1976Ap&SS..39..447L

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 χ_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. *All Rights Reserved*
Copyright © 1976 by D. Reidel Publishing Company, Dordrecht-Holland

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

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 ε_i are independent, have zero mean and a common variance σ^2, a and b are unknown and the frequency f is given.

Adopting the notation

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

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

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

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

we are led to the normal equations

$$\begin{bmatrix} CC & CS \\ CS & SS \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} YC \\ YS \end{bmatrix},$$

and a reduction in the sum of squares of

$$\Delta R(f) = [YC \quad YS]\begin{bmatrix} CC & CS \\ CS & SS \end{bmatrix}^{-1}\begin{bmatrix} YC \\ YS \end{bmatrix} =$$

$$= [YC \quad YS]\begin{bmatrix} SS/D & -CS/D \\ -CS/D & CC/D \end{bmatrix}\begin{bmatrix} YC \\ YS \end{bmatrix}, \tag{1}$$

where

$$D = CC \cdot SS - CS^2.$$

Although for numerical work it is simplest to use an expansion of Equation (1), it would facilitate the statistical description of the LS spectrum if $\Delta R(f)$ could be expressed in the form A^2+B^2. This can be done by fitting

$$y_i = a \cos 2\pi f(t_i - \tau) + b \sin 2\pi f(t_i - \tau)$$

to the data, instead of

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

and choosing τ such that $CS=0$. We then have from Equation (1)

$$\Delta R(f) = [YC \quad YS]\begin{bmatrix} 1/CC & 0 \\ 0 & 1/SS \end{bmatrix}\begin{bmatrix} YC \\ YS \end{bmatrix},$$

where now, e.g.,

$$CC = \sum_{i=1}^{n} \cos^2 2\pi f(t_i - \tau).$$

Hence,

$$\Delta R(f) = \left(\frac{1}{\sqrt{CC}}\, YC\right)^2 + \left(\frac{1}{\sqrt{SS}}\, YS\right)^2 =$$

$$= C^2(f) + S^2(f). \tag{2}$$

When $\Delta R(f)$ is expressed in this compact form the similarity with the usual periodogram formula becomes evident; in fact the periodogram formula is an approximation to this exact formula. By making two assumptions: $CS=0$ for all values of τ and $CC=SS=n/2$, which are both approximately satisfied, Equation (2) can be converted to

$$AR(f) \simeq \left(\sqrt{\frac{2}{n}}\, YC\right)^2 + \left(\sqrt{\frac{2}{n}}\, YS\right)^2,$$

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

which is the formula used in periodogram analysis.

If we let

$$R = \sum_{i=1}^{n} y_i^2,$$

a normalized spectral function can be defined by

$$p(f) = \frac{C^2(f) + S^2(f)}{R}. \tag{3}$$

The values of $p(f)$ lie obviously in the range 0 to 1.

3. Spectrum due to a Sinusoidal Signal

A sinusoidal signal of frequency f_1 can be represented by

$$g_i = a \cos 2\pi f_1(t_i - \tau_1) + b \sin 2\pi f_1(t_i - \tau_1).$$

If the g_i are our observations, that is $y_i = g_i$, we can write for any frequency f_2

$$C(f_2) = \frac{1}{\sqrt{C_2 C_2}}(aC_1C_2 + bS_1C_2),$$

where, e.g.,

$$C_1C_2 = \sum_{i=1}^{n} \cos 2\pi f_1(t_i - \tau_1) \cos 2\pi f_2(t_i - \tau_2).$$

Also

$$S(f_2) = \frac{1}{\sqrt{S_2 S_2}}(aC_1S_2 + bS_1S_2),$$

and the reduction in the sum of squares, for a frequency f_2, is, using Equation (2)

$$\Delta R_G(f_2) = a^2\left(\frac{C_1C_2^2}{C_2C_2} + \frac{C_1S_2^2}{S_2S_2}\right) + b^2\left(\frac{S_1C_2^2}{C_2C_2} + \frac{S_1S_2^2}{S_2S_2}\right) + 2ab\left(\frac{C_1C_2 \cdot S_1C_2}{C_2C_2} + \frac{C_1S_2 \cdot S_1S_2}{S_2S_2}\right).$$

If we now define

$$\varrho_{C_1,C_2} = \frac{C_1C_2}{\sqrt{C_1C_1 \cdot C_2C_2}};$$

similarly define ϱ_{C_1,S_2}, ϱ_{S_1,S_2}, etc. (the reason for this notation will become clear when we discuss the response of the LS spectrum to random noise); and also define

$$A = a\sqrt{C_1C_1} \quad \text{and} \quad B = b\sqrt{S_1S_1},$$

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

we have

$$\Delta R_G(f_2) = A^2(\varrho^2_{c_1,c_2} + \varrho^2_{c_1,s_2}) + B^2(\varrho^2_{s_1,c_2} + \varrho^2_{s_1,s_2}) + \\ + 2AB(\varrho_{c_1,c_2}\,\varrho_{s_1,c_2} + \varrho_{c_1,s_2}\,\varrho_{s_1,s_2}).$$

Now the total sum of squares is given by

$$R = a^2C_1C_1 + b^2S_1S_1 + 2abC_1S_1 = A^2 + B^2$$

if τ_1 was chosen such that $C_1S_1 = 0$.

