



<b>Publication Year</b>	2021
<b>Acceptance in OA</b>	2022-03-15T15:32:59Z
<b>Title</b>	Extending the Z2n and H Statistics to Generic Pulsed Profiles
<b>Authors</b>	BACHETTI, Matteo, PILIA, Maura, Huppenkothen, Daniela, Ransom, Scott M., Curatti, Stefano, RIDOLFI, ALESSANDRO
<b>Publisher's version (DOI)</b>	10.3847/1538-4357/abda4a
<b>Handle</b>	<a href="http://hdl.handle.net/20.500.12386/31589">http://hdl.handle.net/20.500.12386/31589</a>
<b>Journal</b>	THE ASTROPHYSICAL JOURNAL
<b>Volume</b>	909

## Extending the $Z_n^2$ and $H$ statistics to generic pulsed profiles

MATTEO BACHETTI,<sup>1</sup> MAURA PILIA,<sup>1</sup> DANIELA HUPPENKOTHEN,<sup>2</sup> SCOTT M. RANSOM,<sup>3</sup> STEFANO CURATTI,<sup>4</sup> AND  
ALESSANDRO RIDOLFI<sup>1,5</sup>

<sup>1</sup>INAF-Osservatorio Astronomico di Cagliari, via della Scienza 5, I-09047 Selargius (CA), Italy

<sup>2</sup>SRON Netherlands Institute for Space Research, Sorbonnelaan 2, 3584 CA, Utrecht, Netherlands

<sup>3</sup>National Radio Astronomy Observatory, 520 Edgemont Road, Charlottesville, VA 22903, USA

<sup>4</sup>Infora Società Cooperativa, Viale Elmas 142, I-09122 Cagliari, Italy

<sup>5</sup>Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, D-53121 Bonn, Germany

(Received June 1, 2019; Revised January 10, 2019; Accepted March 8, 2021)

Submitted to ApJ

### ABSTRACT

The search for astronomical pulsed signals within noisy data, in the radio band, is usually performed through an initial Fourier analysis to find “candidate” frequencies and then refined through the folding of the time series using trial frequencies close to the candidate. In order to establish the significance of the pulsed profiles found at these trial frequencies, pulsed profiles are evaluated with a chi-squared test, to establish how much they depart from a null hypothesis where the signal is consistent with a flat distribution of noisy measurements. In high-energy astronomy, the chi-squared statistic has widely been replaced by the  $Z_n^2$  statistic and the H-test as they are more sensitive to extra information such as the harmonic content of the pulsed profile. The  $Z_n^2$  statistic and H-test were originally developed for the use with “event data”, composed of arrival times of single photons, leaving it unclear how these methods could be used in radio astronomy. In this paper, we present a version of the  $Z_n^2$  statistic and H-test for pulse profiles with Gaussian uncertainties, appropriate for radio or even optical pulse profiles. We show how these statistical indicators provide better sensitivity to low-significance pulsar candidates with respect to the usual chi-squared method, and a straightforward way to discriminate between pulse profile shapes. Moreover, they provide an additional tool for Radio Frequency Interference (RFI) rejection.

*Keywords:* statistics – pulsars

### 1. INTRODUCTION

In a typical pulsar search, the uncertainty in the flux measurements from a radio telescope are dominated by various sources of noise, both from the sky and the instruments. The pulsar signal can be millions of times weaker than the intrinsic noise of the time series. Therefore, pulsar searches employ Fourier analysis to search for periodicities. This search produces a growing number of candidate pulsations (mainly due to the ever increasing interfering signals from Earth or satellites). The selection of the most promising astrophysical signals amongst them is based on a small number of statistical indicators. Once a candidate pulsar frequency is found, the analysis of candidates typically starts from epoch Folding (EF; Leahy et al. 1983a).

Let  $(t_j, X_j) \forall j = 1, \dots, N$  be pairs of flux measurements  $X_j$  at times  $t_j$ . If we define  $f$  as the candidate pulse frequency and  $\dot{f}, \ddot{f}, \dots$  the frequency derivatives measured at a reference time  $t_{\text{ref}}$ , we calculate the *pulse phase* of each  $X_j$  as

$$\phi_j(t) = \phi_0 + f(t_j - t_{\text{ref}}) + 0.5\dot{f}(t_j - t_{\text{ref}})^2 + \dots \quad (1)$$

where  $\phi_0$  is the pulse phase at  $t_{\text{ref}}$ , which is set to 0 for unknown or candidate pulsars. Given the periodicity involved, one is typically only interested in the fractional part of this phase, distributed between 0 and 1. Since this quantity will always be multiplied by a  $2\pi$  factor in the formulae below, we will follow the convention of including  $2\pi$  in  $\phi$  for consistency: hereafter,  $\phi$  will always mean an angle between 0 and  $2\pi$ . The folded or pulsed profile is then a histogram of these phases falling into

$N_{\text{bin}}$  equal phase bins between 0 and  $2\pi$ , weighted by the flux in each sample<sup>1</sup>:

$$p_i = \sum_{j=1}^N X_j \theta(\phi_j - \phi_{\text{mid},i}) \quad (i = 1, \dots, N_{\text{bin}}) \quad (2)$$

where

$$\phi_{\text{mid},i} = 2\pi \frac{i - 0.5}{N_{\text{bin}}}$$

is the phase corresponding to the middle of bin  $i$  and

$$\theta(x) = \begin{cases} 1 & \text{if } |x| < 0.5/N_{\text{bin}} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Given a pulsed profile  $p_i$ , consisting of  $N_{\text{bin}}$  equally spaced bins, the most common statistical indicator is the Epoch Folding chi-squared statistic (Leahy et al. 1983b; Leahy 1987)

$$\mathcal{S} = \sum_{i=1}^{N_{\text{bin}}} \left( \frac{p_i - \bar{p}}{\sigma_i} \right)^2 \quad (4)$$

where  $\bar{p}$  is the average profile, and  $\sigma_i$  is the standard deviation in each profile bin. In the typical case, the uncertainties  $\sigma_i$  are derived from the statistical uncertainties of the individual flux measurements  $X_j$ . We assume they are distributed according to the Poisson or normal distribution, that the central limit theorem holds and, therefore, that the phase-folded flux measurements  $p_i$  are distributed normally. Also, when comparing the pulse profile obtained folding at different periods, one usually assumes that the  $\sigma_i$  are all equal and obtained by error propagation from the standard deviation of the  $X_j$ :

$$\sigma_i \equiv \sigma = \text{std}(X_j) \sqrt{\left( \frac{N_{\text{samples}}}{N_{\text{bin}}} \right)} \quad (5)$$

If the normality assumption holds, the quantity  $\mathcal{S}$  follows a  $\chi^2_{N_{\text{bin}}-1}$  distribution if the time series is composed of pure noise, and that makes it easy to evaluate the probability of rejecting the null hypothesis. Another quick statistical indicator, also used in searches of non-periodic impulsive phenomena in the time domain (i.e.

Fast Radio Bursts and single pulses), is the simple signal to noise ratio (SNR):

$$\text{SNR} = \frac{\max(p_i) - \bar{p}}{\sigma} \quad (6)$$

where  $\sigma$  is the standard deviation or the median absolute deviation of the profile, and the mean can be substituted by the median depending on the implementation.

In the X-ray and  $\gamma$ -ray bands astronomers have developed other statistical indicators that are more sensitive—better at finding low-amplitude signals—than  $\mathcal{S}$ , provided that the signal is composed of single “events” and the pulsed component can be described as the sum of a relatively small number of sinusoidal harmonics. Moreover, while  $\mathcal{S}$  and SNR only measure any deviation of the pulsed profile from a flat distribution, these new statistical indicators also provide the ability to discriminate between different pulse shapes, as we will discuss below.

Let us assume that we are observing a  $\gamma$ -ray or X-ray pulsar we already know. We know its ephemeris (i.e., how its rotation evolves over time), usually expressed as an initial phase  $\phi_0$  and series of frequency derivatives ( $f, \dot{f}, \ddot{f}, \dots$ ) measured at a reference time  $t_{\text{ref}}$ . Starting from a list of “events” corresponding to time stamps  $t_j$  of photons hitting a detector, one can calculate the phase of each photon  $\phi_j$  as in Eq. 1. The pulsed profile in this case can be obtained as a histogram of the fractional part of these phases.  $p_i$  is now the number of events counted in the phase bin  $i$ . Eq. 4 in this case becomes

$$\mathcal{S}_{\text{bin}} = \sum_{i=1}^{N_{\text{bin}}} \frac{(p_i - \lambda)^2}{\lambda} \quad (7)$$

where  $\lambda$  is the average of the pulsed profile (the equivalent of  $\bar{p}$  in Equation 4), and we used the fact that in a counting experiment the variance in each bin is equal to the Poisson rate.

