Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **459,** 3012–3024 (2016) doi:10.1093/mnras/stw848
Advance Access publication 2016 April 13

Identifying multiple periodicities in sparse photon event time series

Chris Koen★

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

Accepted 2016 April 11. Received 2016 April 8; in original form 2015 December 23

ABSTRACT

The data considered are event times (e.g. photon arrival times, or the occurrence of sharp pulses). The source is multiperiodic, or the data could be multiperiodic because several unresolved sources contribute to the time series. Most events may be unobserved, either because the source is intermittent, or because some events are below the detection limit. The data may also be contaminated by spurious pulses. The problem considered is the determination of the periods in the data. A two-step procedure is proposed: in the first, a likely period is identified; in the second, events associated with this periodicity are removed from the time series. The steps are repeated until the remaining events do not exhibit any periodicity. A number of period-finding methods from the literature are reviewed, and a new maximum likelihood statistic is also introduced. It is shown that the latter is competitive compared to other techniques. The proposed methodology is tested on simulated data. Observations of two rotating radio transients are discussed, but contrary to claims in the literature, no evidence for multiperiodicity could be found.

Key words: methods: statistical – pulsars: general.

1 INTRODUCTION

Let Y_t be a periodic point process, i.e. a delta function which is non-zero at times $t = 0, P, 2P, 3P, \ldots$, and zero at all other times. Four extensions to this basic configuration are considered below: (i) series made up of the superposition of several such processes, each with a distinct period; (ii) processes allowed to be sparse, i.e. the majority of pulses may be unobserved; (iii) admitting measurement errors in the times of the pulses; and (iv) the presence of 'noise' pulses, i.e. non-periodic contaminating pulses. The problem is then to deduce the number of independent periodic processes, and their periods, from the train Y_t of 'noisy' measurements.

For a single periodicity P,

$$Y_j = Y_0 + N_j P + e_j\,, \qquad j = 1, \ldots, N, \tag{1}$$

where Y_0 is a time zero-point, the N_j are integers, and the e_j are measurement errors with mean zero and small variance σ^2. The variance is assumed constant, but not necessarily known. 'Small' here means that $\sigma \ll P$. It therefore follows that each interval $Y_j - Y_0$ is, to a good approximation, an integer multiple of P. The same is true of the differenced sequence of time points:

$$\Delta_1 Y_j = Y_{j+1} - Y_j\,, \qquad j = 1, 2, \ldots, N-1. \tag{2}$$

It is perhaps worth making explicit the distinction between the effects of the measurement errors e_j and those of random 'jitter' in the period P. At first sight these are the same, but in fact equation (1) implicitly assumes a completely stable period. Consider the case

$$P_i = P_0 + \epsilon_i,$$

where P_0 is the mean period and ϵ_i is zero mean white noise with variance σ_ϵ^2. It follows that the true pulse arrival times are

$$\begin{aligned} T_1 &= T_0 + P_1 = T_0 + P_0 + \epsilon_1 \\ T_2 &= T_1 + P_2 = T_0 + 2P_0 + \epsilon_1 + \epsilon_2 \\ \cdots &= \cdots \\ T_j &= T_0 + jP_0 + \sum_{i=1}^{j} \epsilon_i, \end{aligned}$$

i.e. a random walk with drift. Note, in particular, that the variance of T_j is proportional to j. For small σ_ϵ white noise period jitter could therefore be incorporated into (1) by replacing e_j by

$$\begin{aligned} e'_j &= e_j + \sum_{i=1}^{N_j} \epsilon_i \\ e'_j &\sim D(0, \sigma^2 + N_j\sigma_\epsilon^2), \end{aligned} \tag{3}$$

where the last expression indicates that e'_j is distributed with zero mean and variance $\sigma^2 + N_j\sigma_\epsilon^2$. For large jitter the determination of N_j in (1) may become ambiguous. This can be circumvented by working with intervals between pulses (as in equation 2), rather than referring observations to the time zero-point Y_0 in (1). See Nishiguchi & Kobayashi (2000) for an extensive discussion of this point. Zero jitter will be assumed below, although some of the algorithms are resistant to moderate σ_ϵ.

★ E-mail: ckoen@uwc.ac.za

If there are multiple periodicities, then different Y_t correspond to different processes, and some (or even all) of the ΔY_j will not be multiples of any of the periods. This is a considerable complication.

The work in this paper was stimulated primarily by Palliyaguru et al. (2011), who discuss multiperiodicities in pulse arrival times of rotating radio transients (RRATs). A similar problem would be posed by several weak unresolved sources lying in the same field of view of a detector – see, for example, the paper by Chennamangalam & Lorimer (2014), in which it is shown that there may be a very large pulsar population in the direction of the Galactic centre. Two other possible areas of application which come to mind are possible periodicities in the onset of flares in UV Ceti stars and the detection of regularities in period spacings in periodograms of pulsating stars (see e.g. Reed et al. 2010). For references to engineering applications, see Ansari, Zhang & Mahlke (2009). There are other areas in which these methods are used: for example, Le Bot et al. (2013) discuss an application to the study of whale communications.

Pulse widths are not taken into account in what follows, i.e. the time series is treated as a point process, and will be referred to as such.

2 REVIEW OF SOME RELEVANT LITERATURE

Several different approaches to the *singly-periodic* case can be found in the literature.

(i) The periodogram of a train of delta functions located at the times of pulse arrival times is

$$I(\omega) = \frac{1}{N}\left|\sum_{j=1}^{N}\exp[-\mathrm{i}\omega t_j]\right|^2, \tag{4}$$

where ω is a trial frequency (e.g. Fogel & Gavish 1988). Peaks are observed at $\omega = 2\pi/P$ and its harmonics.

(ii) Minimization with respect to the trial period P_t of the least-squares (LS) criterion

$$\sum_j \epsilon_j^2, \tag{5}$$

where the residuals are defined by

$$\epsilon_j = \mathrm{mod}(\Delta Y_j, P_\mathrm{t})/P_\mathrm{t}\,. \tag{6}$$

Sidiropoulos, Swami & Sadler (2005) consider three different forms of the differences ΔY_j in equation (6). The first is the subtraction of successive observation times, as defined in equation (2). Secondly,

$$\Delta_2 Y_j = Y_{2j} - Y_{2j-1} \qquad j = 1, 2, \ldots, N/2. \tag{7}$$

The advantage of this over equation (2) is that the $\Delta_2 Y_j$ are all uncorrelated. The disadvantage is that the number of data for analysis is halved. Thirdly, all possible positive differences:

$$\Delta_3 Y_j = Y_k - Y_i \qquad k = 2, 3, \ldots, N;$$
$$i = 1, 2, \ldots, k-1;\ j = 1, 2, \ldots, N(N-1)/2. \tag{8}$$

The disadvantages of working with the $\Delta_3 Y_j$ are the lack of independence, and computational cost which increases with increasing sample size as N^2. Sidiropoulos et al. (2005) find that the LS procedure using the $\Delta_3 Y_j$ is superior to that based on the $\Delta_1 Y_j$, which again outperforms LS using the $\Delta_2 Y_j$. Furthermore, in their simulations the LS method is not as efficient as the periodogram, particularly for large σ.

(iii) Clarkson (2008) introduced the time differences

$$\Delta_4 Y_j = Y_{N/2+j} - Y_j \qquad j = 1, 2, \ldots, N/2 \tag{9}$$

into the LS methodology described in (ii). He showed that these intervals give considerably more accurate estimates of P than those based on $\Delta_2 Y_j$, while retaining the virtue of independence and low computational cost.

(iv) Clarkson (2008) also considers maximum likelihood estimation (MLE) of P, using a lattice search to obtain the unknown N_j in equation (1). He demonstrates close similarities between this methodology and the periodogram (4). We note in passing that Clarkson implicitly assumes that the error variance σ^2 is known: the implication is that MLE effectively reduces to sum of squares minimization, i.e. LS.

(v) For moderate to small measurement errors in the pulse arrival times (i.e. $\sigma \ll P$), the intervals ΔY_j are close to integer multiples of the true period P. Put differently, P is close to the greatest common denominator (GCD) of the collection of ΔY_j. Casey & Sadler (1996) and Sadler & Casey (1998) presented a generalization of Euclid's method for integer GCD determination, to the case where the GCD might be an arbitrary real number, and applied it to the present problem. The method is computationally very fast, but requires certain tuning parameters. These can be awkward to determine for very sparse data in which very few (or none) of the $\Delta_1 Y \approx P$. The algorithm is also adversely affected by 'noise' pulses: 'These are, in general, quite harmful to the estimation of the gcd' (Casey & Sadler 1996).

(vi) Many methods are based on folding the data with a choice of trial periods, and studying the distributional properties of the scaled residuals ϵ_j defined in equation (6). If P_t is far removed from the correct period, then the ϵ will be distributed uniformly over (0, 1), whereas if $P_\mathrm{t} = P$ all the ϵ will be close to zero or unity. This suggests period determination by finding values of P_t for which the distribution of the ϵ_j are strongly non-uniform. The Kolmogorov–Smirnov (KS) statistic has been used for this purpose in the astronomy literature (e.g. Kawaler 1988).

(vii) Phased data, such as those obtained by folding, are best seen as being *circular*, i.e. the phase points zero and unity are the same. Strictly speaking, uniformity should therefore be tested on the circle with unit circumference, rather than on the line interval [0, 1]. This lies at the basis of many algorithms which have been used in the pulsar literature – see e.g. de Jager, Swanepoel & Raubenheimer (1989), and references therein. The Rayleigh statistic

$$Z = \frac{1}{N}\left[\left(\sum_{j=1}^{N}\cos 2\pi\epsilon_j\right)^2 + \left(\sum_{j=1}^{N}\sin 2\pi\epsilon_j\right)^2\right] \tag{10}$$

is a typical form.

There is also a substantial body of work in the signal processing literature which deals with the identification of *multiple periodicities* in point processes. Techniques are often referred to as 'de-interleaving'. References to a host of different algorithms can be found in Ansari et al. (2009). The methodology generally appears to be aimed at cases with relatively few missing pulses, so that, for example, histograms of the Δ_j will reveal pulse periods directly. A similar method is the 'sequence search algorithm' discussed by Ansari et al. (2009): it essentially relies on finding numbers of close repetitions amongst the $\Delta_3 Y_j$ defined in equation (8). In symbolic form,

$$C(\tau) = \sum_j \mathcal{I}(|\tau - \Delta_3 Y_j| < \delta), \tag{11}$$

where τ is a trial period, δ is a small positive number and $\mathcal{I}$ is the indicator function (equal to unity if its argument is true, zero otherwise). Clearly, if more than say 90 per cent of the pulses are missing, this will not be a very effective technique.

