PRESIDENCY
UNIVERSITY
KOLKATA

On Classification of Light curves of Eclipsing Binary Star System

Author

Angshumita Sarkar

Reg: 19214220017
PG 2 Sem 4
Presidency University
Kolkata

Supervised by
Dr. Atanu Kumar Ghosh

Contents

CHAPTER 1

Introduction

1.1. Prelude

"Two things inspire me to awe- the starry heavens above and the moral universe within"

- **Albert Einstein**

We all have certainly gazed up at the night sky and pondered the twinkling stars at least once in our life. If we think for a while then we shall realize that there is a deeper complexity behind this beauty than what initially meet our eyes. For example, imagining two stars gracefully dancing around each other, intermittently concealing and revealing there luminosity in a celestial waltz .Technically, we call it an **eclipsing binary star system**. And this project aims to study the light curves of such system in more detail.

Astronomers observed unusual changes in the brightness of specific stars over a hundred years ago. They realized these fluctuation where not arbitrary, but rather the two stars revolving around each other in cosmos. The discovery marked the inception of eclipsing binary research. Now we introduce the concept of **Binary Star system** in following part .

1.2. Eclipsing Binaries and other variable stars

Our solar system has one star, the sun. But this is actually bit uncommon in our galaxy. A greater percentage of system containing two or more stars, and in particular **" Binary star system".** One such binary system is visible to the naked eye, Alpha Centauri A & B, as shown in 1.2.1.Variable stars are stars that vary in apparent brightness with time. In fact, all stars are variable at some level of precision, over some timescale. According to E.F. Kallrath, J. & Milone,**[5]** in astronomy there are three basic timescales:

- *dynamic:* (the time it would take for a star to collapse under gravity if radiation and particle pressure were removed), typically tens of minutes;
- *thermal:* (the time to exhaust its stored thermal energy), typically millions of years (as the energy is depleted by the luminosity); and
- *nuclear* (the time to exhaust its nuclear energy), typically, billions of years.

The relevant timescales for variable stars are between the dynamic and the thermal, but certainly much closer to the dynamic. In fact the term "variable star" is usually reserved for stars that vary in brightness by some detectable amount over the interval of the observations. We have no prehistoric record of

such events, but we certainly have ancient records. In the recent past (50 years or so), the observational precision has been of order 0.01 magnitude or more. At present, photometry has, in principle if not usually in practice, improved by an order of magnitude, and at the level of mili magnitudes, most stars will appear variable. For example, [**4**] found in a survey of the galactic cluster NGC 2301 that 56% of 4000 stars were variable at an amplitude of 0.002 magnitude or greater, the precision limit of the survey for the brightest 5 magnitudes of the survey. To keep our present exposition within reasonable bounds, for present purposes, for the most part we will stick to the more classical limit to define a "variable star," namely a star with brightness variation of > 1% or so and over timescales of millennia or less (down to seconds or less). More specifically, in the wider literature variable stars have been held to be variable if they vary in optical wavelengths (~ 0.35 to < 1.0μm) over intervals of decades or less; cf. [**3**].

variable stars are classically assigned to one of three main categories:

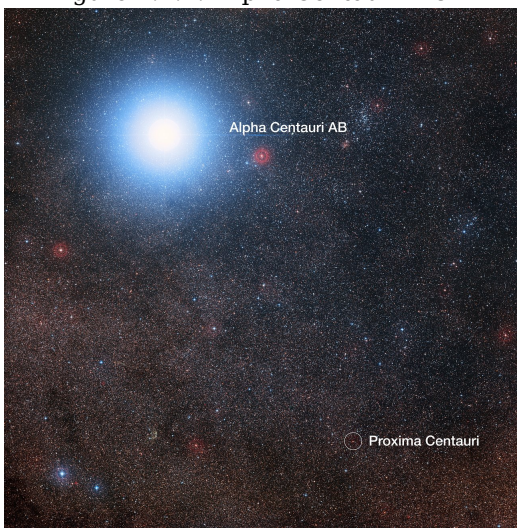
- "geometric" variables;
- "pulsating" variables; and
- "eruptive" variables.

In another classification scheme, a broader distinction was made between "extrinsic" and "intrinsic" variables, with "geometric" variables considered "extrinsic," and the other two "intrinsic." We shall discuss geometric, pulsating, and eruptive variables in sequence. A "geometric" variable varies not due to its own physical behavior but because of changing aspect,1 i.e., the viewable part of a star changes with time. This category includes the EBs and also examples where the eclipse is due to a disk or a planet. It can also include pulsars, which vary mainly because of rotation, and spotted stars in the sense that the spots cause light modulation over a rotation period; spots usually do not last for decades, but there are exceptions [e.g., RW Com, cf. Milone et al. (1980)]. Finally, a cataclysmic variable (see below) may have a large "hump" (due to a hot spot) in its light curve which may be asymmetric due to eclipse by the companion star. Thus the degree to which geometric effects cause the observed variation will differ with the type of system.

1.2.1. Eclipsing Variables: According to E.F. Kallrath, J. & Milone,[**5**] **Eclipsing variables** are periodic (that is, the cycle of variation repeats relatively reliably). This broad grouping was historically divided into three phenomenological classes according to the appearance of the light curves: Algols, β Lyrae systems, and W Ursae Majoris systems. The characteristics of these light curve types are discussed in the following subsections.

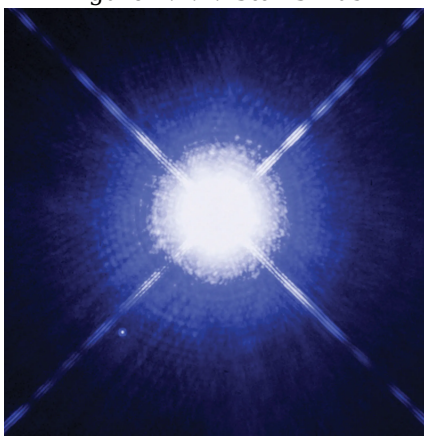
- ***Algol Variable(EA):*** The prototype is β Persei, also known as Algol. In visible passbands, the striking characteristics are approximately constant light outside eclipse and minima that fall and rise abruptly and occupy only a small fraction of the full light curve, typically less than ~15% for each minimum. Typically, the longer the period, the shorter the fraction of light curve taken up by eclipse. The periods range from days to weeks or more in length. Usually such light curves

Figure 1.2.1. Alpha Centauri A &B



suggest little interaction between components. This is often but not generally true because the shapes of light curves in optical bands can be misleading. For example, where the depths of the two minima are very different, the temperatures of the component stars are different,2 and the hotter, bluer star may dominate the light from the system. The light curve may rise near secondary minimum, indicating a "reflection effect," actually a reprocessing of the hotter stars' radiation as it impinges on the atmosphere of its companion, increasing the cooler star's luminosity in the irradiated area – best seen around the secondary minimum. If it were not for the secondary minimum, which may be shallow or even absent in optical passbands, the reflection effect would peak at the phase of mid-secondary minimum. The effect is especially noticeable if the cooler star is significantly larger. If looked at in infrared passbands, the cooler star will contribute relatively more to the combined light and may resemble a β Lyrae-type light curve 1.6.3. In extreme cases, the redder component is so highly evolved that it may be filling its Roche lobe and sending a stream of material toward its companion. This is, in fact, the case with Algol itself. Not all systems with Algol-like light curves will be in this state, however; the components are often two similar stars and not so close to each other that they are distorting each other's shapes. When the stars are far apart, their shapes may be approximated by spheres. It suffices to note that the spherical approximation would not be adequate for all cases. We discuss further this class of eclipsing variable under the EB designation "EA" in the section below. As outlined in Sect. 3.1.6 binary systems in which components are well within their Roche lobes are called "detached" systems and those in

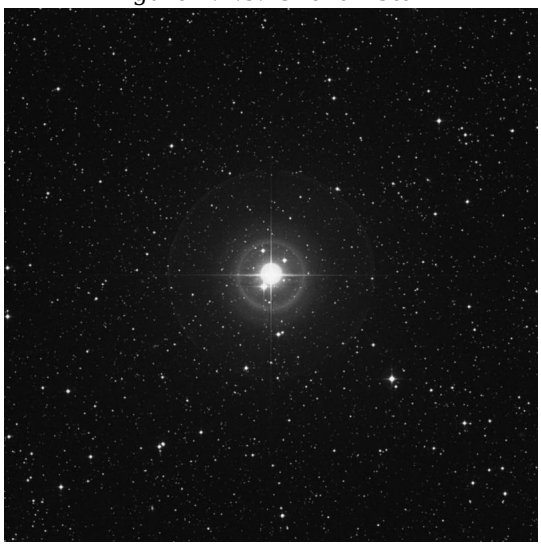
Figure 1.2.2. Star Sirius



which one component fills its Roche lobe are called “semi-detached.” For example1.2.2[**14**]

- ***Beta Lyrae(EB):*** Beta Lyrae, eclipsing binary star, the two component stars of which are so close together that they are greatly distorted by their mutual attraction; they exchange material and share a common atmosphere. The prototype gives its name to the class. The light curve continuously varies across the cycle of variation, and the minima occupy a fairly large proportion of the cycle. The periods are typically days, but when giants or supergiants are involved, the period may be much longer. The important thing is not the cycle length or the scale of the system, but the relative size of the stars to the size of the orbit. The continuous variation of light is partially due to the changing aspects of the stars as they rotate, classically known as the “ellipsoidal variation.” The relative depths of the minima indicate the temperature difference between components; redder passbands tend to show less different depths. The light curves give the (correct) impression that the stars are interacting gravitationally. In fact, the stars are undergoing tidal distortions and their shapes reflect this distortion. Roche geometry is generally used to accurately model these systems’ properties.Beta Lyrae is a member of a class of binary systems known as W Serpentis stars. It is of about third magnitude and lies in the northern constellation Lyra.The variable character of Beta Lyrae was discovered in 1784 by the English amateur astronomer John Goodricke. Its period of about 13 days is increasing by about 19 seconds per year, probably because the stars are steadily losing mass to a continually expanding gaseous ring surrounding them. For example , 1.2.3[**15**]
- ***W Ursae Majoris(EW): W*** Ursae majoris is the variable star designation for a binary star system in northern constellation of Ursa Major. It has an apparent visual magnitude of about 7.9, which is too faint to be seen with the naked eyes. however, it can be viewed with a small