Buccheri et al. (1983) introduced an alternative statistic, now-standard in X- and  $\gamma$ -rays, called  $Z_n^2$ , which is generally more sensitive for pulsed signals whose pulsed profile can be described by a small number of sinusoidal harmonics<sup>2</sup>. If we detect  $N$  photons and their pulse phase is  $\phi_i$  ( $i = 1..N$ ), we define  $Z_n^2$  as

$$Z_n^2 = \frac{2}{N} \sum_{k=1}^n \left[ \left( \sum_{j=1}^N \cos k\phi_j \right)^2 + \left( \sum_{j=1}^N \sin k\phi_j \right)^2 \right] \quad (8)$$

<sup>2</sup> Note that this is close but not equivalent to the incoherent harmonic summing used in radio searches, (See, e.g. Ransom et al. 2002; Lorimer & Kramer 2012)

<sup>1</sup> Whereas this is the most common way to fold radio data, some software packages like PRESTO use a slightly different approach: they assume that each sample is finite in duration and “drizzle” it over the appropriate pulse phase bins. This slightly different approach does not substantively affect the results of this paper, but see Appendix E for more details.

This formula evaluates how well the distribution of pulse phases is described by a series of  $n$  sinusoidal harmonics of increasing order. The statistic in the case of a single harmonic,  $Z_1^2$ , is also known as the Rayleigh test (Mardia 1975; Gibson et al. 1982). Similarly to  $\mathcal{S}$ , this statistic can be used for pulsar searches as well. Given a certain number of trial pulse periods, one can calculate  $Z_n^2$  and select the periods that give the highest value for the statistic.  $Z_n^2$  follows a  $\chi_{2n}^2$  distribution for noise powers (i.e. the statistic’s values obtained with trial period values far from the real period), so that it is straightforward to set “detection” levels or calculate the probability of rejecting the null hypothesis given a  $Z_n^2$  value.

To show why the  $Z_n^2$  statistic is more sensitive than  $\mathcal{S}_{bin}$  for a small number of sinusoidal components, we can follow Leahy et al. (1983b) and define the “quality factor” of the statistic as

$$Q = \frac{S_{sig}}{S_{thr} - \nu} \quad (9)$$

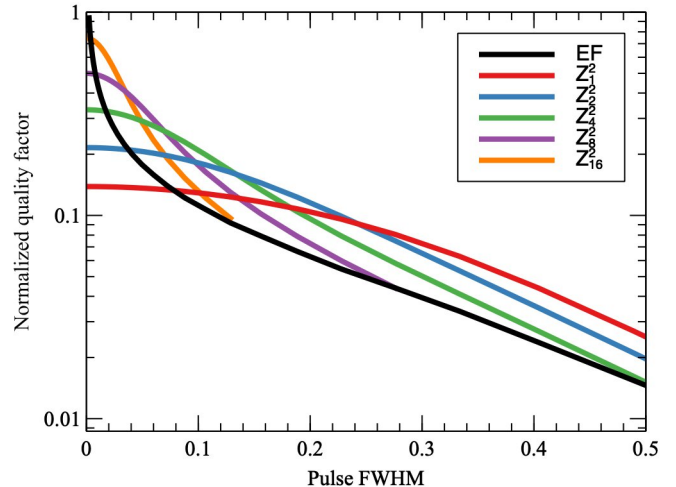
where  $S_{sig}$  is the statistic’s value of a given pulsed profile,  $S_{thr}$  is the “threshold” statistic value indicating a given small p-value for noise powers, and  $\nu$  is the number of degrees of freedom. Leahy et al. (1983b) compares Epoch Folding and the Rayleigh test ( $Z_1^2$ ) using square profiles with different duty cycles (i.e. duration of the “high” signal value), showing that  $\mathcal{S}$  is more sensitive for profiles with shorter duty cycle (duration), while the Rayleigh test is more sensitive for profiles with longer duty cycle. In the next section, we will use a similar approach to show that  $Z_n^2$  is more sensitive than  $\mathcal{S}$  in all cases where the signal can be described by a relatively small number of harmonics.

Finally, one can, in a single test, determine if the signal is described by a sum of harmonics, and what is the best number of harmonics to describe the signal. This is known as the  $H$  test, and it is basically a summary of multiple  $Z_n^2$  values, properly normalized so that they can be compared (de Jager et al. 1989; de Jager & Büsching 2010):

$$H = \max(Z_m^2 - 4m + 4), \quad m = 1, 20 \quad (10)$$

with the  $m$  corresponding to the maximum usually indicated with  $M$ .

While these statistical indicators were developed for Poisson-distributed data (photon-counting experiments), they can be adapted to the case of pulsed profiles obtained by, for example, radio telescopes, through a few simple prescriptions. This paper explains how to adapt these tests to generic pulsed profiles. We do this in two steps: 1) in Section 2 we test that the binned version of the  $Z_n^2$  statistic proposed by Huppenkothen et al. (2019)



**Figure 1.** Quality factors for  $\mathcal{S}$  and  $Z_{n,bin}^2$  for binned pulsed profiles consisting of a single Gaussian with increasing width (in units of pulse phase, from 0 to 1). Following (Leahy et al. 1983b), we normalized the quality factor by  $Nf^2$ , where  $f$  is the ratio between the pulsed area and the total area under the pulsed profile including the DC level.

has the required statistical properties both with white noise and with pulsed signals; 2) in Section 3 we extend this formulation by introducing a version of  $Z_n^2$  for pulsed profiles with Gaussian uncertainties.

## 2. EXTENSION OF THE Z STATISTIC FOR BINNED POISSON DATA

As we have seen, Eq. 8 defines the  $Z_n^2$  statistic for unbinned event data. Huppenkothen et al. 2019 introduced a version of the  $Z_n^2$  test for binned pulsed profiles obtained from X-ray event data):

$$Z_{n,bin}^2 \approx \frac{2}{\sum_i p_i} \sum_{k=1}^n \left[ \left( \sum_{i=1}^{N_{bin}} p_i \cos k\phi_i \right)^2 + \left( \sum_{i=1}^{N_{bin}} p_i \sin k\phi_i \right)^2 \right] \quad (11)$$

where  $p_i$ , working as a weight, corresponds to the number of photons in a given profile bin. We know that we need at least two bins for each cycle to sample a sinusoid (Nyquist limit), so we expect the approximation to make little sense if  $N_{bin} < 2n$ . Also, it is reasonable to expect a better approximation as we sample each harmonic cycle with more and more bins. Huppenkothen et al. (2019) prudently recommend to bin the pulsed profile with a number of bins at least 10 times larger than the number of harmonics  $n$ , but give no justification.

The effect of binning can be quantified through simple analytical manipulations (developed in Appendix B).

Calling

$$A_k = \sum_{j=1}^N \cos k\phi_j \quad (12)$$

$$B_k = \sum_{j=1}^N \sin k\phi_j \quad (13)$$

we can express  $Z_n^2$  as

$$Z_n^2 = \frac{2}{N} \sum_{k=1}^n [A_k^2 + B_k^2] \quad (14)$$

One can demonstrate (see Appendix B) that the binned approximation gives

$$Z_{n,\text{bin}}^2 \approx \frac{2}{N} \sum_{k=1}^n \left( \frac{N_{\text{bin}}}{\pi k} \sin \frac{\pi k}{N_{\text{bin}}} \right)^2 [A_k^2 + B_k^2] \quad (15)$$

Comparing it to Eq. 11, we see that this formula only differs by the factors

$$C(k) = \left( \frac{N_{\text{bin}}}{\pi k} \sin \frac{\pi k}{N_{\text{bin}}} \right)^2 \quad (16)$$

that go to 1 for  $k \ll N_{\text{bin}}$  and lead to a loss of sensitivity<sup>3</sup> of  $\sim 40\%$  for  $k = N_{\text{bin}}/2$ .

