# "Astronomical Time Series Analysis"

***End-Semester Report of***
***6$^{th}$ Semester Mini Project***
*FOR THE DEGREE OF*

**BACHELOR OF TECHNOLOGY**

*IN*

**INFORMATION TECHNOLOGY**

*BY*

Shreyansh Chaudhary(IIT2016068)
Shreyansh Dwivedi(IWM2016501)
Sashank Mishra(IIT2016515)
Pradeep Gangwar(IHM2016501)
Rishi Shukla(IIT2016507)

*UNDER THE SUPERVISION OF*

Dr. Pavan Chakraborty IIIT-ALLAHABAD

**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD**

**May, 2019**

# CANDIDATES’ DELARATION

We hereby declare that the work presented in this project report entitled “**Astronomical Time Series Analysis**”, submitted end-semester report of 6th Semester report of B.Tech. (IT) at Indian Institute of Information Technology, Allahabad, is an authenticated record of our original work carried out from January 2019 to May 2019 under the guidance of **Dr. Pavan Chakraborty**. Due acknowledgements has been made in the text to all other material used. The project was done in full compliance with the requirements and constraints of the prescribed curriculum.

Place:Allahabad

Shreyansh Chaudhary(IIT2016068)
ShreyanshDwivedi(IWM2016501)
Sashank Mishra(IIT2016515)
Pradeep Gangwar(IHM2016501)
Rishi Shukla(IIT2016507)

# CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:

Place: Allahabad

**Dr. Pavan Chakraborty**
Professor
IIIT-Allahabad

# Table of Contents

# ASTRONOMICAL TIME SERIES ANALYSIS

**Abstract**

This report introduces to a basic application of Statistics in Astronomical Data for Statisticians and Computer Scientists.
In this project, we aim to statistically analyze the various aspects of the time-series data generated from astronomical sources namely LINEAR and LIGO Dataset. The components we aim to analyse from the project are:

- Forecasting
- Seasonality and Trend Detection
- Power Spectral Density Estimation
- Frequency Estimation

Keywords- Time Series, Power Spectrum, ARIMA, Astrophysics.

## 1. Introduction

The recent advancements in Astronomy and Astrophysics are mostly obtained from the interpretation of the signals, encoded in the light received from distant objects. The data obtained in astronomical observations is quite noisy, and therefore various statistical methods are developed to tackle the problems in astronomy. These methods include least-squares estimation, and some contemporary methods such as nested sampling.

The data obtained in astronomical observations is mainly of three types: Image Data, Spectral Data and Time Series and Functional Data. The image and spectra are a kind of 'raw' data, which provide a basis of deriving several data types. For the various light sources present in the sky, the astronomers are interested in analysing the brightness over a period of time. Such kind of analysis gives rise to the time series or light curve for each object.

In the time domain, we deal with some periodic phenomena such as binary orbits, stellar rotation and pulsation. Also, we try to examine the transient phenomena such as gamma ray bursts, solar flares, supernovae explosions and other powerful explosions.

Through this project, we aim to analyse the different time series data obtained from various astronomical sources, using various statistical techniques. We aim to analyse the seasonality and trend of the data. Also, we deal with the forecasting using ARIMA model, and try to generate the Power Spectral Density using Lomb-Scargle Periodogram. In this way, this project will serve as an introduction in the area of Time Domain Astronomy.

## 2. Motivation

Through this project, we plan on exploring the domain of astrophysics. There is a huge amount of data that can be analyzed for different patterns and hence opens up possibilities for various predictions based on the humongous data.

In previous semesters, we have never come across such a topic and hence in order to start, we are taking a time series and analyzing the same using the various statistical approaches to analyze the same in order to gain some insight regarding the project.

This will prove to be the first step on our path to exploring new possibilities in the above mentioned field.