Similar remarks apply to the statistic devised by Nishiguchi & Kobayashi (2000):

$$D(\tau) = \left| \sum_{k=2}^{N} \sum_{\ell=1}^{k-1} \mathcal{I}(|\tau - (Y_k - Y_\ell)| < \delta) \exp(2\pi \mathrm{i} Y_k/\tau) \right|, \qquad (12)$$

which is related to both the periodogram (4) and the criterion (11). As demonstrated by the authors, $D(\tau)$ is efficient at discounting multiples of the true period P.

By contrast to the applications reviewed above, this paper is aimed at situations where very few, if any, values of the ΔY_j may be repeated. Note that all the single-period algorithms discussed above, with the exception of the GCD, could, in principle, be applied also to multiperiodic data, although the performance may be degraded. Orsi, Moore & Mahoney (1999) studied the efficacy of the periodogram in the mutiperiodic context, but only for complete data, or allowing for a 1 per cent probability for individual pulses to be missing. As will be shown below, it none the less still works well even if a substantial fraction of pulses is missing.

If the fraction p of observed pulses is small, less than about 10 per cent, then the number of pairs of consecutively observed pulses is roughly binomially distributed if pulses are randomly censored. The expected number of instances of pulses separated by time P (as opposed to multiples of P) is only $\sim p^2 N_{\rm tot}$, where $N_{\rm tot}$ is the total number of pulse periods spanned by the observations. Thus, for example, if $p = 0.03$ and $N_{\rm tot} = 1000$, on average there will be a single pulse separation of length P. More generally, the expected number of pulse pairs is $p(N_{\rm tot}p - 1)$, regardless of the value of p.

A simple modification will improve the efficiency of $C(\tau)$ and $D(\tau)$ (equations 11 and 12) in cases with many missing pulses:

$$C'(\tau) = \sum_{j} \sum_{m=1}^{M(\tau)} \mathcal{I}(|m\tau - \Delta_3 Y_j|/\tau < \delta)$$

$$D'(\tau) = \left| \sum_{k=2}^{N} \sum_{\ell=1}^{k-1} \sum_{m=1}^{M(\tau)} \mathcal{I}(|m\tau - (Y_k - Y_\ell)|/\tau < \delta) \exp(2\mathrm{i}\pi Y_k/\tau) \right|, \qquad (13)$$

where M is chosen so that

$$M\tau \le Y_N - Y_1 < (M+1)\tau.$$

Note that searching over a grid of trial values τ will be computationally demanding: the required resolution is set by the largest value $Y_N - Y_1$ of $\Delta_3 Y_j$ (see Section 4).

Fig. 1 contains an illustration, using simulated data with two periods, 2 and 3.13 time units. Pulse detection probabilities of $p = 0.07$ were assumed for both, giving 35 and 22 data points, respectively, over a baseline $T = 1000$. Comparison of the two panels shows the efficacy of D' in suppressing subharmonics.

The de-interleaving approach below is similar to the standard method used in multiperiodic asteroseismic data analysis: repeated application of the two-step procedure of (i) identifying the most promising candidate period, followed by (ii) prewhitening it from the data. Step (i) is explored in the next three sections of the paper, while a prewhitening algorithm is described in Section 8.

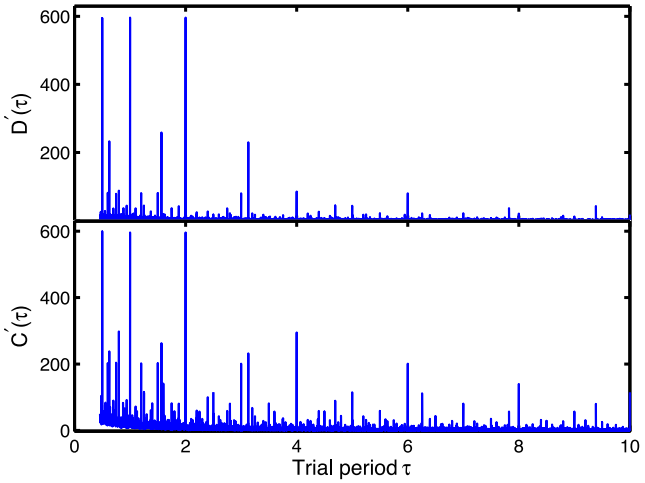

Figure 1. Illustrative simulation results for the equations (13) variants of the statistics in equations (11) and (12). The fraction of observed pulses is 7 per cent, and the true periods are $P = 2$ and $P = 3.13$. Note that periods $P/2$, $P/3$, ... fit the data equally as well as the true periods.

3 CANDIDATE PERIOD-FINDING STATISTICS

Below, period-finding is performed by testing for non-uniformity, as in (vi) and (vii) above. This has the virtue that the necessary calculations are simple, and that non-uniformity 'periodograms' can be constructed with relatively little computational effort for many trial periods. Seven statistics were tried, five designed to test for non-uniformity on the interval [0, 1], and two for testing for circular non-uniformity.

(i) The KS statistic. This has been used in a number of previous studies, e.g. Kawaler (1988), Winget et al. (1991), Silvotti et al. (1999) and Provencal et al. (2012). The cumulative distribution of the ϵ_j is tested against the null

$$F(x) = x\,, \qquad 0 \le x \le 1.$$

The ϵ_j are ordered from small to large to give the order statistics $\epsilon_{(1)}, \epsilon_{(2)}, \ldots, \epsilon_{(N)}$. The KS statistic is

$$D = \sqrt{N} \max_j |\epsilon_{(j)} - j/N|$$

and significance levels can be calculated from the well-known formula

$$\Pr(D \le x) = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} \exp\left[-\frac{(2k-1)^2 \pi^2}{8x^2} \right].$$

It is noted in passing that a circularized form of the KS statistic, the Kuiper (1960) statistic, has been applied to X-ray count data by Paltani (2004). Protheroe (1985) has shown that his circular uniformity test [described in (vi) below] is more powerful than the Kuiper test if deviations from uniformity are very localized, as in the present case.

(ii) The Anderson–Darling (AD) statistic generally has better power than the KS statistic (e.g. D'Agostino & Stephens 1986). In the present context, the statistic is

$$AD = -N - \frac{1}{N} \sum_{j=1}^{N} (2j-1)[\log(\epsilon_{(j)}) + \log(1 - \epsilon_{(n+1-j)})].$$

Percentage points for testing specifically for uniformity can be found in Rahman, Pearson & Heien (2006); for convenience, these are repeated in Table 1.

Table 1. Percentiles of three of the statistics discussed in Section 3. Values for the AD statistic are from Rahman et al. (2006); other results were obtained by simulation. Means and standard deviations of the likelihood ratio statistic were calculated from the results for a few different sample sizes – see Fig. 2.

Percentage	90	95	97.5	99	99.5	99.9
			AD			
	1.94	2.50	3.09	3.90	4.54	6.03
			PR			
$N = 500$	11.16	11.20	11.23	11.27	11.30	11.36
$N = 1000$	12.51	12.54	12.56	12.59	12.61	12.65
			Likelihood ratio statistic Λ			
(Mean)	3.35	4.78	6.24	8.16	9.60	12.82
(Std. dev.)	0.01	0.02	0.07	0.10	0.16	0.19

(iii) Degroote et al. (2009) proposed the 'matrix test', which is based on counting the number of ϵ_j in the intervals $[0, \delta)$ and $(1 - \delta, 1]$, where $0 < \delta \ll 1$. For trial periods, P_t very different from the true period this number will be of the order of $N\delta$; for a singly periodic point process with small timing errors ($\sigma \ll \delta$) it will be close to N if P_t is close to the true period.

It is instructive to compare the Degroote et al. (2009) statistic to the modified sequence search form $C'(\tau)$ in equation (13). Note that for given τ and ΔY_j,

$$|m\tau - \Delta Y_j| < \delta$$

will be true for at most one integer m if $\delta \ll 1$. It follows that

$$\sum_{m=1}^{M(\tau)} \mathcal{I}(|m\tau - \Delta_3 Y_j| < \delta) \equiv \mathcal{I}\left(\min_m |m\tau - \Delta_3 Y_j| < \delta\right).$$

However,

$$\min_m |m\tau - \Delta Y_j| \equiv \min(d_j, 1 - d_j),$$

where

$$d_j = \mathrm{mod}(\Delta Y_j, \tau) = \tau\epsilon_j$$

and hence an equivalent form of the modified sequence search statistic is

$$C'(\tau) = \sum_j \mathcal{I}[\min(d_j, 1 - d_j) < \delta]. \quad (14)$$

The Degroote et al. (2009) statistic is

$$G(\tau) = \sum_j \mathcal{I}[\min(\epsilon_j, 1 - \epsilon_j) < \delta] = \sum_j \mathcal{I}[\min(d_j, 1 - d_j)/\tau < \delta]. \quad (15)$$

The only difference between C' and G is that the residuals are scaled by the trial period τ in the latter.

In implementations, it is convenient to make the transformation (20) below, shifting the origin to 0.5, so that

$$G(\tau) = \sum_j \mathcal{I}\left(|\epsilon_j - 0.5| < \delta\right). \quad (16)$$

The statistic has a binomial distribution with parameters 2δ and N. The pth percentile is obtainable by solving

$$B(1 - 2\delta; N - x, x + 1) = p$$

for x (B being the regularized incomplete beta function). Alternatively, the significance level α corresponding to an observed x is given by

$$\alpha = 1 - B(1 - 2\delta; N - x, x + 1).$$

For large N, the distribution of $G(\tau)$ can be approximated by the normal with mean and variance both equal to $2\delta N$.

(iv) Rayner & Rayner (2001) performed a power study of the Neyman (1937) test statistics for uniformity, and suggested that the fourth-order statistic may be particularly useful. The statistics are calculated as follows:

$$g_1 = 2\sqrt{3}\sum_j (\epsilon_j - 0.5)$$

$$g_2 = \sqrt{5}\sum_j [6(\epsilon_j - 0.5)^2 - 0.5]$$

$$g_3 = \sqrt{7}\sum_j [20(\epsilon_j - 0.5)^3 - 3(\epsilon_j - 0.5)]$$

$$g_4 = 3\sum_j [70(\epsilon_j - 0.5)^4 - 15(\epsilon_j - 0.5)^2 + 0.375]$$

and the statistic Ψ_r of the order of r is given by

$$\Psi_r = \frac{1}{N}\sum_{k=1}^{r} g_k^2.$$

For large samples Ψ_r is distributed as χ^2_r.