Assuming a pulsed profile of the kind

$$p_i = \lambda \left[ 1 + \sum_{l=1}^m a_l \sin l(\phi_i - \phi_{0,l}) \right] \quad (17)$$

(i.e. a DC level  $\lambda$  plus a finite number  $m$  of sinusoidal harmonics with random phases) we can define a statistic  $S_{\text{sig}}$  as

$$S_{\text{sig}} \approx \frac{\lambda N_{\text{bin}}}{2} \sum_{l=1}^m a_l^2. \quad (18)$$

It can be shown (See Appendices D and C) that the statistic  $S_{\text{sig}}$  has exactly the same value for  $\mathcal{S}$  and  $Z_{n,\text{bin}}^2$  provided that  $n \geq m$  and one can neglect the binning ( $N_{\text{bin}} \gg 2m$ ) and reduces to the  $(1/2)NA^2$  (where  $N = \lambda N_{\text{bin}}$  is the total number of photons) given by Leahy et al. (1983b), for  $m = 1$ .

Taking Eq. 9 and comparing the sensitivity of  $\mathcal{S}$  and  $Z_{n,\text{bin}}^2$ , we find the results plotted in Figure 1. Here we created Gaussian pulsed profiles with increasing width (and so, a decreasing number of harmonics) and calculated the following statistic:

- $Z_{n,\text{bin}}^2$  for  $n=1, \dots, 20$
- $\mathcal{S}$  using an increasing number of bins to avoid the sensitivity loss when  $N_{\text{bin}}/n \sim 2$

We used a number of bins that oversampled the Gaussian peak (two points inside  $1 - \sigma$ ) in order not to decrease the sensitivity of any of the methods. Therefore, in Figure 1,  $N_{\text{bin}}$  increases going from right to left. As expected,  $Z_{n,\text{bin}}^2$  is *always* more sensitive than  $\mathcal{S}$  when we use a sufficient number of sinusoidal components. The reason is that the  $Z_{n,\text{bin}}^2$  threshold level for any given  $n$  is insensitive to the number of bins, if not for small corrections at small  $N_{\text{bin}}$ , while  $\mathcal{S}$ 's threshold increases with  $N_{\text{bin}}$ . To detect very sharp signals (or high-order harmonics) we need a large number of profile bins, and  $\mathcal{S}$  is systematically noisier than  $Z_{n,\text{bin}}^2$ . Since the signal level  $S_{\text{sig}}$  is the same for  $Z_{n,\text{bin}}^2$  and  $\mathcal{S}$  (Eq. 18), increasing quadratically with signal amplitude, and the denominator increases slowly with increasing number of degrees of freedom (approximately with the square root), the significance of very strong signals will be very similar using the two methods.

$Z_{n,\text{bin}}^2$  can be used to calculate the H statistic, a single test which allows to investigate the occurrence of signals of multiple sinusoids whose number is not known a priori (Eq. 10). Our generalization of the  $Z_n^2$  test to binned data substantially confirms what was described by de Jager et al. (1989) for the unbinned statistic, and validates the use of the H test for binned profiles. In the next section, we further extend this method to generic profiles (i.e. obtained by any measurements and not just particle/photon counts), provided that their measurement error is approximately Gaussian and constant throughout the profile. Hereafter, we will run the  $Z_{n,\text{bin}}^2$  analysis using multiple  $n$ , and we will be using the  $H$ -test to decide for each profile what number of harmonics  $M$  gives the best detection. Therefore, when we talk about the  $Z_{n,\text{bin}}^2$  significance or the  $H$ -test significance, we will be referring to the significance of  $Z_{n,\text{bin}}^2$  with  $n = M$ .

### 3. THE Z STATISTIC IN THE CASE FOR NORMALLY DISTRIBUTED DATA

Let us start with a set of  $N$  observations  $\{X_j\}_j^N$  at times  $t_j$ , as defined in Section 1, distributed around a mean flux  $\mu_x$  with a variance of  $\sigma_x^2$ . We assume here that the variance is the same for all bins,  $\sigma_{x,j}^2 = \sigma_x^2$ , but will relax that requirement later. We aim to use the  $Z^2$  statistic to compare to a null hypothesis of a constant flux, hence we also assume that the measurements are distributed around a constant flux value  $\mu_x = \text{constant}$ . Thus, measurements are drawn from a normal distribution,  $x_j \sim \mathcal{N}(\mu_x, \sigma_x^2)$ .

<sup>3</sup> This is equivalent to the loss of sensitivity in the power density spectrum due to sampling. See van der Klis (1989), Appendix B

We now compute a pulse profile by distributing fluxes into  $N_{\text{bin}}$  phase bins, and sum the fluxes within each phase bin to obtain phase-dependent fluxes

$$p_i = \sum_{j=iN_{\text{fmpb}}}^{(i+1)N_{\text{fmpb}}} x_j \quad \forall i, i = 0, \dots, N_{\text{bin}} - 1. \quad (19)$$

Here,  $N_{\text{fmpb}}$  is the number of flux measurements in each phase bin,  $N_{\text{fmpb}} = \frac{N}{N_{\text{bin}}}$ . Because the sum of Gaussian variables is also distributed as a Gaussian,  $p_i$ , too, are distributed normally, with the same mean  $\mu_p = \sum_{j=1}^{N_{\text{fmpb}}} \mu_x = \mu_x N_{\text{fmpb}}$ . The variance of the summed random variable can also be calculated as the sum of variances:

$$\sigma_p^2 = \sum_{j=1}^{N_{\text{fmpb}}} \sigma_x^2 = \sigma_x^2 N_{\text{fmpb}}.$$

We define the  $Z_{n,\text{gauss}}$  statistic as

$$Z_{n,\text{gauss}}^2 = \frac{1}{K} \sum_{k=1}^n \left[ \left( \sum_{i=1}^{N_{\text{bin}}} p_i \cos k\phi_i \right)^2 + \left( \sum_{i=1}^{N_{\text{bin}}} p_i \sin k\phi_i \right)^2 \right] \quad (20)$$

where  $K = \sigma_p^2 N_{\text{bin}} / 2$  (see below),  $p_i$  are phase-folded fluxes in each of  $N_{\text{bin}}$  phase bins. Each phase bin is characterized by a phase  $\phi_i$  measured at the mid-point of the phase bin. Finally, in this general case, we compute the  $Z^2$  statistic over  $n$  harmonics. In the following, we will assume  $n = 1$  for much of this section, but will show in the end how this case generalizes to  $n > 1$ .

In the unbinned Poisson case for which  $Z_n^2$  was initially defined, the phases  $\phi_j$  are uniform random variables across the interval  $[0, 2\pi]$ . For the case of summed Gaussian fluxes, the phases  $\phi_i$  are not random variables, but rather real numbers on a regular grid over the same interval spanning 0 to  $2\pi$ . It follows that the sine and cosine of the phases  $\sin \phi_i$  and  $\cos \phi_i$  are also not random variables, but real numbers over the interval  $[-1, 1]$ . In the following, we will first derive the appropriate distributions for the cosine term, but because trigonometric identities hold, the sine term will produce random variables drawn from the same statistical distribution.

Because  $\cos \phi_i$  is not a random variable and  $p_i$  is Gaussian, the product will also be distributed normally, with

$$p_i \cos \phi_i \sim \mathcal{N}(\mu_p \cos \phi_i, \sigma_p^2 \cos^2 \phi_i)$$

The sum of the product of summed fluxes  $p_i$  and cosine terms amounts to another sum of random variables drawn from a number of independent and identically distributed

Gaussians, each of which has been scaled by  $\cos \phi_i$ . We define the sum as

$$A = \sum_{i=1}^{N_{\text{bin}}} p_i \cos k\phi_i$$

This sum, too, yields a Gaussian distribution for which mean and variance are defined as

$$\begin{aligned} \mu_A &= \sum_{i=1}^{N_{\text{bin}}} \mu_p \cos \phi_i \\ \sigma_A^2 &= \sum_{i=1}^{N_{\text{bin}}} \sigma_p^2 \cos^2(\phi_i). \end{aligned} \quad (21)$$

Note that because  $\sum_{i=1}^{N_{\text{bin}}} \cos \phi_i = 0$ , it follows that  $\mu_A = 0$ . Similarly, for a variance that is constant across all profile bins, we can take  $\sigma_p^2$  outside of the sum, and calculate the value for  $\sum_{i=1}^{N_{\text{bin}}} \cos^2(\phi_i)$  as a Riemann sum:

$$\sum_{i=1}^{N_{\text{bin}}} \cos^2(\phi_i) = \frac{\pi}{\Delta\phi}$$

where  $\Delta\phi$  is the size of a phase bin,  $\Delta\phi = \frac{2\pi}{N_{\text{bin}}}$ . Thus, we find that

$$\sigma_A^2 = \sigma_p^2 \frac{\pi N_{\text{bin}}}{2\pi} = \sigma_p^2 \frac{N_{\text{bin}}}{2} \quad (22)$$

Based on this result, we can now determine the statistical distribution for  $A^2$ . In particular, for a Gaussian random variable  $X$  with a standard distribution  $\sigma$ , the variable normalized by its standard deviation is distributed following a standard chi-square distribution with one degree of freedom:  $\frac{X^2}{\sigma^2} \sim \chi_1^2$ . It thus follows that we can define a variable based on  $A^2$ , where  $A$  is defined above, that follows a standard  $\chi_1^2$  distribution.

