MNRAS **000**, 1–10 (2017) Preprint 13 April 2017 Compiled using MNRAS LATEX style file v3.0

Methods for detecting variability in noisy data

Krzysztof Suberlak,[1]★ Željko Ivezić[1]
[1]*Department of Astronomy, University of Washington, Seattle, WA, United States*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We compare various methods used to detect variability in time and frequency domain. In time domain, we consider the how well can a χ^2 test distinguish noise from signal (AstroML 10.1.1). Then we test the maximum-likelihood method of parameter estimation for a Gaussian Distribution (AstroML 5.6.1). Defining a signal to noise ratio of variability detection in a given time series improves the quality of our prediction. In frequency domain, we assess what is the weakest signal detectable by methods related to calculating power over frequency grid (periodogram, power spectral distribution). We consider what influences the accuracy of variability detection in frequency domain - how well can we tackle irregularly sampled time series with heteroscedastic, uncorrelated errors, or what frequency grid to choose.

1 INTRODUCTION

2 MOTIVATION

3 TIME DOMAIN METHODS

3.1 χ^2 test : can we distinguish variability from noise?

What is the minimum variability amplitude that we can measure? How can we distinguish pure noise from a genuine intrinsic variability? In this section we derive the variance of a χ^2_{DOF} distribution, to which we often compare the observed data to test whether it is consistent with Gaussian noise. We find that for N data points, the standard deviation of χ^2_{DOF} is $\sqrt{2/N}$ (eq. 10). We then show that for a time-series with zero mean, and homoscedastic errors $e_i = \sigma$, we can express its variance in terms of a χ^2_{DOF} distribution (eq. 9). We use this result in a concrete case of a harmonic function, for which variance is known analytically (eq. 12), and requiring a 3σ detection, we compare eqs. 9 and 13 to find the smallest detectable amplitude using the χ^2_{DOF} method. This becomes an introduction to the method of more advanced Bayesian parameter estimation (see chapter 5.6.1 in Ivezic+2014). Calculating a full posterior pdf we describe the distribution of underlying variability by σ_0 (Fig. 5) . It turns out that we can characterize the quality of detection by calculating the signal-to-noise ratio for the posterior pdf (Fig. 8) . This method is able to robustly detect much smaller amplitudes than traditional χ^2_{DOF} (Fig. 10)

A simplest way is to compare our time series to a χ^2 distribution. If observed points $\{x_i\}$ are drawn from a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, then with $z_i = x_i - \mu/\sigma$ the quantity $Q = \sum_{i=1}^{N} z_i^2$, follows a χ^2 distribution with N degrees of freedom (eq. 3.58 AstroML) [text too much from there : change !] Often one defines a χ^2 per degree of freedom : $\chi^2_{DOF} = \chi^2(Q/N)$. The mean value of χ^2_{DOF} is 1. For a continuous smoothly varying probability function that describes the parent distribution, the mean is the first raw moment:

$$\mu = \int_{-\infty}^{\infty} xp(x)dx \tag{1}$$

and the variance is the second raw moment:

$$Var = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \int_{-\infty}^{\infty} x^2 p(x)dx - \mu^2 \tag{2}$$

We denote by $\langle x \rangle$ the expectation value of x. Thus the above can be written as : $Var = \sigma^2 = \langle x^2 \rangle - \langle x \rangle^2$.

The standard deviation of $\chi^2_{DOF} = \sqrt{2/N}$. We can prove by first considering the variance of χ^2_{DOF}:

$$Var(\chi^2_{DOF}) = Var\left(\frac{1}{N}\sum \frac{x_i^2}{\sigma^2}\right) = Var\left(\frac{v}{N\sigma^2}\right) \tag{3}$$

Now $Var(\alpha x) = \alpha^2 Var(x)$. Thus

$$Var\left(\frac{v}{N\sigma^2}\right) = \frac{Var(v)}{N^2\sigma^4} \tag{4}$$

$Var(v) = \langle v^2 \rangle - \langle v \rangle^2$. We find first the second component: the mean $\langle v \rangle = \langle \sum x_i^2 \rangle = \sum \langle x_i^2 \rangle$ (no cross terms). And since $Var(x_i) = \langle x_i^2 \rangle = \sigma^2$, $\langle v \rangle = \sum \sigma^2 = N\sigma^2$, so that $\langle v \rangle^2 = N^2\sigma^4$.

The first component is less straightforward:

$$\langle v^2 \rangle = \langle (\sum x_i^2)^2 \rangle = \langle \sum x_i^2 \sum x_j^2 \rangle = \sum\sum \langle x_i^2 x_j^2 \rangle \tag{5}$$

The expectation value of $x_i^2 x_j^2$ involves N center terms and $N^2 - N$ cross terms :

$(x_1^2+x_2^2+x_3^2+...)(x_1^2+x_2^2+x_3^2+...) = x_1^4+x_2^4+...+x_1^2x_2^2+x_1^2x_3^2...$

Thus $\langle x_i^2 x_j^2 \rangle = (N^2 - N) \int x^2 P(x)dx \int x^2 P(x)dx + N \int x^4 P(x)dx$.

Since $A = 0$, $P(x)$ is a Gaussian.
We evaluate

$$\int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} dx \tag{6}$$

substituting $u = x/\sqrt{2}\sigma$ and use the standard result $\int u^2 \exp(-\alpha u^2) du = \sqrt(\pi)/2\alpha^{3/2}$, so that : $\int x^2 P(x) dx = \sigma^2$.
similarly, for

$$\int_{-\infty}^{\infty} \frac{x^4}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} dx \tag{7}$$

with identical substitution and the standard result $\int u^4 \exp(-\alpha u^2) du = 3\sqrt(\pi)/4\alpha^{5/2}$ we obtain : $\int x^4 P(x) dx = 3\sigma^4$.
Therefore:

$$\langle v^2 \rangle = (N^2 - N)\sigma^4 + N3\sigma^4 = N^2\sigma^4 + 2N\sigma^4 \tag{8}$$

Finally,

$$Var\left(\frac{v}{N\sigma^2}\right) = \frac{Var(v)}{N^2\sigma^4} = \frac{1}{N^2\sigma^4}(N^2\sigma^4 + 2N\sigma^4 - N^2\sigma^4) = \frac{2}{N} \tag{9}$$

Thus the standard deviation of χ^2_{DOF} , being the square root of variance by definition, is equal to $\sqrt{2/N}$.

$$\sigma\left(\chi^2_{DOF}\right) = \sqrt{2/N} \tag{10}$$

Therefore, if the observed points x_i originate from a Gaussian distribution, then we would expect χ^2_{DOF} to be centered around 1, with a standard deviation $\sqrt{2/N}$. This means that the plot of χ^2_{DOF} will get narrower as we sample more points from the original series. For $N \to \infty$, $\chi^2_{DOF} \to \delta(1)$.

