The most sensitive period search we can do on event data

G. Bélanger, ESA, ESAC

Presented on 2019 May 22 (Last updated: May 21, 2019)

This talk was presented on Wednesday May 30, 2019, at the 14th workshop of the IACHEC held from May 19 to 23 in Shonan Village, Japan. The subject is that of period searches on event data, and introduces two new sensitive periodogram statistics that can be used for this purpose: $\mathcal{R}_k^2$—the generalised modified Rayleigh statistic, and $\mathcal{Z}^2$—the modified Z^2 statistic. In addition to being the periodogram of choice for weak periodic signals, we show that the $\mathcal{R}_k^2$ can be used for highly refined pulse profile analysis. The details are presented in Belanger [2016].

Contents

A first look at the data 2

Periodogram analysis 3

Fast Fourier Transform periodogram 3

The Rayleigh periodogram 4

The modified Rayleigh periodogram 5

The new $\mathcal{R}_k^2$ and $\mathcal{Z}^2$ periodograms 5

Harmonic decomposition of peaked pulse profiles 5

Closing remarks 6

A first look at the data

YOU HAVE A data set made up of a collection of discretely detected events, each with a precisely measured time of arrival. These can be ultra high energy neutrinos, they can be X-rays from a blazar, or they can be cars driving over a pressure-sensitive line set on the ground across a street of a busy intersection downtown Tokyo. You are interested in how these events are distributed in time. What can you do with these measurements?

First, you'd probably want to look at them on a timeline just to get a general feeling of what they look like. So you make a time series by grouping the events in time bins that are wide enough to reveal the underlying structure , but not too wide as to hide potentially interesting features. And the time series could look something like this:

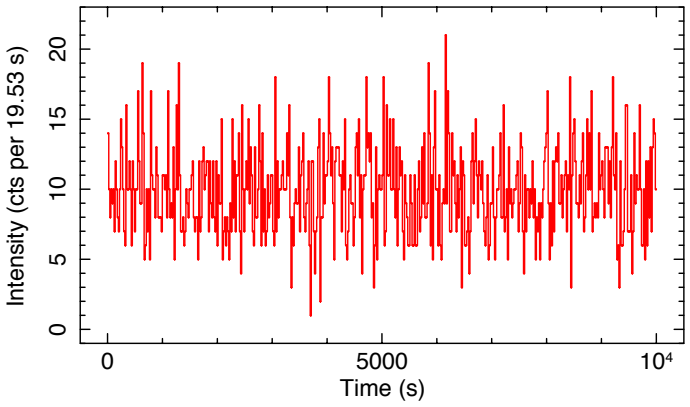

Figure 1: White noise time series with total duration of 10 ks binned on 512 time bins (δt=19.531 s).

There's something else we could—and should—look at right away, and that's the distribution of *waiting* times (interarrival times). This will immediately tell us if there's something odd, like some kind of regular dead time that would imprint a structure that isn't inherit to the data. We just measure the time between each event and draw up their distribution as shown in Figure 2.

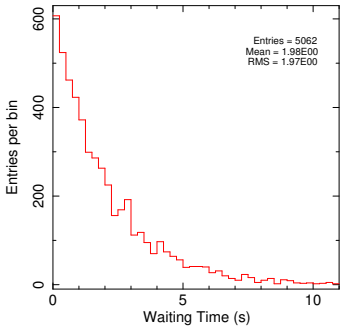

Figure 2: Distribution of waiting or inter-arrival times between events of which there are a total of 5063, and although the distribution is shown up to 11, the longest waiting time is longer than 20 seconds. Such is the nature of the exponential distribution.

This shows us that there doesn't seem to be anything that is out of place, and that the waiting times are distributed as expected following a roughly exponential distribution, with a slightly wider tail due to the variability; this distribution is a pure exponential only for homogeneous Poisson processes—white noise of constant mean rate.

