Fish for a needle in Galaxy: Alogrithm for Transient Signal Detection via Simulation for GNOME

Yuzhe Zhang*

August 6, 2019

* *Undergraduate Student, Department of Modern Physics, University of Science and Technology of China(USTC), Hefei, China*

Preface

Fish for a needle in Galaxy

'Fish for a needle in the ocean' is a Chinese proverb, expressing the difficulty in looking for tininess in a grant area. For us, the 'needle' is exotic physic, and the 'ocean' is the universe. However, other than 'universe', I prefer to use 'galaxy' in the title, because 'galaxy' means 'river of sliver' in Chinese [1]. So now we are fishing for the signal in this sliver river. I wish this romance could be a relief for us during the long waiting before we were truly stabbed by the detection of physic in the dark.

In this report, I will first share my understanding of data simulation and analysis. Then comes a thorough description of algorithm applied in each section of the work. The report will also include the introduction of realization of the algorithm in Python. Due to the limited period of my work, there would be much to be modified or improved, which will also be explained in details in the report. I hope it could be a framework for the next step.

My e-mail address is zyzoli@mail.ustc.edu.cn. Please contact if any problem or idea with this report. I will keep on updating this report according to feedback. Updates will be recorded in update log in appendix near the end of this report. You can request the up-to-date report or any data, figures, codes shown in the report from me.

Yu
Summer, 2019
in Krakow

[1] This romantic name came from imagination inspired by milky way, the hazy band of light in starry nights.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Intro to Simulation

1.1.1 What is Simulation

Simulation means imitation[1, wiki]. Here in this report, data simulation is defined as generating data based on the characteristics of a model. Maybe it's appropriate to use 'modeling and simulation'. But from my point of view, 'simulation' alone has included the procedure of modeling. Let's waste no time in vocabular game and move on. Imagine that the existence of Laplace's demon is authentic. In traditional physic experiments, we are trying every best to figure out how this demon calculate the movement of the world. While in simulation, it's us who make the rules. It seems ridicules that these once rule seekers (us) try to become rule makers. This ridicules misunderstanding is definitely wrong, but actually helps in pushing forward our work. So here comes the question:

1.1.2 Why do simulation

This is not only a question to be answered, but also a question to be questioned. My first response to this question is, why not do simulation? We used to examine our theory in the laboratory or on paper. Simulation could be the third kind, especially beneficial when we could not push forward lab and paper work.
If Galileo and Newton were given a PC with MATLAB or Mathematica, physic might be mainly developed in virtual lab and discrete math. (Just joking. I think they would first dismantling the PC and then become engineers and develop computer science.)
I found three major problems hindering us from the detection of real signal. One is huge background noise. It's easy to understand the first problem: previous physicists discovered signals above the noise, then the signals beneath the noise are left to us. The second problem is nonlinear relationships in processing. We frequently come into nonlinear transformation or nonlinear equations, which make perfect theoretical derivation impossible. About this point, I would explain it in details later. Please forget this, since it's quite ambiguous for now. The third barrier is little knowledge about exotic physic, or dark matter and dark energy. We have so little idea about its form of interaction that we have to make many assumptions and then examine these.
Hence, my second response to 'why do simulation' is that, we are forced to do it. We have to rely on simulating the real world to test our methods, correct our predictions or assumptions, and finally, have some expectations in mind.
What's more, I hope such artificial detections help us adapt to the joy of discovery gradually, preventing us from being over-exciting when real signal is detected.

