A Report

on

“Time-Series Analysis of R Scuti(variable) star magnitude”

By

Kaushal

submitted in the partial fulfilment

of the course

PHYS 601 (Statistical Methods & Computation)

under the guidance of

Prof. H. P. Singh

Department of Physics & Astrophysics

University of Delhi
Delhi-110007

Contents

1 Introduction

Mathematics, Physics, and Engineering are very successful in understanding phenomena of the natural world and building technology upon this based on the first principle modeling. However, for complex systems like those appearing in the fields of biology and medicine, this approach is not feasible and an understanding of the behavior can only be based upon the analysis of the measured data of the dynamics, the so-called time series. Time series analysis has different roots in Mathematics, Physics, and Engineering. The approaches differ by their basic assumptions. While in Mathematics linear stochastic systems were one of the centers of interest, in Physics nonlinear deterministic systems were investigated. While the different strains of the methodological developments and concepts evolved independently in different disciplines for many years, during the past decade, enhanced cross-fertilization between the different disciplines took place, for instance, by the development of methods for nonlinear stochastic systems.

Astronomical time series are somewhat different if to compare with standard time series often used in other branches of science and businesses. The random, often sparse and gapped nature of astronomical observational sequences makes most of techniques of the standard statistical analyis unusable. These lectures introduce some of the analytical, statistical and nu- merical methods which can be used in the context of astronomical data processing.

2 R Scuti Star

This variable star has been discovered by Edward Pigott in 1795 (Pigott & Englefield, 1797). It is a yellow giant variable star of the RV Tauri type, and varies semi-regularly between 5th and 8th magnitude.It is a star in the constellation of Scutum.

Various observers have determined primary periods between 140 and 146 days. Normally, the star's brightness varies between mag 4.8 and mag 6.0, but every fourth or fifth minimum drps to mag 8.0 or fainter. Superimposed over the primary period is at least one other period. According to Burnham, this star was in vestigated by McLaughlin who found that the star's spectrum is normally about G2, and it oscillates similar to a Cepheid variable, besides that the radial velocity of its oscillation is much more irregular. In particular, near the deep minimum reddens to about M3, and shows titanium-oxide bands, which are typical for red-giant spectra. During the consequent rise, hydrogen emission lines appear in the spectrum, which gradually turn to absorption lines as the star reaches its maximum. Spectral studies indicate that the various layers of the star ar expanding and contracting at different rates, and that the star is really huge in linear dimension: at least about 100 times the diameter of our sun. Absolute magnitude was estimated at -4.5 to -5.0, and its mass about 20-30 times that of our sun. This results in a distance estimate of 2,500 to 3,000 light-years.

Other spectral classifications have been G0e Ia near its maximum, K0p Ib near the minimum, in the Moscow General Catalogue of Variable Stars of 1970, and G8 to M3 by J.S. Glasby (1969).

The star shows a proper motion of about 0.06 arc seconds per year, and is receding from us at about 26.5 miles per second (42.4 km/s).

R Scuti is situated near the northern edge of a rich Milky Way star cloud, the Scutum Cloud. It lies just about 1deg south of Beta Scuti, and about 1deg NW of bright open star cluster M11.

3 Correlation

In statistics, dependence refers to any statistical relationship between two random variables or two sets of data. Correlation refers to any of a broad class of statistical relationships involving dependence.

Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling; however, statistical dependence is not sufficient to demonstrate the presence of such a causal relationship.

Formally, dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. In loose usage, correlation can refer to any departure of two or more random variables from independence, but technically it refers to any of several more specialized types of relationship between mean values. There are several correlation coefficients, often denoted or r, measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may exist even if one is a nonlinear function of the other). Other correlation coefficients have been developed to be more robust than the Pearson correlation.

3.1 Cross-correlation

Cross-correlations are statistical measures that indicate how one time-series is related to other or how one part of a time-series data is related to other part of the data. The Cross-correlation function between x_t and y_t is defined by

$$\gamma = \mathrm{Cov}[x_t, y_t] = \mathrm{E}[(x_t - \mu)(y_t - \alpha)]$$

And Cross-correlation coefficient is defined by

$$\rho_k = \frac{\mathrm{E}[(x_t - \mu)(y_t - \alpha)]}{\sqrt{\mathrm{E}[(x_t - \mu)^2 (y_t - \alpha)^2]}} \tag{1}$$

Cross-correlation coefficient is dimensionless and normalized, and will take values b/n -1 and $+1$.

