# LEAST-SQUARES FREQUENCY ANALYSIS OF UNEQUALLY SPACED DATA

N. R. LOMB
*School of Physics, University of Sydney, N.S.W., Australia*

(Received 15 May, 1975)

**Abstract.** The statistical properties of least-squares frequency analysis of unequally spaced data are examined. It is shown that, in the least-squares spectrum of gaussian noise, the reduction in the sum of squares at a particular frequency is a $\chi^2_2$ variable. The reductions at different frequencies are not independent, as there is a correlation between the height of the spectrum at any two frequencies, $f_1$ and $f_2$, which is equal to the mean height of the spectrum due to a sinusoidal signal of frequency $f_1$, at the frequency $f_2$. These correlations reduce the distortion in the spectrum of a signal affected by noise. Some numerical illustrations of the properties of least-squares frequency spectra are also given.

## 1. Introduction

In astronomy – especially in the field of variable stars – it is often necessary to analyse data for unknown periodicities. For data obtained at uniformly spaced intervals, standard methods of analysis are available, such as Fourier methods based on the Fast Fourier Transform and the recently developed Method of Maximum Entropy. Unfortunately, in most ground based astronomical work uniform spacing is impossible to achieve. Observations are necessarily limited to night time and are further restricted by the weather, availability of telescope time and the position of the object under observation. Even within each night of observation the data are rarely equally spaced.

The spectrum of a set of non-uniform data is far more complex than the spectrum of a set of uniform data, for there is no frequency region, as there is in the analysis of equally spaced data, in which a period is unambiguously defined. Each true peak in the spectrum gives rise to a number of other peaks (aliases) of various heights, distributed throughout the spectrum. As a consequence no more than one period can be determined for any one calculation of the spectrum because of possible confusion with the alias structure of the major peak. Subsequent periods have to be found by successively subtracting the previously found periodicities from the data and calculating the 'prewhitened' spectrum.

The most commonly used method of calculating the spectrum of non-uniformly spaced data is periodogram analysis. It ignores the non-equal spacing and involves calculating the normal Fourier power spectrum, as if the data were equally spaced, though, of course, without recourse to the Fast Fourier Transform algorithm. It has been used, for example, by Wehlau and Leung (1964). A slightly modified form of periodogram analysis has been devised by Gray and Desikachary (1973), in which prewhitening is carried out in the frequency domain instead of the time domain. However, with unequally spaced data the Fourier power spectrum has no well-defined

*Astrophysics and Space Science* **39** (1976) 447–462. 

properties. Even in the simplest possible case of noise-free data containing one sinusoidal periodicity the highest peak does not necessarily occur at the correct period. The sole justification for the use of periodogram analysis is that, as will be shown later, it provides a reasonably good approximation to the spectrum obtained by fitting sine waves by least-squares to the data and plotting the reduction in the sum of the residuals against frequency. This least squares (or LS) spectrum (Barning, 1963) provides the best measure of the power contributed by the different frequencies to the overall variance of the data and can be regarded as the natural extension of Fourier methods to non-uniform data. It reduces to the Fourier power spectrum in the limit of equal spacing.

The statistics and behaviour of the LS spectrum will be investigated in this paper. An elaborate scheme of least-squares frequency analysis has been put forward by Vaníček (1971), in which for each trial frequency a least-squares solution is made simultaneously for the amplitudes of all known constituents of the data and the amplitude and phase of the sine wave with the trial frequency. This scheme will not be considered here as, under most circumstances, it provides only a marginal improvement to the accuracy of the simple LS spectrum and also, it would greatly increase the complexity of the discussion. However, it is felt that at least some of the results obtained for the LS spectrum could be applied to Vaníček's method. Some of the questions that will be asked about the LS spectrum are: What is the probability distribution of the height of the spectrum at a given frequency if the data consists of noise with a gaussian distribution? Considering that a sinusoidal periodicity in the data gives rise to a number of alias peaks, are there any correlations between the heights of noise peaks at different frequencies? How much does the presence of noise distort the spectrum due to a sinusoidal signal?

## 2. Formulae for the LS Spectrum

Given a set of $n$ observations $y_i$, $i=1, 2, \ldots, n$, with zero mean and obtained at times $t_i$, we can set up the model

$$y_i + \varepsilon_i = a \cos 2\pi f t_i + b \sin 2\pi f t_i,$$

where the errors $\varepsilon_i$ are independent, have zero mean and a common variance $\sigma^2$, $a$ and $b$ are unknown and the frequency $f$ is given.

Adopting the notation

$$CC = \sum_{i=1}^{n} \cos^2 2\pi f t_i, \qquad SS = \sum_{i=1}^{n} \sin^2 2\pi f t_i,$$

$$CS = \sum_{i=1}^{n} \cos 2\pi f t_i \sin 2\pi f t_i,$$

$$YC = \sum_{i=1}^{n} y_i \cos 2\pi f t_i, \qquad YS = \sum_{i=1}^{n} y_i \sin 2\pi f t_i,$$