(v) Swanepoel & de Beer (1990) developed a statistic to test for narrow periodic pulses: calculate the absolute values of all $M = N(N-1)/2$ distinct phase differences, i.e.

$$W_{ij} = |\epsilon_i - \epsilon_j|, \quad i = 1, 2, \ldots, N-1; j = i+1, \ldots, N. \quad (17)$$

Arrange the W_{ij} in ascending order, giving the set of order statistics $W_{(1)}, W_{(2)}, \ldots, W_{(M)}$. Now define

$$\beta = \left[\frac{M}{2^{1/3}N^{2/3}}\right] \approx \left[(N/2)^{4/3}\right],$$

where the square brackets indicate rounding down to the nearest integer (the 'floor' operation). Then, the Swanepoel & de Beer (1990) statistic is

$$\mathrm{SB} = 2^{1/3}N^{2/3}W_{(\beta)}.$$

The statistic SB differs from all others discussed here in that deviations from uniformity lead to *small* values of the statistic. The authors provided a table with percentiles for $N \leq 170$. For larger values of N, percentiles are given by

$$C_N(p) = 0.5 + 2^{-1/3}N^{-2/3}[0.25 - \Phi^{-1}(p)],$$

where Φ^{-1} is the inverse standard Gaussian cumulative distribution function. [For example, $\Phi^{-1}(0.95) = 1.645$.]

(vi) A second narrow-pulse statistic is due to Protheroe (1985):

$$\mathrm{PR} = \frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} V_{ij}^{-1}, \quad (18)$$

where

$$V_{ij} = \frac{1}{N} + \frac{1}{2} - \left|W_{ij} - \frac{1}{2}\right|$$

with W_{ij} defined as in equation (17). Note that V_{ij} transforms the spacing W_{ij} between ϵ_i and ϵ_j to reflect the fact that these are phases. Protheroe (1985) provides a table of percentiles, together

with interpolation formulae, for $N \leq 200$. Simulated percentiles for $N = 500$ and $N = 1000$ are given in Table 1.

(vii) It is possible to tailor-make a statistic for the problem studied in this paper. If the data are folded with respect to an incorrect period, the scaled ϵ_j are uniformly distributed on [0, 1]. If the data are folded with respect to a true period, then roughly half the ϵ_j associated with the periodicity will be very close to zero, and the other half very close to unity. The remainder of the ϵ (due to either other periodicities, or to noise pulses) will be uniformly spread over [0, 1]. This means that the distribution of the scaled residuals ϵ_j is a mixture of three distributions, the two half-Gaussians

$$f_1(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{1-\epsilon}{\sigma}\right)^2\right]$$
$$f_2(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\epsilon}{\sigma}\right)^2\right] \tag{19}$$

and a uniform distribution on [0, 1]. By making the transformation

$$\xi_j = \begin{cases} \epsilon_j + 0.5 & \epsilon_j < 0.5 \\ 1 - \epsilon_j & \epsilon_j > 0.5, \end{cases} \tag{20}$$

the ξ_j follow a mixture of a uniform distribution on [0, 1] and a single Gaussian with mean 0.5 and standard deviation σ. It will be assumed below that this transformation has been made, and that the distributions of the ϵ is of the mixture form

$$f(\epsilon) = (1-\alpha) + \frac{\alpha}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\epsilon - 0.5}{\sigma}\right)^2\right], \tag{21}$$

when folding is done with respect to a true period. In equation (21), the factor α represents the fraction of the total data which are associated with the folding period.

The proposed test quantity is the likelihood ratio statistic

$$\Lambda = 2\{\max[\log L(H1)] - \max[\log L(H0)]\},$$

where $L(H0)$ and $L(H1)$ are, respectively, the statistical likelihoods under the null and alternative hypotheses. Under the null hypothesis, the ϵ_j are uniformly distributed over the unit interval, and

$$L(H0) = \prod_j 1 = 1,$$

while equation (21) holds under the alternative. It follows that

$$\Lambda = 2 \max_{\alpha,\sigma} \sum_j \log f(\epsilon_j) \tag{22}$$

with $f(\epsilon)$ given by equation (21).

In many settings likelihood ratio statistics have χ^2 distributions with known degrees of freedom. However, in the present case regularity conditions are not satisfied – the null hypothesis is essentially

$$H0: \ \alpha = 0,$$

i.e. the 'nuisance parameter' σ is unspecified under the null hypothesis. The distribution of Λ in such cases is often non-standard. In the present instance, the distribution of Λ was determined by simulation. Since the results for sample sizes in the range 50–1 000 differed little (see Fig. 2), only mean percentiles are reported in Table 1.

4 PRACTICALITIES

There are a number of issues which are of practical importance in implementation of the theory above:

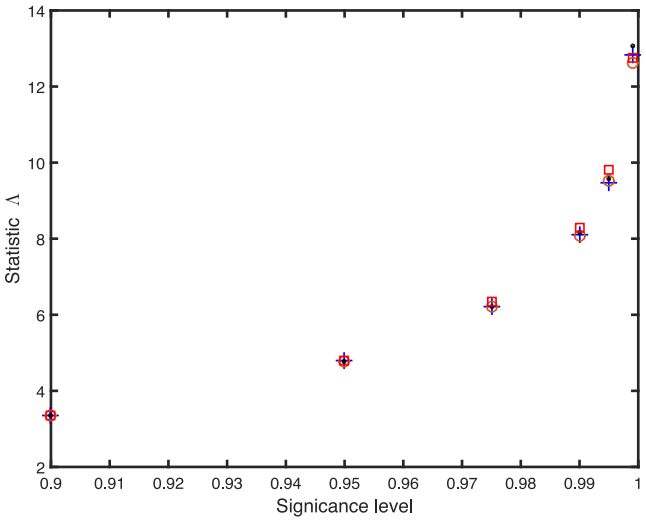

Figure 2. The likelihood ratio statistic Λ for various percentage points. Values are shown for sample sizes 50 (plus signs), 200 (open circles), 500 (filled dots) and 1000 (squares).

(i) In the case of a single periodicity, the length of the shortest interval ΔY_j (as defined in equation 2) provides an approximate upper limit on the period. (In the absence of measurement errors the limit would be exact). The situation is more complex if there is multiperiodicity: since pulses at Y_{j+1} and Y_j may be from different processes, ΔY_j is not necessarily a multiple of *any* of the periods.

(ii) If ΔY_j is a multiple of a specific period P_k, then the longer ΔY_j, the more accurate the estimate of P_k:

$$P_k \approx \Delta Y_j / N_{jk} = P_k + (e_{j+1} - e_j)/N_{jk}, \tag{23}$$

where N_{jk} is the closest integer number of times P_k fits into the interval ΔY_j. Clearly, the larger ΔY_j, the larger N_{jk}, and the smaller the error term in equation (23).

(iii) An implication of point (i) is that when working with intervals between events, then it may be necessary to test the entire collection of intervals

$$\Delta_3 Y_j = Y_k - Y_i, \qquad i = 1, 2, \ldots, N-1;$$
$$k = i+1, i+2, \ldots, N$$

in order to ensure that all periodicities are detected.

(iv) Let

$$\Delta Y_j / P = n_j + r,$$

where n_j is an integer and $0 \leq r < 1$. Changing P by some small amount δP generally changes the remainder r by

$$|\delta r| = \frac{\Delta Y_j}{P^2} \delta P. \tag{24}$$

Clearly, in a grid search through trial periods, 'phase' changes $|\delta r|$ should be small, less than say 0.05.

(vi) Equation (24) implies that δr increases with increasing ΔY_j, and decreases with increasing P. This suggests that the period resolution is determined by the largest value of ΔY, and that it improves with decreasing period.

(vii) For small P, computation time may be exorbitant at the required resolution. Increasing δP, i.e. using a coarse resolution, is not an option, as phases corresponding to large ΔY will be randomly spread over (0, 1) if the trial period differs slightly from the true value. This suggests placing an upper limit on values of ΔY_j which are used in the computations.

(viii) Of course, using only smaller ΔY_j will compromise the accuracy of the period determination, but this is easily remedied by a second, more accurate, round of restricted period searches around promising candidate values.

(ix) Equation (24) furthermore suggests that the period search grid should not be regularly spaced, but that trial periods should be calculated from

$$\delta P = \frac{P^2}{\Delta Y_{\max}} \delta r_{\max}, \tag{25}$$

where $\delta r_{\max}$ is the maximal allowable phase change corresponding to the maximum interval $Y_{\max}$.

(x) The choice

$$\Delta Y_{\max} = nP, \tag{26}$$

for some fixed n, seems reasonable: the maximum interval should cover n cycles. In that case,

$$\delta P = kP\ , \qquad k = \delta r_{\max}/n \ll 1,$$

and it follows that a grid of M trial period values would be

$$P_0, (1+k)P_0, (1+k)^2 P_0, \ldots, (1+k)^M P_0.$$

If it is required to cover the interval $[P_0, P_e]$, then the number of trial periods would be

$$M \approx \frac{1}{k} \log \frac{P_e}{P_0} = \frac{n}{r_{\max}} \log \frac{P_e}{P_0}. \tag{27}$$

As an example, $M = 3000$ trial periods are needed to cover the interval [0.05, 1] with $\delta r_{\max} = 0.05$ and $n = 50$; this increases to $M = 5300$ if $P_e = 10$.

(xi) Only time intervals satisfying

$$\Delta Y_j \geq P \tag{28}$$

should of course be included in the analysis.

(xii) Since $\nu = 1/P$ implies that $|\delta\nu| \approx |\delta P|/P^2$, (25) implies a regularly spaced *frequency* search grid with

$$\delta\nu = \frac{\delta r_{\max}}{\Delta Y_{\max}}. \tag{29}$$

If the periodogram is used for period searching, the appropriate frequency spacing is therefore $\delta r_{\max}/(Y_N - Y_0)$.

(xiii) There is no natural lower limit for candidate periods. If there is only one periodicity, and no noise pulses, then $P \leq \min_j \Delta_1 Y_j$. In general, if any period P fits the data, then so will P/n, for integer n. This implies that if a sequence of spectral peaks of similar strength are seen at multiples of frequency ν_0, then the correct 'alias' is the lowest value ν_0.

5 POWER OF THE STATISTICS

At a first glance, it may seem that there are numerous parameters which need to be taken into account in a comparison of the various test statistics. However, the problem can be considerably simplified. First, it is assumed that the time points are folded with respect to a period close to the true value: in practice this can be accomplished by using a fine enough period search grid. There are then only three relevant parameters – the number of data belonging to the periodic process; the number of 'noise' data (including points belonging to other periodicities); and the measurement error variance. (In the applications below, it is convenient to replace the former two numbers by the total number of data, and the fraction α belonging to the periodic series.) Note that the effects of non-constant periods can also be studied, at least for slight variability, by inflating the measurement errors.