3.2 Auto-correlation

Autocorrelations are statistical measures that indicate how a time-series is related to itself over time. Actually computing the correlation between two distinct data sets, the original series and the same series moved forward in time a specified number of periods. A graph of the correlation values is called "correlogram".

The covariance between x_t and its value x_{t+k} separated by lag k is called the auto-covariance at lag k and is defined by

$$\gamma_k = \mathrm{Cov}[x_t, x_{t+k}] = \mathrm{E}[(x_t - \mu)(x_{t+k} - \mu)]$$

And similarly the autocorrelation at lag k is defined by

$$\rho_k = \frac{\mathrm{E}[(x_t - \mu)(x_{t+k} - \mu)]}{\sqrt{\mathrm{E}[(x_t - \mu)^2 (x_{t+k} - \mu)^2]}} \tag{2}$$

4 Fourier Analysis: Fourier Series and Transforms

Periodic phenomena involving waves, rotating machines (harmonic motion), or other repetitive driving forces are described by periodic functions. Fourier series are a basic tool for solving ordinary differential equations (ODEs) and partial differential equations (PDEs) with periodic boundary conditions. Fourier Transforms deals with non-periodic phenomena and signals. The common name for the field is Fourier analysis.

Fourier Analysis has to do with breaking a signal(function) into simpler constituent parts. While Synthesis deals with resampling a signal from its constituent parts. The two things go together. One without another is nothing. Break some signal into it's constituent parts, take those parts, maybe modify them (depending upomn which part is more important in the original signal and which is not). The process of doing these steps are the two aspects of Fourier Analysis. Both these operations (Analysis and Synthesis) are accomplished by Linear operations. Linear operations implies here Integrals and Series. Because of this, one often thinks that Fourier Analysis is a part of "The Study of Linear Systems". But it is not necessarily true.

4.1 Fourier Series: Analysis of Periodic Phenomena

It is occassionaly helpful to classify periodic phenomena as:

- Periodicity in Time: Harmonic Motion (e.g. A pendulum)
- Periodicity in Space: Physical quantity distributed over a region (space) with symmetry (Consequence of the symmetry) (e.g. Distribution of Heat on a circular ring, here the physical quantity is Temperature and region is circular ring. $T(x, y)$ is periodic in space. "Heat Distribution over a ring" is the problem that Fourier himself considered).

Therefore, often Fourier Analysis is seen associated with the problems of symmetry. For mathematical description of periodicity in time, we often use "Frequency": Number of repeations(cycles) of the pattern in one second. While, for the phenomena that is periodic in space, we use "Period": Physical measurement of how long (big) the pattern is before it repeats. Periodicity in time and periodicity in space are not completely separate phemomena. The two notions come together in, e.g. wave motion (regularly repeating pattern that changes with time).

Now, where again the two description comes, we have frequency(denoted by ν) for periodicity in time that's the number of times the pattern repeats in one second(cycles/second)and wavelength(denoted by λ, fix the time and see how the pattern is distributed in space) for periodicity in space that's the length of one complete pattern. Periodicity in time and periodicity in sapce may be two complete different phenomena that have nothing to do with each other. But for wave motion (regularly repeating pattern over time and space), there is a relationship between the two notions. This relation is governed by the velocity and given as:

$$distance = rate \times time$$

In case of wave motion:

$v = velocity\ (rate)\ of\ the\ wave$

$\lambda = distance\ travelled\ by\ the\ wave\ in\ one\ cycle$

$\nu = number\ of\ cycles\ in\ one\ sec$

Therefore

$$\lambda = v \times \frac{1}{\nu}$$

or

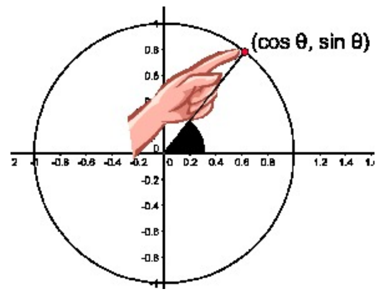
$$v = \nu \times \lambda$$

Hence for a wave with fixed velocity, the above equation exhibits a reciprocal relationship between frequency and wavelength or in other words reciprocal relationship between the two domains of Fourier Analysis, the frequency domain and spatial domain.

