpha > 0)$ probability of a false positive (which is known as the family-wise error rate), we employ the multiple testing correction method, Holm–Bonferroni method [9]:

The Holm–Bonferroni Procedure

1. Say that in total, K hypothesis tests were performed with p values. Sort the p values from smallest to largest to get: $p_{(1)}, \ldots, p_{(k)}$
2. If $p_{(1)} \geq \frac{\alpha}{K}$, none of the tests are significant. Otherwise, continue.
3. Test the second p value. If $p_{(2)} \geq \frac{\alpha}{K-1}$, then the procedure is stopped and no further p values are significant. Otherwise, continue testing till all p values are significant or the ith p value is such that:

$$p_{(i)} \geq \frac{\alpha}{K - i + 1}.$$

If the p value is significant after the Holm–Bonferroni correction, then we can claim that the interval contains the desired signal. If not, the interval does not contain the true signal and is cleaned by setting it to 0. If an IMF only has one change point (and thus cannot have a before, during, and after interval), then $\max(\sigma^2_{\text{before}}, \sigma^2_{\text{after}})$ is replaced with σ^2_{after}. If there are no change points, then the entire IMF is set to 0. If there are no change points in an IMF, then there is no identifiable local signal so the entire IMF is set to 0. We note here that if a segment of an IMF is considered significant, that segment is included in its entirety. It is likely that one can further improve the performance by including a smoothing or SURE-based cleaning procedure to clean the significant segments in addition to setting nonsignificant segments to zero. However, to do this, careful work must be put into determining the appropriate level of cleaning for each IMF, a nontrivial question which may require stronger assumptions.

Algorithm 1 Cleaning of nonlinear signal $X(t)$ using `LCDSC`

Require: Nonlinear signal $X(t)$, number of EEMD replications n_{EEMD}, Type I Error control level α, sensitivity parameter γ

Ensure: Cleaned IMFS, $CIMF_i(t)$

```
function LCDSC($X(t), n_{EEMD}, \alpha$)
    for $l \leftarrow 1$ to $n_{EEMD}$ do
        $\tilde{X}(t) \leftarrow X(t) + w(t)$
        $EMDIMF_{l,i}(t) \leftarrow$ EMD Cleaned IMF
    end for
    $n_{IMF} \leftarrow$ Number of IMFs generated
    for $i \leftarrow 1$ to $n_{IMF}$ do
        $EEMDIMF_i \leftarrow \sum_{l=1}^{n_{EEMD}} (EMDIMF_{l,i})/n_{EEMD}$
        $\{S_{i,1}, \ldots. S_{i,n_{seg(i)}}\} \leftarrow$ Segments of $EEMDIMF_i$ identified by the change point detection algorithm with mBIC. $n_{seg(i)}$ is the number of segments.
    end for
    $n_{tests} \leftarrow n_{seg(1)} + \cdots + n_{seg(n_{IMF})}$
    for $i \leftarrow 1$ to $n_{IMF}$ do
        if $n_{seg(i)} = 1$ then
            $CS_1(t) \leftarrow 0$
        end if
        if $n_{seg(i)} = 2$ then
            if $\sigma_{i,1}$ significantly larger than $\gamma\sigma_{i,2}$ after Bonferroni Holm then
                $CS_2(t) \leftarrow 0$
            end if
            if $\sigma_{i,2}$ significantly larger than $\gamma\sigma_{i,1}$ after Bonferroni Holm then
                $CS_1(t) \leftarrow 0$
            end if
        end if
        if $n_{seg(i)} > 2$ then
            for $l \leftarrow \{2, \ldots, n_{seg(i)} - 1\}$ do
                if $\sigma_{i,l}$ not significantly larger than $\gamma Max(\sigma_{i,l-1}, \sigma_{i,l+1})$ after Bonferroni Holm then
                    $CS_{i,l}(t) \leftarrow 0$
                end if
            end for
        end if
        $CIMF_i(t) \leftarrow$ Concatinate the cleaned segments $\{CS_{i,1}, \ldots. CS_{i,n_{seg(i)}}\}$
    end for
    return $\{CIMF_1(t), \ldots CIMF_{n_{IMF}}(t)\}$
end function
```

3 Simulation

3.1 Simulation 1: Doppler Signal

To demonstrate this signal cleaning procedure, we take a synthetic example where a Doppler signal is hidden in the midst of Gaussian white noise. The Doppler is a classic example of a nonlinear signal with variable frequency, exactly the kinds of signals that the flexible EMD algorithm is well suited for. We will refer to this as the local Doppler example. For Simulation 1, we will use a local Doppler of length 2000 with the Doppler occurring during the middle of the signal:

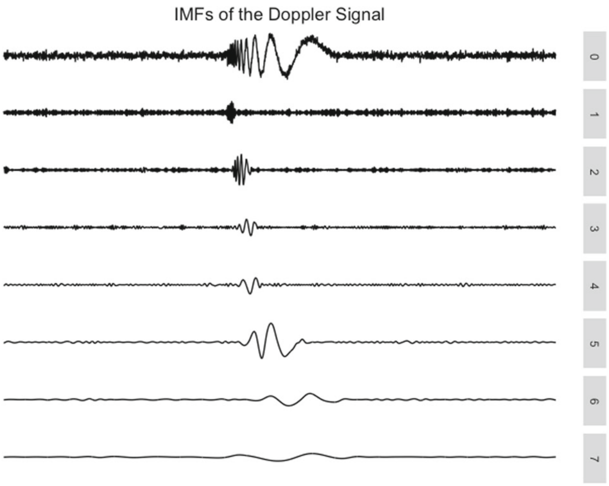

Fig. 2 EEMD of the local Doppler signal. In the EEMD, none of the IMFs are purely signal or noise necessitating a local signal cleaning procedure

$$X(t) = \begin{cases} R(t) & \text{if } t < 1500, t > 1500 \\ S(\frac{t-1000}{1500}) + R(t) & \text{if } 1000 \leq t \leq 1500. \end{cases} \tag{16}$$

