Astronomy and Computing 1 (2013) 5–16

Contents lists available at SciVerse ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

Irregular time series in astronomy and the use of the Lomb–Scargle periodogram

R. Vio[a], M. Diaz-Trigo[b], P. Andreani[b,c,*]

[a] Chip Computers Consulting s.r.l., Viale Don L. Sturzo 82, S. Liberale di Marcon, 30020 Venice, Italy

[b] ESO, Karl Schwarzschild strasse 2, 85748 Garching, Germany

[c] INAF-Osservatorio Astronomico di Trieste, via Tiepolo 11, 34143 Trieste, Italy

ARTICLE INFO

Article history:
Received 28 August 2012
Accepted 13 December 2012

Keywords:
Methods: data analysis
Methods: statistical

ABSTRACT

Detection of a signal hidden by noise within a time series is an important problem in many astronomical searches, i.e. for light curves containing the contributions of periodic/semi-periodic components due to rotating objects and all other astrophysical time-dependent phenomena. One of the most popular tools for use in such studies is the *periodogram*, whose use in an astronomical context is often not trivial. The *optimal* statistical properties of the periodogram are lost in the case of irregular sampling of signals, which is a common situation in astronomical experiments. Parts of these properties are recovered by the *Lomb–Scargle* (LS) technique, but at the price of theoretical difficulties, that can make its use unclear, and of algorithms that require the development of dedicated software if a fast implementation is necessary. Such problems would be irrelevant if the LS periodogram could be used to significantly improve the results obtained by approximated but simpler techniques. In this work we show that in many astronomical applications, simpler techniques provide results similar to those obtainable with the LS periodogram. The meaning of the *Nyquist frequency* is also discussed in the case of irregular sampling.

1. Introduction

The search for characteristic frequencies in astrophysical phenomena requires a careful analysis of the data with appropriate statistical tools. Given the simplicity of its use and the wide availability of efficient related software, one of the most popular techniques for looking for periodicities within a time series is the periodogram technique. In astronomical applications, however, the use of this technique is not trivial. In fact, this tool exhibits its *optimal* properties only in the case of signals sampled on a regular time grid, a common situation in engineering applications but not always in astronomical experiments. The analysis of a periodogram in the case of irregular sampling is often limited by the possibilities for fully fixing its statistical properties. This is an old problem (see e.g. Gottlieb et al., 1975) and there have been many attempts to solve it. A partial solution has been found in the *Lomb–Scargle* (LS) approach (Lomb, 1976; Scargle, 1982), but at the price of theoretical difficulties that make its use unclear and, if a fast implementation is needed (e.g. in the case of very long time series), the necessity of dedicated software. Of course, this would not constitute a relevant issue if the LS periodogram could be used to notably improve the results obtainable by the statistical analysis of a time series. In this paper we argue that in astronomical applications, often this is not the case. We show how the negligible improvements obtained with LS are offset by the ease of interpretation and clarity of the results provided by simpler techniques, which do not demand high computing power and/or complicated algorithms.

In Section 2 the statistical analysis of sampled signals is addressed in the case of a regular sampling, where the mathematical notation and formalism are also outlined. The problems and advantages of an irregular sampling are analyzed in Section 3. The real advantage of the LS periodogram with respect to an approximated but simpler technique is considered in Section 4 on the basis of theoretical arguments as well as numerical experiments based on synthetic data and an experimental time series. Finally, Section 5 derives our conclusions.

2. Statistical analysis of regularly sampled signals

If a signal $x(t)$ is sampled on a regular time grid with a constant time step Δt, a time series $\{x_j\}_{j=0}^{N-1} \equiv (x_0, x_1, \ldots, x_{N-1})$ is obtained[1]. Often the main problem is testing whether $x(t)$ is due only

* Corresponding author at: ESO, Karl Schwarzschild strasse 2, 85748 Garching, Germany.
E-mail address: pandrean@eso.org (P. Andreani).

[1] Typically it is assumed that $\Delta t = 1$.

2213-1337/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.ascom.2012.12.001

to a noise $n(t)$ or whether some other component $s(t)$ is present, i.e. $x_j = s_j + n_j$. The most popular approach consists of computing the *periodogram* $\{p_k\}_{k=0}^{N-1}$ for a set of N equispaced frequencies $\{f_k\}_{k=0}^{N-1} \equiv \{k/N\}$: $p_k = \frac{1}{N} |\widehat{x}_k|^2$ with the *discrete Fourier transform* (DFT) of $\{x_j\}$ being

$$\widehat{x}_k = \sum_{j=0}^{N-1} x_j \mathrm{e}^{-i2\pi kj/N}, \quad k = 0, 1, \ldots, N-1; \tag{1}$$

and $\{f_k\}$ being the *Fourier frequencies*. The original time series $\{x_j\}$ can be recovered from $\{\widehat{x}_k\}$ via

$$x_j = \frac{1}{N} \sum_{k=0}^{N-1} \widehat{x}_k \mathrm{e}^{i2\pi kj/N}, \quad j = 0, 1, \ldots, N-1. \tag{2}$$

In the case where $\{x_j\}$ is only noise with $\{n_j\}$ a zero-mean, Gaussian, white-noise stationary process with standard deviation σ_n, from Eq. (1) it can be readily verified that, independently of k, $\widehat{p}_k/\sigma_n^2$ is given by the sum of two squared independent, zero-mean, unit-variance, Gaussian random quantities. As a consequence, the corresponding *probability density function* (PDF) is the exponential distribution. Moreover, whenever $k \neq k'$ with $k, k' = 0, 1, \ldots, N/2$, p_k is independent of $p_{k'}$. Hence, the probability α that at least one of the p_k is expected to exceed a level L_{Fa} is

$$\alpha = 1 - \left[1 - \mathrm{e}^{-p_k/\sigma_n^2}\right]^{N^*}. \tag{3}$$

Through this quantity it is possible to fix a detection threshold L_{Fa},

$$L_{\mathrm{Fa}} = -\sigma_n^2 \ln\left[1 - (1-\alpha)^{1/N^*}\right], \tag{4}$$

corresponding to the level that one or more peaks due to the noise would exceed with a pre-fixed probability α when a number N^* of (*statistically independent*) frequencies are inspected. Threshold L_{Fa} is called the *level of false alarm*.