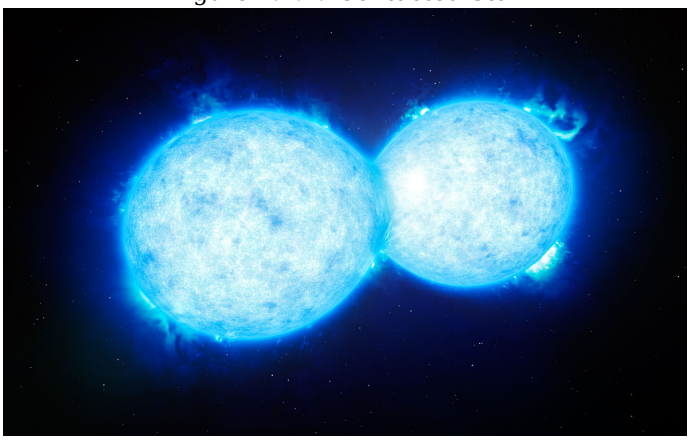
Figure 1.2.3. Sheliak Star



telescope.The prototype is an eclipsing binary with period less than a day, characteristic of the class. Like β Lyrae stars, the light curve varies continuously, but the depths of the minima are usually similar, but rarely exactly identical. Systems that exhibit these light curves are thought to arise from binaries in physical contact, not through a stream, but through an actual neck of material that bridges the small distance between the inward pointing edges of the components. Such systems are known as "over-contact" or if just barely touching, as "contact" systems. Although some astronomers use the term contact to refer to both contact and over-contact systems, here we will use the term exclusively in its narrower meaning. Roche geometry is used generally for the accurate modeling of the components of these systems. There are two sub classes of WUMa systems, about which capable astronomers argue endlessly: A-type and W-type systems. In A-type systems the more massive star is larger and hotter; in W-type systems, the more massive star is larger but cooler than its companion. Although both types of systems may exhibit asymmetries in light curves, the W-type tends to exhibit more of this sort of behavior. There may be a difference in depth of up to 0.1 magnitude.1.2.4[**13**]

- ***Pulsating Variable(PUL):*** Pulsating variables undergo variations in radius due to intrinsic variation of temperature and pressure. They may be strictly periodic, as in RR Lyrae stars or Cepheid variables, or merely cyclic, as in RV Tauri or Mira variables. The period of variation may be very rapid – minutes for some high-temperature variables – to years for the Miras.The General Catalogue of Variable Stars [**12**] lists the following types: α Cygni; β Cephei; Cepheids; W Virginis; δ Scuti; Irregular; Mira; PV Telescopii; RR Lyrae; RV Tauri; Semiregular; SX

Figure 1.2.4. Contacted Star



Phoenicis; and ZZ Ceti; most of them have assigned subtypes, which we omit here.. Of greatest interest to those outside the variable star community are the RR Lyrae variables, with approximately constant luminosities, and the Cepheids, with luminosities that increase with period. Such stars are considered to be "standard candles," and may be used to determine the distance on any ensemble in which they are found. RR Lyrae stars are giant stars that have periods of about half a day. Cepheids are supergiant stars that have periods from 1 to tens of days. Both are found in the field and in globular clusters, but RR Lyrae stars are much more common. There are two types of Cepheids, the classical Cepheids, members of Population I that are younger and are associated with the galactic plane, and Population II Cepheids, found in the galactic halo and in globular clusters. The realization that there are two types of Cepheids by Walter Baade led to a revised (primarily, extragalactic) distance scale (by a factor of 2). Closely related to the Cepheids and RR Lyrae objects are the δ Scuti stars (so designated by Harlan J. Smith 1955) and their globular cluster-resident cousins, the SX Phoenicis stars. These are subgiants or dwarfs (luminosity classes IV and V, respectively) that have periods of pulsation that are typically small fractions of a day. All three groups are found in the "instability strip" on the Hertzsprung–Russell diagram (luminosity or absolute magnitude versus spectral type, stellar temperature or color index: basically brightness plotted against color), where no stable stars are found.

1.3. Types of Binary Star system on the basis of visibility

Here we classify them on the basis of from the basis of visibility. [**11**]

1.3.1. Visual Binaries: When we can see both stars, with bare eyes , means we can resolve them separately they are called visual binaries. This

classification is not robust, as differentiating both star depends on different eyesight and better equipment.

But beyond this, we also know that stars comes in different varieties, depending on their temperature to their different stages in evaluation. Therefore, the possible combinations that exist for the two types of stars that comprise a binary system are even more numerous

1.3.2. Spectroscopic Binaries: When a binary system is too distant, or when the stars are too close together, the stellar bodies cannot be resolved separately by a telescope as with visual binaries. However, astronomers can use a phenomenon called Doppler shift to distinguish the stars in these binary systems, which are called spectroscopic binaries.The detection and analysis of spectroscopic binaries is not subject to geometrical resolution limits as are angular measurements. With sufficient light gathering power, it is possible to investigate spectroscopic binaries even in nearby galaxies and to derive the luminosity ratio and mass ratio.

1.3.3. *Eclipsing Binaries:* According to E.F. Kallrath, J. & Milone, a variable star observer measures a time-dependent flux, the display of which versus time or phase (the repeated folding of the time into the period of variation) is known as the light curve. Whereas eruptive, pulsating, rotating, and cataclysmic variables are said to be intrinsic variables caused by different physical mechanisms, EBs are extrinsic variables requiring models including both astrophysics and geometry. As we have indicated, an eclipsing variable is a binary system whose orbital motion is in a plane sufficiently edge-on to the observer for eclipses to occur. The smaller the orbit relative to the sizes of the stars, the greater the likelihood of eclipses. For a special subgroup of EBs (so-called over-contact binaries, with a common envelope) eclipses may occur, although perhaps not perceptibly, even if the inclination is as small as 35$\circ$ .These binaries usually have orbital periods of less than 10 days and in most cases less than 1 day. Among the exceptions are some rare cases of hot and/or developed systems. The longest period EB known at present is ε Aurigae with an orbital period of 27.1 years.According to Kepler's third law this binary has an orbit relatively large10 compared to the sizes of the components. Historically, considerations concerning the likelihood of eclipses lead to a connection between EBs and "close binaries." In the early days, a "close binary" was defined as a binary with component radii not small compared to the stars' separation.

B studies often involve the combination of photometric (light curve) and spectroscopic (mainly, radial velocity curve) data. Analysis of the light curve yields, in principle, the orbital inclination and eccentricity, relative stellar sizes and shapes, the mass ratio in a few cases, the ratio of surface brightness, and brightness distributions of the components among other quantities. If radial velocities are available, the masses and semi-major axis may also be determinable. Many other parameters describing the system and component stars may be determined, in principle, if the light curve data have high enough precision and the stars do not differ greatly from the assumed model. The prediction of the information content of particular light curves has been a major topic of

concern in binary star studies; the exposition of this topic is an important component of the present work also.

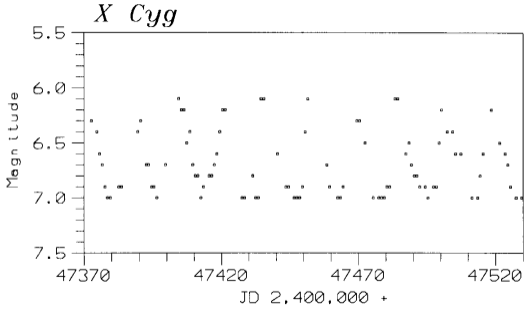
So, we can see that stellar evolution is much more complicated in binary system due to the influence each star has on each other. There are so many fascinating system out there, not just binary system, but triple star system and beyond. As complex they are , many of them harbor planets.

In our study and in our data set The data set under our study contains observations on 1318 new variable stars[**7**] covering 0.25 square degree region of the Galactic plane centered on Galactic coordinates (latitude, longitude) of (330.94, −2.28) degree. Subjective study[**7**]based on the appearance of observed LCs of the stars hinted at four possible groups, viz. Algol type (EA), Beta Lyrae (EB), W Ursae Majoris (EW) and un-categorized pulsating stars (PUL).

1.4. Why Binary stars are important?

Binary stars are important, first, because they are numerous. Latham et al. [**16**]conclude that the frequency of spectroscopic binaries detected in the galactic halo is not significantly different from that in the disk, despite differences in kinematic properties and chemical composition. The observed frequency is approximately 20%; the actual frequency is higher because many binaries remain undetected. In the solar neighborhood, where we have the benefit of proximity so that proper motion variations can be detected, the frequency is more than 50% – and several stars are in fact multiple systems. The second reason why binaries are important is that they are the primary source of our knowledge of the fundamental properties of stars. For example, the direct determination of the mass of any astronomical object requires measurable gravitational interaction between at least two objects (galaxy–galaxy, star–star, star–planet, planet–satellite). In galaxy–galaxy interactions, the distances and separations are so large that no detectable motion on the plane of the sky is possible. In star–planet interactions the objects contrast so greatly in brightness that outside the solar system only the highest possible – and until recently rarely attained – precision can resolve the objects. Typically in the latter case, only the star's motions are detectable, and the properties of that star must be assumed, mainly on the basis of binary star studies, in order to deduce the properties of the planet. In star–star interactions, the variations in position and velocity caused by orbital motion are detectable for a wide range of stellar separations and up to at least a factor of 5 in brightness. It is often the case that both stars may be studied in any of several ways, depending on their distances, brightness, and motions. Other basic properties of stars and of the systems they constitute can be determined through analysis of observational data, depending on the observational technique by which the interaction is studied. The four main types of binaries described by the observational technique are visual, astrometric, spectroscopic, and EB systems. We discuss each type in turn. (reference from E.F. Kallrath, J. & Milone 2009)

Figure 1.5.1. Heliocentric Julian Date Vs. magnitude curve



1.5. Concept of phase & Light Curve

In any periodic process, the same cycle repeats over and over. If we want to know what is happening at any moment, it does not matter which cycle we are observing, because every cycle is exactly the same.What does matter is which part of the cycle we are observing. So if a star is perfectly periodic, then its variation depends only on where it is in its cycle, a quantity called the **phase.** In case of astronomical data, we shall denote the time points in terms of phase. So how we exactly calculate phase of a star?

The given time points are converted into standard phases (always lie between 0 and 1) with the help of the following

$$\text{decimal portion of } \left(\frac{t - t_0}{P}\right)$$

where t is the time of having measurement on brightness of the star in the metric Heliocentric Julian Day (HJD), t_0 is the time of the first observed maximum brightness and P is the period of the star in days.

Note that, a phase of 1 (start of next cycle) is really the same as a phase of 0 (start of this cycle) or a phase of 3, or 17, or 256. Any two phases which differ by an integer are really the same phase.For example, Phase 1.5 is the same as phase 0.5.

