REGRESSION, THE PERIODOGRAM, AND THE LOMB-SCARGLE PERIODOGRAM

Barry G. Quinn

Department of Statistics
Faculty of Science and Engineering
Macquarie University
Sydney, NSW, 2109, Australia
email:barry.quinn@mq.edu.au

ABSTRACT

In [1], Lomb developed a nonlinear regression approach to estimating the frequency of a noisy sinusoid when the measurement times were not equispaced, and a method for correcting the times so that the resulting regression sum of squares appeared very similar to the usual periodogram. Scargle [2] suggested that the usual periodogram be discarded, and replaced by the new version, which has become known as the Lomb-Scargle periodogram. In this paper, we extend Lomb's development to include a 'DC' term. We show why it is important to include this term, especially when the times are irregular or the frequency low.

Index Terms— Lomb-Scargle, periodogram, sinusoids, DC term

1. INTRODUCTION

The most general model for a noisy single sinusoid measured at non-equidistant times $t_1, t_2, \ldots, t_N$ is

$$X_n = \mu + \alpha\cos(\omega t_n) + \beta\sin(\omega t_n) + \varepsilon_n. \tag{1}$$

In the seminal article [1], Lomb rejected the periodogram approach to estimating frequency, which depended on the times being equispaced, and developed a nonlinear regression approach, together with an ingenious method of correcting the times t_n so that the resulting regression sum of squares appeared very similar to the usual periodogram. His approach, and the formula stated in [2], have become known as the Lomb-Scargle periodogram, and are in standard use in astronomy. There have also been numerous articles (e.g. [3]) in the engineering literature, extending the approach to damped sinusoids and investigating applications.

In this paper, we revisit [1], and extend his development to include μ, the 'DC' term. We develop the regression sum of squares for (1) , and re-examine the equidistant times case. Finally, we show why it is important to incorporate μ, especially when the times are irregular or the frequency low. It has been known for some time [4] that the usual periodogram is not applicable when estimating a frequency that is low, and that a regression approach should be used.

Note that the frequency $\omega = 2\pi f$ is measured in radians per unit time, and so f is measured in cycles per unit time, rather than Hz.

2. NONLINEAR REGRESSION

The least squares estimators of μ, α, β and ω are found by minimizing

$$S(\mu,\alpha,\beta,\omega) = \sum_{n=1}^{N}\left\{X_n - \mu - \alpha\cos(\omega t_n) - \beta\sin(\omega t_n)\right\}^2. \tag{2}$$

For fixed ω, this is just a linear regression, and the least squares estimators are given by

$$\begin{bmatrix}\widehat{\mu}\\ \widehat{\alpha}\\ \widehat{\beta}\end{bmatrix} = D^{-1}C,$$

where D is symmetric,

$$D = N^{-1}\begin{bmatrix} N & \sum_{n=1}^{N}\cos(\omega t_n) & \sum_{n=1}^{N}\sin(\omega t_n)\\ & \sum_{n=1}^{N}\cos^2(\omega t_n) & \sum_{n=1}^{N}\sin(\omega t_n)\cos(\omega t_n)\\ & & \sum_{n=1}^{N}\sin^2(\omega t_n)\end{bmatrix}$$

$$C = N^{-1}\begin{bmatrix}\sum_{n=1}^{N}X_n\\ \sum_{n=1}^{N}X_n\cos(\omega t_n)\\ \sum_{n=1}^{N}X_n\sin(\omega t_n)\end{bmatrix}.$$

The residual sum of squares is then given by

$$\sum_{n=1}^{N}X_n^2 - N\left(\widehat{\mu}\overline{X} + \widehat{\alpha}C_2 + \widehat{\beta}C_3\right),$$

where C_i denotes the ith element of C, and the regression sum of squares is then

$$\sum_{n=1}^{N}X_n^2 - N\overline{X}^2 - \left\{\sum_{n=1}^{N}X_n^2 - N\left(\widehat{\mu}\overline{X} + \widehat{\alpha}C_2 + \widehat{\beta}C_3\right)\right\}$$
$$= N\left(\widehat{\mu}\overline{X} + \widehat{\alpha}C_2 + \widehat{\beta}C_3\right) - N\overline{X}^2.$$