$S(t)$ is the Doppler signal from [7] rescaled to occur between [1000,2000]:

$$S(t) = 7(t(1-t)^{0.5} \sin(2\pi(1+0.05)/(t+0.05)). \tag{17}$$

The goal of the this simulation would be to have the algorithm:

- Identify when the signal started and ended (time points: 1000–1500)
- Clean the Signal that was isolated

As can be seen in Fig. 2, the Doppler signal in the middle is expressed in most IMFs with the lower IMFs expressing the higher frequency parts of the signal and the latter IMFs expressing the lower frequency sections. Moreover, no single IMF is ever purely signal or purely noise necessitating a local change point detection and signal cleaning. All figures were generated using R 4.2.2.

Running the change point detection algorithm in Fig. 3 at an $\alpha = 0.05$ type I error level and $\gamma = 1$ identifies many locations at which a change in the signal was detected. While IMFs 1, 3–6 correctly identify two changes, one when the Doppler signal starts within their IMF and one when it ends, in IMFs 7–12, many spurious change points are detected that are not necessarily due to the Doppler signal. To remove these, the F-test cleaning step is performed.

The resulting cleaned signal in Figs. 4 and 5 illustrates how all of the change points outside of the duration of the Doppler signal were deemed nonsignificant by Holm–Bonferroni and set to zero. Not only does this provide a good estimation of the shape of the Doppler signal, matching the general sinusoidal shape and increasing frequency,

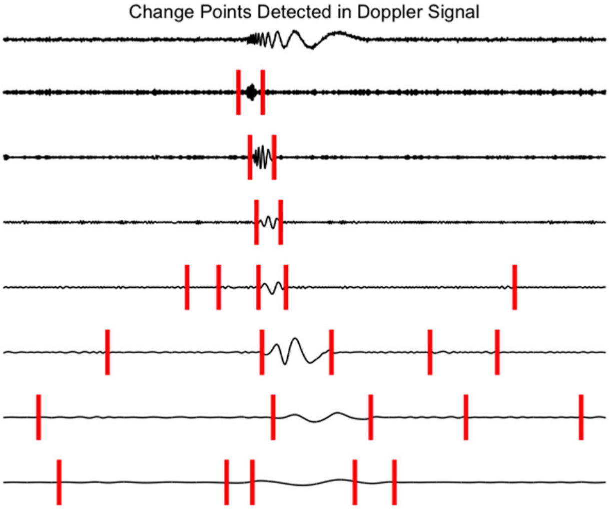

Fig. 3 Change points that were detected in the local Doppler signal in Fig. 2 when employing the normal likelihood ratio objective function and the Modified Bayesian Information Criterion over-fitting penalty

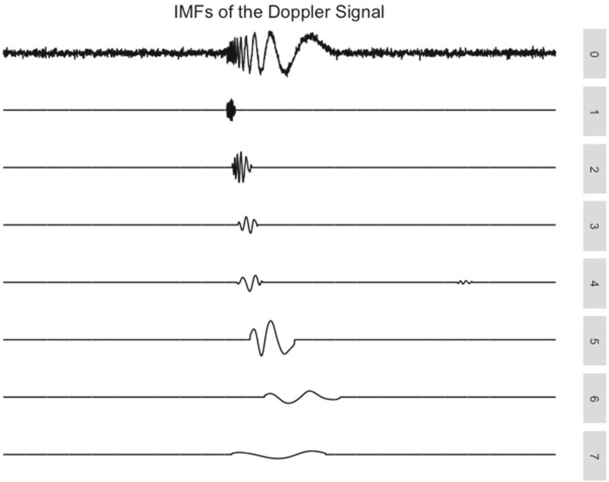

Fig. 4 IMFs in Fig. 3 after each section that was identified by the change point detection algorithm was cleaned using the F-test/Hole–Bonferroni procedure with $\gamma = 1$. Notice how the basis functions are set to 1 when the signal is not present within the basis function

but LCDSC provides a good estimate of when the Doppler signal starts, as the first nonzero point in IMF1 is at point 1010, only 1% of the way into the start of the Doppler signal.

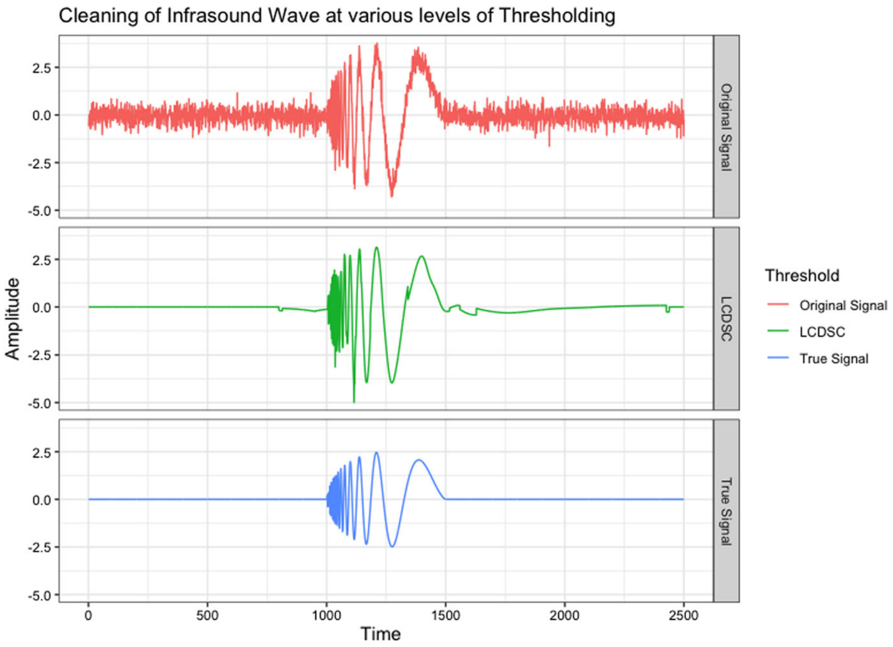

Fig. 5 Comparison of the original signal with the cleaned signal. The LCDSC recovers much of the original signal. It performs especially well at cleaning the signal to closely match the true start and end points