If

$$A^2 = \left( \sum_{i=1}^{N_{\text{bin}}} p_i \cos k\phi_i \right)^2$$

then the variable

$$Q = \frac{A^2}{\sigma_A^2} = \frac{\left( \sum_{i=1}^{N_{\text{bin}}} p_i \cos k\phi_i \right)^2}{\sigma_p^2 \frac{N_{\text{bin}}}{2}} \quad (23)$$

follows a standard  $\chi_1^2$  distribution. We can treat the second term in the outer sum of the  $Z^2$  statistic,

$$B = \sum_{i=1}^{N_{\text{bin}}} p_i \sin k\phi_i$$

in the same way as  $A$ . Because trigonometric identities hold, we obtain another  $\chi_1^2$  distributed variable for  $B^2$ .

Adding these (properly normalized) variables together yields the final  $Z^2$  statistic. Because the sum of  $\chi_1^2$  distributed variables is a  $\chi_2^2$  distributed variable with degrees of freedom  $df = 2$ , we find that  $Z_1^2 \sim \chi_2^2$  assuming the correct normalization constant  $K$ , defined for the case of constant data variances,  $\sigma_{P,j} = \sigma_P$  as

$$K = \sigma_P^2 \frac{N_{\text{bin}}}{2} \quad (24)$$

Extending this result to  $n > 1$  in order to account for multiple harmonics is straightforward: each harmonic adds a variable to the sum drawn from  $\chi_2^2$ . Because the sum of  $k$  random variables, each drawn from a  $\chi_2^2$  distribution, is another  $\chi^2$  distribution with  $df = 2n$ , we find that

$$Z_{n,\text{gauss}} \sim \chi_{2n}^2. \quad (25)$$

### 3.1. Extension to data sets where the uncertainty varies with phase

Let us assume a case where  $\sigma_p^2 = \sigma_{p,i}^2$ , i.e. the summed fluxes are still drawn from a normal distribution, but each with a different variance,

$$p_i \sim \mathcal{N}(\mu_x, \sigma_{p,i}^2).$$

This change in assumptions alters one important step of the calculation: in Eq. 22, we have assumed that the variance of each profile bin is identical, which made it possible to take that term out of the sum and compute the sum of all  $\cos \phi_i$  terms independently. For unequal variances, this no longer holds, and normalization for the  $Z_n^2$  statistic instead becomes

$$K = \sum_{i=1}^{N_{\text{bin}}} \sigma_{p,i}^2 \cdot \cos^2(\phi_i) \quad (26)$$

Note that this normalization is a generalization of Eq. 24.

The distributions presented above are *exact*. For both unbinned and binned Poisson data, the assumption that we can write  $Z_n^2 \sim \chi_{2n}^2$  crucially rests on the validity of the Central Limit Theorem for this case. In the case of unbinned Poisson data, the phases  $\phi_j$  are uniformly distributed random variables, and the cosine of these phases is distributed as

$$\cos \phi_j \sim \text{Arcsin}(-1, 1)$$

where  $\text{Arcsin}(-1, 1)$  is a scaled arcsine distribution (and equivalently,  $\sin \phi_j$  also produces an arcsine distribution). This distribution is heavily non-Gaussian, and the

sum  $\sum_{j=1}^N \cos \phi_j$  for  $N$  photons will only approximate a Gaussian distribution if  $N$  is reasonably large.

Similarly, when binned Poisson event rates are used in Eq. 20 as  $p_i$  instead of normally distributed measurements, the resulting product  $p_i \cos \phi_i$  will not be drawn from any analytically known probability distribution. As a result,  $Z_n^2$  will only be distributed as the expected  $\chi_{2n}^2$  distribution if either the count rates  $p_i$  in each bin are large enough that the Poisson counts approximate a normal distribution, or there are a sufficient number of bins  $i$  that the sum of bins will again tend to a Gaussian.

Also note that in the binned Poisson case with mean and variance equal to  $\lambda$ , Eq. 24 yields

$$K = \sigma_p^2 \frac{N_{\text{bin}}}{2} = \frac{\lambda N_{\text{bin}}}{2} = \frac{\sum_i p_i}{2}$$

that confirms Eq. 11.

### 3.2. Extension to unevenly sampled Gaussian-distributed measurements

In practice, light curves—especially in the optical wavelength regime—are often subject to observing constraints, leading to unevenly sampled time series. In calculating  $Z^2$ , this translates into phase-binned fluxes  $p_i$  for which the number of measurements in each bin differs for each bin  $i$ ,  $N_{\text{fmpb}} = N_{\text{fmpb},i}$ . In these cases, it is advisable to calculate  $Z^2$  for the *average* of fluxes falling into the same phase bin, rather than the *sum*, as we have done above.

In this case, we define

$$p_i = \frac{1}{N_{\text{fmpb},i}} \sum_{j=1}^{N_{\text{fmpb},i}} x_j \quad \forall i, i = 0, \dots, N_{\text{bin}} - 1 \quad (27)$$

In this case, the mean of the resulting variables  $p_i$  correspond to the mean of the individual measurements,  $\mu_p = \mu_x$ , and the variance becomes

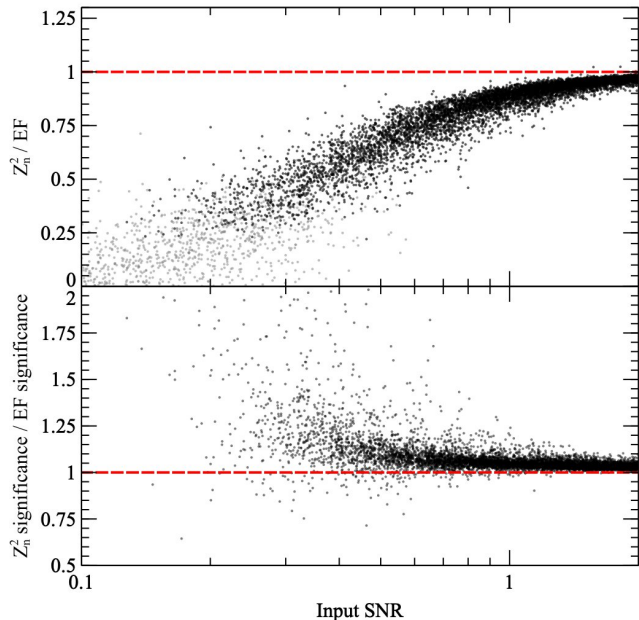
$$\sigma_{p,i}^2 = \frac{1}{N_{\text{fmpb},i}} \sigma_x^2. \quad (28)$$

The remainder of the procedure can be calculated using the same formalism as for the case described in Section 3.1.

## 4. TEST THE GAUSSIAN Z AND H TESTS: SIMULATIONS

Let us now test whether this method produces predictable values of  $Z_n^2$  and H, and compare them with the other statistical indicators used in radio astronomy.

We simulated 10000 pulsed profiles with 1024 bins with Gaussian errors, varying all parameters (input SNR, peak



**Figure 2.** Comparison of the raw statistical value (top) and the detection significance (bottom) using the standard epoch folding statistic (Eq. 4) and the  $Z_{n,\text{gauss}}$  (Eq. 20). In the signal-dominated regime (high significance) the values from  $\mathcal{S}$  and  $Z_{n,\text{gauss}}$ , and their significances, approach asymptotically. However, at low SNR the  $\mathcal{S}$  gives higher values of the statistic, but  $Z_n^2$  is more sensitive, yielding a higher significance (see Section 2). Grey points indicate values below detection level (these points have invalid significance ratio and are absent in the bottom plot.)

width, baseline, noise, etc.). The large number of bins was chosen to minimize binning effects for high-order harmonics (See Section 2, Eq. 15). We evaluated the profiles with the following statistic:

- the Epoch folding statistic (Eq. 4), rebinning the pulsed profile to 8, 16, 32, ... 256 bins. As the EF statistic is noisier for a higher number of bins, but on the other hand it might miss sharp profiles for a small number of bins, we tried various combinations to be sure we were not underplaying this technique.
- the  $Z_n^2$  and  $H$  statistics (Equations 20 and 10).