There is a trick used in any first course in statistics that reduces the above problem to a 2-dimensional rather than 3-dimensional problem. We write (2) as

$$\sum_{n=1}^{N}\left\{X_n - \nu - \alpha\left\{\cos(\omega t_n) - D_{12}\right\} - \beta\left\{\sin(\omega t_n) - D_{13}\right\}\right\}^2,$$

where $\nu = \mu + \alpha D_{12} + \beta D_{13}$ and D_{ij} denotes the (i,j)th element of D, and the residual sum of squares is then

$$\sum_{n=1}^{N} \left(X_n - \overline{X}\right)^2 - N\left(\widehat{\alpha}\widetilde{C}_1 + \widehat{\beta}\widetilde{C}_2\right),$$

where

$$\left[\begin{array}{c} \widehat{\alpha} \\ \widehat{\beta} \end{array}\right] = \widetilde{D}^{-1}\widetilde{C},$$

$$\widetilde{D} = N^{-1}\left[\begin{array}{cc} D_{22} - ND_{12}^2 & D_{23} - ND_{12}D_{13} \\ D_{23} - ND_{12}D_{13} & D_{33} - ND_{13}^2 \end{array}\right] \quad (3)$$

$$\widetilde{C} = N^{-1}\left[\begin{array}{c} \sum_{n=1}^{N}\left(X_n - \overline{X}\right)\cos\left(\omega t_n\right) \\ \sum_{n=1}^{N}\left(X_n - \overline{X}\right)\sin\left(\omega t_n\right) \end{array}\right]$$

$$= N^{-1}\left[\begin{array}{c} C_2 - N\overline{X}D_{12} \\ C_3 - N\overline{X}D_{12} \end{array}\right].$$

The regression sum of squares, as a function of ω, is then

$$P(\omega) = N\left(\widehat{\alpha}\widetilde{C}_1 + \widehat{\beta}\widetilde{C}_2\right) \quad (4)$$

$$= N\widetilde{C}'\widetilde{D}^{-1}\widetilde{C},$$

and it is this function that is maximized so as to estimate ω.

3. THE REGRESSION SUM OF SQUARES AND PERIODOGRAM FOR EQUISPACED DATA

When $t_n = n - 1$, much of the above is simplified, for then D is

$$N^{-1}\left[\begin{array}{ccc} N & \sum_{n=0}^{N-1}\cos\left(\omega n\right) & \sum_{n=0}^{N-1}\sin\left(\omega n\right) \\ & \sum_{n=0}^{N-1}\cos^2\left(\omega n\right) & \sum_{n=0}^{N-1}\sin\left(\omega n\right)\cos\left(\omega n\right) \\ & & \sum_{n=0}^{N-1}\sin^2\left(\omega n\right) \end{array}\right]$$

$$= N^{-1}\left[\begin{array}{ccc} N & \operatorname{Re} g(\omega) & \operatorname{Im} g(\omega) \\ & \left\{N + \operatorname{Re} g(2\omega)\right\}/2 & \frac{1}{2}\operatorname{Im} g(2\omega) \\ & & \left\{N - \operatorname{Re} g(2\omega)\right\}/2 \end{array}\right],$$

where

$$g(\omega) = \sum_{n=0}^{N-1} e^{j\omega n} = \frac{e^{j\omega N} - 1}{e^{j\omega} - 1}.$$

