A 'road test' of ANOVA versus DFT and LS as a period-finding algorithm

C A Engelbrecht[1] and F A M Frescura[2]
[1]Department of Physics, University of Johannesburg, Kingsway, Auckland Park, Johannesburg 2092, South Africa
[2]School of Physics, University of the Witwatersrand, Private Bag 3, WITS 2050, Johannesburg, South Africa

E-mail: chrise@uj.ac.za

Abstract. The mathematical properties of harmonic functions have brought Fourier-based algorithms into popular use as period-finding tools in astronomy, specifically in asteroseismology. Recently published work has indicated that the ANOVA approach could outperform these traditional approaches in certain cases. Very little practical application of ANOVA has appeared in the literature, though. Given the rapidly growing body of time-domain data in astronomy and the considerable importance of some of the conclusions that have been reached on the basis of these data, the recently published results prompt a closer look at ANOVA as a competitor to DFT and Lomb-Scargle (LS)-based methods in asteroseismology. In this paper, we present and analyse the main findings of a comparative 'road-test' of ANOVA, DFT and Lomb-Scargle algorithms, applied to a variety of observing scenarios, including typical ground-based, time-domain data of pulsating stars.

1. Introduction

Aided comprehensively by the data obtained from various space-based missions, the techniques of asteroseismology have profoundly affected our understanding of stellar structure and evolution. The accurate determination of pulsation frequencies, amplitudes and phases is essential for the successful application of asteroseismology to this end. The most commonly used techniques for period determination in variable stars are the Discrete Fourier Transform (DFT) adapted for non-equally-spaced time series [1], and the Lomb-Scargle (LS) periodogram [2]. Both of these techniques rely on certain mathematical properties of the harmonic functions for their efficiency as period-finding tools. There are various well-developed alternatives to Fourier-based methods. These have been summarised in a recent review [3]. The most prominent alternative period-finding strategy identifies the most probable periodicity occurring in a time series by minimising the dispersion in phase bins associated with the respective periods inspected (instead of calculating transforms of time series or fitting functions to them, as the Fourier-based methods do). These are called 'dispersion-based methods' or, occasionally, 'entropy-based methods'. The original dispersion/entropy-based method is the 'String-length' method introduced by Lafler & Kinman [4] for the purpose of dealing with light curves which deviate substantially from sinusoidal shapes. Schwarzenberg-Czerny [5] pioneered a thorough series of investigations into the statistical techniques embodied in the analysis of variance and their potential for the accurate detection of periodic variations in time series. We shall refer to Schwarzenberg-Czerny's formulation of these techniques as 'ANOVA' in what follows.

A recent study by Graham et al. [6] found that the ANOVA method, as described in [5], was exceptionally successful at retrieving periods in time series generated by large astronomical surveys, specifically: the MACHO, CRTS and ASAS surveys. Since the authors of the present paper probe the detection of low-amplitude pulsations in main-sequence stars, an exercise where it is essential to optimise the period-finding techniques employed, the findings reported in [6] have prompted us to investigate the efficacy of ANOVA in our own time series. We decided to measure this efficacy by exposing various sets of artificial data – with properties that are similar to the various datasets that we have at our disposal for our studies of pulsating stars – to three period-finding methods: the ANOVA method, the standard DFT algorithm described in [1] and the LS described in [2]. The following sections describe our simulated time series and the results of applying the three period-finding methods to these time series.

2. Simulated light curves

Three different observing scenarios were simulated. The following parameters are common to the three simulated time series: the total time-span of the observations is 110 days and a total of 1000 observations are taken during this time-span. These choices are typical of actual observing scenarios that produce useful asteroseismological results for pulsating early B stars. A previous study [7] by the present authors compared the efficacy of the Generalised Lomb-Scargle (GLS) method for period-finding, with DFT and LS methods. For ease of comparison, we have retained (except for computer-generated random errors) the simulated time series used in [7].