## 3. Problem Formulation

We have formulated a problem of analyzing various Time Series data, obtained from different astronomical sources, such as LIGO, LINEAR, RR-Lyrae and others. We aim to apply and compare different statistical methods, starting with the seasonality and trend detection. Then we go on to generation of Power Spectra, starting with the Fourier transform, to the more advanced versions, such as Lomb-Seargle Periodogram. Later on, we deal with the forecasting of the time-series data using **Auto Regressive Integrated Moving Average(ARIMA)** model. We then plan to obtain various insights on it.

## 4. Literature Review

### *4.1. Time Series*

A Time Series consists of various data points plotted over a course of time. In general view, it consists of sequence of data, evenly spaced over in time. Therefore, it can be classified as a discrete data. The time series are mostly used in statistics, signal processing, econometrics, weather forecasting, astronomy and many other fields.

The Time Series Analysis basically comprises of the application of statistical techniques to analyse the time series data and extract useful information from it. Time series forecasting is the use of a model to predict future values based on previously observed values. The various analysis that can be done on the time series are as follows:

- Forecasting
- Signal Detection and Estimation
- Clustering
- Classification
- Curve Fitting
- Segmentation

### *4.2. Power Spectral Density*

The power spectrum $S_{xx}(f)$ of a time series $x(t)$ describes the distribution of power into frequency components composing the signal. When the energy of the signal is concentrated around a finite time interval i.e. if its total energy is finite, we compute the Energy Spectral Density.

It applies to signals existing over all time, or over a time period large enough that it could as well have been over an infinite time interval.

The power spectrum answers the question How much of the signal is at a frequency **w**?. We have seen that periodic signals give peaks at a fundamental and its harmonics; quasi periodic signals give peaks at linear combinations of two or more irrationally related frequencies (often giving the appearance of a main sequence and sidebands).

### *4.3. Transformation Functions*

#### *4.3.1. Fourier Transform*

A fast Fourier transform (FFT) is an algorithm that computes the Discrete Fourier transform(DFT) of a sequence. The DFT is obtained by decomposing a sequence of values into components of different frequencies which is often a slow process in the practical world. An FFT rapidly computes such transformations and as a result, it manages to reduce the complexity of computing the DFT.

#### *4.3.2. Welch's Method*

This method proposed by Welch is often used as an estimator of the PSD. The method consists of dividing the time series data into (possibly overlapping) segments, computing a modified periodogram of each segment, and then averaging the PSD estimates resulting in Welch's PSD estimate.

Although overlap between segments introduces redundant information, this effect is diminished by the use of a non-rectangular window, which reduces the importance or weight given to the end samples of segments
A Window Function is a mathematical function that is zero-valued outside of some chosen interval, normally symmetric around the middle of the interval, usually near a maximum in the middle, and usually tapering away from the middle. When another function is multiplied by a window function, the product is zero-valued outside the interval: all that is left is the part where they overlap, the "view through the window". In actual practice, the segment of data within the window is first isolated, and then only that data is multiplied by the window function values.
There are many types of windows present, but the 2 most common of them are:

- Hanning Window - This window is used in order to smoothen the discontinuities at the beginning and end of the sampled signal.
- Top Hat Window - Top Hat Window is typically employed on data where frequency peaks are distinct and well separated from each other.

The Hanning Window is preferred when the frequency peaks are not guaranteed to be well separated as it is less likely to cause individual peaks to be lost in the spectrum whereas if the frequency peaks are distinct and well separated from each other Top Hat Window is employed.