It is important to check this carefully because any kind of measurement effects that impose a temporal structure onto the data need to be identified in order to be taken into account when we treat and analyse the data we have collected.

Now that we have checked this, and found that things look generally good, we can move forward. And one of the most effective ways to look carefully at the details of the distribution in time of a collection of measurements is to make a periodogram.

Periodogram analysis

A PERIODOGRAM DECOMPOSES the time-domain signal into a frequency-domain signal. Each of the frequencies that are accessible in the data can be tested for periodicity in the arrival times of events. So if there is any kind of temporal structure in these data we will see it as an excess of *power* at the one or several frequencies carrying information about this structure. And this is where we get to the essence of this talk: how can we access as much information as possible with the periodogram statistic we are using.

There are three things we need from a powerful and reliable periodogram statistic: it must (1) be able to use each event's arrival time in order to access all variability timescales, (2) allow for oversampling in order to explore frequency space without restrictions, and (3) take into account the oscillation in the mean, variance, and covariance of the Fourier components as a function of frequency. Let me explain what this means.

Fast Fourier Transform periodogram

The most well know and commonly used periodogram is made using the fast Fourier transform (FFT). To run the FFT, the time series of events must be binned (in bins of equal size), and the bin time (or number of bins), determines the maximum frequency that can be tested; it's called the Nyquist frequency, and it is given by $\nu_{\rm N} = 1/2\delta t$, where δt is the bin time.[1] The lowest frequency that can be tested is always determined by the length of the time series, as we obviously cannot test a period that is longer than that; so, $\nu_{\rm min} = 1/T$. Here is the FFT periodogram of the time series of Figure 1:

[1] It is important to highlight that the corresponding limit for arrival times would be defined based on the shortest of the waiting time between two events in the list.

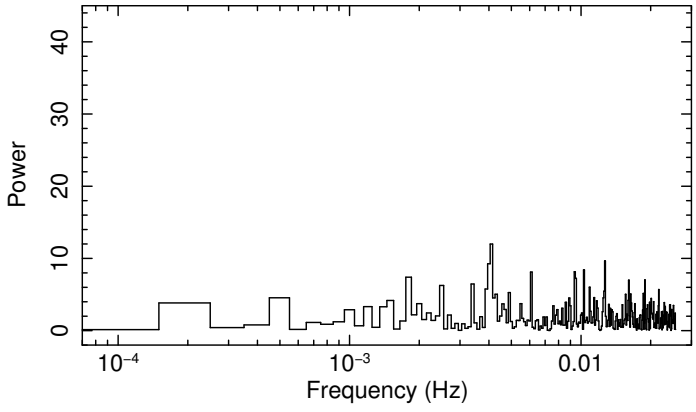

Figure 3: The FFT periodogram computed on 512 time bins (δt=19.531 s) and thus 256 real frequencies with $\nu_{\rm min} = \delta f = 10^{-4}$ Hz and $\nu_{\rm max} = 0.0256$ Hz.

Each of the frequencies that were tested are called independent Fourier frequencies because they are an integer multiple of $1/T$. The step size between independent frequencies is $1/T$. But aren't there an infinite number of frequencies between any two of these independent frequencies? Is there no way to test for those? What if the periodic signal happen to fall exactly between two independent

frequencies? We need to be able to *oversample* which means to test frequencies that cover the space between independent frequencies.

The Rayleigh periodogram

This we can do easily with an event data periodogram like the classic Rayleigh statistic [Leahy et al., 1983] that will, if we oversample finely enough, reveal peaks that could have been missed with the standard FFT periodogram.[2] As is obviously the case here:

