Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **493,** 48–54 (2020)
Advance Access publication 2020 January 22

doi:10.1093/mnras/staa190

Significance levels of common frequencies extracted from multiple data sets

Chris Koen★

Department of Statistics, University of the Western Cape, Private Bag X17, Bellville 7535, Cape Town, South Africa

Accepted 2020 January 17. Received 2020 January 17; in original form 2019 October 15

ABSTRACT

Large monitoring campaigns, particularly those using multiple filters, have produced replicated time series of observations for literally millions of stars. The search for periodicities in such replicated data can be facilitated by comparing the periodograms of the various time series. In particular, frequency spectra can be searched for common peaks. The sensitivity of this procedure to various parameters (e.g. the time base of the data, length of the frequency interval searched, number of replicate series, etc.) is explored. Two additional statistics that could sharpen results are also discussed: the closeness (in frequency) of peaks identified as common to all data sets, and the sum of the ranks of the peaks. Analytical expressions for the distributions of these two statistics are presented. The method is illustrated by showing that a 'dubious' periodicity in an 'Asteroid Terrestrial-impact Last Alert System' data set is highly significant.

Key words: methods: data analysis – methods: statistical.

1 INTRODUCTION

The type of data considered in this paper is a low-amplitude sinusoidal signal embedded in white noise:

$$y(t_{kj}) = A_k \cos(2\pi\nu_0 t_{kj} + \phi_k) + e(t_{kj}); \quad j = 1, 2, \ldots, N_k; \quad k = 1, 2, \ldots, K, \tag{1}$$

where the A_k are amplitudes, ν_0 is the (common) frequency, ϕ_k are the phases of the signal, and e is uncorrelated noise. The K time series in equation (1) are each observed in N_k time points t_{kj}, which will in general be irregularly spaced. In the simulations presented below, the noise is conveniently assumed zero-mean Gaussian with standard deviation σ_k,

$$e(t_{kj}) \sim N(0, \sigma_k^2),$$

but the specific distribution is not too important. Note also that the noise standard deviations, amplitudes, and number of observations per time series may differ, although it will be assumed in the theoretical treatment below that

$$A_1 = A_2 = \cdots A_K \qquad \sigma_1 = \sigma_2 = \cdots \sigma_K \qquad N_1 = N_2 = \cdots N_K \ .$$

This assumption is made for the sake of expositional clarity, not out of necessity. In fact, the units of the different time series need not be the same; for example, intensity measurements from different parts of the electromagnetic spectrum (radio, optical, X-ray, etc.) and/or quantities derived from spectra (radial velocities, equivalent widths, etc.) can all be included.

★ E-mail: ckoen@uwc.ac.za

The analyses below will be performed in the frequency domain. A central role will therefore be played by spectral transformations of the data. The amplitude spectrum S of data $\{ y(t_j)\}$ is defined in terms of the periodogram

$$I(\nu) = \frac{1}{N}\left|\sum_{j=1}^{N}[y(t_j) - \overline{y}]\exp(-2\pi i \nu t_j)\right|^2 \tag{2}$$

as

$$H(\nu) = \frac{2}{N}\sqrt{I(\nu)}\ . \tag{3}$$

In equations (2) and (3), ν is the frequency, $\overline{y}$ is the mean of the time series, and $i = \sqrt{-1}$. Searching for sinusoidal signals embedded in a noisy time series usually involves plotting I or H against ν, to see whether there are power or amplitude excesses at any frequencies. The hypothesis test

H0: The time series is pure noise

H1: There is a signal in the time series (4)

is usually performed by comparing, in some way, the height of the largest spectral peak to the level of the rest of the spectrum – see e.g. Frescura, Engelbrecht & Frank (2008). A somewhat different scenario is considered in this paper, namely establishing significance levels for the presence of sinusoidal signals when more than one independent realization of the time series is available, as in equation (1).

Fig. 1 shows the amplitude spectra of three independent simulated data sets of the form equation (1). In each data set, $\nu_0 = 0.05\,\mathrm{d}^{-1}$, $A = 0.007$ mag, $\sigma = 0.02$ mag, and $N = 150$. In order to obtain

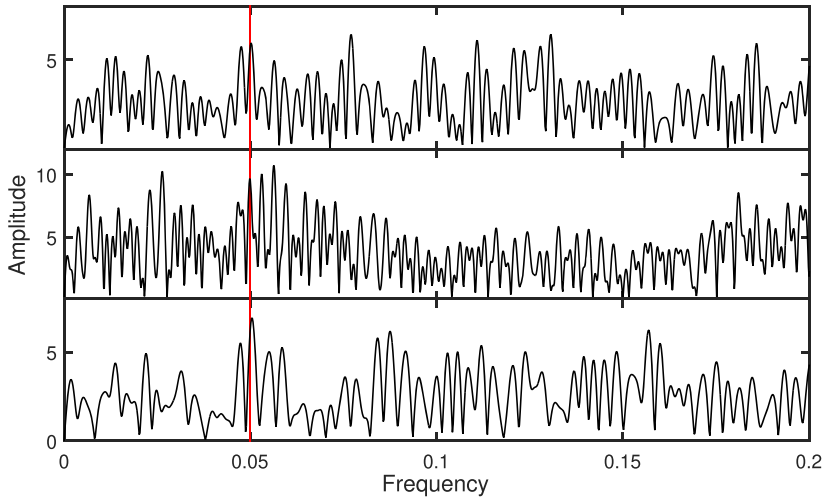

Figure 1. Amplitude spectra of three simulated data sets, each consisting of a sinusoid with amplitude 7 mmag with superposed white noise with $\sigma = 20$ mmag. The (red) vertical line shows the position of the sinusoid frequency $\nu_0 = 0.05\,\mathrm{d}^{-1}$.