The normalized spectrum then becomes, by use of Equation (3),

$$p_G(f_2) = \frac{A^2}{A^2 + B^2}(\varrho^2_{c_1,c_2} + \varrho^2_{c_1,s_2}) + \frac{B^2}{A^2 + B^2}(\varrho^2_{s_1,c_2} + \varrho^2_{s_1,s_2}) + \\ + \frac{2AB}{A^2 + B^2}(\varrho_{c_1,c_2}\,\varrho_{s_1,c_2} + \varrho_{c_1,s_2}\,\varrho_{s_1,s_2}). \tag{4}$$

This value for $p_G(f_2)$ will vary slightly as the ratio of A to B changes, that is as the phase of the signal is varied. To find the mean value let us put

$$\sin^2\alpha = \frac{A^2}{A^2 + B^2} \quad \text{and} \quad \cos^2\alpha = \frac{B^2}{A^2 + B^2}.$$

Then Equation (4) becomes

$$p_G(f_2) = \sin^2\alpha(\varrho^2_{c_1,c_2} + \varrho^2_{c_1,s_2}) + \cos^2\alpha(\varrho^2_{s_1,c_2} + \varrho^2_{s_1,s_2}) + \\ + \sin 2\alpha(\varrho_{c_1,c_2}\cdot\varrho_{s_1,c_2} + \varrho_{c_1,s_2}\cdot\varrho_{s_1,s_2});$$

and the mean value as the phase of the signal is varied is given by

$$\bar{p}_G(f_2) = \tfrac{1}{2}(\varrho^2_{c_1,c_2} + \varrho^2_{c_1,s_2} + \varrho^2_{s_1,c_2} + \varrho^2_{s_1,s_2}). \tag{5}$$

Equations (4) and (5) completely describe the spectrum due to a sine wave. They are, however, rather complex, so to get a qualitative picture of the shape of the spectrum of a sine wave it is necessary to simplify them by making some approximations. By making the approximations that $CS = 0$ for all values of τ and that $CC = SS = n/2$ we reach the stage of approximation represented by periodogram analysis (Section 2). It can then be shown that Equations (4) and (5) reduce to

$$P_G(f_2) = \bar{p}_G(f_2) \propto |W(f_2 - f_1) + W(f_2 + f_1)|^2,$$

where $W(f)$ is the Fourier transform of the observing window, which is a function that equals 1 whenever $t \in \{t_1, t_2, \ldots, t_n\}$.

4. Spectrum of Random Noise

If the series u_i, $i = 1, 2, \ldots, n$, constitute a random sample from a normally distributed population with mean zero and variance σ^2 and we take $y_i = u_i$ then

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

$$C(f) = \frac{1}{\sqrt{CC}} YC = \frac{1}{\sqrt{CC}} \sum_{i=1}^{n} u_i \cos 2\pi f(t_i - \tau)$$

is also normally distributed and its mean or expected value is given by

$$E[C(f)] = \frac{1}{\sqrt{CC}} \sum_{i=1}^{n} E(u_i) \cos 2\pi f(t_i - \tau) = 0;$$

and its variance, by

$$E[C^2(f)] = \frac{1}{CC} \sum_{i=1}^{n} E(u_i^2) \cos^2 2\pi f(t_i - \tau).$$

In these equations we have ignored terms involving $E(u_i)$, since these are equal to zero and so

$$E[C^2(f)] = E(u_i^2) = \sigma^2.$$

The function $S(f)$ is normally distributed with zero mean and variance of σ^2. The covariance of $C(f)$ and $S(f)$ is given by

$$E[C(f) \cdot S(f)] = \frac{1}{\sqrt{CC}} \frac{1}{\sqrt{SS}} \sum_{i=1}^{n} E(u_i^2) \cos 2\pi f(t_i - \tau) \sin 2\pi f(t_i - \tau) =$$

$$= \frac{\sigma^2}{\sqrt{CC}\sqrt{SS}} CS = 0,$$

since τ was chosen so that $CS = 0$. Thus $C(f)$ and $S(f)$ are independent and $AR_N(f) = C^2(f) + S^2(f)$ is σ^2 times a χ^2 variate with 2 degrees of freedom.

From the above result it would seem that the spectrum of random noise is a set of peaks, the heights of which are governed by the χ_2^2-distribution. However, we found when discussing the spectrum of a sine curve that each true peak gives rise to a number of other peaks (aliases). Consequently, it would be reasonable to suppose that each noise peak would be related to some other peaks in the spectrum.

