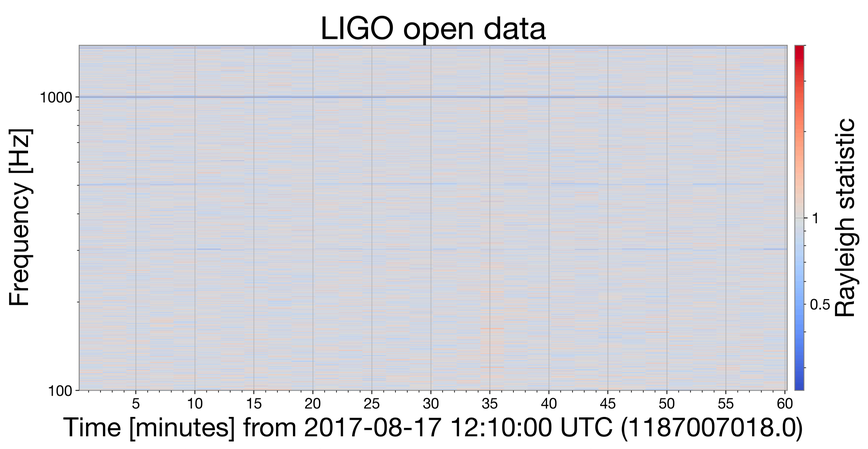
Improvement of the method for Gaussianity evaluation for observation data of a GW-detector

Shunsei Yamamura, Hirotaka Yuzurihara, Takahiro Yamamoto, Takashi Uchiyama *Phys. S. U. T , ICRR* yamamura@icrr.u-tokyo.ac.jp

Motivation

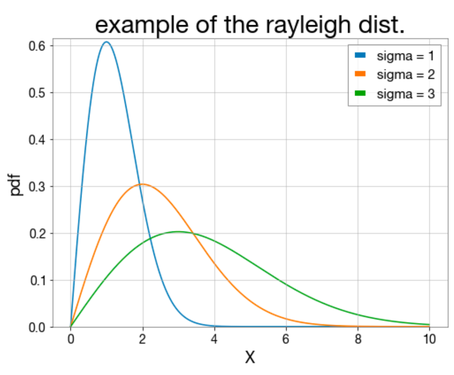
Observation data of GW detector is dominated by a lot of noise. "Gaussianity" is a nature of data that represents how close the data is to Gaussian noise. So, evaluating it will help to select which data to analyze.
The current tool for testing Gaussianity cannot quantitatively evaluate it. Therefore, our goal is to make a tool that can do that. In order to realize the goal, we have developed a method using the p-value of a statistical test.



The left figure is a time vs. frequency map which is calculated by rayleigh_spectrogram (the current tool). The used data is the open data at LIGO.

The red and blue color means it is apart from gaussian noise.

Rayleigh distribution



Definition

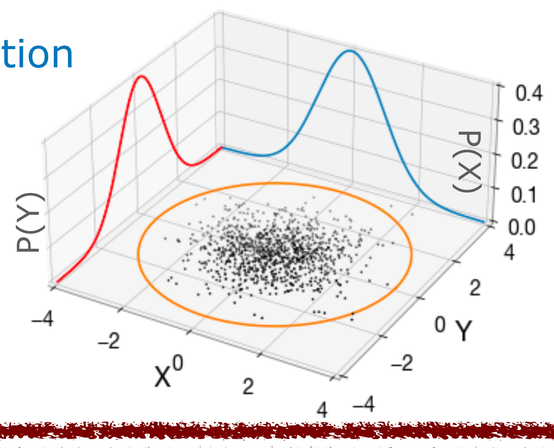
The probability density function of the Rayleigh distribution is

$$f(r;\sigma) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} \quad , \quad r \geq 0$$

where σ is the parameter.

Relation to the normal distribution

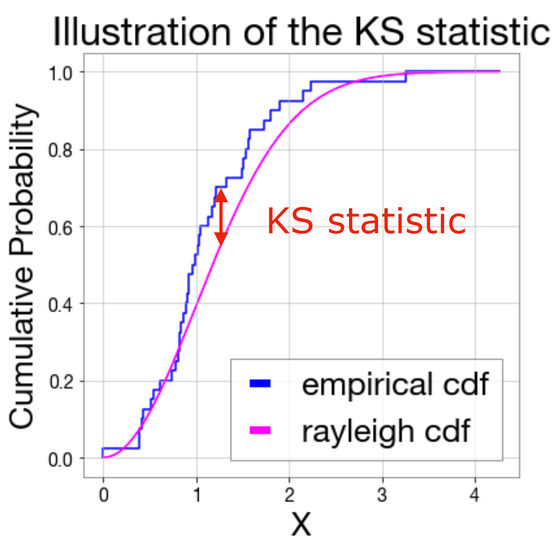
When X and Y follow the normal distribution whose mean is 0 and variance is σ^2 independently, $R = \sqrt{X^2 + Y^2}$ follows the Rayleigh distribution.
R geometrically represents the distance from the origin.