In order to compare the performance of the different statistics it is therefore sufficient to consider a variety of values of N, α and σ, and to compare the efficiency with which the different statistics can detect deviations from the null hypothesis

$$H_0 : \alpha = 0$$

with α defined as in equation (21). This can be done by (i) assuming a particular parameter combination; (ii) generating N data elements in accordance with the mixture model (21); (iii) calculating each of the statistics; (iv) testing at (say) the 5 per cent level whether H_0 is accepted, and noting the result; (v) applying steps (ii)–(iv) many (typically a few tens of thousands) times and (vi) comparing the rejection rates of the different statistics.

It is interesting that the periodogram can be treated in the same way as the uniformity test statistics: note that, from equations (4) and (10),

$$I(\omega) \equiv Z(\omega),$$

i.e. each periodogram ordinate is just the Rayleigh statistic calculated for that particular frequency. Large values of $I(\omega)$ correspond to frequencies for which the ϵ_j in (10) are not uniformly distributed. The Rayleigh statistic is therefore added to the list of statistics for which power properties will be studied below. Significance levels are given by

$$\Pr(Z \geq x) \approx \exp(-x)$$

provided $N \gtrsim 50$ (e.g. Fisher 1993), i.e. under the null hypothesis of uniform ϵ_j the Rayleigh statistic is exponentially distributed with unit mean. It is noteworthy that Z has good power against unimodal alternatives, i.e. exactly the situation studied in this paper.

In general, the Degroote G and likelihood ratio Λ statistics greatly outperform the other statistics considered. It is not difficult to see why: G and Λ contrast the distribution of points in very narrow intervals (2δ and a few times σ, respectively) centred on 0.5, with the distribution over the remainder of the (0, 1) interval. By contrast, the precise location of deviations from uniformity remains unspecified by the other statistics. Next best are PR and SB, both devised to detect narrow-pulse deviations from uniformity. The Ψ_4 statistic is indeed superior to $\Psi_1 - \Psi_3$, and has power quite similar to that of the Rayleigh statistic. The KS statistic is slightly inferior to Ψ_4 and Z, while the AD has least power by some margin.

The statistics G and Λ, and to a lesser extent PR, really come into their own for small σ, i.e. if the concentration of points at (transformed) phase 0.5 is very narrow. This is easily seen for the Degroote statistic. As an illustration, consider a situation in which only 5 per cent of points are associated with the periodicity. Provided $\sigma \ll \delta \ll 1$, then the mean value of G will be $(0.05 + 0.95 \times 2\delta)N$; if e.g. $\delta = 0.005$, then G will be about six times the null hypothesis value of $0.02N$. As a corollary, provided σ is small, very small fractions of periodic points can give highly significant detections.

The advantage of Λ over G is, of course, that the effective bin size parameter (σ) in the former is estimated from the data, rather than having to be specified (δ). Of course, if there is extraneous information about the measurement errors, δ can be chosen accordingly, rather than having to rely on guesswork.

6 RECOMMENDED STATISTICS

It may seem obvious from the discussion in the preceding section that the test statistics of choice should be Degroote G and/or the

likelihood ratio statistic Λ. Both have drawbacks, though. An injudicious choice of the free parameter δ used in the calculation of G can lead to periodicities being missed – see e.g. Fig. 13. This can be avoided by evaluating G for a few different values of δ, although establishing significance levels is then not as straightforward as described in Section 7. Use of Λ, on the other hand, does not require specification of any parameters – the analogue of δ, namely σ in equation (21), is derived from the data. However, calculation of equation (22) is currently computationally expensive.

In practice, the frequency interval searched, and the frequency resolution, will determine the total computational load. If these are moderate, then Λ should be favoured. The statistic G could be calculated as a matter of course. The Rayleigh statistic, although not as powerful, is useful for graphically displaying the quality of the model fits, as will be demonstrated below.

7 SIGNIFICANCE LEVELS OF SPECTRA

Although the distributions of most of the statistics discussed in Section 3 are known, a spectrum consists of many evaluations of the chosen statistic, and these are not independent. Simulation is therefore the easiest way of determining the significance of the most extreme value in a spectrum. The recipe is the following:

(i) Determine the height of the highest spectrum peak (or the depth of the deepest minimum, in the case of the SB statistic).

(ii) Generate an artificial data set by randomly 'jittering' each of the time points. (In the implementations in the next section of the paper, points were shifted randomly in time by amounts which were distributed as zero-mean Gaussians with standard deviations of order unity.)

(iii) Calculate the spectrum of the data generated in (ii) and determine the maximum value over all frequencies (or minimum, for the SB statistic).

(iv) Repeat steps (ii) and (iii) many times (10 000 in the applications below).

(v) Determine the percentage of simulated values which are more extreme than the value from (i); this is the estimated significance level.

8 PREWHITENING AN IDENTIFIED PERIOD

Procedures described in this section of the paper are illustrated by analysing a simulated data set with four periodicities ($P_1 = 2.222$, $P_2 = 3.456$, $P_3 = 4.765\,43$ and $P_4 = 5.1$). The time baseline is $T = 3\,000$; only 6 per cent of the pulses in this interval are detected, giving $n_1 = 81$, $n_2 = 50$, $n_3 = 37$ and $n_4 = 32$ in the particular simulation analysed. The assumed Gaussian measurement error standard deviation is 0.004. 50 noise pulses are distributed randomly throughout the observation interval; these constitute 20 per cent of the 'observations'.

Perhaps the simplest prewhitening algorithm can be based on the periodogram. For a given frequency ω, the exponential term in equation (4) effectively assigns a phase angle in complex space to each observations. Maxima of $I(\omega)$ occur at those frequencies where many phase angles are aligned. This allows the particular observations contributing to a periodogram maximum to be easily identified, and hence to be removed from the data set.

The details of the procedure are illustrated by using the example data set. Fig. 3 shows the spectrum (4) of the data: it is dominated by the three peaks at frequencies 0.449 99, 0.9001 and 1.3501. Although the latter peak is marginally the highest, $\nu_1 = 0.4500$ is

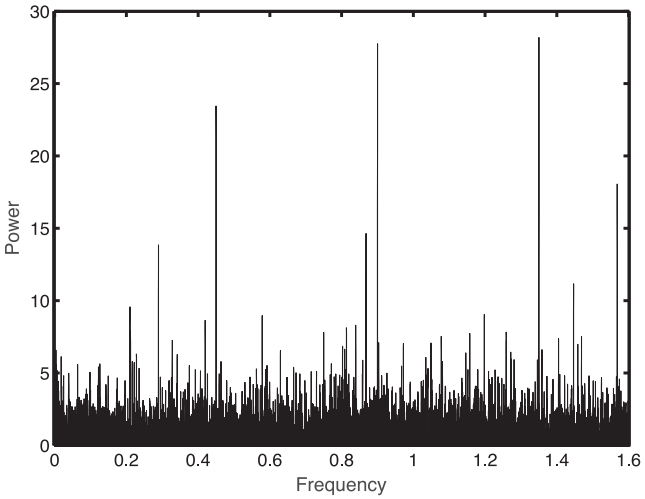

Figure 3. The periodogram of an example data set of 250 points. The 'signal' part of the simulated data set consists of four different periodic sequences, each only 6 per cent complete. 20 per cent of the data points are randomly distributed 'noise'.

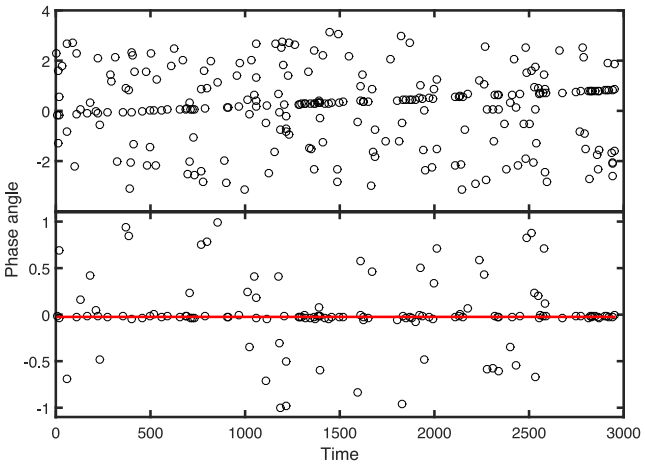

Figure 4. Top panel: the phase angles of each of the 250 data points, calculated with respect to a periodogram peak frequency, and plotted against the time of the 'observation' Note the overdensity of points, consisting primarily of points belonging to the periodicity responsible for the periodogram peak. Bottom panel: as in the top panel, but after adjusting the frequency used to calculate the phase angles in such a way that a robust regression fit (red line) has a zero slope.

selected in accordance with remark (xiii) of Section 4. The top panel of Fig. 4 displays the phase angles of each of the data points. There is a slight slope to the overdensity of points with phase angles near zero; this indicates that the correct period is slightly different from $1/0.459\,99 = 2.2223$ (see Appendix A).

A refinement of the period can be obtained by adjustment of the frequency so that the slope of the overdensity of points is zero. This can be accomplished by fitting a straight line to the phase angle versus timing plot, and changing ν until the slope is zero. This task is aided by two devices: first, data points with phase angles far from the angles making up the overdensity are excluded from the regression. [In the bottom panel of Fig. 4 only points with angles in the range $(-1, 1)$ are taken into account.] Secondly, in order to downweight the influence of remaining points not overtly associated with the overdensity, robust regression is used. The line

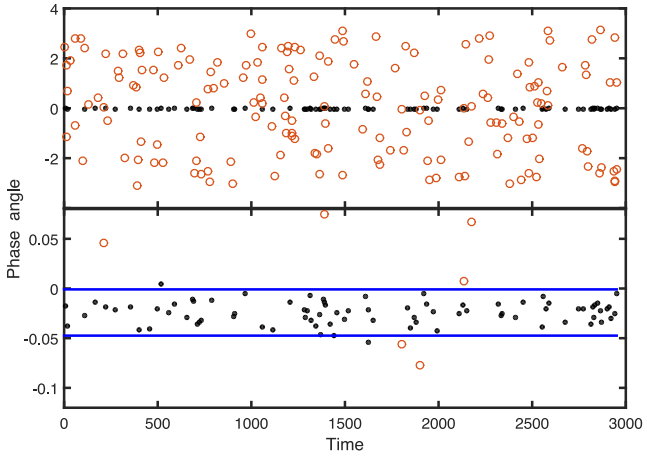

Figure 5. The phase angles of each of the 250 data points, calculated with respect to the frequency determined as in the bottom panel of Fig. 4. Filled dots are points belonging to the periodic sequence with frequency close to that used to calculate phase angles; open circles denote the remainder of the points. The detail in the lower panel also shows the $\pm 2\hat{\sigma}$ area associated with timing measurement errors.