Now there is another general result worth mentioning. Namely, for a time series with mean $\langle x \rangle = 0$ (such as x_i), $z_i = x_i - \mu/\sigma = x_i/\sigma$, so that $\chi^2_{DOF} = \sum_{i=1}^{N} (1/N) z_i^2 = \sum_{i=1}^{N} = \langle x_i^2 \rangle/\sigma^2$. Now since $Var(x) = \langle x^2 \rangle - \langle x \rangle^2$, this means that for a 0 mean time series,

$$\chi^2_{DOF} = Var/\sigma^2 \tag{11}$$

Therefore we can express the variance of a concrete time series in terms of χ^2_{DOF}.

Consider a harmonic time series. For $y(t) = A(\sin(\omega t)$, with Gaussian errors $\epsilon \sim \mathcal{N}(0, \sigma)$, sampled by $N = 100$ points, $x_i = y_i + \epsilon_i$, the variance is $Var(x) = Var(y + \epsilon) = Var(y) + Var(\epsilon) = Var(A\sin(x)) + \sigma^2$. Now from the definition, $Var(A\sin(x)) = \langle A^2 \sin^2(x) \rangle - \langle A \sin(x) \rangle^2 = A^2/2$, because $\langle \sin^2(x) \rangle = 1/2$ and $\langle sin(x) \rangle = 0$ due to symmetry. Thus for a sinusoidal time series,

$$Var(x) = A^2/2 + \sigma^2 \tag{12}$$

Given that variance of time series, we can usefully link the χ^2_{DOF} to the amplitude of time series by the above result. We can ask what is the amplitude for which χ^2_{DOF} would have over three sigma departure from 1. How significant is such departure? The probability that χ^2_{DOF} departs by more than 3 standard deviations from the mean is 0.001.

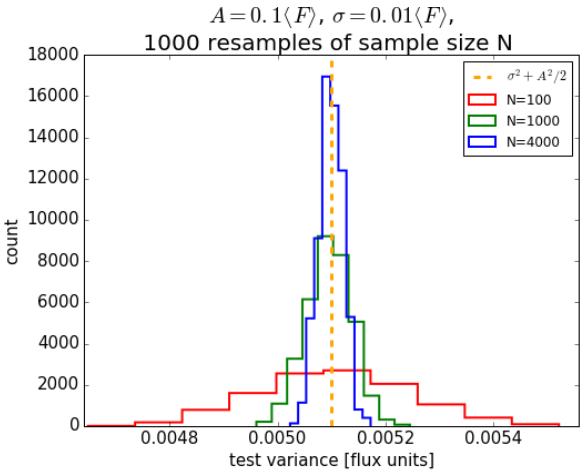

Figure 1. We simulated the $y = A\sin(t) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma)$ time series drawing N samples at equal time intervals. For all realizations, $A = 0.1$, and $\sigma = 0.1A$. For each N, we perform 1000 realizations of the time series, for which we calculate the variance. The vertical orange dashed line is the theoretical approximation of $Var(y)$ by $V_{theory} = A^2/2 + \sigma^2$. We see that the higher the number of points (samplings from the parent distribution), the smaller spread around the theoretical value, i.e. the better the approximation. Note that the histogram is centered on V_{theory}, spanning only 9% of V_{theory} in each direction along x-axis.

Carrying on the calculation, if we require at least $\chi^2_{DOF} = 1 + 3\sigma(\chi^2_{DOF})$, i.e.

$$\chi^2_{DOF} = 1 + 3\sqrt{2/N} \tag{13}$$

and using $\chi^2_{DOF} = Var(x)/\sigma^2$ (eq.9), we have $Var(x)/\sigma^2 = 1 + 3\sqrt{2/N}$. Since for the sinusoidal time series, $Var(x) = A^2/2 + \sigma^2$ (eq. 12), we have:

$$\frac{A^2/2 + \sigma^2}{\sigma^2} > 1 + 3\sqrt{2/N} \tag{14}$$

thus $A^2 > 6\sqrt{2}\sigma^2/\sqrt{N}$, so that $A > 2.9\sigma/N^{1/4}$. Therefore, the minimum detectable amplitude at the 3σ level is

$$A_{min} = \frac{2.9\sigma}{N^{1/4}} \tag{15}$$

We illustrate these calculations with the following example. We simulate $y = A\sin(t) + \epsilon$, and show that $Var(x)$ is indeed $A^2/2 + \sigma^2$ (Fig. 1) We also show that the χ^2_{DOF} of that time series is $\propto \sqrt{2/N}$ (Fig. 2).

3.2 What we have so far : how good χ^2 can be ?

In assessing the usefulness of considering χ^2 - like quantities, we found in Sec. 3.1 that in order to achieve a 3σ detection for a zero-mean harmonic time series parametrized by $y = A\sin(t) + \epsilon$, we need $A_{min} = \frac{2.9\sigma}{N^{1/4}}$ (Eq. 15). On Fig. 8 we compare that approach to the fully Bayesian calculation of Sec. 3.3, where a quality of detection is described by a signal-to-noise ratio (SN).

Now if we want to distinguish time series from noise,

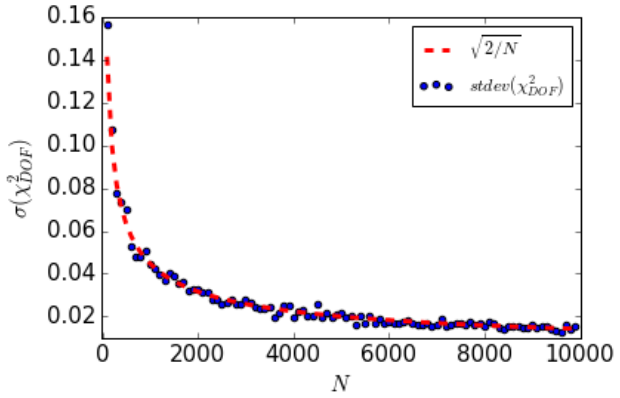

Figure 2. The standard deviation of values of χ^2_{DOF} for time series $y = A\sin(t) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma)$. We verify that the standard deviation for the χ^2_{DOF} of many realizations of time series is proportional to the $\sqrt{2/N}$, where N is the size of the sample. We explore N over the grid of 100 uniformly spaced values between 100 and 9900. For each N we simulate 100 realizations of the time series, and for each realization we calculate the χ^2_{DOF}. Then we calculate the standard deviation of the 100 values of χ^2_{DOF} per N. As expected, the standard deviation of χ^2_{DOF} follows the relation $\chi^2_{DOF} \sim \sqrt{2/N}$, as in Eq. 9