3.2 Simulation 2: Doppler Signal-Comparison Study

To compare the performance of our algorithm, we extend our Doppler simulation from Simulation 1 and compare our performance against other EEMD signal cleaning techniques. These techniques come in two general varieties. Techniques 2–5 in Table 1 are based on identifying some subset of the IMFs as containing only noise and cleaning the signal by completely removing the noise IMFs. With some of the cleaning procedures, the user must pre-specify how many basis functions to set to zero or clean through a trial-and-error process. To account for any possible variability in performance due to these subjective judgments, we will come up with an upper-bound for the performance of each algorithm by computing the best possible set of IMFs for each of the algorithms in question.

As for the wavelet hard thresholding (WHT) and wavelet interval thresholding (WIT) cleaning techniques, these are based on performing a wavelet-like thresholding on each of the IMFs [17]. These compute the base noise level within each IMF and perform a hard or soft thresholding if the IMF lies within the expected noise band. While this method does not suffer from a subjective choice of IMF removal, it assumes that the true signal occurs throughout the entire duration of the signal, leading to a biased estimation of the base noise level.

The data model for the simulation will utilize the local Doppler model with the middle containing our desired signal but with the total signal length T at differing values:

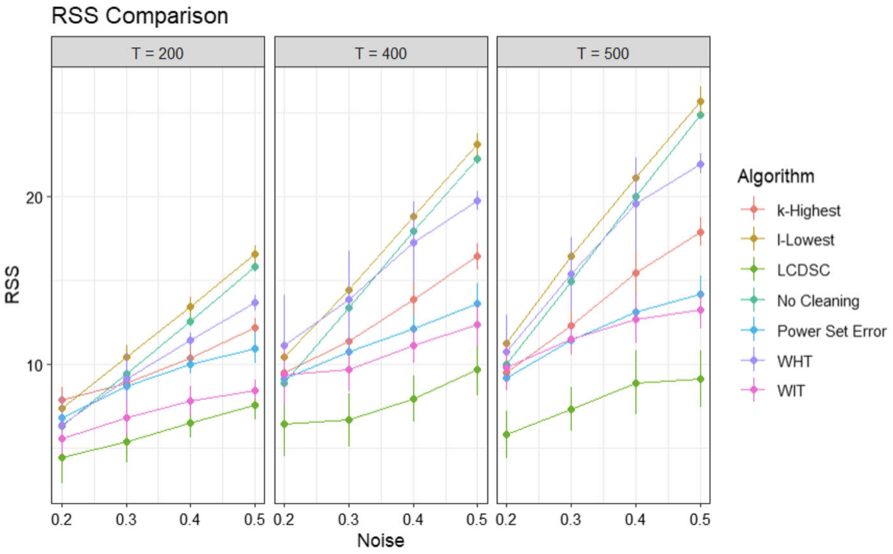

Fig. 6 RSS Comparison of common cleaning methods versus LCDSC. The center point represents the mean RSS across 20 replicates and the bar represents one standard deviation from the center. From this, we observe that LCDSC performs better than competing signal cleaning methods, able to create the closest representation of the true signal. Note that this is RSS and not MSE so it is entirely expected that as the signal gets longer, the RSS should also increase

$$X(t) = \begin{cases} R(t) & \text{if } t < \frac{2}{5}T, t > \frac{3}{5}T \\ S(t) + R(t) & \text{if } \frac{2}{5}T \leq t \leq \frac{3}{5}T. \end{cases} \tag{18}$$

T is tested at 1000, 2000, and 2500 time steps. $R(t)$ will again be Gaussian white noise but with the noise level varying from 0.2 to 0.5. The cleaned signal is then compared to the underlying Doppler signal and error computed in terms of residual sum of squares (RSS) as this corresponds to the total power difference between the estimated and the cleaned signal:

$$\text{RSS} = \sum_{t=1}^{T} (X(t) - \text{Cleaned}(t))^2. \tag{19}$$

At each level of noise and signal length, 20 replicates of the simulation were performed.

The results in Fig. 6 illustrate that across a wide scale of noise levels and sample sizes, the LCDSC performs well at local signal cleaning, uniformly outperforming other non-local signal cleaning techniques.

3.3 Simulation 3: Comparison Study—What if the Signal is Not Local?

While the LCDSC is built for the problem of local signal detection and cleaning, it is important to determine its performance as the duration of true signal is