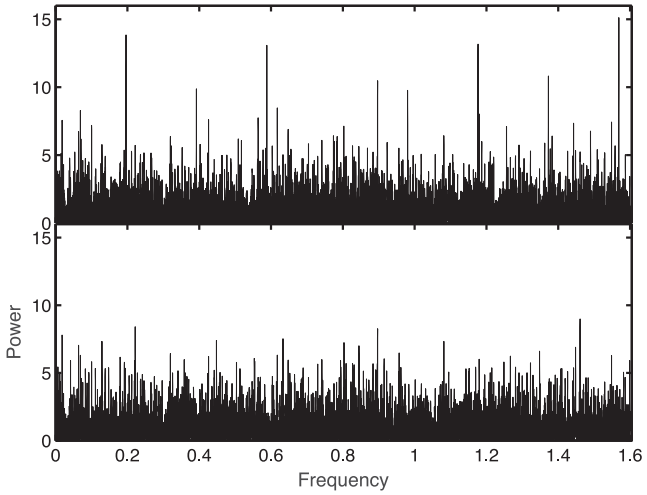

Figure 6. Top panel: the periodogram of the 89 points left after prewhitening three periods from the simulated example data set. Bottom panel: the periodogram of the 58 points left after prewhitening four periods from the data.

plotted in the bottom panel of Fig. 4 was calculated using iteratively reweighted LS, with a bisquare weighting function – it corresponds to the improved frequency 0.450 045 ($P = 2.222\,001$).

The slope of fitted lines will also be close to zero for frequencies ν far from ν_1 [i.e. phase angles uniformly distributed over $(-\pi, \pi)$]. When refining ν_1, this can be guarded against by restricting the search to a relatively narrow interval around the value ν_1 at which the periodogram reaches a maximum. Furthermore, it is expected that the residual scatter will be larger in such cases, hence it can also be required that the scatter remains within a specified bound during the frequency refinement.

Phase angles calculated using the refined frequency are plotted in the top panel of Fig. 5, with an expanded view of the central portion in the bottom panel. There is clearly a very good separation between points with $P_1 = 2.222$ (solid dots) and the remainder (open circles). In order to decide exactly which data to prewhiten, a mixture model is fitted to the data in the restricted phase angle interval $(-0.5, +0.5)$. It is assumed that the distribution of phase angles has two components: a Gaussian with mean μ and variance σ^2, and a uniform distribution over $(-0.5, 0.5)$. Assuming that a fraction α of angles belongs to the former distribution, the full probability density function of phase angles is

$$f(x) = \frac{\alpha}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] + (1-\alpha). \tag{30}$$

Estimates of μ, σ and α can be found by numerical maximization of the log likelihood function

$$\mathcal{L} = \sum_j \log f(x_j). \tag{31}$$

[Note the close similarity to the mixture model (21) used to derive the Λ statistic of Section 3; in the case of equation (30) though, the mean is not known, and must be estimated.]

In the present case, $\hat{\mu} = -0.024$, $\hat{\sigma} = 0.012$ and $\hat{\alpha} = 0.76$ are obtained. The parallel lines in the bottom panel of Fig. 5 enclose those angles in the interval $(\hat{\mu} - 2\hat{\sigma}, \hat{\mu} + 2\hat{\sigma})$: it includes 79 of the 81 points with $P = P_1$, and no points from either the other periodicities, or the 50 noise points.

The 79 points can now be removed from the data set, and the analysis repeated afresh. The largest periodogram peak for the remaining points is at $\nu = 1.4468$ ($P = 0.6912$), with the second and third ranked peaks at frequencies 0.2894 and 0.8681, corresponding to periods 3.457 and 1.152. The ratio of the three frequencies is very close to 1:3:5, so that the lowest is adopted as the correct 'alias'. This is refined to 0.289 352 ($P = 3.456\,00$). In this instance, $\hat{\sigma} = 0.006$, and the interval $(\hat{\mu} - 2\hat{\sigma}, \hat{\mu} + 2\hat{\sigma})$ contains 48 of the 50 phase angles belonging to the series with $P = P_2 = 3.4560$.

The largest periodogram peak of the remaining 123 data points is at $\nu = 1.5686$ (corresponding to $P_4/8$), but the pattern of peaks is otherwise dominated by multiples of $\nu = 0.2098$ ($P = 4.7664$), so this is the frequency focused on. The refined value is $\nu = 0.209\,8444$ ($P = 4.765436$); $\hat{\sigma} = 0.0052$, with an estimated 71 per cent of the phase angles in the interval $(-1, 1)$ belonging to the Gaussian distribution (as opposed to the uniform distribution). Of the 37 data elements forming the sequence with $P_3 = 4.765\,43$, 34 lie within $2\hat{\sigma}$ of $\hat{\mu}$, and are prewhitened from the data set.

The periodogram of the residuals is dominated by peaks at multiples of $\nu = 0.196\,09$ ($P = 5.0996$), with the largest at $\nu = 1.5686 \approx 8 \times 0.1961$ (top panel of Fig. 6). The peak height is significant at the 0.3 per cent level, as established from 10 000 simulations, as outlined in Section 7. The improved frequency of 0.196 0788 corresponds to a period $P = 5.099\,99$, which can be compared to $P_5 = 5.100\,00$. Of the 32 points associated with this periodicity, 31 are in the interval $\hat{\mu} \pm 2\hat{\sigma} = -0.016 \pm 2 \times 0.0055$. The periodogram of the 50 noise pulses, and the 8 misidentified data points from the four periodic sequences, is plotted in the bottom panel of Fig. 6. There are no large peaks in the latter plot, and the large reduction in power compared to the upper panel of the figure is noteworthy. The significance level of the largest peak is 75 per cent (from 10 000 simulations of 58 points of similar time spacing).

Clearly, superior results would have been obtained by using a slightly more liberal interval for the selection of points to be prewhitened, e.g. $\hat{\mu} \pm 2.5\hat{\sigma}$. Another, more general lesson learned is that periodograms need to be carefully searched for the lowest frequency in a sequence of peaks induced by the same periodicity. The latter point will be particularly important if very many data sets

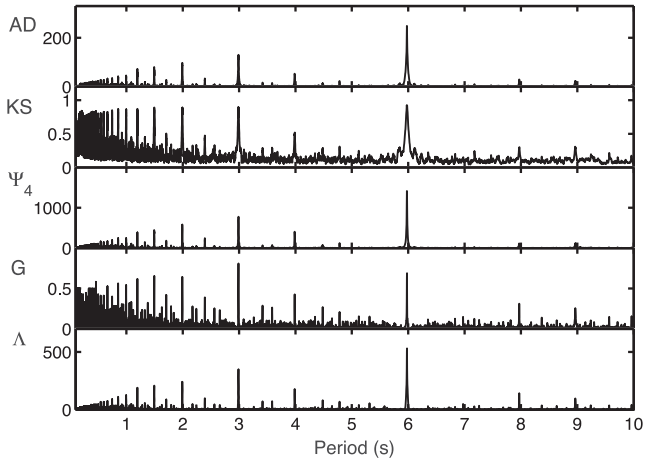

Figure 7. Uniform distribution test statistics, over a range of trial folding periods, for timing for the RRAT PSR J0847−4316. From top to bottom: AD, KS, Neyman (fourth order), Degroote and the likelihood ratio statistic (22). Statistics were calculated for the intervals $\Delta_1 Y$ between detection times, including only those intervals up to $n = 100$ times the trial period [see point (x) in Section 4].

are to be processed, i.e. if operation of the algorithm will not be supervised.

A final remark: if a spurious pulse does arrive at a time very close to that associated with a particular periodicity, then within the framework of the point processes it will be impossible to tell that it does not belong to that periodicity.

9 APPLICATIONS TO RRATS DATA

Fig. 7 demonstrates the application of five of the test statistics discussed in Section 3, to observations of the RRAT PSR J0847−4316. The number of detections of the object was 151, spanning 6.8 years. With such a long time baseline the frequency resolution is $\sim$4.05 $\times$ 10^{-4} d^{-1} (4.6 $\times$ 10^{-9} Hz). In order to cover the period range of 0.1–10 s, i.e. frequency range 8640–864 000 d^{-1}, at an adequate resolution, the data would need to be folded with respect to at least $\sim 2 \times 10^9$ trial periods, and preferably a factor of 10 or so more (i.e. $\delta r_{\max} < 0.1$ in equation 29). This computing time demand is easily sidestepped by initially working only with the differenced data $\Delta_1 Y$, and for any given trial period P using only those intervals $\Delta_1 Y \leq 100P$: for $\delta r_{\max} = 0.04$, only 11 520 trial periods are required.

The only free parameter associated with any of the statistics in Fig. 7 is the width 2δ of the interval used to calculate the Degroote statistic G: the value $\delta = 0.01$ was used throughout the results which follow. Inspection of the figure very clearly supports the presence of an $\sim$6 s periodicity in the data, The statistic Λ is a maximum at $P_* = 5.978\,66$ s; other parameters of interest are $\sigma = 0.018$, $\alpha = 0.71$ (i.e. 71 per cent of the intervals are associated with the period P_*). The test statistics can now be calculated at a much finer period resolution, centred on P_*. This is done in two rounds: first, the statistics are calculated for 1001 trail periods, at a spacing of $\delta P = 6.8 \times 10^{-10}$ d (59 μs). This gives $P_* = 5.977\,4816$ s, $\sigma = 0.0024$, $\alpha = 0.78$. Decreasing δP by a further factor of 40 leads to closely similar results – $P_* = 5.977\,4890$ s, $\sigma = 0.0025$ and $\alpha = 0.81$. The statistics as calculated for the second refinement are plotted in Fig. 8. Interestingly, the more conventional AD and KS statistics show local *minima* near the periods where the other three statistics reach maxima. The superior resolution of the Λ statistic is evident.

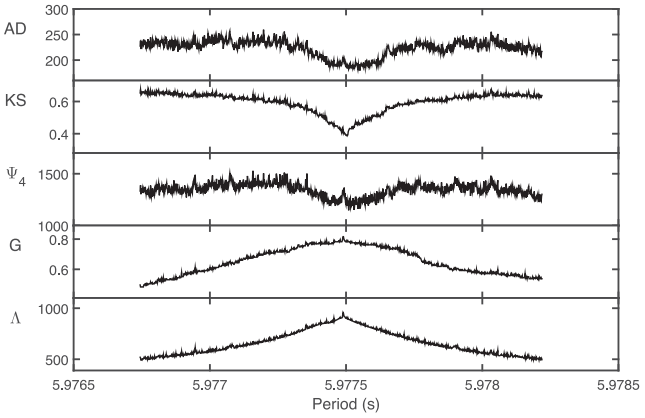

Figure 8. As in Fig. 7, but focusing on a small period range around the 'best' period extracted from Fig. 7, and including all intervals $\Delta_1 Y$ between successive detections.

