University of Massachusetts Dartmouth

Department of Mathematics

An Exploration of Statistical Methods in Determining the Gaussianity of LIGO Detector Data

A Data Science project
by

Ashish Thomas Mathew
(02101455)

Project Advisors:

Dr. Sarah Caudill
University of Massachusetts Dartmouth

Dr. Melissa Lopez
National Institute for Subatomic Physics (Nikhef)

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

May, 2025

Abstract

The Advanced Laser Interferometer Gravitational-wave Observatory (LIGO) is one of the most sophisticated instruments ever built, capable of measuring motion 10,000 times smaller than a proton. It was the first device to detect a gravitational wave (GW) event. While analyzing the GW data from LIGO, the detector noise is approximated to be stationary and Gaussian. However, due to its high sensitivity, a common issue it faces is its susceptibility to glitches: transient, non-Gaussian noise bursts that contaminate the time series data obtained from the detector. This project aims to develop a tool to test if the distribution of the amplitudes of the detector noise is Gaussian or not as a way to assess whether it is clean up to a defined threshold, hence improving the quality of the analysis performed. Taking the null hypothesis that the distribution is Gaussian, we study statistical approaches of normality testing, namely the Shapiro-Wilk test, the Kolmogorov-Smirnov test, and the Anderson-Darling test, to test this null hypothesis against the alternative of the distribution being non-Gaussian. We found that of the three tests, the Shapiro-Wilk test performed the best in detecting non-Gaussian samples with an accuracy of 72.9% and a precision of 91.41%. In comparison, the Kolmogorov-Smirnov test had an accuracy of 61.77% and a precision of 100%, while the Anderson-Darling test had an accuracy of 61.30% and a precision of 100% when detecting non-Gaussianity.

1 Introduction

Nearly a century after Einstein's prediction of the existence of gravitational waves (GWs) in 1916, the first direct gravitational wave detection was achieved by the Advanced Laser Interferometer Gravitational-Wave Observatory (aLIGO) [1] and Virgo [2] interferometers during the binary black hole merger event GW150914 on September 14, 2015. The LIGO Livingston (L1) interferometer underwent several upgrades following this run - effectively a massive redesign - enhancing its sensitivity by 15% to 25% [3]. This improvement was evident during the O2 observing run during which L1, in conjunction with its Hanford (H1) counterpart and Virgo (V1), detected eleven new gravitational wave signals [4]. Subsequently, during the O3 run, all three detectors operated at their best possible sensitivity, leading to the first single-detector GW detection, GW190425, achieved by LIGO Livingston [5]. There have been over 90 GW events recorded at a high level of confidence since the LIGO-Virgo collaboration's inception to date [6].

A consequence of these detectors' sensitivities is their susceptibility to noise. During observation runs, various noise sources, such as seismic noise, suspension thermal noise, and sensing noise, affect the data collected by the interferometers. Among these, the most problematic are **glitches**: transient events caused by non-astrophysical phenomena such as anthropogenic noise, weather conditions, or instrument malfunctions [7, 8]. Glitches manifest as localized bursts of excess power in interferometer time series data and often do not have well-defined sources. They can occur at energy levels and frequencies that overlap with GW signals, thereby mimicking them and increasing the number of false positive detections. During the first half of the third observing run (O3a) H1 and L1 recorded glitches rates of 0.29 min^{-1} and 1.10 min^{-1} respectively, which rose to 0.32 min^{-1} and 1.17 min^{-1} during the second half (O3b) [6]. A notable example of glitches posing such an issue was during the binary neutron star merger event GW170817 [9], during which instrumental noise transients were detected before the event's coalescence time, complicating its detection and subsequent analysis.

The main focus of this project is the idea of Gaussianity: the property of a sample distribution of data being normally distributed, also known as a Gaussian distribution. A **normal distribution** or **Gaussian distribution** is a continuous probability distribution for a random variable where most of the data points are located around the mean, with lesser amounts of data further away. The Gaussian distribution is symmetric about the mean μ, which is equal to zero, with the standard deviation σ being equal to 1. This distribution is important as it can approximate many real world phenomena, such as the distribution of heights and weights of a country's population. It is also the basis for many statistical tests and methods. Figure 1 shows an example of a Gaussian distribution. Any variation from this distribution is considered non-Gaussian.