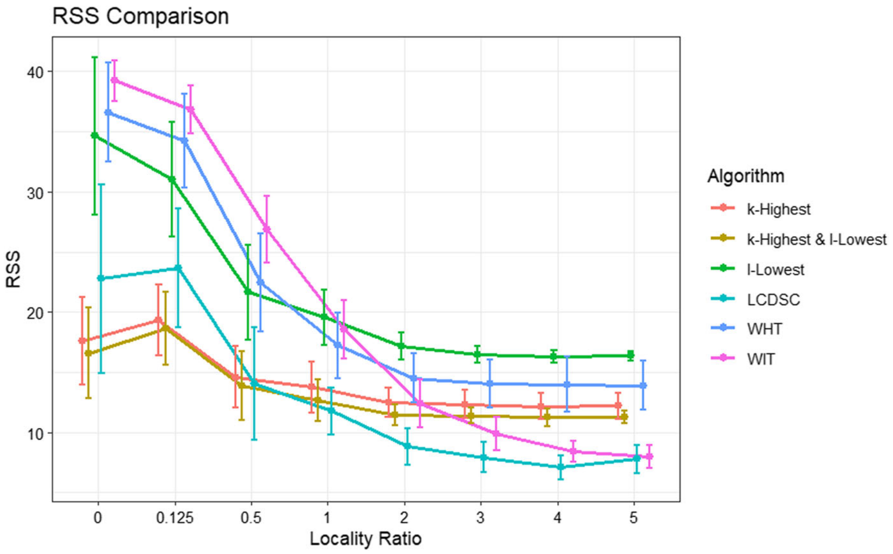

Fig. 7 Changes in residual sum of squares as the locality ratio is increased. When the noise ratio is low, the LCDSC performs slightly worse than k-Highest and k-Highest & L-lowest, but once the noise ratio increases above 1, the LCDSC becomes the best performing method. No cleaning was not plotted as it had a much higher error than all the others and power set was near equivalent to k-Highest & l-Lowest

increased or decreased. We can express how local our signal is in terms of a "locality Ratio":

$$\text{locality Ratio} = \frac{\text{len}(A)}{T - \text{len}(A)}. \tag{20}$$

$\text{len}(A)$ is the length of the interval A when the true signal is being expressed and T is the total length of the noisy signal. We vary the locality Ratio between 0 and 5, making the local signal cleaning problem increasingly local and favorable to LCDSC.

Figure 7 illustrates that when the locality ratio is at or below one, then LCDSC is competitive with the best performing method such as k-Highest. However, once the locality ratio goes beyond one, LCDSC becomes the dominant signal cleaning technique followed by WIT. This gives us a rough guide for when to start considering a signal cleaning problem local or global. When the noise ratio is below one, it can be better to clean with global cleaning methods, whereas local cleaning methods are better when the ratio is greater than one, while global methods may be preferable when the locality ratio is less than 0.5.

3.4 Simulation 4: Additional Simulations—Distinguishing Consecutive Signals

In previous simulations, we have focused our attention on examples where we have one true that is preceded and followed by white noise. However, because our algorithm makes no assumptions about the number of true signals, it is also useful in situations where we are interested in isolating multiple true signals. To demonstrate this, we will consider the situations where we have two Doppler signals separated by white noise of length δ:

$$X(t) = \begin{cases} R(t,\sigma) & \text{if } t < 500 \\ S(\frac{t-500}{1000}) + R(t,\sigma) & \text{if } 500 \leq t < 1000 \\ R(t,\sigma) & \text{if } 1000 \leq t < 1000+\delta \\ S(\frac{t-1000+\delta}{1500+\delta}) + R(t,\sigma) & \text{if } 1000+\delta \leq t < 1500+\delta \\ R(t,\sigma) & \text{if } 1500+\delta \leq t < 2000+\delta \end{cases} . \tag{21}$$

Here, $\delta \in \mathcal{N}$ controls the gap between the two Doppler signals and $R(t,\sigma) \sim N(0,\sigma^2)$ controls the standard deviation of the Gaussian background noise. As we will see as δ is decreased and σ is increased, it will become progressively difficult to distinguish the Doppler signals from each other.

Looking at an example of such a signal when ($\delta = 500, \sigma = 0.25$), we can see in Fig. 8 a plot of the IMFs, the instantaneous amplitudes, and the cleaned IMFs. From this, we can already notice several properties. (1) the increases in instantaneous amplitudes relative to the white noise is most apparent in the middle IMFs (3–5 for this example). This is because the amplitude of the Doppler signal is highest in the middle frequencies. Thus, we should expect the middle IMFs to be most distinguishable from background noise while the smallest and largest IMFs do no show clear spikes in amplitudes. (2) The Doppler signal is expressed at later and later time points as the IMF number increases. This is due to a direct property of the IMF decomposition and the Doppler Signal. Higher number IMFs express lower frequency signals and the Doppler signal increases in frequency over time. Thus, while they will not necessarily be occurring at the same time, there should still be two discernible spikes separated by a gap in the cleaned IMFs.

Thus, for this simulation, we will be evaluating how many IMFs when cleaned yield two clear spikes with at least 50% of the space in between, identified as noise and set to zero. So in the example of Fig. 8, when ($\delta = 500, \sigma = 0.25$), IMFs 1–6 exhibit the desired criteria while IMFs 7–9 do not.

In these simulation results in Fig. 3.4, we have generated 50 signals with σ drawn from Uniform[0.05,1] and δ from Uniform[10, 500]. These signals are then decomposed into their constituent IMFs and then each IMF is cleaned using our algorithm. If the IMF identifies, via human inspection, two clear spikes and more than 50% of the space in between set to 0, we say that we have successfully cleaned both signals. The results of this separability study are shown in Fig. 3.4. Here, we see that IMF1 is only separable when there is lower than 0.25-—0.5 standard deviations of noise. Likewise, IMF 2 also exhibits problems with separability when there is a high noise level, albeit with issues now occurring when above 0.75. IMFs 2–6 seem to be separable regardless of the level of background noise or gap size. But around IMF 7–10, the separability of the IMFs seems to fall again, except unlike IMFs 1–2, the fall in separability seems to occur uniformly until only 1 or 2 out of 50 simulations show separable IMFs. From this, we can say that our algorithm into cleaning the signals into separable chunks