1.1.3 Game of Probability

We know two games of probability: gambling and quantum mechanism. In data processing, we are not talking about quantum mechanism (neither gambling). We would focus on the science closely related to gambling: probability theory and statistics.
For signal detection, I think it's reasonable for us to reach on this: the results should always come with confidence interval, like 'we are 99% sure that, we find a signal within a certain time range and within a certain energy range', or 'false alarm rate estimated to be less than 1 event per 203 000 years, equivalent to a

significance greater than 5.1σ'[Observation of Gravitational Waves from a Binary Black Hole Merger]. In other words, what we are searching for should be a probability density distribution, which indicates the confidence that we find the signal in any interval.
As to how-to-do, I recommend Monte Carlo method. We repeat the experiment in simulation, with known input paraments. The result processed with statistics method would show us the probability density distribution.
Take excess power analysis as an example. How do we know the probability that a signal is buried beneath somewhere under the noise? We generate random noise of which characteristics match real noise's. Then we insert a signal with certain amplitude, FWHM or any other factors. We repeat this simulated experiment, and finally discover how the result is like.
However, Monte Carlo method is blamed for its low convergence speed. Anyone should be careful with this before simulation.

1.1.4 What is 'zero result' indeed

Let's think about one more question. You may have heard about this, 'zero result is also a result'. When I heard about this first in high school, I thought this is nothing more than a relief for people like us. Does 'zero result' mean zero or blank? Of course NOT. By announcing 'zero result', the scientists are actually answering such a question: at which level of confidence, it's impossible to detect signal with some certain characteristic. 'Zero result' not only sets the upper limit for detecting, but also implies the orientation of the next-step research.

1.1.5 What's a Good Simulation for GNOME

to be written

1.2 For Developer

This section is written for program developers.
The algorithms are all realized in Python 3.7. When it comes to large amount of calculations, I will display the running time of the program. My laptop is HP ENVY x360 Convertible 13-ag0xx, coming in AMD Ryzen 5 2500U with Radeon Vega Mobile Gfx 2.00GHz, 8GB RAM, 256GB SSD, Windows 10 Version 1903. All tests are run in Best Performance Mode when charging. My Python IDE is PyCharm, setting heap size to 890 MB.
I always trying to avoid using some uncommon Python packages for three reasons. First of all, I tried every method to install some LIGO packages but failed anyway. I am afraid that this could happen on others' PC. The next reason is that some packages requires Python 2.7, which would not be maintained soon. The last reason is that I found no package catering to GNOME data processing. It's better for comprehension to write the package on our own. However, LIGO packages are still important reference for us. I am jealous that they have so many experts in data processing.
As a student major in physics, I am far from professional in programing, not to mention that it has been just 2 months since I started with Python. I am deeply dependent on my little experience with C, C++, MATLAB and Mathematica. Codes are written in the most simple way without Python skill. If you are a skillful programmer or even professional, you can skip this section. If you are a non-professional like me, please read the following tips.
Always turn to common or well-known packages for help
When it comes to calculation, try to find a wheel in Python packages, instead of building the wheel on your own. Recommendation: NumPy, Matplotlib, Pandas. **Be careful with program running time**
When your program consumes extremely long time(1 minute is extreme for a single section in GNOME data processing), program would probably end up in failure. Don't blame your PC first. Try to use functions from common packages. Try to use vectorization or parallel operations instead of loop structure.
10^8 threshold
For my laptop and Python IDE, 10^8 is a useful standard. For any amount of calculations or data, things are completely different in 10^7 and 10^8. When it reaches 10^8, it's usual that we go into memory error and heavy cooling burden.
Python IDE
If you have just started with Python, I recommend PyCharm for your Python IDE choice. Its Community Edition(Free) is adequate for us. It's comfortable and efficient to develop codes in a good IDE.

Chapter 2

A Brief Introduction to Discrete Fourier Transfrom (DFT)

more to be written here

Chapter 3

Signal Generating

more to be written here

Chapter 4

Excess Power Analysis

more to be written here

Chapter 5

Single-station Event Identification

5.1 Intro

Finally we are at detection stage. To distinguish signal detection in single- and multi- station situation, **Identification** is deployed here to emphasize that we are identifying whether a signal is authentic under a certain threshold. While in multi-station detection, **Locating and Examination** is used to fully describe two unique steps in detection. Single-station signal identification constructs the basis for locating and examining signals in multi-station network.
Let's focus on a single station first.
The problem at this stage is simple: We know 'power' distribution over time and frequency from excess power analysis. How to pick up 'exotic signals'?

