Precision Asteroseismology
Proceedings IAU Symposium No. 301, 2013
J. A. Guzik, W. J. Chaplin, G. Handler & A. Pigulski, eds.

doi:10.1017/S1743921313014129

Extracting oscillation frequencies from data: various approaches

C. A. Engelbrecht

Department of Physics, University of Johannesburg, PO Box 524, Auckland Park 2006, South Africa
email: chrise@uj.ac.za

Abstract. Asteroseismology depends absolutely on the detection of authentic pulsation signatures in stars. A variety of mathematical and statistical tools have been developed to extract such signatures from photometric and spectroscopic time series. The earliest tools were developed on the platform of Fourier analysis, and Fourier-based methodology still plays a major part in the detection of pulsation signatures in the present day. Alternative approaches have been gaining ground in recent years. This article offers a brief but broad review of the various methodologies for detecting authentic periodic signals that have been developed over the past few decades, including examples of their pitfalls and successes.

Keywords. methods: data analysis, methods: statistical, stars: oscillations

1. Introduction

Asteroseismology has vastly expanded the scope, range and depth of our understanding of stellar structure and evolution in recent years. As but one prominent example of this, Bedding *et al.* (2011) found a very powerful correlation between the spacing of g-modes and the dominant nuclear energy source (i.e. helium-core or hydrogen-shell fusion) in red giants. Such breakthroughs depend critically on the accuracy of the oscillation parameters (frequencies, amplitudes and phases) that are determined from the analysis of observational data. Fourier-based methods of analysis have been used successfully for over a hundred years, starting with the work of Schuster (1897). However, these methods have their pitfalls and other mathematical and statistical points of departure have received increasing attention as asteroseismology has matured to its present strength. This article briefly reviews practically the full range of asteroseismologically relevant methods that have been presented in the general literature. Their basic premises, strengths and weaknesses are briefly summarized and examples of their successful applications are presented.

The scope of published methods may be naturally classified under the two headings of Fourier-based methods and Dispersion/Entropy-based methods, respectively. These two classes will be discussed separately, as sections 2 and 3. The impact of Bayesian approaches on existing methods of period-finding is discussed in section 4. Section 5 contains a summary of ancillary issues of importance and some suggestions for further exploration of this topic.

At the outset, it might be useful to emphasise that the *spectrum* of a time series represents the respective *probability amplitudes* (quantified via one or other measure statistic) of regular variations present in the time series, as a function of variation frequency. This article considers the various methods available for computing such a spectrum.

2. Fourier-based methods

The popularity of Fourier-based methods is due to the mathematical properties of harmonic functions, which make them attractive as the basis of algorithms designed to search for periodicity in the frequency domain. The original Fourier-based method is the Fourier transform, which applies to a continuous function in the time domain. Measurements taken of real signals are, of necessity, not continuous but discrete. Real measurements are also not infinitely long in total extent. The adaptation of the Fourier transform to discrete data (i.e. a discrete sampling of a supposed underlying continuous function) constitutes the Discrete Fourier Transform (DFT), originally applied in cases where the discrete sampling is exactly equally-spaced in the time domain. Both the original DFT and the efficiently-computable Fast Fourier Transform (FFT) can be attributed to Gauss (Heideman *et al.* 1984). Astronomical measurements are almost never exactly equally-spaced. Therefore, a modification to the equally-spaced DFT for application to non-equally-spaced data (which also have a finite total extent), was proposed by Deeming (1975). As pointed out by Ferraz-Mello (1981), amongst others, there is no mathematical foundation supporting the Deeming formula. Although it appears to work well in many cases, it remains, in its essence, a heuristic device. Given the importance of accuracy and precision of period detections in asteroseismology, attention should be given to alternatives to the raw Deeming formula that provide answers with a higher level of confidence. Ferraz-Mello (1981) introduced the so-called Date-Compensated DFT which introduces a non-zero constant mean into the calculated transform. The major advantage of this refinement of the Deeming formula for a non-equally-spaced DFT (called the Deeming DFT in the remainder of this article) is a more accurate estimation of the *amplitudes* of harmonic signals present in the observed data, especially when the number of observations is relatively small. Ferraz-Mello also introduced a weighting algorithm to deal with observations that do not all have exactly the same measurement error. The Deeming DFT and its refinements have been used very extensively in the practice of asteroseismology over many decades, through to the present. It works very well when applied to quasi-equally-spaced data (like the data in the *Kepler* database). However, as stated already, it remains a heuristic device and it should always be used with circumspection.