### *4.3.3. Lomb-Scargle Periodogram*

The Lomb-Scargle periodogram is a standard method to search for periodicity in unevenly sampled time series data. It is a well-known algorithm for detecting and characterizing periodicity in unevenly sampled time-series. Since, it can handle uneven data well, it is widely used within the astronomy community. The LombScargle periodogram is a method that allows efficient computation of a Fourier-like power spectrum estimator from such unevenly sampled data, resulting in an intuitive means of determining the period of oscillation.
The Lomb-Scargle periodogram corresponds to a single sinusoidal model, $\mathbf{y(t) = a\ sin(wt) + b\ cos(wt)}$ , where $t$ is time and $w$ is angular frequency ($= 2\pi f$). The model is linear with respect to coefficients $a$ and $b$, and non-linear only with respect to frequency $w$. A Lomb Scargle periodogram measures the power $\mathbf{P_{LS}(w)}$

$$\mathbf{P_{LS}(f) = \frac{1}{2}\{\frac{(\sum_n g_n \cos(2\pi f[t_n - \tau]))^2}{(\sum_n \cos^2(2\pi f[t_n - \tau]))} + \frac{(\sum_n g_n \sin(2\pi f[t_n - \tau]))^2}{(\sum_n \sin^2(2\pi f[t_n - \tau]))}\}}$$

where $\tau$ is specified for each f to ensure time-shift in-variance:

$$\tau = \frac{1}{4\pi f}\arctan(\frac{\sum_n \sin(4\pi f t_n)}{\sum_n \cos(4\pi f t_n)})$$

*4.4. Time Series Decomposition*

Decomposition provides a useful abstract model for thinking of a time series as a combination of level, trend, seasonality, and noise components, and for better understanding problems during time series analysis and forecasting.

A useful abstraction for selecting forecasting methods is to break a time series down into systematic and unsystematic components. A given time series is thought to consist of three systematic components including level, trend, seasonality, and one non-systematic component called noise.

These components are defined as follows:

- Level - The average value in the series
- Trend - The increasing or decreasing value in the series.
- Seasonality - The repeating short-term cycle in the series.
- Noise - The random variation in the series.

A series is thought to be an aggregate or combination of these four components, either as an Additive model or a Multiplicative model. An additive model suggests that the components are added together as follows:

$$\mathbf{y(t) = Level + Trend + Seasonality + Noise}$$

A multiplicative model suggests that the components are multiplied together as follows:

$$\mathbf{y(t) = Level * Trend * Seasonality * Noise}$$

*4.5. Autocorrelation and Partial Autocorrelation*

The sample autocorrelation function (ACF) for a series gives correlations between the series $x_t$ and lagged values of the series for lags of 1, 2, 3, and so on. The lagged values can be written as $x_{t-1}$, $x_{t-2}$, $x_{t-3}$, and so on. The ACF gives correlations between $x_t$ and $x_{t-1}$, $x_t$ and $x_{t-2}$, and so on.

Let $x_t$ denote the value of a time series at time $t$. The ACF of the series gives correlations between $x_t$ and $x_{t-h}$ for $h = 1$, 2, 3, etc. Theoretically, the autocorrelation between $x_t$ and $x_{t-h}$ equals

$$\mathbf{ACF} = \frac{\mathbf{Covariance(x_t, x_{t-h})}}{\mathbf{Std.Dev.(x_t) * Std.Dev.(x_{t-h})}}$$

The ACF can be used to identify the possible structure of time series data. That can be tricky going as there often isnt a single clear-cut interpretation of a sample autocorrelation function.

The PACF, on other hand, is a conditional correlation. It is the correlation between two variables under the assumption that we know and take into account the values of some other set of variables. For instance, consider a regression context in which $y$ is the response variable and $x_1$, $x_2$ and $x_3$ are predictor variables. The partial correlation between $y$ and $x_3$ is the correlation between the variables determined taking into account how both $y$ and $x_3$ are related to $x_1$ and $x_2$.

In regression, this partial correlation could be found by correlating the residuals from two different regressions:

- Regression in which we predict $y$ from $x_1$ and $x_2$,
- regression in which we predict $x_3$ from $x_1$ and $x_2$. Basically, we correlate the parts of $y$ and $x_3$ that are not predicted by $x_1$ and $x_2$.

