5-2021

Automated Identification of Lines in Data from Gravitational Wave Detectors

Thomas A. Cruz
The University of Texas Rio Grande Valley

AUTOMATED IDENTIFICATION OF LINES IN DATA FROM GRAVITATIONAL WAVE DETECTORS

A Thesis

by

THOMAS A. CRUZ

Submitted to the Graduate College of
The University of Texas Rio Grande Valley
In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2021

Major Subject: Physics

AUTOMATED IDENTIFICATION OF LINES IN DATA FROM GRAVITATIONAL WAVE DETECTORS

A Thesis
by
THOMAS A. CRUZ

COMMITTEE MEMBERS

Dr. Soumya Mohanty
Chair of Committee

Dr. Soma Mukherjee
Committee Member

Dr. Malik Rakhmanov
Committee Member

May 2021

ABSTRACT

Cruz, Thomas A., Automated Identification of Lines in Data from Gravitational Wave Detectors. Master of Science (MS), May, 2021, 38 pp., 16 figures, 32 references.

On the frontier of gravitational wave (GW) astronomy, the LIGO detectors record vast quantities of data that need to be analyzed constantly for rare and transient GW signals. A foundational problem in LIGO data analysis is the identification of spectral line features in the Power Spectral Density (PSD) of the data. Such line features correspond to high power terrestrial or instrumental signals that must be removed from the data before any search for GW signals can take place. In this study the method developed aims to automate the extraction of the frequencies and bandwidths of the lines, treated as sharp features in the PSD. For this purpose, we use a non-parametric curve fitting method recently developed at UTRGV called SHAPES. Our method automates the determination of line features through window smoothing, estimation through SHAPES and extremum analysis with a program called ALINE. In order to apply SHAPES, which uses B-spline curve-fitting with an adaptive knot sequence, we had to first overcome its limitation on the length of input data. This was achieved by applying SHAPES to overlapping segments of the PSD data along with an exponentially weighted averaging of the SHAPES estimates in the overlap regions. With the estimation groundwork laid, the second step is determining the locations of the lines through adaptive extrema differences in the SHAPES estimate of the PSD with ALINE. Our study shows that it is possible to expand SHAPES, an adaptive estimation tool, to (i) any sized data, thus impacting applications in all statistical endeavors, and (ii) use it to automate the identification of lines in the PSD of LIGO data with the ALINE algorithm.

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Souyma Mohanty for his guidance throughout the entire research process. I had no prior experience with Matlab, the algorithm's coding language, and he's taught and continues to teach me intricacies of the language. I would also like to acknowledge and thank the UTRGV Physics and Astronomy department for offering me an assistantship which has provided funding for this research and my schooling. Lastly, I would like to acknowledge Kimberly Reyes, my significant other, for giving invaluable feedback on the theory, code, and the writing of this thesis.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The point of this research is to devise an algorithm that can sift through large quantities of gravitational wave (GW) data and identify the locations of noisy spectral line features for removal or further evaluation. It is thus important to understand at least two concepts; gravitational waves and splines. GW data resembles other kinds of wave data but is hidden in a seemingly innocuous noise floor. To find GW chirps, it is vital to remove spectral line features that are due to terrestrial noise. The line identification algorithm, or ALINE, is based on a SHAPES estimate of the data which uses splines to fit the data. Sec 1.1 discusses gravitational waves and splines, 1.2 delves into why this research is needed, and 1.3 walks through the algorithms and their design and intent.

1.1 Background

1.1.1 Gravitational Waves

Gravitational waves (GW) are on the frontier of modern day astronomy. Predicted by Albert Einstein's Theory of General Relativity in 1916, the idea that massive accelerating objects would distort space-time like a ripple throughout the cosmos was outlandish at the time. Einstein himself was not confident in the physicality of the mathematics and in his original 1936 *Physical Review* submission of *Do Gravitational Waves Exist?* he concluded the waves were fictitious in nature. Howard Robertson, the Physical Review's anonymous referee, discovered an error with Einstein's conclusion which left Einstein upset and dismissed the journal. Robertson coincidentally befriended Einstein's assistant, Leopold Infeld, and talked with him about Einstein's erroneous conclusion. Infeld consulted Einstein which led him to reformulate his paper under a less controversial title, *On Gravitational Waves* [14]. Proof of GW existence did not come to fruition until 1974 when Arecibo

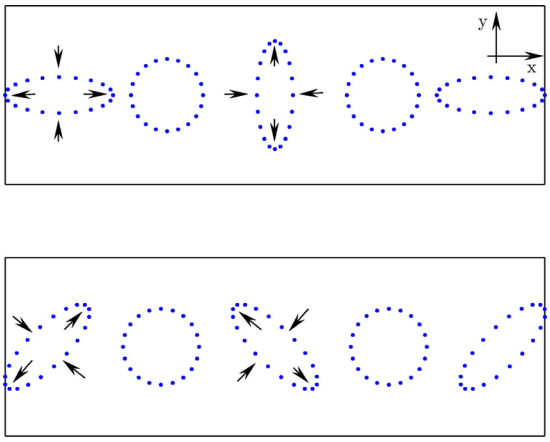

Figure 1.1: Shows how space is warped by gravitational waves.
[15]

Radio Observatory discovered a binary pulsar, a massive accelerating system that could be used to test Einstein's theory [30]. After years of measurement, it was clear that the stars were cycling inwards at the precise rate predicted by the theory. However, despite indirect proof that gravitational waves exist, they had not been directly measured. That is until September 14, 2015 when the Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) physically measured the disturbance in space-time caused by two colliding black holes approximately 1.3 billion light-years away [1]. LIGO is comprised of two US-based detectors in Hanford, Washington and Livingston, Louisiana. Since the first detection, more gravitational wave detectors have come online. This thesis uses data from GW170817 [2], a binary neutron star inspiral, that was detected by LIGO and Advanced Virgo [5]. Second-generation detectors KAGRA [27], a Japan-based detector, , and LIGO-India [28], an India based observatory, are both currently in development with KAGRA attempting to improve its sensitivity and stability to join the other detectors [22] and LIGO-India still in its planning stages.

Since gravitational waves warp space itself by stretching and compressing in perpendicular directions as seen in Figure 1.1, LIGO is designed to measure this change in distance by monitoring two perpendicular laser beams that span 4 km. As an interferometer, LIGO works by measuring the interference pattern of two laser beams and monitoring their phase. Normally, the beams are in phase with each other, however, when a gravitational wave of sufficient strength passes by, the

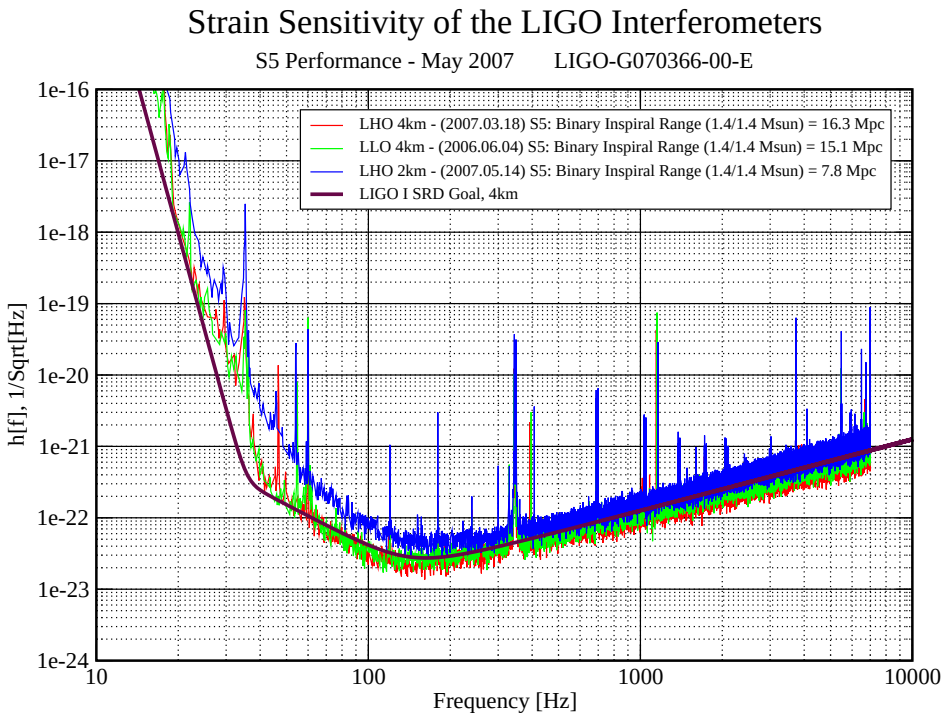

Figure 1.2: Livingston and Hanford LIGO detector sensitivity compared to the design sensitivity. [32]

length of the beams, and thus their phase, will change. The discrepancy in phase will lead to the gravitational wave chirp. These chirps are very faint and not noticeable until the spectral line features are removed.

The data that is analyzed in this thesis is of the form of a Power Spectral Density, or PSD. This was obtained from the whitened Fourier Transformed time series of the space-time strain which the LIGO detectors measure. Since this thesis revolves around isolating approximately stationary noisy features, we can best characterize the noise in the frequency domain, therefore we take the fast Fourier Transform, or `fft` in Matlab, of the strain time-series [3]. Then we should estimate the PSD of the data because the PSD identifies which frequency range variances are most prominent. To estimate the data, we use Welch's overlapped segment averaging estimator, or `pwelch` in Matlab [31]. The PSD will highlight the noisy features prominently such as in Figure 1.2. At its core the LIGO interferometers are measuring waves which experience many different forms of interference,

either from gravitational waves disturbing the space or some other noise source.

The sources of noise include seismic, thermal, photon shot, gas, static charging, laser, and many others [21]. Each of these can be mitigated but not entirely resolved. Seismic noise is due to all forms of terrestrial disturbances affecting the motion of the mirrors, i.e. ground vibrations, wind, storms, or human movements. To counteract the effect of these disturbances LIGO uses active and passive isolation which aims to mitigate contact with the disturbances. Active isolation uses constant monitoring of seismic data to determine an appropriate counter force to apply to the test masses. Passive isolation uses a quadruple pendulum test mass suspension system inside a vacuum chamber to constantly diminish seismic effects [4]. Thermal noise refers to the atomic fluctuations of the mirrors and the suspension material. One way the minute movements are mitigated is the sheer size and mass of the test masses and also the multi-level pendulum which helps with the thermal noise in the suspension system.

1.1.2 Splines

To understand the SHAPES and PSO algorithms that draw the estimate it is vital to cover splines and how they are used. A more complex version of splines, B-splines, are used in SHAPES and are similarly discussed in this section.

