Regression, the periodogram, and the Lomb-Scargle periodogram

MACQUARIE UNIVERSITY | FACULTY OF SCIENCE

Barry Quinn, Macquarie University

Introduction

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

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

In [1], Lomb rejected the periodogram approach to estimating frequency, which depended on the times being equispaced, and developed a least-squares 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. Lomb's function, and that in [2], have become known as the Lomb-Scargle periodogram, and are in standard use in astronomy. There have 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 include a '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 the DC term, 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 $\omega = 2\pi f$ is measured in radians per unit time, and so f is measured in cycles per unit time, and not Hz.

Nonlinear Regression

The least squares estimators of μ, α, β and ω minimise

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

For fixed ω, (1) is just a linear regression, and the least squares estimators are given by

$$\left[\widehat{\mu}\ \widehat{\alpha}\ \widehat{\beta}\right]' = D^{-1}C,$$

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

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

and D is symmetric. The regression sum of squares is

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

If 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, then the regression sum of squares is now

$$P(\omega) = N\left(\widehat{\alpha}\widetilde{C}_1 + \widehat{\beta}\widetilde{C}_2\right) = N\widetilde{C}'\widetilde{D}^{-1}\widetilde{C}, \quad (3)$$

where $\left[\widehat{\alpha}\ \widehat{\beta}\right]' = \widetilde{D}^{-1}\widetilde{C}$,

$$\widetilde{D} = N^{-1}\begin{bmatrix} D_{22} - ND_{12}^2 & D_{23} - ND_{12}D_{13} \\ D_{23} - ND_{12}D_{13} & D_{33} - ND_{13}^2\end{bmatrix} \quad (4)$$

$$\widetilde{C} = N^{-1}\begin{bmatrix} \sum_{n=1}^N \left(X_n - \overline{X}\right)\cos(\omega t_n) \\ \sum_{n=1}^N \left(X_n - \overline{X}\right)\sin(\omega t_n)\end{bmatrix} = N^{-1}\begin{bmatrix} C_2 - N\overline{X}D_{12} \\ C_3 - N\overline{X}D_{12}\end{bmatrix}.$$

and it is this function $P(\omega)$ that is maximized so as to estimate ω.

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

$$N^{-1}\begin{bmatrix} N & \sum_{n=0}^{N-1}\cos(\omega n) & \sum_{n=0}^{N-1}\sin(\omega n) \\ & \sum_{n=0}^{N-1}\cos^2(\omega n) & \sum_{n=0}^{N-1}\sin(\omega n)\cos(\omega n) \\ & & \sum_{n=0}^{N-1}\sin^2(\omega n)\end{bmatrix}$$
$$= N^{-1}\begin{bmatrix} N & \operatorname{Re} g(\omega) & \operatorname{Im} g(\omega) \\ & \left\{N + \operatorname{Re} g(2\omega)\right\}/2 & \frac{1}{2}\operatorname{Im} g(2\omega) \\ & & \left\{N - \operatorname{Re} g(2\omega)\right\}/2\end{bmatrix},$$

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 (3). If ω is one of the so-called *canonical* or *Fourier* frequencies

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

D is diagonal, and

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

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,

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

which has led to the use of (5) or (6) as the statistics used to estimate or detect 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.

Lomb-style Simplification of the Regression Sum of Squares

When the times t_n are equidistant, (5) and (6) may 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 later.

The obvious Jordan-form diagonalization method, not the one that Lomb used, does not result in a useful formula.

We adopt Lomb's approach, instead. The reason that (3) is complicated is that $\widetilde{D}_{12} \ne 0$. Indeed, if $\widetilde{D}_{12} = 0$, then $P(\omega)$ would be

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

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 before will be used to eliminate the DC term. We minimize

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

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. It is easily shown that

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

for some B, where

$$\tan\xi = \frac{\sum_{n=1}^N \sin(2\omega t_n) - 2ND_{12}D_{13}}{\sum_{n=1}^N \cos(2\omega t_n) + ND_{12}^2 - ND_{13}^2}. \quad (7)$$

When $\phi = \xi/2$, the regression sum of squares is then

$$P(\omega) = \frac{\left\{\sum_{n=1}^N (X_n - \overline{X})\cos(\omega t_n - \phi)\right\}^2}{\sum_{n=1}^N \cos^2(\omega t_n - \phi) - NE_1^2} + \frac{\left\{\sum_{n=1}^N (X_n - \overline{X})\sin(\omega t_n - \phi)\right\}^2}{\sum_{n=1}^N \sin^2(\omega t_n - \phi) - NE_2^2} \quad (8)$$

or

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

These formulae should be compared with Lomb's

$$\frac{\left\{\sum_{n=1}^N X_n\cos(\omega t_n - \phi)\right\}^2}{\sum_{n=1}^N \cos^2(\omega t_n - \phi)} + \frac{\left\{\sum_{n=1}^N X_n\sin(\omega t_n - \phi)\right\}^2}{\sum_{n=1}^N \sin^2(\omega t_n - \phi)}, \quad (9)$$

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

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

The differences are in the definitions 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 regression sum of squares given by (3).

In the special case of equispaced data, (8) is easily computed exactly.

Numerical exploration

In the following examples, $\mu = 1, \alpha = 1, \beta = 0, \omega = 2\pi f$. The ε_n were simulated normally distributed with mean 0 and variance 0.2. In the figures, we show $P(\omega)$ given by (3) and (8), which is termed 'Regression' in the legend, the mean-corrected Lomb-Scargle periodogram given by (10), termed 'LS Mean corrected', and the raw version given by (9), termed 'LS Raw'. In Figure 1, where $f = 0.256$, and $N = 100$, time-spacings were independent and uniformly distributed on $(0, 1)$. 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. 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', $f = 0.035$. 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. Only the values very near the true frequency differ. This difference is quite large and could lead to discrepancies, especially if the periodogram is used for detection.