Kolmogorov Smirnov test (KS test)

The KS test is one of the statistical tests. It can compare an empirical distribution with a reference distribution.

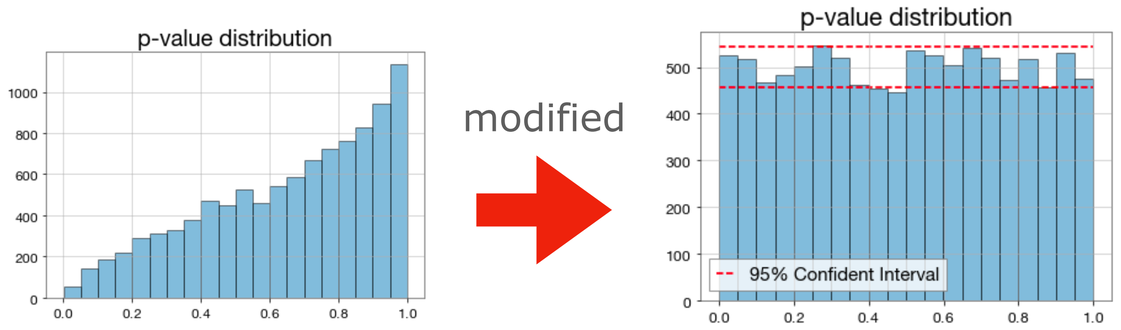
KS statistic : the longest distance between the empirical cumulative distribution function (CDF) of the sample and the CDF of the reference distribution.
This quantity is used to calculate a p-value.



modified KS test

When the parameter of the reference distribution is estimated by the sample, the KS test does not work properly and the distribution of the p-value is biased(below left figure).

→ We modified the test by simulating it 10,000 times and estimating the background distribution of the KS statistic. The p-value is calculated by the estimated distribution.



Method

1. Assume all data is Gaussian noise.(null hypothesis)
2. Divide the data into several segments.
3. Furthermore, divide each segment into several pieces.
4. Calculate amplitude spectrum density(ASD) with each piece. ASD should follow the Rayleigh distribution under the assumption.
5. Estimate the parameter by sample, and perform a modified KS test. Then, a p-value is obtained at each frequency.
6. Repeat steps 4 and 5 with each segment.

Why?

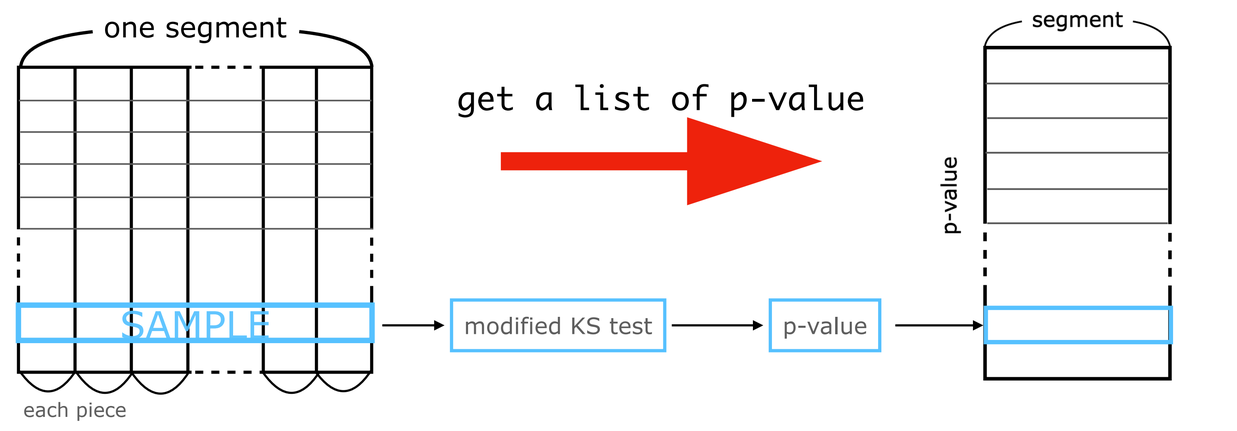
→ Let a_m be a coefficient of the Fourier transform and $a_m = x + iy$.