Definition Splines are piece-wise polynomial functions [11] that are used in non-parametric regression problems because of the flexibility they allow for the models while still satisfying a smoothness condition [29][18]. Splines use polynomials that are connected via breakpoints or knots. Since the polynomials are joined at the knots with conditions enforcing the data to match at the knot points and also that all derivatives up to order $k-2$ are continuous across interior knots, where k is the order of the polynomial pieces, splines can be incredibly smooth curves. Splines are most commonly referenced by the degree of the polynomial pieces instead of the order, for example for $k = 4$ the polynomial pieces are cubic and, as such, the spline would be called a cubic spline.

B-Splines The set of all cubic splines for a given knot sequence comprise a vector space. As such, any cubic spline belonging to this space can be expressed as a linear combination of basis

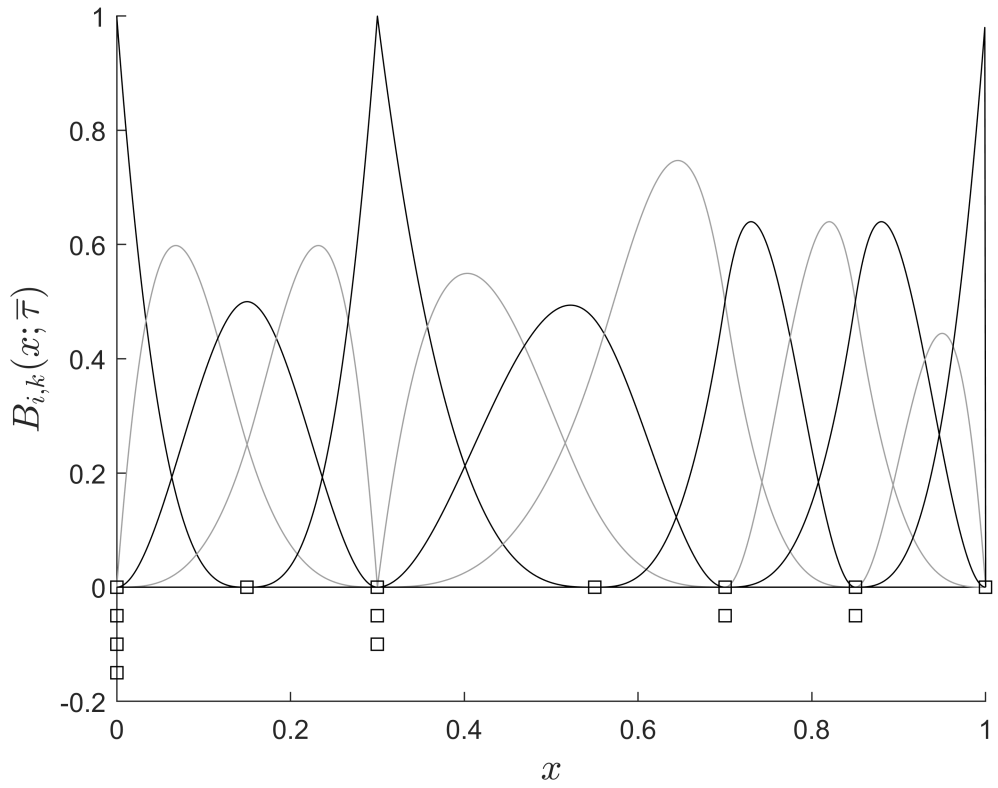

Figure 1.3: Cubic B-Spline functions with 14 arbitrarily placed knots marked by squares. Variations of the B-spline function are pictured in black and grey representing other configurations. Stacked squares indicate repetitions.

functions. A popular choice of basis functions are B-spline functions [11]. For a set of M knots and polynomial order k, the vector space has a dimensionality of $M+k-2$. In the case of cubic splines, the splines used by SHAPES, $k=4$, thus the vector space has dimensionality of $M+2$. SHAPES uses knots than can appear more than once mainly because these repetitions allow for SHAPES to fit discontinuities, see Mohanty and Fahnestock [24] for a more thorough explanation of the explicit usage of B-splines. Figure 1.3 gives a visual description of cubic B-splines with some of them showing discontinuities due to knot repetition.

1.2 Motivations

As mentioned in the gravitational wave section, the data from the detectors is rife with terrestrial noise. The gravitational wave chirps are invisible because of the noisy line features.

Most features are indicative of common frequencies, 60 Hz power supplies, or instrument resonant frequencies, and it has been the subject of other papers [20][7][26][13] to identify the origins of these features, however, there are many features that do not correspond to any known phenomenon [3]. It is the goal of this thesis to devise an algorithm that identifies the spectral line features regardless of their origin with the intent of applying to gravitational wave detector data.

The problem of identifying the features is twofold, the first is a strong estimate and the second is a method that is expansive enough to grab minute features but is focused enough to ignore noise. An estimate that disregards the impact of the noise will lead to an easier time for identification algorithms to sift through the data. The first problem is addressed with the use of the SHAPES algorithm which is a curve fitting method that uses splines to create a solid estimate. The second is an ongoing, constantly evolving problem that is heavily dependant on the incoming data. The ALINE algorithm that handles the fine tuning and identification of the features is given a wide breadth of knobs that can be adjusted to best fit the data. Both the SHAPES and ALINE algorithms are discussed in greater detail in the Sec. 1.3.

1.3 SHAPES and ALINE

1.3.1 SHAPES

SHAPES, or Swarm Heuristics based Adaptive and Penalized Estimation of Splines, is a non-parametric curve fitting method that can fit both continuous and discontinuous curves [24]. It fits a cubic spline with adaptive knot placement and allows knot repetition in order to fit discontinuities. The defining feature of SHAPES is that it allows a fixed number of knots to freely move around to find the best knot number that best fits the segment of data using the lowest AIC, or Akaike Information Criterion [6]. The AIC is a method for comparing models with different number of parameters, or knots in this case, the model with the lowest AIC being selected as the best one . To find the optimal knot placement SHAPES uses PSO, or Particle Swarm Optimization. PSO is an algorithm that is based on the behavior of biological swarms, i.e. a flock of birds, that is tasked with finding the best source of food in an area. The prime idea is that flocks communicate in small local

groups about food or predators, for example if one bird in the swarm finds a small pile of food it will communicate with adjacent birds the size and location of the food cache. If all birds in a swarm do this over a large, global, area then eventually they will be able to find the most amount of food [19]. This behavior is the principal behind PSO, in SHAPES this is used by giving the swarm a knot sequence, on which to grab an estimate of the spline from. Recall from the Splines section that knots are points where spline polynomials are joined. It is the job of the swarm to determine the ideal location for the knots and the best number of knots needed to fit data with the lowest AIC [23]. SHAPES instructs PSO to run on each model and then SHAPES chooses the best model to fit a estimate spline. However, this is where the major problem arises, SHAPES can be used to fit curves that are either simple but long, or complex but short data segments; it is limited by the number of knots. SHAPES indicates the location of knots and where there is multiplicity, or where knots are close enough to each other to be considered a single knot. A major tuning knob available in SHAPES is the regulator gain, or λ, which is where the penalization of SHAPES comes in. SHAPES uses cubic splines because they give the solution to the variational problem of a smoothing spline estimate, or $\hat{f}$ which is defined by equation 1.1.

$$\hat{f} = \arg\min_{f} \left[\sum_{i=0}^{N-1} (y_i - f(x_i))^2 + \lambda \int_0^1 dx (f''(x))^2 \right] \tag{1.1}$$

The first term measures the fidelity of the model to the observations, or how well the model fits, and the second term measures by the average curvature, penalizing the roughness of the model, or how displaced the model fits. The larger the regulator gain, the larger the penalization [24].

The greatest takeaway is that simple data, like a Gaussian curve, can be reduced to a single polynomial with few knots, and that complex curves will require more knots. This idea was the defining motivation for this thesis initially, the hypothesis was "does a large congregation of knots infer the presence of a noisy line feature?" The location of the knots did seem to indicate the presence of a line feature, however, the method was not without flaws. The primary issue was the inconsistency of the locations of knots; seemingly mundane segments with a small line feature had very few knots because the line feature was considered to be part of the noise floor or with

overly complex segments some features did not get enough knots or multiplicity. A new method for identifying line features had to be created, as well as a modification to SHAPES to expand its limits.

1.3.2 ALINE

ALINE, or Automated Line Identification of Non-stationary noise-based Estimates applies differential peak filtering of the SHAPES estimate to determine the location of spectral line features. The primary goal of ALINE is finding lines for removal so that gravitational wave chirps can be identified. ALINE works generally as a User Interface, controlling plot figures, tuning, and the data that is sent to SHAPES. PSO works best with data that is normalized and SHAPES can identify non-noise features for estimation better with windowing. ALINE normalizes the data using the lower 95^{th} percentile of the data with the intention of setting the noise floor close to zero. It then applies a moving average using the Blackman-Harris [16] windowing to the data which smooths the data allowing SHAPES to create an estimate that does not over fit the noise. After giving the normalized and smoothed data to SHAPES, the estimate is formed. The second derivative is taken and the peak extrema of the data are located via a peak-finding algorithm. Lastly, the peak locations and values are filtered through a user-defined threshold scheme which weeds out the noise floor peaks and leaves only the peaks corresponding to line features. A more thorough discussion of ALINE's theory and results will be given in Chapters III and IV, respectively.

The rest of the thesis is organized as follows. The first main chapter, Chapter II, explains the process of expanding SHAPES to allow a larger sample input and it tests the process on the Donoho-Johnstone Benchmark Test functions. The second main chapter, Chapter III, discusses the design intent of ALINE with an in-depth look at how it works and how ALINE can be used. Chapter IV presents and analyzes the results of SHAPES on the test functions and ALINE on GW detector data. The final chapter, Chapter V, concludes with a verdict on how effective SHAPES and ALINE are on their respective tasks3 and sets future goals.

CHAPTER II

EXPANDING SHAPES

SHAPES is a powerful tool for estimation of sharp discontinuous features often present in noisy data. However, it faces a major hurdle in its limitation on the length of the input data [24]. The longer the data segment the worse the estimate becomes without changing the knot sequence. Extending the knot sequence does help for slightly larger data sets, but quickly the computational burden becomes too much. The way we resolved this issue is through segmentation, by taking a large set and breaking it up into small segments on which to run SHAPES. This is the primary topic in Sec. 2.1, Sec. 2.2 discusses how the improved SHAPES runs on the Donoho-Johnstone Benchmark Test Functions.