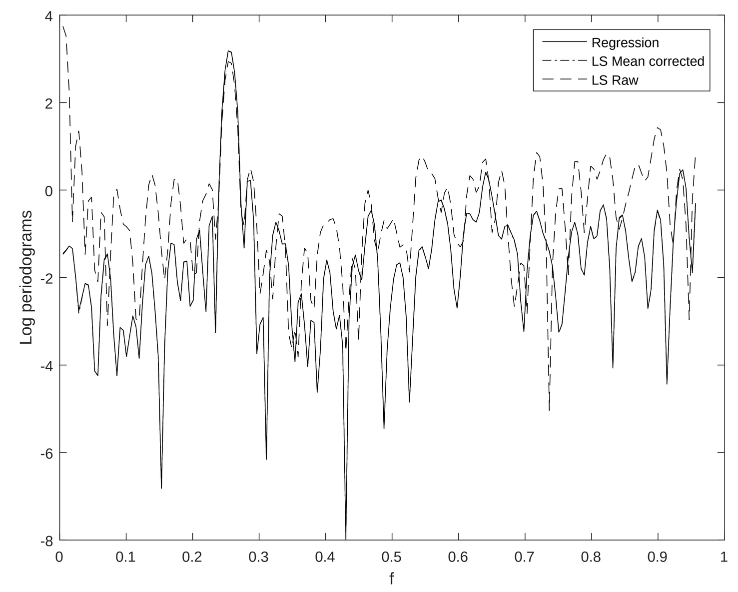

Figure 1: Periodograms

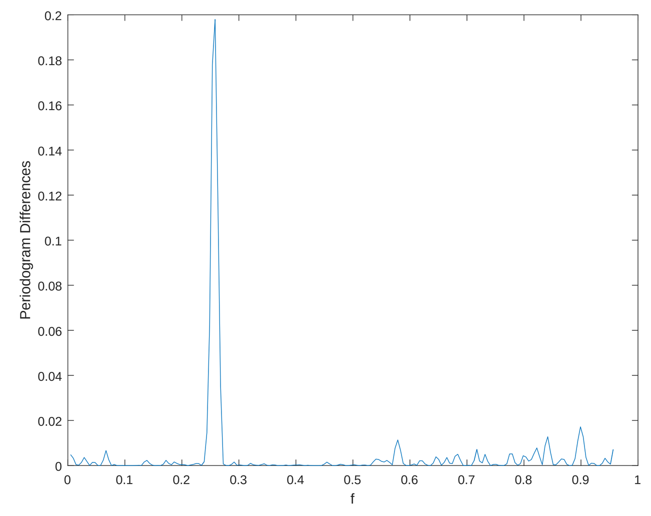

Figure 2: Differences between regression and mean-corrected LS periodograms

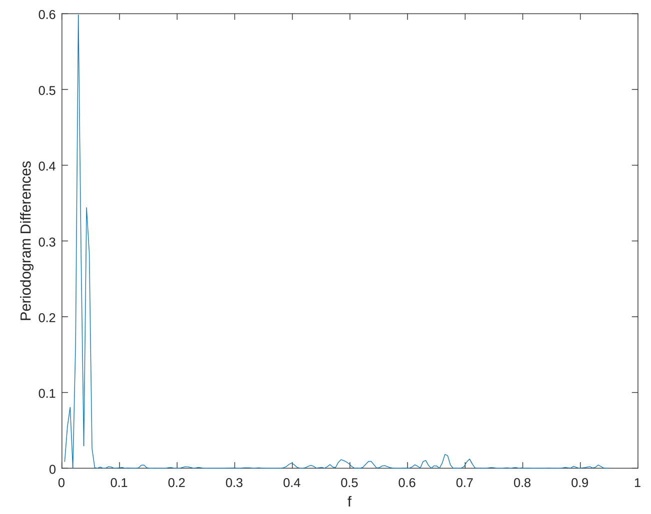

Figure 3: Differences between regression and mean-corrected LS periodograms, low frequency

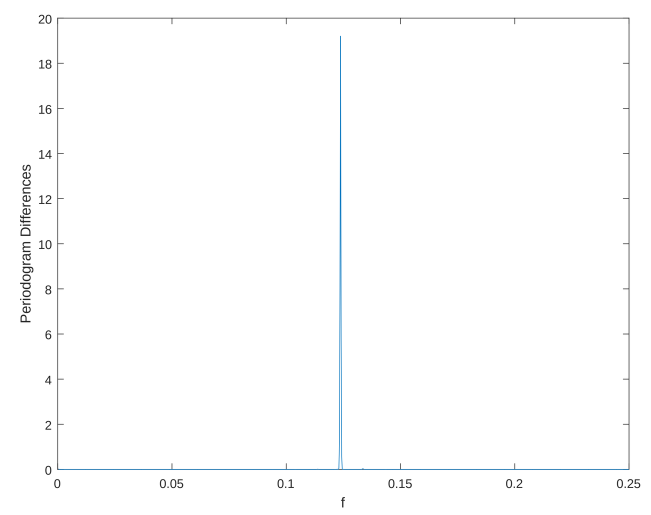

Figure 4: Differences between regression and mean-corrected LS periodograms, unusual spacing

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 complex data.

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.