Consider the correlation between $C(f_1)$ and $C(f_2)$ in a noise spectrum

$$\varrho[C(f_1), C(f_2)] = \frac{E[C(f_1)\, C(f_2)]}{(E[C^2(f_1)]\, E[C^2(f_2)])^{1/2}} =$$

$$= \frac{1}{\sigma^2 \sqrt{C_1 C_1 \cdot C_2 C_2}} \sum_{i=1}^{n} E(u_i^2) \cos 2\pi f_1(t_i - \tau_1) \times$$

$$\times \cos 2\pi f_2(t_i - \tau_2) =$$

$$= \frac{C_1 C_2}{\sqrt{C_1 C_1 \cdot C_2 C_2}},$$

but this has already been defined as ϱ_{C_1, C_2}. So

$$\varrho[C(f_1), C(f_2)] = \varrho_{C_1, C_2};$$

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

and similarly for the correlation between $C(f_1)$ and $S(f_2)$, etc. If we call the correlation between the level of the spectrum at f_1 and f_2, ϱ_{12} then

$$\varrho_{12} = \varrho[p_N(f_1), p_N(f_2)] = \\ = \varrho[\Delta R_N(f_1), \Delta R_N(f_2)].$$

Now, using the fact that

$$E[\Delta R_N(f_1)] = E[\Delta R_N(f_2)] = \sigma^2 E(\chi_2^2) = 2\sigma^2$$

and that

$$E[\Delta R_N(f_1) - 2\sigma^2]^2 = E[\Delta R_N(f_2) - 2\sigma^2]^2 = 4\sigma^4, \tag{6}$$

since the variance of a χ_2^2-variable is equal to 4, we have

$$\varrho_{12} = E[(\Delta R_N(f_1) - 2\sigma^2)(\Delta R_N(f_2) - 2\sigma^2)]/4\sigma^4 = \\ = \frac{1}{4\sigma^4} E[\Delta R_N(f_1) \cdot \Delta R_N(f_2)] - 1.$$

It is shown in the Appendix that

$$E[\Delta R_N(f_1) \cdot \Delta R_N(f_2)] = 4\sigma^4 + 2\sigma^4(\varrho^2_{c_1,c_2} + \varrho^2_{c_1,s_2} + \varrho^2_{s_1,c_2} + \varrho^2_{s_1,s_2});$$

and so we obtain, finally,

$$\varrho_{12} = \tfrac{1}{2}[\varrho^2_{c_1,c_2} + \varrho^2_{c_1,s_2} + \varrho^2_{s_1,c_2} + \varrho^2_{s_1,s_2}] = \bar{p}_G(f_2) \tag{7}$$

from Equation (5). Thus the correlation between the heights of a noise spectrum at frequencies f_1 and f_2 is equal to the mean height of the spectrum of a sine curve of frequency f_1 at frequency f_2. Note that from Equation (6), ϱ_{12} is also the regression coefficient of $p_N(f_1)$ on $p_N(f_2)$ and $p_N(f_2)$ on $p_N(f_1)$.

5. Effect of Noise on the Spectrum of a Sine Wave

Let

$$y_i = g_i + u_i,$$

where again

$$g_i = a \cos 2\pi f_1(t_i - \tau_1) + b \sin 2\pi f_1(t_i - \tau_1)$$

and u_i is normally distributed with $E(u_i) = 0$ and $E(u_i^2) = \sigma^2$. The reduction in the sum of squares at a particular frequency f_2 is given, with the help of Equation (2), by

$$\Delta R(f_2) = \frac{1}{C_2 C_2}(GC_2 + UC_2)^2 + \frac{1}{S_2 S_2}(GS_2 + US_2)^2,$$

where we have used the notation

$$GC_2 = \sum_{i=1}^{n} g_i \cos 2\pi f_2(t_i - \tau_2), \qquad UC_2 = \sum_{i=1}^{n} u_i \cos 2\pi f_2(t_i - \tau_2), \quad \text{etc.}$$

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

On expanding we find that

$$\Delta R(f_2) = \Delta R_G(f_2) + \Delta R_N(f_2) + Ia(f_2),$$

where $\Delta R_G(f_2)$ is the reduction in the sum of squares due to the signal in the absence of noise, $\Delta R_N(f_2)$ is the reduction due to the noise in the absence of a signal and $Ia(f_2)$ is an interaction term between the signal and the noise. $Ia(f_2)$ is given by

$$Ia(f_2) = \frac{2}{C_2 C_2} GC_2 \cdot UC_2 + \frac{2}{S_2 S_2} GS_2 \cdot US_2. \tag{8}$$

Since R, the total sum of squares is equal to $R_G + R_N$ where $R_G = \sum_{i=1}^{n} g_i^2$ and $R_N = \sum_{i=1}^{n} u_i^2$, the normalized spectrum can be written as

$$p(f_2) = \frac{R_G}{R_G + R_N} p_G(f_2) + \frac{R_N}{R_G + R_N} p_N(f_2) + \frac{Ia(f_2)}{R_G + R_N}, \tag{9}$$

where

$$p_G(f_2) = \Delta R_G(f_2)/R_G \quad \text{and} \quad p_N(f_2) = \Delta R_N(f_2)/R_N.$$