2.1 Expansion Methods and Limits

SHAPES works well when the data sets are small, since the knot sequence is tailored to that data size. If the size is increased it then makes sense to alter the knot sequence to accommodate this change. The initial knot sequence

$$N_{knots} = \{5, 6, 7, 8, 9, 10, 12, 14, 16, 18\} \tag{2.1}$$

allows for simple features like in Figure 2.1. The data length is not large, 256 samples, nor are there complex shapes or discontinuities. We can extend this same knot sequence to larger sets if we segment the data segment.

Consider we have an $N = 1,000$ long sample as shown in Figure 2.2 and we want to split it up into segments of $N_s = 250$ samples with an overlap of $N_o = 83$ samples. We would then have six non-overlapping regions of $N_n = 167$ samples. The overlap regions are vital to the success

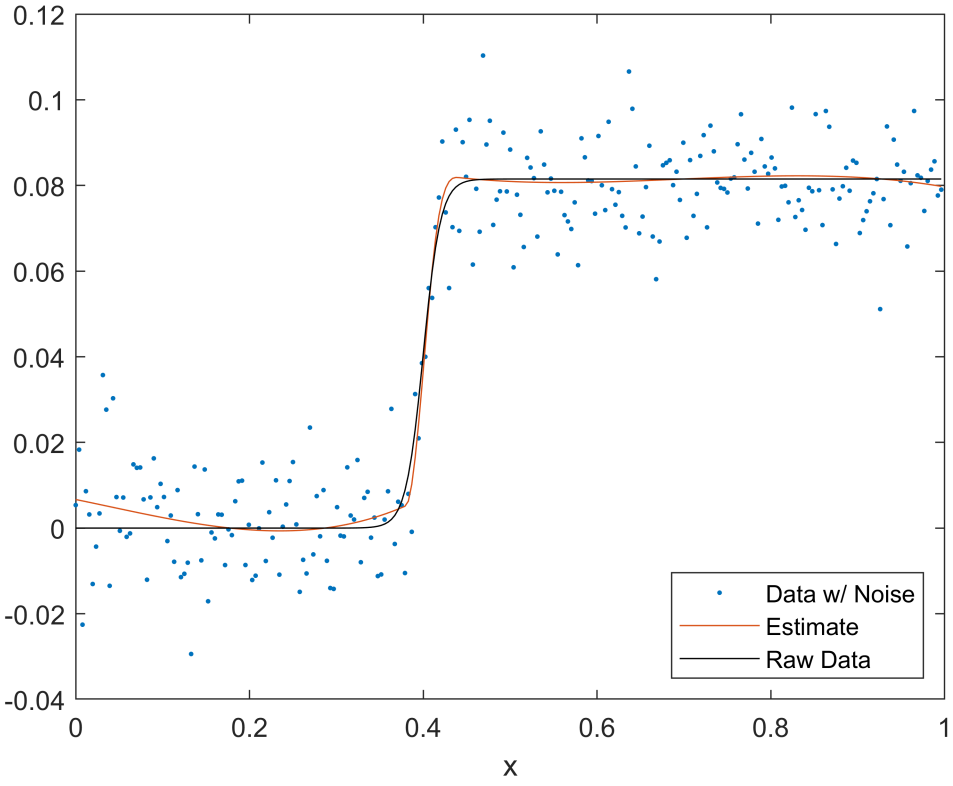

Figure 2.1: Test Data for SHAPES. An illustration of SHAPES on a benchmark function. The blue dots is the data coupled with noise, the red line is the SHAPES estimate, and the black line is the raw data.

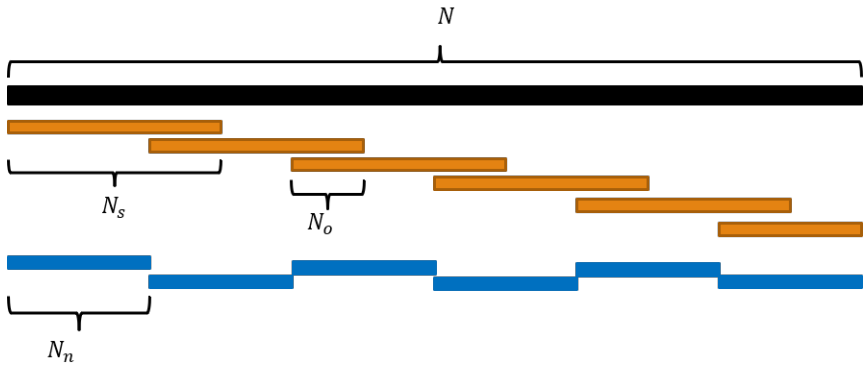

Figure 2.2: Segmentation in SHAPES Example. The black bar, N, is the entire sample set, the whole orange bars, N_s, are the overlapping sample segments, the small orange piece, N_o, is the overlap of segments, and the blue bars, N_n, are the non-overlapping sample segments.

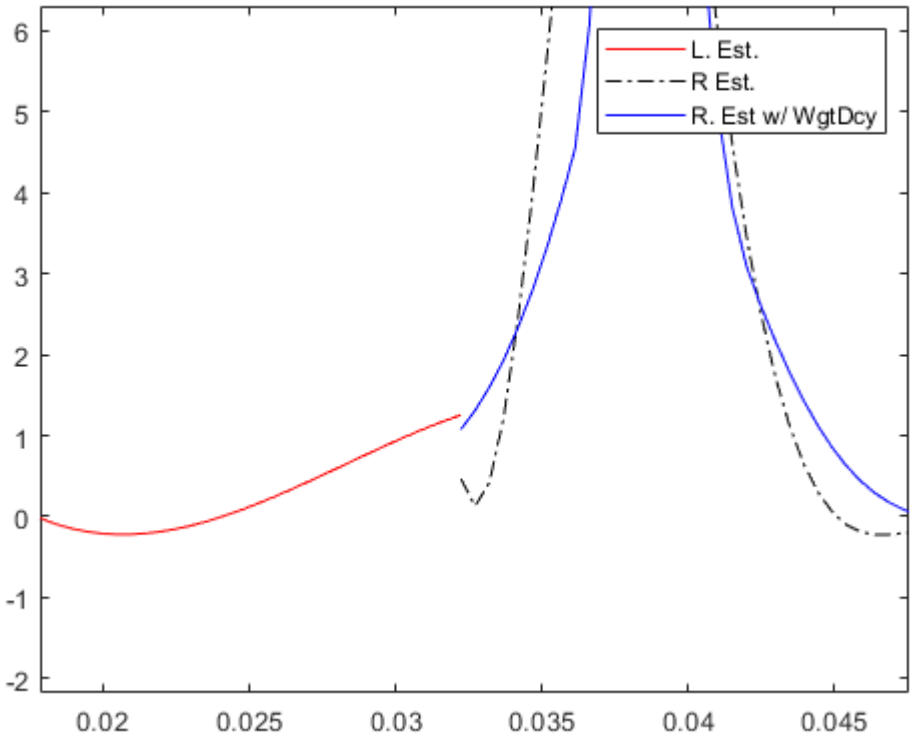

Figure 2.3: Example of Weight Decay Averaging. The solid red and dashed black lines are the estimates before averaging. The solid blue line is the result of applying exponential weight decay which smooths the transition from one segment to another.

of the estimate around the edges of each segment. SHAPES is run on the segments and then an exponential weight decay is applied to the overlap. Starting with the second segment, the last 83 samples of the previous segment and the first 83 samples of the current are averaged using an exponential weighting where the data from the previous decays to zero as it approaches the end of the data, whereas the current segment's data starts from zero and raises exponentially to full values after 83 samples. This system allows the ends of the data between non-overlapping regions to flow seamlessly together as seen in Figure 2.3.

In the `shpsext` algorithm the user can specify the size of the non-overlapping segments, or N_n, i.e. the segments that will be seen, and the size of the overlap, N_o. The size of the overlapping segments, N_s, the number of segments, N_{seg}, and the remainder, r, are calculated with the following equations

$$N_s = N_n + N_o \tag{2.2}$$

$$N_{seg} = \texttt{floor}(\frac{N}{N_n}) \tag{2.3}$$

$$r = N - N_{seg}N_n \tag{2.4}$$

The `floor` function from equation 2.3 is a Matlab function which takes the rounded-down integer part of its argument. The remainder is the amount of samples left after the penultimate segment, it is typically of a smaller size than the other segments unless N is wholly divisible by N_n. If the sample is sufficiently small, it will be added to the penultimate segment.

2.2 Donoho-Johnstone Benchmark Test Functions

The Donoho-Johnstone Benchmark Test Functions [12] are a set of test functions for estimation models to show how well they fit the data, there are four functions in the set; blocks, bumps, Doppler, and heavisine. Each function has unique characteristics that can be used to test different data shapes; the blocks function contains many discontinuities, bumps is comprised of sharp lines features displaced along a flat floor, Doppler has a rapidly oscillating sine wave that decays, and heavisine features discontinuities along an otherwise smooth sine wave.

Applying SHAPES extended, or SHAPES-Ext, to the benchmarks helps tune the parameters of PSO and SHAPES. The bumps function is most notable because the features present are similar to the features present in GW detector data. The data sets were of 2,048 samples in length, substantially larger than the test data in Figure 2.1, and so a perfect environment to optimize the segmentation from Sec. 2.1. The bulk of the tuning involved the size of the segments, the knots model, and the overlap averaging. The size of the segments is dictated by two thresholds, the lower limit of how small a segment should be for PSO and the upper limit of how many knots is enough to accurately find an estimate. The lower limit is around 25 samples, while the upper limit relies heavily on