5.2 Algorithm

5.2.1 Brief Desciption

We will deal with this in statistical way. The key point is to know the probability density distribution over 'power', and therefore choose a threshold for 'Excess' power to pick up 'exotic' signals.

5.2.2 Detailed Description

1. Choose a Certain Set of Parameter for Excess Power analysis
   Let's review on the input parameters in data processing.

   segment length

   segment stride

   welch method number per segment

   average method: default to 'Exponential Moving Average (**EMA**)'

   EMA factor

   EMA window length Here we introduce two more parameters.

   frequency low-pass

   frequency high-pass
   In the last chapter, we kept all frequency bands on spectrum. However, real signals often respond in a narrow frequency band. We can focus on only the sensitive band of the excess power analysis result. Figure 5.1 shows different response in frequency domain after being processed by welch method.

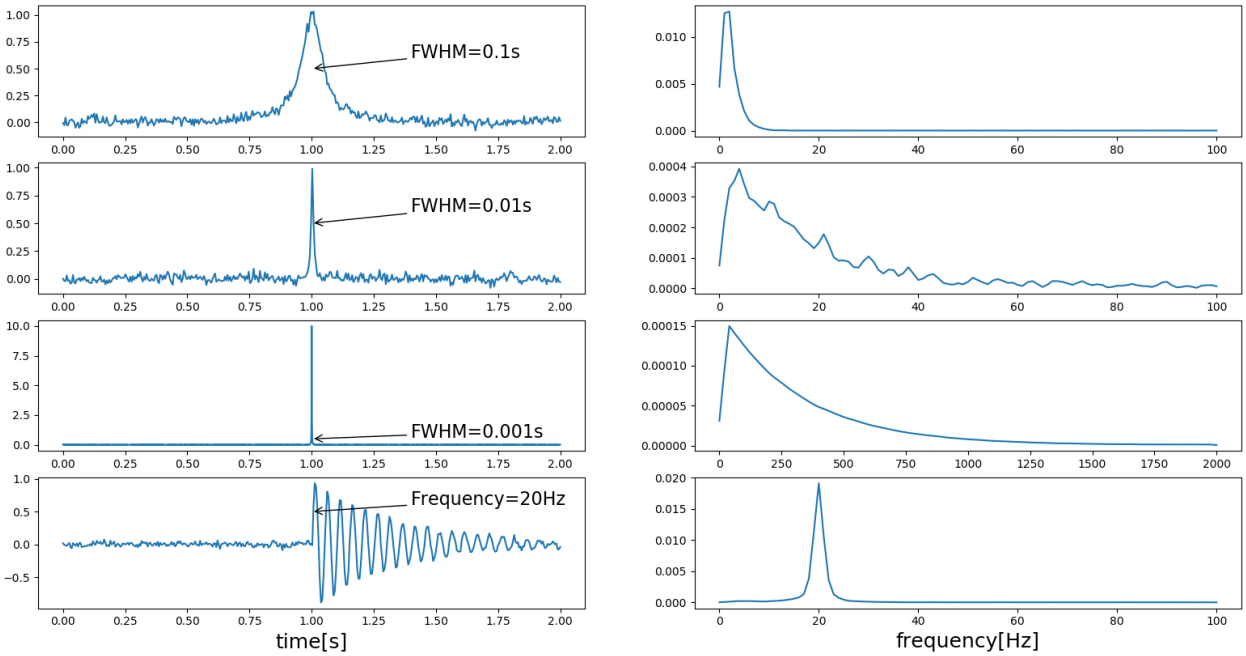


Figure 5.1: Time domain signals (left) and its welch method's result (right)

People always desire to find the 'optimal' parameters for analysis. From my pespective, 'optimal' parameters make real signal significant. For example, if you want to search for a Lorentzian signal, of which FWHM is 0.1s, Figure 5.1 tells you that you'd better focus on frequency from 0 Hz to 5Hz, where its response in frequency domain is at its maximun.
There should be a set of 'optimal' parameters for analysis under noise with certain characteristics, and signal response to 'exotic' physic in the specific sensor. And we can find it out via simulation . But generally speaking, there's no conclusion which can figure out a general rule for choosing 'optimal' parameters.
Actually the choice of parameters is influenced by various factors, including the method in Multi-station Locating and Examination. Further details on 'optimal' parameters will be addressed in Chapter 6. (However, I havn't finish writing this part, so you can't find it in Chapter 6 either.)
Let's assume that we have known the 'optimal' parameters, and move on to the next step.