Now why does the mathematics comes in when the periodicty is kind of a physical phenomena? Becasue, there are some mathematical functions that are simple mathematical functions and are periodic and therefore can be used to model periodic phenomena. Sine and cosine are very general periodic functions and can be used to expressed many complicated and easy periodic and non-periodic function in a simpler way. These are periodic of period 2π.

Sine and cosine are associated with a circle as shown in the adjacent figure, where $cos\ \theta$ is x-coordinate and sin θ is y-coordinate (θ is radian measure). If we go once around the whole circle i.e. if we go from θ to $\theta + 2\pi$, we will end up from where we started with and that implies periodicity in space. Furthermore, it is not just 2π but any mutiple of 2π, positive(anti-clockwise) or negative(clockwise).

Mathematically,



$\sin(\theta + 2\pi n) = \sin(\theta)$ and $\cos(\theta + 2\pi n) = \cos(\theta)$ where $n = 0, \pm 1, \pm 2, ...$

These general periodic functions can be used to model "the most complex phenomena". That is the fundamental point of Fourier Series.

Now, we have $\sin(2\pi f t)$ and $\cos(2\pi f t)$ as model signals, where $\omega = 2\pi f$ anf $f = \frac{1}{T}$ is the fundamental frequency. We can modify and combine these signals to model general periodic signals.

Let us think of the following: One period and many frequencies.

1. $\sin(2\pi t)$, Period=1
   Frequency=1

2. $\sin(4\pi t)$, Period=$\frac{1}{2}$
   Frequency=2

3. $\sin(6\pi t)$, Period=$\frac{1}{3}$
   Frequency=3

$\sin(4\pi t)$ and $\sin(6\pi t)$ can also be considered having period 1, if we think of two and three cycles for these signals respectively as one basic pattern. If we combine (simply sum) these three signals together, the resultant would be $\sin(2\pi t) + \sin(4\pi t) + \sin(6\pi t)$ and it will look like as shown in Figure 1.

Period of the sum is that of the slowest in the signal, in our case it is 1. Although the terms of higher frequencies are repeating more rapidly, but the sum can not go back to the point where it started until the slowest one gets cut up.

One period may fcrequencies: There are three frequencies in the sum i.e. 1, 2 & 3. But adding up together there is only one period. For compicated periodic phenomena, it

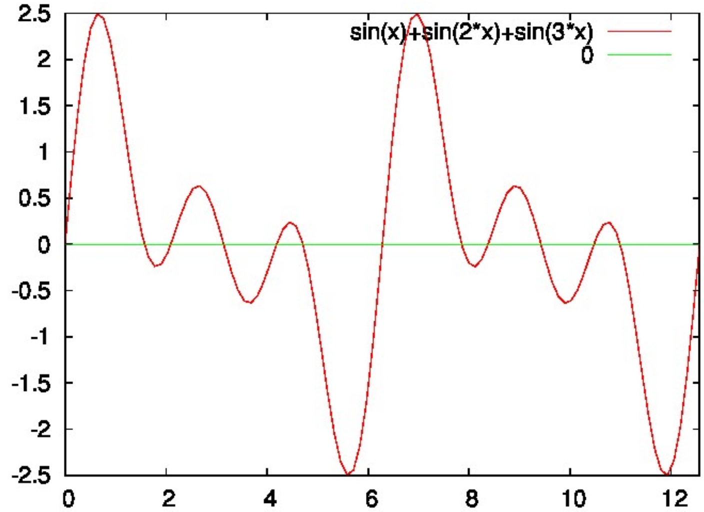


Figure 1: Resultant of combining three Sin pulse: Same period, Three frequencies.

is really better and more revealing to talk in terms of frequencies rather than in terms of period.