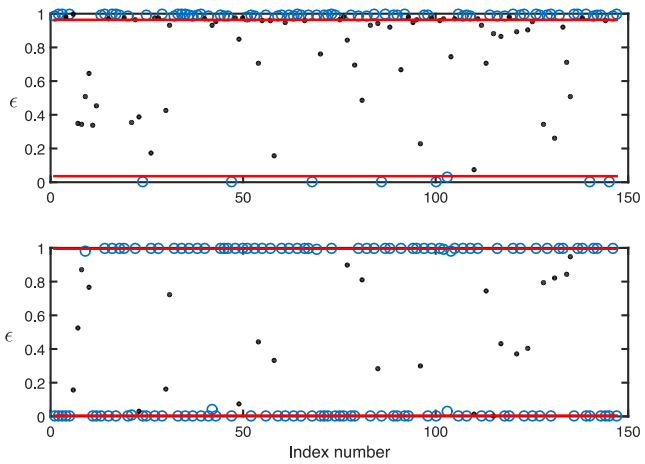

Figure 9. The scaled residuals $\epsilon = \text{modulus}(\Delta_1 Y, P_*)/P_*$ for the intervals $\Delta_1 Y$ and optimal periods P_* corresponding to, respectively, Fig. 7 (top panel) and Fig. 8 (bottom panel). Open circles denote residuals corresponding to $\Delta_1 Y \leq 100 P_*$ (top panel) or $\Delta_1 Y \leq 10\,000 P_*$ (bottom panel); dots denote the residuals corresponding to larger intervals $\Delta_1 Y$. Estimated 2σ offsets from $\epsilon = 0$ and $\epsilon = 1$ are also indicated (in red).

The period spacing $\delta P = 1.5$ μs used to produce Fig. 8 corresponds to the best resolution $P_*^2/\max(\Delta_1 Y)$ possible with the intervals $\Delta_1 Y$. This may compared with the resolution $P_*^2 \delta\nu = 5.977^2 \times 4.6 \times 10^{-9} = 0.16$ μs allowed by the full time span of the observations. The logical next step would be to calculate 'spectra' using intervals such as $\Delta_3 Y$, or the actual times pulse arrival times. Examination of plots of the scaled residuals,

$$\epsilon = \text{modulus}(\Delta_1 Y, P_*)/P_*,$$

reveals that this will not be useful for this data set. The top panel of Fig. 9 is based on the optimal period P_* extracted from Fig. 7. Open circles denote intervals $\Delta_1 Y$ which are shorter than $100 P_*$; these all lie within 2σ (as indicated by the red lines) of zero or unity; the dots correspond to intervals $\Delta_1 Y > 100 P_*$, and these scatter more or less randomly over [0, 1]. This is perfectly consistent, since only those $\Delta_1 Y \leq 100 P_*$ were taken into account in determining $P_* = 5.978\,66$ s. The bottom panel, on the other hand, is based on *all* the $\Delta_1 Y$. In this case, the open circles are those data with $\Delta_1 Y \leq 100\,000 P_*$; with three marginal exceptions, these all lie within the very narrow 2σ intervals from zero or unity. Dots denote

residuals corresponding to the longer intervals $\Delta_1 Y > 100\,000 P_*$; and these again scatter randomly over the unit interval.

The implication of the bottom panel of Fig. 9 is that only the shorter intervals between successive observations are well described by P_*, while the longer intervals are not. This is a clear sign that the period is not constant, and that its evolution is proportional to time. The rate of period change is small enough to almost undetectable over 10^5 or fewer cycles, but large enough to manifest over greater numbers of cycles.

Linear period change models have in fact been fitted by Keane et al. (2011) to this data set, and to all but one of the other data sets analysed by Palliyaguru et al. (2011). Rather than equation (1), the appropriate model for the event times is then

$$Y_j \approx Y_0 + N_j P_0 + \frac{1}{2} N_j (N_j + 1) P_0 \dot{P} + e_j \tag{32}$$

(e.g. Koen 1996), where P_0 is the period at time Y_0, and higher order terms in the period derivative $\dot{P}$ have been neglected.

A point of interest is the number of cycles 'lost' during the span of the observations, due to the lengthening period. The total number of pulse periods of ∼5.9775 s covered was $\sim 3.6 \times 10^7$, and Keane et al. (2011) find that $\dot{P} = 119.94 \times 10^{-15}$. Therefore, the RRAT pulsed

$$\frac{1}{2} N_j (N_j + 1) P_0 \dot{P} \sim 460$$

fewer times than if the period were fixed at P_0. On the other hand, over 100 000 cycles only 0.0036 of a cycle is 'lost', which explains the appearance of the lower panel in Fig. 9.

It is useful to be able to avoid explicit determination of the N_j. This can be done by using the approximation

$$N_j \approx (Y_j - Y_0)/P_0$$

in the term in equation (32) which contains the period derivative. The error incurred is minimal, since

$$\frac{(Y_j - Y_0)/P_0}{N_j} = 1 + \frac{1}{2}(N_j + 1)\dot{P} \equiv 1 + x$$

and $x \lesssim 2.2 \times 10^{-6}$. This means that

$$Y_j \approx Y_0 + N_j P_0 + \frac{1}{2} \frac{(Y_j - Y_0)^2}{P_0} \dot{P} \tag{33}$$

to excellent approximation.

It is possible to use equation (33) to 'correct' the observed values of Y_j for the presence of the period drift:

$$Y'_j \equiv Y_j - \frac{1}{2} \frac{(Y_j - Y_0)^2}{P_0} \dot{P} \approx Y_0 + N_j P_0. \tag{34}$$

The known value of $\dot{P}$ taken from Keane et al.(2011) will now be used in equation (34) to further demonstrate the application of the theory of this paper to the PSR J0847−4316 data.

Fig. 10 shows the periodograms (4) of, respectively, the raw Y_j and adjusted Y'_j, calculated in a grid of ∼600 000 trial periods. It is noteworthy that the presence of the period trend causes the periodicity to be inconspicuous in the raw data. The corresponding scaled residuals can be seen in Fig. 11. The top panel is based on the original Y_j, folded with respect to the optimal period from the top panel in Fig. 10. Residuals in the bottom panel of Fig. 11 were calculated from the Y'_j, using the optimal period from the bottom of Fig. 10. Note that *all* the residuals in the bottom panel cluster around either zero or unity; this confirms that a single period provides a good description of the data, albeit with non-zero period derivative.

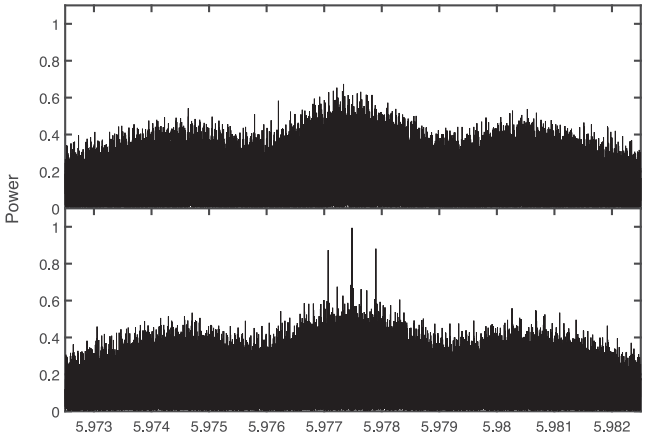

Figure 10. Periodograms of the original PSR J0847−4316 observations (top panel) and of the data adjusted for the changing period (bottom panel).

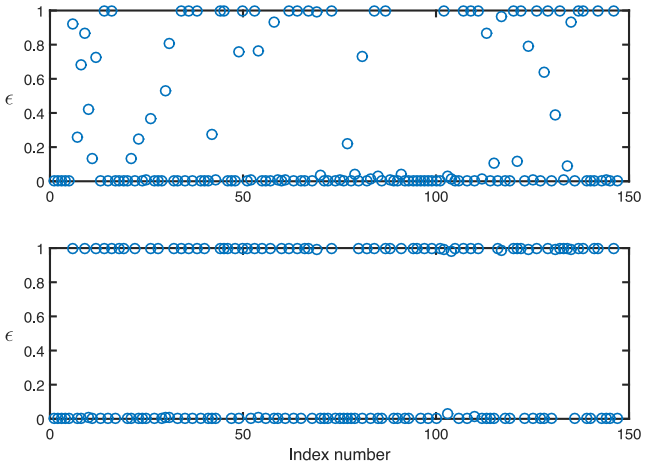

Figure 11. Top panel: the residuals of the original PSR J0847−4316 observations, obtained by folding the data with respect to the 'best' period in the top panel of Fig. 10. Bottom panel: as for the top panel, but residuals were obtained from the adjusted data (as defined in equation 34), folded with respect to the optimal period in the bottom panel of Fig. 10.

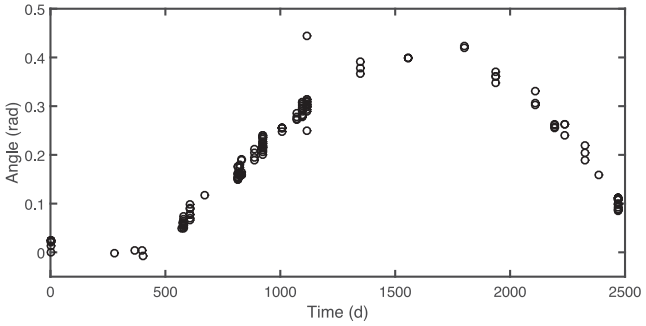

Figure 12. The phase angles, in the complex plane, of the adjusted PSR J0847−4316 event times, plotted against time.

The phase angles of the adjusted observations are plotted in Fig. 12: the shape of the curve traced over time suggests that the rate of period change itself changed abruptly roughly 500 d after monitoring commenced. The presence of an outlying point near $Y - Y_0 \sim 1100$ d is also noted in passing.

As mentioned above, Keane et al. (2011) found linear period changes in data for all the RRATs observed by Palliyaguru et al.