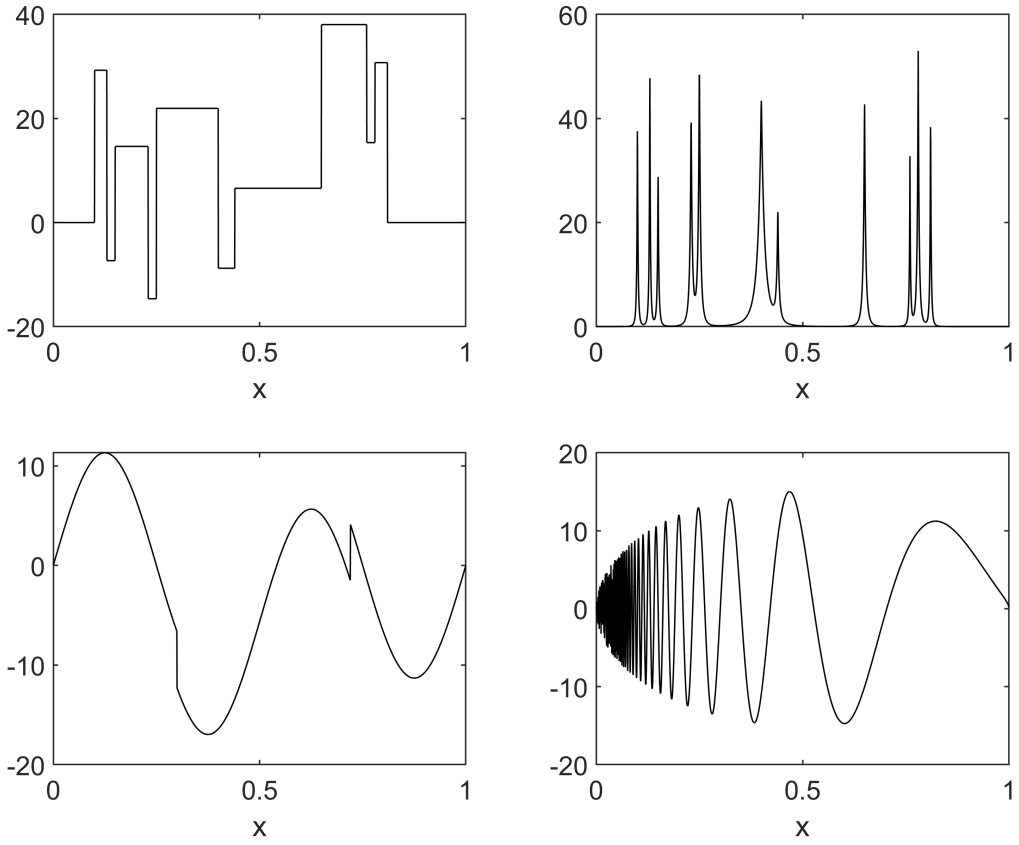

Figure 2.4: The Donoho-Johnstone Benchmark Test Functions. Top Left: Blocks, Top Right: Bumps, Bottom Left: Heavisine, and Bottom Right: Doppler
[12]

the complexity of the data, and thus is much harder to generalize. A singular feature, like the one around 0.65 of the x-axis in Figure 2.4's top-right plot, can be a part of a large-sized segment, approximately 400 samples, with an accurate estimate with no issues. On the other hand, a segment from 0.1 to 0.2, a 200-long sample segment that encompasses multiple features, may struggle with a strong estimate.

The knots model is contingent on the size of the segments and the complexity, the catch-all knot sequence used throughout the research process is as follows

$$N_{knots} = \{5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 28, 30, 33, 36\} \tag{2.5}$$

Chapter III discusses the reason for the expanded knot sequence.

The overlap averaging, from Sec. 2.1, applied exponential weight decay averaging. At each end of a segment, an overlap of specified value is created with the subsequent segment, the two segments must be combined, this is done using the following formula

$$y = \frac{y_R \cdot e^{4(w_R - N_o)/N_o} + y_L \cdot e^{4(-w_L)/N_o}}{\left(e^{4(w_R - N_o)/N_o} + e^{4(-w_L)/N_o}\right)} \tag{2.6}$$

where y is the averaged data point, y_R and y_L are the right and left segments data points, respectively, w_R and w_L are the right and left weight decays, that will be vary from 0 to the overlap value, N_o or from the N_o to 0, respectively.

CHAPTER III

ALINE

ALINE, or Automated Line Identification of Non-stationary noise-based Estimates is an algorithm that determines the location of spectral line features from estimates of data. The program is designed specifically to find spectral line features from GW detector wave data for removal filtering but does have uses outside of this application. The first section, 3.1, focuses on how ALINE is designed, and why it is designed this way. Sec. 3.2 discusses the practical applications of ALINE and areas it can be applied to.

3.1 Spectral Line Features

In order to understand why ALINE is designed the way it is, it is important to delve into what a spectral line feature is and how we can program an algorithm to find them. A spectral line feature is a sharp collection of points indicating a prominent frequency in the data. To be able to detect gravitational wave signals, the detectors of LIGO and similar facilities have to be incredibly sensitive to minute vibrations and, in doing so, pickup noise from all kinds of sources. Some are constant like the power supply or instrument resonant frequencies, others are fleeting like seismic disturbances. Spectral line features are a broad sort, with some consisting of several line features over a short breadth of frequencies, others incredibly sharp, singular features, and others with wide bases. It is the goal of the ALINE algorithm to determine where all of these features are located for extraction.

Despite the numerous differences, all spectral line features share consistencies, notably they all exhibit large values with respect to the noise floor, a rise and eventual fall in value, they also have peaks. Mathematically, this means that the average value of feature is larger than the average

of the noise floor, their derivative's value will be larger than the floor and their derivatives will be zero at their peaks. It is simple to determine where a peak in a spectral line feature is but this also poses an important question, where is the location of a line? Is it the peak, or is it a range? Should the range start at the points of inflection or where the feature leaves the noise floor? ALINE is designed to take the difference of the estimate and then find the peaks of that estimate difference. There will be two sets of peaks, maximum and minimum peaks, and each one corresponds to the point of inflection from a feature going up or coming down. From this we have locations of when a feature is going up and coming down, the space between the extrema is the line feature. However, as seen in Figure 3.1 if we were to filter between all of those pairs of peaks we would have features that only go half up, i.e. the whole line feature was not removed, only the points between inflections. Instead, we want where the second difference's peaks are located, these would indicate where the difference peaks start, or when there is a significant change in the estimate, for example when a line feature emerges from the noise floor.

To avoid over-saturation of peaks, a threshold scheme is devised that will disregard all peaks below a user-specified difference between points. The first condition is defined as follows: If

$$|y_{p,max}(i) - y_{p,min}(i)| \geq T \tag{3.1}$$

and

$$|y_{p,min}(i) - y_{p,max}(i+1)| \geq T \tag{3.2}$$

Then a feature exists in the range $(x_{p,max}(i), x_{p,min}(i+1))$. Where $(x,y)_{p,max}(i)$ corresponds to the i^{th} iteration of x-y maximum peak coordinates, indicated by red stars in Figure 3.1, and $(x,y)_{p,min}$ corresponds to the i^{th} iteration of x-y minimum peak coordinates, indicated by blue stars in Figure 3.1.

The T-value controls ultimately what ranges are visible and is the most important tool for determining the location of line features. The value is currently user-specified but will, in future versions, derive from statistical analysis of the noise floor. A range check condition is created to

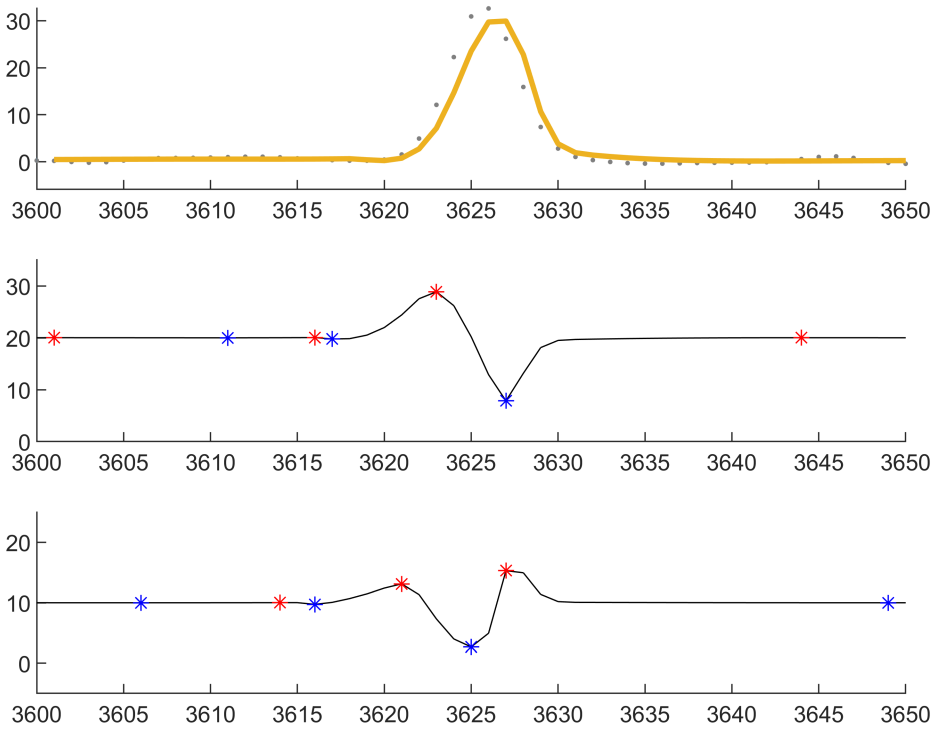

Figure 3.1: SHAPES Estimate with peak 1^{st} and 2^{nd} differences of Hanford detector data from GW170817. Top plot: estimate is the solid line, and dots are raw data. Middle and Bottom plots: differences is black lines, red stars are maximum peaks, and blue stars are minimum peaks. Middle plot is of first difference, bottom is of second difference

adjust the range size to better fit a spectral line feature and is defined as follows: If

$$|x_{p,max}(i+1) - x_{p,min}(i+1)| \geq D \tag{3.3}$$

Then the range must be reduced to

$$\Big(x_{p,max}(i), x_{p,max}(i+1) + floor(\frac{1}{\eta}|(x_{p,max}(i+1) - x_{p,min}(i+1)))\Big). \tag{3.4}$$

The value, D, is the minimum separation distance for a pair of subsequent peaks (max. and then min.) when the x-values are integers, it is determined by the user. η controls how much the separation distance will be between $x_{p,max}(i+1)$ and the new range indicator. Setting $\eta = 2$ will have peaks that pass the condition use half of the separation distance between said peaks. A second 'range indicator check' is introduced to create better visual clarity and also fix the size of the ranges, it is defined as follows

$$|y_{p,max}(i+1) - y_{p,min}(i+1)| \geq T \tag{3.5}$$

and

$$|y_{p,min}(i+1) - y_{p,max}(i+2)| \geq T \tag{3.6}$$

Then a feature exists in the range $(x_{p,max}(i), x_{p,min}(i+2))$. This condition looks at the subsequent pair of peaks before the program starts the next iteration to verify if the next set of peaks corresponds to a feature. If the condition is passed then the range is increased to encompass both the first and second line features. The result of these conditions can be seen in Sec. 4.2.

3.2 Practical Applications

ALINE, as an idea, is a powerful tool, it identifies the location of sharp features automatically. This has obvious implications in GW detector data because the spectral line features require filtering to discover chirps in the data and the location of the features is not always the same. Identifying exactly where to filter is incredibly handy; however, sometimes identifying the location is the prime

goal.