As we modified the frequencies in the last example, similarly we can modify the amplitude as well as the phase of each one of these signals. To model a complicated signal of period 1, we can modify the amplitude, the frequency and the phases of $\sin(2\pi f t)$ and add up the results. Mathematically, it can be written as:

$$\sum_{k=1}^{n} A_k \sin(2\pi k t + \phi_k) \tag{3}$$

where $A_k' s$ are introduced to modify the amplitudes, k term in the bracket to add the higher frequencies and ϕ_k to modify the phase of the signals. The longest period of the sum is when $k = 1$. Terms with higher frequencies in Equation 3 are called **harmonics** because of their connection with musical phenomena. The harmonics have shorter period with higher frequencies but the period of the sum is 1, because the whole pattern can not repeat itself until the longest period is repeated.

but

$$\sin(2\pi kt + \phi_k) = \sin(2\pi kt)\cos(\phi_k) + \cos(2\pi kt)\sin(\phi_k)$$

using above expression in Equation 3, we obtain:

$$\sum_{k=1}^{n}(a_k \cos(2\pi kt) + b_k \sin(2\pi kt))$$

where $a_k = A_k \cos(\phi_k)$ and $b_k = A_k \sin(\phi_k)$.

Above expression can be combined with a constant term to shift the whole thing for the purpose of generality. And the expression becomes:

$$\frac{a_0}{2} + \sum_{k=1}^{n}(a_k \cos(2\pi kt) + b_k \sin(2\pi kt)) \tag{4}$$

The first term of Equation 4 is referred to as "DC component" in Signal Analysis. The coefficients $a_k's$ and $b_k's$ contains the information about the phase. By far better is to use to represent sine and cosine via complex exponentials.

$$\cos(2\pi kt) = \frac{e^{2\pi ikt} + e^{2\pi ikt}}{2} \qquad \& \qquad \sin(2\pi kt) = \frac{e^{2\pi ikt} - e^{2\pi ikt}}{2i}$$

Thus Equation 4 yields:

$$\sum_{k=-n}^{n} C_k e^{2\pi ikt} \tag{5}$$

where $C_k's$ are complex numbers and can be expressed in terms of $a_k's$ and $b_k's$. But these are not just any arbitrary complex numbers, they satisfy symmetry property i.e. $C_{-k} = C_k^*$. And because $C_k's$ satisfy this property, the signals represented by the complex exponentials in Equation 5 are real.

Now if we are given a periodic function $f(t)$ of period 1, can we write $f(t)$ as a linear combination of basic trignometric quantities as

$$f(t) = \sum_{k=-n}^{n} C_k e^{2\pi ikt} \tag{6}$$

Suppose we can do this, then what are the coefficients $C_k's$ in Equation 6 ? Equation 6 can be re-written as:

$$\begin{aligned}
f(t) &= .... + C_m e^{2\pi imt} + ... \\
C_m e^{2\pi imt} &= f(t) - \sum_{k \neq m} C_k e^{2\pi ikt} \\
C_m &= f(t)e^{-2\pi imt} - \sum_{k \neq m} C_k e^{2\pi i(k-m)t}
\end{aligned}$$

Now integrating above expression over time over the whole period from 0 to 1 yields

$$\begin{aligned}
\int_0^1 C_m dt &= \int_0^1 f(t)e^{-2\pi imt} dt - \int_0^1 \sum_{k \neq m} C_k e^{2\pi i(k-m)t} dt \\
\int_0^1 C_m dt &= \int_0^1 f(t)e^{-2\pi imt} dt - \sum_{k \neq m} C_k \underbrace{\int_0^1 e^{2\pi i(k-m)t} dt}_{\delta(k-m)}
\end{aligned}$$

Coefficients $C_k's$ and $C_m's$ are independent of time, therefore itegration in the last term of above expression can be taken inside the summation. The last term in the above equation is 0. Therfore:

$$C_m = \int_0^1 f(t)e^{-2\pi imt} dt \tag{7}$$

If $f(t)$ is known, then the coefficients are given by Equation 7. To represent any general periodic phenomena, we have to consider infinite sum (It takes high frequencies to make sharp corners). Thus, a complex-valued signal $f(t)$ that is periodic with period P can be written in the form of an infinite complex Fourier Series.

$$f(t) = \sum_{k=-\infty}^{\infty} C_k e^{2\pi ikt}$$

Here k is the fundamental frequecy. So k can be replaced by $\omega_0 = 2\pi f_0$, where $f_0 = \frac{1}{P}$ is the fundamental frequency of the signal. Thus the expression becomes

$$f(t) = \sum_{m=-\infty}^{\infty} C_m e^{2\pi imf_0t} \tag{8}$$

And the coefficients are given by

$$C_m = \frac{1}{P}\int_{-P/2}^{P/2} f(s)e^{-2\pi imf_0s}ds \tag{9}$$

4.2 Aperiodic Signals and Fourier Transforms

A signal $f_P(t)$ of period P and fundamental frequency $f_0 = \frac{1}{P}$ can be represented by the Fourier Series as:

$$f_P(t) = \sum_{m=-\infty}^{\infty} \left[\frac{1}{P}\int_{-P/2}^{P/2} f(s)e^{-2\pi imf_0s}ds\right].e^{2\pi imf_0t} \tag{10}$$