The regression sum of squares is then given by (4). Now, if ω is one of the so-called *canonical* or *Fourier* frequencies

$$\left\{2\pi k/N; 0 \le k \le \lfloor (N-1)/2 \rfloor\right\},$$

there is considerable simplification, for then D is diagonal, and

$$P(\omega) = \frac{2}{N}\left[\left\{\sum_{n=0}^{N-1}\left(X_n - \overline{X}\right)\cos\left(\omega n\right)\right\}^2 + \left\{\sum_{n=0}^{N-1}\left(X_n - \overline{X}\right)\sin\left(\omega n\right)\right\}^2\right]$$

$$= \frac{2}{N}\left|\sum_{n=0}^{N-1}\left(X_n - \overline{X}\right)e^{-j\omega n}\right|^2 \quad (5)$$

which further reduces when $k \ge 1$ to

$$\frac{2}{N}\left|\sum_{n=0}^{N-1} X_n e^{-j\omega n}\right|^2. \quad (6)$$

Moreover, when ω is *not* a Fourier frequency,

$$D = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{array}\right] + O\left(N^{-1}\right),$$

which has led to the use of (5) or (6) as the statistics used to estimate a 'hidden' frequency. There are several things wrong with doing this, however. Firstly, the periodogram is routinely used when N is small, and, secondly, when the true frequency ω is 'small', neither approximation, and especially (6), is accurate enough at low frequency to produce consistent estimators of ω, since $g(2\omega)$ may be quite large.

4. LOMB-STYLE SIMPLIFICATION OF THE REGRESSION SUM OF SQUARES

When the time t_n are equidistant, the forms of the periodogram in (5) and (6) are appealing because of their simplicity and the ability to be computed using fast FFT-based methods. The motivation behind [1, 2] was, for the general case, to obtain a periodogram-like form for the regression sum of squares. However, it appears that Lomb and others believed that the term μ (the 'DC' term), could be eliminated by mean-correction of $\{X_n\}$ at the outset. This can lead to large errors in certain cases, for example when N is small, ω is small, or the time-sampling unusual. Indeed even if ω is not small, exclusion of the times at which the sinusoidal component is negative could lead to biases. This is illustrated in section 6.

We start by examining the obvious diagonalization method, which is not the one that Lomb used. Write $\widetilde{D}$ in Jordan form as

$$\widetilde{D} = Q\Lambda Q'$$

$$Q = \left[\begin{array}{cc} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{array}\right]$$

$$\Lambda = \left[\begin{array}{cc} \lambda_1 & 0 \\ 0 & \lambda_2 \end{array}\right],$$

where, without loss of generality, we assume that $\lambda_1 > \lambda_2$. Then it is easily shown that, with $\widetilde{D}$ defined by (3),

$$\cos(2\phi) = \frac{\widetilde{D}_{11} - \widetilde{D}_{22}}{\lambda_1 - \lambda_2} \quad (7)$$

$$\sin(2\phi) = \frac{2\widetilde{D}_{12}}{\lambda_1 - \lambda_2} \quad (8)$$

and so

$$\tan(2\phi) = \frac{2\widetilde{D}_{12}}{\widetilde{D}_{11} - \widetilde{D}_{22}}.$$

In solving for ϕ, care should be taken to ensure that the solution conforms with the signs of (7) and (8). Hence $P(\omega)$ in (4) becomes

$$N\left(Q'\widetilde{C}\right)'\left[\begin{array}{cc} \lambda_1^{-1} & 0 \\ 0 & \lambda_2^{-1} \end{array}\right]Q'\widetilde{C},$$

where

$$Q'\widetilde{C} = \left[\begin{array}{cc} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{array}\right]N^{-1}\left[\begin{array}{c} \sum_{n=1}^{N}\left(X_n - \overline{X}\right)\cos\left(\omega t_n\right) \\ \sum_{j=1}^{N}\left(X_n - \overline{X}\right)\sin\left(\omega t_n\right) \end{array}\right]$$

$$= N^{-1}\left[\begin{array}{c} \sum_{n=1}^{N}\left(X_n - \overline{X}\right)\cos\left(\omega t_n + \phi\right) \\ \sum_{n=1}^{N}\left(X_n - \overline{X}\right)\sin\left(\omega t_n + \phi\right) \end{array}\right].$$