In neuro-anatomy it is important to understand the difference between voluntary and involuntary actions in the brain for use in prosthesis. A way to quantify the difference is by analyzing the action potentials of the two types of movements and identifying the length of time it takes for the brain to respond. It can be measured by attaching nodes to a subject's muscles and monitoring the electromyograms or EMGs for spikes of activity. Similar to GW data, activity is not static, constantly shifts, and sometimes has random interference which hinders ones ability to determine what the source of a feature is. Applying ALINE to this example can help determine exactly where a spike is which can be cross-referenced with biological feedback to fine-tune a prosthetic device for finer motor movement [9].

ALINE can be used in analytic chemistry to identify emission lines in laser-induced breakdown spectroscopy [10] or in astronomy, identifying stellar spectra of diffuse interstellar bands or DIBs. DIBs are absorption features that are observed in the spectra of astronomical objects, believed to be caused by interstellar mediums [17]. The bands span from Near-Infrared (NIR) to Near-Ultraviolet (NUV) wavelengths and are ubiquitous in spectra across the galaxy. The problem with DIBs is their unknown origin, or carriers, which are unique configurations of elements like C_{60}^{+}, one of the few confirmed carriers of several DIBs [8]. Determining the origin of these DIBs requires precise modeling of stellar line lists to estimate the wavelengths of the spectral lines [25]. ALINE specializes in identifying the location of bands of spectral line features and can easily be used on such data to improve the line lists.

CHAPTER IV

RESULTS

This chapter discusses the results of the research in detail. Sec. 4.1 covers how effective SHAPES-Ext takes estimates on Donoho-Johnstone benchmark test functions and Sec. 4.2 delves into how well ALINE can identify the location of lines in GW detector data.

4.1 SHAPES-Ext on Benchmarks

The first part of the research involved expanding SHAPES to be able to comfortably run on large data sets, this was done by segmenting the data in sizable pieces with overlaps for SHAPES to run on independently. The segment-wise estimates were then connected via a weighted exponential averaging by using the overlap of each segment. The Donoho-Johnstone Benchmark Test functions, or DJBM, were used to test SHAPES-Ext and the segmentation. All benchmark functions had pseudo-random normally distributed noise, or `randn`, added to their data to simulate noisy PSDs.

4.1.1 Blocks

Figure 4.1 shows how well SHAPES performs on discontinuities. Here, the colored segments are the non-overlapping segments with averaged points, the black line is the original test function before the noise realization and the grey points are the result of realized Gaussian noise. The general shape of the blocks function is detected by SHAPES and the sharpness of most blocks is replicated by the estimate. The denser regions of knots, or the black triangles on Fig. 4.1, normally congregate where the blocks are located because of the sharp discontinuity. The narrower blocks are more rounded because SHAPES has not worked out if the small collection of points is part of a curve or a discontinuity. The rounding is often accompanied by an early departure from the model, most notably at the 900 block, likely due to the particular noise realization with some points extending

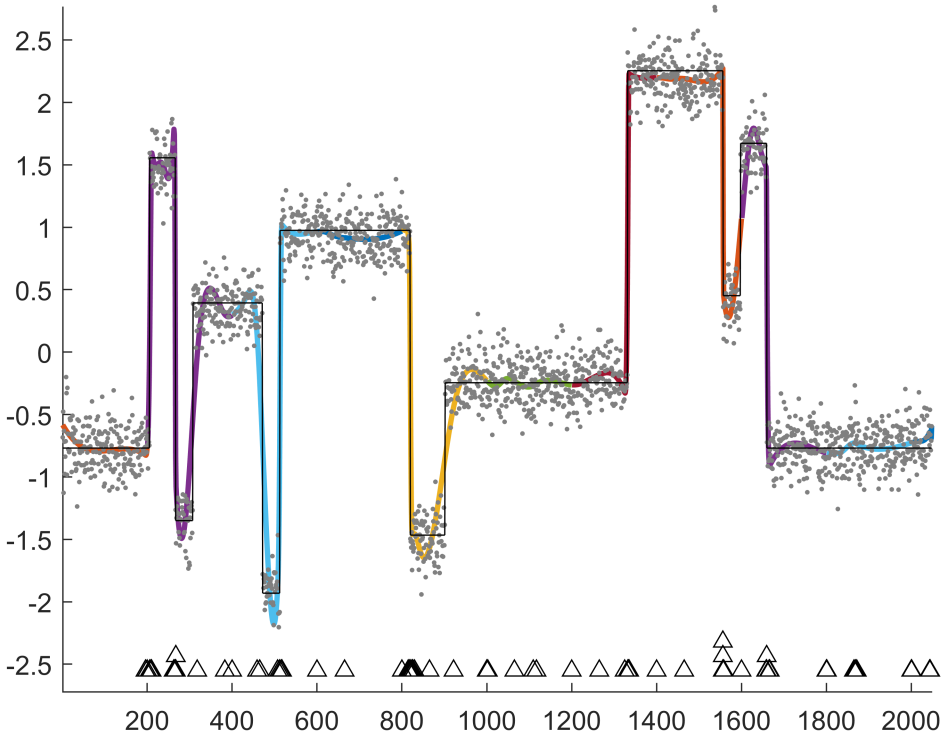

Figure 4.1: DJBM Blocks function SHAPES estimate with $N_n = 200$, $\lambda = 0.05$ with a signal-to-noise ratio, or SNR, of 400. The colored lines are the SHAPES estimated non-overlapping segments, grey dots are raw data with Gaussian white noise, black lines are the raw data, and black triangles are the knot locations. Stacked knots are repeated knots, indicating higher multiplicity due to discontinuities.

far towards the next block. Counterexamples exist to substantiate this claim; the end of segment 1, the 1,350 block, and 1,650 block all have lower noise realizations on block edges and accordingly, trace the block model to a higher fidelity.

4.1.2 Bumps

Figure 4.2 gives preliminary information on SHAPES' ability to capture lines. SHAPES excels in this test with an estimate that picks up each feature exactly with no visible disconnect from the raw data. What can also be seen is the lack of an overfit by SHAPES, it identifies the superfluousness of capturing every single point at the peaks of the line features, and distinguishes noise floor peaks from spectral line features. The stellar performance of SHAPES on the bumps

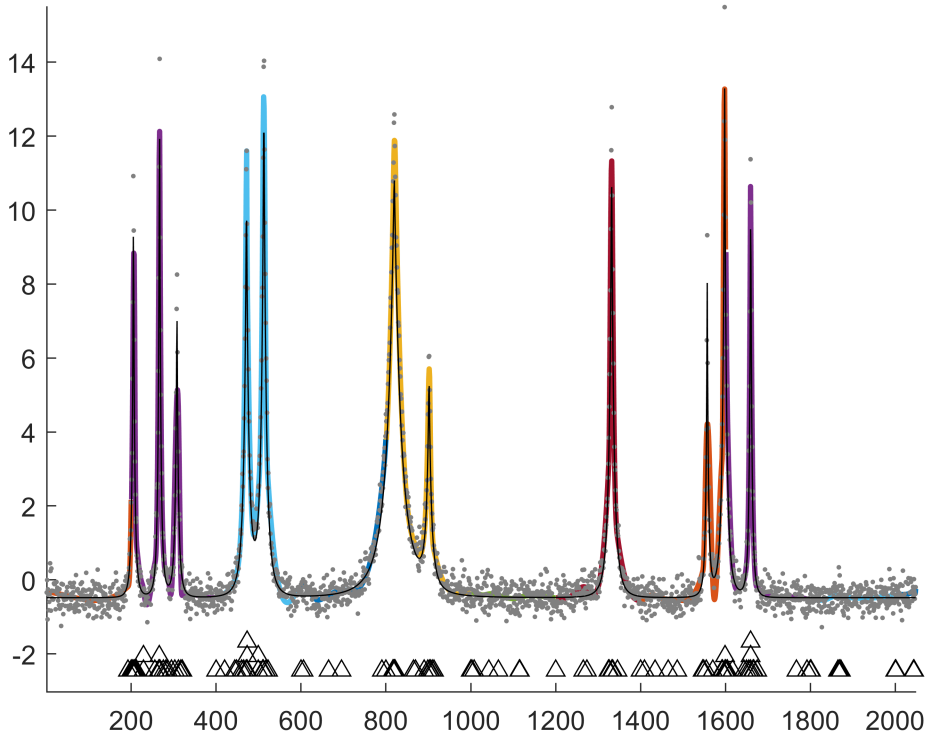

Figure 4.2: DJBM Bumps function SHAPES estimate with $N_n = 200$, $\lambda = 0.05$ and $SNR = 400$. The colored lines are the SHAPES estimated non-overlapping segments, grey dots are raw data with Gaussian white noise, black lines are the raw data, and black triangles are the knot locations.

function solidifies SHAPES as an estimation tool capable enough to be used on GW detector data. Bumps on average has more knots than the blocks function, with several dense regions coalescing under line features which might be due to the type of discontinuity. On segments with multiple features, SHAPES does have a slight tendency to exaggerate the important of noise; at sample 550 when the light blue curve is meeting the noise floor, the estimate dips farther than the noise tends. The tendency is more prevalent in higher-knot number segments since higher-knots nearly always indicate a complexity in the data.