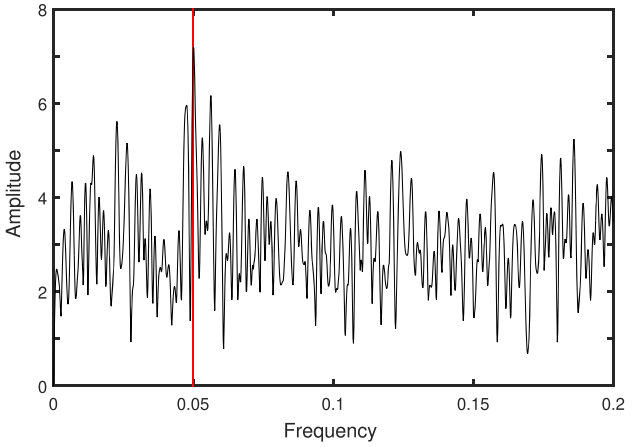

Figure 2. The average of the three spectra in Fig. 1. The (red) vertical line shows the position of the sinusoid frequency $\nu_0 = 0.05\,\mathrm{d}^{-1}$.

realistic irregular data spacings, observation times of a few stars were taken from the 'Asteroid Terrestrial-impact Last Alert System' (ATLAS) variable star catalogue (see Heinze et al. 2018). The time spans covered by the three data sets are 506, 658, and 501 d. The red vertical line in the Figure marks the position of the signal frequency ν_0. The highest peak in the bottom spectrum is indeed at ν_0, but in both the other two spectra, the peak at ν_0 is ranked fourth. The most likely conclusion drawn from Fig. 1 is that there is no evidence for a periodicity in any of the three spectra.

Fig. 2 demonstrates that coadding the spectra does not greatly improve the situation. Although the tallest peak is at the correct frequency, its height (7.2 mmag) is not markedly in excess of all other peaks, and only 2.4 times the mean noise level (3.06 mmag).

Nonetheless, as will be demonstrated below, it is possible to correctly identify ν_0 from the three spectra in Fig. 1, and with high significance. This is essentially done by comparing the positions of peaks extracted from each of the spectra. The first step in the analysis is therefore to identify the positions (frequencies) of all peaks, and then to search for peak positions which are closely similar across all spectra.

It is well known that the frequency resolution of the periodogram is $\sim 1/\Delta T$, where ΔT is the time base covered by the observations (e.g. Kovács 1981). Experimentation gave good results with peak widths taken to be $0.8/\Delta T$. If there are K data sets, peaks are considered coincident if there is some overlap of each peak with *all* others. In set-theoretic notation,

$$P_1 \cap P_2 \cap \cdots \cap P_K \neq \emptyset \tag{5}$$

is required, where for each peak $P_k = [P_{0k} - 0.4/\Delta T_k,\ P_{0k} + 0.4/\Delta T_k]$, P_{0k} being the frequency of maximum power in spectrum k.

The next section of the paper presents the results of simulations for the noise only case ($A = 0$ in equation 1). This paves the way for the significance level determinations demonstrated in Section 3.

2 NULL HYPOTHESIS SIMULATIONS

It is, of course, possible to obtain chance coincidences of spectral peak positions. In this section of the paper, the impact of various properties of the time series on such false alarm probabilities is studied. Aside from the number K of independent time series and the number of observations N, included in the study are the following: the mean time baseline $\overline{\Delta T}$, the frequency interval $(0,\nu_{\mathrm{E}})$ searched, the number of peaks L from each spectrum, which is taken into consideration (starting with the highest), and the spacing of the measurements. As far as the latter is concerned, the ATLAS spacings are supplemented by random exponentially distributed intervals between measurements.

Table 1 summarizes the results of an extensive study – each line is the outcome of at least 10 000 simulations. Briefly:

1 Comparing results for different K, the probability of finding peaks coincidences decreases rapidly with K. This is no surprise – finding four peaks aligned by chance is clearly more rare than finding two or three.

2 The larger ν_{E}, the smaller p_0. The reason is that if the top ranked L peaks are spread over a wider frequency interval $(0,\ \nu_E)$, the probability of a chance peak coincidence is reduced.

3 Not surprisingly, the probability of finding peak coincidences increases with L, the number of peaks from each data set, which is taken into account.

Table 1. Null hypothesis simulation results. Meanings of the symbols are: K – number of spectra compared; L – the peak rank to which each spectrum is searched; N – number of observations in the simulated time series; ν_E – the upper limit of the frequency range over which the spectrum is calculated; $\overline{\Delta T}$ – mean (over the K time series) time span of the observations; A, E – observation spacing, either resembling ATLAS or with exponentially distributed times between observations; p_0 – the probability of finding at least one spectral peak alignment.