2. Creating Background Reference
   As long as we know the characteristics of the noise, we can simulate it and apply excess power analysis. After statistic methods, we will get the probability density distribution over 'power'. This distribution would be the reference to set threshold for 'exotic'.

5.2.3 Reconstruct Original Signals from Result

to be completed

Chapter 6

Multi-station Event Locating and Examination

6.1 Intro

Thanks to the construction of global network, we have the confidence in discovering or denying exotic physics. LIGO earned a Nobel Prize with only two stations. The error rate decreases exponentially with the number of stations, so I think we can expect three or even more prizes coming soon.

6.2 Assumptions

Let's take Earth as the reference system.
Below are listed some commonly used assumptions. Of course we can revise or even abandon these assumptions when necessary.

Assumption 1.a (Plane Wave Like)**.** *The perturbation from 'exotic physic' can be described as a plane wave like impact.*

Assumption 1.b (Constant Velocity)**.** *The velocity of perturbation is constant.*

Assumption 1.a and 1.b decide the distance and time interval of signals between stations.
In reality, the perturbation is not likely to be perfect plane wave. Thinking in common sense, there can't be a kind of matter which interacts with the environment while keeping itself unchanged. So the key point is, to which extent does this 'exotic' change after interacting with atoms? We have to assume that the change on the perturbation can be ignored from human's or Earth's perspective. Otherwise Assumption 1.a and 1.b would be in contradiction with themselves.
Two more assumptions are also necessary.

Assumption 3.a (Cosine Amplitude)**.** *The amplitude of the signal is proportional to the projection of the plane wave's direction on sensor's sensitive axis.*

Assumption 3.b (Linear Combination)**.** *The amplitude of the signal is the linear combination of 'exotic' and others.*

Assumption 3.a and 3.b decide the amplitude of the signals. Of course we can make more complicated relation than linear and test them by adding several new functions in the program.
These rational assumptions are not necessarily true, especially when we are looking for 'exotic'. We are just carrying on research under these assumptions for now.

6.3 Algorithm

6.3.1 Brief Desciption

From now on, we will be looking for the trace of a perturbation or so-called event in all stations.
more to be written in Brief Desciption