A graph of observed magnitude versus corresponding phase is called a phase diagram. Phase diagram for phase ranging from 0.0 to 1.0 shows one complete cycle of the corresponding star. A phase of 0 is the same as a phase of 1, -1 or 2. The standard phase always lies between 0 and 1, subtracting 1 gives the previous-cycle phase within -1 and 0, or adding +1 gives the next-cycle phase. In 1.5.1 we can see in phase vs. magnitude curve has a periodicity.

1.6. Concept of Brightness Vs. Phase plot of eclipsing binary star system

Binary star orbits can be short or very long. Some stars separated by tens or 100 of billion kilometers , can take centuries to orbit each other, while some are close they may only take days. One binary star, e.g 4U 1820-30 , and

Figure 1.6.1. Phase Vs. Luminosity graph

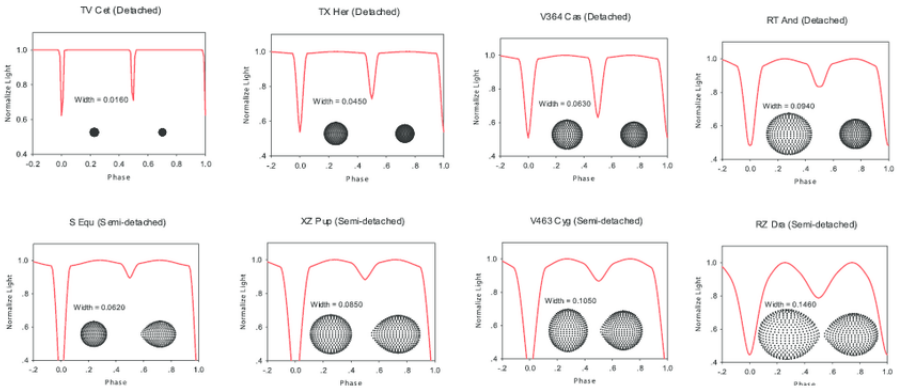
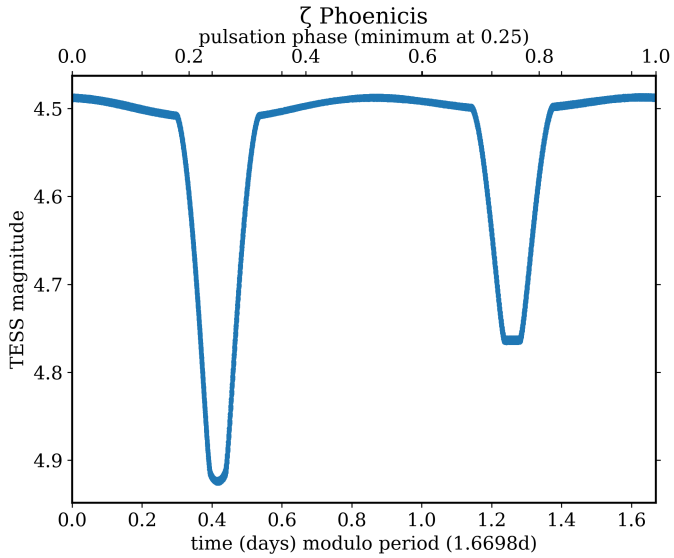


Figure 1.6.2. Algol Stars Phase Vs. Luminosity curve



it comprised of neutron star and white dwarf. They orbit each other in 685 seconds.

Normally the stars are too far away to resolve both stars. Instead we can only see the combined light of both stars. Binary stars orbits can be The light curve, which is a plot of the intensity of light over time, shows two dips in brightness if we consider one period, as shown in the 1.6.1.

Examples of prototypical light curves are shown in 1.6.2,1.6.3,1.6.4. They correspond to the classical categories, discussed above, of “Algol,” “β Lyrae,” and “W UMa” light curves, also known as EA, EB, and EW light curves, respectively.

Beta Lyre Phase Vs. luminosity curve is given below,1.6.3

Figure 1.6.3. Beta Lyre phase Vs. Luminosity Curve

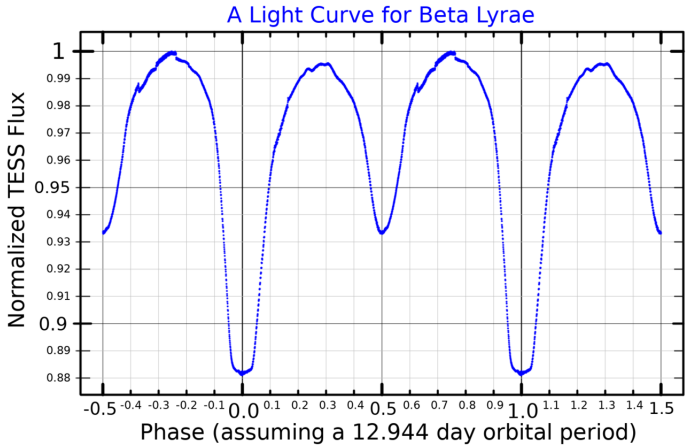
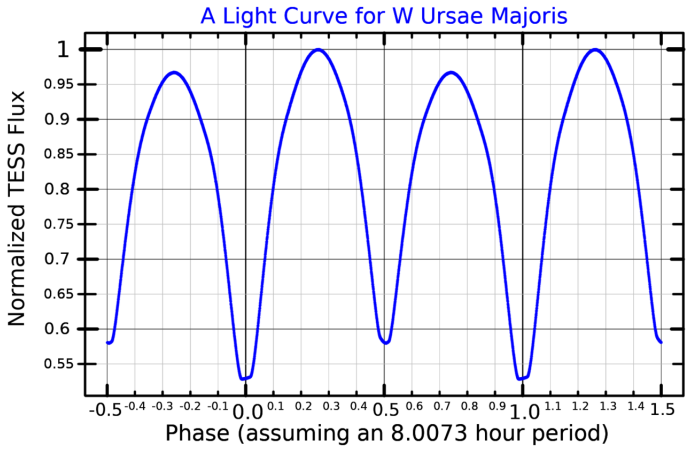


Figure 1.6.4. W Ursae Majoris Phase Vs. Brightness curve



W Ursae Majoris Phase Vs. Brightness curve is given below,1.6.4

The EA light curves are typically almost flat-topped, suggesting that effects due to the proximity of the components are small, with a large difference between the depths of the two minima. Indeed, in some wavelengths the secondary minimum may be undetectable, and there may be an increase in light near the expected phase of secondary minimum due to the “reflection effect.”

The EB light curves, on the other hand, are continuously variable (the “ellipsoidal variation”11), characteristic of tidally distorted components, and with a large difference in depths of minima indicating components of quite different surface brightness.

Finally, the EW (or W UMa) light curve is also continuously variable, but with only a small difference in the depths of the minima. The variation outside the eclipse in the latter two types is indeed due to proximity effects (mainly

the tidally distorted shapes of the stars), but the EB light curves arise from detached12 or semi-detached binaries, whereas EW systems are over-contact.13 The expression “EA light curve,” on the other hand, is somewhat misleading. Judged by the light curve, the system may look undistorted, but only in light from the visible (or, as infrared astronomers refer to it, the “optical”) part of the spectrum. In the infrared, for example, Algol itself presents a continuously variable light curve and a fairly deep secondary minimum. This reveals quite clearly that the bluer, hotter component in the system is relatively small and undistorted, and its radiation enhances the bright inner face of its companion.

1.7. What can be derived from eclipsing binaries

According to Kallrath, Josef and Milone, Eugene F and Wilson,

- We can roughly estimate the dimensions of binary stars and their orbits.
- The analysis of photo metric light curve can provide orbital inclination, relative quantities such as the radii, ratio of luminosity, stellar figures and perhaps photo metric mass ratio.
- We also can roughly estimate the distance or parallaxes of stars. They can also tell us about the stellar evolution of the stars.

1.8. The Database & The Data Acquisition

1.8.1. Photoelectric Photometry. The development of photoelectric devices has occasioned considerable changes in all types of instrument in which intensities are measured or compared. These devices are directly sensitive to flux and the optical systems used with them must be designed accordingly. At very low intensities the photo-multiplier cell allows a light flux of only a few quanta per second to be measured and the theoretical limitations and practical arrangements are discussed. In the ultra-violet spectrum, the photoelectric technique must be revised, but efficient detectors for the whole region are known and some are described. Infra-red detection requires recourse to the ‘internal photoelectric effect in semiconductors, and the properties of the most important detectors of the class are briefly described.

A characteristic advantage of photoelectric detectors is that their output is linearly related to the intensity of the incident light and is in a form suitable for electronic data processing by analogue or digital methods. Some examples in industrial and astronomical photometry which make use of this property are described, and there are some notes on precision photo-electric photometry as used in the maintenance and use of standards of radiation.[**1**]

1.8.2. Two-Star Photometers. The introduction of computer data acquisition and instrument control in the late 1960s and early 1970s made possible important new advances in high-speed stellar photometry, notably at McDonald Observatory and at Princeton Observatory . This technology permitted routine observations of such rapid events as white-dwarf oscillations, flickering of cataclysmic variables , and lunar occultations . Significant research in

these areas is possible with telescopes of less than one-meter aperture. Unfortunately, however, many such telescopes are located at sites where photometric weather conditions are rare. An example is the 0.9-m reflector of the Louisiana State University (LSU) Observatory, located about 40 km north of Baton Rouge. A photometer that incorporates a second channel to monitor a nearby constant comparison star makes precision high-speed photometry possible on many nights that are clear but not of photometric quality because of variable sky transparency. Such conditions occur often at the LSU Observatory, and of course can also occur even at good sites like Kitt Peak National Observatory (KPNO). Astronomers at several observatories have constructed two-star photometers. Descriptions of these instruments have been given by Blitzstein (see Wood 1967), Warner (1971), Warner and Nather (1972), Goudis and Meaburn (1973), Geyer and Hoffmann (1975), Furlani and Sedmak (1977), Bernacca et al. (1978), and De Biase et al. (1978). However, relatively little detailed information on the successful use of these instruments in the two-star mode for astronomical observations, or on data-reduction techniques, has appeared in the literature.[**2**]Most depend on having separate light path detectors, or electronics for the different channels. The one with which we are most familiar does not. It is the Rapid Alternate Detection System (RADS), used at the University of Calgary's Rothney Astrophysical Observatory since about 1981 (Milone et al. 1982).