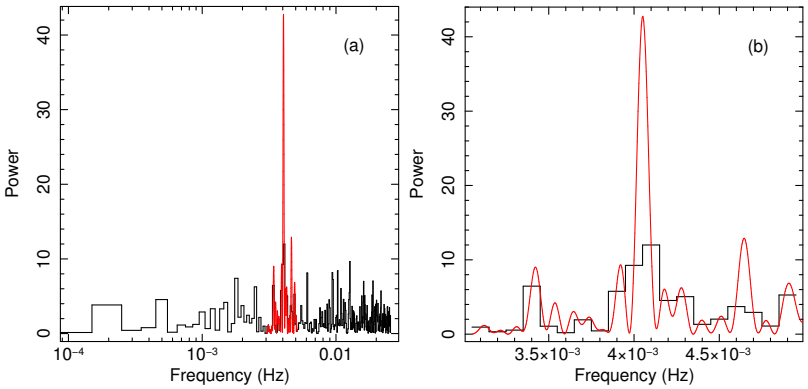

Figure 4: Comparison of the FFT and R^2 statistic on simulated data that are of white noise (duration $T = 10$ ks, mean rate $\mu = 0.5\,\mathrm{s}^{-1}$), with a 10% pulsed fraction (1 of 10 events is modulated) for a sinusoid at 0.00405 Hz (≈ 247 s). The Rayleigh periodogram in red is around ± 1 IFS of the peak (0.003–0.005 Hz), with sampling of 21 frequencies per IFS. Panel (a) shows the full range, and panel (b) shows the range of the R^2 periodogram. This example with a period between two independent frequencies was picked to clearly illustrate the important difference in sensitivity that can be achieved in some cases.

And it is immensely clear that this periododic signal which is totally obvious in the Rayleigh periodogram, was simply not present in the FFT. And the reason for that is just that it happened to be precisely between two independent frequencies. Moreover, as you can see, the Rayleigh periodogram permits us to test any frequency over any arbitrary range of frequencies. But what if it was computed for the entire frequency range?

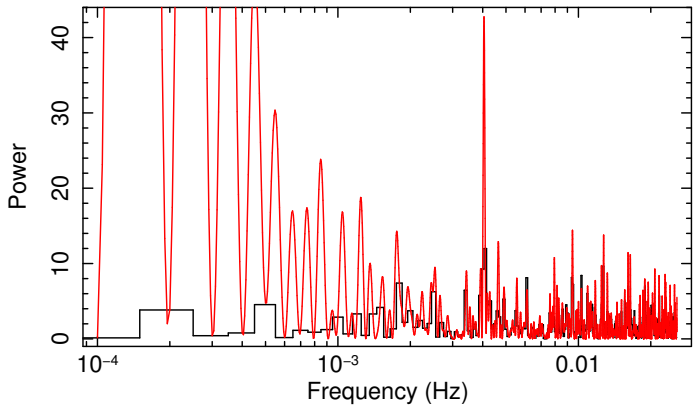

Figure 5: Artifacts in the R^2 periodogram shown on a truncated linear scale, visibly growing in a power-law fashion toward lower frequencies, with R^2 estimates between independent frequencies deviating noticeably from the FFT estimates below $\approx 3 \times 10^{-3}$ Hz.

All this stuff that you see at low frequencies is statistical noise. It's not real. What I mean by that is that the statistical noise, these huge peaks of power in the periodogram, are artefacts that arise from treating all frequencies as if they were independent frequencies even though they are not.

[2] From a data set comprising N events, the Rayleigh power at a given frequency ν (or period P), is calculated by converting each arrival time, t_i, to a phase, ϕ_i, given by the fractional part of $2\pi t_i \nu$ (or $2\pi t_i / P$), and computing

$$R^2 = \frac{2}{N}\left[\left(\sum_{i=1}^{N}\cos\phi_i\right)^2 + \left(\sum_{i=1}^{N}\sin\phi_i\right)^2\right]$$

The modified Rayleigh periodogram

The good news is that once we've understood that, we can include the appropriate correction in the calculation of power analytically right in the equations used to compute the periodogram. And this is what we get when we do:

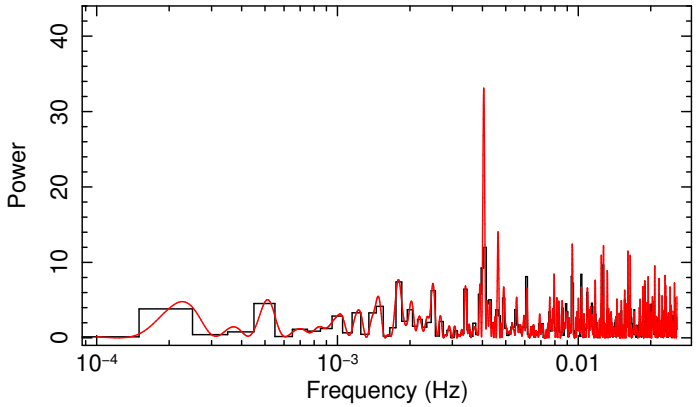

Figure 6: The $\mathcal{R}_1^2$ statistic applied to the same data, with the periodic signal clearly detected at the same frequency (0.004052 Hz) but at a somewhat lower power from the more accurate calculation (33.1 instead of 42.8, and a probability of 10^{-8} instead of 10^{-10} of arising from a noise fluctuation).

What we see if we look carefully, is that the estimates of power in this modified Rayleigh periodogram agree with the FFT at the independent frequencies, that it clearly detects the periodic signal just like the R^2 does, and that it doesn't suffer from the fake power peaks at low frequencies that completely disqualifies the R^2 periodogram from being able to make any statement about what could be happening in this region of frequency space. Let's go one step further.

The new $\mathcal{R}_k^2$ and $\mathcal{Z}^2$ periodograms

What if we are not looking for a weak or sinusoidal signal? What if we're trying to characterise a signal whose pulse profile is skewed or highly peaked, like it is for the Crab pulsar? This is why the Z^2 test [Buccheri et al., 1983] was devised: to sum the power of higher harmonics beyond just the fundamental. The Z^2 test sums over several R^2 periodograms for different harmonics. But obviously, this means that it will suffer not only in the same way as the R^2 does, but much more, because the artefacts will be combined with each additional harmonic that is used in the sum.

The solution is a generalised modified Rayleigh statistic:[3]

$$\mathcal{R}_k^2 = \begin{pmatrix} C_k - \langle C_k \rangle \\ S_k - \langle S_k \rangle \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \sigma^2{}_{C_k} & \sigma_{C_k S_k} \\ \sigma_{C_k S_k} & \sigma^2{}_{S_k} \end{pmatrix}^{-1} \begin{pmatrix} C_k - \langle C_k \rangle \\ S_k - \langle S_k \rangle \end{pmatrix} \tag{1}$$

that can be used to construct a modified Z^2 statistic, $\mathcal{Z}^2$, as:

$$\mathcal{Z}^2 = \sum \mathcal{R}_k^2. \tag{2}$$

Harmonic decomposition of peaked pulse profiles

Equipped with these, we can do something like this: We can compute the power contained in the Crab's pulse at different harmonics

[3] The dependency on the harmonic is carried by the variable k in the argument of the sine and cosine functions. The terms C_k and S_k are defined as:

$$C_k = \frac{1}{N}\sum_{i=1}^{N} \cos k\phi_i \quad \text{and} \quad S_k = \frac{1}{N}\sum_{i=1}^{N} \sin k\phi_i,$$

and the other terms are given by

$$\langle C_k \rangle = \frac{1}{k\omega T}\left[\sin k\omega t\right]_{t_1}^{t_2},$$

$$\langle S_k \rangle = \frac{-1}{k\omega T}\left[\cos k\omega t\right]_{t_1}^{t_2},$$

$$\sigma^2_{C_k} = \frac{1}{2N}\left(1 + \frac{1}{k\omega T}\left[\sin k\omega t \cos k\omega t\right]_{t_1}^{t_2}\right) - \langle C_k \rangle^2,$$

$$\sigma^2_{S_k} = \frac{1}{2N}\left(1 - \frac{1}{k\omega T}\left[\sin k\omega t \cos k\omega t\right]_{t_1}^{t_2}\right) - \langle S_k \rangle^2,$$

$$\sigma_{C_k S_k} = \frac{1}{2k\omega TN}\left[\sin^2 k\omega t\right]_{t_1}^{t_2} - \langle C_k \rangle \langle S_k \rangle.$$