The inner transformation

$$F(k) = \int_{-\infty}^{\infty} f(t)e^{-i2\pi ft}dt \tag{11}$$

is called the **Fourier transform** of signal $f(t)$, and the outer transform

$$f(t) = \int_{-\infty}^{\infty} F(k)e^{i2\pi ft}dk$$

is called the **Inverse Fourier Transform**.

4.3 Discrete and Fast Fourier Transforms

In practice, for many signals, we only sample the value of the signal at discrete times, although in reality the signal continues between these sampling times. In such cases we can approximate the integrals involved in calculation of the Fourier transforms in the same way as one does in numerical integration in calculus, using left-handed rectangles, trapezoids, Simpson's rule, etc. We use the simplest approximation, which is equivalent to assuming that the signal is constant between the sampling times (and rectangles' areas approximate the area under the function).

So suppose that the sampling period is T_s , with the sampling frequency $f_s = \frac{1}{T_s}$, so that the signal's sample is given in the form of a finite sequence,

$$x_k = x(kT_s), \qquad k = 0, 1, 2, ...N - 1 \tag{12}$$

and we interpret it as a periodicsignal with period

$$P = \frac{1}{f_0} = NT_s = \frac{N}{f_s} \tag{13}$$

The integral in Equation 8 approximating the Fourier transform of the signal $f(t)$ at discrete frequencies $m = 0, 1, 2, ...N - 1$; can now be, in turn, approximated by the sum

$$\begin{aligned} X_m &= X(mf_0) \\ &= \frac{1}{P}\sum_{k=0}^{N-1} x(kT_s)e^{-2\pi i m f_0 k T_s}.T_s \\ &= \frac{1}{N}\sum_{k=0}^{N-1} x_k e^{-2\pi i m k/N} \end{aligned} \tag{14}$$

The sequence

$$X_m, \qquad m = 0, 1, 2, ...N - 1$$

is called the **Discrete Fourier Transform (DFT)** of the signal sample $x_k,\ k = 0, 1, 2, ...N-1$.

Note that the calculation of the DFT via formula 12 calls for N^2 multiplications,

$$x_k.e^{-2\pi imk/N}, \qquad m, k = 0, 1, 2.........N - 1$$

One often says that the formula's computational (algorithmic) complexity is of the order N^2 . This computational complexity, however, can be dramatically reduced by cleverly grouping terms in the Equation 14. The technique, which usually is called the **fast Fourier transform (FFT)**, was known to Carl Friedrich Gauss at the beginning of the nineteenth century, but was rediscovered and popularized by Cooley and Tukey in 1965. This technique will be explained in the special case when the signals sample size N is a power of 2.

So assume that $N = 2^n$, and let $\omega_N = e^{-2\pi i/N}$. The complex number ω_N is called a complex N'th root of unity because $\omega_N^N = 1$. Obviously, for $M = N/2$, we have

$$\omega_{2M}^{(2k)m} = \omega_M^{km} \qquad , \qquad \omega_M^{M+m} = \omega_M^m \quad \text{and} \quad \omega_{2M}^{M+m} = -\omega_{2M}^m \tag{15}$$

The crucial observation is to recognize that the sum in Equation 14 can be split into two pieces:

$$X_m = \frac{1}{2}(X_m^{even} + X_m^{odd}.\omega_{2M}^m), \tag{16}$$

where

$$X_m^{even} = \frac{1}{M}\sum_{k=0}^{M-1} x_{2k}\omega_m^{km} \qquad \text{and} \qquad X_m^{odd} = \frac{1}{M}\sum_{k=0}^{M-1} x_{2k+1}\omega_m^{km} \tag{17}$$