This system 1.8.1 employs a single pulse-counting detector and a swiveled secondary mirror which is driven by the dial-in settings of a function generator. The amplitude of the throw, the duty cycle, and the delay time for mirror settling can be entered separately for each of four positions. The delay time depends on the aperture, because a smaller aperture requires more stability for the image as the mirror ringing dies down. The sum of the duty cycles and time delay determines the chopping frequency, which may be as high as 20 or 30 Hz; the system is usually operated closer to 1 Hz, however, because of the overhead caused by image motion and the delays at each station. The chopping line of the mirror may be rotated to coincide with the line between two stars.

1.8.3. Data source and variables.

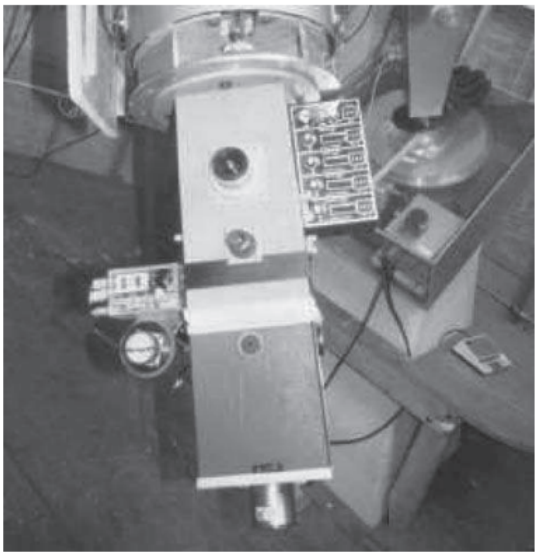
1.8.3.1. *Data:* This data set is taken from "Miller et. al..(2010)"[**7**] [**8**]where they have carried out observations of 1318 new variable stars covering 0.25 square degree region of the galactic plane centered on galactic coordinates of (330.94-2.28) degree.

1.8.3.2. *Variables:*

1st: data file consists the following variables

- ID - 1:1318 variables star number
- chip - CCD chip and sub chip
- RAh - Right ascension (J2000) (h)
- RAm - Right ascension (J2000) (min)
- RAs - Right ascension (J2000) (s)
- DE- Declination sign (J2000)
- DEd - Declination (J2000) (deg)

Figure 1.8.1. RADS instrument. Shown is the controller for the RADS. It consists of a function generator which controls the successive positions of the secondary mirror within a cycle involving successive settings on the program star, sky near the comparison star, the comparison star, and the sky near the program star



- DEm - Declination (J2000) (arcmin)
- DEs - Declination (J2000) (arcsec)
- Rcmag - WFI Cousins R magnitude (mag)
- B-Rc - WFI Johnson-Cousins B-R colour index (mag)
- Rc-Ic - WFI Cousins R-I colour index (mag)
- Per - Period (d)
- Type - Variable star type
- FileName - Name of the file containing the light curve in subdirectory photo

2nd: data set have

- HJD - Heliocentric Julian Date
- df/f - Relative flux variation in R-band
- e_df/f - ? Uncertainty in relative flux variation

for the i^{th}Star system. where i=1,2,...., 1318. For all stars the observed LC s(Relative flux variation on R-band over time measures in Heliocentric Julian Date) are unevenly spaces in different length (Ranging from 130-264 except one LC having length 5) with values at different time points and the period varying from several hours to several weeks.

1.9. Outline of this project

In this project we look at the light curves of eclipsing binary stars, building on the work of past researchers in this field. The abundance of the data from recent sky survey requires new analysis methods, which is where astrostatistics plays a key-role. In this project we are specifically studying how to classify light curves from eclipsing binary star system using different clustering techniques. By applying new approach to classify light curves we can reveal new insights and connection, providing better understanding.

In chapter 1 we are discussing the the basic concept related to binary star system ,phases, phase Vs. brightness curve, Database and data acquisition process . In chapter 2. we will discuss the approach done by Modak.et.al(2019) using interpolation and binning and it's implementation on the data set. We will also discuss the problem and setbacks with this approach. In chapter 3 we introduce our own approach to aid the problems discussed in the previous one by implementing a parametric model and further implementing it to the model to get a better result than the previous.

CHAPTER 2

Naive Approach

2.1. Introduction

In the study , Modak.et.al.[**8**] they present an advance approach to classifying variable stars following k-mediods clustering method,for 1318 light curve. Following their work in this chapter we further explore the classification of variable stars.

2.2. Data pre processing using interpolation & binning:

From the paper Modak et. al. [**8**] we perform the data pre processing in the data set.

Step: 1: As there are different lengths of the Light Curves having values at different time points, comparison of the Light Curves are only possible in terms of observations over each cycle. To tell exactly what the shape of a cycle is, all the cycles could be superimposed on top of each other. Hence each data point can be plotted, but instead of plotting the time, we would like to plot "how far it is into the cycle". That way, all the cycles will be "folded" on top of each other, and we may have enough data to give us an accurate picture of what the cycle looks like. For a LC "how far it is into the cycle" is termed as its phase. So if the period of a star is known, and constant, it is possible to define phase the fraction of the star's variability cycle which has elapsed. Using 1.5 on page 11

$$\text{decimal portion of } \left(\frac{t - t_0}{P}\right)$$

where, the symbols follows it's usual meaning.

Step: 2: Step 1 gives the i^{th} LC over phase interval [0, pi], with pi as the maximum of standard phases for the i^{th} Light Curve. For clustering these times series, using the distance measure, we need to get a full cycle phase over 0 to 1 for each star having values at the same and equidistant phase points. For this purpose we follow the following steps [**8**]:

(1) After standardizing using the time points in days converted into phases over the range of [0,1]. which results in 1318 star variable with light curves of different length over different phase range.

(2) we also consider the next phase for each stars. That means we are considering two phases (the standard phase and the immediate next one) for each star. i.e. we are considering the phase [0,1] and phase

[1,2], the immediate next phase. For the next phase the light curve value will remain same as they are light curve of same stars, which will be same same for every phase interval. Now we have total phase value of two periods ranging from [0,2), for $i = 1, 2, ..., .1318$ and their light curves.

(3) For the sake of analysis we need equidistant points. So, we fix the no. of equidistant points for each stars to be $l^* = 272$(why?) so that there is no loss of information and a length of 272 is not too large to increase the computational burden.

(4) Finally, piece wise linear interpolation is applied to the observations for given phase point[**9**].

Step: 3: We fit the linear spline to the Light Curves from step (ii) at l ´ evenly spaced phases over [0,1] using the following formulas:

(1) For star with the i^{th} LC, we have l_i values of the brightness function y against different values of phase point $p\prime$ as $y_j = y(p\prime_j), j = 1, ..., l_i$with $p'_j < p'_{j+1}$ for $j = 1, ..., l_i - 1$.

(2) The interpolation function joins the following $(l_i - 1)$ linear functions

$$g_j(p') = a_j y_j + b_j y_{j+1}, for p' \in [p'_j, p'_{j+1}], j = 1, ..., l_i - 1$$

a_j**:** and $b_j$are constants which satisfy the following: $g_j(p'_j) = y_j$and $g_j(p'_{j+1}) = y_{j+1}$for $j = 1, ..., l_i - 1$ i.e.

$$a_j = \frac{p'_{j+1} - p'}{p'_{j+1} - p'_j}$$

and

$$b_j = 1 - a_j$$

Step: 4: Finally, the lengths of all Light Curves are fixed at $l' = 272$, after a trade-off between extraction of relevant information from the Light Curves and the interpolation error in approximating the Light Curves, which produced the best possible clustering results .

2.3. Concept of k-medoids and Silhouette Plot

2.3.1. K-Medoids algorithm: The K-medoids algorithm[**6**] similar to the K-means algorithm, except when fitting the centers c1, c2, ..., ck , we restrict our attention to the points themselves.We start with an initial guess for c_1, c_2, ..., c_k (e.g., pick K points at random over the range of x_1, x_2, ..., x_k), then repeat:

(1) Minimize over C: for each $i = 1, .., n$, find the cluster center c_k closest to x_i , and let $C(i) = k$

(2) Minimize over $c_1, c_2, ..., c_k$: for each $k = 1, ..., K$, let $c_k = x^*_k$, the medoid of points in cluster k, i.e., the point x_i in cluster k that minimizes that minimizes $\sum_{C(j)=k} ||x_j - x_i||_2^2$

Stop when the within cluster variation doesn't change.

2.3.2. Silhouette Width: For each observation i, the Silhouette Width is defined as follows[**10**]:

Let us first define the numbers $s(i)$ in the case of dissimilarities. Take any object i in the data set, and denote by A the cluster to which it has been assigned. When cluster A contains other objects apart from i, then we compute,

$a(i)$ =Average dissimilarity of i to all objects of C (another cluster).

$d(i, C)$ =Average dissimilarity to i to all other objects of C.

After computing $d(i, c)$ for all cluster $C \neq A$,we select the smallest of those numbers and denote it by,

$b(i) = \underset{C \neq A}{minimum} d(i, C)$

The number $s(i)$is obtained by combining $a(i)$ and $b(i)$ as follows:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & if, a(i) < b(i) \\ 0 & if, a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & if, a(i) > b(i) \end{cases}$$

In one formula:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

When cluster A contains only a single object it is unclear how $a(i)$ should be defined, and then we simply set $s(i)$ equal to zero. This choice is of course arbitrary, but a value of zero appears to be most neutral. Indeed, from the above definition we easily see that

$$-1 \leq s(i) \leq 1$$

for each object i.

As a tool of clustering diagnostic we use Silhouette Coefficient defined as

$$SC = \underset{k}{max} \bar{s_k}$$

SC	Interpretation
0.71-1.00	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial
≤0.25	No substantial structure has been found

2.4. Implementation & Results

(1) Dendogram for the data:
(2) For K=2 the Silhouette plot :
(3) For K=3 the silhouette plot:
(4) Average Silhouette plot for every K:

In this approach we find the average silhouette with is 0.62 for k=2 in k-means approach. So, A reasonable structure has been found. There are 2 clusters in the data.