Thus

$$P(\omega) = \frac{\left\{\sum_{n=1}^{N}\left(X_n - \overline{X}\right)\cos\left(\omega t_n + \phi\right)\right\}^2}{N\lambda_1} + \frac{\left\{\sum_{n=1}^{N}\left(X_n - \overline{X}\right)\sin\left(\omega t_n + \phi\right)\right\}^2}{N\lambda_2},$$

which has a simpler form than (4). However, even though the numerators may be computed using existing algorithms, the forms of the denominators are quite complicated, since

$$\lambda_1 = \frac{\widetilde{D}_{11} + \widetilde{D}_{22} + \sqrt{\Delta}}{2}$$
$$\lambda_2 = \frac{\widetilde{D}_{11} + \widetilde{D}_{22} - \sqrt{\Delta}}{2},$$

where

$$\Delta = \left(\widetilde{D}_{11} - \widetilde{D}_{22}\right)^2 + 4\widetilde{D}_{12}^2.$$

We adopt Lomb's approach, instead. The reason that (4) is complicated is that $\widetilde{D}_{12} \neq 0$. Indeed, if $\widetilde{D}_{12} = 0$, then $P(\omega)$ would be

$$\frac{\left\{\sum_{n=1}^N \left(X_n - \overline{X}\right)\cos(\omega t_n)\right\}^2}{N\widetilde{D}_{11}} + \frac{\left\{\sum_{n=1}^N \left(X_n - \overline{X}\right)\sin(\omega t_n)\right\}^2}{N\widetilde{D}_{22}}.$$

We thus write (1) as

$$X_n = \mu + A\cos(\omega t_n - \phi) + B\sin(\omega t_n - \phi) + \varepsilon_n,$$

with $\phi = \omega\tau$ yet to be determined. The same method as in section 2 will be used to eliminate the DC term. We minimize

$$\sum_{n=1}^N \left[X_n - \nu - A\left\{\cos(\omega t_n - \phi) - E_1\right\} - B\left\{\sin(\omega t_n) - E_2\right\}\right]^2,$$

where

$$E_1 = N^{-1}\sum_{n=1}^N \cos(\omega t_n - \phi)$$
$$E_2 = N^{-1}\sum_{n=1}^N \sin(\omega t_n - \phi),$$

with respect to ν, A and B, for fixed ω, choosing ϕ so as to make the columns of the design matrix orthogonal, i.e. so that the analog of $\widetilde{D}_{12}$ is 0. Note that E_1 and E_2 depend on ϕ. Now

$$\begin{aligned}
&\sum_{n=1}^N \left\{\sin(\omega t_n - \phi) - E_2\right\}\left\{\cos(\omega t_n - \phi) - E_1\right\} \\
&= \frac{1}{2}\sum_{n=1}^N \sin\left(2\left(\omega t_n - \phi\right)\right) - NE_1E_2 \\
&= \frac{1}{2}\sum_{n=1}^N \sin(2\omega t_n)\cos(2\phi) - \sum_{n=1}^N \cos(2\omega t_n)\sin(2\phi) \\
&- N^{-1}\left\{\sum_{n=1}^N \cos(\omega t_n)\cos\phi + \sum_{n=1}^N \sin(\omega t_n)\sin\phi\right\} \\
&\times \left\{\sum_{n=1}^N \sin(\omega t_n)\cos\phi - \sum_{n=1}^N \cos(\omega t_n)\sin\phi\right\} \\
&= -B\sin(2\phi - \xi),
\end{aligned}$$

where

$$2B\sin\xi = \sum_{n=1}^N \sin(2\omega t_n) - 2N^{-1}\sum_{n=1}^N \sin(\omega t_n)\sum_{n=1}^N \cos(\omega t_n)$$
$$\begin{aligned}
2B\cos\xi &= \sum_{n=1}^N \cos(2\omega t_n) - N^{-1}\left\{\sum_{n=1}^N \sin(\omega t_n)\right\}^2 \\
&+ N^{-1}\left\{\sum_{n=1}^N \cos(\omega t_n)\right\}^2,
\end{aligned}$$