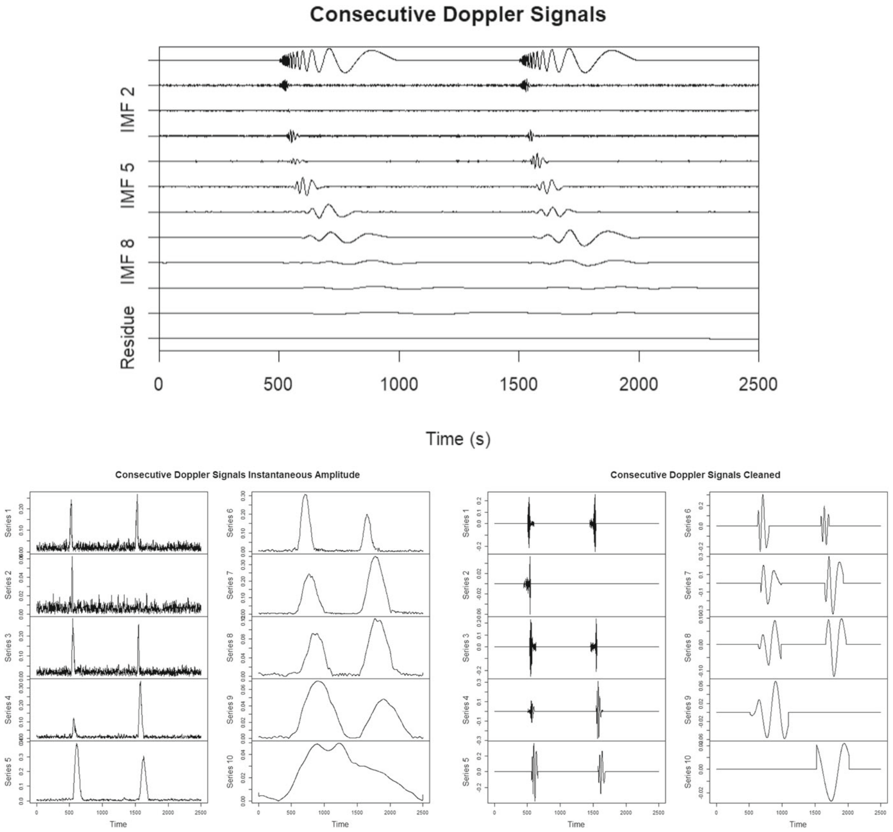

Fig. 8 Top: Raw signal and IMFs for the consecutive Dopplers (Bottom Left): Instantaneous amplitudes for the IMFs. (Bottom Right): Cleaned IMFs

depend on the level of background nose (especially for high-frequency IMFs), the IMF number, but is fairly robust to changes in the gap size.

4 Application

4.1 Application: Detection of Gliding Events in Acoustic Explosions

On October 28, 2014, an Antares rocket operated by Orbital Sciences Corporation exploded shortly after takeoff [24]. The resulting explosion was powerful enough that acoustic shockwave arrivals were observed at stations over 2000 km away from the launch site. At the time, 226 acoustic and atmospheric stations from the Transportable USArray network were located within range of the explosion, resulting in arrivals from the explosion being picked up by the array's infrasound sensors. Many of these arrivals exhibited characteristics of dispersive waves at the infrasound level (<20 Hz). This is of interest as dispersive waves were only recognized recently in the infrasound

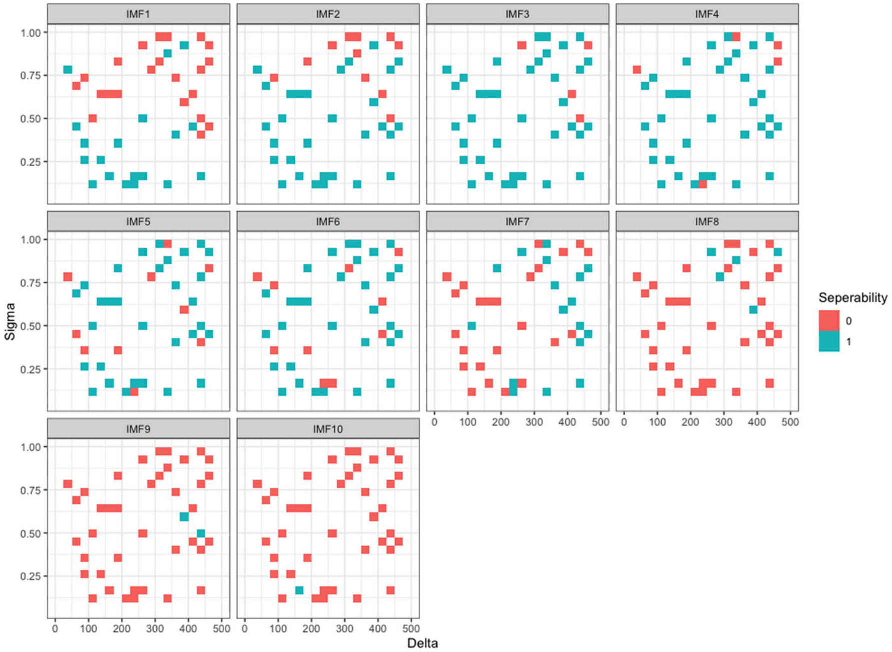

Fig. 9 An investigation of how LCDSC can separate multiple consecutive Doppler Signals. Delta controls the gap between the Dopplers and Sigma controls the level of background noise. If half or more of the time between Dopplers was correctly cleaned as noise, then it was deemed success separated. Separability was best seen in low IMF numbers, particularly less than or equal to 6 and low sigmas. Delta had comparatively less effect on separability

domain [23] and because the Antares explosion was one of the largest demonstrations to date of the existence of infrasound dispersive waves [27]. These dispersive waves are a result of the arrivals being reflected at different heights in the troposphere as well as being influenced by atmospheric conditions such as temperature and wind speed. This makes studying infrasound arrivals important tools in evaluating atmospheric density models [27].

Isolating these dispersive waves can be complicated due to the relatively short time periods when the explosion was detected as well as the complex weather and atmospheric factors affecting recording conditions at each sensor.