6.3.2 Detailed Description

1. Baiscs
   To address the problem clearly, we will begin to dicuss in spherical coordinates frame.

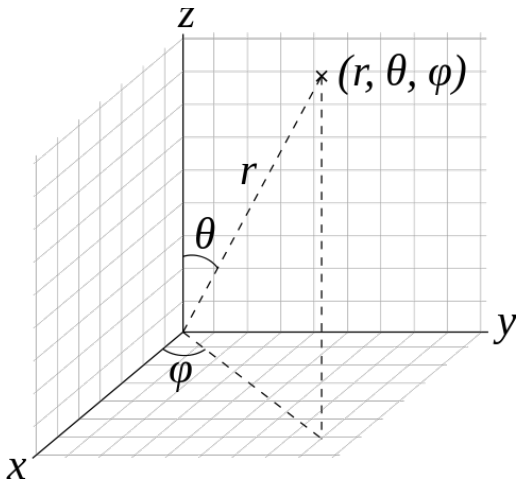
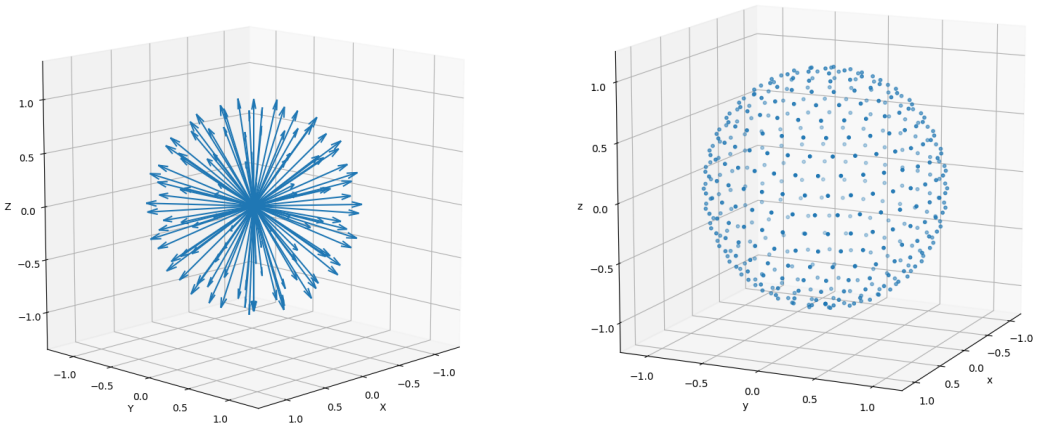


Figure 6.1: Spherical coordinates(r, θ,φ). From wiki

According to Assumption 1.a and 1.b, as long as we know the **locations of all the stations** on Earth, and assume the velocity as $\vec{v}$ (constant), it's simple to calculate **time intervals** of signals between stations caused by the same event. The calculation formulas will be illustrated later in realiziation section.

2. Generate reference
   Let's do some exercises first. Choose directions as shown in Figure 6.2. Since we know the **locations of all the stations** on Earth, we can calculate time intervals between stations in all these different directions.
   What we care about is the relative time interval between stations, so we can choose the time as **0** when one station first receive the perturbation. And by choosing a proper speed, we can always make it **1** the time when the last station receive the perturbation .
   The we save the directions and its time intervals as file. Basically, this is what can perform as reference for time intervals.

3. Locate Events
   After generating the reference, we can get started to seek events.
   We can infer from Assumption 1.a that, one perturbation would leave one signal in a sensor. By picking up one signal from one stations, we get a set of signals. We can call this **potential event**, which means it could arise from the same perturbation, or merely noise, or mixture of them two.
   Let's determine the problem we are coping with now. We have this **potential event**, and we know the time intervals between signals in this event, so we are trying to locate the direction of the perturbation which caused this **potential event**.
   Since we are looing for its direction, we can let alone the speed of the perturbation at present. We repeat the linear transform to transform the time intervals into [0, 1], which means the time is set as **0** when the first station received the perturbation, and time is set as **1** when the last station received the



(a) directions in vector form

(b) directions in scatter form

Figure 6.2: Directions. Vectors indicate the directions, while making it hard to recognize directions when there's too many. Scatters indicate the directions in a more fresh way.

perturbation.
We can look into the reference, and try to find a match between time intervals in the **potential event** and in reference. Of course there would not be a perfect match. But there could be some approximate match, which indicates a **real event**. There also could be no approximate match, which indicates a **fake event**.
I realized the following test in the program: Input the direction of a perturbation, and obtain a set of time intervals, which can be defined as 'to-be-located intervals'. Compare it with reference, and pick up three directions in the reference which match the 'to-be-located intervals' best. We can see the result in Figure 6.3. We can further locate the direction of the perturbation by generating a finer reference.

4. Examine the Result
   As mentioned before, we are play the game of probability. We have to calculate the confidence or uncertainty of the perturbation direction.
   Additionally, we have to take signal amplitude and perturbation speed into examination.
   Sadly, there's no conclusion to tell us how to calculate the confidence based on these factors. This part is to be completed.