and so the columns becomes orthogonal when $\phi = \xi/2$, i.e. when

$$\tan(2\phi) = \frac{\sum_{n=1}^N \sin(2\omega t_n) - 2ND_{12}D_{13}}{\sum_{n=1}^N \cos(2\omega t_n) + ND_{12}^2 - ND_{13}^2}. \tag{9}$$

The regression sum of squares is then

$$P(\omega) = \frac{\left\{\sum_{n=1}^N \left(X_n - \overline{X}\right)\cos(\omega t_n - \phi)\right\}^2}{\sum_{n=1}^N \cos^2(\omega t_n - \phi) - NE_1^2} + \frac{\left\{\sum_{n=1}^N \left(X_n - \overline{X}\right)\sin(\omega t_n - \phi)\right\}^2}{\sum_{n=1}^N \sin^2(\omega t_n - \phi) - NE_2^2} \tag{10}$$

or

$$\frac{\left\{\sum_{n=1}^N X_n\cos(\omega t_n - \phi) - N\overline{X}E_1\right\}^2}{\sum_{n=1}^N \cos^2(\omega t_n - \phi) - NE_1^2} + \frac{\left\{\sum_{n=1}^N X_n\sin(\omega t_n - \phi) - N\overline{X}E_2\right\}^2}{\sum_{n=1}^N \sin^2(\omega t_n - \phi) - NE_2^2}. \tag{11}$$

These formulae should be compared with Lomb's

$$\frac{\left\{\sum_{n=1}^N X_n\cos(\omega t_n - \phi)\right\}^2}{\sum_{n=1}^N \cos^2(\omega t_n - \phi)} + \frac{\left\{\sum_{n=1}^N X_n\sin(\omega t_n - \phi)\right\}^2}{\sum_{n=1}^N \sin^2(\omega t_n - \phi)}, \tag{12}$$

or what has been suggested to be used, the mean-corrected form

$$\frac{\left\{\sum_{n=1}^N \left(X_n - \overline{X}\right)\cos(\omega t_n - \phi)\right\}^2}{\sum_{n=1}^N \cos^2(\omega t_n - \phi)} + \frac{\left\{\sum_{n=1}^N \left(X_n - \overline{X}\right)\sin(\omega t_n - \phi)\right\}^2}{\sum_{n=1}^N \sin^2(\omega t_n - \phi)}. \tag{13}$$

The differences are in the definition of ϕ and the denominator terms, but these may be quite substantial if E_1 or E_2 are significant. Finally, we note that [3] has raised the question about computational problems in computing ϕ. For these reasons, although the expressions for $P(\omega)$ are elegant, it might be better from the computational point of view just to use the regressions sum of squares given by (4).

5. SPECIAL CASE: EQUISPACED DATA

When $t_n = n - 1$,

$$D_{12} = N^{-1}\operatorname{Re} g(\omega), D_{13} = N^{-1}\operatorname{Im} g(\omega)$$
$$\sum_{n=1}^N \cos(2\omega t_n) = \operatorname{Re} g(2\omega),$$
$$\sum_{n=1}^N \sin(2\omega t_n) = \operatorname{Im} g(2\omega)$$
$$E_1 = D_{12}\cos\phi + D_{13}\sin\phi,$$
$$E_2 = D_{13}\cos\phi - D_{12}\sin\phi$$
$$\sum_{n=1}^N \cos\left\{2\left(\omega t_n - \phi\right)\right\} = \operatorname{Re}\left\{e^{-j\phi}g(2\omega)\right\}.$$