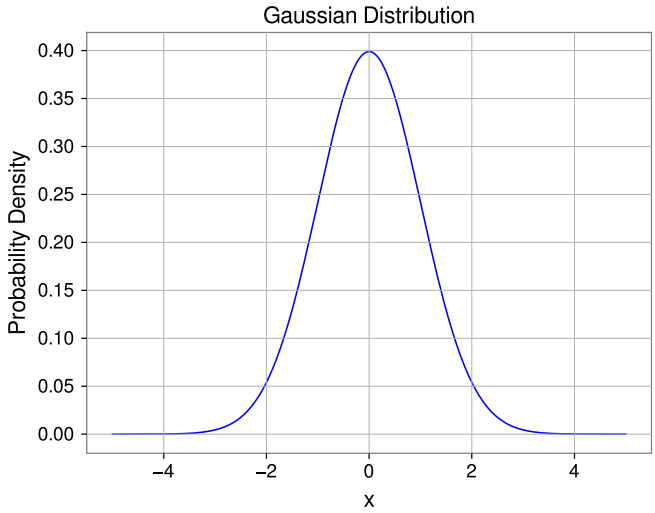


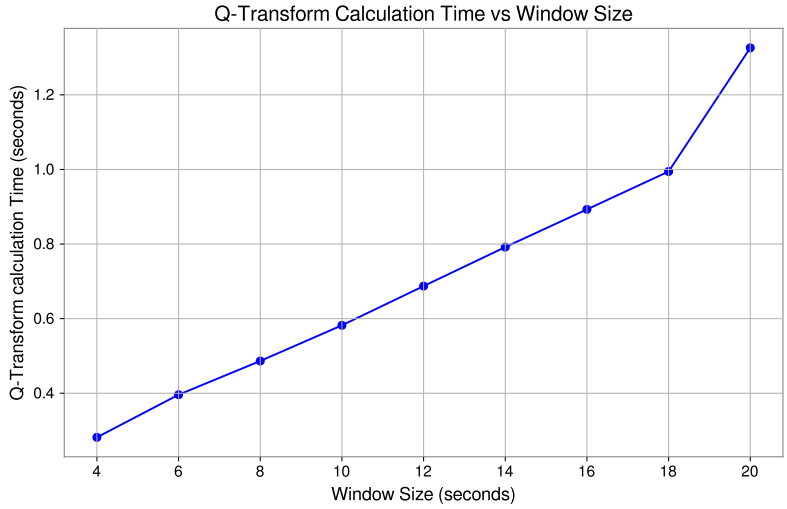
Figure 1: Example of a Gaussian distribution

All these noise sources collectively produce time series signals which can be treated as stochastic processes with their corresponding joint probability distributions and statistical properties [7]. In gravitational wave data analysis the noise floor of the detectors is approximated to be stationary and Gaussian. In the event of a glitch, the noise has some structure with a high signal-to-noise ratio (SNR), making it *non-Gaussian*. This project aims to develop a tool to test if the distribution of the amplitudes of detector noise is Gaussian or not to help improve the quality of the analysis performed. The null hypothesis (h_0) here is that the distribution is Gaussian, indicating that the signal is clean and the alternative hypothesis is that the distribution is non-Gaussian.

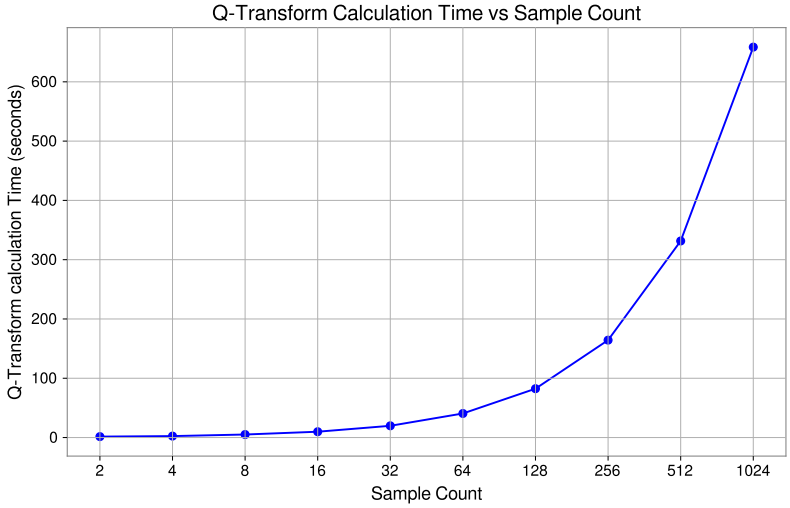
This study focuses on applying (1) the Shapiro-Wilk test, (2) the Kolmogorov-Smirnov test, and (3) the Anderson-Darling test on preprocessed samples of the L1 interferometer data to determine its normality and assess how well these tests differentiate between Gaussian (clean) and non-Gaussian (glitched) samples. This experiment also includes studying the distributions, durations, and frequency ranges, with particular emphasis on failure points for each statistical test across various glitch types. Furthermore, considering the frequency bands at which each glitch type occurs, this study also examines how each statistical test performs under band-pass filtering at various frequency ranges.

This project explores statistical hypothesis testing on time-amplitude domain data as a faster alternative to the current glitch detection methods working in the time-frequency domain. Detecting and mitigating the effects of glitches in interferometric strain data remains an active area

of research within astrophysical data analysis, with several techniques proposed and implemented for the same [10–12]. However, many of these are computationally expensive, slowing down the process of assessing of the signals for glitch activity. The more popular solutions incorporate the *Q-transform* [13, 14], a modification of the standard Fourier transform where the analysis window scales inversely with frequency. The Q-transform is used as inputs to statistical tests to obtain a p-value depicting the statistical significance of excess power in the data. Assuming the background data to be Gaussian, the excess power that does not follow a Gaussian distribution suggests the presence of a glitch. While this method is effective for post-detection run analysis, it is much slower when scaled up to work on multiple samples. Figures 2a and 2b illustrate the time taken to compute the Q-transform while varying the sample window size and the number of samples respectively.



(a) A plot of the Q-transform calculation time against the sample window for a single



(b) Interference pattern

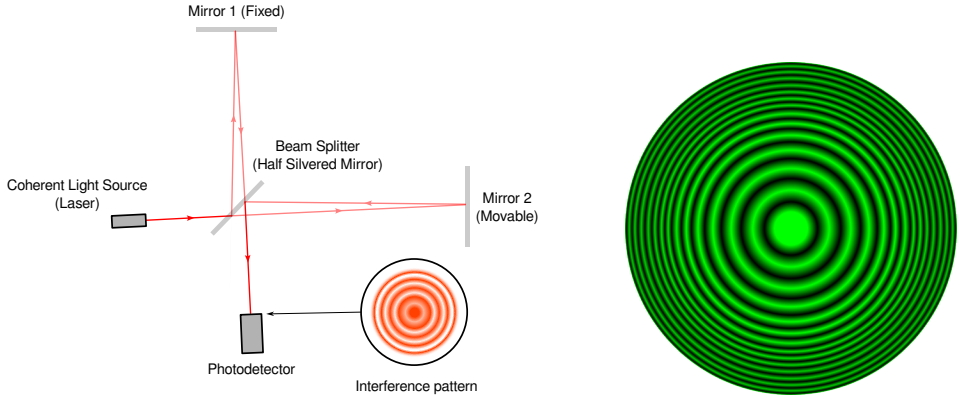
Figure 2: Illustrations of a basic Michelson interferometer [15, 16] and what an interference pattern would look like.