we may use two other approaches coming from information theory, that quantify the difference in information content between the two models; in our case : noise model , and the time series model. The two most common definitions of information content used are Bayesian Information Criterion (BIC), and Akaike Information Criterion (AIC):

$$BIC \equiv -2\ln[L^0(M)] + k\ln(N) \tag{16}$$

$$AIC \equiv -2\ln[L^0(M)] + 2k + \frac{2k(k+1)}{N-k-1} \tag{17}$$

where $L^0(M)$ is the maximum value of the data likelihood, k is the number of model parameters, and N is the number of data points. We prove in Appendix A that $-2\ln[L^0(M)] = \chi^2_{DOF}$, and thus we can compare BIC and AIC of noise and harmonic models. If we require $\Delta BIC = \Delta AIC = 10$, then we can find (see Appendix A), what is the minimum amplitude that can cause such change in information content, and in this way become distinguishable from noise.

Therefore, to detect variability in harmonic time series described by $y = Asin(\omega t)$, we can :

a) require that it's $\chi^2_{DOF,series}$ has at least 3σ departure from pure noise ($\chi^2_{DOF,noise} = 1$), so that we have: $\chi^2_{DOF,series} = V/\sigma^2 = 1 + 3\sqrt{2/N}$, we find:

$$A_{min}/\sigma = 2.9/N^{1/4} \tag{18}$$

b) require that the $\Delta BIC > 10$, so that:

$$A_{min}/\sigma = \left(\frac{6\ln N + 20}{N}\right)^{1/2} \tag{19}$$

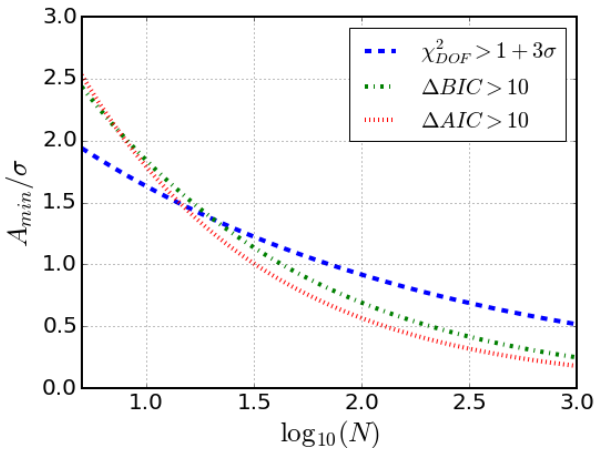

Figure 3. Minimum amplitude required to distinguish a harmonic time series $Asin(\omega t)$ from noise. For $\log_{10} N > 1.5$ ($N > 30$) the BIC and AIC criteria allow a detection of smaller A then a χ^2_{DOF}-based 3σ departure.

c) require that the $\Delta AIC > 10$, so that:

$$A_{min}/\sigma = \sqrt{32/N} \tag{20}$$

A comparison between these three complementary ways of assessing the minimum detectable amplitude is shown on Fig. 3

3.3 Beyond χ^2 : intrinsic variability with Bayesian approach

As described in detail in Chapter 5 of Ivezic+2014, it is possible to estimate parameters for an underlying Gaussian Distribution given the observed data $\{x_i\}$. In summary, there are few ways to think about what kind of distribution $\{x_i\}$, together with associated errors, may come from.

First, $\{x_i\}$ could be measurements of a constant quantity, eg. length of a ruler, and the sought mean μ would correspond to our best estimate of the ruler length. If errors are known and heteroscedastic, then we also measure them individually : σ_i is a measured quantity, and the likelihood for obtaining data $\{x_i\}$ given μ and σ_i is :

$$p(\{x_i\}|\mu, I) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma_i^2}\right) \tag{21}$$

(eq. 5.47). We thus seek a one-dimensional posterior pdf for μ.

Second, if $\{x_i\}$ corresponds to measurements of a constant quantity, where measurement errors are negligible compared to the underlying spread σ of that quantity, we seek a two-dimensional posterior pdf for μ and σ. The likelihood function is a product of Gaussians for each measurement point :

$$p(\{x_i\}|\mu, \sigma, I) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \tag{22}$$

(eq. 5.52)

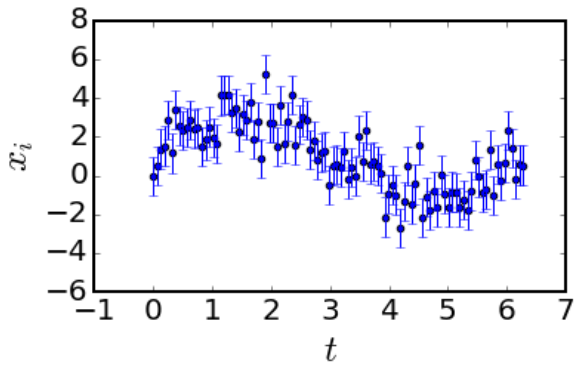

Figure 4. A simulated time series with $N = 100$ evenly spaced in time samples from a process $x(t) = A\sin(t)+\mu+n$, where $\mu = 1$, $A = 2$, noise is $n \sim \mathcal{N}(0, \sigma_0 = 1)$, and all errors are $e = \sigma_0$. Theoretical variance of $x(t)$ is $Var(x) = A^2/2 + \sigma_0^2$, i.e. here $Var(x) = 3$.

Finally, if $\{x_i\}$ corresponds to measurements of a constant quantity, where measurement errors are known, and non-negligible, but they originate from a Gaussian distribution (width of which for each point is given by e_i), we seek a two-dimensional posterior pdf for μ and σ, where σ and e_i are coupled. The likelihood function is, as before, a product of Gaussians for each measurement point, but here the intrinsic spread of the measured quantity σ is coupled to the width of the error distribution e_i:

$$p(\{x_i\}|\mu,\sigma,I) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi(\sigma^2+e_i^2)}} \exp\left(\frac{-(x_i-\mu)^2}{2(\sigma^2+e_i^2)}\right) \quad (23)$$

(eq. 5.63)

We consider $\{x_i\}$ to be measurements of a brightness of an astrophysical source, with an underlying mean μ. Given measurement errors e_i, a detection of nonzero intrinsic σ would mean that the object is more variable than what could be explained solely by measurement errors. It does not explain or describe the nature of variability - whether it is a damped random walk, sinusoid, or some other more complicated periodic or aperiodic behavior. The main question that estimate of σ can answer is whether points $\{x_i\}$ with errors e_i could come from a constant distribution, or are they too spread out for that to be the case.