However, this problem is well suited for LCDSC. First, each infrasound is relatively quick (on the order of a few seconds within the 24-h monitoring of the USArray sensors). Second, as seen in Fig. 10, one of the canonical features of an infrasound dispersive wave is the presence of a “gliding" or steadily increasing frequency in the signal. This makes infrasound dispersive waves display gliding similar to a Doppler signal reversed, which the LCDSC has performed well at cleaning.

Performing LCDSC on the signal from one of the acoustic stations, we do indeed observe in Fig. 11 that LCDSC cleans the signal well especially compared to WIT which has made very little change to the signal due to the large period of noise throwing off the estimation WIT’s baseline noise estimation.

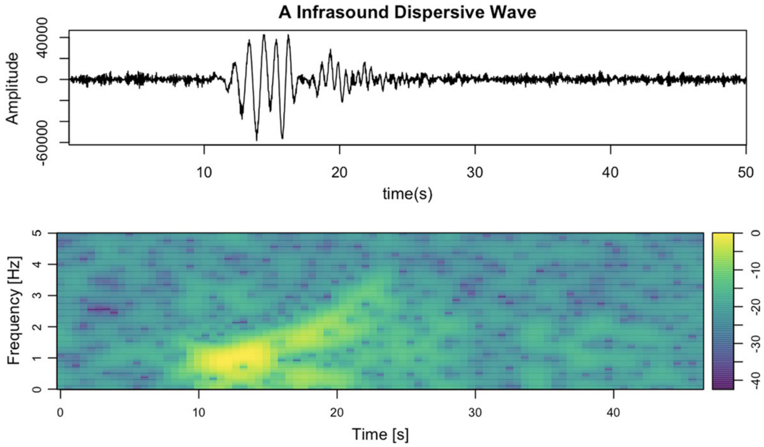

Fig. 10 A canonical example of an infrasound dispersive wave. Note the increase in frequency over the duration of the signal in the spectrogram plot

Moreover, in Fig. 12, by increasing the threshold value, γ, we can clean the signal further and further until only the acoustic explosions are singled out. This occurs when gamma is around 2. This informs us that dispersive waves seem to lead to at least a 2 times increase in power in all of the IMFs.

5 Limitations

While we seek to demonstrate the utility of our `LCDSC` algorithm in this study, there are two large limitations. First comes from the run time of the underlying EEMD procedure. While the EMD is relatively quick with some results indicating that it has similar run time complexity to the fast Fourier transform [31], the EEMD requires multiple iterations of EMD and is thus considerably slower. We noted that it was common for signals of length greater than 20,000 to take over an hour on our computing hardware even when asking for less than 100 EEMD replicates. Thus, we recommend that in situations when one would like to perform a local change point detection on a signal of this length of greater, it may be prudent to split the signal into pieces and have these analyzed separately, perhaps even in parallel. This is the approach taken in CUDA [4] and Open-MP implementations [3] of EMD/EEMD.

The second limitation concerns the optimality of this procedure. While we have attempted to base this algorithm using known guarantees from the statistical change point detection literature, there are still gaps between the assumptions used to design the change point detection process and what is known about theoretical properties of the EMD/EEMD. As an example, the mBIC, which we employed in our change point detection procedure, is derived using an asymptotic argument to approximate a Bayes factor-based model selection procedure for discrete Gaussian processes. However, as described above, the EEMD does not scale well with signal length, so it is conceivable that the abundance of seeming false positives in Fig. 3 could be due to this application of an asymptotic penalization term in a non-asymptotic situation. Based on this and

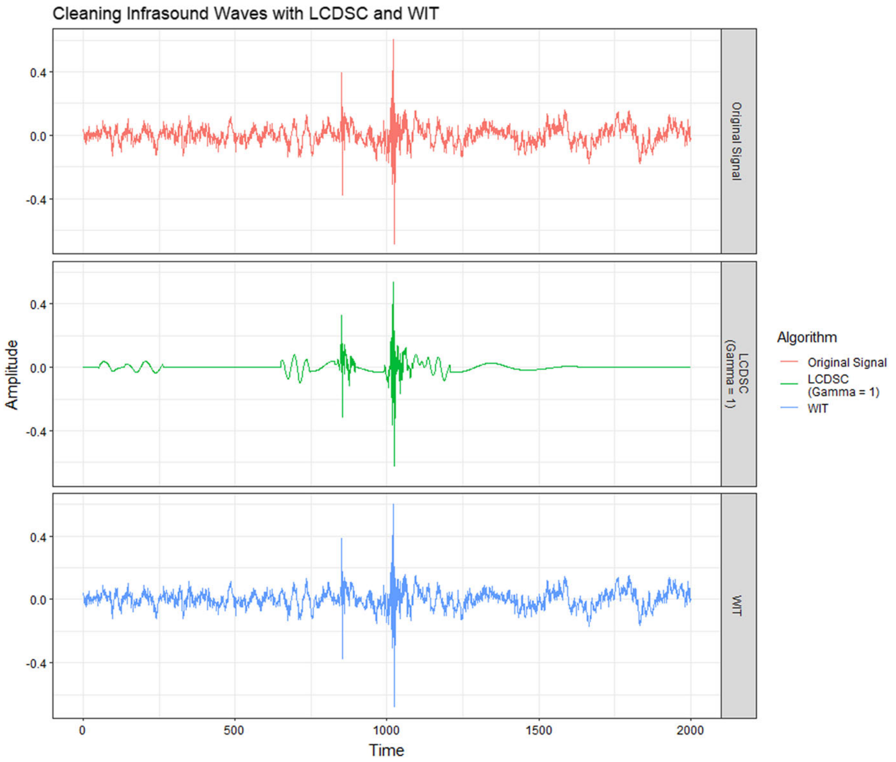

Fig. 11 Comparison of uncleaned gliding signal with LCDSC cleaned signal and WIT cleaned signal. Note that LCDSC performs a better job at cleaning the signal than WIT, especially in helping to isolate the two spikes between 800 and 1000 that represent the infrasound dispersive wave

the relative dearth of known theoretical properties of the EMD/EEMD compared to the competing Fourier transform and wavelet transform, this algorithm should be viewed not as a guaranteed optimal signal cleaning procedure or change point detection procedure, but rather as a first step in working toward the development of an optimal cleaning procedure for local EMD signals.