This report is organized as follows: Section 2 discusses how interferometer data is obtained and describes the properties of the time series data used in this study. It also outlines the process of acquiring clean and glitched samples and the preprocessing applied. Section 3 describes in detail the statistical tests of Gaussianity used to assess the sample data. The experimental procedures and results are presented in Sections 4.1 and 5. Finally, Section 6 summarizes the findings of our experiments, discussing the strengths and limitations of the methods, followed by the future scope in Section 7.

2 Data Acquisition and Conditioning

The Advanced LIGO and Virgo interferometers are large-scale, heavily modified versions of the Michelson interferometer (Figure 3a), originally invented by American physicist Albert A. Michelson in 1887 [6, 17]. LIGO operates in a vacuum, using a laser beam of light split into two orthogonal parts with the help of a half-silvered mirror mounted on horizontal seismometer suspensions. The orthogonally split beams are then sent through two arms of the interferometer, known as Fabry-Pérot cavities. Each of these arms is 4 kilometers long, consisting of mirrors at each end. The first of the mirrors is fixed in place, while the second one is movable with the help of precise micrometer drives. The laser beams, when reflected by the mirrors on either end of the arms, are combined again at the beam splitter and sent to photoelectric detectors.

The advanced LIGO interferometer is designed such that if the beams travel equal distances, the recombination would lead to a destructive interference pattern, resulting in no light coming out of the instrument [18]. When a disturbance of either astrophysical or terrestrial origin is detected, an infinitesimally small change occurs in lengths of the detector arms, in the order of $10^{-8}m$ [19]. One of the arms is stretched, and the other arm is compressed in the perpendicular direction, altering the lengths of the reflected and combined beams, misaligning them and resulting in an interference pattern (Figure 3b). This interference pattern, coupled with results from several other sensors, provides detailed information on the *strain* of the GW or glitch event being observed.



(a) Construction of an Interferometer (b) Interference pattern

Figure 3: Illustrations of a basic Michelson interferometer [15, 16] and what an interference pattern would look like.

The `GWpy` Python package is widely used in this project. It provides a suite of tools to access and condition detector strain data from the Gravitational-Wave Open Science Centre (GWOSC) database [20]. This data is in the time domain, as timestamps in the GPS time system at nanosecond precision, and records the amplitude of the calibrated noise event as a differential change in lengths of the interferometer arms. GWpy handles detector data using the `TimeSeries` object, which is built upon `numpy` arrays. This allows compatibility with most of the core numpy utilities along with custom functions for signal processing, tabular data filtering, and visualization.

We use the strain data from L1 during O3a for this project due to the high rate of occurrence of glitches. Figure 4 shows the steps taken to acquire signal data from the detector and condition it for statistical testing.

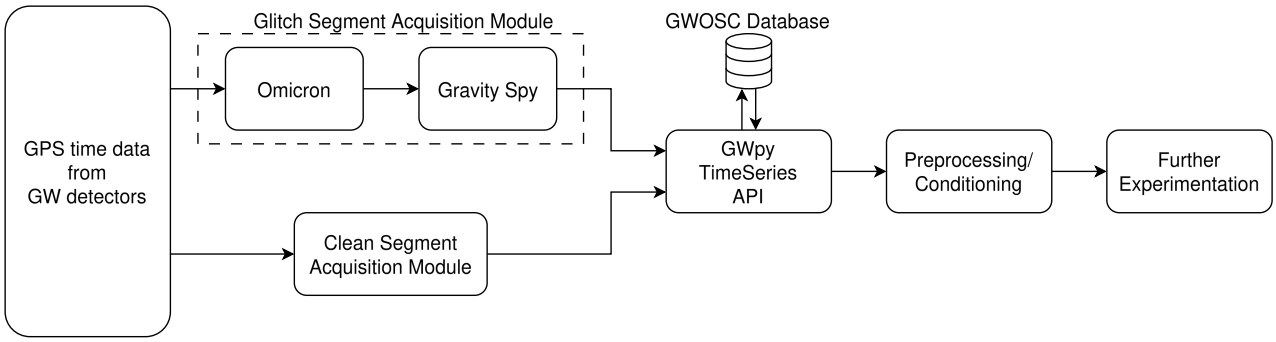


Figure 4: The Data Acquisition and Conditioning Pipeline

There are two main types of detector samples that we receive from the detectors: **Glitched detector data** and **Clean detector data**. The following sections will go a bit deeper into each of these types of data and how they are obtained.

2.1 Glitch data

The GPS times of glitch occurrences are obtained using *Gravity Spy* [21], a large-scale citizen-science project that combines astrophysics, machine learning, and human efforts to classify glitches in GW interferometer data. The *Omicron* transient search algorithm is used by Gravity Spy to generate q-transform spectrograms and calculate SNR of the time series samples [10]. This algorithm is crucial in identifying the most useful samples for data classification and analysis. Based on the morphological characteristics from the spectrograms, a total of 22 glitch classes were identified with an SNR above 7.5 and peak frequencies between 10 Hz and 2048 Hz.

In the O3a data for L1, Fast_Scattering and Tomte glitches are the most prevalent, while Chirp, 1080Lines and Wandering_line glitches have fewer samples, as shown in Table 1. The No_Glitch class represents glitch samples that lack significant traits or energy levels and do not fit in with the other classes morphologically. Hence, for our study, we do not consider this glitch class.

Glitch Class	Count	Glitch Class	Count
Fast_Scattering	21749	Whistle	896
Tomte	18708	Low_Frequency_Lines	788
Blip_Low_Frequency	7549	Scratchy	207
Scattered_Light	5398	Repeating_Blips	164
No_Glitch	5358	Violin_Mode	164
Extremely_Loud	4319	Paired_Doves	155
Koi_Fish	4268	Light_Modulation	72
1400Ripples	2363	Helix	21
Blip	1947	Wandering_Line	20
Power_Line	1189	1080Lines	9
Low_Frequency_Burst	1187	Chirp	6

Table 1: Glitch counts per class for LIGO Livingston (L1) during the O3a run.