4.1.3 Heavisine

Figure 4.3 highlights how SHAPES handles smooth curves with discontinuous pieces. SHAPES estimates the heavisine, similar to the bumps result, quite well, with the discontinuous

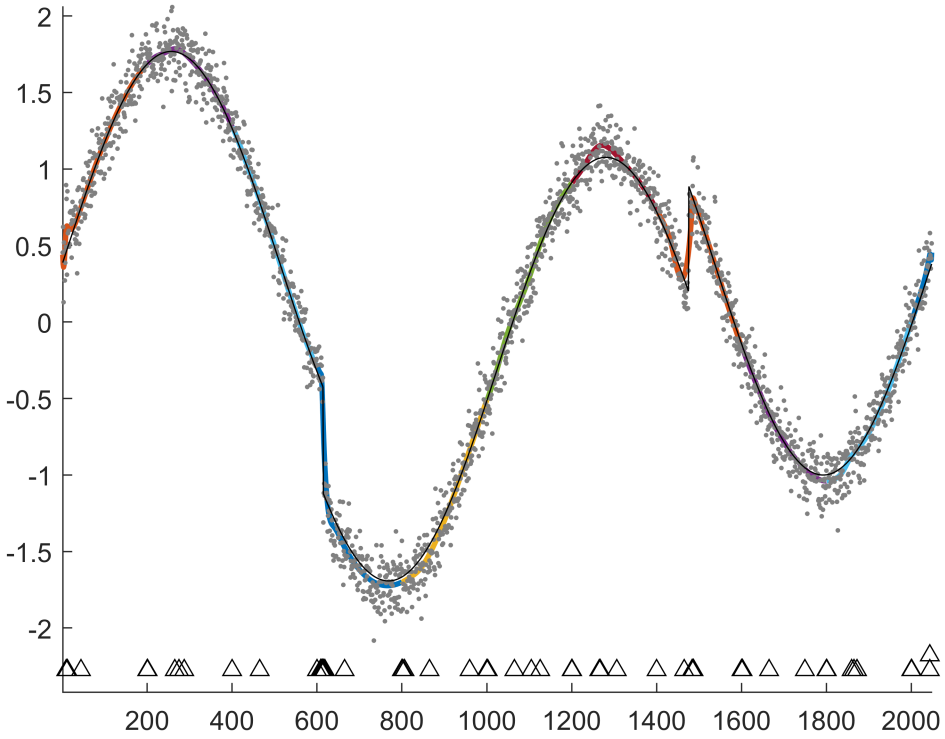

Figure 4.3: DJBM Heavisine function SHAPES estimate with $N_n = 200$, $\lambda = 0.05$ and $SNR = 400$. The colored lines are the SHAPES estimated non-overlapping segments, grey dots are raw data with Gaussian white noise, black lines are the raw data, and black triangles are the knot locations.

portions matching perfectly with the data. The function has far fewer knots than all other test functions because of the simplicity of the function. Most knots congregate at the extrema of the curves or at the two discontinuous segments, this behavior is consistent with the previous functions. The high-knot tendency referenced before is evident here in the second red segment around 1,300 samples; the segment features a discontinuity inclining SHAPES to apply more knots resulting in the estimate peaking further into the noise realization. Similar segments devoid of higher knot numbers, like those around 250 and 1,800 samples, have high reaching noise but more closely follow the model.

4.1.4 Doppler

Figure 4.4 showcases SHAPES on a Doppler function. The Doppler function has some features that make it difficult to estimate, the abundance of sharp features at the beginning being the most notable. To address the issue, the size of the segments was adjusted to better capture those features, however, there are simply too many features to reasonably pick up each one. A more drastic change would be allow much higher knot numbers per segment, this would allow the earlier segments to pickup more features with more knots available to place. Despite these fixes in place, most of the features are not caught below 150 samples due to the sheer amount of oscillations. Sec. 5.3 discusses future adjustments to the expanded SHAPES which could fix this issue. Its important to keep in mind the size of the segments when considering the knot density, the lower the segment sizes the more dense the knots will appear. However, the knots of this function follow as one would expect with the previous examples in mind; where the features are close and narrow in breadth or at extrema, there is a denser population of knots. The changes to the size of the segments help the early rapidly oscillating parts of the Doppler function but hinder the estimate later on. Between the 1,500 and 1,900 samples, the estimate is attempting to fit the noise because the segments do not span a large curve and since the data is so tightly compacted. The segments are unaware of what the segments around it are fitting nor what is the 'big picture.' Implementing a 'neighborhood' system whereby adjacent segments could merge or disconnect for a superior fit could be a worthwhile pursuit.

4.2 ALINE on GW Detector Data

The second part of the research focuses on applying SHAPES-Ext to GW detector data and then using the second difference of the estimate and its peaks to determine the location of line features for filtering. GW170817 data, the second three-way detection of gravitational waves, is used as samples for testing, specifcally from the Hanford and Virgo detectors.

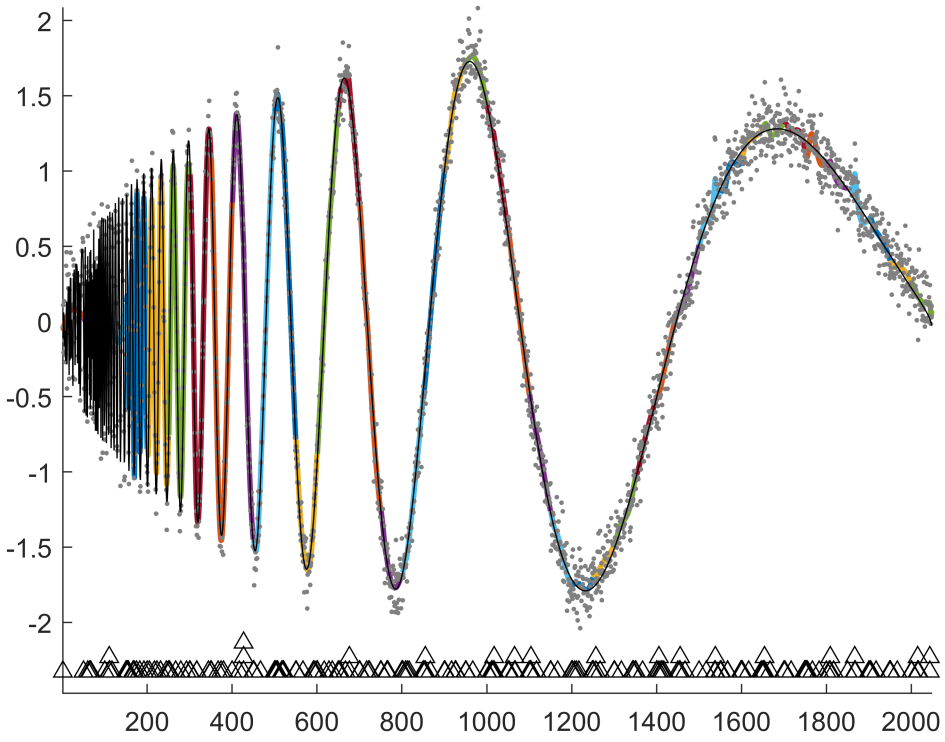

Figure 4.4: DJBM Doppler function SHAPES estimate with $N_n = 50$, $\lambda = 0.05$ and $SNR = 400$. The colored lines are the SHAPES estimated non-overlapping segments, grey dots are raw data with Gaussian white noise, black lines are the raw data, and black triangles are the knot locations.

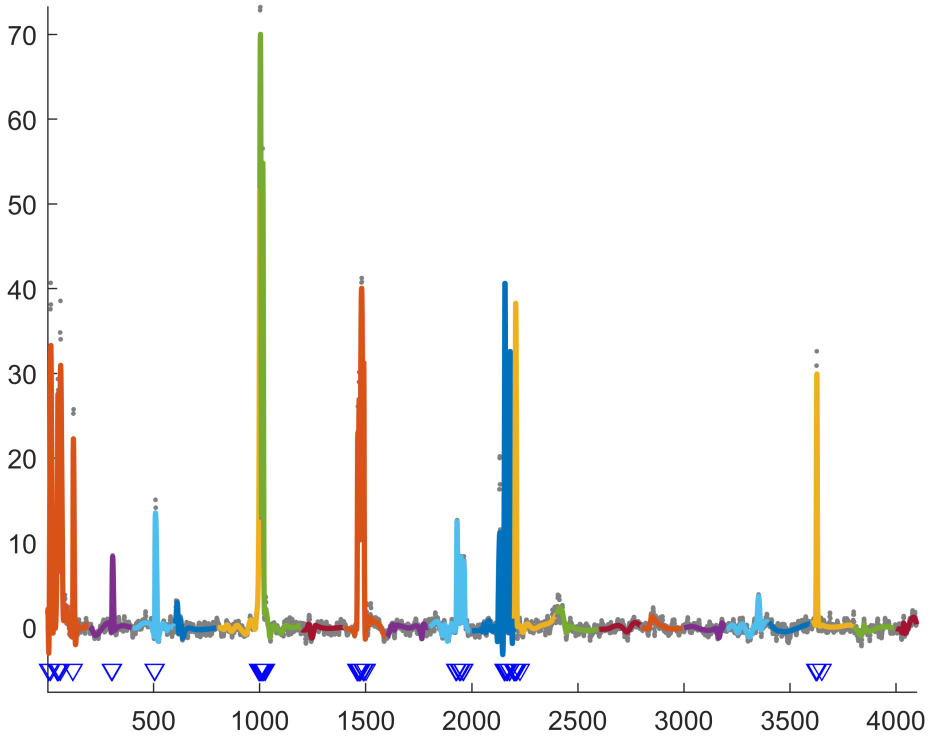

Figure 4.5: Hanford Detector Data from GW170817 estimated by SHAPES with line identification by ALINE. The colored lines are the non-overlapping samples sized at $N_n = 200$, the raw data is represented by the grey dots, and the blue triangles indicate the range where a spectral line feature is present between. $\lambda = 0.05$, window size $w_{sz} = 12$, and the peak second difference threshold, $T = 1.75$.

4.2.1 Hanford

In Figure 4.5 ALINE is successfully able to capture all prominent features with the threshold scheme of $T = 1.75$ for the peak second difference. Smaller features like those around 2,400 and 3,400 could be identified with a lower threshold value. However, the ranges identified do not completely encompass the line and there appears to be too many ranges for a single feature, i.e. at 1,000 samples.

To simplify the figure and the ranges, a change to the threshold scheme is made that checks if the peak second difference is larger than T to indicate the presence of a line and if the next peak difference is larger than . The range identified should be a combination of both of the two ranges.

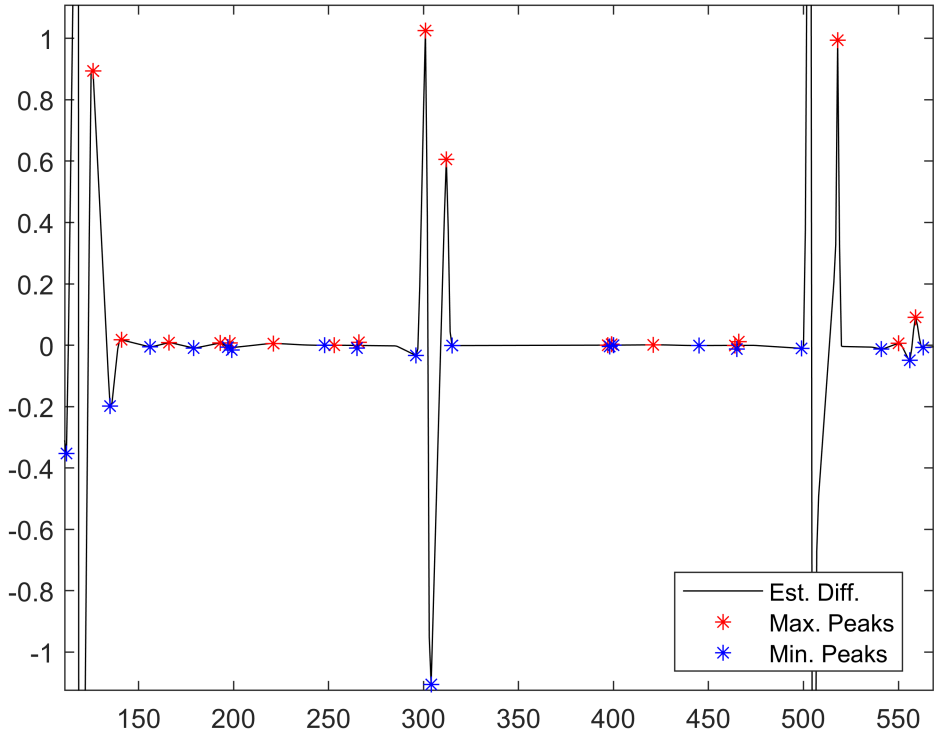

Figure 4.6: Peak Second Difference of Hanford whitened data centered at line feature 300. $N_n = 200$, $\lambda = 0.05$, and $w_{sz} = 12$. Red stars indicate the maximum local peaks, blue stars indicate the minimum local peaks, and the black solid line is the second difference of the estimate.

