Regression, the periodogram, and the Lomb-Scargle periodogram

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

\[
X_n = \mu + \alpha \cos(\omega t_n) + \beta \sin(\omega t_n) + \epsilon_n. \tag{1}
\]

In [1], Lomb rejected the periodogram approach to estimating frequency, which depended on the times being equispaced, and developed a least-squares approach, altogether 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 at the frequency low. It has been shown 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 \( \mu, \alpha, \beta \) and \( \omega \) minimise

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

(2)

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

\[
\hat{\mu} = \frac{\sum N \omega t_n}{N \sum \omega t_n^2}, \quad \hat{\alpha} = \frac{\sum X_n \omega t_n}{N \sum \omega t_n^2}, \quad \hat{\beta} = \frac{\sum X_n \omega t_n^2}{N \sum \omega t_n^2}, \quad \hat{\omega} = \frac{\sum X_n \omega t_n^2}{N \sum \omega t_n^2} - \frac{\sum \omega t_n^2}{N \sum \omega t_n^2}.
\]

The regression sum of squares is then given by (5). If \( \omega \) is one of the so-called canonical Fourier frequencies (2\(k\pi/N \) for \( k/N \leq 1/2 \)), the periodogram, and

\[
D = \sum_{n=1}^{N} X_n \omega t_n^2,
\]

is diagonal, and

\[
P(\omega) = \sum_{n=1}^{N} (X_n - \mu - \alpha \cos(\omega t_n) - \beta \sin(\omega t_n))^2
\]

(5)

which further reduces when \( k \geq 1/2 \) to

\[
\sum_{n=1}^{N} X_n^2 - \sum_{n=1}^{N} X_n \omega t_n + \sum_{n=1}^{N} \omega t_n^2 \sum_{n=1}^{N} \omega t_n^2 - \sum_{n=1}^{N} \omega t_n^2.
\]

Moreover, when \( \omega \) is not a Fourier frequency, the periodogram, and

\[
D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/6 \end{bmatrix} + \omega \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}
\]

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 essentially used when \( N \) is small, and, secondly, when the true frequency \( \omega \) is small, ‘nearly’ approximation, and especially \( \epsilon \), is accurate enough at low frequency to produce consistent estimators of \( \omega \), since \( g(2\pi) \) 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 \( \mu \) (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, \( \omega \) is small, or the time-sampling unusual. Indeed even if \( \omega \) is not small, omission of the times at which the sine- and cosine 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 \( D \) is not identity. Indeed, \( D = \mu \) is zero, then \( P(\omega) \) would be

\[
\sum_{n=1}^{N} (X_n - \mu - \alpha \cos(\omega t_n) - \beta \sin(\omega t_n))^2.
\]

We thus write (1) as

\[
N \sum_{n=1}^{N} \cos(\omega t_n) \sum_{n=1}^{N} \sin(\omega t_n), \quad N \sum_{n=1}^{N} \cos(\omega t_n), \quad N \sum_{n=1}^{N} \sin(\omega t_n), \quad N \sum_{n=1}^{N} \cos(\omega t_n) \sum_{n=1}^{N} \sin(\omega t_n).
\]

\[
\tilde{D} = D - \sum_{n=1}^{N} \cos(\omega t_n) \sum_{n=1}^{N} \sin(\omega t_n), \quad \tilde{D} - \sum_{n=1}^{N} \cos(\omega t_n), \quad \tilde{D} - \sum_{n=1}^{N} \sin(\omega t_n), \quad \tilde{D} - \sum_{n=1}^{N} \cos(\omega t_n) \sum_{n=1}^{N} \sin(\omega t_n).
\]

\[
\tilde{D} = N \sum_{n=1}^{N} \cos(\omega t_n) \sum_{n=1}^{N} \sin(\omega t_n), \quad N \sum_{n=1}^{N} \cos(\omega t_n), \quad N \sum_{n=1}^{N} \sin(\omega t_n), \quad N \sum_{n=1}^{N} \cos(\omega t_n) \sum_{n=1}^{N} \sin(\omega t_n).
\]

Numerical exploration

In the following examples, \( \mu = 1, \alpha = 1, \beta = 0, \omega = 2\pi f \). The \( \mu \)’s were simulated normally distributed with mean 0 and variance 0.2. In the figures, we show \( P(\omega) \) as given by (1) 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 \( f = 0.25 \), 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. Figure 2, we show the actual differences between the Regression and Lomb-Scargle mean-corrected versions. Noticeable are the differences near \( f = 0 \) and \( 0.12 \). For the other two cases, we show only the difference between the Regression and Lomb-Scargle mean-corrected versions. Figure 3 repeats the first experiment, but with ’low frequency’, \( f = 0.003 \). The mean-corrected Lomb-Scargle periodogram is quite different from the uncorrected version. Figure 4 is for the case where \( N = 1024 \) and \( f = 0.1259 \), but with integer spacings for which all of the \( \cos(\omega t_n) \)s 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.

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