Figure 2.4.1. Dendogram

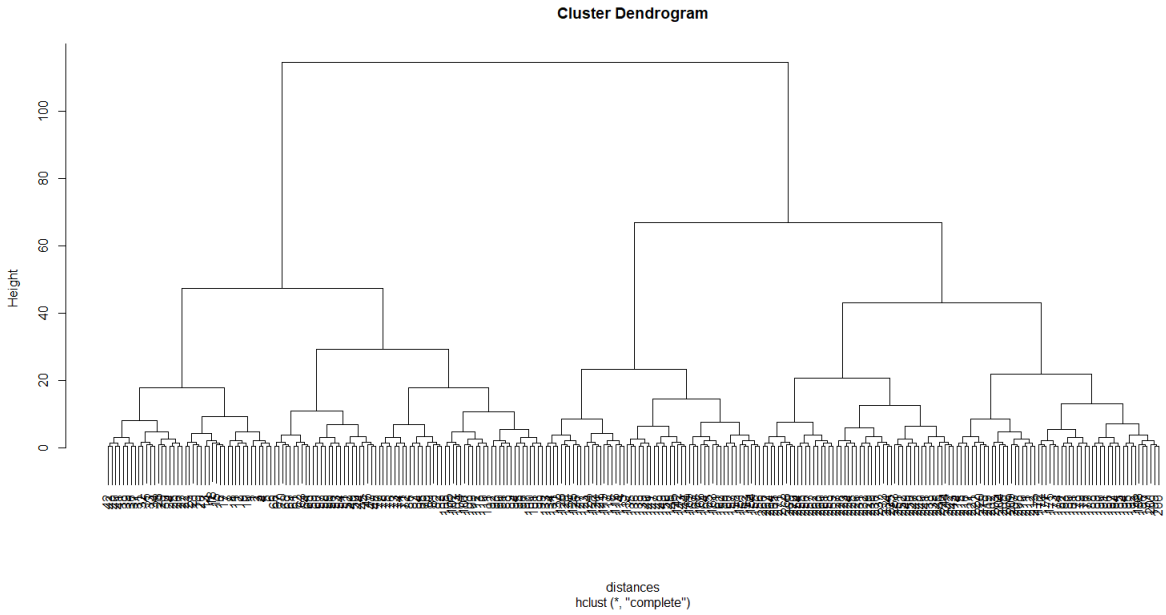
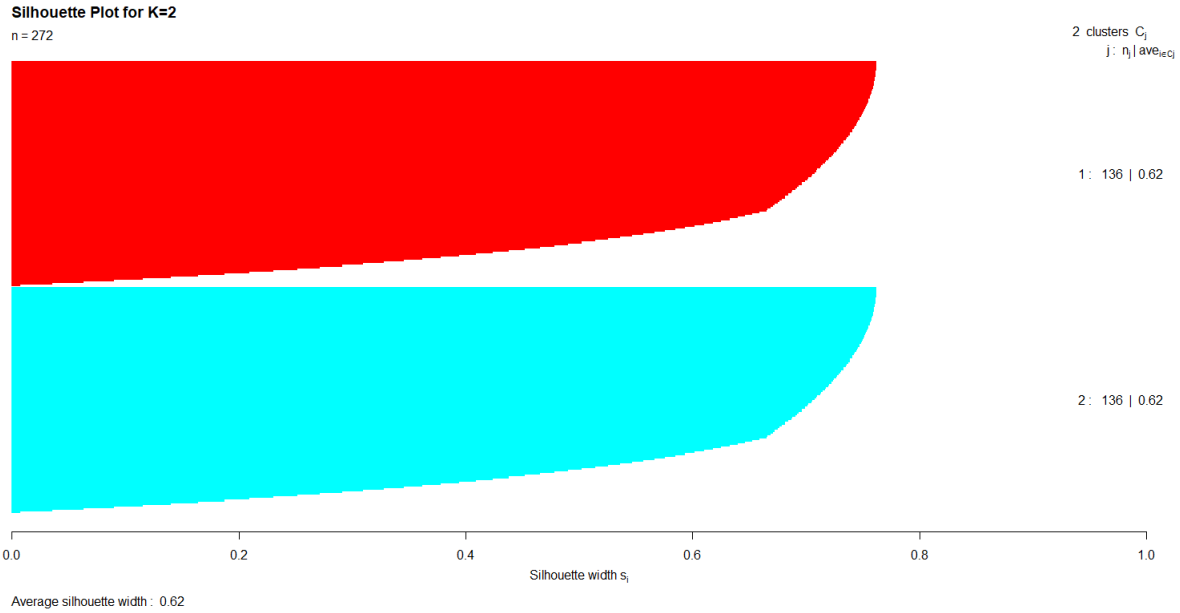


Figure 2.4.2. For k=2 silhouette plot



2.5. Problem with this approach

(1) If we assume the 272 data generated from the interpolation formulae then we are assuming a specific form (linear Spline) for the light

Figure 2.4.3. For k=3 silhouette plot

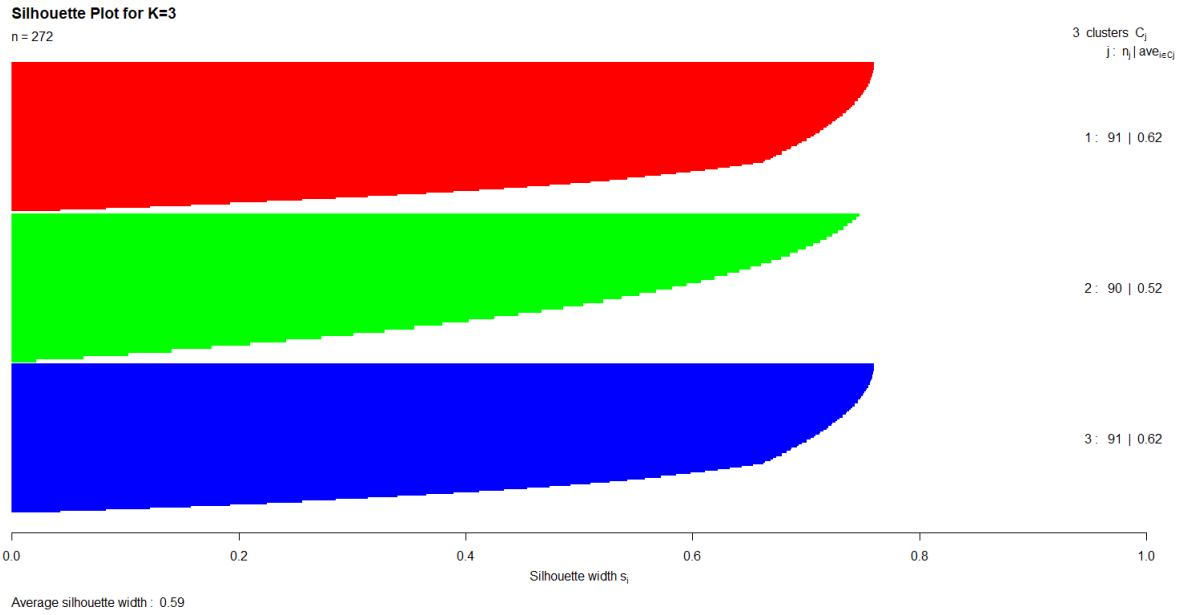
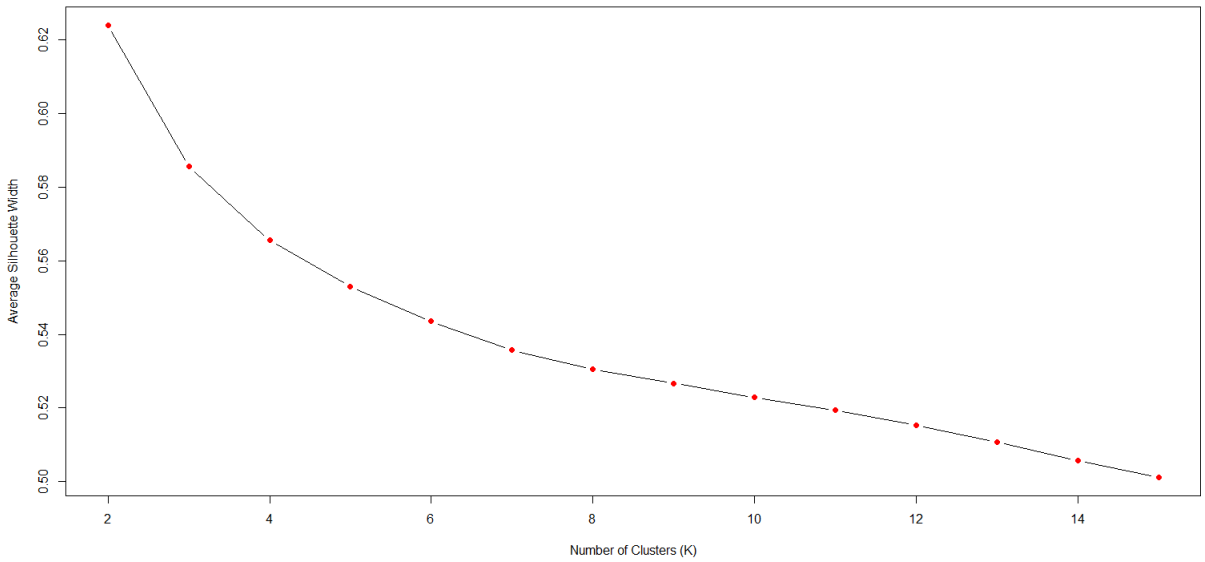


Figure 2.4.4. Average Silhouette plot for every K



curve. **Then assuming Light curves to be unknown does not make sense**.

(2) We are clustering 272 sample points in a way such that they are independent. **This is completely ignoring the mutual dependency of the 272 data points corresponding to every star**.
(3) **Why extending the phase form [0,1] to [0,2] ?** The Y values are taken to be periodic and hence we repeat the same Y observations. However, the actual Y values are never observed and what we get to see are Y values contaminated with errors. So, instead of repeating the Y values it would be better to use a new series of observations, with separate set of errors.

CHAPTER 3

Our Approach

3.1. Phase Extension: A new Idea

We assume the observed data are always contaminated with errors.

Hence, when, $t \in [0, 1]$

$$Y_t = \mu_t + \epsilon_t$$

where, μ_t =true signal,

Y_t =Observed Signal

ϵ_t =error

However, when $t^{'} \in [1, 2]$

we assume,

$$Y_{t'} = Y_t + \epsilon_{t'}$$

where, $\epsilon_t' \sim N(0, 0.00001)$. This means

$$Y_{t'} = \mu_t + \epsilon_t + \epsilon_{t'}$$

$$= \mu_t + \eta_{t'}$$

Hence, between [0,1] we observe $\mu_t + \epsilon_t$ and in [1,2] we observe $\mu_t + \eta_t$.

3.2. Parametric Model

We note that the given data satisfies the setup of functional data; the light curves corresponding to every star form a different function, which are unknown to us. We only have certain observations (not equidistant) corresponding to different light curves. The traditional approach assume non-parametric versions of the light curves.