An illustration of the second difference is pictured in Figure 4.6 which is a closeup of the peak difference plot for the line feature at 300 samples. If the difference between a max. peak and the subsequent min. peak is larger than T, it will be identified as a feature with a range between the x-locations of those peaks, this is the first threshold scheme. A second threshold scheme is put in place to expand the range if the following set of peaks passes a threshold of half of T, since a line feature's second derivative is indicated by a max. peak, followed by a min. peak, then another max. peak as shown in Figure 4.6. The size of the second threshold is half of T because the min. peak following the second max. peak is located at the floor of the function. If the feature passes both thresholds, the new range will be the x-locations of the two max. peaks. Figure 4.7 showcases the new threshold scheme. The modified scheme not only increases precision by expanding the range where suitable, but also improves visual clarity.

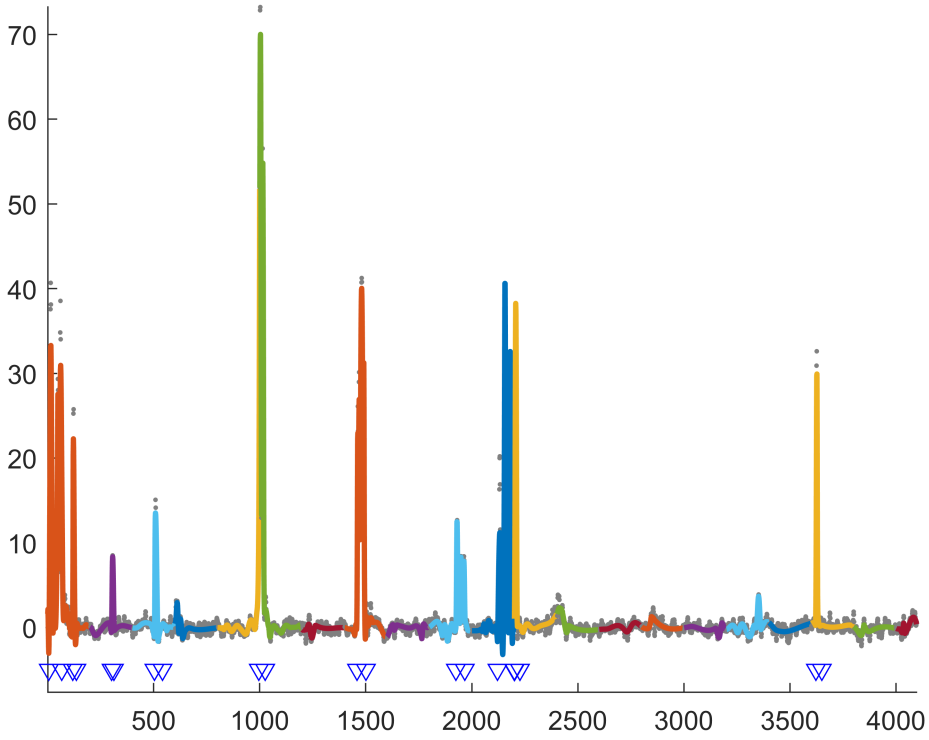

Figure 4.7: Hanford Detector Data from GW170817 estimated by SHAPES with line identification by ALINE. The colored lines are the non-overlapping samples sized at $N_n = 200$, the raw data is represented by the grey dots, and the blue triangles indicate the range where a spectral line feature is present between. $\lambda = 0.05$, window size $w_{sz} = 12$, and the peak second difference threshold, $T = 1.00$.

4.2.2 Virgo

In this thesis, the Virgo detector mentioned refers to the Advanced Virgo Detector which is still fairly new, and so the Virgo data, from GW170817, contains many more spectral line features. This is due to the higher frequency resolution than that of Hanford's to counteract the noisier Virgo data. In Figure 4.8 we can see from the top plot that not all the features are estimated by SHAPES, mostly those with values less than 4 or 5. The feature around 700, in the dark blue, is one of the most egregious examples with one line featured but not the other, other similar features are apparent throughout the example. The reason for this mistake is unknown but must be caused by an error in the base SHAPES algorithm. One proposed and enacted solution was to lower the window size from 12, the window for Hanford's data, to 3 which made the features slightly more prominent with small curves, but not sharp lines as the raw data shows. Another test with no windowing yielded worse results with more features disappearing.

ALINE is able to capture most features but appears to have trouble with clumped line features, this is most likely due to the lower threshold allowing too much through. There are some issues with the range on the wider singular features, for example at 1,500, the two indicators are stacked very close to each other whereas the other feature on that segment is not identified. Since the method for identification relies on the features having second differences that are higher than a threshold, smoother lines can be disregarded despite clearly appearing in an estimate. This issue was discovered early on in research, first with the Hanford data when the estimates for the smaller features were not as sharp as they should be. Eventually this problem was resolved when the estimates sharpened the weaker features, but the features in Virgo are naturally wider in breadth which resurfaced the problem. A solution to this issue is two kinds of thresholds, a line would need to be pass one or the other to be considered a feature; the first would be the current threshold scheme with peak second differences, the second would be a value or a peak first difference threshold based on the noise floor of the data set.

The bottom plot of Figure 4.8 gives the knot numbers for each segment. This plot was instrumental in understanding the knot sequence SHAPES requires for a data set. Simpler data

sets originally saw a knot sequence like equation 2.1 which worked well because the samples were smooth or short, but as the data sets grew more complex with sharper or denser features, the need for an expanded knot sequence became apparent. The knot sequence for the Virgo data may need an extension seeing as the 12th and 13th segments, corresponding to the 2,500 sample feature, hit the max knot value for the sequence. The greatest inference one can take away from the bottom plot is that higher knot numbers indicate the presence of a line granted they are high enough, i.e. segment 6 has just as many knots as segment 7 but does not show a segment in the estimate whereas segment 7 clearly has a sharp feature. Likewise, segment 17 has a feature and a half but only uses 6 knots for them. Another indicator, in conjunction with knot numbers, could prove to be a potent alternative method.

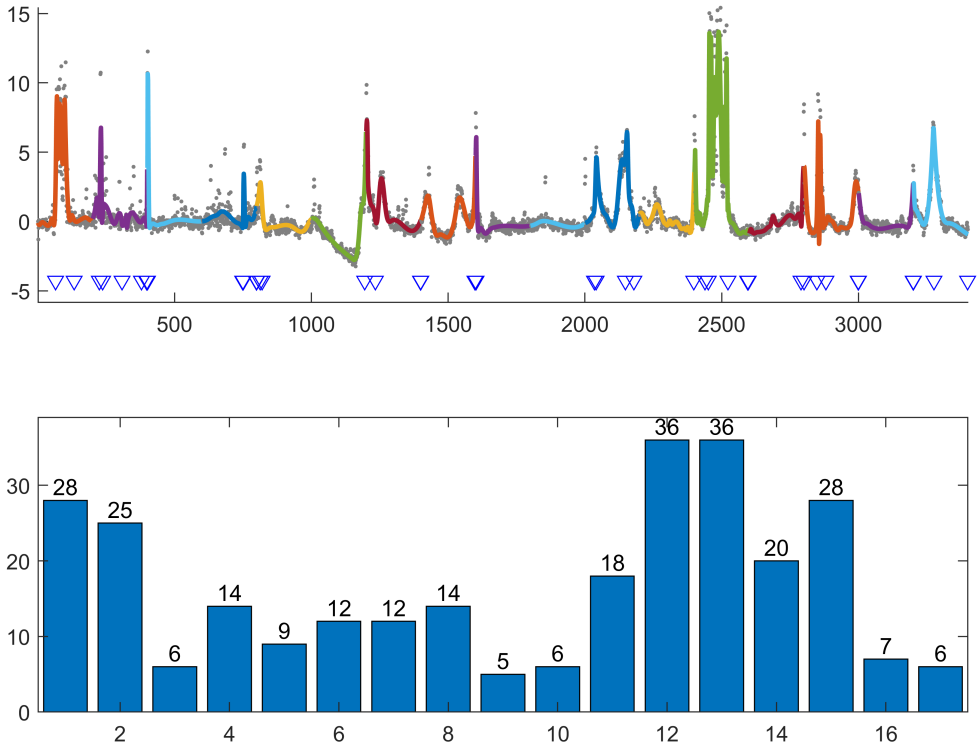

Figure 4.8: For the top plot: Virgo Detector Data estimated by SHAPES and with line identification by ALINE. The colored lines are the non-overlapping samples sized at $N_n = 200$, the raw data is represented by the grey dots, and the blue triangles indicate the range where a spectral line feature is present between. $\lambda = 0.05$, window size $w_{sz} = 3$, and the peak second difference threshold, $T = 0.05$. For the bottom plot: Segment Knot numbers for each segment of the Virgo Estimate.

CHAPTER V

CONCLUSIONS

The final chapter concludes the research with notes on how well SHAPES works on large data sets in Sec. 5.1, ALINE's efficiency in capturing line features in Sec. 5.2, and what the future holds for both algorithms in Sec. 5.3.