Assigning a *statistical significance* to suspected oscillation frequencies harvested from a Deeming DFT is problematic, as the statistical conclusions that attach to purely equally-spaced time series do not apply to non-equally-spaced data. This makes it impossible to attach statistical significance *a priori* to peaks in Deeming DFT periodograms. Scargle (1982) attempted to address this particular problem by deriving an alternative algorithm to the Deeming DFT, as discussed below. Breger *et al.* (1993) offered a helpful rule of thumb to estimate the statistical significance of periodogram peaks obtained with a Deeming DFT: if the harmonic wave associated with a frequency peak has an amplitude greater than 4 times the residual noise level in the periodogram (see section 5.4 of Aerts *et al.* (2010) for a detailed discussion), it is statistically significant. It is unfortunate that this commonly used rule of thumb is being called a "four-sigma detection" with growing regularity, since this might create considerable confusion in the future. The terms "four-sigma detection", "five-sigma detection", etc. have a very well-established meaning in the theory of normal distributions and such phrases are commonly used in many areas of modern science. The Breger rule of thumb means something totally different and it should rather be referred to as the "4:1 rule" or something similar. Useful as the rule has proven to be, the brief experiments conducted by Koen (2010) show that the quantitative statistical meaning of the "4:1 rule" can vary wildly from one dataset to the next. The present author has independently found similar results (unpublished). The discussion in

Aerts *et al.* (2010) reminds the reader of the very particular and limited context wherein Breger found the "4:1 rule" to be valid. Considerable caution is required when applying it in general.

Scargle's seminal paper (Scargle 1982) followed on foundations laid by Barning (1963), Vaníček (1969) and Lomb (1976). Scargle claimed that the algorithm that he put forward (commonly called the Lomb-Scargle periodogram or LS periodogram for short) retrieves the statistical properties of the classical (equally-spaced) DFT; therefore, the probability of obtaining a peak of height X in the LS periodogram of a time series consisting of normally-distributed noise is proportional to $\exp(-X)$. However, this claim encounters an insurmountable practical difficulty, as has been pointed out repeatedly (see, for example, Horne & Baliunas (1986), Koen (1990), Frescura *et al.* (2008) and Scargle's own recognition of the problem in his paper): when computing the periodogram for more than one frequency, using non-equally-spaced data, one requires the exact number of statistically independent frequencies in the set of frequency values at which the periodogram is being calculated. There is no way to determine this number (see Horne & Baliunas 1986 and the follow-up work by Frescura *et al.* 2008). Various workers have concluded that Monte Carlo procedures are the only reliable way of determining the statistical significance of peaks in a LS periodogram (for instance, see the papers just quoted). An interesting and thorough discussion of this question appears in Vio *et al.* (2010).

Mathematically, the Lomb-Scargle periodogram calculates the goodness-of-fit of a harmonic (i.e. sinusoidal) function compared with the data, for a selected grid of frequencies. It therefore works particularly well at extracting the frequencies of small-amplitude oscillations in general (since small excursions from an equilibrium state are always well approximated by a harmonic variation). It has been used with considerable success for more than thirty years and remains one of the most widely-used methods of extracting oscillation frequencies from astronomical data. It should be noted that the LS periodogram and the Deeming DFT do not always produce similar results; a good example of discrepant results is discussed in the very thorough study of the LS periodogram by Vio *et al.* (2013). The LS periodogram is generally less susceptible to aliasing than the Deeming DFT is – see e.g. Reegen (2007). Furthermore, Frescura *et al.* (2008) demonstrated that it is vitally important to "oversample" the frequency grid when calculating the LS periodogram.