However, in our case, we know that the light curves are periodic with respect to the phases we shall assume a parametric form for each light curve. To aid in understanding this pattern, we include a random observation's phase vs. light curve (3.2.1).This graph illustrates the typical sinusoidal nature of the light curve across the different phases.

More specifically, for the i^{th} Light curve, we assume,

$$Y_t^{(i)} = f_i(x_t) = a_i + b_i sin(\gamma_i x_t) + \epsilon_t^{(i)}$$

$\forall t = 1, 2, ..., n_i, \forall i = 1, 2, ..., 1318$

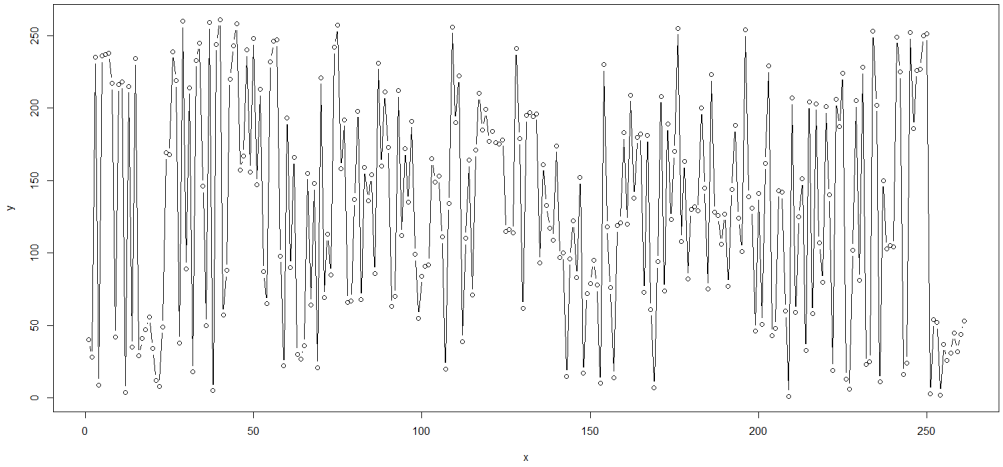
where,

x_t = the phase of the i^{th} star

$f(x_i)$ =represents the observed light curve.

a_i = is a constant that shifts the function vertically

Figure 3.2.1. Phase Vs. Light curve graph of a random observation in the data



b_i =Scaling factor

γ_i = $Scaling$factor that adjust the frequency of the sine wave to match the periodicity of the light curve.$\forall i = 1(1)1318$

$\epsilon_t^{(i)} = error$

By employing this model, it becomes possible to not only describe the observed data with greater precision but also to make predictions and derive further insights into the behavior of the variable being studied. The fitting process and subsequent analysis are detailed in the following sections.

3.3. Estimating the parameter

Nonlinear regression models are of the form:

$$Y_i = f(X_i, \gamma) + \epsilon_i \tag{3.3.1}$$

An observation $Y_i$is still the sum of a mean response $f(X_i, \gamma)$ given by the nonlinear response function $f(X, \gamma)$ and the error term ϵ_i. The error terms usually are assumed to have expectation zero, constant variance, and to be uncorrelated, just as for linear regression models. Often, a normal error model is utilized and assumes that the error terms are independent normal random variables with constant variance.

The parameter vector in the response function $f(X, \gamma)$is now denoted by γrather than βas a remainder that the response function here is nonlinear in the parameter. We present now two examples of nonlinear regression models that are widely used in practice.

3.3.1. Nonlinear Regression: To obtain the normal equation for a nonlinear regression model:

$$Y_i = f(X_i, \gamma) + \epsilon_i \tag{3.3.2}$$

we need to minimize the least squares criterion Q:

$$Q = \sum_{i=1}^{n} [Y_i - f(X_i, \gamma)]^2 \tag{3.3.3}$$

with respect to $\gamma_0, \gamma_1, ..., \gamma_{p-1}$.The partial derivative of $Q$with respect to $\gamma_k$is:

$$\frac{\partial Q}{\partial \gamma_k} = \sum_{i=1}^{n} -2[Y_i - f(X_i, \gamma)] \left[\frac{\partial f(X_i, \gamma)}{\partial \gamma_k} \right] \tag{3.3.4}$$

When the p partial derivatives are each set equal to 0 and the parameters γ_k are replaced by the least squares estimates g_k, we obtain after some simplification the p normal equations:

$$\sum_{i=1}^{n} Y_i \left[\frac{\partial f(X_i, \gamma)}{\partial \gamma_k} \right]_{\gamma=g} - \sum_{i=1}^{n} f(X_i, g) \left[\frac{\partial f(X_i, \gamma)}{\partial \gamma_k} \right]_{\gamma=g} = 0, k = 0, 1, ...p - 1$$

where g is the vector of the least square estimates g_k :

$$g = \begin{bmatrix} g_0 \\ g_1 \\ : \\ g_{p-1} \end{bmatrix}$$

The normal equations for nonlinear regression models are nonlinear in the parameter estimates g_k and are usually difficult to solve, even in the simplest of cases. Hence, numerical search procedures are ordinarily required to obtain a solution of the non normal equations iteratively.

3.3.1.1. *Numerical Search-Gauss-Newton Method:* In many nonlinear regression problems, it is more practical to find the least squares estimates by direct numerical search procedures rather than by first obtaining the normal equations and then using numerical methods to find the solution for these equations iteratively. The major statistical computer packages employ one or more direct numerical search procedures for solving nonlinear regression problems. We now explain one of these direct numerical search methods. The **Gauss-Newton method**, also called the linearization method, uses a Taylor series expansion to approximate the nonlinear regression model with linear terms and then employs ordinary least squares to estimate the parameters. Iteration of these steps generally leads to a solution to the nonlinear regression problem.

The Gauss-Newton method begins with initial or starting values for the regression parameters $\gamma_0, \gamma_1, ..., \gamma_{p-1}$.We denote these by $g_0^{(0)}, g_1^{(0)}, ..., g_{p-1}^{(0)}$, where the superscript in parentheses denotes the iteration number. The starting values $g_k^{(0)}$may be obtained from previous or related studies, theoretical expectations, or a preliminary search for parameter values that leads to low criterion value in Q.

Once the starting values for the parameters have been obtained, we approximate the mean responses $f(X_i, \gamma)$ for the n cases by the linear terms in the Taylor series expansion around the starting values $g_k^{(0)}$. We obtain for the ith case:

$$f(X_i,\gamma) \approx f(X_i,,g^{(0)}) + \sum_{i=1}^{p-1}\left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_k}\right]_{\gamma=g^{(0)}} (\gamma_k - g_k^{(0)}) \tag{3.3.5}$$

,where;

$$g_{px1}^{(0)} = \begin{bmatrix} g_0^{(0)} \\ g_1^{(0)} \\ : \\ g_{p-1}^{(0)} \end{bmatrix} \tag{3.3.6}$$

the $g^{(0)} is$the vector of the parameter starting values.

Let us now simplify the notation as follows:

$$f_1^{(0)} = f(X_i, g^{(0)})$$

$$\beta_k^{(0)} = \gamma_k - g_k^{(0)}$$

$$D_{ik}^{(0)} = \left[\frac{\partial f(X_i,\gamma)}{\partial \gamma_k}\right]_{\gamma=g^{(0)}}$$

The Taylor approximation for the mean response for the i th case then becomes in this notation :

$$f(X_i,\gamma) \approx f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)}\beta_k^{(0)}$$

and an approximation to the non linear regression model:

$$Y_i = f(X_i,\gamma) + \epsilon_i$$

is:

$$Y_i \approx f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)}\beta_k^{(0)} + \epsilon_i$$

when we shift the $f_i^{(0)}$term to the left and denote the difference $Y_i - f_i^{(0)}$by $Y_i^{(0)}$,we obtain the following linear regression model approximation:

$$Y_i^{(0)} \approx \sum_{k=0}^{p-1} D_{ik}^{(0)}\beta_k^{(0)} + \epsilon_i, i = 1, ..., n$$

where:

$$Y_i^{(0)} = Y_i - f_i^{(0)}$$

Note that the linear regression model approximation is of the form:

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

We shall represent the regression model approximation in3.3.5 matrix form as follows:

$$Y^{(0)} \approx D^{(0)}\beta^{(0)} + \epsilon$$

where:

$$Y^{(0)}_{nx1} = \begin{bmatrix} Y_1 - f^{(0)}_1 \\ : \\ Y_n - f^{(0)}_n \end{bmatrix} \tag{3.3.7}$$

$$D^{(0)}_{nxp} = \begin{bmatrix} D^{(0)}_{10} & \dots & D^{(0)}_{1,p-1} \\ : & & : \\ D^{(0)}_{n0} & \dots & D^{(0)}_{n,p-1} \end{bmatrix} \tag{3.3.8}$$

$$\beta^{(0)}_{px1} = \begin{bmatrix} \beta^{(0)}_0 \\ : \\ \beta^{(0)}_{p-1} \end{bmatrix} \tag{3.3.9}$$

$$\epsilon_{nx1} = \begin{bmatrix} \epsilon_1 \\ : \\ \epsilon_n \end{bmatrix}$$

Note again that the approximation model is precisely in the form of the general linear regression model , with the D matrix of partial derivatives now playing the role of the X matrix . We can therefore estimate the parameters $\beta^{(0)}$ by ordinary least squares :

$$b^{(0)} = (D^{(0)\prime} D^{(0)})^{-1} D^{(0)\prime} Y^{(0)}$$

where $b^{(0)}$is the vector of the least squares estimated regression coefficients.As we noted earlier, an ordinary multiple regression computer program can be used to obtain the estimated regression coefficients $b^{(0)}_k$ with a specification of no intercept. We then use these least squares estimates to obtain revised estimated regression coefficients $g^{(0)}_k$by means of 3.3.7:

$$g^{(1)}_k = g^{(0)}_k + b^{(0)}_k$$

where $g^{(1)}_k$denotes the revised estimate of $\gamma_k$at the end of the first iteration. In matrix form, we represent the revision process as follows:

$$g^{(1)} = g^{(0)} + b^{(0)}$$

3.3.2. Estimating In the Model: our non-linear equation here is

$$\underset{\sim}{Y} = f(\underset{\sim}{X}, \underset{\sim}{\gamma}) + \underset{\sim}{\epsilon}$$

where $\underset{\sim}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ : \\ y_{n_i} \end{pmatrix}, n_i = varies$ from 136 to 264 (except on with 5 observation)

= It is the observed light curve for the stars.