6 Conclusion and Discussion

Here, we provided a demonstration of the utility of LCDSC for the problem of local change point detection and signal cleaning. While other EEMD signal cleaning algorithms can exhibit drawbacks when there are long periods of no signal, our LCDSC does not suffer from the same disadvantage. This makes it ideal for the cleaning of short-term signals such as acoustic shock waves. We believe that the future development of EEMD signal decomposition will benefit greatly from the further development of methods based on local changes in basis functions.

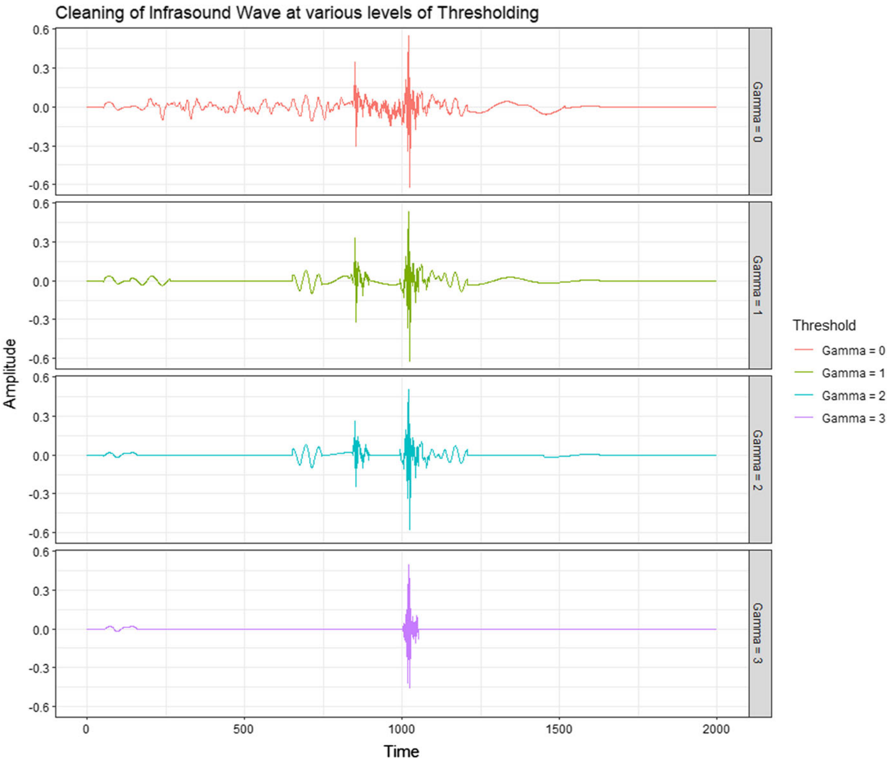

Fig. 12 Comparison of LCDSC signal cleaning as γ is increased. As γ is increased, this results in a sparser and sparser signal cleaning, with $\gamma = 2$ most cleanly isolating the dispersive wave. This indicates that the dispersive wave can be identified in each IMF as a twofold increase in SNR compared to background noise

Acknowledgements This work was supported by the National Science Foundation under Grant Nos. DMS-1613112, IIS-1633212, DMS-1916237, DMS-1929298, and DMS-2152289. All authors have no competing interests to declare that are relevant to the content of this article.

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

1. H. Akaike, A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974). https://doi.org/10.1109/TAC.1974.1100705
2. D.C. Bowman, J.M. Lees, The Hilbert–Huang transform: a high resolution spectral method for nonlinear and nonstationary time series. Seismol. Res. Lett. **84**(6), 1074–1080 (2013). https://doi.org/10.1785/0220130025
3. L.-W. Chang, M.-T. Lo, N. Anssari, K.-H. Hsu, N.E. Huang, W.-m.W. Hwu, Parallel implementation of multi-dimensional ensemble empirical mode decomposition, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011), pp. 1621–1624. https://doi.org/10.1109/ICASSP.2011.5946808
4. D. Chen, L. Wang, G. Ouyang, X. Li, Massively parallel neural signal processing on a many-core platform. Comput. Sci. Eng. **13**(6), 42–51 (2011). https://doi.org/10.1109/MCSE.2011.20