To test the viability of the method, we simulate a sinusoidal time series, $x(t) = A\sin(t)+\mu+e$, sampling at $N = 100$ equally spaced points, setting $\mu = 1$, $A = 1$, $e \sim \mathcal{N}(0, \sigma_0 = 0.1)$(see Fig. 4). Variance of such time series is theoretically $\sigma^2 = Var(x) = Var(A\sin(t)) + Var(\mathcal{N}) = A^2/2 + \sigma_0^2$ (as derived in text above Eq. 12)

We simulate $N_{sim} = 1000$ of realizations (drawing from random Gaussian noise $\mathcal{N}(0, \sigma_0 = 0.1)$ at each iteration).

One method to distinguish the time series from Gaussian noise is to calculate the χ^2_{dof}, which would converge to $\mathcal{N}(1, \sqrt{2/N})$ for a lack of intrinsic variability (see Fig. 1) . In theory, standard deviation of χ^2_{dof} is $\sqrt{2/N}$, and Fig. 8 illustrates that we need to set amplitude to as high as 2 to distinguish it from noise.

For each realization of time series we calculate the full posterior pdf for σ_0 and μ.

Each pdf is sampled by a grid of 70 points (we call the grid μ_{int} and σ_{int}) with minimum set at $min(\sigma_{int} = 0$, and maximum at $max(max(\sigma_{boot}), \sigma_{st.dev.}(x_i)$, where $max(\sigma_{boot})$ is the maximum of the boostrapped resample of the histogram of σ resulting from the approximate method, and $\sigma_{st.dev.}(x_i)$ is the standard deviation of the input data.

The posterior pdfs $p(\mu_{int})$ and $p(\sigma_{int})$ are shown on Fig. 5, which highlights the differences between the mean $\bar{\sigma_{int}}$, the value of the pdf at that point $p(\bar{\sigma_{int}})$, the maximum value of $max(p(\sigma_{int}))$, and the theoretical standard deviation - a square root of variance of the time series:

$$\bar{\sigma_{int}} = \frac{\sum p(\sigma_{int})\sigma_{int}}{\sum p(\sigma_{int})} \quad (24)$$

$$Var(\sigma_{int}) = \frac{\sum p(\sigma_{int})\sigma_{int}^2}{\sum p(\sigma_{int})} - \bar{\sigma_{int}}^2 \quad (25)$$

We can quantify the accuracy of our estimate of the intrinsic variability (σ_0) by describing the shape of the resulting posterior pdf $p(\sigma)$.

Therefore, for each lightcurve for which we calculate $p(\sigma)$, we also calculate :

- the expectation value of σ (i.e. mean):

$$\langle\sigma_{int}\rangle = \frac{\sum \sigma_{int} p(\sigma_{int})}{\sum p(\sigma_{int})} \quad (26)$$

- standard deviation of σ, $st.dev.(\sigma)$, found from the square root of Variance (see Appendix B) :

$$st.dev.(\sigma) = \left(\frac{\sum \sigma_{int}^2 p(\sigma_{int})}{\sum p(\sigma_{int})} - \langle\sigma_{int}\rangle^2\right)^{1/2} \quad (27)$$

- weighted median(σ) (the weighted 50-th percentile)
- the robust interquartile range assuming Gaussian origin: $\sigma_G = 0.741 * (percentile(75) - percentile(25))$ using weighted percentiles
- assuming $p(\sigma)$ normalized : the probability for σ to lie between $\langle\sigma\rangle \pm 2\, st.dev(\sigma)$. This allows to asses the Gaussianity of the pdf, by comparison to such probability for a pure Gaussian, which is $p(Gauss) = 0.95449973610364158$.
For $m = |\sigma - \langle\sigma\rangle| < 2 st.dev(\sigma)$,

$$p(2\, st.dev(\sigma)) = \sum p(\sigma[m])\Delta\sigma \quad (28)$$

where
$\Delta\sigma = \sigma[1] - \sigma[0]$, i.e. the grid spacing. If $p(2\, st.dev) > p(Gauss)$ then we consider the pdf to be Gaussian.

Calculating the weighted interquartile range is non-trivial : indeed it is not readily implemented in available software packages. Traditionally, given an array V, a q-th percentile of V is the value q/100 of the way from minimum to maximum of the sorted array V. Fig. 6 shows an example of a weighted dataset, and Fig. 7 illustrates steps taken in finding a weighted percentile, required to find σ_G and median(σ) above.

Using the mean and variance of σ we can define the signal-to-noise (S/N) ratio, taking $S = \hat{\sigma}$ and the noise as $N = \sqrt{Var(\sigma)}$. We thus χ^2_{dof} and S/N are shown on Fig. 8 for a sampling of 200 amplitudes between 0.01 and 2.0.

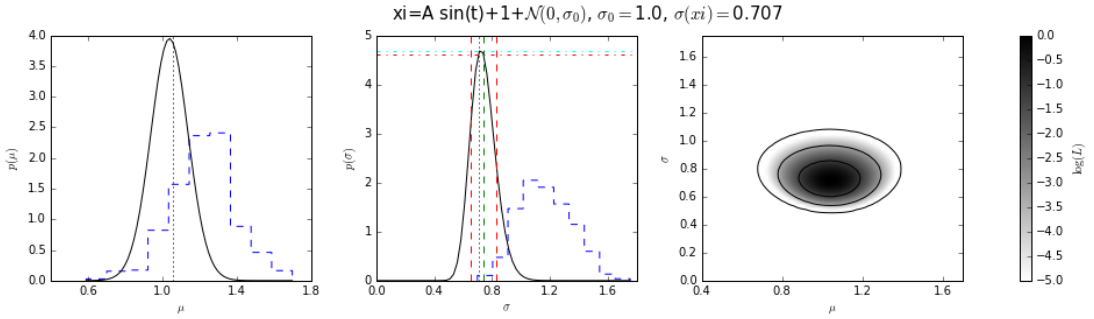

Figure 5. A realization of time series $x(t) = A\sin(t)+\mu+\mathcal{N}(0,\sigma_0)$, with $\mu = 1$, $A = 1$, $\sigma_0 = 1$. Vertical black dotted line indicates the location of the expected intrinsic standard deviation given by $stdev(x) = \sqrt{Var(x)} = \sqrt{A^2/2}$, so that for $A = 1$, $stdev(x) = \sqrt{0.5} = 0.707$. Vertical green line is the mean $\bar{\sigma_{int}}$ (Eq. 26), and vertical red lines indicate the standard deviation $\pm 1\sigma_{\sigma_{int}}$ level, where $\sigma_{\sigma_{int}} = \sqrt{Var(\sigma_{int})}$ (see Eq. 27). The horizontal red line is the value of the posterior pdf at the mean : $p(\bar{\sigma_{int}})$, whereas the horizontal blue line is the maximum value of $p(\sigma_{int})$. We use the mean : $\bar{\sigma_{int}}$, to define signal , and the standard deviation : $\sigma_{\sigma_{int}}$ to define noise. Thus $S/N = \bar{\sigma_{int}}/\sigma_{\sigma_{int}}$