$f(\underset{\sim}{X}, \underset{\sim}{\gamma}) = a_i + b_i sin(\gamma_i x_i), \forall i = 1(1)1318$, $\underset{\sim}{X} = Phase$of the star,$\underset{\sim}{\gamma} = \begin{pmatrix} a_i \\ b_i \\ \gamma_i \end{pmatrix}$

$\underset{\sim}{\epsilon} = \left(\left(\epsilon_{in_i}\right)\right), \forall i = 1(1)1318, \forall n_i = depending$ on the observation.

where $\epsilon_{ij} \sim N(0, 0.000001)$
we need to minimize the least squares criterion Q:

$$Q = \sum_{i=1}^{n} [Y_i - f(X_i, \gamma)]^2 3.3.3$$

The partial derivative of $Q$with respect to $\underset{\sim}{\gamma}$is:

$$\frac{\partial Q}{\partial a} = \sum_{i=1}^{j} (y_i - a - bsin(\gamma x_i)) = 0 \tag{3.3.10}$$

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{j} sin(\gamma x_i)(y_i - a - bsin(\gamma x_i)) = 0 \tag{3.3.11}$$

$$\frac{\partial Q}{\partial \gamma} = \sum_{i=1}^{j} (y_i - a - bsin(\gamma x_i)) bx_i cos(\gamma x_i) = 0 \tag{3.3.12}$$

Then we solve these 3 normal equation using Gauss-Newton method . The computation is done in R- software here.

Here is the glimpse of the $\underset{\sim}{\hat{\gamma}}$for first 10 stars out of 1318 stars.

	a_i	b_i	c_i
1	99.770988	0.03853139	2.346781
2	1.135201	1.63915990	2.912423
3	95.998171	0.02416219	3.143315
4	1.363225	1.22299804	3.048303
5	100.856572	0.01004527	2.965450
6	99.885962	0.02800547	2.398336
7	1.008861	2.04489236	3.043426
8	97.611656	0.01357525	3.185803
9	98.225108	0.02078968	3.401398
10	98.718997	0.01725762	2.708231

Now we have in total 1318 functions $f(x_i) = \hat{a_i} + \hat{b_i} sin(\hat{\gamma_i} x_i)$ for each stars.

3.4. Calculating Distance Matrix

Suppose the i^{th} and j^{th}fitted light curve are given by $f_i(x)$ & $f_j(x)$ respectively. Then the distance between the i^{th}& j^{th} light curves, given by

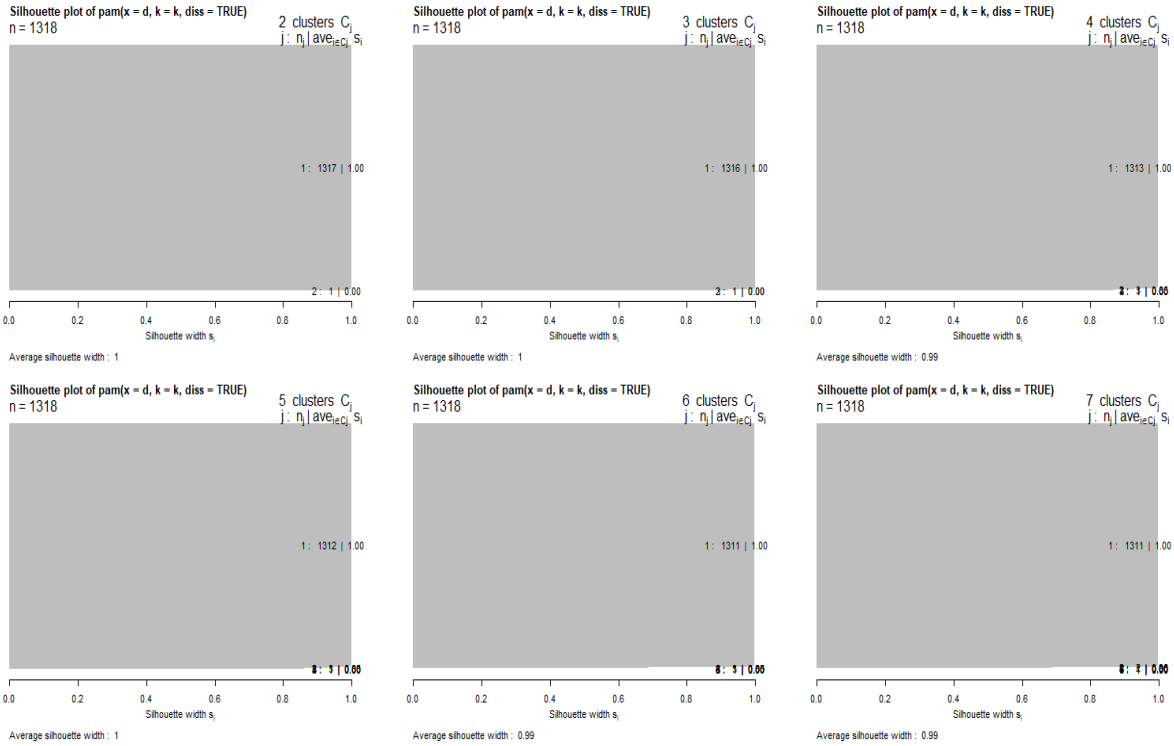
$$d_{ij} = \int_0^2 (f_i - f_j)^2 dt$$

$$= \int_0^2 [(\hat{\alpha_i} - \hat{\alpha_j}) + (\hat{\beta_i} sin(\hat{\gamma_i} t) - \hat{\beta_j} sin(\hat{\gamma_j} t)]^2 dt$$

3.5. Clustering Partitioning around medoids (PAM)

Our final objective is to cluster the 1318 estimated light curves. In this setup, we could have used $K-$means algorithm for the clustering purpose. However, that would require defining the average light curve for each cluster. Hence we avoid that task by implementing partitioning around mediod (PAM) algorithm, where the cluster center is selected as one among the given light curves. The choice of K, the number of clusters, however, still remains to be decided. Traditionally, in the absence of preconceived value of K, the average silhouette width can be considered as an objective criterion for selecting K. But in our case, as we shall see, using silhouette width gives misleading conclusions. In figure 3.5.1 and figure 3.5.2, we create the silhouette plot for K ranging from 2 to 20 using PAM. We find that for $K = 1, 2, ..., 15$, the average silhouette width is 1 and in each case, there is only major cluster and other $K - 1$ clusters each containing single observation. When we further increase K, we find although the average silhouette width slightly goes down, but there are many clusters with single observations. The reason for these high values of average silhouette width is too many clusters with single observations where the within cluster dissimilarity becomes 0. On inspecting all the plots, we find that there are around 16 to 18 outliers each one of which are forming single clusters. The remaining data are then clustered primarily into two clusters. In this case, we shall consider a clustering to be valid if there are no negative silhouette coefficient.

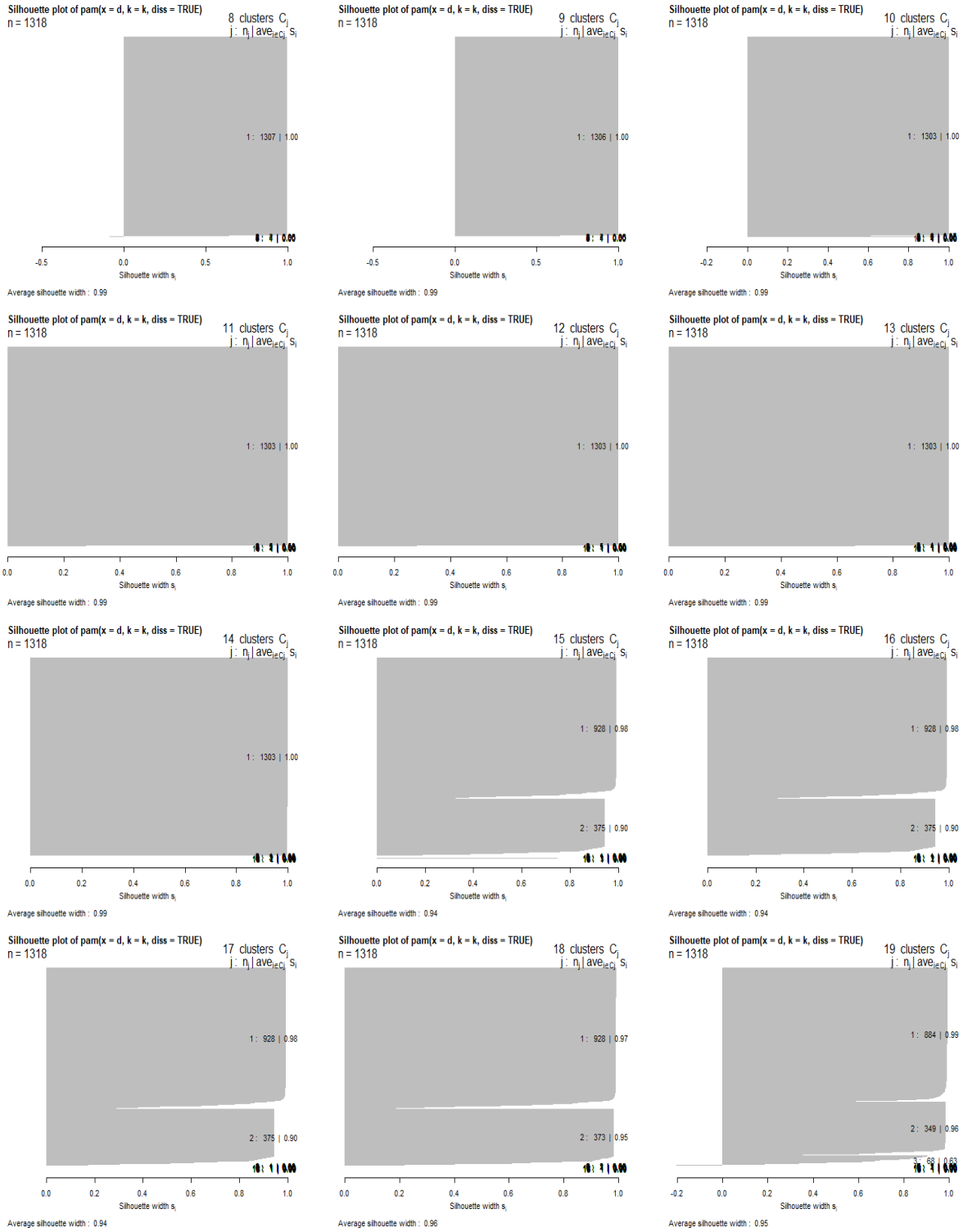
Figure 3.5.1. Silhouette plot for k= 2to 7



3.6. Conclusion

This parametric approach helps us to classify the light curves of 1318 eclipsing binary stars. In this approach, we identified two significant clusters which is consistent with the findings of all the previous works in this regard. However, we also detected the presence of some outliers in the data effectively. Hence we can conclude, that the given 1318 stars can be categorized primarily into two groups and there are around 18 binary stars which can be considered as extreme observations.