The following three realistic observing scenarios were chosen:

i) Scenario 1: A strictly equally-spaced time sampling over the 110 days, typical of what is achieved with a space-based telescope making uninterrupted observations of the same object. The observing protocol of the original mission design of the *Kepler* space telescope is similar to this scenario (although *Kepler* – in its 'Long-Cadence' observing mode - would obtain approximately 5200 equally-spaced observations over a 110-day period, instead of the chosen value of 1000 in this simulation). With the chosen numbers, the equal time-spacing between successive observations is 0.11 days, or 2.64 hours. This implies a Nyquist frequency of just over 4.5 cycles per day (c/d). All of the calculated periodograms in this study were therefore terminated at a frequency value of 4.5 c/d. This equally-spaced observing scenario is perfectly suited to the DFT method and we expected the DFT periodograms to perform at least as well as the LS and ANOVA periodograms in this scenario (if not better).

ii) Scenario 2: A typical ground-based observing campaign at a single site, consisting of five full weeks spread over the 110-day time-span and consisting of varying numbers of photometric nights per week. This time series typifies the result of small-telescope observing allocations of the South African Astronomical Observatory at its Sutherland site in the Northern Cape. In this simulation, a total of 17 photometric nights were simulated among the 35 days of allocated observing time. This ~50% duty cycle is typical of actual conditions at Sutherland. The actual sampling times were adapted to accommodate the advance of sidereal time over the 110 days of calendar time, which is accompanied by an earlier rising time of target stars each night. Within a night, the observations were taken at equal time intervals of 7.49 minutes. This number was determined by requiring 1000 observations to be taken in 17 photometric nights. Again, this time-spacing is typical of ground-based, observer-driven observations of pulsating early B stars.

iii) Scenario 3: A ground-based, pre-programmed survey project with no observer input, observing the target star a few times each day (and moving to other survey locations in

between). A random component was introduced in selecting the actual times of observation on any particular day. On most nights, the total number of observations taken per night was 8, 9 or 10. The spacing between successive observations was *not equal* and varied by more than an order of magnitude.

The simulated light curves shown in figure 1 illustrate the nature of the sampling regimes for the three scenarios. The method used to construct *signal strengths* is described after the figure.

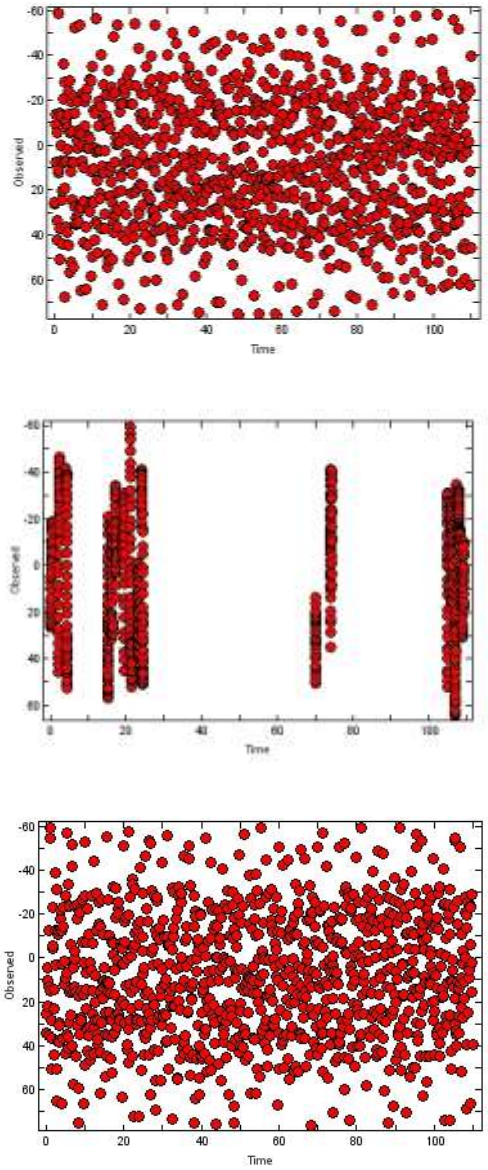

Figure 1. A sample of light curves generated for Scenarios 1, 2 and 3 (from top to bottom) respectively. Since the signal has a random noise component, the light curves will appear slightly differently every time they are calculated.