In Equation (9) the first term is constant, while the statistical behaviour of the second term was discussed in the previous section. Let us now discuss the third term in the equation. As UC_2 and US_2 are normally distributed and independent (as was shown in the previous section), $Ia(f_2)$ is a normal variable with zero mean and variance given, using Equation (8), by

$$E[Ia(f_2)]^2 = \sigma^2 \left(\frac{4}{C_2 C_2^2} GC_2^2 \, C_2 C_2 + \frac{4}{S_2 S_2^2} GS_2^2 \, S_2 S_2 \right) = 4 \Delta R_G(f_2) \sigma^2.$$

$Ia(f_1)$ can be written in a slightly simpler form than Equation (8) as $GC_1 = aC_1 C_1$ and $GS_1 = bS_1 S_1$ and some of the factors cancel. Thus

$$Ia(f_1) = 2(aUC_1 + bUS_1); \tag{10}$$

and similarly to $Ia(f_2)$ it is a normal variable with zero mean and variance given by

$$E[Ia(f_1)]^2 = 4R_G \sigma^2.$$

To find the correlation between $Ia(f_1)$ and $Ia(f_2)$ we need their covariance which by use of Equations (8) and (10) becomes

$$\begin{aligned} E[Ia(f_1)\, Ia(f_2)] &= 4 \left[\frac{1}{C_2 C_2} GC_2(aC_1 C_2 + bS_1 C_2) + \right. \\ &\quad \left. + \frac{1}{S_2 S_2} GS_2(aC_1 S_2 + bS_1 S_2) \right] \sigma^2 = \\ &= 4 \left(\frac{1}{C_2 C_2} GC_2^2 + \frac{1}{S_2 S_2} GS_2^2 \right) \sigma^2 \\ &= 4 \Delta R_G(f_2) \sigma^2. \end{aligned}$$

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

We can now find the correlation between $Ia(f_1)$ and $Ia(f_2)$. It is given by

$$\varrho[Ia(f_1), Ia(f_2)] = \frac{E[Ia(f_1)\, Ia(f_2)]}{(E[Ia(f_1)]^2\, E[Ia(f_2)]^2)^{1/2}} = \\ = \frac{4\Delta R_G(f_2)\sigma^2}{4\sigma^2\sqrt{R_G\, \Delta R_G(f_2)}} = \sqrt{p_G(f_2)}.$$

Also the regression coefficient of $Ia(f_2)$ on $Ia(f_1)$ is

$$\beta[Ia(f_2), Ia(f_1)] = \frac{E[Ia(f_2)\, Ia(f_1)]}{E[Ia(f_1)]^2} = \\ = \frac{4\Delta R_G(f_2)\sigma^2}{4R_G\sigma^2} = p_G(f_2). \tag{11}$$

As we now know the statistical behaviour of all three terms of Equation (9), we can consider the statistical behaviour of $p(f_2)$ itself. Specifically, we want the expectation value of $p(f_2)$ given that $p(f_1)$ is affected by noise.

Let

$$p(f_1) = \frac{R_G}{R_G + R_N} + \frac{R_N}{R_G + R_N}\, x + \frac{y}{R_G + R_N} = Z, \tag{12}$$

obtaining if in Equation (9) we have taken $p_N(f_1) = x$ and $Ia(f_1) = y$. Then

$$E[p(f_2) \text{ given } p(f_1) = Z] = \\ = \frac{R_G}{R_G + R_N}\, p_G(f_2) + \frac{R_N}{R_G + R_N}\, E[p_N(f_2) \text{ given } p_N(f_1) = x] + \\ + \frac{1}{R_G + R_N}\, E[Ia(f_2) \text{ given } Ia(f_1) = y] = \\ = \frac{R_G}{R_G + R_N}\, p_G(f_2) + \frac{R_N}{R_G + R_N}\, [\varrho_{12}(x - \bar{x}) + \bar{x}] + \frac{y p_G(f_2)}{R_G + R_N},$$

where we have used Equation (11) and the fact proved in the previous section that ϱ_{12} is the regression coefficient of $p_N(f_2)$ on $p_N(f_1)$. The contents of the square brackets can be rewritten as $\varrho_{12}x + (1 - \varrho_{12})\bar{x}$. It is clear that the second term can only assume values that are small compared to the possible values of the first term, consequently we will make the approximation that the second term equals zero. The accuracy of this approximation will obviously increase as ϱ_{12} approaches 1. From Equation (7) we know that $\varrho_{12} = \bar{p}_G(f_2)$, which is approximately equal to $p_G(f_2)$. Thus the term in square brackets reduces to $\sim p_G(f_2)x$, and so

$$E[p(f_2) \text{ given } p(f_1) = Z] \approx \frac{R_G}{R_G + R_N}\, p_G(f_2) + \frac{R_N}{R_G + R_N}\, x p_G(f_2) + \\ + \frac{y p_G(f_2)}{R_G + R_N}. \tag{13}$$