Using the glitch GPS times, the time series information is obtained from the GWOSC database using the `TimeSeries.fetch_open_data()` API provided by GWpy. Here, a time window of 10 seconds is taken on either side of each of the GPS times with the glitch in the center, sampled at a rate of 4096 Hz. The TimeSeries objects obtained are then conditioned for statistical testing and cropped down to a 1-second time window around the glitch for further use.

2.2 Clean Samples

The process of finding gaps of clean detector noise is done using `gwtrigfind`, a package developed to search for event triggers files from GW detectors, in conjunction with `EventTable`, provided by GWpy.

Taking the start and end GPS times of the O3a run, gwtrigfind is used to find the file path containing Omicron triggers from the *L1:GDS-CALIB_STRAIN* channel of the L1 detector. EventTable is then used to load all the trigger data, containing information on the start and end GPS times of the events. Taking the time frames between the end and start times of successive events, we obtain the gaps between glitches/GW triggers, which do not have a significant amount of noise activity. The sizes of the gaps are clamped between 7 and 30 seconds because too short of a gap could lead to the inclusion of glitches in the sample, as the time series data may not have enough time to stabilize after a glitch; and too long of a time gap could be indicative of detector malfunction or a period of time when it is not operational.

The GPS start and end times of the clean segments are then used to calculate their Q-transform and corresponding p-values. The GPS time intervals with a p-value greater than 0.95 are considered to be clean segments of data, and are used to obtain the corresponding TimeSeries data from GWpy. The TimeSeries data is then, similar to the glitch data, conditioned for statistical testing and cropped down to 1-second samples for further use.

2.3 Data Conditioning

The strain readings obtained from a GW detector are usually a combination of the GW signal and detector noise [22–24]. Considering the strain output to be $s(t)$ with $n(t)$ as the noise and $h(t)$ being the possible signal received, the GW detector output is given by:

$$s(t) = n(t) + h(t) \tag{1}$$

A lot of the preprocessing in this project involves the use of Fourier transforms. Given the signal $x(t)$ we get its corresponding Fourier transform as

$$\tilde{x}(f) = \mathcal{F}\{x(t)\}(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi i f t} dt \tag{2}$$

and the inverse Fourier Transform given by

$$x(t) = \mathcal{F}^{-1}\{\tilde{x}(f)\}(t) = \int_{-\infty}^{\infty} \tilde{x}(f) e^{2\pi i f t} df \tag{3}$$

2.3.1 Power Spectral Density and Amplitude Spectral Density

In most cases, the noise in the GW detectors is stationary [8], i.e. the characteristics of the noise do not change over time, and the components are, for the most part, uncorrelated, hence keeping it Gaussian. As a result, the statistical properties of the noise can be described by an auto-correlation matrix described by

$$C_n(\tau) = \langle n_a(t+\tau) n_b(t) \rangle - \langle n_a(t+\tau) \rangle \langle n_b(t) \rangle, \tag{4}$$

The Wiener-Khinchin theorem [25] states that the Fourier transform of the correlation matrix of a signal gives us the **power spectral density** (PSD) $S_n(f)$. This would give us the following equation:

$$S_n(f) = 2 \int_{-\infty}^{\infty} C_n(\tau) e^{-2\pi i f \tau} d\tau S_n(f) = 2\tilde{C}_n(\tau) \tag{5}$$

The PSD provides a representation of how power is distributed across different frequency bands in a noise signal[24, 26], allowing for the identification of dominant frequency components and their corresponding amplitudes. Considering the frequency domain representation of the noise signal, the ensemble average of the auto-correlation can be expressed as

$$\langle \tilde{n}(f) \tilde{n}^*(f') \rangle = \frac{1}{2} \delta(f - f') S_n(f). \tag{6}$$

Here $\langle \cdot \rangle$ represents the ensemble average, $\tilde{\cdot}$ represents the Fourier transform from 2, and $\cdot^*$ represents the complex conjugate. Aside from Power Spectral Density, another way to chartacterize the noise of a detector is with the **Amplitude Spectral Density (ASD)** values, given by $\sqrt{S_n(f)}$. The ASD measures the amplitude of the signal at each frequency, and it is often used to compare the noise levels of different detectors or to assess a detector's sensitivity to specific frequencies. The ASD is normally plotted on a logarithmic scale, with frequency on the x-axis and the ASD on the y-axis, allowing for easy identification of noise peaks and their corresponding frequencies.

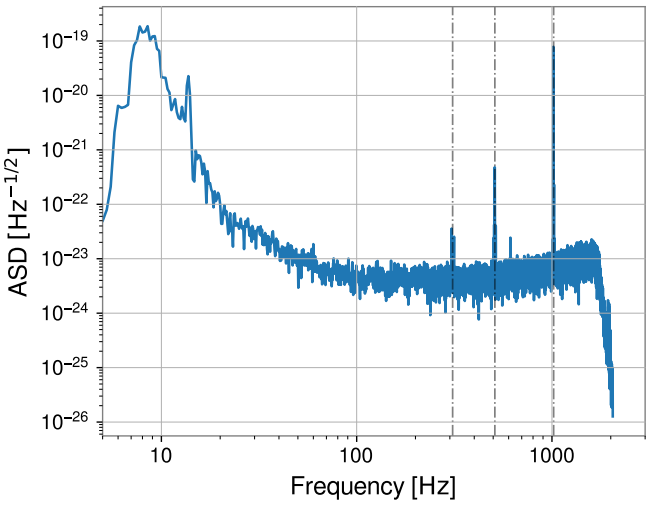


Figure 5: ASD plot for a Tomte Glitch.