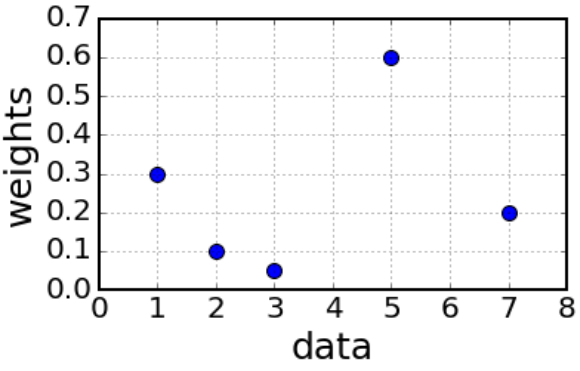

Figure 6. A dataset used to illustrate calculation of the weighted interquartile range (see Fig. 7). We plot data and the corresponding weights. A real example could correspond for instance to the pdf $p(\sigma)$ considered in this section, where data is σ, and weights are $p(\sigma)$, so that calculating weighted median or weighted percentile (for σ_G) for σ amounts to a problem of finding a weighted percen

Given amplitude A, fixing $\sigma_0 = 1.0$, for each histogram of N_{sim} realizations of $x(t, A)$ we calculate the completeness as a fraction of detections above a given S/N threshold. Thus sampling over A we compute completeness curves $S/N > 3$, and $S/N > 5$, shown on Fig. 9

As a sanity check, we plot on scatter plots χ^2 vs. A (Fig. 11) , S/N vs A (Fig. 12), and $\hat{\sigma}$ vs. A (Fig. 13). Since the plot would have very narrow horizontal 'strips' with only one value of A per 1000 values of either of χ^2, S/N, or $\hat{\sigma}$, we add a small offset to A, so that $A_1 = A + \mathcal{N}(0, 0.005)$

We also perform null test to verify the way in which the AstroML Bayesian method responds to pure error. We simulate 1000 times time series $x(t) = A\sin(t) + 1 + \mathcal{N}(0,\sigma_0)$, with null amplitude : $A = 0$, and Gaussian noise sampled from a distribution centered on 0 with width of $\sigma_0 = 1$. We assume homoscedastic errors, where $e_i = \sigma_0$ for each point. We show on Fig. 14 normalized cumulative sums for the same quantities as what is shown on previous figures.

With two sources of variance, χ^2_{dof} method is the most popular , but it heavily underperforms. In method of AstroML 5.8 we assume Gaussian origin of variations, but the type of variations is not known a-priori as we observe a random object. So whether we assume Gaussian origin, or sinusoidal (like for Lomb-Scargle), in the worst case we are equally wrong.

We test the performance of AstroML 5.8 method against Lomb-Scargle 10.14.

4 FREQUENCY DOMAIN METHODS

4.1 Periodograms

4.2 Choosing the frequency grid

The choice of frequency grid can vastly affect our ability to detect the signal in the noisy, undersampled data. We could calculate either power either in angular frequency grid ω [radians], or the ordinary frequency $f = \omega/2\pi$ [1 / seconds] . We consider three different approaches to the problem. The smallest frequency that we can detect in a regularly sampled time series may correspond to largest time interval (the baseline) $\Delta T = t_{max} - t_{min}$, so that $f_{min} = 1/T$. This is adopted by AstroML and EBellm, but Astropy reaches to frequencies $2S$ times smaller, where S, termed as the number of samples across a typical peak, is set to 5 by default. Thus :

$$f_{min}^{AstroML} = f_{min}^{EBellm} = 2Sf_{min}^{AstroPy} = \frac{1}{\Delta T} \tag{29}$$

The maximum frequency depends on the shortest time interval between data. For evenly sampled data, where Δt is equal to the time interval between data points, the maximum detectable frequency would be $f_{max} = 1/\Delta t$, also called Nyquist frequency. For the more realistic case of unevenly sampled data (as in our illustration, see Fig. 15) AstroML and EBellm use a pseudo-Nyquist frequency of $f_{max} = 1/(2median(\Delta t))$, where Δt is the time difference between consecutive observations (see Ivezic+2014, Debosscher+2007, Eyer&Bartholdi 1999). Another approach, used in AstroPy, is to make the maximum frequency dependant

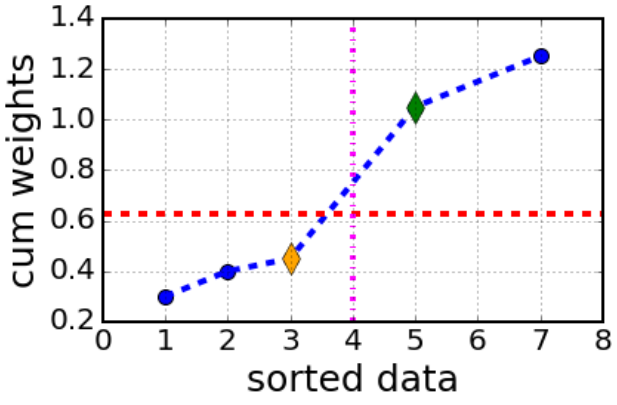

Figure 7. An illustration of calculating the weighted percentile. This can be used to calculate eg. the median, which is the 50-th percentile , or the Gaussian robust width (interquartile range σ). According to SciPy percentile specification, "Given a vector V, a q-th percentile of V is the value q/100 of the way from minimum to maximum of the sorted array V". For the data with weights , we need to find the percentile considering the data together with associated weights - if some datapoint had much more weight than others, then the percentile would be preferentially skewed towards that point. Classically, for integer weights one could simply replicate the given datapoints as many times as weights would indicate: for d = [1,2,3], and w=[3,2,4], we could say that it is equivalent to d = [1,1,1,2,2,3,3,3,3]. However ,this approach fails with non-integer weights. Our method allows for non-integer weights. Consider a short list of data with weights, shown on Fig. 6 , d = [2, 5, 1, 7, 3], w = [0.1, 0.2, 0.3, 0.1, 0.05], we sort them into tuples according to the data : $[(d_i, w_i)]$ = [(1, 0.3), (2, 0.1), (3, 0.05), (5, 0.2), (7, 0.1)], so that weights w_i are sorted acccording to the data values d_i. Then we calculate the cumulative sum of sorted weights, and find which w_i are below the fracpoint. Fracpoint f, a generalization of midpoint, is a fraction of the total sum of the weights corresponding to the q-th percentile: $f = (q/100)\sum w_i$. For $q = 50$, $f = 0.5 * (0.1 + 0.2 + 0.3 + 0.1 + 0.05) = 0.625$. This is marked by the horizontal red dashed line. The last point in the cumulative sum smaller than f is marked by the orange diamond, while the next point in cumulative sum is marked by the green diamond. If f is close to the orange diamond, then the q-th percentile becomes the datapoint corresponding to the mean distance between the two diamond points.