Scenarios 1, 2 and 3 determined the *time* vectors in the simulated light curves. The *signal* vectors were calculated as follows: As a first step, we added three sinusoids with respective amplitudes, frequencies and phase differences typical of low-amplitude p-modes in pulsating B stars. The value of

this cumulative 'clean' signal was evaluated at the calculated time values obtained in each of the respective scenarios. Secondly, an underlying mean signal level as well as a random error (different for each individual signal value (i.e. for each individual 'observation')) were added to these 'clean' signal values. The magnitudes of the random errors were chosen to correspond with the typical mean errors and error variances found in small-telescope photometry at Sutherland. The calculated light curves were then fed into the mathematical algorithms for calculating DFT, LS and ANOVA periodograms and the results were compared. These results are discussed in the following section.

3. Retrieving periods

The ability of the DFT, LS and ANOVA periodograms to retrieve the input frequencies and amplitudes of the periodic variations present in the simulated light curves was explored for various signal-to-noise (SNR) ratios and differing degrees of homogeneity of the signal errors. A brief summary of our results is presented in this section.

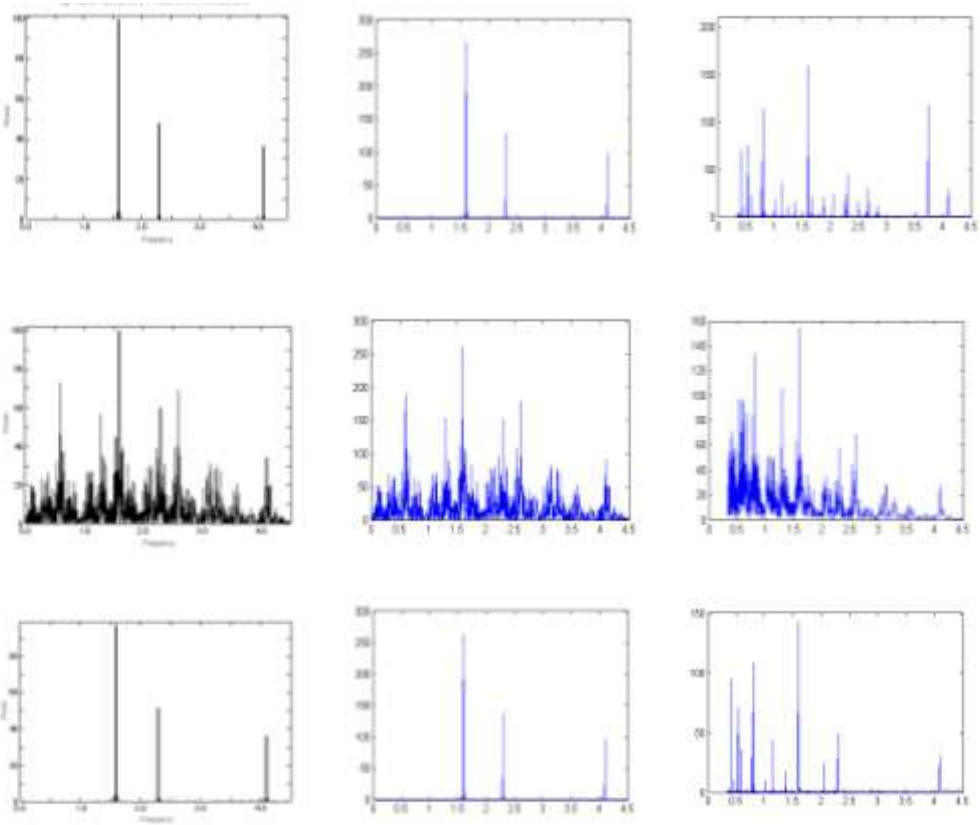

Figure 2. Periodograms obtained with DFT, LS and ANOVA algorithms (from left to right in each row), for result A. First row: scenario 1 (equal time spacing); second row: scenario 2 (ground-based observer); third row: scenario 3 (ground-based survey).