Figure 5 shows an example of the ASD plot of a Tomte glitch event. To obtain the root-mean-square strain noise at a given frequency band, we integrate over the squares of the ASD readings over the frequency band of interest and take its square root. Observing the strong spectral lines at around the 300, 512, and 1024 Hz ranges, marked by dotted lines, there is a big chance that these are of instrumental origin. The problem with this plot, however, is that it does not capture the glitch signal well because glitches are relatively weak, highly transient and easily overpowered by the instrument noise. To better visualize the glitch, we would need to **whiten** the data and view its time-amplitude plot.

2.3.2 Whitening

Whitening is the process of supressing low-frequency and spectral line noise from the data to allow for a clearer view of the underlying signals at sensitive frequency ranges. Whitening is one of the first steps in astrophysical data analysis, especially when the noise is highly non-stationary and varies significantly over time. By whitening the data, we effectively reduce the impact of noise on our analysis, improving the sensitivity of our detection methods.

Whitening is done by taking the inverse of the ASD and multiplying it with the original time series data.

$$\tilde{n}_{\text{whitened}}(f) = \frac{\tilde{n}(f)}{ASD} = \frac{\tilde{n}(f)}{\sqrt{S_n(f)}} \tag{7}$$

$$n_{\text{whitened}}(t) = \mathcal{F}^{-1}\left\{\tilde{n}_{\text{whitened}}(f)\right\} \tag{8}$$

This effectively flattens the power across all frequencies, allowing for a clearer view of the glitch signal. The resulting whitened data can then be used for further analysis, such as statistical testing or machine learning classification.

In figure 6 we see a Tomte glitch signal before and after whitening. In the first plot we do not clearly see the effects of the glitch on the time series data. However, after whitening the data, we can clearly see the glitch signal in the time-amplitude plot. The Q-transform allows a view of the non-stationary portion of the glitch signal, which shows a clear peak in the normalized energy coinciding with the glitch event. For a clean signal we would not see any significant peaks in the whitened time series or non-stationary parts in the Q-transform.

With this, the received signal can now be used for our statistical testing.

2.3.3 Scaling

Some of our statistical tests, particularly the Kolmogorov-Smirnov test, is highly affected by the values of the data. We need to ensure that the comparisons being done in these tests are not biased due to the absolute values of the data points. Hence, this requires for the data to be scaled down to the same range of a standard normal distribution. This is achieved by scaling the data to have a mean of 0 and a standard deviation of 1. This is also known as **standardization** or **z-score normalization**.

To achieve this the `StandardScaler` class from the `sklearn.preprocessing` module is used. This module works by transposing the data to have a zero mean, calculates the standard deviation, and scales the data points down such that the standard deviation is 1. The **standard score** or

z-score of a sample point, which tells us how many standard deviations it is from the mean, is given by

$$z = \frac{x - \mu}{\sigma}, \tag{9}$$

where x is the original data point, μ is the mean, and σ is the standard deviation of the data. The z value obtained can be treated as the new data point after scaling and is particularly useful in this case as it does not change the position of the data point in its calculation.

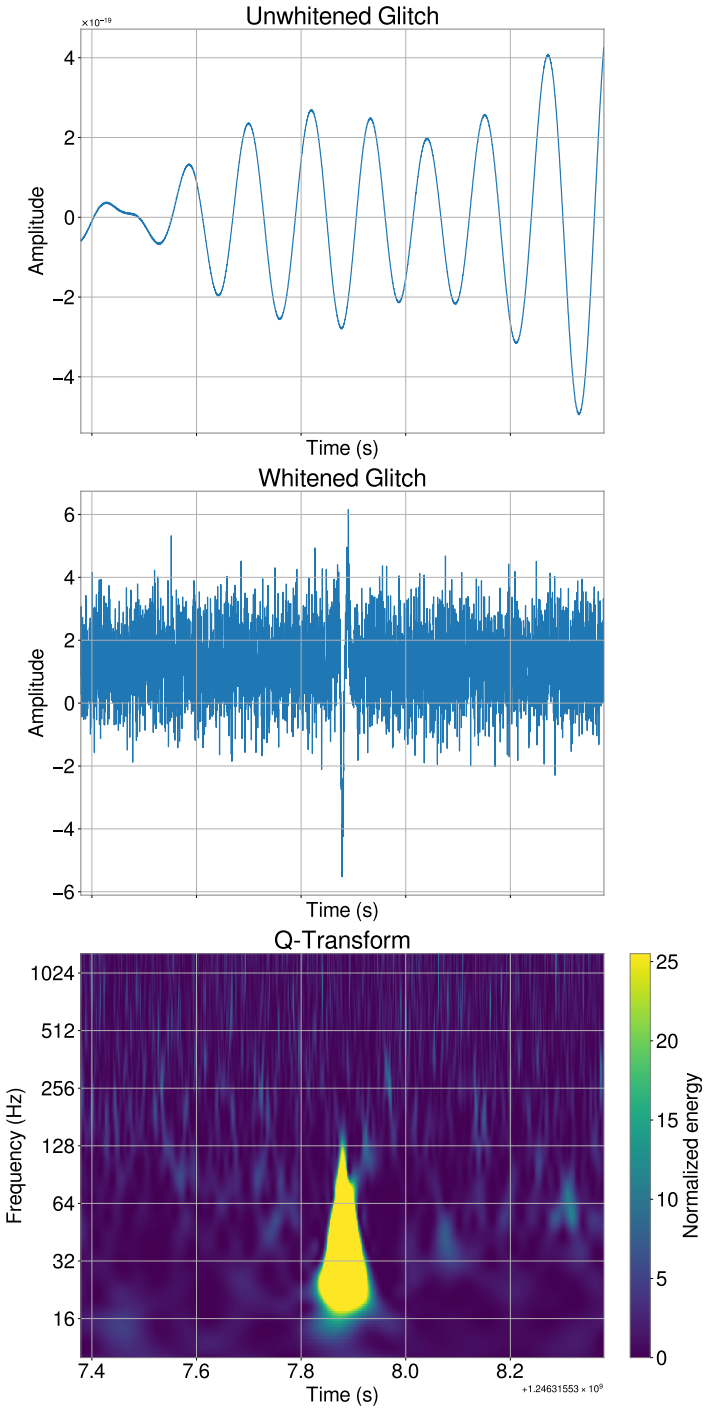


Figure 6: Example of a Tomte glitch signal before and after whitening.