More formally, we can define the partial correlation just described as

$$\mathbf{PACF} = \frac{\mathbf{Covariance(y, x_3|x_1, x_2)}}{\sqrt{\mathbf{Variance(y|x_1, x_2) * Variance(x_3|x_1, x_2)}}}$$

### *4.6. ARIMA modelling*

### *4.6.1. Auto Regressive(AR) model*

An autoregressive (AR) model is a representation of a type of random process; as such, it is used to describe certain time-varying processes in nature, economics, etc. It specifies that the output variable depends linearly on its own previous values. The notation **AR(p)** indicates an autoregressive model of order p. The AR(p) model is defined as:

$$\mathbf{X_t} = \mathbf{c} + \sum_{\mathbf{i=1}}^{\mathbf{p}} \phi_{\mathbf{i}} \mathbf{X_{t-i}} + \varepsilon_{\mathbf{t}}$$

### *4.6.2. Moving Average(MA) model*

In time series analysis, the moving-average model(MA model), is a common approach for modeling univariate time series. It specifies that the output variable depends linearly on the current and various past values of a stochastic (imperfectly predictable) term.Together with the AR model,it is a special case and key component of ARIMA model of time series.The notation MA(q) refers to the moving average model of order q:

$$\mathbf{X_t} = \mu + \epsilon_{\mathbf{t}} + \theta_{\mathbf{1}} \epsilon_{\mathbf{t-1}} + \cdots + \theta_{\mathbf{q}} \varepsilon_{\mathbf{t-q}}$$

### *4.6.3. ARIMA model*

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

In this $x(t)$, we express as a function of past value(s) of x and/or past errors (as well as a present time error). When we forecast a value past the end of the series, we might need values from the observed series on the right side of the equation or we might, in theory, need values that arent yet observed.

Any ARIMA model can be converted to an infinite order MA model:

$$x_t - \mu = w_t + \Psi_1 w_{t-1} + \Psi_2 w_{t-2} + \cdots + \Psi_k w_{t-k} + \dots \quad (1)$$

$$= \sum_{j=0}^{\infty} \Psi_j w_{t-j} \text{ where} \Psi_0 \equiv 1 \quad (2)$$

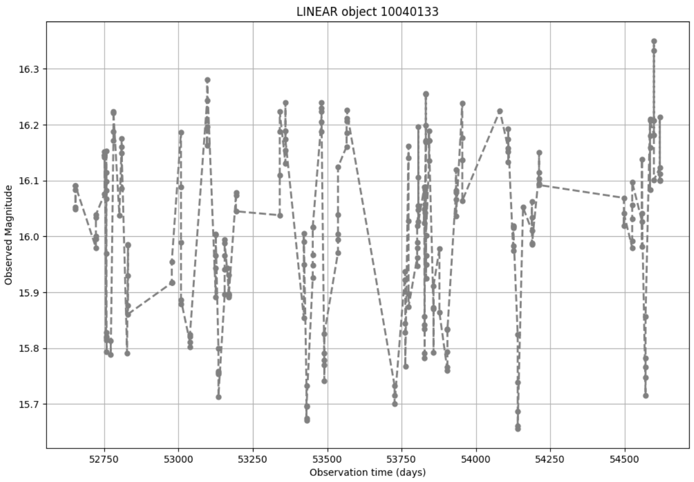
## 5. Proposed Methodology

### *5.1. Data set Description*

### *5.1.1. LINEAR*

LINEAR (Lincoln Near-Earth Asteroid Research) is a project collaborated to detect and catalog the near-Earth asteroids or Near-Earth objects (NEO). The data set contains over 5 billion photometric measurements for about 25 million objects, mostly stars. We use Sloan Digital Sky Survey (SDSS) data from the overlapping ∼10,000 deg2 of sky to recalibrate LINEAR photometry and achieve errors of 0.03 mag for sources not limited by photon statistics with errors of 0.2 mag at r∼18. With its 200 observations per object on average, LINEAR data provide time domain information for the brightest four magnitudes of the SDSS survey.