and using Equation 13, we get

$$X_{m+M} = \frac{1}{2}(X_m^{even} - X_m^{odd}.\omega_{2M}^m) \tag{18}$$

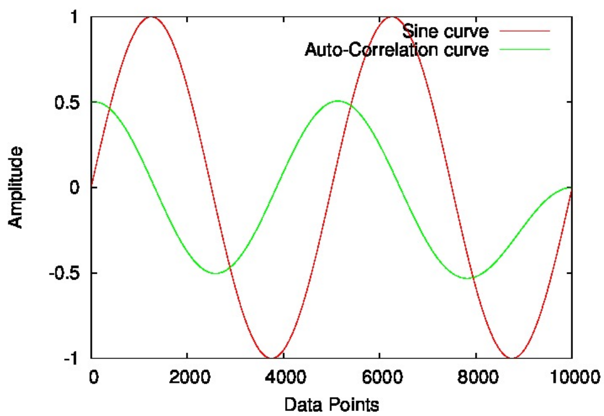
As a result, only values $X_m,\ m = 0, 1, 2......, M - 1 = \frac{N}{2} - 1$, have to be calculated by laborious multiplication. Thus the computational complexity using the above technique is redued to

$$CC(n) = 2^{n-1}\log_2 2^n = \frac{1}{2}\log_2 N \tag{19}$$

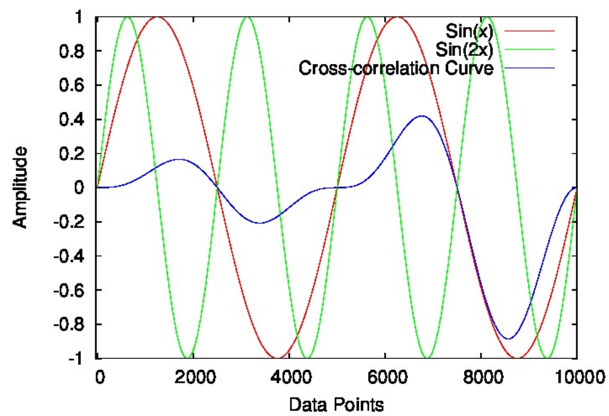
which is a major improvement over the N^2 -order of the computational complexity of the straightforward calculation of the DFT.

5 Data Analysis

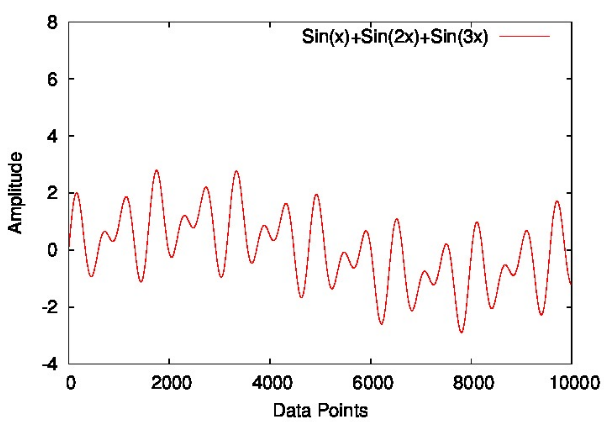
5.1 Sine Pulse for Demonstration



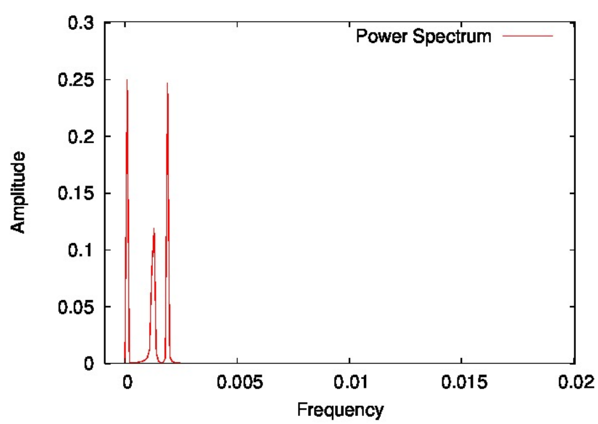
(a) Sine wave and its Autocorrelation function.



(b) Cross-correlation of sin(x) and sin(2x).