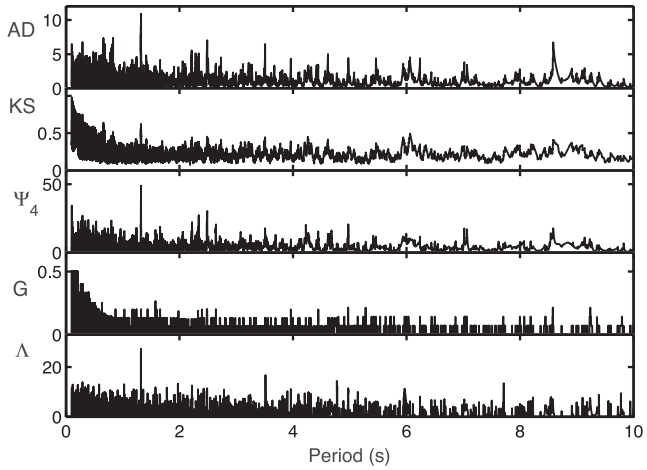

Figure 13. Uniform distribution test statistics, over a range of trial folding periods, for photon arrival times of RRAT J1754−30. From top to bottom: AD, KS, Neyman (fourth order), Degroote and the likelihood ratio statistic (22). Statistics were calculated for the intervals $\Delta_1 Y$ between detection times, including only those intervals up to $n = 200$ times the trial period [see point (x) in Section 4].

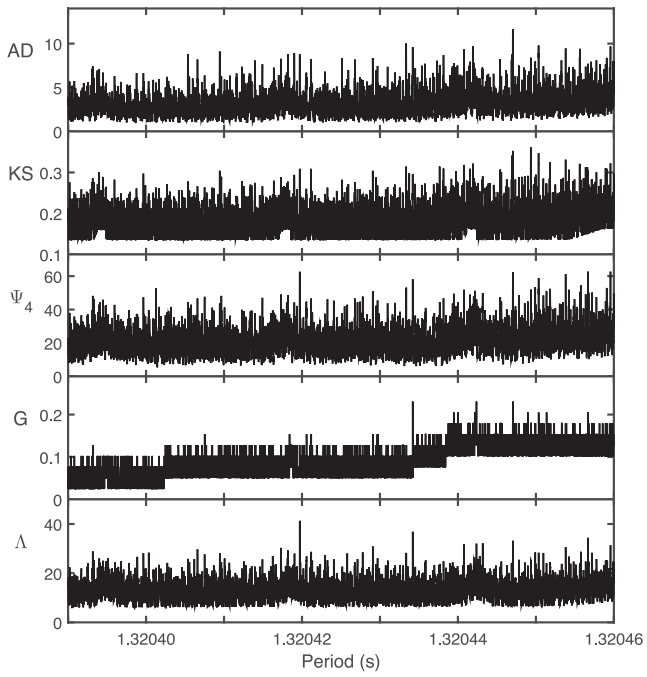

Figure 14. As in Fig. 13, but focusing on a small period range around the 'best' period extracted from Fig. 13, and including all intervals $\Delta_1 Y$ between successive detections.

(2011), with a single exception. Keane et al. (2011) could find no entirely satisfactory model for event times of PSR J1754−30. There are only 40 pulse arrival times for this RRAT, spanning 6.5 years. The period resolution is therefore similar to that of PSR J0847−4316. The uniformity test statistics for the intervals $\Delta_1 Y \leq 200P$ are plotted in Fig. 13. The best period, according to the Λ statistic, is 1.320 43 s; other parameters are $\sigma = 0.029$ and $\alpha = 0.19$.

Fig. 14 shows the results of calculating the statistics over a small period interval, centred on the best period from Fig. 13. All 39 intervals $\Delta_1 Y$ between successive events were taken into account,

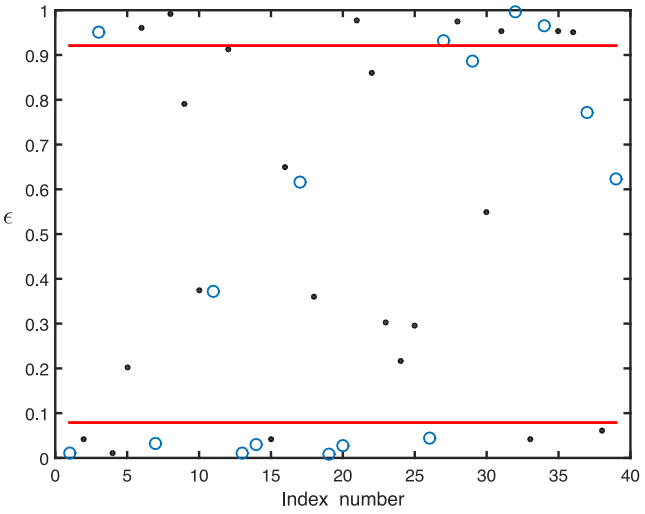

Figure 15. The scaled residuals $\epsilon = \text{modulus}(\Delta_1 Y, P_*)/P_*$ for the intervals $\Delta_1 Y$ and optimal periods P_* corresponding to Fig. 14. Open circles denote residuals corresponding to $\Delta_1 Y \leq 1000 P_*$; dots denote the residuals corresponding to larger intervals $\Delta_1 Y$. Estimated 2σ offsets from $\epsilon = 0$ and $\epsilon = 1$ are also indicated (in red).

and a very fine period grid $\Delta P = 8 \times 10^{-10}$ s was used. The Λ statistic reaches a maximum at $P_* = 1.320\,4197$ s, with $\sigma = 0.040$ and $\alpha = 0.53$.

Two points are perhaps worth remarking on. First, comparison of Figs 8 and 14 gives the impression that the period resolution in the latter is considerably better. This is only a partial truth – period resolution is proportional to P^2 and inversely proportional to $\max(\Delta_1 Y)$, which, with the differences in P_*, gives $\delta P(\text{J1754})/\delta P(\text{J0847}) \approx 0.03$. In reality, features in Fig. 8 are smeared because of the presence of the systematic period change. The true resolution inherent in the PSR J0847−4316 data is more clearly demonstrated by the bottom panel of Fig. 10.

The second point is the disappointing performance of the Degroote statistic in Fig. 14. This can be ascribed to the large value $\sigma = 0.040$, as compared to the default value $\delta = 0.01$ – see Fig. 15. Fig. 15 also demonstrates the uncomfortable fact that for some relatively short intervals, $\Delta_1 Y$ is *not* very close to an integer multiple of P_*: for these short intervals this cannot be ascribed to secular period changes, but must be due to shorter time-scale effects.

It is also noteworthy that the two shortest intervals between successive events are 2.3386 and 4.7743 s; since the first three multiples of P_* are 1.3204, 2.6408, 3.9612 and 5.2816 s, clearly these two intervals either involve spurious events or events not associated with the primary periodicity.

In principle,periods can be much more accurately estimated directly from the sequence of time points Y_j than from intervals. However, as pointed out above, the long time baseline implies a very fine period resolution, which implies that the test statistics need to be calculated for a great many trial periods. This is currently not feasible for all the statistics; in what follows, only the periodogram is used.

The periodogram of the 40 observations of PSR J1754−30 can be seen in Fig. 16. The highest peak is at $P_* = 1.321\,956\,5077$ s, with 1.319 4442 s a close second. The scatter is substantial though, as illustrated in Fig. 17. Transforming the residuals as in equation (20) gives a standard deviation $\sigma_\xi = 0.153$. The isolated point at index number 17 is associated with the minimum interval 2.3386 s

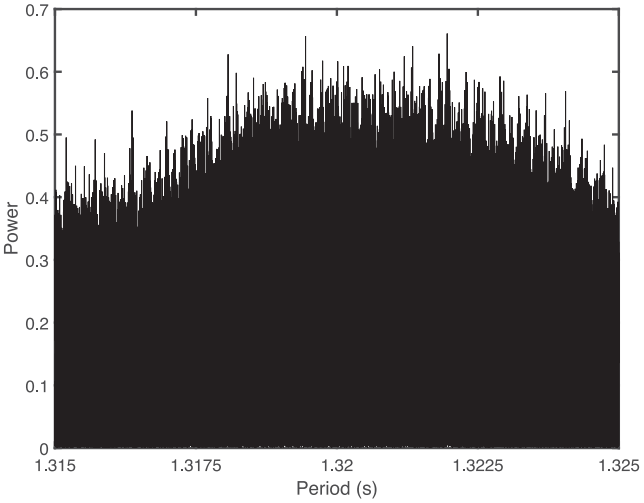

Figure 16. The periodogram of the J1754−30 pulse times, in the vicinity of 1.32 s.

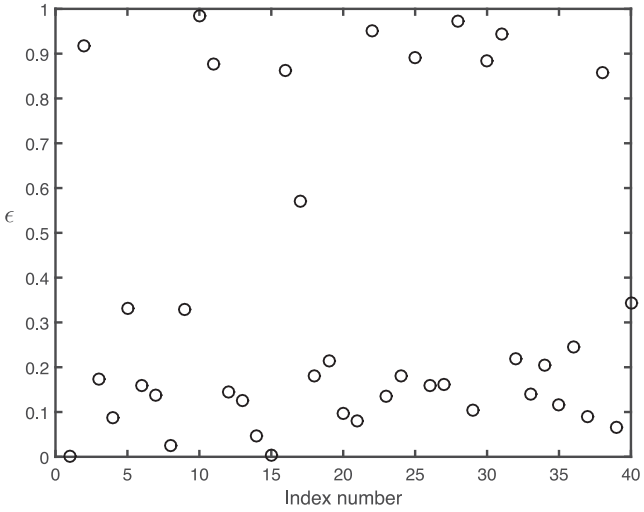

Figure 17. The scaled residuals obtained when folding the J1754−30 timings with respect to the optimal period $P_* = 1.321\,9565$ s (from Fig. 16).

referred to above. Leaving this out does not affect P_*, but reduces σ_ξ to 0.131.

Examining subsets of the data is interesting. The first nine events, covering 835 d, are well described by a period of 1.320 82 s ($\sigma_\xi = 0.024$); events 12–23 (excluding the anomalous observations 17) cover 523 d, and have $P_* = 1.321\,09$ ($\sigma_\xi = 0.025$); pulses 24–35 cover 104 d, with $P_* = 1.319\,54$ s and $\sigma_\xi = 0.035$; pulses 30–40 (excluding an anomalous event 38) cover 769 d, and can be modelled by $P_* = 1.318\,57$ s, $\sigma_\xi = 0.034$.

The results in the preceding paragraph suggest that all but perhaps four of the pulses belong to a process with a period ∼1.32 s, but that the period may not be perfectly constant. Fitting a definitive model to the PSR J1754−30 pulse data is outside the scope of this paper.

10 AN INTERESTING ASIDE

The paper is concluded by demonstrating the ease with which RRAT-type data can be checked for the presence of a planet-sized body orbiting the neutron star, provided the light travel time Δt of the pulsar emission is smaller than its pulse period. The latter requirement can be written as

$$\Delta t = a_1 \sin i/c = \alpha P, \tag{35}$$

where a_1 is the distance from the pulsar to the centre of mass of the pulsar–planet system, i is the orbital inclination, P is the pulsar period and $\alpha < 1$. Use will be made of the phase angles defined by the periodogram (4), evaluated in the 'best period' P.