5. Further Thinking
   Now we've got a big network around the world. I believe that perturbations coming from different directions will have their unique sets of time intervals, which is the 'fingerprint' of the direction. However, this is not necessarily true, especially when few stations response to the event. For example, when we get only three station, we can only tell the direction in the plane constructed by the stations.
   There's one more thing to be cautious with: we can never find the direction perfectly in nature for two reasons. Firstly, the reference is created via interpolation in directions. Secondly, when in excess power analysis, we divide the signal series into segments. We have no means to locate the signal accurately in this segment. The second reason is the fundamental reason. Even with perfect device and precise measurement, it's impossible to locate the signal. To deal with this, we could choose a smaller segment length in excess power analysis at the expense of frequency resolution. Or do finer analysis in excess power analysis at the expense of huge calculations and our brain cells.
   There are more factors in preventing as from find the perfect match, like the uncertainty in time-sync, flaws in previous assumptions and so on.

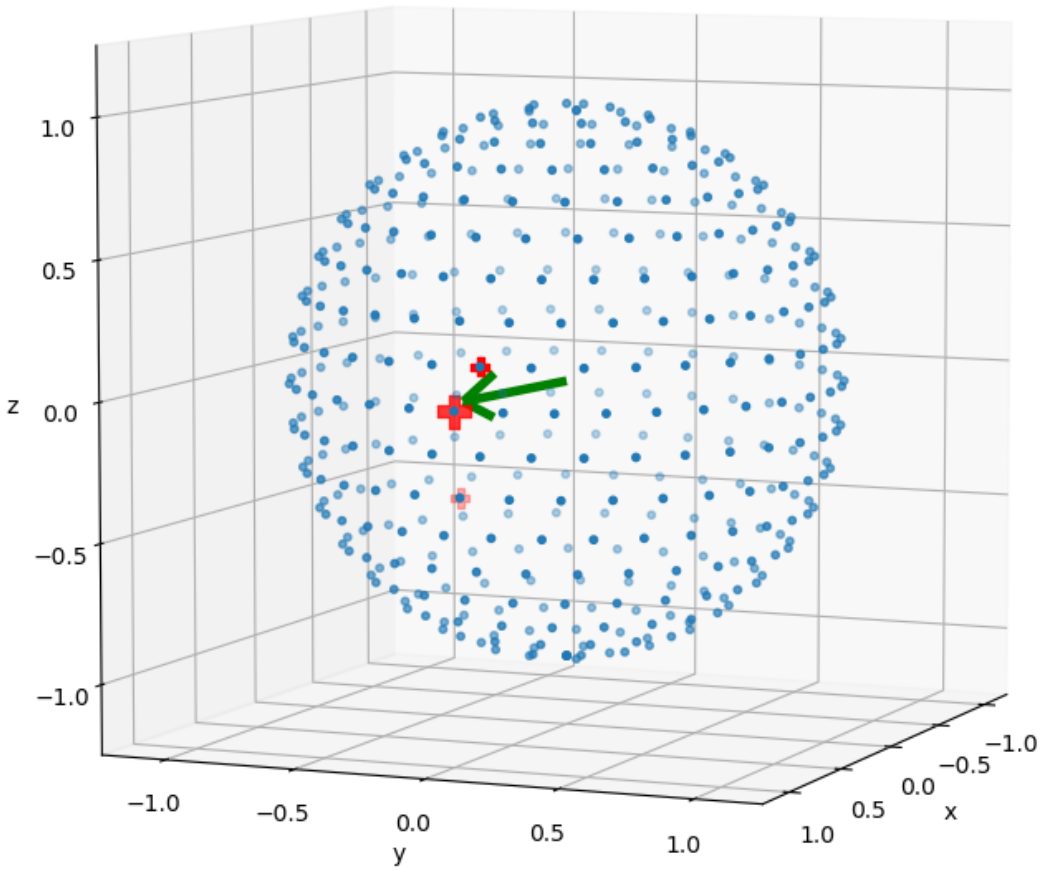


Figure 6.3: The green vector is real direction of perturbation. Red dots indicate three best-match directions in the reference.