Thus (10) is easily computed exactly. The numerator terms in (11) are most likely best computed using

$$\sum_{n=1}^N X_n e^{j(\omega t_n - \phi)} = e^{-j\phi}\sum_{n=1}^N X_n e^{j\omega t_n}.$$

6. NUMERICAL EXPLORATION

In the following examples, we have simulated $\{X_n\}$ according to (1) with $\mu = 1, \alpha = 1, \beta = 0, \omega = 2\pi f$. In all cases, the ε_n were simulated normally distributed with mean 0 and variance 0.2. In the figures, we show $P(\omega)$ given by (4) and (10), which is termed 'Regression' in the legend, the mean-corrected Lomb-Scargle periodogram given by (13), termed 'LS Mean corrected', and the raw

version given by (12), termed 'LS Raw'. In Figure 1, where $f = 0.256$, and $N = 100$, time-spacings were generated that were independent and uniformly distributed on $(0, 1)$, and the Regression and Lomb-Scargle mean-corrected versions are nearly indistinguishable, but very different from the raw version. In Figure 2, we show the actual differences between the Regression and Lomb-Scargle mean-corrected versions in this case. Noticeable are the differences near f and 0. For the other two cases, we show only the difference between the Regression and Lomb-Scargle mean-corrected versions, as they are similar, and very different from the uncorrected version. Figure 3 repeats the first experiment, but with 'low frequency', $f = 0.035$. It is seen there that the mean-corrected Lomb-Scargle periodogram is quite different from the Regression periodogram. Figure 4 is for the case where $N = 1024$ and $f = 0.1238$, but with integer spacings for which all of the times where $\cos(\omega t) < 0$ have been excluded. It appears that only the values very near the true frequency differ. However, this difference is quite large and could lead to discrepancies, especially if the periodogram is used for detection.