A substantial overhaul of the classical LS periodogram is contained in the work of Cumming *et al.* (1999) and Zechmeister & Kürster (2009). Cumming *et al.* introduced a floating mean into the LS periodogram calculation, meaning that a constant term is included in the fit for each trial frequency. Their detailed study led them to conclude that "allowing the mean to float is crucial if the number of observations is small, the sampling is uneven, or there is a period comparable to the duration of the observations or longer". Uneven sampling is almost ubiquitous in the data used in asteroseismology, hence their conclusions deserve careful scrutiny. These authors caution that periodic signals could be totally missed, or their amplitudes miscalculated, when the floating mean is omitted. Zechmeister & Kürster added a weighting procedure, to accommodate observations that do not all have exactly the same precision. They chose the name "Generalised Lomb-Scargle periodogram" for their more refined procedure, abbreviated as GLS (although they point out one previous use of the same phrase in a different context). They concluded that, compared to the classical LS periodogram, GLS provides a more accurate frequency determination, is less susceptible to aliasing, and gives a much better determination of the spectral intensity. A recent, comprehensive comparison of various period-finding methods by Graham *et al.* (2013a) provides strong endorsement for the accuracy of the GLS algorithm.

A detailed study of an alternative approach for calculating statistical significances of detected periodicities in unequally-spaced data was presented as the SigSpec algorithm by Reegen (2007), who claimed period-finding accuracy on a par with LS but with less susceptibility to aliasing. The most prominent new feature of SigSpec is its inclusion of *phase* information, while it also incorporates the floating mean and weighting refinements of GLS. SigSpec has been widely used in the literature, often in conjunction with other period-finding methods. Zechmeister & Kürster (2009) demonstrated many equivalences between the GLS and SigSpec approaches to period-finding.

A quite different methodology for period-finding, SparSpec, was presented by Bourguignon *et al.* (2007). The authors explain that the name "SparSpec" derives from the method's character as a "multisine fitting [...] addressed as the *sparse* representation of data in an overcomplete dictionary of frequencies". They model the given data using the sum of an arbitrarily large number of pure frequencies, discretised on a fixed grid, and seek that particular representation that produces the fewest non-zero amplitudes. Following various detailed tests, the authors conclude that SparSpec: i) correctly locates the frequencies embedded in some test data while classical methods fail to do so; ii) outperforms the familiar sequential prewhitening methods in countering sampling artifacts; iii) accurately estimates both frequencies and amplitudes; and iv) is less sensitive to low-frequency perturbations, e.g. to those caused by orbital movements, than familiar methods. Bourguignon *et al.* (2007) actively encourage the community to apply their code and test its efficacy. Bourguignon *et al.*'s (2007) SparSpec paper includes a thorough application of the CLEAN and CLEANest methods. CLEAN refers to the algorithm presented by Roberts *et al.* (1987), with the picturesquely phrased aim to "undo the damage inflicted" by the incomplete sampling of the physical signal. Essentially, the aim of the CLEAN approach is to subtract the spectral window from the so-called "dirty spectrum". Foster (1995) followed this up with CLEANest, specifically tailored to deal with very long time series. Foster (1996) contains a thorough study of wavelet theory applied to period-finding in time series, with the specific aim to treat *variable* periods and/or amplitudes.

In summary, the Fourier-based methods readily available for period-finding at present include the Deeming DFT (often employed in the form of the Period04 package), the classical LS periodogram (available in various packages but also very simple to self-code), GLS, SigSpec and SparSpec. A positive feature of the various modern alternatives to the classical methods is their common presentation to potential users in a very accessible format. The SigSpec and SparSpec codes are readily available on the web, together with substantial user manuals. Refinements to classical methods, like GLS, are simple enough to self-code, with the algorithms readily provided in the source papers.

3. Dispersion/Entropy-based methods

A useful overview of the mathematical basis for these methods appears in Aerts *et al.* (2010). In brief terms, these methods identify the most probable period (or frequency) of a regular variation present in a time series by minimising the scatter in phase bins associated with each respective period in a chosen grid (instead of calculating transforms of time series or fitting functions to them, as the Fourier-based methods do). The original dispersion/entropy-based method ("D/E method" in what follows) is the "String-length" (STR) method introduced by Lafler & Kinman (1965), expressly to deal with light curves that deviate substantially from sinusoidal shapes. The meaning of their test statistic Θ is quite transparent: For a time series consisting of measurements m_j,

https://doi.org/10.1017/S1743921313014129 Published online by Cambridge University Press

with mean $\bar{M}$,

$$\Theta = \frac{\sum_i (m_i - m_{i+1})^2}{\sum_i (m_i - \bar{M})^2} \tag{3.1}$$

This simple algorithm for a dispersion-based statistic was followed by a succession of increasingly sophisticated approaches, continuing to the present day (see below). However, the simple STR approach has also received continuing attention. Dworetsky (1983) pointed out the advantages offered by STR when dealing with sparsely sampled signals, while Clarke (2002) refined the original STR statistic, to render it independent of sample size. In summary: STR handles non-sinusoidal signals rather well, but the statistic produces a plot that is rich in false peaks ("aliases" of a kind) and the method is computing-intensive.