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

The variance of the first term is zero, the variance of the second term can easily be shown to be approximately a function of $1-p_G^2(f_2)$, and the variance of the third term is a function of $(1-p_G(f_2))p_G(f_2)$. Using Equation (12), Equation (13) becomes

$$E[p(f_2) \text{ given } p(f_1) = Z] \simeq p_G(f_2)\, Z;$$

and finally we find that

$$E\left[\frac{p(f_2)}{p(f_1)}\right] = p_G(f_2).$$

The variance of $p(f_2)/p(f_1)$ is a function of $1-p_G^2(f_2)$ and $(1-p_G(f_2))p_G(f_2)$ and, consequently, approaches zero as $p_G(f_2)$ approaches 1.

Thus the spectrum of a signal which has been affected by noise, after normalization by the height of the highest peak, should closely resemble the noise-free spectrum, especially for the higher aliases.

6. Numerical Examples

A number of examples has been calculated in order to illustrate some of the properties of the LS spectrum that have been found analytically. The examples have been made realistic by basing them on actual observations: radial velocity measurements of the two Beta Canis Majoris stars, β Centauri and α Virginis. The data on β Cen are from Lomb (1975) and consist of 38 measurements obtained over only $1^d.5$, while the α Virginis data, which are taken from Struve and Ebbighausen (1934), are made up of 72 measurements distributed at the two ends of an interval of nearly 5 years.

Example No. 1 – Figure 1 shows the LS spectrum of a sinusoidal periodicity of 8.6 cycles/day frequency (top) and the LS spectrum of a sinusoidal periodicity of 6.4 cycles/day frequency (bottom), both sinusoids sampled at the same times as the β Cen velocities. Points of interest about the figure are that there is no symmetry about the highest peak in either spectrum and that the difference between the two spectra is much more than a simple translation in frequency.

It was shown in Section 2 that the formula used in periodogram analysis gives an approximate value of the reduction in the sum of squares. For Figure 2 the same spectra as in Figure 1 have been plotted, but this time calculated using the approximations of periodogram analysis. On comparison of the two figures it is seen that they are close to being identical; the only differences between them are the heights of the largest peaks. The highest peaks in the top and bottom curves in Figure 2 have heights 93% and 107% respectively, instead of the 100% they must have, by definition, in an LS spectrum. Although such small inaccuracies do not seem important in this case where the correct peaks are well defined, in other cases, where the differences between the heights of peaks are small and the spectra are affected by noise, they could be very disturbing.

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

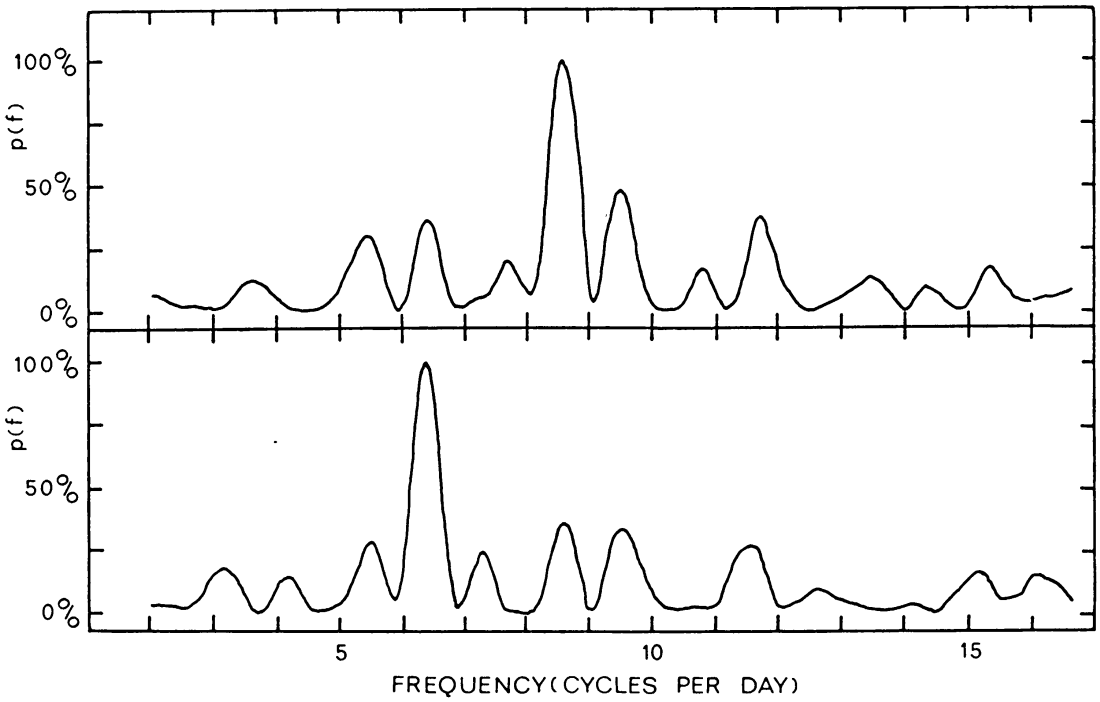

Fig. 1. *Top:* the LS spectrum of a sinusoidal periodicity of 8.6 cycles/day frequency; *Bottom:* the LS spectrum of a sinusoidal periodicity of 6.4 cycles/day frequency. Both sinusoids are sampled at the same time as the β Cen radial velocities.