3 Methods

Our main objective is to determine the Gaussianity of time series data obtained from the LIGO Livingston detector and use this as a criterion to determine the presence of a glitch in it. If a sample is found to be non-Gaussian, there is a high chance of a glitch being present. From the previous section, we have seen how whitening the time series data brings out the glitch characteristics when plotted. Taking the same glitch example as before, if we were to treat this whitened data as a population of points, ignoring the temporal component, we notice that visually, they do not follow a Gaussian distribution. Similarly, when observing a sample of clean data, the signal is closer to that of a Gaussian distribution than the glitched signal. However, to quantify this, we need to apply certain statistical tests to the data.

To achieve this, we employ three statistical tests: the Shapiro-Wilk test, the Kolmogorov-Smirnov test, and the Anderson-Darling test. Each of these tests have their own strengths and weaknesses, and are used to assess the normality of the data in different ways. Each of these tests give us a statistic along with a p-value, the latter of which will be used to determine the significance of the result.

3.1 The Shapiro-Wilk Test

The Shapiro-Wilk test is a parametric test, meaning that it assumes the data follows a specific distribution. The null hypothesis in this case is that the sample data is normally distributed. The original research paper [27] does a good job of explaining how this test works. The idea is to calculate a test statistic W based on the ratio of the best estimate of the sample data's variance to its actual variance. This was the first test that could detect departures from normality due to either the skewness or the kurtosis of the data.

If we were to consider $x_1, x_2, \ldots, x_n$, to be a collection of ordered sample points from a population, ranked smallest to largest, with the sample mean $\bar{x}$ and $a_1, a_2, \ldots, a_n$ being constants computed from the expected values of the order statistics of a normal distribution, then the test statistic W is given by:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{10}$$

Taking $m = (m_1, m_2, \ldots, m_n)$ to be the expected values of the order statistics of a normal distribution such that $m_i = \mathbb{E}[z_i]$ where z_i is the i-th order statistic of the normal distribution, and V the covariance matrix of the ordered statistics, the constants a_i, which make up the **optimal weight vector** $a \in \mathbb{R}^n$, can be computed as

$$a_i = \frac{m^T V^{-1}}{||V^{-1} m||} \tag{11}$$

$$= \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}} \tag{12}$$

The test statistic W lies in the range $[0, 1]$, with values closer to 1 indicating that the sample data is more likely to be normally distributed. The null hypothesis H_0 is rejected if W is significantly less than 1, indicating that the sample data is not normally distributed.

The p-value of the statistic is usually used to determine the significance of the result. If the p-value is less than the significance level defined, the null hypothesis is rejected, indicating that the sample data is non-Gaussian and likely contains a glitch.

The Shapiro-Wilk test, though powerful, has some limitations. It is sensitive to the sample size and may not perform well with small or large samples. It works ideally for sample sizes $n \leq 2000$. The `shapiro()` function from `scipy.stats` works well for $n \leq 5000$ after which the p-value calculations' accuracy drops. Additionally, for this test to work the data must be continuous and univariate.

3.2 The Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) test is a non-parametric test, originally proposed in the 1930s, that is used to decide whether a sample comes from a particular type of distribution [28, 29]. compare the empirical distribution function of a sample with another reference probability distribution . This test has two versions: the **one-sample KS test** which is used to compare an empirical sample distribution with a reference theoretical distribution, and the **two-sample KS test** which compares the empirical distributions of two samples. We use the one-sample KS test to compare our detector data against a standard normal distribution.

The idea of this test is to check the difference in shape between the distributions being studied, which is done by measuring the maximum vertical distance between the **empirical cumulative distribution function (EDF)** of the sample and the **cumulative distribution function (CDF)** of the reference normal distribution. Taking a sample of ordered points $X_1, X_2, \ldots, X_N$ as N ordered points from a population, the ECDF of the reference function, a normal distribution in this case, is given by

$$F_N = \frac{n(i)}{N} \tag{13}$$

where $n(i)$ is the number of points less than or equal to X_i in the sample. This is in the form of a step function increasing at steps of $1/N$ at each ordered point [30]. Taking the CDF of the theoretical CDF of the distribution being tested as F, and the Empirical the KS test statistic D_n is defined as

$$D_n = \sup_{1 \leq i \leq N} \left(F(X_i) - \frac{i-1}{N}, \frac{i}{N} - F(X_i) \right) \tag{14}$$

If D_n is significantly large, it indicates that the ECDF of the sample data significantly differs from the CDF of the normal distribution. A critical value is used to determine the significance of this result. The null hypothesis is rejected if D_n is greater than the critical value, indicating that the sample data is not normally distributed.

3.3 The Anderson Darling test

The Anderson-Darling (AD) test is a goodness-of-fit test used to determine whether a sample of data is drawn from a specific distribution, most commonly a normal distribution. It is a modification of the Kolmogorov-Smirnov and Cramer-von Mises tests with more sensitivity to differences in the tails of the distribution [30, 31].

Similar to the Kolmogorov-Smirnov test, this test compares the ECDF of the sample data with the CDF of a theoretical reference distribution by taking a distance function between them. The main difference here is that the full range of the data is used to calculate the distance function rather than just the largest distance value [31]. Here, taking n to be the number of elements in the sample and $w(x)$ to be the weighting function the squared area between the sample stepped CDF F_n and the reference distribution F is calculated as

$$n \int_{-\infty}^{\infty} \left(F_n(x) - F(x)\right)^2 w(x) dF(x) \tag{15}$$

The weighting function $w(x)$ is defined as

$$w(x) = \frac{1}{F(x)(1 - F(x))} \tag{16}$$

Taking the above into consideration, the distance A^2 [32] is given by

$$A^2 = n \int_{-\infty}^{\infty} \frac{\left(F_n(x) - F(x)\right)^2}{F(x)(1 - F(x))} \, dF(x) \tag{17}$$

To obtain random samples from a distribution function, the CDF is to be calculated and random samples x are to be taken from a uniform distribution $x \in [0, 1]$. Hence, the above equation can be rewritten as

$$A^2 = -n - S \tag{18}$$

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left(\frac{2i - 1}{N}\right) \left(\ln F(X_i) + \ln(1 - F(X_{N-i+1}))\right), \tag{19}$$

which is effectively a weighted cross-product of the samples, allowing for a comparison of the sample data with the reference distribution.