Not long after Lafler & Kinman's introduction of STR, Jurkevich (1971) proposed what he called "a statistically more natural" approach, containing a set of statistics with a stronger foundation in formal statistical theory. This procedure was generalised by Stellingwerf (1978) in the format that has been very widely used as "Phase Dispersion Minimisation" or PDM. Similarly to STR, PDM produces very noisy plots of the PDM statistic against frequency (this type of plot is the D/E methods' equivalent of the Fourier-based methods' periodogram).

An important development in D/E methods is the "Analysis of Variance" (or AoV) algorithm proposed by Schwarzenberg-Czerny (1989; SC89). This algorithm is rooted very deeply in fundamental statistical theory and Schwarzenberg-Czerny obtains very impressive results when applying AoV to a variety of test cases. SC89 also makes pertinent statements about STR and PDM. The AoV method appears to be remarkably more powerful at extracting periodicities from time series (at a given significance level) than many other methods. SC89 also points out that Fourier-based methods are superior to D/E methods for the detection of sinusoidal signals; however, the converse is true when dealing with substantially non-sinusoidal signals (e.g. sharp pulses). Schwarzenberg-Czerny (1996; SC96) extends AoV to the so-called "multi-harmonic" case. He exploits a correspondence between Fourier series and series of complex polynomials to set up an algorithm which appears exceptionally powerful at damping aliases (and therefore at detecting weaker physical signals in the data). The comparative study by Graham *et al.* (2013a), mentioned earlier, concluded that the SC96 methodology (labelled AoVMHW) has great merit as an accurate and sensitive period-finding procedure. Baluev (2009) conducted a meticulous statistical study of AoVMHW, focusing on the treatment of non-sinusoidal light curves, and confirmed the merits of AoVMHW as a versatile and powerful period-finding method. Baluev (2008) constitutes a similar treatment of the LS periodogram. The work in both of these papers is carried further in Baluev (2013a), where the Von Mises function is exploited to obtain thought-provoking results. Baluev (2013a) also supplies a weblink to the C++ code for his algorithm with an invitation for its widespread use.

A couple of period-finding algorithms based on various definitions of *entropy* have been introduced recently. Cincotta *et al.* (1995) use the Shannon entropy to select an optimal period. The analytical theory behind SE is presented in Cincotta *et al.* (1999). They test their algorithm with simulated data containing non-sinusoidal variability and find that their method (SE, for Shannon entropy) is more sensitive than the classical periodograms at period detection and that it is very good at resolving closely spaced frequencies. As with the other D/E methods, the actual shape of a periodic variation in the data is not important and the method is well suited to the detection of non-sinusoidal variations. One substantial advantage of the method is its mathematical simplicity and

https://doi.org/10.1017/S1743921313014129 Published online by Cambridge University Press

rapid computability. However, it is quite vulnerable to gap aliasing. On the contrary, the "Conditional Entropy" (CE) method introduced by Graham *et al.* (2013b) is found to be "particularly robust against common aliasing issues". The authors back these claims up with persuasive test results. Essentially, CE is a simple but effective modification of SE. The authors put CE through a robust test run in their paper comparing various period-finding methods (Graham *et al.* 2013a), where it emerges as a very powerful tool for the accurate detection of periods in large survey databases. Huijse *et al.* (2012) present a useful summary of some other D/E methods while introducing their CKP (Correntropy Kernelised Periodogram) algorithm.

In summary, the following D/E-methods have been well-tested in the past or recently added to the range of options: STR, PDM, AoV, AoVMHW, SE, CE and CKP.