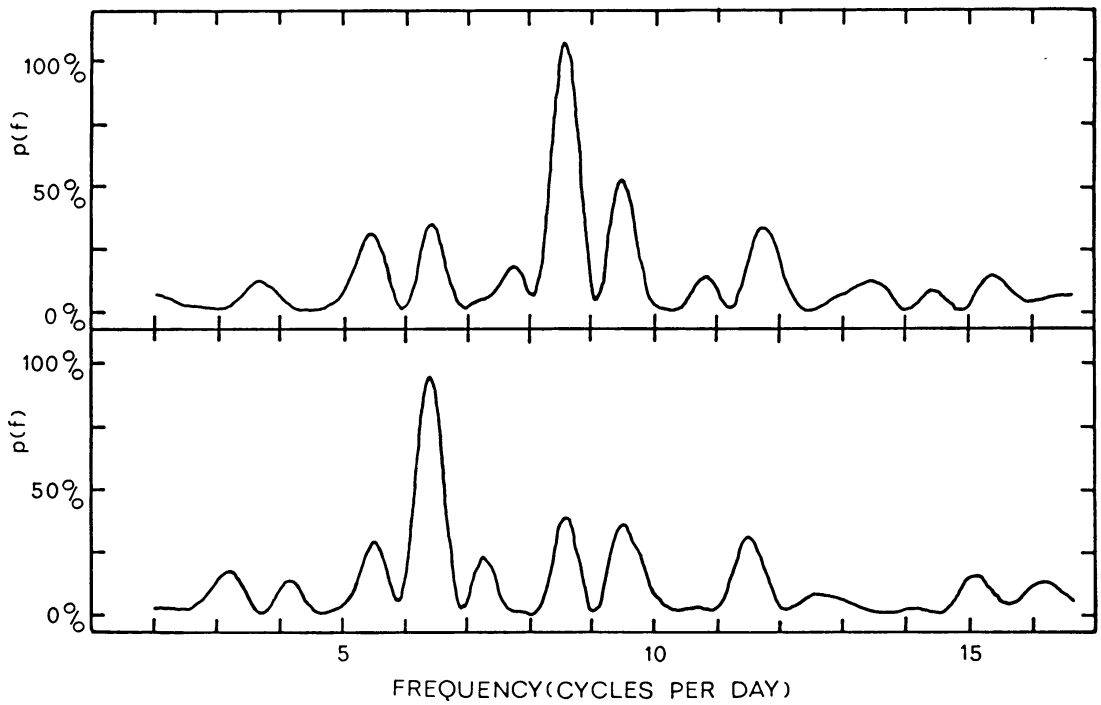

Fig. 2. *Top:* the periodogram of a sinusoidal periodicity of 8.6 cycles/day frequency; *Bottom:* the periodogram of a sinusoidal periodicity of 6.4 cycles/day frequency. Both sinusoids are sampled at the same times as the β Cen radial velocities.

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

Example No. 2 – To test the response of an LS spectrum to noise the β Cen data were again used. The observed velocities were replaced by gaussian noise and the spectrum calculated in the same frequency region as for Figures 1 and 2, that is from 2.0 to 16.66 cycles/day. This was repeated eleven times, each time with a different sequence of random (strictly, quasi-random) noise. Table I indicates the average level and the height of the highest peak in each spectrum.

Using the theory given in Section 4, we can calculate the theoretically predicted mean level. The reduction in the sum of squares at a particular frequency, $\Delta R_N(f)$, was shown to be σ^2 times a χ^2 variable with 2 degrees of freedom, where σ^2 is the population variance. The population variance cannot, of course, be found directly, and so σ^2 will be taken to be equal to the observed variance. This is an acceptable approximation, provided the observed variance is based on a reasonable number of points (say, greater than 30). Under these conditions $p_N(f)$ is also a χ_2^2 variable, multiplied by

$$\frac{\sigma^2}{R_N} = \frac{\sigma^2}{(n-1)\sigma^2} = \frac{1}{n-1},$$

where n is the number of observations. For the mean level in the spectrum we then have

$$E[p_N(f)] = \frac{1}{n-1} E(\chi_2^2) = \frac{2}{n-1}.$$

In the present case $2/(n-1)$ is equal to 0.054 or 5.4%. This predicted value compares favourably with the observed values for the average level listed in Table I.