5.1 SHAPES Estimate reliability

The SHAPES estimates on the DJBM functions were fairly reliable in fitting the estimate, with only the start of the Doppler function missing some features. With the Hanford data, the estimate managed to capture every feature. The Virgo data has been shown to be difficult and is surely in need of further tuning, but despite this most features from the data were estimated successfully. Overall with the six examples tested, SHAPES excelled in finding an estimate for four and most of the others which is remarkable given the varied shapes and lengths of the functions. Our results show that the extended version of SHAPES is a reliable estimation tool for data samples similar to the benchmark functions and GW detector data when the frequency resolution is in a suitable range. The expansion methods laid out in this paper worked well to allow SHAPES-Ext to work on any-sized data set, further tuning is only needed for the base SHAPES algorithm to capture additional features.

5.2 ALINE's Efficiency

ALINE was successfully able to identify all visible features from the estimate. However, the algorithm is only as strong as the estimate allows and appeared to miss key features in the Virgo data. Careful manipulation of the threshold value should allow all spectral line features to be identified because the threshold is entirely based on the input data values. The addition of a

secondary threshold scheme helped solidify the range of some features that seemed inconsistent with only the first threshold as well as helped with the visual clarity of the ranges in the figures.

5.3 The Future

In this paper we've discussed how SHAPES can be expanded to estimate data of any input size and also how to identify spectral line features from that estimate. However, the research does not end here, future endeavors aim to resolve the issues uncovered by this study.

Current points of investigation include the modification of the SHAPES-Ext to estimate dense populations of features, like the Doppler test function. A possible solution is allowing SHAPES to run on the data twice, once to determine the amount of knots required for segments and a second time that uses the information of the previous run to pick a more optimal knot sequence. For example, a simple data set like bumps, does not require a large amount of knots for most of its segments. SHAPES could find the knot numbers of each segment and create a unique knot sequence for each segment tuned to their ideal set of knots. For the segments with sharp spectral features, the knots sequences would be, on average, higher than the segments between the features.

A similar solution would aim to create an adaptive segmentation algorithm, whereby segments with high knot numbers will be separated until the segment only contains the feature. This solution removes the need for threshold scheming but relies more heavily on accurate knot numbers, based on the assumption that higher knot numbers or multiplicity suggest the presence of a feature, which if preliminary investigations are any indication, is not a strong basis. Further investigations may be needed to reach a concrete verdict. A mentioned alternative could be the 'neighborhood' system that adaptively merges or disconnects segments based on if the estimate of a different sized segment is better suited to the model. The solutions all carry varying strengths but most often the burden is whether the computational expenditure is worth the effort.

The other major problem discovered is the missing of features in the Virgo estimate. Adjusting the main parameters; window size, segment size, and regulator gain, did not modify the results by a noticeable amount whereas with other estimates they did. This begs the question, what variable needs to be a parameter with Virgo-like data sets? Future investigations may need to alter the PSO

parameters which went unchanged throughout the experiment as suggested by the literature [24].

As discussed in 4.2.2, a multi-threshold system is currently being investigated as a means of resolving identification issues with smooth features. A vetting system could be created that is based on knot numbers that checks itself with a threshold scheme so there is multiple ways a line could be identified. This vetting system lends its way to a final point of study worth investigating, expanding the ability of ALINE to identify specific shapes of curves, and not merely lines.

BIBLIOGRAPHY

[1] B. P. ABBOTT, R. ABBOTT, T. ABBOTT, M. ABERNATHY, F. ACERNESE, K. ACKLEY, C. ADAMS, T. ADAMS, P. ADDESSO, R. ADHIKARI, ET AL., *Observation of gravitational waves from a binary black hole merger*, Physical review letters, 116 (2016), p. 061102.

[2] B. P. ABBOTT, R. ABBOTT, T. ABBOTT, F. ACERNESE, K. ACKLEY, C. ADAMS, T. ADAMS, P. ADDESSO, R. ADHIKARI, V. ADYA, ET AL., *Gw170817: observation of gravitational waves from a binary neutron star inspiral*, Physical Review Letters, 119 (2017), p. 161101.

[3] B. P. ABBOTT, R. ABBOTT, T. D. ABBOTT, S. ABRAHAM, F. ACERNESE, K. ACKLEY, C. ADAMS, V. B. ADYA, C. AFFELDT, M. AGATHOS, ET AL., *A guide to ligo–virgo detector noise and extraction of transient gravitational-wave signals*, Classical and Quantum Gravity, 37 (2020), p. 055002.

[4] R. ABBOTT, R. ADHIKARI, G. ALLEN, S. COWLEY, E. DAW, D. DEBRA, J. GIAIME, G. HAMMOND, M. HAMMOND, C. HARDHAM, ET AL., *Seismic isolation for advanced ligo*, Classical and Quantum Gravity, 19 (2002), p. 1591.

[5] F. A. ACERNESE, M. AGATHOS, K. AGATSUMA, D. AISA, N. ALLEMANDOU, A. ALLOCCA, J. AMARNI, P. ASTONE, G. BALESTRI, G. BALLARDIN, ET AL., *Advanced virgo: a second-generation interferometric gravitational wave detector*, Classical and Quantum Gravity, 32 (2014), p. 024001.

[6] H. AKAIKE, *Information theory and an extension of the maximum likelihood principle*, in Selected papers of hirotugu akaike, Springer, 1998, pp. 199–213.

[7] S. BISCOVEANU, C.-J. HASTER, S. VITALE, AND J. DAVIES, *Quantifying the effect of power spectral density uncertainty on gravitational-wave parameter estimation for compact binary sources*, Physical Review D, 102 (2020), p. 023008.

[8] E. K. CAMPBELL, M. HOLZ, D. GERLICH, AND J. P. MAIER, *Laboratory confirmation of c 60+ as the carrier of two diffuse interstellar bands*, Nature, 523 (2015), pp. 322–323.

[9] E. A. CLANCY, S. BOUCHARD, AND D. RANCOURT, *Estimation and application of emg amplitude during dynamic contractions*, IEEE Engineering in Medicine and Biology Magazine, 20 (2001), pp. 47–54.

[10] D. A. CREMERS, R. A. MULTARI, AND A. K. KNIGHT, *Laser-induced breakdown spectroscopy*, Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation, (2006), pp. 1–28.

[11] C. DE BOOR, *A practical guide to splines*, vol. 27, springer-verlag New York, 1978.

[12] D. DONOHO AND I. JOHNSTONE, *Ideal spatial adaptation via wavelet shrinkage. biometrika. to appear*, tech. rep., Also Tech. Report, Department of Statistics, Stanford University, 1992.

[13] M. C. EDWARDS, R. MEYER, AND N. CHRISTENSEN, *Bayesian semiparametric power spectral density estimation with applications in gravitational wave data analysis*, Physical Review D, 92 (2015), p. 064011.

[14] A. EINSTEIN AND N. ROSEN, *On gravitational waves*, Journal of the Franklin Institute, 223 (1937), pp. 43–54.

[15] G. HAMMOND, S. HILD, AND M. PITKIN, *Advanced technologies for future ground-based, laser-interferometric gravitational wave detectors*, Journal of Modern Optics, 61 (2014), pp. S10–S45.

[16] F. J. HARRIS, *On the use of windows for harmonic analysis with the discrete fourier transform*, Proceedings of the IEEE, 66 (1978), pp. 51–83.

[17] G. H. HERBIG, *The diffuse interstellar bands*, Annual Review of Astronomy and Astrophysics, 33 (1995), pp. 19–73.

[18] W. HÄRDLE, *Applied Nonparametric Regression*, Econometric Society Monographs, Cambridge University Press, 1990.

[19] J. KENNEDY AND R. EBERHART, *Particle swarm optimization*, in Proceedings of ICNN'95-international conference on neural networks, vol. 4, IEEE, 1995, pp. 1942–1948.

[20] T. B. LITTENBERG AND N. J. CORNISH, *Bayesian inference for spectral estimation of gravitational wave detector noise*, Physical Review D, 91 (2015), p. 084034.

[21] D. V. MARTYNOV, E. HALL, B. ABBOTT, R. ABBOTT, T. ABBOTT, C. ADAMS, R. ADHIKARI, R. ANDERSON, S. ANDERSON, K. ARAI, ET AL., *Sensitivity of the advanced ligo detectors at the beginning of gravitational wave astronomy*, Physical Review D, 93 (2016), p. 112004.

[22] S. MIYOKI, *Current status of kagra*, in Ground-based and Airborne Telescopes VIII, vol. 11445, International Society for Optics and Photonics, 2020, p. 114450Z.

[23] S. D. MOHANTY, *Swarm intelligence methods for statistical regression*, CRC Press, 2018.

[24] S. D. MOHANTY AND E. FAHNESTOCK, *Adaptive spline fitting with particle swarm optimization*, Computational Statistics, 36 (2021), pp. 155–191.

[25] A. MONREAL-IBERO AND R. LALLEMENT, *Measuring diffuse interstellar bands with cool stars-improved line lists to model background stellar spectra*, Astronomy & Astrophysics, 599 (2017), p. A74.

[26] C. J. MOORE, C. P. BERRY, A. J. CHUA, AND J. R. GAIR, *Improving gravitational-wave parameter estimation using gaussian process regression*, Physical Review D, 93 (2016), p. 064001.

[27] K. SOMIYA, *Detector configuration of kagra–the japanese cryogenic gravitational-wave detector*, Classical and Quantum Gravity, 29 (2012), p. 124007.

[28] C. UNNIKRISHNAN, *Indigo and ligo-india: scope and plans for gravitational wave research and precision metrology in india*, International Journal of Modern Physics D, 22 (2013), p. 1341010.

[29] E. J. WEGMAN AND I. W. WRIGHT, *Splines in statistics*, Journal of the American Statistical Association, 78 (1983), pp. 351–365.

[30] J. M. WEISBERG, J. H. TAYLOR, AND L. A. FOWLER, *Gravitational waves from an orbiting pulsar*, Scientific American, 245 (1981), pp. 74–83.

[31] P. WELCH, *The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms*, IEEE Transactions on audio and electroacoustics, 15 (1967), pp. 70–73.

[32] J. ZWEIZIG, *Strain sensitivity of the ligo interferometers, s5 performance - may 2007*.

BIOGRAPHICAL SKETCH

Thomas Cruz was was born in Arlington, Texas. After receiving his high school diploma at W.B. Ray High School in 2015 he attended Texas A & M University-Kingsville where he obtained a Bachelors in Physics with minors in Mathematics and Music in 2018. After graduating he spent 1 year studying Statistics, Analytics, Computing, and Modeling at the same university. Soon after he earned a Master of Science in Physics at The University of Texas Rio Grande Valley in May 2021. During his schooling he shared his passion for physics by working as a teaching assistant and laboratory instructor at both universities. Outside of academia he enjoys playing piano and video games. He can be contacted at thomas.cruz03@utrgv.edu.