on the number of points, so that $f_{max} = f_{min} + (Ny_F)/(2\Delta T)$, where N is the number of data points, and y_F is the multiple of the average Nyquist frequency:

$$f_{max}^{AstroML} = f_{max}^{EBellm} = \frac{1}{median(\Delta T)} \tag{30}$$

$$\frac{f_{max}^{AstroML}}{f_{max}^{AstroPy}} = \frac{S\Delta T}{median(\Delta t)(1 + Sy_F N)} \tag{31}$$

The density of frequency grid should depend on the cadence and time span of the data. With unlimited computational resources, one would use the highest possible number of gridpoints. However, realistically we are limited by computational time, and thus we aim to use as little grid points as possible to robustly detect the peak freqyency. Debosscher+2007 chose to express the frequency step as a fraction η of the smallest detectable frequency: $\Delta f_1 = \eta f_{min}$

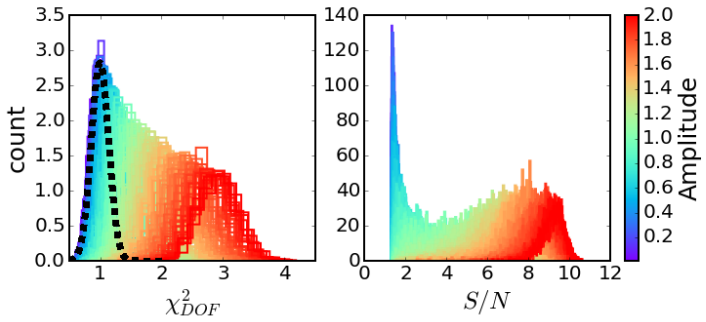

Figure 8. For the time series $x(t) = A\sin(t) + \mu + \mathcal{N}(0, \sigma_0)$, shown on Fig. 4, we assume $\sigma_0 = 1.0$, $\mu = 1$, and errors $e_i = \sigma_0$ and sample 200 values of amplitude A evenly spaced between $A = 0.01$ (blue) and 2.0 (red). We use 1000 boostraps in calculating $p(\sigma)$. We perform 1000 simulations (realizations of time series) per the value of Amplitude. Thus per value of A we have 1000 realizations of the time series, and we calculate χ^2 per amplitude. On the left panel the thick dashed black curve corresponds to a Gaussian with a mean 1, width $\sqrt{(2/N) = 0.141}$ ($\mathcal{N}(1, 0.141)$). The right panel shows how the S/N of the posterior intrinsic σ distribution increases as we increase the amplitude of variability. Importantly, the S/N of $p(\sigma)$ increases more rapidly than the χ^2_{DOF}, showing that using the Bayesian method is quicker to show signs of variability for small amplitudes.

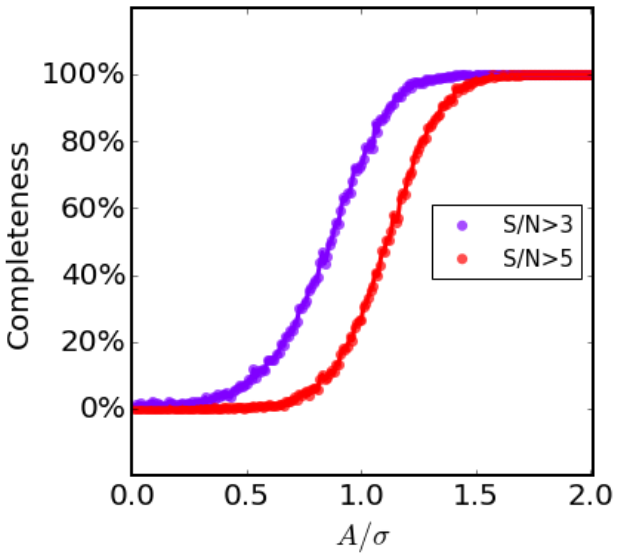

Figure 9. Completeness curves for $S/N > 3$ or $S/N > 5$, as shown by violet and red curves, respectively. We simulate time series $x(t) = A\sin(t) + \mu + \mathcal{N}(0, \sigma_0)$, iterating 1000 times over each amplitude. For each iteration we calculate χ^2_{DOF} and S/N, thus there are 1000 values of each per amplitude (see Fig. 8 for the histograms). Assume that all values of S/N for a given A is stored as a 1000-element array of S/N. For each value of A, completeness corresponds to the fraction of S/N that is bigger than the chosen S/N_{cut} (either 3 or 5). This corresponds to the ratio of the area of the histogram of S/N values above S/N_{cut} to that below. Thus, for a given A, $f(S/N_{cut}) = (count(S/N > S/N_{cut}))/1000$. Practically, this implies that if we require $S/N > 3$, we reach 90% completeness level at the A/σ level of 0.2

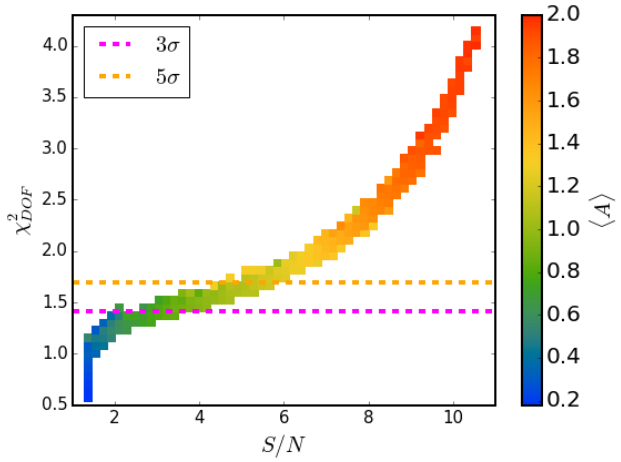

Figure 10. Collapsing information from Fig. 8 onto one plot, color represents number density. The standard deviation for χ^2 distribution is $\sigma = \sqrt{(2/N)}$. For $N = 100$, $\sigma = 0.14$, so that $1 + 3\sigma = 1.42$, and $1 + 5\sigma = 1.70$. Horizontal lines mark 3 and 5 σ limits. For each value of amplitude A we make 1000 realizations of time series with N=100 points. For each realization, we calculate the AstroML 5.8 σ, as well as properties of the PDF : $p(\sigma)$, that yields S/N for each time series. We also calculate for each realization the value of χ^2_{DOF} against the hypothesis of no variability (model y=1) . The aim is to compare how does χ^2_{DOF} compare to S/N in selecting variable objects for a given (input) amplitude. Since for each A we have 1000 realizations of time series, we can plot the histogram of χ^2_{DOF} or S/N , which is done in figure above.