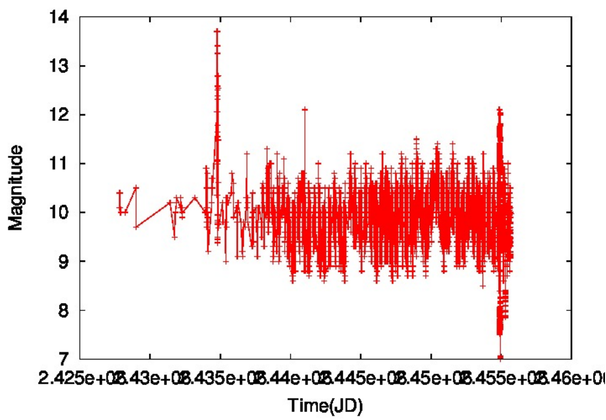
(c) Combination of three sinusoidal signals.



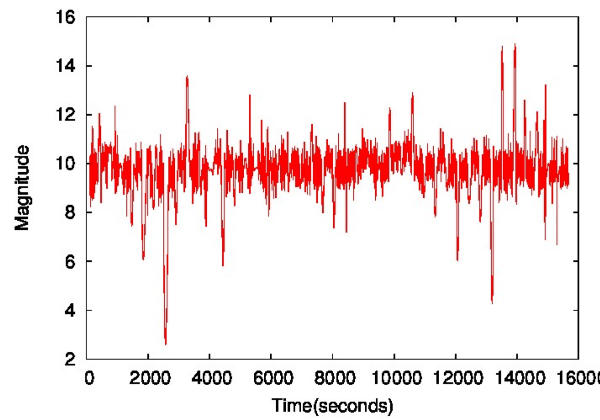
(d) Power spectrum of sin(x)+sin(2x)+sin(3x).

Figure 2: Sinusoidal waves, their correlation and power spectrum

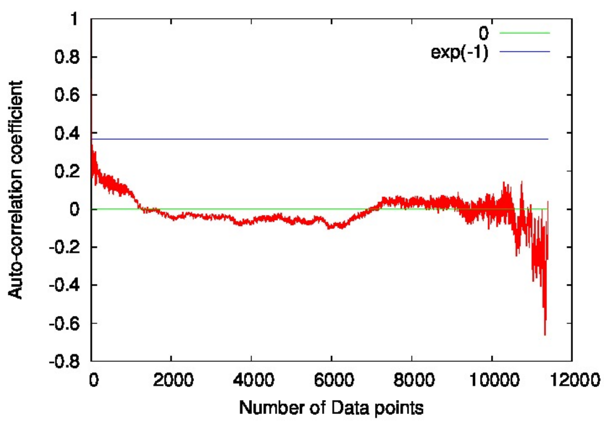
5.2 Varying Magnitude of R Scuti Star: Time-Seriesed Data



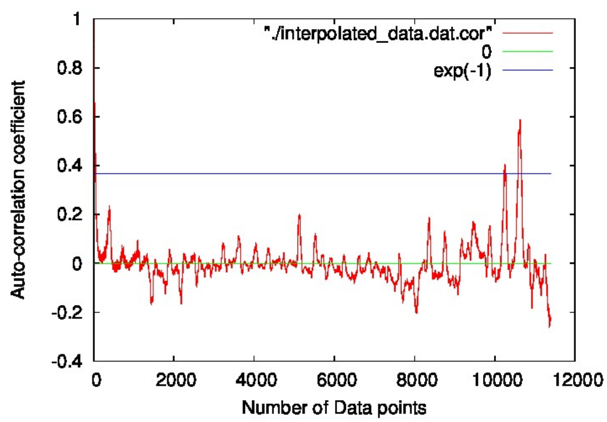
(a) Variation of R Scuti's magnitude over Time.



(b) Interpolated data of R Scuti's magnitude.



(c) Auto-correlation of the original data.



(d) Auto-correlation of the interpolated data.

Figure 3: R Scuti's magnitude variation and it's Auto-correlation

6 Bibliography

1. http://en.wikipedia.org/wiki/RScuti

2. A First course in Statistics for Signal Analysis (Wojbor A. Woyczynski)

3. Applied Time-Series: Analysis and Forecasting (T. M. J. A. Cooray)

4. http://www.aavso.org/vsotsrsct

5. http://www.aavso.org/lcotw/r-scuti

6. http://www.mpipks-dresden.mpg.de/ tisean/Tisean3.0.1/index.html

7. Practical implementation of nonlinear time series methods: The TISEAN package (Rainer Hegger, Holger Kantz, Thomas Schreiber)