The results are plotted in Figure 2. As the plot shows, the  $Z_n^2$  statistic has equivalent sensitivity to the EF statistic in the signal-dominated regime, with the advantage that it allows to classify the signals based on their harmonic content. However, the  $Z_n^2$  is *more* sensitive at low significance. This is because the noise spectrum of  $Z_n^2$  is less noisy (For  $2n < n_{\text{bin}}$ , which is always the case), and a more or less equivalent signal power will “stand out” more clearly against the noise.

By using the  $H$  test, we can run the  $Z_n^2$  for multiple values of  $n$  and obtain the best number of harmonics to describe the profile.

Thanks to these tests, we are confident that the method described in this Section, with the formulation described in Section 2, allows one to apply the  $Z_n^2$  and  $H$  statistics to pulsar searches conducted with a non-counting instrument producing data with normally distributed uncertainties, for example in the radio or optical band.

## 5. APPLYING THE METHOD TO REAL DATA

The final and most important test is applying the methods described in Section 3 to real astronomical data. For this, we used data from a test observation of PSR B0331+45 taken at the Sardinia Radio Telescope (SRT; Prandoni et al. 2017). PSR B0331+45 (Dewey et al. 1985) is a relatively slow pulsar ( $p_{\text{spin}} \sim 269$  ms), with a very sharp profile. We selected a 30-min observation at L-band (1300-1800 MHz) performed on UT 2016-01-09 as part of routine calibration procedures for a pulsar survey. The data were acquired using the pulsar DFB3 backend<sup>4</sup>) with a frequency resolution of 2 MHz and a time resolution of 128 ms.

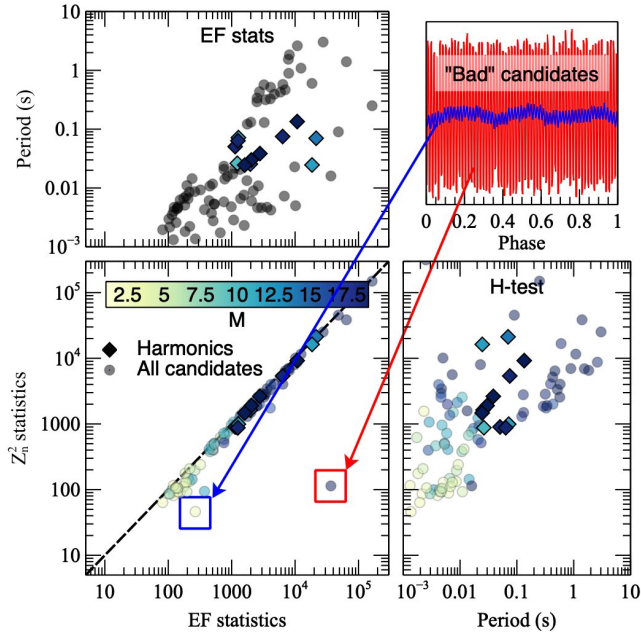
### 5.1. Data processing and pulsar search

We analyzed the data with PULSAR\_MINER<sup>5</sup>, a pipeline based on PRESTO (Ransom 2011; Ransom et al. 2002). The observation was plagued by strong RFI, both impulsive and periodic, broad- and narrow-band, which could not be eliminated completely using PRESTO’s tool `rfifind`. A standard blind search was performed, under the hypothesis that we did not know anything about the pulsar we are looking for. The data were dedispersed with 980 dispersion measure (DM) trials, in the range 2 to 100  $\text{pc cm}^{-3}$ . We then used `accelsearch` to search for periodicities from 1 ms to 20 s within each dedispersed time series. The candidate periods were sifted in order to purge them from the most prominent false positives. The final candidates from the search are also saved as text files (with extension `.bestprof`), containing the pulse profile and information on the detection including the period and period derivatives, the reference time, the detection significance, an estimate of the noise of the profile, etc. Finally, we applied the methods described in Section 3 to the profiles of the candidates obtained from this search.

### 5.2. Estimating the noise

<sup>4</sup> [www.jb.man.ac.uk/pulsar/observing/DFB.pdf](http://www.jb.man.ac.uk/pulsar/observing/DFB.pdf)

<sup>5</sup> [https://github.com/alex88ridolfi/PULSAR\\_MINER](https://github.com/alex88ridolfi/PULSAR_MINER)



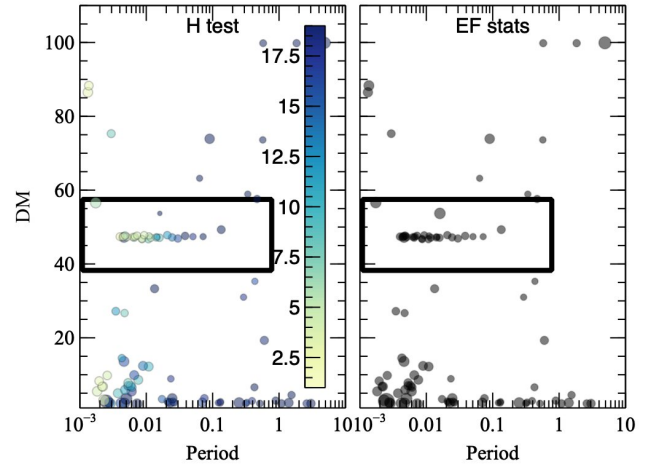
**Figure 3.** Raw values of the statistics from  $Z_{n,\text{gauss}}$  and  $\mathcal{S}$  for a pulsar search at the Sardinia Radio Telescope. Colors indicate the number of harmonics corresponding to the best detection in the  $Z_n^2$  search, and diamonds indicate the harmonics of pulsar candidates (we maintained the color coding of these candidates in the EF panel for an easier comparison). The  $\mathcal{S}$  statistic shows significantly higher values of significance only for signals with highly distorted pulse shapes, as expected, due to the high number of harmonics needed to describe such shapes. Two examples of such signals are shown, folded, in the upper right panel (note that both pulses have a fast square wave from RFI superimposed to the profile).

Arguably the most important part of applying the method is an estimate of the profile noise,  $\sigma_p^2$ . The standard deviation of the profile includes contributions from both the random noise of the data and the signal variability. Separating these two components from a folded profile is in general complicated and requires subtraction of a model of the profile. However, PRESTO correctly calculates the noise a priori, from the variability of the data before folding. We recommend that people using different software, in particular custom-made software, do a similar procedure to calculate the standard deviation from Eq. 5 or, equivalently, the profile variance  $\text{VAR}(p)$  starting from the data variance  $\text{VAR}(d)$ :

$$\text{VAR}(p) = \text{VAR}(d) \frac{N_{\text{samples}}}{N_{\text{bin}}} \quad (29)$$

where  $N_{\text{samples}}$  is the number of samples in the light curve.

### 5.3. Analysis of search candidates



**Figure 4.** Standard DM versus period plot of the candidates of the pulsar search in Figure 3, using  $\mathcal{S}$  and  $Z_{n,\text{gauss}}$ . The signals being highly significant, the significance with the two methods is mostly equivalent. However, the  $Z_{n,\text{gauss}}$  allows to easily classify the candidates in terms of shape, and penalizes profiles containing strong square-wave RFI.

Finally, we applied the statistic to the pulsed profiles obtained in Section 5.1. In Figures 3 and 4 we show the results. As expected, similarly to what was found using simulated data, the values of  $Z_{n,\text{gauss}}$  and  $\mathcal{S}$  are generally very close for candidates with high SNR. We checked the candidates where this was not true and found that, invariably, these candidates had very distorted pulsed profiles, usually with the profile being or containing a square wave (Figure 3). This tells us that a comparison between  $\mathcal{S}$  and  $Z_{n,\text{gauss}}$  statistics is a good diagnostic to eliminate some of the strong RFI still present among the sifted candidates. Figure 4 shows the candidates in a DM vs period plot, where the color-coding represents the number of harmonics needed to describe the profile, while the size represents the significance of the candidates as calculated using the H test. This, again, turns out to be a useful way to plot candidates: the candidates corresponding to higher harmonics of the pulsar show up at more and more sinusoidal candidates at the same DM.