Result A: Firstly, we considered SNR values which are typical of ground-based asteroseismology of pulsating B stars, with a homogeneous error distribution across the entire observing epoch of 110 days. Initial periodograms (i.e. before any prewhitening) for Result A are shown in figure 2. For equal time spacing (scenario 1: the first row of periodograms in the figure), the DFT and LS periodograms deliver quite clear-cut results: they retrieve the three input frequencies (and their associated amplitudes) with ease and without any ambiguity. In comparison, the periodogram produced with the ANOVA algorithm appears beset with a multitude of spurious peaks. However, closer inspection

reveals that these are the sub-harmonics of the three primary frequencies from which the signal is constructed. Measures can be taken to account for the sub-harmonics when searching for periods with the ANOVA algorithm. Still, the DFT and LS methods appear simpler and less prone to mis-interpretation in the equally-spaced scenario.

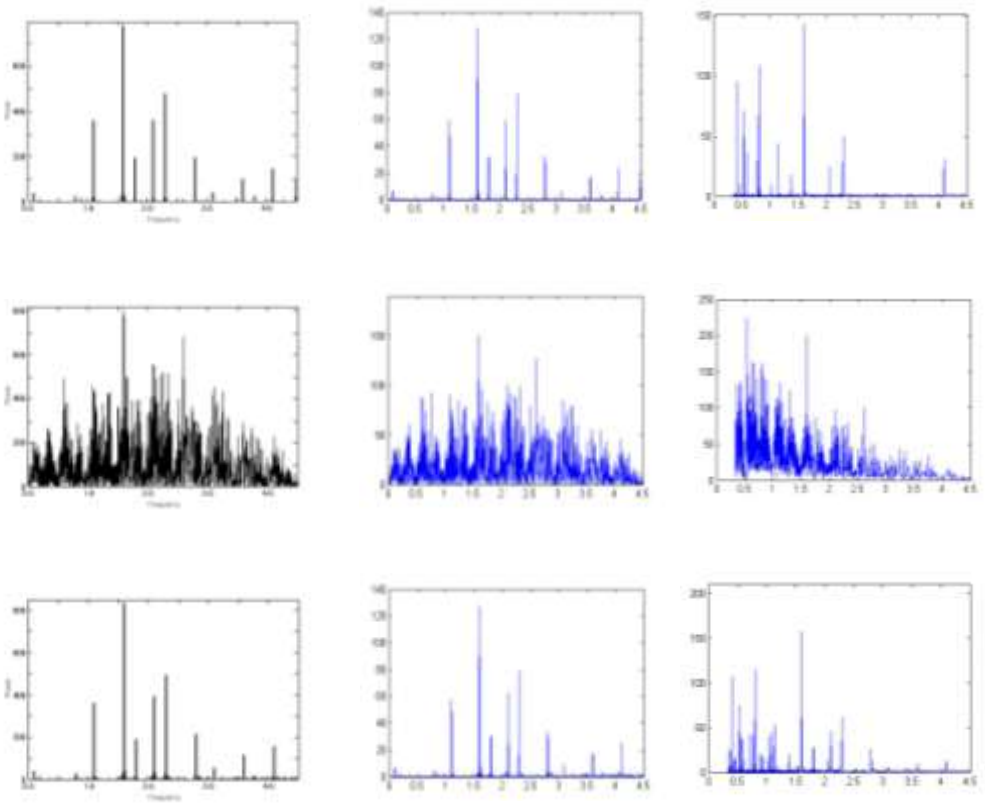

Figure 3. Periodograms obtained with DFT, LS and ANOVA algorithms respectively, for Result B. The arrows in the third column of graphs indicate the positions of the periods corresponding with the input signals, as for figure 2. The three rows correspond to scenarios 1, 2 and 3 respectively, as for figure 2.

The simulation of single-site observer-controlled observing produces the much more noisy periodograms in row 2 of figure 2. The one-day aliasing peaks are clearly seen in the DFT and LS periodograms, and a prewhitening procedure (not shown, but performed by the authors) succeeds in retrieving all three input periods without any ambiguity, removing practically all the noise in the periodogram as well. It is far more problematic extracting sub-harmonics from the ANOVA plots in this scenario. It does not appear useful to employ ANOVA in this context. Turning to row 3 of figure 2, which shows the periodograms for the scenario 3 (ground-based survey), it is surprising that the ANOVA periodogram fares better for this non-equally-spaced scenario, compared to scenario 1. This seems to be the context for which ANOVA is best suited as an alternative to DFT and LS methods.