Figure 1: Original Data Plot



### *5.1.2. LIGO*

We put into use the time series released from LIGO (Laser Interferometer Gravitational Observatory) Hanford, sampled at 4096 Hz. The data is calibrated such that the the wave signals have units of dimensionless strain ($\Delta$L/L).

### *5.2. Technologies Used*

- Numpy
- AstroML
- Pandas
- Astropy
- Scipy
- Tkinter

### *5.3. Process Performed*

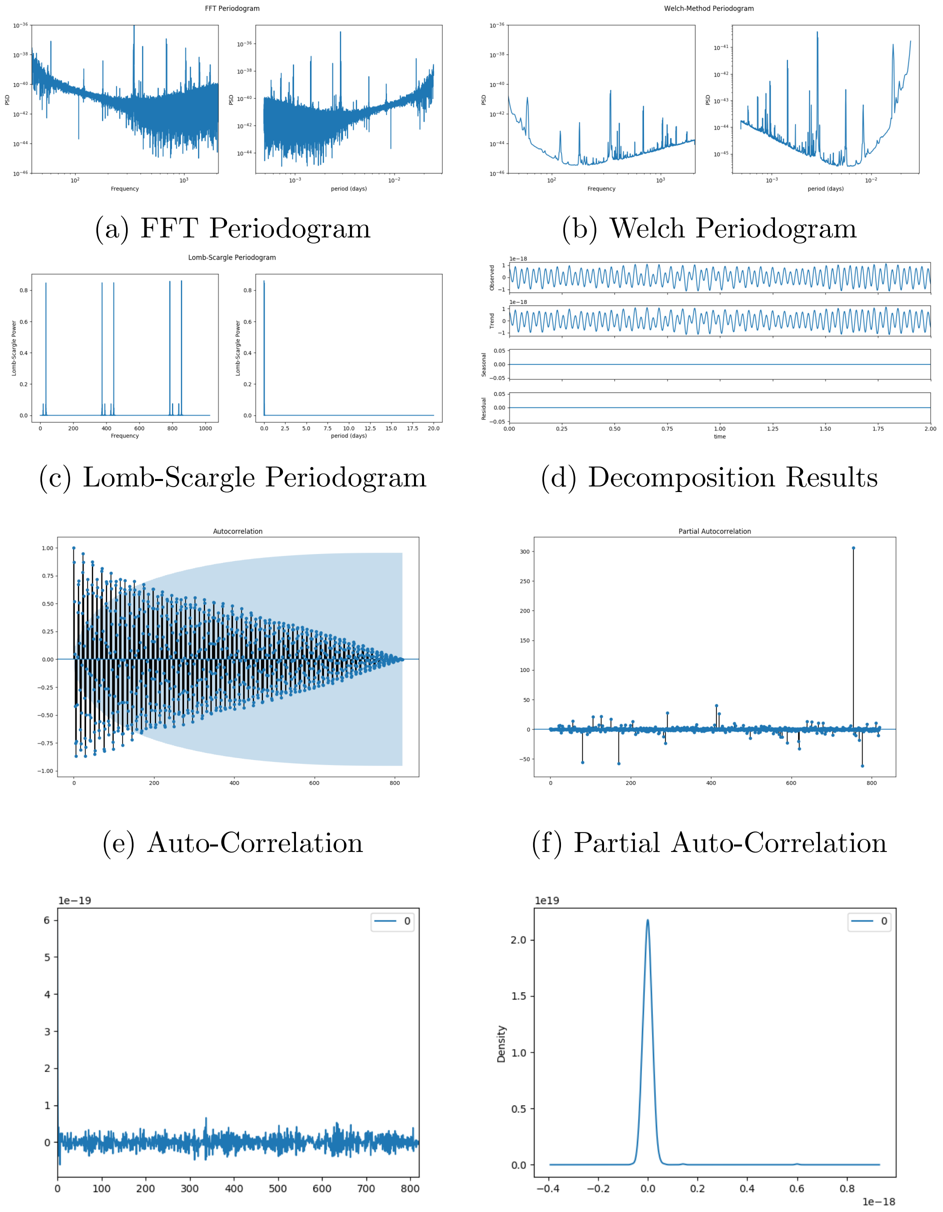
Initially we have imported the datasets from their respective official archives. The LIGO dataset has evenly-sampled time periods, while the LINEAR is unevenly sampled. Therefore, have dealt with them in two separate ways.