In the case of the mean and standard deviation of the sample distribution being known with the distribution size, n being greater than 5, the significance levels are 1%, 2.5%, 5%, 10% and 15%, with the critical values for each significant level being pre-computed through Monte-Carlo simulations. The null hypothesis is rejected if the calculated statistic is greater than the critical value at a selected significance level, indicating that the sample data is not normally distributed.

For all the tests used here, the p-value obtained along with the calculated statistics are used to determine Gaussianity of our sample data as opposed to the statistics themselves. This allows us to better understand the results, as they provide a direct measure of their significance. For example, if we took the Kolmogorov-Smirnov test, the calculated statistic is a distance measure, which can be difficult to interpret in terms of the distribution being Gaussian or not. The p-value, on the other hand, gives us a direct measure of the probability of the test statistic being as extreme as calculated hence, eliminating the need to determine a critical value for each statistic. Additionally, setting different critical values for each test would lead to vastly different results, making it difficult to compare the results. The p-values standardize the results hence allowing for like-to-like comparison of the result.

If the p-value is less than a predetermined significance level, 0.05 in this case, we reject the null hypothesis and conclude that the sample data is not normally distributed, hence containing a glitch. Appendix 8.1 discusses p-values and the discourse and reasoning behind the selection of the 0.05 threshold.

Along with studying how effective these tests are at determining the Gaussianity of GW detector data, we will also be validating the claim in [33], which states that the Shapiro-Wilks test is the most powerful, followed by the Anderson darling test and the Kolmogorov-Smirnov test.

4 Experimentation

The implementation of the proposed statistical tests on the whitened detector TimeSeries is fairly straightforward, with the `scipy.stats` package providing dedicated functions for each. Treating each point in the TimeSeries as a sample of a distribution, we can use the corresponding functions to obtain the relevant statistics and p-values.

4.1 Using the complete frequency range

5 Experimentation with a band pass filter applied

6 Conclusions

7 Future Scope

8 Appendix

8.1 P-values and Selecting a Threshold

8.2 Selecting a Suitable Scaling Method

8.3 Experimentation with Time Series Anomaly Detection

References

[1] B. P. Abbott, R. Abbott, et al. "Observation of Gravitational Waves from a Binary Black Hole Merger". In: *Physical Review Letters* 116.6 (Feb. 2016). ISSN: 1079-7114. DOI: 10.1103/physrevlett.116.061102. URL: http://dx.doi.org/10.1103/PhysRevLett.116.061102.

[2] Timothée Accadia, Fausto Acernese, et al. "Virgo: A laser interferometer to detect gravitational waves". In: *Journal of Instrumentation* 7 (Mar. 2012), P03012. DOI: 10.1088/1748-0221/7/03/P03012.

[3] Andrew Grant. "Advanced LIGO ramps up, with slight improvements". en. In: *Physics Today* 2016.11 (Nov. 2016), p. 12128. ISSN: 19450699. DOI: 10.1063/pt.5.9074. URL: https://pubs.aip.org/physicstoday/online/12128.

[4] The LIGO Scientific Collaboration, the Virgo Collaboration, et al. "GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs". In: *Physical Review X* 9.3 (Sept. 2019). arXiv:1811.12907 [astro-ph], p. 031040. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.9.031040. URL: http://arxiv.org/abs/1811.12907.

[5] R. Abbott et al. *GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run*. arXiv:2010.14527 [gr-qc]. Mar. 2021. DOI: 10.1103/PhysRevX.11.021053. URL: http://arxiv.org/abs/2010.14527.

[6] The LIGO Scientific Collaboration, the Virgo Collaboration, et al. *GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run*. arXiv:2111.03606 [gr-qc]. Oct. 2023. DOI: 10.1103/PhysRevX.13.041039. URL: http://arxiv.org/abs/2111.03606.

[7] The LIGO Scientific Collaboration, the Virgo Collaboration, et al. "A guide to LIGO-Virgo detector noise and extraction of transient gravitational-wave signals". In: *Classical and Quantum Gravity* 37.5 (Mar. 2020). arXiv:1908.11170 [gr-qc], p. 055002. ISSN: 0264-9381, 1361-6382. DOI: 10.1088/1361-6382/ab685e. URL: http://arxiv.org/abs/1908.11170.

[8] The LIGO Scientific Collaboration and the Virgo Collaboration. "Characterization of transient noise in Advanced LIGO relevant to gravitational wave signal GW150914". In: *Classical and Quantum Gravity* 33.13 (July 2016). arXiv:1602.03844 [gr-qc], p. 134001. ISSN: 0264-9381, 1361-6382. DOI: 10.1088/0264-9381/33/13/134001. URL: http://arxiv.org/abs/1602.03844.

[9] The LIGO Scientific Collaboration and The Virgo Collaboration. "GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral". In: *Physical Review Letters* 119.16 (Oct. 2017). arXiv:1710.05832 [gr-qc], p. 161101. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.119.161101. URL: http://arxiv.org/abs/1710.05832.

[10] Florent Robinet et al. "Omicron: A tool to characterize transient noise in gravitational-wave detectors". In: *SoftwareX* 12 (2020), p. 100620. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2020.100620. URL: https://www.sciencedirect.com/science/article/pii/S2352711020303332.