Result B: The vulnerability of the respective period-finding methods to heavy aliasing effects was tested by removing every second day's data from the equally-spaced time series (scenario 1) and similarly thinning out the time series for the other two scenarios by 50% in a quasi-regular way. Initial periodograms (i.e. before prewhitening) for Result B are shown in figure 3. When compared with figure 2, it is clear that all nine periodograms are detrimentally affected by the 'aliasing noise' introduced by the deliberate, 'gap-generating', thinning procedure used. The primary input signal is still unambiguously identified in eight of the nine periodograms. However, for scenario 2 (middle

row), the ANOVA periodogram fails to identify even the strongest input signal as the primary peak. ANOVA appears at least as vulnerable to aliasing effects as the Fourier-based methods are, if not more so.

Result C: Vulnerability to variable background noise levels was tested by dividing the time series into three blocks - with the mean noise levels chosen to lie in the ratio 4:2:1 among the blocks. The highest noise level was chosen to be larger than two of the three input amplitudes of the harmonic signal in the simulated time series. Initial periodograms (before prewhitening) for Result C are shown in figure 4 (for scenario 2 only). Comparison with figure 2 indicates that none of the three period-finding algorithms shows any significant vulnerability to *variable* background noise levels across an observing season. However, note that the *signal-to-noise ratios* of the dominant peaks in the plots in figure 4 are indeed lower than in their counterparts in figure 2.

4. Discussion and conclusions

The three experiments described as Result A, Result B and Result C in the previous section lend a measure of support to the period-finding performance attributed to ANOVA in [6]. It would appear prudent to apply this period-finding algorithm in parallel to the DFT and LS algorithms, particularly in situations where aliasing effects might be compromising the latter methods. Also, ANOVA might be of particular value when analysing sparsely time series (corresponding to scenario 3 in this paper). However, the ANOVA method does not appear to be a viable alternative for finding periods in time series that are typical of the single-site, observer-controlled example (scenario 2) discussed in this work.

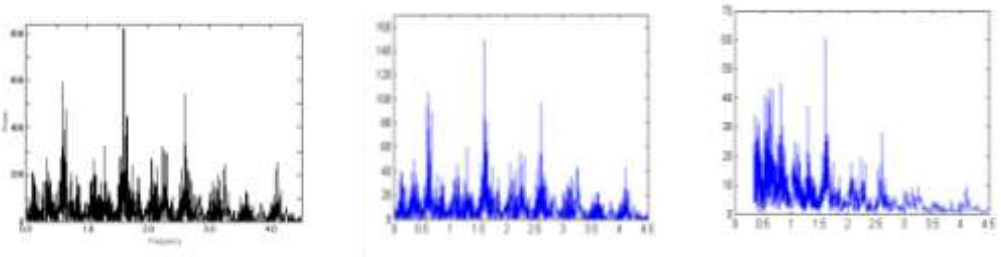

Figure 4. Periodograms obtained with DFT, LS and ANOVA algorithms respectively, for Result C. Only scenario 2 is shown, as the other scenarios add nothing new.

References

[1] Deeming T J 1975 *Ap&SS* **36** 137
[2] Scargle J D 1982 *ApJ* **263** 835
[3] Engelbrecht C A 2014 *Precision Asteroseismology: Proc. IAU Symposium 301 (Wroclaw, Poland, 19-23 August 2013)* eds J A Guzik, W J Chaplin et al (Cambridge: CUP) pp 77-84
[4] Lafler J and Kinman T D 1965 *ApJS* **11** 216
[5] Schwarzenberg-Czerny A 1989 *MNRAS* **241** 153
[6] Graham M J, Drake A J, Djorgovski S G, Mahabal A A, Donalek C, Duan V, Maker A 2013 *MNRAS* **434**, 3423
[7] Engelbrecht C A and Frescura F A M 2015 *Proc. 59th Annual Conference of the SA Institute of Physics (Johannesburg, South Africa, 7-11 July 2014)* eds C A Engelbrecht and S Karataglidis, pp 324-329