For the LIGO data, we have extracted a section of the data for our purpose. The data is then decomposed into their respective seasonality and trends so as to gain better insight into the time series. For the Power Spectral Density, all the possible methods (viz. FFT, Welch's method and Lomb-Scargle method) have been applied. For the ARIMA modelling, we first obtained the plot of the auto correlation and partial autocorrelation. The ACF gives us an estimate of the order of the MA model, while the PACF gives an estimate of the AR modelling. The results of both of them are combined so as to create the required ARIMA model.

For the LINEAR data, we have only the option of applying the Lomb-scargle method so as to create the required Power Spectral Density. Also, the ACF and PACF have been obtained for the data.

Finally, a Graphical User Interface has been created so as to present the results in an interactive manner.

### *5.4. Results*



(a) FFT Periodogram

(b) Welch Periodogram

(c) Lomb-Scargle Periodogram

(d) Decomposition Results

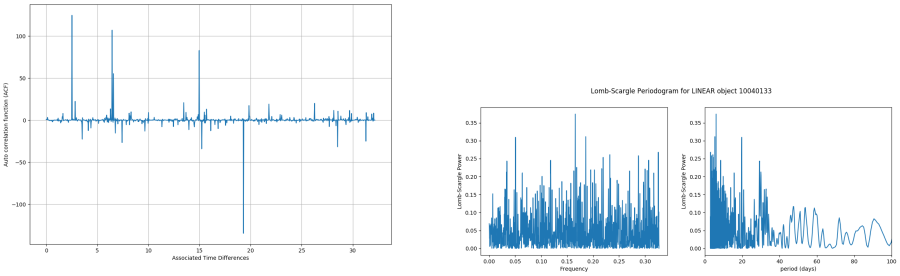
(e) Auto-Correlation

(f) Partial Auto-Correlation

(g) line plot of the Residual Errors of ARIMA Model

(h) Density plot of the residual error values

Figure 2: Graphical results based on LIGO



(a) Auto Correlation function

(b) Lomb-Scargle Periodogram

Figure 3: Graphical results based on LINEAR

## 6. CONCLUSION

Through this project, we aim to analyse and experiment over the astronomical data, using the time series data generated over several experiments and observations. We aim to explore and analyse different method of statistical techniques of time series analysis viz. Decomposition, Spectral Analysis and Forecasting. Our aim is to deal with big, uneven data, and apply different statistical methods to generate insight into the data.

## REFERENCES

[1] James P. Long, Rafael S. De Souza, Statistical Methods in Astronomy

[2] Petre Stoica, Randolph Moses, Spectral Analysis of Signals

[3] Jacob VanderPlas, Andrew J. Connolly, Zeljko Ivezic, Alex Gray, Introduction to astroML: Machine Learning for Astrophysics

[4] R. Vio, N.R. Kristensen, H. Madsen, W. Wamsteker, Time series analysis in Astronomy: limits and potentialities

[5] B. P. Abbott, R. Abbott, R. Adhikari et aI., UGO: the Laser Interferometer Gravitational-Wave Observatory

[6] B. Sesar, J. S. Stuart, Z. Ivezic, D. P. Morgan, A. C. Becker, and P. Wozniak, Exploring the Variable Sky with LINEAR. I. Photometric Recalibration with the Sloan Digital Sky Survey