To be able to perform significance tests on an LS spectrum it is necessary to know the probability distribution of the height of the highest peak in the spectrum. This is impossible to obtain analytically with any degree of accuracy, due chiefly to the correla-

TABLE I

Response to random noise of an LS spectrum (using β Cen R.V. data)

Run	Av. level (%)	Highest peak (%)
1	4.8	15
2	6.1	25
3	4.3	16
4	4.7	24
5	3.3	17
6	7.1	20
7	4.5	14
8	4.9	15
9	3.6	13
10	2.7	13
11	2.8	12

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

tion between the levels of an LS noise spectrum at different frequencies. In some cases it may be worthwhile to establish the probability distribution numerically by calculating the spectra of different sequences of quasi-random noise as in this example. However, instead of eleven calculations of the spectrum, at least a hundred would be necessary.

Example No. 3 – A model of the observed short period variation in the α Vir 1934 velocities (Shobbrook *et al.*, 1972) was set up. This model consisted of a sine wave with a period of $0^d.173\,790$ and an amplitude of 9.1 km s^{-1} plus gaussian noise of 7.1 km s^{-1} standard deviation. The spectra of ten such sets of data were calculated; each set of data contained a different sequence of quasi-random noise. Table II gives the percentage heights of the four nearest aliases to the main period for each of the ten spectra. The heights have been standardized by taking the height of the main peak as 100%. In two of the spectra the highest peak is not at the frequency of the true period. In those cases the heights of the peaks normalized by the height of the highest peak are also given (in brackets). For comparison Table II also gives the heights of the aliases for noise-free spectra of sine waves with periods of P_1, P_2 and P_3.

There is a good match between spectra 1 to 5 and the noise-free spectrum of the correct period, P_3. For spectra 6, 7 and 8 the highest peak is still at P_3, but there is little resemblance to the noise-free spectrum of P_3. The highest peaks for spectra 9

TABLE II

Effect of random noise on the spectrum of a sine wave (using α Vir R.V. data)

	% heights of peaks				
Run	P1	P2	P3	P4	P5
1	61	87	100	90	72
2	63	88	100	88	65
3	58	84	100	93	75
4	58	85	100	91	72
5	59	86	100	92	78
6	69	92	100	86	67
7	55	81	100	98	83
8	58	83	100	96	78
9	81(80)	101(100)	100(99)	82(81)	63(62)
10	40(38)	73(70)	100(96)	104(100)	90(87)

Noise-free spectra

Period	P1	P2	P3	P4	P5
P_3	60	87	100	90	71
P_2	90	100	87	61	39
P_1	31	60	90	100	93

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

and 10 are at P_2 and P_4, respectively. Neither of these spectra has any resemblance to its appropriate noise-free spectrum and in both cases the height of the peak at P_3 is only a few percent less than the height of the highest peak.

From these results the following conclusions can be drawn:

(i) Even with a low signal to noise ratio (in this case 1.3), there is a reasonable probability of a satisfactory match between an observed spectrum and a noise-free spectrum. This is in agreement with the predictions of Section 5, which shows that due to correlation between noise at different frequencies, noise has less effect on a spectrum than could otherwise be expected.

(ii) If there is a satisfactory match between an observed spectrum and a noise-free spectrum of period P, then P is the true period.

(iii) There is a fairly large probability that the highest peak in a spectrum is the correct peak, even with a low signal to noise ratio.

Appendix: Expectation Value of the Product of the Reductions in the Sum of Squares due to Noise at Two Frequencies

Using Equation (2)

$$E[\Delta R_N(f_1)\,\Delta R_N(f_2)] = E[(C^2(f_1) + S^2(f_1))(C^2(f_2) + S^2(f_2))] = \\ = E[C^2(f_1)\,C^2(f_2)] + E[C^2(f_1)\,S^2(f_2)] + \\ + E[S^2(f_1)\,C^2(f_2)] + E[S^2(f_1)\,S^2(f_2)]. \qquad \text{(A1)}$$

As shown in Section 4, when dealing with noise from a normally distributed population, with mean zero and variance σ^2, each of the $C(f_1)$, $C(f_2)$, $S(f_1)$ and $S(f_2)$ are normal variables with mean zero and variance σ^2. The correlation coefficients between $C(f_1)$ and $S(f_1)$, and $C(f_2)$ and $S(f_2)$ are both zero, while the correlation coefficients between $C(f_1)$ and $C(f_2)$, $C(f_1)$ and $S(f_2)$, $S(f_1)$ and $C(f_2)$ and $S(f_1)$ and $S(f_2)$ have been defined as ϱ_{C_1,C_2}, ϱ_{C_1,S_2}, ϱ_{S_1,C_2} and ϱ_{S_1,S_2} respectively.