[11] Duncan M. Macleod et al. "GWpy: A Python package for gravitational-wave astrophysics". In: *SoftwareX* 13 (2021), p. 100657. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2021.100657. URL: https://www.sciencedirect.com/science/article/pii/S2352711021000029.

[12] D. Davis et al. "Subtracting glitches from gravitational-wave detector data during the third observing run". In: *Classical and Quantum Gravity* 39.24 (Dec. 2022). arXiv:2207.03429 [astro-ph], p. 245013. ISSN: 0264-9381, 1361-6382. DOI: 10.1088/1361-6382/aca238. URL: http://arxiv.org/abs/2207.03429.

[13] S. Chatterji et al. "Multiresolution techniques for the detection of gravitational-wave bursts". In: *Classical and Quantum Gravity* 21.20 (Oct. 2004). arXiv:gr-qc/0412119, S1809–S1818. ISSN: 0264-9381, 1361-6382. DOI: 10.1088/0264-9381/21/20/024. URL: http://arxiv.org/abs/gr-qc/0412119.

[14] Leah Vazsonyi and Derek Davis. "Identifying glitches near gravitational-wave signals from compact binary coalescences using the Q-transform". In: *Classical and Quantum Gravity* 40.3 (Feb. 2023). arXiv:2208.12338 [astro-ph], p. 035008. ISSN: 0264-9381, 1361-6382. DOI: 10.1088/1361-6382/acafd2. URL: http://arxiv.org/abs/2208.12338.

[15] Stannered. *Simple Michelson interferometer diagram*. 2007. URL: https://commons.wikimedia.org/wiki/File:Interferometer.svg.

[16] WiredSense. *The Michelson Interferometer - A Laser Lab Alignment Guide*. WiredSense Tutorials. URL: https://www.wiredsense.com/tutorial/the-michelson-interferometer-a-laser-lab-alignment-guide.

[17] R. Weiss et al. *Quarterly Progress Report No. 105*. Apr. 1972. URL: https://dspace.mit.edu/handle/1721.1/56271.

[18] The LIGO Scientific Collaboration. *What is an Interferometer?* URL: https://www.ligo.caltech.edu/page/what-is-interferometer.

[19] Sudarshan Ghonge et al. *Assessing and Mitigating the Impact of Glitches on Gravitational-Wave Parameter Estimation: a Model Agnostic Approach*. arXiv:2311.09159 [gr-qc]. Oct. 2024. DOI: 10.48550/arXiv.2311.09159. URL: http://arxiv.org/abs/2311.09159.

[20] D. M. Macleod et al. "GWpy: A Python package for gravitational-wave astrophysics". In: *SoftwareX* 13 (2021), p. 100657. ISSN: 2352-7110. DOI: 10.1016/j.softx.2021.100657. URL: https://www.sciencedirect.com/science/article/pii/S2352711021000029.

[21] M Zevin et al. "Gravity Spy: integrating advanced LIGO detector characterization, machine learning, and citizen science". In: *Classical and Quantum Gravity* 34.6 (Feb. 2017), p. 064003. ISSN: 1361-6382. DOI: 10.1088/1361-6382/aa5cea. URL: http://dx.doi.org/10.1088/1361-6382/aa5cea.

[22] Curt Cutler and Eanna Flanagan. "Gravitational Waves from Mergin Compact Binaries: How Accurately Can One Extract the Binary's Parameters from the Inspiral Waveform?" In: *Physical Review D* 49.6 (Mar. 1994). arXiv:gr-qc/9402014, pp. 2658–2697. ISSN: 0556-2821. DOI: 10.1103/PhysRevD.49.2658. URL: http://arxiv.org/abs/gr-qc/9402014.

[23] Christopher J. Moore, Robert H. Cole, and Christopher P. L. Berry. "Gravitational-wave sensitivity curves". In: *Classical and Quantum Gravity* 32.1 (Jan. 2015). arXiv:1408.0740 [gr-qc], p. 015014. ISSN: 0264-9381, 1361-6382. DOI: 10.1088/0264-9381/32/1/015014. URL: http://arxiv.org/abs/1408.0740.

[24] M. Lopez. "Exploring the Frontier of Transient Gravitational Wave Detection - Unleashing the Power of Machine Learning". PhD thesis. Utrecht University, 2025.

[25] Christopher Chatfield. *The Analysis of Time Series—An Introduction*. 4th ed. London: Chapman and Hall, 1989, pp. 94–95. ISBN: 0-412-31820-2.

[26] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. 3rd. USA: Prentice Hall Press, 2009. ISBN: 0131988425.

[27] S. S. Shapiro and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)". In: *Biometrika* 52.3/4 (1965), pp. 591–611. ISSN: 00063444, 14643510. URL: http://www.jstor.org/stable/2333709.

[28] Frank J. Massey. "The Kolmogorov-Smirnov Test for Goodness of Fit". In: *Journal of the American Statistical Association* 46.253 (1951), pp. 68–78. ISSN: 01621459, 1537274X. URL: http://www.jstor.org/stable/2280095.

[29] I. M. Chakravarti, R. G. Laha, and J. Roy. *Handbook of Methods of Applied Statistics, Volume I*. John Wiley and Sons, 1967, pp. 392–394.

[30] William F. Guthrie. *NIST/SEMATECH e-Handbook of Statistical Methods (NIST Handbook 151)*. en. 2020. DOI: 10.18434/M32189. URL: https://www.itl.nist.gov/div898/handbook/.

[31] Michael J De Smith. "StatsRef.com — The sourcebook for statistics". In: (2025). URL: https://www.statsref.com/.

[32] T. W. Anderson and D. A. Darling. "A Test of Goodness of Fit". In: *Journal of the American Statistical Association* 49.268 (1954), pp. 765–769. DOI: 10.2307/2281537.

[33] Nornadiah Razali and Yap Bee Wah. "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests". In: *Journal of Statistical Modeling and Analytics* 2.1 (2011), pp. 21–33.