K	L	N	ν_E	$\overline{\Delta T}$	Spacing	p_0
2	5	150	0.1	582	A	0.52
2	5	150	0.1	575	E	0.55
2	5	150	0.2	582	A	0.29
2	5	150	0.1	1202	A	0.28
2	5	150	0.1	2930	A	0.13
2	5	150	0.5	582	A	0.13
2	5	250	0.1	624	A	0.51
2	10	150	0.1	582	A	0.96
2	10	150	0.1	572	E	0.97
3	5	150	0.1	555	A	0.071
3	5	150	0.1	561	E	0.075
3	5	150	0.2	555	A	0.020
3	5	150	0.1	1135	A	0.017
3	5	150	0.5	555	A	0.0030
3	5	250	0.1	605	A	0.067
3	10	150	0.1	555	A	0.46
3	10	150	0.1	568	E	0.47
4	5	150	0.1	546	A	0.0081
4	5	150	0.1	586	E	0.0070
4	5	150	0.2	561	A	0.0011
4	5	150	0.1	1117	A	7.9E-4
4	5	150	0.5	561	A	1.0E-4
4	5	250	0.1	587	A	0.0063
4	10	150	0.1	546	A	0.12
4	10	150	0.1	553	E	0.12

4 The probability p_0 of a spurious peak alignment decreases with increasing $\overline{\Delta T}$. This follows because the spectral resolution improves, spectral peaks are narrower, and the probability of a chance coincidence is therefore reduced.

5 The number of observations N in each data set has little influence on p_0. The same is true of the two types of data spacing considered.

3 CALCULATING p-VALUES

The probabilities in the last column of Table 1 provide a first step towards obtaining significance levels for peak coincidences.

Two further independent statistics are useful for improving the estimation of peak-coincidence p-values. The first is

$$S = \sum_{k=1}^{K} r_k,$$

where r_k is the rank of the peak amongst peaks in spectrum k. For example, there is a single peak coincidence in Fig. 1, near $\nu = 0.05$; the ranks of the three peaks are $r_1 = 4$, $r_2 = 4$, and $r_3 = 1$, giving $S = 9$. In the absence of any signal, the ranks r_j will be completely random between unity and L, whereas, in general, the r_k will tend to be smaller in the presence of signals, hence S will also be small.

An expression for the probability function of S is available, on recognizing that its genesis is similar to a so-called 'urn' problem. Imagine an urn, with L balls inside, numbered from one to L. Now draw, with replacement, K balls from the urn, and note the sum S of the K numbers drawn. The probability function P_S of S is given by

$$P_S(x) = L^{-K} \sum_{r=0}^{M} (-1)^r \binom{K}{r} \binom{x - rL - 1}{K - 1} \quad K \le x \le KL, \quad (6)$$

where

$$M = \min\left(K, \frac{x-K}{L}\right)$$

(Charalambides 2005). The cumulative probability function is

$$F_S(x) = L^{-K} \sum_{r=0}^{M} (-1)^r \binom{K}{r} \binom{x - rL}{K}. \quad (7)$$

Fig. 3 compares equation (6) with the results of one of the simulation experiments reported in Table 1.

The second statistic is the width w of the interval covered by the K coincident-peak frequencies. Again, this is expected to be smaller if peaks are due to a sinusoidal signal rather than a chance near-alignment of peaks due to noise. Analytical expressions for the probability density function (PDF) f_w and the cumulative distribution function (CDF) F_w of w are derived in Appendix A, for the special case where all ΔT are equal. The situation is considerably more complicated if the time baselines of the different data sets are different, and it is unclear if simple expressions for f_w and F_w exist for this general case.

Results for one of the simulations in Table 1 are displayed in Fig. 4. The dotted line in the bottom panel is F_w derived in the Appendix, calculated using the mean of the three values of $\ell = 0.4/\Delta T$. It evidently provides a good approximation of the empirical CDF in the lower tail, which is of greatest interest.

Returning to Fig. 1, there is a single coincidence of peaks amongst the three spectra, close to the true signal frequency of $0.05\,\mathrm{d}^{-1}$. According to the relevant entry in Table 1 ($K = 3$, $L = 5$, $N = 150$, $\nu_E = 0.2\,\mathrm{d}^{-1}$, $\overline{\Delta T} = 555$ d), the probability of obtaining a peak coincidence by chance is $p_0 = 0.02$. The sum of ranks is $S = 9$. Using equation (5), the probability of obtaining $S \le 9$, given that there is a peak coincidence, is $p_1 = 0.58$. Finally, $w = 6.86 \times 10^{-4}\,\mathrm{d}^{-1}$, with $p_2 = F_w(0.00069) = 0.23$. It may be concluded that observing this particular configuration of spectral peaks, or any more convincing, is $p_0 p_1 p_2 = 0.0026$.

It is also possible to compare the spectra two at a time. The results of doing so are summarized in Table 2. The only comparison that is significant at the conventional 5 per cent level is that of spectra 1 and 3 ($p = 0.036$).