4. The impact of Bayesian approaches

Asteroseismology relies on the meticulous matching of theoretical calculations of pulsation parameters with the values of those parameters derived from observational data. Besides the actual *values* of frequencies, amplitudes, phases and modal indices we derive from data, a key quantity is how *reliable* the derived values are. Many procedures exist for estimating the statistical significance of data-derived quantities. Bayesian statistics offers a methodology for estimating data-derived quantities as well as modeled quantities. An informative example of estimating modeled quantities is found in Bazot *et al.* (2012), while Brewer & Stello (2009) present a good example of assigning probabilities to data-derived quantities. Marsh *et al.* (2008) present an application to solar oscillations. Bourguignon & Cartanfan (2008) demonstrate efficient period extraction with Bayesian methods when traditional methods fail, while White *et al.* (2010) offer a balanced comparison of Bayesian and Fourier methods and point out the former's advantage in avoiding spurious aliases. Wang *et al.* (2012) find distinct advantages in applying Bayesian methods to non-sinusoidal light curves. Stoica *et al.* (2009) (also see He *et al.* 2009) combine their RIAA method with a Bayesian analysis.

5. Ancillary issues and suggestions

The thorough study by Cumming *et al.* (1999) makes important conclusions regarding the appropriate *normalization* of the LS periodogram. Koen (2006, 2010b) explores the interpretation of the *Nyquist frequency* for irregularly spaced time series. Eyer & Bartholdi (1999) and Pelt (2009) also consider the topic. Süveges (2012) presents an interesting treatment of the *statistical significance* of periodogram peaks. Further treatments of various aspects of period-finding may be found in Jetsu & Pelt (1999), Palmer (2009), Pelt *et al.* (2011) and Leroy (2012).

The content of the many published papers on period-finding in astronomical time series is very diverse. Meticulous statistical studies of this issue are rare; however, a few authors have each produced a substantial set of papers that explore important questions regarding the statistical content and meaning of various aspects of observed time series. The reader is referred to the following papers of these authors (in alphabetic order) as an overview: Baluev (2008, 2009, 2012, 2013a, 2013b), Koen (1990, 1999, 2000, 2006, 2009, 2010b), Koen & Lombard (1993), Schwarzenberg-Czerny (1989, 1991, 1996, 1997, 1998, 1999). The work of Stahn & Gizon (2008) on the analysis of time series containing many gaps is also worth consulting.

The aim of this article has been to be as inclusive as possible in pointing the reader to the full scope of methods for extracting oscillation frequencies from time series. There

are many important features that attend to the respective methods and many subtle points that thread through them. A detailed study of the papers cited here (and their references) is recommended for further elucidation.

The organisers are thanked for the invitation to compile this review. Financial support from the University of Johannesburg/DHET and the IAU is gratefully acknowledged.