The terms $\langle C_k \rangle$ and $\langle S_k \rangle$ are the expectation values, $\sigma^2_{C_k}$ and $\sigma^2_{S_k}$ are the variances, and $\sigma_{C_k S_k}$ is the covariance of C_k and S_k.

using the $\mathcal{R}_k^2$ statistic. When we do this, we find that most of the power comes from the second, third, and fifth harmonics, but that there's hardly any power in the fourth, and that there's actually more power in the third and fifth than in the fundamental.

It's very interesting. And we simply would never be able to know this without the help of such a periodogram statistic. But the most critical point is that because the power is computed correctly at every frequency, we can trust the result to be accurate and reliable. Isn't this always the most important? What's the point of getting any result at all if it's not reliable and accurate. Here's what this looks like graphically:

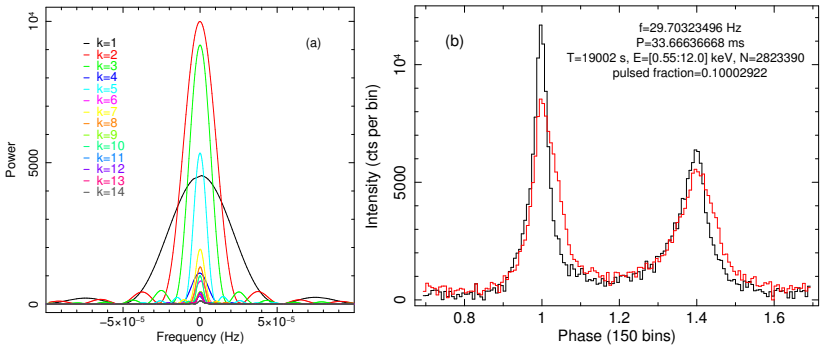

Figure 7: Multi-harmonic periodogram of the Crab pulsar's X-ray emission centered on the pulse frequency in Panel (a) and phasograms resulting from folding the arrival times on our best estimate of the pulse frequency compared to the frequency derived from the radio ephemerides in Panel (b). The data is from an *XMM-Newton* observation (ID 0611181501-003 on 2012 February 24–25) with an elapsed time of 19,002 s, using the Epic PN Timing/FastBurst data comprising 2823390 events in the range 0.55–12.0 keV (mean rate 148.6 s^{-1}).

Closing remarks

We will close on this, and just leave you with the following concluding remarks:

- Event data are quantitatively different from a collection of measurements in which there is an inherent binning due to the measurement process or the quantity being measured. It carries temporal information.
- To extract as much as we can from this temporal information the events carry we need to use powerful and reliable tools, that include periodogram statistics.
- A powerful and reliable periodogram statistic must be able to use each event's arrival time in order to access all variability timescales; allow for oversampling in order to explore frequency space without restrictions; and correct for the testing of non-independent frequencies.
- In light of this, the periodogram statistics of choice are the new generalised modified Rayleigh $\mathcal{R}_k^2$ statistic and the $\mathcal{Z}^2$ test derived from it.

You can find more details on this in Belanger [2016]. Thank you for listening.

References

G Belanger. On More Sensitive Periodogram Statistics. *ApJ*, 822(1): 14, May 2016.

Buccheri et al. Search for pulsed gamma-ray emission from radio pulsars in the COS-B data. *A&A*, 128:245, nov 1983.

D. A Leahy, R. F. Elsner, and M. C. Weisskopf. On searches for periodic pulsed emission - The Rayleigh test compared to epoch folding. *ApJ*, 272:256, sep 1983. DOI: 10.1086/161288.