If the search had been to weaker peaks, specifically $L = 10$, then for $K = 3$, $p_0 = 0.14$, and $\Pr(S = 9) = 0.084$, and hence the significance level would be $p_0 p_1 p_2 = 0.0027$. If only two spectra are compared, the probability of obtaining a spurious peak match rises to $p_0 = 0.76$.

4 POWER FUNCTIONS

The power of a statistical test is defined as the probability of rejecting the null hypothesis when it is indeed false. In the present context, the alternative hypothesis in equation (4) is not precise enough, and 'power' here will rather be the probability of correctly identifying ν_0, or an alias of it.

Fig. 5 illustrates that why it is necessary to allow the possibility that an alias of ν_0 is identified. Each panel of the diagram is based on 10 000 simulations of K data sets with irregular time spacings, with each simulated data set consisting of a sinusoid (frequency $\nu_0 = 0.02\,\mathrm{d}^{-1}$ and amplitude $A = 0.01$ mag) plus white noise with

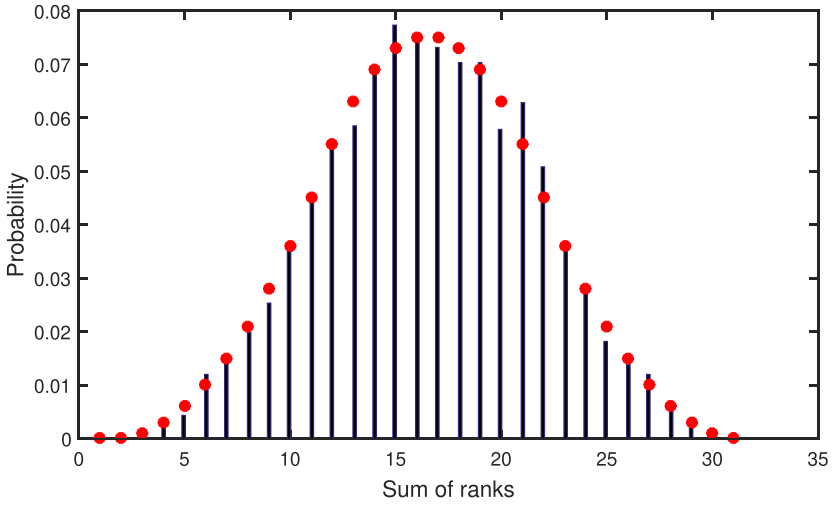

Figure 3. Distribution of the sum of ranks for one of the simulations reported in Table 1 ($K = 3$, $L = 10$, $\nu_E = 0.1$, $N = 150$, and $\overline{\Delta T} = 568$). The bars show the result of 10 000 simulations, while the filled circles denote the probability function (equation 6).

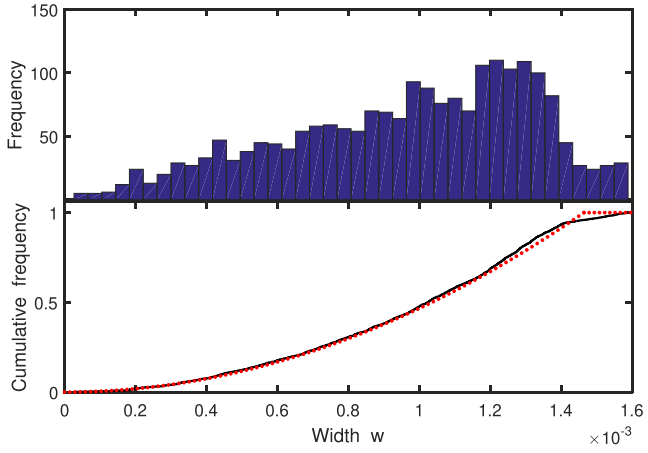

Figure 4. The distribution of the frequency interval widths defined by the positions of three near-coincident peaks, for one of the simulations reported in Table 1 ($K = 3$, $L = 10$, $\nu_E = 0.1$, $N = 150$, and $\overline{\Delta T} = 568$). The top panel shows a histogram estimate of the PDF. In the bottom panel, the solid (black) line is the empirical CDF based on the data in the top panel, while the (red) dotted line shows the theoretical result (equation A3).

$\sigma = 0.02$ mag. The spectra of each set of K data sets is searched for peak coincidences. Typically, more than one such coincidence is found. Only that with the smallest p-value is retained. In Fig. 5, the further requirement $p < 0.01$ is imposed, i.e. effectively the hypothesis test is carried out at the 1 per cent level; this leaves 6609 ($K = 2$), 9075 ($K = 3$), and 8917 ($K = 4$) estimates of ν_0. Of these, 56 ($K = 2$), 4 ($K = 3$), and 0 ($K = 4$) are neither close to ν_0 nor to any nearby alias: Put another way, the percentages of incorrect frequency identifications amongst those significant at the 1 per cent level are 0.85 per cent, 0.04 per cent, and 0 per cent for $K = 2$, 3, and 4, respectively. Only 'correct' frequency estimates are shown in Fig. 5.