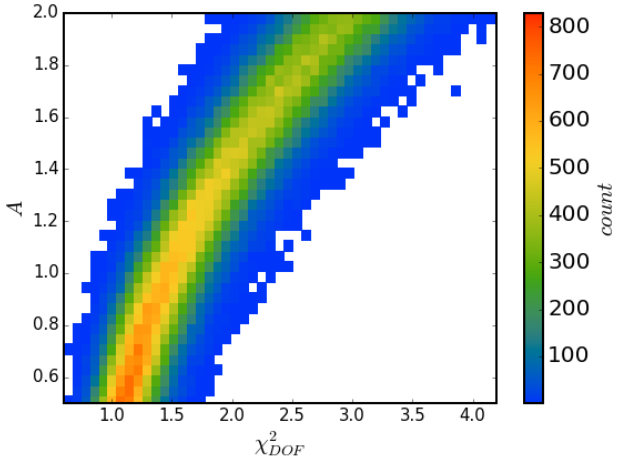

Figure 11. A sanity check : representing the same information as the left panel of Fig. 8 as a heatmap.

(in their treatment, $\eta = 0.1$). The frequency step should not be larger than the standard deviation of the location of the peak : $\Delta f_1 < \sigma_f$, for otherwise the main periodogram peak may be missed by a too coarse frequency grid. Choosing the frequency step, is in other words choosing the number of frequency bins in a linearly spaced grid. This can be a proportional to n times the minimum frequency fits in the frequency span: $N = n(f_{max} - f_{min})/f_{min}$. We can see

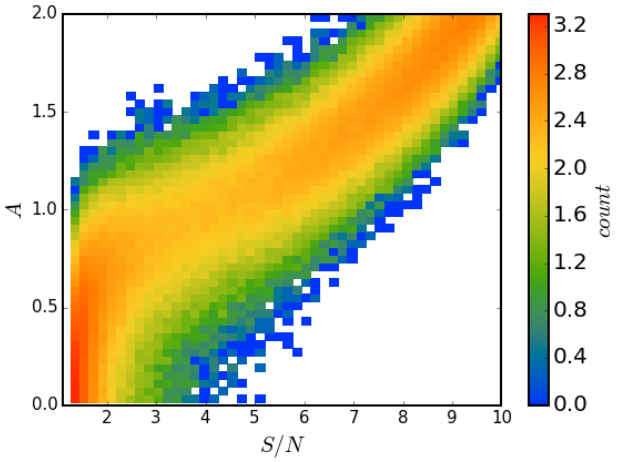

Figure 12. A sanity check: representing the same information as the right panel of Fig. 8 as a heatmap.

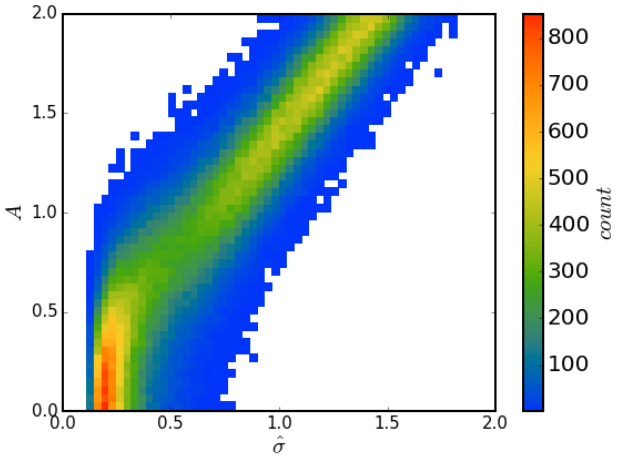

Figure 13. A sanity check: S/N values are calculated as the ratio of the mean σ ($\hat{\sigma}$), to the standard deviation of the likelihood $p(\sigma)$. We see that above A ≈ 1 the Bayesian AstroML method of detecting intrinsic variability becomes sensitive to the input amplitude of the harmonic time series ($x(t) = A\sin(t) + \mu + N(0, \sigma_0)$, with unit mean : $\mu = 1$, and input intrinsic variability : $\sigma_0 = 1.0$, evaluated over $t \in (0, 2\pi)$, sampled by $N = 100$ points).

that $n = 1/\eta$, so that the first choice of Δf is equivalent to $N_1 = (f_{max} - f_{min})/\Delta f_1 = 10(f_{max} - f_{min})/f_{min}$ bins. E. Bellm (priv. comm.) chose $n = 5$ (iPTF Summer School 2016 hands-on exercises), whereas Ivezic+2014, following Debosscher+2007, chose $n = 10$. In other words, using expressions above for f_{min} and f_{max} we arrive at following explicit formulations:

$$N_{bins}^{AstroML} = 2 * N_{bins}^{EBellm} = \frac{5\Delta T}{median(\Delta t)} - 10 \tag{32}$$

In AstroPy the number of bins is set as proportional to the number of datapoints:

$$N_{bins}^{AstroPy} = 0.5 S_{yF} N \tag{33}$$

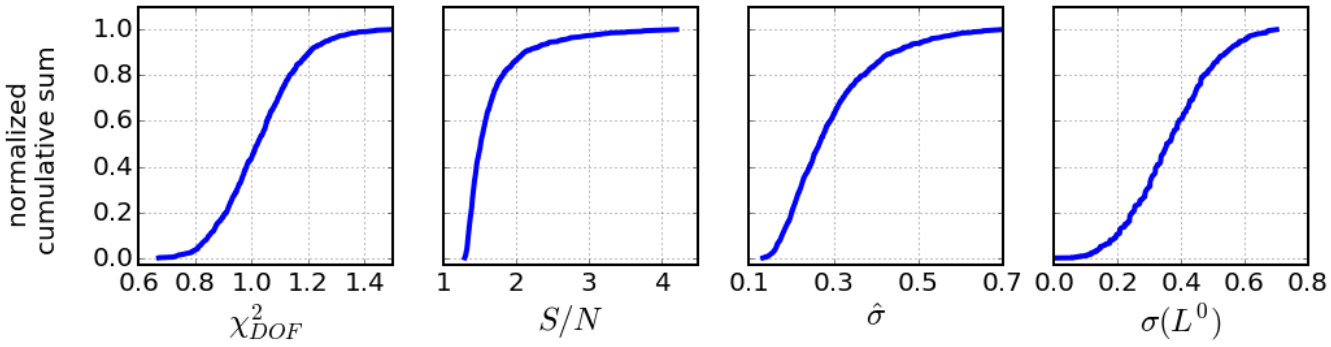

Figure 14. A sanity check: setting A=0, we perform 1000 realizations of harmonic time series with Gaussian unit noise (as Figs. 8,9,10,11,12,13). From left to right,panels show cumulative sums of χ^2_{DOF}, S/N ratio ($= \hat{\sigma}$ / $st.dev.[p(\sigma)]$ per iteration), the mean σ ($\hat{\sigma}$), $\sigma(L^0)$: σ at the 2-D maximum of the full log-likelihood. [more info : why do these plots make sense ?]