## 6. CONCLUSIONS

In this paper, we outlined a method to apply statistical tests developed for counting experiments to the folded profiles of radio pulsars. We did this in two steps: 1) we showed that the  $Z_n^2$  statistic can be easily applied to folded profiles from high-energy pulsars, obtained by counting the events falling at different pulse phases; 2) we applied a normalization to radio pulsar profiles that preserves the signal-to-noise ratio and creates profiles with the correct statistical properties to apply the binned

version of  $Z_n^2$  developed at step 1. We demonstrate how the  $Z_{n,\text{gauss}}$  statistic can be a great tool to characterize the candidates from radio pulsar searches, being at once more sensitive to low-significance signals and better at discriminating between pulse shapes. We applied the method to a real pulsar search, showing how the method is also a great tool to eliminate RFI candidates. The  $Z_{n,\text{gauss}}$  statistic is available in the open-source software packages for astronomical time series analysis PRESTO (Ransom 2011), Stingray (Huppenkothen et al. 2016, 2019), and HENDRICS (Bachetti 2018).

## ACKNOWLEDGMENTS

This work was conducted in the framework of CLOPS – Citizen Computing Pulsar Search, a project supported by *POR FESR Sardegna 2014 – 2020 Asse 1 Azione 1.1.3* (code RICERCA\_1C-181), call for proposal "Aiuti per Progetti di Ricerca e Sviluppo 2017" managed by Sardegna Ricerche. The authors wish to thank the referee for useful comments, Paul Demorest and Walter Brisken for the development of PRESTO's effective degree of freedom correction described in Appendix E. D.H. is supported by the Women In Science Excel (WISE) programme of the Netherlands Organisation for Scientific Research (NWO). AR gratefully acknowledges financial support by the research grant "iPeska" (P.I. Andrea Possenti) funded under the INAF national call Prin-SKA/CTA approved with the Presidential Decree 70/2016, and continuing valuable support from the Max-Planck Society.

*Software:* Stingray (Huppenkothen et al. 2016, 2019), HENDRICS (Bachetti 2018), astropy (Astropy Collaboration et al. 2018), PINT (Luo et al. 2019), ATNF pulsar catalogue (Manchester et al. 2005), PRESTO (Ransom 2011), Veusz<sup>6</sup>

## REFERENCES

- Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, *The Astronomical Journal*, 156, 123, doi: [10.3847/1538-3881/aabc4f](https://doi.org/10.3847/1538-3881/aabc4f)
- Bachetti, M. 2018, *Astrophysics Source Code Library*, ascl:1805.019
- Buccheri, R., Bennett, K., Bignami, G. F., et al. 1983, *A&A*, 128, 245. [http://adsabs.harvard.edu/cgi-bin/nph-data\\_query?bibcode=1983A%26A...128..245B&link\\_type=ABSTRACT](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1983A%26A...128..245B&link_type=ABSTRACT)
- de Jager, O. C., & Büsching, I. 2010, *Astronomy and Astrophysics*, 517, L9, doi: [10.1051/0004-6361/201014362](https://doi.org/10.1051/0004-6361/201014362)
- de Jager, O. C., Raubenheimer, B. C., & Swanepoel, J. W. H. 1989, 221, 180. [http://adsabs.harvard.edu/cgi-bin/nph-data\\_query?bibcode=1989A%26A...221..180D&link\\_type=ABSTRACT](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1989A%26A...221..180D&link_type=ABSTRACT)
- Dewey, R. J., Taylor, J. H., Weisberg, J. M., & Stokes, G. H. 1985, *The Astrophysical Journal Letters*, 294, L25, doi: [10.1086/184502](https://doi.org/10.1086/184502)
- Gibson, A. I., Harrison, A. B., Kirkman, I. W., et al. 1982, *Nature*, 296, 833, doi: [10.1038/296833a0](https://doi.org/10.1038/296833a0)
- Huppenkothen, D., Bachetti, M., Stevens, A. L., Migliari, S., & Balm, P. 2016, *Astrophysics Source Code Library*, ascl:1608.001
- Huppenkothen, D., Bachetti, M., Stevens, A. L., et al. 2019, *ApJ*, 881, 39, doi: [10.3847/1538-4357/ab258d](https://doi.org/10.3847/1538-4357/ab258d)
- Leahy, D. A. 1987, 180, 275. [http://adsabs.harvard.edu/cgi-bin/nph-data\\_query?bibcode=1987A%26A...180..275L&link\\_type=ABSTRACT](http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1987A%26A...180..275L&link_type=ABSTRACT)
- Leahy, D. A., Darbro, W., Elsner, R. F., et al. 1983a, *ApJ*, 266, 160, doi: [10.1086/160766](https://doi.org/10.1086/160766)
- Leahy, D. A., Elsner, R. F., & Weisskopf, M. C. 1983b, *ApJ*, 272, 256, doi: [10.1086/161288](https://doi.org/10.1086/161288)
- Lorimer, D. R., & Kramer, M. 2012, *Handbook of Pulsar Astronomy*
- Luo, J., Ransom, S., Demorest, P., et al. 2019, *Astrophysics Source Code Library*, ascl:1902.007. <http://adsabs.harvard.edu/abs/2019ascl.soft02007L>
- Manchester, R. N., Hobbs, G. B., Teoh, A., & Hobbs, M. 2005, *The Astronomical Journal*, 129, 1993, doi: [10.1086/428488](https://doi.org/10.1086/428488)
- Mardia, K. V. 1975, *Journal of the Royal Statistical Society: Series B (Methodological)*, 37, 349, doi: [10.1111/j.2517-6161.1975.tb01550.x](https://doi.org/10.1111/j.2517-6161.1975.tb01550.x)