The probability of correctly extracting ν_0 (or a close alias) from time series with $\sigma = 0.02$ mag white noise is plotted in Fig. 6. Time intervals between observations were taken to be exponentially distributed, with $\overline{\Delta T} \approx 550$ d. The highest $L = 5$ peaks over the range $(0, \nu_E = 0.1\,\mathrm{d}^{-1}]$ in each spectrum were taken into account.

Perhaps the biggest surprise in Fig. 6 is the good performance of the two-peak comparison when testing at the 5 per cent level. The smaller probability of identifying the correct frequency when comparing four spectra is likewise a surprise. These results are put into proper context by Fig. 7, which shows rather high probabilities of an incorrect frequency determination when testing at the 5 per cent level for $K = 2$. For $K = 4$, on the other hand, frequencies found significant are quite unlikely to be wrong.

When testing the frequency at the 1 per cent level, 'powers' of $K = 3$ and $K = 4$ are very similar. The probabilities of an incorrect determination are also comparable for amplitudes larger than about 9 mmag. This suggests that there are situations where it is sufficient to search for alignments of three peaks, with little to be gained by attempting to match four peaks. (Note though, that Fig. 5 shows that the correct alias is more likely to be selected if $K = 4$). For $K = 2$, when testing at the 1 per cent level, the 'power' is excellent for amplitudes 12 mmag or so, with the probability of a spurious frequency identification being at the 0.5 per cent level or lower.

5 TWO APPLICATIONS TO REAL DATA

The first data sets analysed in this section, observations of the star ATO 129.0947−26.1809, were extracted from the ATLAS variable star catalogue (Heinze et al. 2018). The brightness of the star was measured 84 times through the c (cyan) filter and 151 times through the o (orange) filter. Single outliers were removed from each of the two data sets. The respective time baselines covered were $\Delta T_1 = 477$ (c) and $\Delta T_2 = 558$ (o) d.

Amplitude spectra of the two data sets can be seen in Fig. 8. The ATLAS period is given as 0.054255 d ($\nu = 18.4315\,\mathrm{d}^{-1}$), but it is classified as a 'dubious' ('probably not real') variable. The highest peak in the o data spectrum is indeed at $18.431\,25\,\mathrm{d}^{-1}$.

There are no peak correspondences between the two spectra for $L < 7$, so the two limits $L = 10$ and $L = 20$ are investigated. Results are in Table 3. For $L = 10$, there is a single correspondence, at $\nu = 18.4312\,\mathrm{d}^{-1}$. In order to evaluate the significance level, 12 000 permutations of each data set were performed, and the peak correspondences were searched for. At least one peak correspondence was found in 62 of the 12 000 synthetic data sets, i.e. $p_0 = 0.0052$. The probability p_2 of the observed frequency range was evaluated

Table 2. Parameters extracted from the amplitude spectra in Fig. 1. Spectra are numbered (1–3) from top to bottom.

Spectra compared	Mean frequency	Range w	Sum of ranks	p-value
1,2	0.0500	3.8E-4	8	0.096
1,3	0.0504	3.1E-4	5	0.036
2,3	0.0502	6.9E-4	5	0.091
1,2,3	0.0502	6.9E-4	9	0.0026

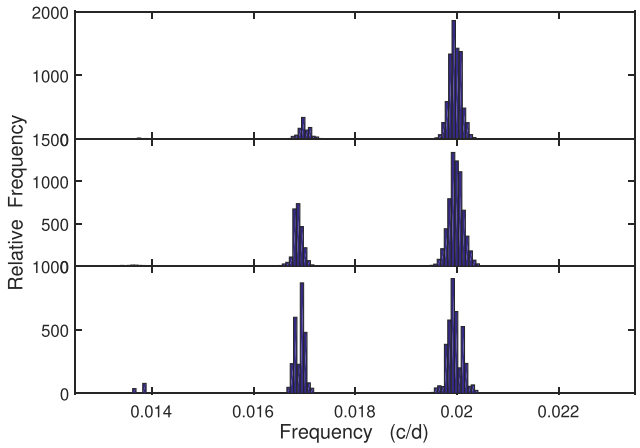

Figure 5. Estimates of ν_0 from periodogram peak correspondences for $K = 4$ (top), $K = 3$ (middle), and $K = 2$ (bottom). The results in each of the panel are based on 10 000 simulations. Note the obvious alias near $0.017\,\mathrm{d}^{-1}$ of $\nu_0 = 0.02\,\mathrm{d}^{-1}$. Given the uneven data spacings, aliasing is to be expected.