For a periodic component with amplitude A, phase ϕ_l and frequency f_l (in units of $1/\Delta t$) in the set of the Fourier frequencies $\{f_k\}$, $s_j = A\sin(2\pi f_l t_j + \phi_l)$, the periodogram will show a prominent peak at $k = l$. Indeed, since $\widehat{x}_{N-k+1}$ is the complex conjugate of $\widehat{x}_k$, then $\cos[2\pi(N-k+1)j] = \cos[2\pi kj]$ and $\sin[2\pi(N-k+1)j] = -\sin[2\pi kj]$. Hence, Eq. (2) can be written in the form (Chu, 2008)

$$x_j = \frac{1}{N} \sum_{k=0}^{N-1} a_k \cos\frac{2\pi kj}{N} + b_k \sin\frac{2\pi kj}{N}, \tag{5}$$

where

$$a_k = \sum_{j=0}^{N-1} x_j \cos\frac{2\pi kj}{N}; \tag{6}$$

$$b_k = \sum_{j=0}^{N-1} x_j \sin\frac{2\pi kj}{N}, \tag{7}$$

or

$$a_k = \frac{\widehat{x}_k + \widehat{x}_{N-k+1}}{2}; \tag{8}$$

$$b_k = i\frac{\widehat{x}_k - \widehat{x}_{N-k+1}}{2}. \tag{9}$$

Now, since

$$s_j = a_l \cos\frac{2\pi lj}{N} + b_l \sin\frac{2\pi lj}{N}, \tag{10}$$

only the coefficients $\widehat{x}_l$ and $\widehat{x}_{N-l}$ and hence only $\widehat{p}_l = (a_l^2 + b_l^2)/N$ will be different from zero. More generally, if $x_j = A\sin(2\pi f_l^* t_j + \phi) + n_j$, with f_l^* close but not identical to the Fourier frequency f_l, the periodogram takes the form of a squared "sinc" function centered at f_l^*. Also in this case, it is expected that $p_l > L_{\mathrm{Fa}}$ for small values of α (typically 0.05 or 0.01). If $s(t)$ is semi-periodic or even non-periodic, the situation is more complicated since more peaks are expected, but the basic idea does not change.

Regular sampling has many advantages, among them:

- The sine and cosine modes corresponding to the *Fourier frequencies* constitute an orthonormal basis for signal $\{x_j\}$. This makes operations such as noise filtering, separation and/or detection of components of interest easier.
- The spectrogram can be shown to derive from the least-squares fit of model (5) to the observed signal (see e.g. see Vio et al., 2010). This provides a physical interpretation of the quantity p_k as energy associated with the component at frequency f_k.
- Under the pure noise hypothesis, $x_j = n_j$ and, independently of k, a_k and b_k are uncorrelated (independent) Gaussian quantities. As a consequence p_k contains all the available information. In other words, the use of the joint distribution of a_k and b_k does not provide any advantage with respect to the use of p_k. Moreover, the quantities $\{p_k\}_{k=0}^{N/2}$ are mutually independent and have a known PDF. All of these facts permit the development of simple and effective detection techniques.
- Quite efficient algorithms are available for the computation of $\{p_k\}$.

At the same time, however, it is necessary to stress that:

- The Fourier frequencies have no particular physical meaning. They constitute kinds of *natural frequencies* that, however, are intrinsic to the sampling characteristics and not to the signal under analysis. This implies that the frequency of interest could not belong to such a set.
- If x_j contains a sinusoidal component with frequency $f_u > f_{\mathrm{Ny}} = 0.5$ (in units of $1/\Delta t$), the periodogram will show a peak in correspondence to a frequency $f = \mathrm{mod}\,(f_u, 2\pi) < f_{\mathrm{Ny}}$.[2] This puts an upper limit f_{Ny}, the so called *Nyquist frequency*, on the maximal frequency that can be detected in a time series.

In conclusion, a regular sampling simplifies the analysis of the data as well as the development of efficient algorithms. However, especially in the context of exploratory data analysis, it suffers of some annoying limitations.

3. Periodogram analysis of irregularly sampled signals

3.1. Statistical issues

In astronomy, often the experimental conditions do not permit a regular sampling of signals and this leads to the following. First, it is no longer possible to define a set of *natural* frequencies (such as the Fourier frequencies) for which to compute the periodogram. Hence, there is no reason for the number N of frequencies to be equal to the number M of the sampling time instants $t_0, t_1, \ldots, t_{M-1}$. Therefore, we write the transformation corresponding to that given by Eq. (1) in the general form

$$\widehat{x}_f = \sum_{j=0}^{M-1} x_{t_j} \mathrm{e}^{-i2\pi f t_j}, \tag{11}$$

where, without loss of generality, we have $t_1 = 0$. The spectrogram is still defined as $p_f = \left|\widehat{x}_f\right|^2/M$. Similarly, Eqs. (8)–(9) become

[2] The function $z = \mathrm{mod}(x, y)$ provides the remainder z from the division of x by y.