Consider the first term on the right-hand side of Equation (A1). For simplicity let us put

$$x = C(f_1)/\sigma, \qquad y = C(f_2)/\sigma \quad \text{and} \quad \varrho = \varrho_{C_1,C_2}.$$

Then

$$E[C^2(f_1)\,C^2(f_2)] = \sigma^4 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2y^2 g(x, y, \varrho)\,\mathrm{d}x\,\mathrm{d}y, \qquad \text{(A2)}$$

where $g(x, y, \varrho)$ is the bivariate normal probability function for variables x and y, with zero mean, unit variance and correlation ϱ. Equation 26.3.2 of the *Handbook of Mathematical Functions* (Abramowitz and Stegun, 1964) gives

$$g(x, y, \varrho) = (1 - \varrho^2)^{-1/2} Z(x) Z\left(\frac{y - \varrho x}{\sqrt{1 - \varrho^2}}\right), \qquad \text{(A3)}$$

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

where $Z(x)$ is the normal probability function for a variable with zero mean and unit variance.

Change variables to

$$X = x, \qquad Y = \frac{y - \varrho x}{\sqrt{1 - \varrho^2}}. \tag{A4}$$

Then

$$x = X, \qquad y = (1 - \varrho^2)^{1/2} Y + \varrho X;$$

and, accordingly,

$$x^2 y^2 = (1 - \varrho^2) X^2 Y^2 + \varrho^2 X^4 + 2\varrho(1 - \varrho^2)^{1/2} X^3 Y. \tag{A5}$$

The Jacobian of the transformation is given by

$$J = \begin{vmatrix} \frac{\partial x}{\partial X} & \frac{\partial x}{\partial Y} \\ \frac{\partial y}{\partial X} & \frac{\partial y}{\partial Y} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ \varrho & (1 - \varrho^2)^{1/2} \end{vmatrix} = (1 - \varrho^2)^{1/2}. \tag{A6}$$

Using relations (A3), (A4), (A5) and (A6), Equation (A2) becomes

$$E[C^2(f_1) \cdot C^2(f_2)] = \sigma^4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{(1 - \varrho^2) X^2 Y^2 + \varrho^2 X^4 + 2\varrho(1 - \varrho^2)^{1/2} X^3 Y\} Z(X) Z(Y) \, dX \, dY =$$
$$= \sigma^4 (1 - \varrho^2 + 3\varrho^2),$$

since for a normal distribution with zero mean and unit variance

$$E(x^2) = 1, \qquad E(x^3) = E(x) = 0 \quad \text{and} \quad E(x^4) = 3.$$

Thus

$$E(C^2(f_1) \cdot C^2(f_2)] = \sigma^4 (1 + 2\varrho^2) = \sigma^4 (1 + 2\varrho^2_{c_1, c_2}), \tag{A7}$$

as we had put

$$\varrho = \varrho_{c_1, c_2}.$$

Similarly,

$$E[C^2(f_1) \, S^2(f_2)] = \sigma^4 (1 + 2\varrho^2_{c_1, s_2}), \tag{A8}$$

$$E[S^2(f_1) \, C^2(f_2)] = \sigma^4 (1 + 2\varrho^2_{s_1, c_2}), \tag{A9}$$

and

$$E[S^2(f_1) \, S^2(f_2)] = \sigma^4 (1 + 2\varrho^2_{s_1, s_2}). \tag{A10}$$

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System

1976Ap&SS..39..447L

On substituting Equations (A7) to (A10) in Equation (A1), we obtain

$$E[\Delta R_N(f_1)\,\Delta R_N(f_2)] = 4\sigma^4 + 2\sigma^4(\varrho^2_{c_1,c_2} + \varrho^2_{c_1,s_2} + \varrho^2_{s_1,c_2} + \varrho^2_{s_1,s_2}).$$

Acknowledgements

I would like to thank Dr D. Herbison-Evans for introducing me to least-squares frequency analysis and for reading a preliminary version of this paper. The financial support of a Commonwealth Postgraduate Research Award during the period most of this work was carried out, is gratefully acknowledged.

References

Abramowitz, M. and Stegun, I. A.: 1964, *Handbook of mathematical functions*, National Bureau of Standards, Washington, D.C.
Barning, F. J. M.: 1963, *Bull. Astron. Inst. Neth.* **17**, 22.
Gray, D. F. and Desikachary, K.: 1973, *Astrophys. J.* **181**, 523.
Lomb, N. R.: 1975, *Monthly Notices Roy. Astron. Soc.*, in press.
Shobbrook, R. R., Lomb, N. R., and Herbison-Evans, D.: 1972, *Monthly Notices Roy. Astron. Soc.* **156**, 165.
Struve, O. and Ebbighausen, E.: 1934, *Astrophys. J.* **80**, 365.
Vaníček, P.: 1971, *Astrophys. Space Sci.* **12**, 10.
Wehlau, W. and Leung, K.-C.: 1964, *Astrophys. J.* **139**, 843.

© Kluwer Academic Publishers • Provided by the NASA Astrophysics Data System