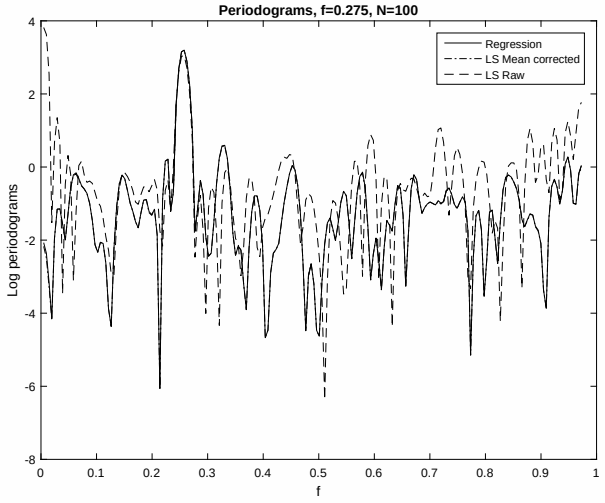

Fig.1 Periodograms

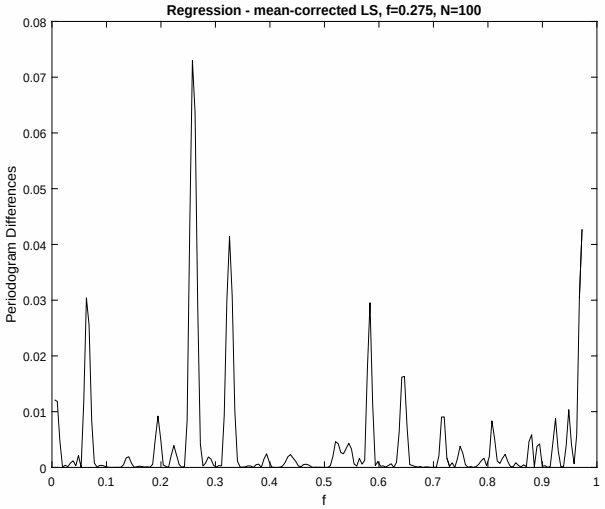

Fig. 2 Differences between regression and mean-corrected LS periodograms

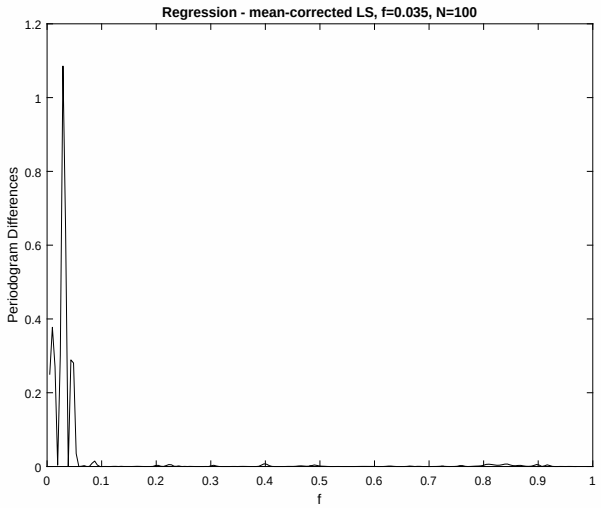

Fig. 3 Differences between regression and mean-corrected LS periodograms, low frequency

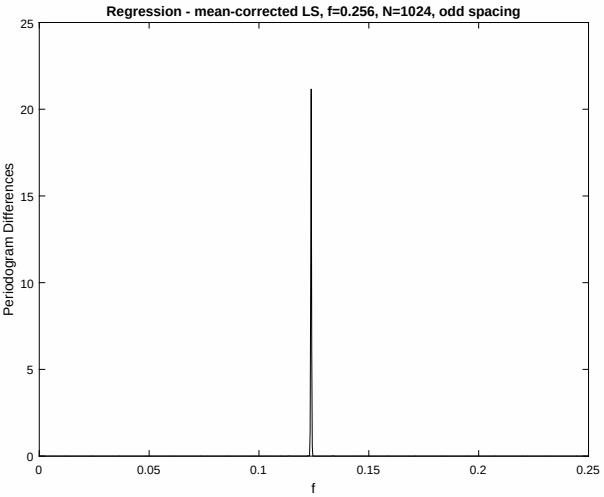

Fig. 4 Differences between regression and mean-corrected LS periodograms, odd spacing

7. CONCLUSION

The Lomb-Scargle periodogram has been extended to include an unknown DC term. Rather than mean-correcting the data, the DC offset has been included as a parameter to be estimated, and a simple formula derived. The development may be readily extended to the complex data case. What has not been done is an asymptotic analysis of the maximizer of the extended Lomb-Scargle periodogram, in the style of [5]. This may be difficult to do unless it is assumed that $\{\varepsilon_n\}$ is white. However, the more realistic assumption is that $\varepsilon_n = e_{t_n}$, where $\{e_t\}$ is a continuous-time stochastic process with some unknown continuous spectral density. To the author's knowledge, a rigorous central limit theorem has not been developed for the Lomb-Scargle periodogram maximizer, even when $\mu = 0$ and $\{\varepsilon_n\}$ is Gaussian and white.

8. REFERENCES

[1] N.R. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophysics and Space Science*, vol. 39, no. 2, pp. 447–462, 1976.

[2] J.D. Scargle, "Studies in astronomical time series analysis. ii - statistical aspects of spectral analysis of unevenly spaced data," *Astrophysical Journal*, vol. 263, pp. 835–853, December 1982.

[3] Petre Stoica, Jian Li, and Hao He, "Spectral analysis of nonuniformly sampled data: A new approach versus the periodogram," *Signal Processing, IEEE Transactions on*, vol. 57, no. 3, pp. 843–858, March 2009.

[4] E. J. Hannan and B. G. Quinn, "The resolution of closely adjacent spectral lines," *Journal of Time Series Analysis*, vol. 10, no. 1, pp. 13–31, 1989.

[5] E.J. Hannan, "The estimation of frequency," *J. App. Prob*, vol. 10, pp. 510–519, 1973.