5. X. Chen, X. Zhang, J. Zhou, K. Zhou, Rolling bearings fault diagnosis based on tree heuristic feature selection and the dependent feature vector combined with rough sets. Appl. Sci. **9**(6), 1161 (2019). https://doi.org/10.3390/app9061161
6. X. Chen, B. Cui, Efficient modeling of fiber optic gyroscope drift using improved EEMD and extreme learning machine. Signal Process. (2016). https://doi.org/10.1016/j.sigpro.2016.03.016
7. D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3), 425–455 (1994). https://doi.org/10.1093/biomet/81.3.425
8. S. Gaci, A new ensemble empirical mode decomposition (EEMD) denoising method for seismic signals. Energy Procedia **97**, 84–91 (2016). https://doi.org/10.1016/j.egypro.2016.10.026
9. S. Holm, A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**(2), 65–70 (1979)
10. M. Hotradat, K. Balasundaram, S. Masse, K. Nair, K. Nanthakumar, K. Umapathy, Empirical mode decomposition based ECG features in classifying and tracking ventricular arrhythmias. Comput. Biol. Med. **112**, 103379 (2019). https://doi.org/10.1016/j.compbiomed.2019.103379
11. N.E. Huang, Z. Shen, S.R. Long, A new view of nonlinear water waves: the Hilbert spectrum. Annu. Rev. Fluid Mech. **31**(1), 417–457 (1999). https://doi.org/10.1146/annurev.fluid.31.1.417
12. N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.-C. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **454**(1971), 903–995 (1998). https://doi.org/10.1098/rspa.1998.0193
13. N.E. Huang, M.-L.C. Wu, S.R. Long, S.S.P. Shen, W. Qu, P. Gloersen, K.L. Fan, A confidence limit for the empirical mode decomposition and Hilbert spectral analysis. Proc. R. Soc. Ser. A **459**, 2317–2345 (2003)
14. N. E. Huang, S. S. P. Shen, *Hilbert-Huang Transform and Its Applications* (World Scientific, 2005). https://doi.org/10.1142/5862
15. Z. Huimin, S. Meng, D. Wu, Y. Xinhua, A new feature extraction method based on EEMD and multiscale fuzzy entropy for motor bearing. Entropy (2017). https://doi.org/10.3390/e19010014
16. C. Inclán, G.C. Tiao, Use of cumulative sums of squares for retrospective detection of changes of variance. J. Am. Stat. Assoc. **89**, 913–923 (1994)
17. Y. Kopsinis, M. Stephen, Development of EMD-based denoising methods inspired by wavelet thresholding. IEEE Trans. Signal Process. **57**, 1351–1362 (2009). https://doi.org/10.1109/TSP.2009.2013885
18. Y. Lei, M.J. Zuo, Fault diagnosis of rotating machinery using an improved HHT based on EEMD and sensitive IMFs. Meas. Sci. Technol. **20**(12), 125701 (2009). https://doi.org/10.1088/0957-0233/20/12/125701
19. T. Li, M. Zhou, C. Guo, M. Luo, J. Wu, F. Pan, Q. Tao, T. He, Forecasting crude oil price using EEMD and RVM with adaptive PSO-based kernels. Energies (2016). https://doi.org/10.3390/en9121014
20. D. Liu, X. Yang, G. Wang, J. Ma, Y. Liu, C.K. Peng, J. Zhang, J. Fang, HHT based cardiopulmonary coupling analysis for sleep apnea detection. Sleep Med. (2012). https://doi.org/10.1016/j.sleep.2011.10.035
21. G. Liu, Y. Luan, An adaptive integrated algorithm for noninvasive fetal ECG separation and noise reduction based on ICA-EEMD-WS. Med. Biol. Eng. Comput. **53**(11), 1113–1127 (2015). https://doi.org/10.1007/s11517-015-1389-1
22. M. Lozano, J.A. Fiz, R. Jané, Performance evaluation of the Hilbert–Huang transform for respiratory sound analysis and its application to continuous adventitious sound characterization. Signal Process. **120**, 99–116 (2016). https://doi.org/10.1016/j.sigpro.2015.09.005
23. P.T. Negraru, E.T. Herrin, On infrasound waveguides and dispersion. Seismol. Res. Lett. **80**(4), 565–571 (2009). https://doi.org/10.1785/gssrl.80.4.565
24. K. Northon, NASA Statement Regarding Oct. 28 Orbital Sciences Corp. Launch Mishap (2015). https://www.nasa.gov/press/2014/october/nasa-statement-regarding-oct-28-orbital-sciences-corp-launch-mishap
25. G. Schwarz, Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978). https://doi.org/10.1214/aos/1176344136
26. C. Truong, L. Oudre, N. Vayatis, Selective review of offline change point detection methods. Signal Process. **167**, 107299 (2020). https://doi.org/10.1016/j.sigpro.2019.107299
27. J. Vergoz, The Antares explosion observed by the USArray: an unprecedented collection of infrasound phases recorded from the same event. Infrasound Monit. Atmosp. Stud. (2018). https://doi.org/10.1007/978-3-319-75140-5_9

28. T. Wang, M. Zhang, Q. Yu, H. Zhang, Comparing the applications of EMD and EEMD on time-frequency analysis of seismic signal. J. Appl. Geophys. **83**, 29–34 (2012). https://doi.org/10.1016/j.jappgeo.2012.05.002
29. W. Wang, D. Xu, X. Chen, Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. Water Resour. Manag. **29**, 2655–2675 (2015). https://doi.org/10.1007/s11269-015-0962-6
30. X. Wang, C. Liu, F. Bi, X. Bi, K. Shao, Fault diagnosis of diesel engine based on adaptive wavelet packets and EEMD-fractal dimension. Mech. Syst. Signal Process. **41**(1), 581–597 (2013). https://doi.org/10.1016/j.ymssp.2013.07.009
31. Y.-H. Wang, C.-H. Yeh, H.-W.V. Young, K. Hu, M.-T. Lo, On the computational complexity of the empirical mode decomposition algorithm. Physica A **400**, 159–167 (2014). https://doi.org/10.1016/j.physa.2014.01.020
32. Y.-X. Wu, Q.-B. Wu, J.-Q. Zhu, Improved EEMD-based crude oil price forecasting using LSTM networks. Physica A **516**, 114–124 (2019). https://doi.org/10.1016/j.physa.2018.09.120
33. Z. Wu, N.E. Huang, Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv. Adapt. Data Anal. **1**, 1–41 (2009)
34. N.R. Zhang, D.O. Siegmund, A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics **63**(1), 22–32 (2007). https://doi.org/10.1111/j.1541-0420.2006.00662.x
35. J. Zheng, H. Pan, S. Yang, J. Cheng, Adaptive parameterless empirical wavelet transform based time-frequency analysis method and its application to rotor rubbing fault diagnosis. Signal Process. **130**, 305–314 (2017). https://doi.org/10.1016/j.sigpro.2016.07.023

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.