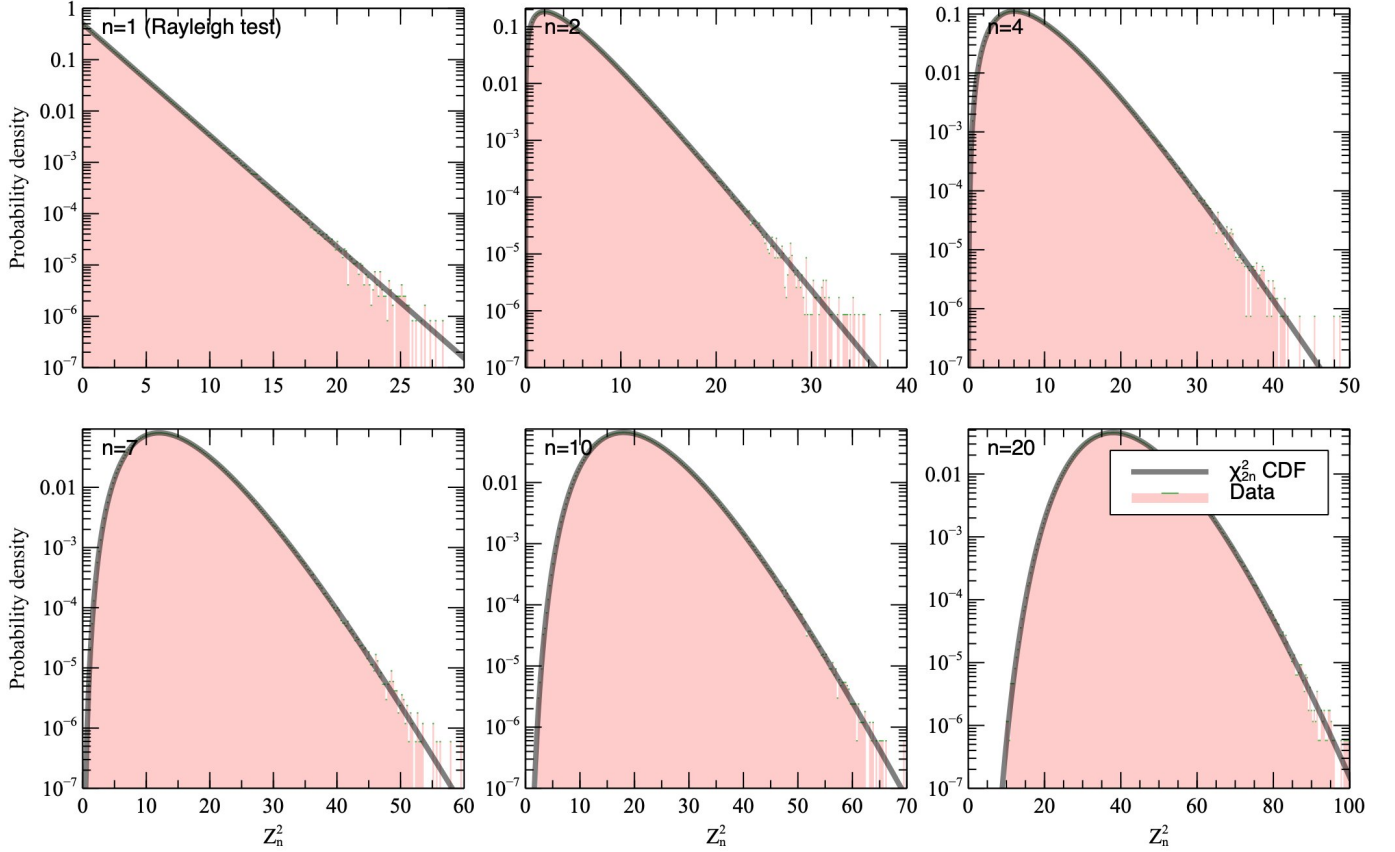
<sup>6</sup> <https://veusz.github.io>

Prandoni, I., Murgia, M., Tarchi, A., et al. 2017, *A&A*, 608,  
A40, doi: [10.1051/0004-6361/201630243](https://doi.org/10.1051/0004-6361/201630243)

Ransom, S. 2011, *Astrophysics Source Code Library*,  
ascl:1107.017

Ransom, S. M., Eikenberry, S. S., & Middleditch, J. 2002,  
astro-ph

van der Klis, M. 1989, in *Timing Neutron Stars: Proceedings  
of the NATO Advanced Study Institute on Timing  
Neutron Stars Held April 4-15, 27*



**Figure 5.** Statistical distribution of the  $Z_n^2$  values for white noise data, with different values of  $n$ , compared to the expected  $\chi_{2n}^2$  distributions. We only selected values of  $Z_n^2$  obtained starting from pulsed profiles with least  $2n$  bins (one needs at least two bins to contain a sinusoidal harmonic; this is analogous to the Nyquist theorem for the Fourier Transform).

## APPENDIX

### A. STATISTICAL PROPERTIES OF THE BINNED Z STATISTIC

We verify here that the  $Z_{n,\text{bin}}^2$  values agree closely with the expected  $\chi_{2n}^2$  distribution in the case of white noise.

To measure this agreement, we used 10,000,000 simulations of binned profiles with different numbers of bins (between 4 and 1024) and total number of photons (between 10 and 10,000). For each value of  $n$ , we selected all simulations with  $N_{\text{bin}} > 2n$  (as per the Nyquist argument in Section 2). Then, we compared the distribution of the  $Z_{n,\text{bin}}^2$  values with the expected  $\chi_{2n}^2$  distribution, in particular in the tails where the values are more important for significance estimates. From this first test, it is clear that the approximation yields the expected probability distribution for  $Z_n^2$  for all values of  $n$  if  $N_{\text{bin}} > 2n$  (see Figure 5).

### B. WHAT IS THE DIFFERENCE BETWEEN THE Z STATISTIC FOR BINNED AND UNBINNED EVENT DATA?

The difference between the  $Z_n^2$  formulations in Equations 8 and 11 is the fact that the phase of each event is approximated with the phase at the middle of a profile bin. Looking at Eq. 8, we see that it is of the form

$$Z_n^2 = \frac{2}{N} \sum_{k=1}^n [A_k^2 + B_k^2] \quad (\text{B1})$$

where

$$A_k = \sum_{j=1}^N \cos k\phi_j \quad (\text{B2})$$

$$B_k = \sum_{j=1}^N \sin k\phi_j \quad (\text{B3})$$

Now, let us express each phase as the sum of the phase at the center of each bin and an additional (usually small) term

$$\phi_j = \phi_{b(j)} + \epsilon_j \quad (\text{B4})$$

Using the standard trigonometric formulas for the cosine and sine of sums of angles,  $A_k$  and  $B_k$  become

$$A_k = \sum_{b=1}^{N_{\text{bin}}} \left[ \cos k\phi_b \sum_{i=1}^{w_b} \cos k\epsilon_i - \sin k\phi_b \sum_{i=1}^{w_b} \sin k\epsilon_i \right] \quad (\text{B5})$$

$$B_k = \sum_{b=1}^{N_{\text{bin}}} \left[ \sin k\phi_b \sum_{i=1}^{w_b} \cos k\epsilon_i + \cos k\phi_b \sum_{i=1}^{w_b} \sin k\epsilon_i \right] \quad (\text{B6})$$

Let us now take a closer look at the inner sums. We are referring the  $\epsilon_i$  to the center of the profile bins, so it is reasonable to assume that, in general, they will be equally distributed in a range  $-\pi/N_{\text{bin}} \leq \epsilon_i \leq \pi/N_{\text{bin}}$  around the bin centers<sup>7</sup>. Therefore,

$$\sum_{i=1}^{w_b} \sin k\epsilon_i \approx 0 \quad (\text{B7})$$

and

$$\sum_{i=1}^{w_b} \cos k\epsilon_i \approx w_b \text{mean}(\cos k\epsilon_i) \quad (\text{B8})$$

$$= w_b \frac{1}{2\alpha_k} \int_{-\alpha_k}^{\alpha_k} \cos k\epsilon_i d\epsilon_i \quad (\text{B9})$$

where

$$\alpha_k = \frac{\pi}{N_{\text{bin}}}$$

so that

$$\sum_{i=1}^{w_b} \cos k\epsilon_i \approx w_b \frac{\sin k\alpha_k}{k\alpha_k}. \quad (\text{B10})$$

Concluding, Equations B5 and B6 become

$$A \approx \frac{\sin k\alpha_k}{k\alpha_k} \sum_{b=1}^{N_{\text{bin}}} w_b \cos k\phi_b \quad (\text{B11})$$

$$B \approx \frac{\sin k\alpha_k}{k\alpha_k} \sum_{b=1}^{N_{\text{bin}}} w_b \sin k\phi_b \quad (\text{B12})$$

and Eq. B1 becomes

$$Z_n^2 \approx \frac{2}{N} \sum_{k=1}^n \left( \frac{\sin k\alpha_k}{k\alpha_k} \right)^2 \left[ \left( \sum_{b=1}^{N_{\text{bin}}} w_b \cos k\phi_b \right)^2 + \left( \sum_{b=1}^{N_{\text{bin}}} w_b \sin k\phi_b \right)^2 \right] \quad (\text{B13})$$

<sup>7</sup> This assumption breaks down for pulse profiles with very sharp features well above the noise level, but this implies high significance in any case

Comparing it to Eq. 11, we see that this formula only differs by the factor

$$C(k) = \left( \frac{\sin k\alpha_k}{k\alpha_k} \right)^2 \quad (\text{B14})$$

$$= \left( \frac{N_{\text{bin}}}{\pi k} \sin \frac{\pi k}{N_{\text{bin}}} \right)^2 \quad (\text{B15})$$

that goes to 1 for  $k/N_{\text{bin}} \ll 1$ , as expected, and drops to  $(2/\pi)^2 = 0.405$  for  $k = N_{\text{bin}}/2$ .

This is the  $Z_n^2$  equivalent of the loss of sensitivity of Fourier Power Density Spectra (PDS) at high frequencies due to sampling: a light curve of length  $T$  sampled with  $N$  bins is equivalent to convolving a continuous light curve with a ‘‘binning window’’, non-zero only between times  $t - 0.5\Delta t$  and  $t + 0.5\Delta t$ , where  $\Delta t = T/N$  is the sampling time. The Fourier transform of this sampling window is (See eq. 2.19 from [van der Klis 1989](#))

$$B(f) = \frac{\sin \pi f T / N}{\pi f T / N} \quad (\text{B16})$$

and its square, similarly to Eq. B15, goes from a maximum of 1 at low frequencies to  $(2/\pi)^2$  at the Nyquist frequency, acting as a low-pass filter.