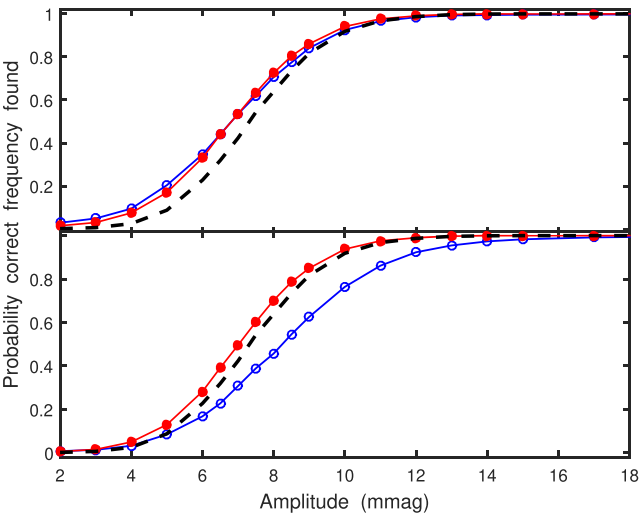

Figure 6. The probability of correctly identifying ν_0 (or a close alias), as a function of the signal amplitude. The open circles, the dots, and the broken line respectively indicate results for $K = 2$, $K = 3$, and $K = 4$. Top panel: frequencies tested at the 5 per cent level. Bottom panel: frequencies tested at the 1 per cent level. Other relevant parameter values are $N = 150$ and $L = 5$.

both from the simulation results and from equation (A3), while the probability p_1 of the sum of ranks follows from equation (7). The overall p-value is $p = p_0 p_1 p_2 = 0.00012$, where the larger of the two values of p_2 was used. It follows that the periodicity in the ATO 129.0947−26.1809 data is highly significant.

For $L = 20$, there are three peak correspondences. The first is, of course, the same as was found for $L = 10$. The other two, less significant correspondences, appear to be at $\sim 1\,\mathrm{d}^{-1}$ aliases of $\nu = 18.312\,\mathrm{d}^{-1}$ (Table 3). For $L = 20$, $p_0 = 0.015$ was obtained from 5000 permutations of the data. Overall p-values are listed in Table 3.

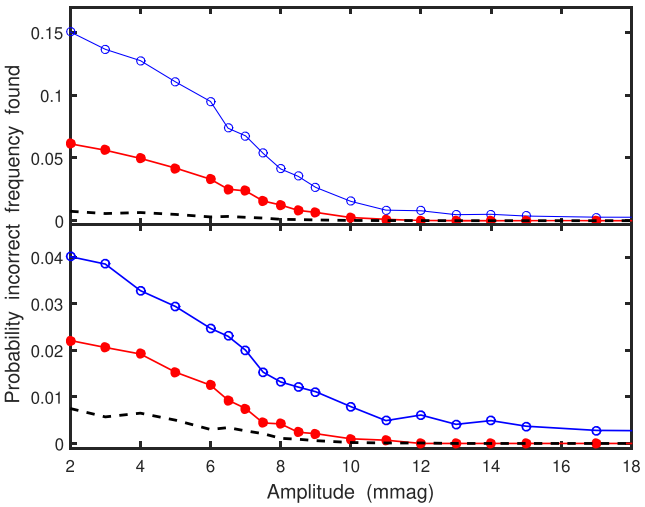

Figure 7. The probability that an incorrect frequency is found to be significant. The open circles, the dots, and the broken line respectively indicate results for $K = 2$, $K = 3$, and $K = 4$. Top panel: frequencies tested at the 5 per cent level. Bottom panel: frequencies tested at the 1 per cent level. Other relevant parameter values are $N = 150$ and $L = 5$.

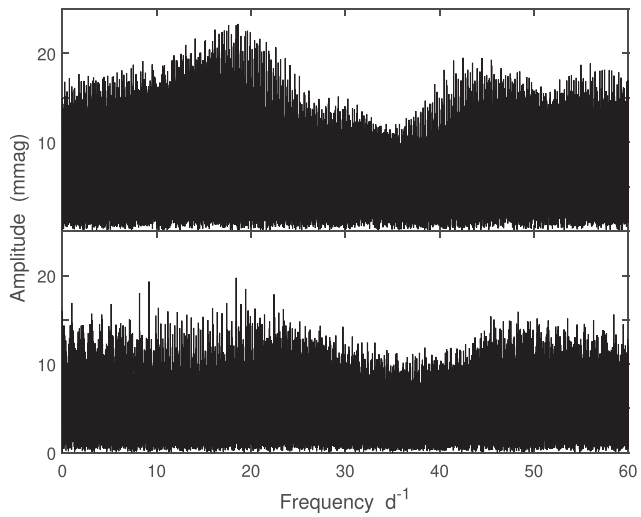

Figure 8. Amplitude spectra of ATLAS observations of ATO 129.0947−26.1809. Top panel: c filter. Bottom panel: o filter.

The second example analysis is of radial velocity measurements of ϵ Eridani (HD 22049). Mawet et al. (2019) extensively discuss the evidence for an exoplanet associated with the star and provide two new sets of radial velocities (their tables 5 and 6). Since the second of these data sets spans a time period of only 3.3 yr, whereas the

Table 3. An analysis of frequency spectra of ATLAS observations of ATO 129.0947−26.1809. Two probabilities are given for the differences in the frequencies (i.e. range) obtained, respectively, from the *c* and *o* data sets. The first was calculated from spectra of permuted data, while the second follows from equation (A3).