References

Aerts, C., Christensen-Dalsgaard, J., & Kurtz, D. W. 2010, *Asteroseismology* (Springer)
Baluev, R. V. 2008, *MNRAS*, 385, 1279
Baluev, R. V. 2009, *MNRAS*, 395, 1541
Baluev, R. V. 2012, *MNRAS*, 422, 2372
Baluev, R. V. 2013a, *MNRAS*, 431, 1167
Baluev, R. V. 2013b, *MNRAS*, 436, 807
Barning, F. J. M. 1963, *BAN*, 17, 22
Bazot, M., Bourguignon, S., & Christensen-Dalsgaard, J. 2012, *MNRAS*, 427, 1847
Bedding, T. R., Mosser, B., Huber, D., *et al.* 2011, *Nature*, 471, 608
Bourguignon, S., Carfantan, H., & Böhm, T. 2007, *A&A*, 462, 379
Bourguignon, S., Carfantan, H., & Böhm, T. 2008, *Statistical Methodology*, 5, 318
Breger, M., Stich, J., Garrido, R., *et al.* 1993, *A&A*, 271, 482
Brewer, D. J. & Stello, D. 2009, *MNRAS*, 395, 2226
Cincotta, P. M., Mendez, M., & Nunez, J. A. 1995, *ApJ*, 449, 231
Cincotta, P. M., Helmi, A., Mendez, M., Nunez, J. A., & Vucetich, H. 1999, *MNRAS*, 302, 582
Clarke, D. 2002, *A&A*, 386, 763
Cumming, A., Marcy, G. W., & Butler, R. P. 1999, *ApJ*, 526, 890
Deeming, T. J. 1975, *Ap&SS*, 36, 137
Dworetsky, M. M. 1983, *MNRAS*, 203, 917
Eyer, L. & Bartholdi, P. 1999, *A&AS*, 135, 1
Ferraz-Mello, S. 1981, *AJ*, 86, 619
Foster, G. 1995, *AJ*, 109, 1889
Foster, G. 1996, *AJ*, 112, 1709
Frescura, F. A. M., Engelbrecht, C. A., & Frank, B. S. 2008, *MNRAS*, 388, 1693
Graham, M. J., Drake, A. J., Djorgovski, S. G., *et al.* 2013a, *MNRAS*, 434, 3423
Graham, M. J., Drake, A. J., Djorgovski, S. G., Mahabal, A. A., & Donalek, C. 2013b, *MNRAS*, 434, 2629
He, H., Li, J., & Stoica, P. 2009, *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop*, p. 375
Heideman, M. T., Johnson, D. H., & Burrus, C. S. 1984, *IEEE ASSP Magazine*, 1, 14
Horne, J. H. & Baliunas, S. L. 1986, *ApJ*, 302, 757
Huijse, P., Estévez, P. A., Protopapas, P., Zegers, P., & Príncipe, J. C. 2012, *IEEE Transactions on Signal Processing*, 60, 5135
Jetsu, L. & Pelt, J. 1999, *A&AS*, 139, 629
Jurkevich, I. 1971, *Ap&SS*, 13, 154
Koen, C. 1990, *ApJ*, 348, 700
Koen, C. 1999, *MNRAS*, 309, 769
Koen, C. 2000, *MNRAS*, 316, 613
Koen, C. 2006, *MNRAS*, 371, 1390
Koen, C. 2009, *MNRAS*, 392, 190
Koen, C. 2010a, *Ap&SS*, 329, 267
Koen, C. 2010b, *MNRAS*, 401, 586
Koen, C. & Lombard, F. 1993, *MNRAS*, 263, 287
Lafler, J. & Kinman, T. D. 1965, *ApJS*, 11, 216
Leroy, B. 2012, *A&A*, 545, A50
Lomb, N. R. 1976, *Ap&SS*, 39, 447

Marsh, M. S., Ireland, J., & Kucera, T. 2008, *ApJ*, 681, 672
Palmer, D. M. 2009, *ApJ*, 695, 496
Pelt, J. 2009, *Baltic Astronomy*, 18, 83
Pelt, J., Olspert, N., Mantere, M. J., & Tuominen, I. 2011, *A&A*, 535, A23
Reegen, P. 2007, *A&A*, 467, 1353
Roberts, D. H., Lehar, J., & Dreher, J. W. 1987, *AJ*, 93, 968
Scargle, J. D. 1982, *ApJ*, 263, 835
Schuster, A. 1897, *Terrestrial Magnetism*, 3, 14
Schwarzenberg-Czerny, A. 1989, *MNRAS*, 241, 153
Schwarzenberg-Czerny, A. 1991, *MNRAS*, 253, 198
Schwarzenberg-Czerny, A. 1996, *ApJ*, 460, L107
Schwarzenberg-Czerny, A. 1997, *ApJ*, 489, 941
Schwarzenberg-Czerny, A. 1998, *MNRAS*, 301, 831
Schwarzenberg-Czerny, A. 1999, *ApJ*, 516, 315
Stahn, Th. & Gizon, L. 2008, *Solar Phys.*, 251, 31
Stoica, P., Li, J., & He, H. 2009, *IEEE Transactions on Signal Processing*, 57, 843
Stellingwerf, R. F. 1978, *ApJ*, 224, 953
Süveges, M. 2012, arXiv: 1212.0645
Vaníček, P. 1969, *Ap&SS*, 4, 387
Vio, R., Andreani, P., & Biggs, A. 2010, *A&A*, 519, A85
Vio, R., Diaz-Trigo, M., & Andreani, P. 2013, *Astronomy & Computing*, 1, 5
Wang, Y., Khardon, R., & Protopapas, P. 2012, *ApJ*, 756, 67
White, T. R., Brewer, B. J., Bedding, T. R., Stello, D., & Kjeldsen, H. 2010, *CoAst*, 161, 39
Zechmeister, M. & Kürster, M. 2009, *A&A*, 496, 577