### C. THE BINNED $\mathcal{S}$ OF A COMPOSITION OF SINUSOIDAL SIGNALS

Let us consider a pulsed profile composed of a sum of  $l$  sinusoidal harmonics with random phases  $\phi_{0,l}$ . Neglecting binning effects (i.e. choosing a number of bins  $N_{\text{bin}} \gg m$ ), the profile can be described as:

$$p_i = \lambda \left[ 1 + \sum_{l=1}^m a_l \sin l(\phi_i - \phi_{0,l}) \right] \quad (\text{C17})$$

The  $\mathcal{S}$  statistic in Eq. 7 gives

$$\mathcal{S} = \frac{1}{\lambda} \sum_{i=1}^{N_{\text{bin}}} (p_i - \lambda)^2 \quad (\text{C18})$$

$$= \frac{1}{\lambda} \sum_{i=1}^{N_{\text{bin}}} \left[ \sum_{l=1}^m a_l \sin l(\phi_i - \phi_{0,l}) \right]^2 \quad (\text{C19})$$

$$= \frac{1}{\lambda} \sum_{i=1}^{N_{\text{bin}}} \left[ \sum_{l=1}^m a_l \sin l\phi_i \cos l\phi_{0,l} - \sin l\phi_{0,l} \cos l\phi_i \right]^2 \quad (\text{C20})$$

We now manipulate the trigonometric functions and note that: the sum of any odd power of sinusoidal functions integrates to zero over a full cycle, so that only the terms with squares survive; the sinusoidal harmonics form an orthogonal base, so that any integral of products of different harmonics integrates to zero as well. Then, we get

$$\mathcal{S} = \frac{1}{\lambda} \sum_{l=1}^m a_l \left( \cos^2 l\phi_{0,l} \sum_{i=1}^{N_{\text{bin}}} \sin^2 l\phi_i + \sin^2 l\phi_{0,l} \sum_{i=1}^{N_{\text{bin}}} \cos^2 l\phi_i \right) \quad (\text{C21})$$

The integral of the square of a sinusoidal harmonic over a pulse profile gives  $N_{\text{bin}}/2$ , so it is easy to obtain

$$\mathcal{S}_{\text{sig}} \approx \frac{\lambda N_{\text{bin}}}{2} \sum_{l=1}^n a_l^2 \quad (\text{C22})$$

### D. THE BINNED $Z$ STATISTIC OF A COMPOSITION OF SINUSOIDAL SIGNALS

Again, we start from the profile in Eq. C17 and calculate  $Z_{n,\text{bin}}^2$ . We assume that  $n \geq m$ , i.e. the  $Z_{n,\text{bin}}^2$  formula contains enough harmonics to describe the signal completely. Also,  $N_{\text{bin}} \gg n$ , so that we can neglect the effect of binning.

As we did previously, we divide the  $Z_{n,\text{bin}}^2$  formula in two terms

$$Z_{n,\text{bin}}^2 = \frac{2}{\sum_i p_i} \sum_{k=1}^n [A_k^2 + B_k^2] \quad (\text{D23})$$

where we use the profile bins  $p_i$  as weights ( $p_i$  in Eq. 11) and

$$A_k = \lambda \sum_{i=1}^{N_{\text{bin}}} \cos k\phi_i \left[ 1 + \sum_l a_l \sin l(\phi_i - \phi_{0,l}) \right] \quad (\text{D24})$$

$$B_k = \lambda \sum_{i=1}^{N_{\text{bin}}} \sin k\phi_i \left[ 1 + \sum_l a_l \sin l(\phi_i - \phi_{0,l}) \right] \quad (\text{D25})$$

Applying the same arguments of Appendix C (all sums of odd powers of sinusoidal functions over a full profile go to zero, orthogonality of sinusoidal harmonics), it is easy to get

$$\begin{aligned} A_k &= \lambda \sum_i \cos k\phi_i \sum_l a_l (\sin l\phi_i \cos l\phi_{0,l} - \cos l\phi_i \sin l\phi_{0,l}) \\ &= \lambda a_k \sum_i \cos k\phi_i (\sin k\phi_i \cos k\phi_{0,k} - \cos k\phi_i \sin k\phi_{0,k}) \\ &= \lambda a_k \left[ \cos k\phi_{0,k} \sum_i \cos k\phi_i \sin k\phi_i - \sin k\phi_{0,k} \sum_i \cos^2 k\phi_i \right] \\ &= -\frac{N_{\text{bin}}\lambda}{2} a_k \sin k\phi_{0,k} \end{aligned} \quad (\text{D26})$$

$$B_k = -\frac{N_{\text{bin}}\lambda}{2} a_k \cos k\phi_{0,k} \quad (\text{D27})$$

And finally

$$Z_{n,\text{bin}}^2 = \frac{2}{N_{\text{bin}}\lambda} \sum_k \left( \frac{N_{\text{bin}}\lambda a_k^2}{2} \right)^2 \left[ (\sin k\phi_{0,k})^2 + (\cos k\phi_{0,k})^2 \right] \quad (\text{D28})$$

$$= \frac{N_{\text{bin}}}{2\lambda} \sum_{k=1}^n a_k^2 \quad (\text{D29})$$

## E. INTER-BIN PULSE PROFILE CORRELATIONS DUE TO FOLDING METHODOLOGY

As discussed in §1 near Eq. 2, there is an alternative method to fold a time series that is arguably better when the time series bins are comprised of integrated samples or events. In that case, instead of each time series bin being treated as a delta function in time, which can be placed at a specific phase in a pulse profile, each time series bin has a beginning and end in both time and in pulse phase. The time series values are then spread proportionally over the corresponding pulse profile bins, effectively “drizzling” the integrated data over the corresponding portions of the accumulating pulse profile. This is, in fact, how the program `prepfold` from PRESTO folds data.

Because the time series values are spread proportionally, neighboring bins of the pulse profile end up slightly correlated. The amount of correlation depends on the ratio of the duration in phase of a pulse profile bin,  $\Delta\phi_p = 1/N_{\text{bin}}$ , to the duration of a time series bin  $\Delta t$  in units of pulse phase,  $\Delta\phi_t = \Delta t/P$ , where  $P$  is the folding period. The ratio, which we will call

$$\psi = \frac{\Delta\phi_p}{\Delta\phi_t} = \frac{P}{N_{\text{bin}}\Delta t}, \quad (\text{E30})$$

describes how much time resolution you have in your time series compared to the rapidity of the searched-for pulsations.

In the limit of small  $\psi$  (i.e. values near zero), the pulse profile bins become perfectly correlated (and will therefore never show any pulsations), whereas for large  $\psi$ , the pulse profile bins are basically uncorrelated and we effectively

reproduce the folding methodology of Eq. 2. Much pulsar searching happens in the middle range where  $\psi$  is between 1–100.

The result of these correlations is that the effective number of degrees of freedom  $\nu$  in the folded profile, for a  $\chi^2$  sensitivity calculation, for instance, decreases below  $N_{\text{bin}} - 1$ . Alternatively, you can think of the effect as decreasing the variance in, or a smoothing of, the resulting pulse profile.

We have developed a semi-analytic correction  $C$  (with help from Paul Demorest and Walter Brisken) to the relevant statistics due to the correlations of the form:

$$C = a\psi (1 + \psi^b)^{(-1/b)}, \quad (\text{E31})$$

so that the effective number of degrees of freedom  $\nu_{\text{eff}}$  to use in statistical tests is  $\nu_{\text{eff}} = C\nu = C(N_{\text{bin}} - 1)$ .

We performed a large number of simulations over a wide range of  $\psi$ , using time series of pure Gaussian noise and the folding code in PRESTO, to determine both the validity of Eq. E31 as well as the values of  $a$  and  $b$ :  $a = 0.96$  and  $b = 1.806$ . The correction is good to a fractional error of less than a few percent as long as  $\psi \gtrsim 0.5$ . There is a small dependence on  $N_{\text{bin}}$  which becomes apparent when  $\psi \lesssim 0.7$ .

To correct the measured noise level  $\sigma_{\text{meas}}$  in the profile (for estimating a signal-to-noise ratio or flux density via the radiometer equation, for example), dividing by the square-root of  $C$  will inflate  $\sigma_{\text{meas}}$  appropriately:  $\sigma_{\text{corr}} = \sigma_{\text{meas}}/\sqrt{C}$ , where  $\sigma_{\text{corr}}$  is the corrected standard deviation of the profile noise.