L	Mean frequency	Sum ranks	Prob. (ranksum)	Frequency range	Prob. (range)	Overall *p*
10	18.4312	9	0.36	9.99E-5	0.011, 0.065	1.2E-4
20	18.4312	9	0.090	9.99E-5	0.049, 0.065	8.8E-5
–	19.4338	17	0.34	1.50E-4	0.11, 0.097	5.6E-4
–	17.4286	19	0.43	1.40E-4	0.098, 0.090	6.3E-4

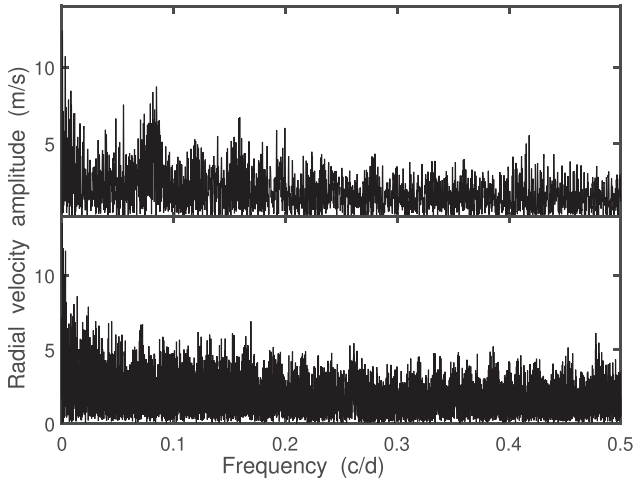

Figure 9. Amplitude spectra of two sets of radial velocity measurements of ϵ Eri. Top panel: recent Keck/HIRES data (Mawet et al. 2019). Bottom panel: Lick/Hamilton Spectrograph data (Howard & Fulton 2016).

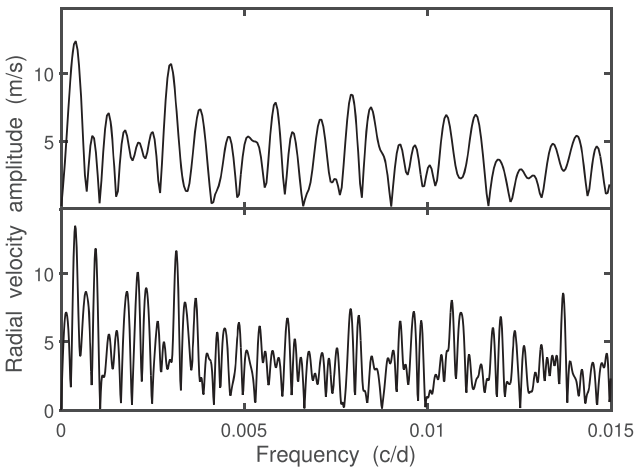

Figure 10. The low frequency sections of the spectra in Fig. 9.

periodicity of interest is ∼7.4 yr, it is replaced by the considerably more extensive radial velocities of the star collected by Howard & Fulton (2016). Amplitude spectra of the two data sets (Mawet et al. 2019, table 5: $N = 91$, $\Delta T = 7.45$ yr and Howard & Fulton 2016 : $N = 176$, $\Delta T = 24.11$ yr) are plotted in Fig. 9. Fig. 10 shows low frequency details of the spectra.

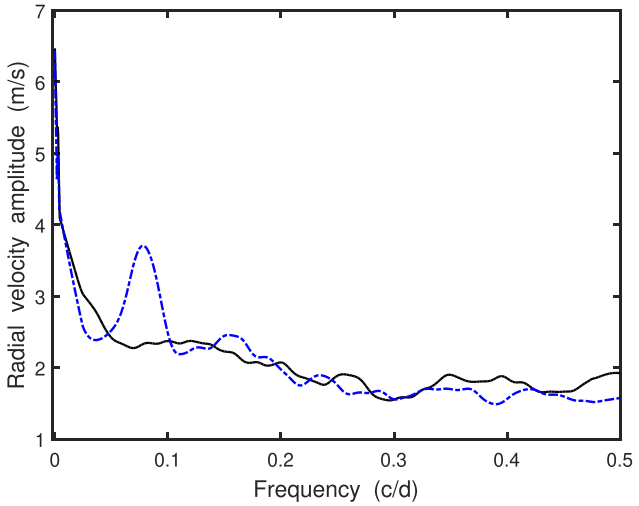

Figure 11. Smoothed versions of the two spectra in Fig. 9. The thick blue line: Keck/HIRES data. The thin black line: Lick/Hamilton Spectrograph data.

Inspection of Figs 9 and 10 reveals that both spectra are dominated by low frequency features. For $\nu_E = 0.5\,d^{-1}$ (i.e. a short period limit of 2 d), and $L = 5$, there are two peak coincidences between the two spectra, both at very low frequencies – $3.95 \times 10^{-4}\,d^{-1}$ ($P = 6.94$ yr) and $3.06 \times 10^{-3}\,d^{-1}$ ($P = 327$ d). The two frequencies are not independent: If the two data sets are prewhitened by the lower (more significant) frequency, the spectra of the residuals show no coincidences whatsoever for $L \leq 10$. We therefore proceed only with $f = 3.95 \times 10^{-4}\,d^{-1}$ – see Table 4.