Kepler's law is

$$P_{\rm O} = \frac{(a_1 + a_2)^3}{M_1 + M_2} \approx a_2^3/M_1, \tag{36}$$

where subscripts 1 and 2, respectively, refer to the pulsar and the small body, and $P_{\rm O}$ is the orbital period. Furthermore,

$$m_1 a_1 = m_2 a_2. \tag{37}$$

Equations (35)– (37) can be combined to find

$$M_2 \sin i \approx c M_1^{2/3} P_{\rm O}^{-2/3} \Delta t. \tag{38}$$

Some of the units in equation (38) are a little peculiar: Δt and $P_{\rm O}$ are measured in years, and c in au/yr. The form

$$M_2 \sin i \approx 0.102 M_1^{2/3} P_{\rm O}^{-2/3} \Delta t \tag{39}$$

with M_1 and M_2 in $\mathrm{M}_\odot$, $P_{\rm O}$ in d, and Δt in s, is more useful.

A good estimate of the most likely values of $P_{\rm O}$ can be obtained from the phase angles

$$\psi_j = \tan^{-1}\left\{\tan\left[\frac{2\pi(t_j - t_1)}{P}\right]\right\},$$

where the inverse tan function returns values in the interval $(-\pi, \pi)$. The conventional periodogram

$$I_{\rm C}(\omega) = \frac{1}{N}\left|\sum_{j=1}^{N}(\psi_j - \overline{\psi})\mathrm{e}^{-\mathrm{i}\omega(t_j - t_1)}\right|^2 \tag{40}$$

of the phase angles of PSR J1754−30 is plotted in the top panel of Fig. 18. The period $P = 1.321\,956\,5077$ s, from Section 9, has been used, and the suspect observation 17 has been omitted. The maximum value of $I_{\rm C}$ is reached in a frequency of 0.5064 d^{-1}, i.e. $P_{\rm O} = 1.975$ d. Given the sparsity of the 39 values of ψ over the observation interval, there is, of course, dreadful spectral leakage – prewhitening $P_{\rm O}$ from the phase angles leads to the residual periodogram in the middle panel of Fig. 18. It is clear that $P_{\rm O}$ can account very well for the power in the top panel periodogram.

A sinusoid with period $P_{\rm O}$ fitted to the ψ_j has an LS-estimated amplitude of 0.68. The interval $[-\pi, \pi]$ represents one full period P; therefore, $\alpha = 0.68/2\pi = 0.11$. Substituting $P_{\rm O} = 1.975$ d, $P = 1.322$ s, $M_1 = 1.5\,\mathrm{M}_\odot$ and $\Delta t = \alpha P = 0.145$ s into equation (39), $M_2 \sin i \approx 0.012\,\mathrm{M}_\odot$ is obtained.

The phase angles are shown phased with respect to $P_{\rm O}$ in the bottom panel of Fig. 18. Although this looks quite convincing the periodogram peak in the top panel of Fig. 18 should be evaluated formally. Significance levels are easily ascertained, under the null hypothesis H_0 that there is no time dependence in the ψ_j:

(i) Permute the phase angles – under the null hypothesis, this gives a statistically equivalent data set.

(ii) Calculate the periodogram of the permuted phase angles, and note the height of the tallest peak.

(iii) Repeat steps (i)–(ii) many times, for different permutations.

(iv) The collection of periodogram maxima represents the distribution of largest peak values. Comparison of the actual peak value to the distribution gives the significance level.

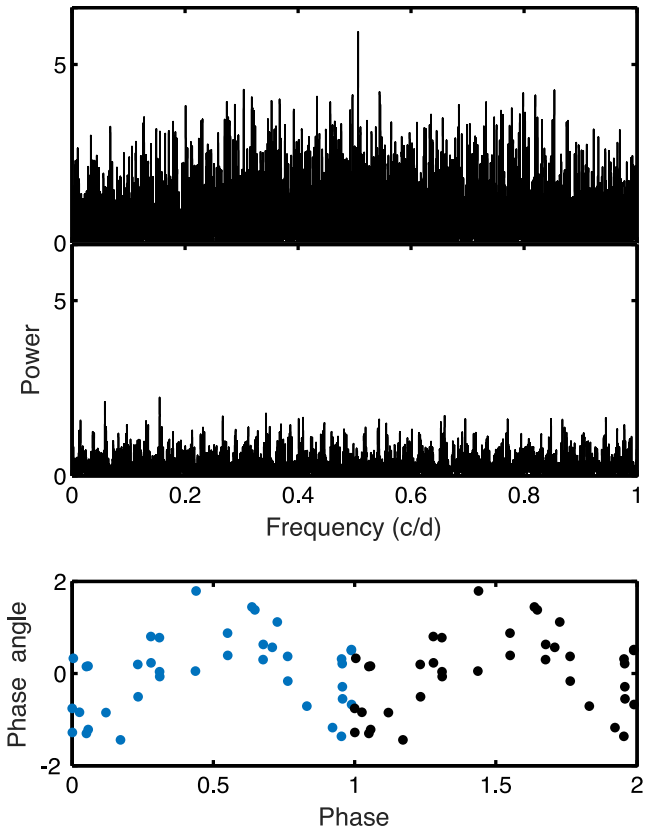

Figure 18. Top panel: the standard periodogram (see equation 40) of the PSR J1754−30 phase angles. Middle panel: the periodogram of the residuals after prewhitening a frequency of 0.5064 d^{-1} from the phase angles. Bottom panel: the phase angles phased with respect to the 0.5064 d^{-1} frequency.

In the case of the PSR J1754−30 data the maximum in the top panel of Fig. 18 is 5.92; 14.7 per cent of permuted data showed periodogram maxima in excess of this value, i.e. the peak in Fig. 18 is not significant.

ACKNOWLEDGEMENTS

The author is grateful to Dr Nipuni Palliyaguru for sharing the RRAT observations discussed in this paper, and to the South African National Research Foundation for a grant.

REFERENCES

Ansari A., Zhang D., Mahlke S., 2009, Proc. 7th IEEE Symp. on Application Specific Processors. IEEE, Piscataway, NJ, p. 88
Casey S. D., Sadler B. M., 1996, IEEE Trans. Signal Process., 44, 2260
Chennamangalam J., Lorimer D. R., 2014, MNRAS, 440, L86
Clarkson I. V. L., 2008, IEEE Trans. Signal Process., 56, 1779
D'Agostino R. B., Stephens M. A., 1986, Goodness-of-Fit Techniques. Marcel Dekker, Inc., New York
Degroote P. et al., 2009, A&A, 506, 471
De Jager O. C., Swanepoel J. W. H., Raubenheimer B. C., 1989, A&A, 221, 180
Fisher N. I., 1993, Statistical Analysis of Circular Data. Cambridge Univ. Press, Cambridge
Fogel E., Gavish M., 1988, Proc. Int. Conf. Acoust., Speech, Signal Process., 4, 2348
Kawaler S., 1988, in Christensen-Dalsgaard J., Frandsen S., eds, Proc. IAU Symp. 123, Advances in Helio- and Asteroseismology. Reidel, Dordrecht, p. 329
Keane E. F., Kramer M., Lyne A. G., Stappers B. W., McLaughlin M. A., 2011, MNRAS, 415, 3065
Koen C., 1996, MNRAS, 283, 471
Kuiper N. H., 1960, Ned. Akad. Wet. A, 63, 38
Le Bot O., Bonnel J., Mars J. I., Gervaise C., 2013, J. Acoust. Soc. Am., 133, 3312
Neyman J., 1937, Skand. Aktuarietidskr., 20, 150
Nishiguchi K., Kobayashi M., 2000, IEEE Trans. Aerosp. Electron. Syst., 36, 407
Orsi R. J., Moore J. B., Mahoney R. E., 1999, IEEE Trans. Signal Process., 47, 1646
Palliyaguru N. T. et al., 2011, MNRAS, 417, 1871
Paltani S., 2004, A&A, 420, 789
Protheroe R. J., 1985, Proc. 19th Int. Cosmic Ray Conf., Program Sessions, Vol. 3. Washington, DC, p. 485
Provencal J. et al., 2012, ApJ, 751, 91
Rahman M., Pearson L. M., Heien H. C., 2006, Bull. Malays. Math. Sci. Soc., 29, 11
Rayner G. D., Rayner J. C. W., 2001, J. Appl. Math. Dec. Sci., 5, 181
Reed M. et al., 2010, MNRAS, 414, 2885
Sadler B. M., Casey S. D., 1998, IEEE Trans. Signal Process., 46, 2990
Sidiropoulos N. D., Swami A., Sadler B. M., 2005, IEEE Trans. Signal. Process., 53, 733
Silvotti R., Dreizler S., Handler G., Jiang X. J., 1999, A&A, 342, 745
Swanepoel J. W. H., de Beer C. F., 1990, ApJ, 350, 754
Winget D. E. et al., 1991, ApJ, 378, 326

APPENDIX A: PERIODOGRAM PHASE ANGLES

Let P_* be one of the periods present in the data. Only for those observations, t_j corresponding to the period P_* will

$$t_j = m_j P_* + \phi_* + e_j,$$

where the m_j are integers, ϕ_* is a fixed phase and e_j is a small zero-mean 'noise' component. The periodogram is calculated from a sum over terms $\exp(\mathrm{i}\omega t_k)$; for $\omega = 2\pi/P_*$, terms corresponding to P_* reduce to

$$\exp(\mathrm{i}\omega t_j) = \exp[2\pi\mathrm{i}(\phi_* + e_j)/P_*].$$

It follows that the angles in the complex plane associated with these measurements scatter around $2\pi\phi_*/P_*$, whereas for time points *not* associated with P_*, phase angles will be uniformly spread over $(0, 2\pi)$.

For slightly misspecified trial periods $P_\mathrm{t} = P_* + \delta$,

$$t_j/P_\mathrm{t} \approx m_j(1 - \delta/P_*) + \phi_*/P_\mathrm{t} + e_j/P_\mathrm{t} \equiv m_j(1 - \delta/P_*) + \phi' + e'_j$$

and hence

$$\exp(\mathrm{i}2\pi t_j/P_\mathrm{t}) \approx \exp[\mathrm{i}2\pi(-m_j\delta/P_* + \phi' + e'_j)].$$

The phase angle corresponding to t_j is

$$\psi_j = 2\pi(\phi' + e'_j) - 2\pi\delta m_j/P_*, \qquad \text{(A1)}$$

i.e. the angles are linear functions of m_j, or approximately linear functions of t_j. This point is illustrated in Fig. 4.

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