Figure 3.5.2. Silhouette Plot for k=8 to 19



Acknowledgment:

It’s my pleasure to acknowledge assistance of a number of people without whom this project would not have been possible. First I would like to express my gratitude to **Dr. Atanu Kumar Ghosh** (Project Supervisor), Assistant Professor of our Statistics Department; for providing invaluable guidance & support. I pay my my humble regard & gratitude to Associate Professor and Head of the department Statistics , **Shree Biswajit Roy**, for extending his helping throughout the journey. After the completion of the project I can confidently say that, this experience not only enriched me with the knowledge but also has matured the thought and vision.

Angshumita Sarkar

Bibliography

[1] H J J Braddick. Photoelectric photometry. *Reports on Progress in Physics*, 23(1):154, jan 1960.

[2] ALBERT D Grauer and HOWARD E Bond. Two-star high-speed photometry. *Publications of the Astronomical Society of the Pacific*, 93(553):388, 1981.

[3] Gerhart Hoffmeister. *Petrarkistische Lyrik*. Springer, 1973.

[4] Steve B Howell, Cassandra VanOutryve, John L Tonry, Mark E Everett, and Raelin Schneider. A search for variable stars and planetary occultations in ngc 2301. ii. variability. *Publications of the Astronomical Society of the Pacific*, 117(837):1187, 2005.

[5] E.F. Kallrath, J. Milone. *Eclipsing Binary Stars: Modeling and Analysis*. Astronomy and Astrophysics Library. Springer New York, 2009.

[6] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[7] VR Miller, MD Albrow, C Afonso, and Th Henning. 1318 new variable stars in a 0.25 square degree region of the galactic plane. *Astronomy & Astrophysics*, 519:A12, 2010.

[8] Soumita Modak, Tanuka Chattopadhyay, and Asis Kumar Chattopadhyay. Unsupervised classification of eclipsing binary light curves through k-medoids clustering. *Journal of Applied Statistics*, 2019.

[9] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. Numerical recipes in c, 1992.

[10] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[11] Helen Rovithis-Livaniou. Study of eclipsing binaries: Light curves o-c diagrams interpretation. *Galaxies*, 8(4), 2020.

[12] Wayne H Warren Jr. General catalogue of variable stars 4 th edition (kholopov et al. 1985 {88), year = 1988, ranking = rank3,.

[13] Wikipedia contributors. W ursae majoris — Wikipedia, the free encyclopedia, 2023. [Online; accessed 4-May-2024].

[14] Wikipedia contributors. Algol — Wikipedia, the free encyclopedia, 2024. [Online; accessed 4-May-2024].

[15] Wikipedia contributors. Beta lyrae — Wikipedia, the free encyclopedia, 2024. [Online; accessed 4-May-2024].

[16] H Zinnecker, BA Wilking, A Duquennoy, and M Mayor. Binaries as tracers of stellar formation, 1992.

•

Appendix(R code):

```
# Reading the file called table2.DAT
data<-read.table("C:\\Users\\ANGSHUMITA\\Desktop\\project work\\J_A+A_519_A12.tar\\J_A+A_5
#renaming the column in file 1
# naming the variable
ID<-data$V1 ;chip<-data$V2; RAh<-data$V3; RAm<-data$V4 ;RAs<-data$V5 ;DEd<-data$V6 ;DEm<-d
# making a dataframe
df2<-data.frame(ID,chip,RAh,RAm,RAs,DEd,DEm,DEs,Rcmag,B_RC,RC_IC,Period,type,Filename)
#making a data frame with the working variable
####
add <- paste0("C:\\Users\\ANGSHUMITA\\Desktop\\phot\\", df2$Filename)  #pasting the file i
 #### # Data processing in Naive approch
DAT <- list(NULL)
for(i in 1:length(add))
{
DAT[[i]] <- read.table(add[i])
print(paste("Successful for", i)) #seeing the missing values and filling tit with null val
}
# replacing the missing info in a particular column with "0000" to read the data #vizualiz

#step (i)
LOL <- list(NULL)
for(i in 1:length(add))
{
LOL[[i]] <- ((DAT[[i]][,1]-floor(DAT[[i]][, 1]))/df2$Period[i])%%1# t-t_0/p_i
}
#step (ii)
LOL1 <- list(NULL)
for(i in 1:length(add))
{
LOL1[[i]] <- (LOL[[i]]+1)# extending p=[0,1]->p=[0,2)
}
#step (iii)
gol <- function(p,i)
{

p_order=c(LOL[[i]],LOL1[[i]])
```

```r
j <- max(unlist(lapply(p_order, function(x) which(x < p))))
pj= p_order[j]
pj1=p_order[j+1]
aj <- (pj1 - p) / (pj1 - pj)
bj <- 1-aj
y=c(DAT[[i]]$V2,DAT[[i]]$V2)
yj=y[j]
yj1=y[j+1]
return(aj*yj + bj*yj1)
}
p_new1<-NULL
for(i in 1:1318)
{
p_new1[i]=min(LOL[[i]])
}
 max(p_new1)+0.0001
p_dash=seq(max(p_new1)+0.0001,1,length=272) p_new <- NULL for(i in 1:1318) { p_new[[i]]<-g
p_data=data.frame(p_new)
distances <- dist(p_data)
hclust(distances, method = "complete") ;plot(hclust(distances, method = "complete"))

library(cluster)
k <- 3  # Number of clusters
kmedoids_result <- pam(distances,  k=k) # Output of k-means clustering
print(kmedoids_result)
 sil_width <- silhouette(kmedoids_result$cluster, distances) # Create silhouette plot
plot(sil_width, col = rainbow(3), main="Silhouette Plot for K=3 ",border = NA)
k <- 2  # Number of clusters
kmedoids_result2 <- pam(distances,  k=2) # Output of k-means clustering
print(kmedoids_result2)
sil_width2 <- silhouette(kmedoids_result2$cluster, distances) # Create silhouette plot
plot(sil_width2, col = rainbow(2),main="Silhouette Plot for K=2 ", border = NA)

k_values <- 2:15  # Example range from 2 to 15 clusters
avg_sil_widths <- numeric(length(k_values))
for (k in k_values)
{
   km <- pam(distances, k=k)
        sil_width <- silhouette(km$cluster, distances)
     # Compute the average silhouette width for this K
avg_sil_widths[k - min(k_values) + 1] <- mean(sil_width[, "sil_width"])
}
plot(k_values, avg_sil_widths, type = "b", xlab = "Number of Clusters (K)", ylab = "Averag
points(k_values, avg_sil_widths, col = "red", pch = 19)
##############################################################################################
```

```r
# our approch
Iterative_fun=function(initial,i)
{
##G_0=NULL
l=length(DAT[[i]]$V2)
e_i=rnorm(l,0,0.001)
x=c(LOL[[i]],LOL1[[i]])
y=c(DAT[[i]]$V2,DAT[[i]]$V2+e_i)
z=initial[1]+initial[2]*sin(x*initial[3])

d_0=matrix(rep(0,2*1),2*l,3)
for (j in 1:length(x))
 {
c_1=sum(y[j]-initial[1]-initial[2]*sin(initial[3]*x[j]))
c_2=sum(sin(initial[3]*x[j])*(y[j]-initial[1]-G_0[2]*sin(initial[3]*x[j])))
c_3=sum(initial[2]*x[j]*cos(initial[3]*x[j])*(y[j]-initial[1]-initial[2]*sin(initial[3]*x[
d_0[j,1] <-c_1
d_0[j,2]<-c_2
   d_0[j,3]<-c_3    # Assigning values incrementally y_0=y[j]-z[j]
}
Y_0=matrix(y_0,2*l,1)

model=lm(Y_0~d_0-1)
B_0=coef(model)
return(B_0)
}
#Estimating the parameters in the model
results_matrix <- matrix(NA, ncol = 5, nrow = 1318)
G_0=c(1,2,3)

for( i in 1:659)
{
   tolerance <- 1e-6
max_iterations <- 100
g_current <- G_0
converged <- FALSE
  iterations <- 0
while (!converged && iterations < max_iterations)
  {
     # Calculate b for current g
     b <- Iterative_fun(g_current,i)
  g_next <- g_current + b
if (max(abs(g_next - g_current) < tolerance))
 {
       converged <- TRUE
     }
```

```
else {
       # Update current g for next iteration
       g_current <- g_next
     }
          # Increment iteration counter
     iterations <- iterations + 1
   }
results_matrix[i, ] <- c(i, g_current, iterations)
}
for( i in 660:1318)
{
   tolerance <- 1e-6
max_iterations <- 100
g_current <- G_0
converged <- FALSE
  iterations <- 0
while (!converged && iterations < max_iterations)
  {
     # Calculate b for current g
     b <- Iterative_fun(g_current,i)
  g_next <- g_current + b
if (max(abs(g_next - g_current) < tolerance))
 {
       converged <- TRUE
     }
else {
       # Update current g for next iteration
       g_current <- g_next
     }
          # Increment iteration counter
     iterations <- iterations + 1
   }
results_matrix[i, ] <- c(i, g_current, iterations)
}

print(results_matrix)
 dim(results_matrix)
a_i=results_matrix[i=1:1318,2]
b_i=results_matrix[i=1:1318,3]
g_i=results_matrix[i=1:1318,4]


#distance matrix


 ##integrate(phi,lower=0,upper=2)
```

```
d=matrix(0,nrow=1318,ncol=1318)
for(i in  1:1317)
{
for (j in (i+1):1318)
{

phi=function(x)
{
((a_i[i]-a_i[j])+b_i[i]*sin(g_i[i]*x)-b_i[j]*sin(g_i[j]*x))^2
}
d[i,j]=integrate(phi,0,2, subdivisions = 1000)$val
}
}
for(i in  2:1318)
{
   for (j in 1:i-1)
{
d[i,j]=d[j,i]
}
}

#clustering using PAM
library(cluster)
par(mfrow=c(2,3))
k_values <- 2:20
k=19
kmedoidsresult<-pam(d,diss=TRUE,k=k)
plot(kmedoidsresult)
```