In order to find the peak correspondence probability p_0, the same recipe as in the case of ATO 129.0947−26.1809 could be followed. At least one peak coincidence is obtained in only 247 out of the 12 000 permutations, i.e. $p_0 = 0.002$. However, there is an implicit assumption in the use of the permutation method, namely that the time dependence of the data can be adequately mimicked by the random rearrangement of the observations amongst the times of observation. In the frequency domain, this is tantamount to assuming that the overall shape of the spectrum is flat – something which clearly does not apply in the case of the radial velocity measurements. Fig. 11 shows the results of smoothing the two

Table 4. An analysis of frequency spectra of two sets of radial velocity measurements of ϵ Eri.

L	Mean frequency	Sum ranks	Prob. (ranksum)	Frequency range	Prob. (range)	Overall *p*
5	3.954E-4	2	0.13	1.81E-5	0.094	0.006

spectra; the plot demonstrates the substantial power excess, and hence inflated likelihood of substantial peaks, at low frequencies.

Fortunately, a modification of the premutation method, which delivers more realistc spectra, is fairly easy: The flat spectra of permuted data are simply multiplied by the smooth functions in Fig. 11, so that the simulated have the same overall shapes as the observed spectra. The importance of this correction is manifested by the increase of p_0 to 0.51. The reason for the large change is not difficult to find: Effectively, the largest peaks will overwhelmingly be found in a narrow range at very low frequencies so that the probability of peak coincidences is greatly inflated.

The spectral shape also affects the distribution of the sum of peak ranks, increasing the probabilities of small sums relative to larger values. Thus $p_1 = 0.13$ instead of 0.04 for the flat-spectrum case. The statistic w is uniformly distributed over (0, $1.925E-4$), giving $p_2 = 0.094$ for the observed value of 1.808×10^{-5} d^{-1}. The overall significance level for the peak coincidence is 0.006, i.e. highly significant.

6 DISCUSSION

An intriguing possibility raised by the theory of this paper is applications to single data sets, by separating these into multiple sets. This should be particularly useful in cases where there might be multiple weak periodicities present, as such data present problems for the usual approaches based on contrasting peak heights with noise levels. The results above suggest that separating time series into $K = 3$ subsets may be particularly fruitful.

ACKNOWLEDGEMENTS

The scientific editor made a useful suggestion which led to the inclusion of the second analysis in Section 5.

REFERENCES

Charalambides C. A., 2005, Combinatorial Methods in Discrete Distributions. John Wiley & Sons Inc., Hoboken, New Jersey, USA
Frescura F. A. M., Engelbrecht C. A., Frank B. S., 2008, MNRAS, 388, 1693
Heinze A. N. et al., 2018, AJ, 156, 241
Howard A. W., Fulton B. J., 2016, PASP, 128, 114401
Kovács G., 1981, Ap&SS, 78, 175
Mawet D. et al., 2019, AJ, 157, 33

APPENDIX A: THE PROBABILITY DENSITY FUNCTION OF THE FREQUENCY INTERVAL SPANNED BY K NEAR-COINCIDENT SPECTRAL PEAKS

It is given that the K peaks are coincident, i.e. equation (4) holds, where $P_j = [P_{0j} - \ell, P_{0j} + \ell] = [a_j, b_j]$ ($j = 1, 2, \cdots, K$). It is required to find, in the first instance, the cumulative distribution function (CDF) of the interval spanned by the P_{0j}, i.e.

$$w = \max_j P_{0j} - \min_j P_{0j} \ .$$

For convenience, assign the index $j = 1$ to the smallest P_{0j}, i.e. $a_1 < a_j$ ($j = 2, 3, \ldots, K$) also. Since all intervals P_j ($j \geq 2$) must overlap P_1, it follows that all a_j ($j > 2$) lie in the interval P_1. In fact, each of the a_j ($j > 2$) is uniformly randomly distributed in $[a_1, b_1]$. The maximum of the a_j, i.e. the upper order statistic $a_{(K)}$, then has CDF

$$F_y(y) = [F_a(y)]^{K-1}, \tag{A1}$$

where F_a is the CDF of a uniform distribution on $[a_1, b_1]$. The latter is easily shown to be

$$F_a(y) \begin{cases} 0 & y < a_1 \\ (y - a_1)(b_1 - a_1) & a_1 \leq y \leq b_1. \\ 1 & y > b_1 \end{cases} \tag{A2}$$

From the definition of w,

$$w = a_{(K)} - a_1$$

and hence, from equations (A1) and (A2),

$$\begin{aligned} F_w(w) = F_y(w + a_1) &= \begin{cases} 0 & w < 0 \\ [w/(b_1 - a_1)]^{K-1} & 0 \leq w \leq b_1 - a_1 \\ 1 & w > b_1 - a_1 \end{cases} \\ &= \begin{cases} 0 & w < 0 \\ (w/2\ell)^{K-1} & 0 \leq w \leq 2\ell. \\ 1 & w > 2\ell \end{cases} \end{aligned} \tag{A3}$$

The PDF of w follows immediately as

$$f_w(w) = \begin{cases} 0 & w < 0 \\ (K-1)w^{K-2}/(2\ell)^{K-1} & 0 \leq w \leq 2\ell. \\ 0 & w > 2\ell \end{cases} \tag{A4}$$

This paper has been typeset from a TEX/LATEX file prepared by the author.