1. x and y follow the normal distribution independently.
2. ASD $\propto |a_m| = \sqrt{x^2 + y^2}$

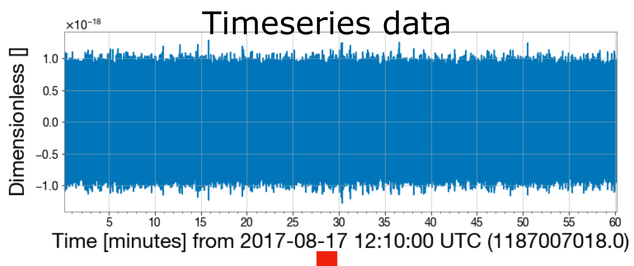
definition of Fourier transform in numpy

$$A_k = \sum_{m=0}^{n-1} a_m \exp\left\{-2\pi i \frac{mk}{n}\right\}$$

From these two things, ASD follows the Rayleigh distribution.

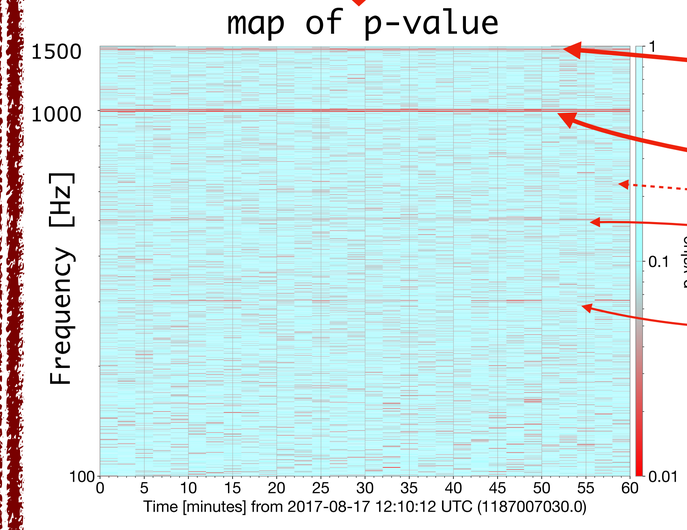


Results



our tool
input : Timeseries
output: p-value at each time and frequency

map of p-value



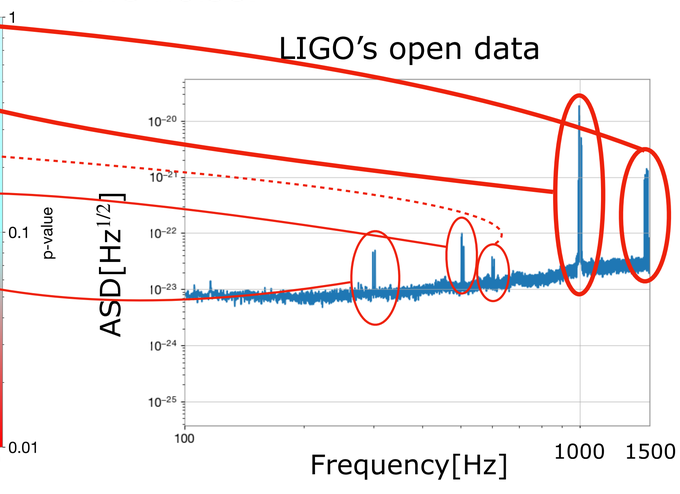
● map of p-value

For demonstration, we applied the tool to one hour of LIGO's open data (same as the figure in the Motivation part).

BLUE
→ cannot reject the null hypothesis
RED
→ reject the null hypothesis
(not Gaussian noise!)

✦ We cannot reject the hypothesis for the most parts.

✦ Several red lines can be seen. These correspond to the large line noise.



Line noise can be detected by this tool because the frequency component does not follow the Rayleigh distribution.

Summary and Future work

- We are developing a tool to check "Gaussianity" of data.
- The method is newly proposed.

- We want to make this a GUI tool so that people can analyze the data easily.
- Furthermore, we have to verify whether the results are valid.
  - ✦ for example, apply to ideal data of Gaussian noise, or data including the transient noise.