Thus all approaches, regardless of how grid spacing, start and end points are defined, result in a linearly spaced grid : $f_{grid} = linear\ space(f_{min}, f_{max}, N)$.

5 SUMMARY

APPENDIX A: PROOF OF AIC AND BIC

We prove that both *BIC* and *AIC* are related to χ^2_{DOF}, and that for a harmonic time series the difference between model and noise information criteria is a function of amplitude, number of points, and measurement error :

$$\Delta BIC, \Delta AIC(\text{model} - \text{noise}) = f(A, N, \sigma) \tag{A1}$$

Let us start from a definition of Bayesian Information Criterion. Given model M,

$$BIC \equiv -2\ln[L^0(M)] + k\ln(N) \tag{A2}$$

where $L^0(M)$ is the maximum value of the data likelihood, k is the number of model parameters, and N is the number of data points. We aim to compare the BIC between the harmonic model and the pure noise model. First, note for a combined likelihood $\mathcal{L}$ for any set of measurements $y_1, y_2, ..., y_N$, we can often assume that each value of y_i is sampled with a Gaussian likelihood:

$$\begin{aligned}\mathcal{L}(y_1, y_2, ..., y_N) &= \mathcal{L}(y_1)\mathcal{L}(y_2)...\mathcal{L}(y_N) \\ &= \prod \mathcal{L}(y_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp -\frac{(y_i - \bar{y_i})^2}{2\sigma_i^2}\end{aligned} \tag{A3}$$

Now, we replace the mean values $\bar{y_i}$ with the theoretically 'expected' values $y_{i,exp}$ and we assume that measurement errors are homoscedastic, so that $\sigma_i = \sigma$:

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp -\frac{\left(y_i - y_{i,\bar{exp}}\right)^2}{2\sigma_i^2} \\ &= \frac{1}{\sqrt{2\pi\sigma^{2N}}} \exp\left\{\sum_{i=1}^{N} -\frac{\left(y_i - y_{i,\bar{exp}}\right)^2}{2\sigma_i^2}\right\}\end{aligned} \tag{A4}$$

Now we rename part of the sum :

$$\sum_{i=1}^{N} -\frac{\left(y_i - y_{i,\bar{exp}}\right)^2}{\sigma_i^2} \equiv \chi^2 \tag{A5}$$

so that :

$$\mathcal{L} = \frac{1}{\sqrt{2\pi\sigma^{2N}}} e^{-\chi^2/2} \tag{A6}$$

Thus if we drop normalization coefficients, we find that the log-likelihood can be expressed with χ^2 :

$$\ln \mathcal{L}^0(M) \propto \ln e^{-\chi^2/2} = -\chi^2/2 \tag{A7}$$

$$BIC = -\chi^2 + k\ln N \tag{A8}$$

Now the advantage of choosing harmonic model over noise is :

$$\begin{aligned}\Delta BIC &= BIC(harmonic) - BIC(noise) \\ &= -\chi^2_\omega + k_\omega \ln N - (-\chi^2_0 + k_0 \ln N) \\ &= \chi^2_0 - \chi^2_\omega - (k_0 - k_\omega)\ln N\end{aligned} \tag{A9}$$

(Eq. 10.54 in (Ivezić et al. 2014))

Now we show how χ^2 can be linked for a harmonic model to the Lomb-Scargle Periodogram. For a harmonic model, $y(t_j) = a\sin(\omega t) + b\cos(\omega t)$. Thus :

$$\begin{aligned}\therefore \chi^2_\omega &= \frac{1}{\sigma^2} \sum_{j=1}^{N} \left(y_j - a_0\sin(\omega t_j) - b_0\cos(\omega t_{j1})\right)^2 \frac{1}{\sigma^2} \\ &= \sum_{j=1}^{N} (\, y_j^2 - 2a_0 y_j \sin - 2b_0 y_j \cos \\ &\qquad + 2a_0 b_0 \sin\cos + a_0^2 \sin^2 + b_0^2 \cos^2\,)\end{aligned} \tag{A10}$$

We can rewrite this expression in terms of $V = \sum y_j^2$, $I =$

$$AIC \equiv -2\ln[L^0(M)] + 2k + \frac{2k(k+1)}{N-k-1} \tag{A11}$$

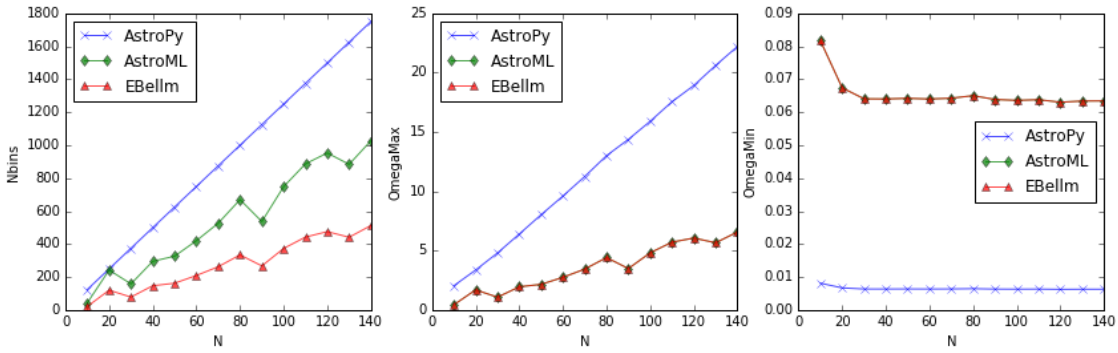

Figure 15. We simulate time series with N observation times randomly chosen between 0 and 100, varying N from 10 to 150 in intervals of 10. As we explain in the text, the left panel shows that AstroPy chooses higher number of gridpoints for the same N than AstroML or EBellm. The middle and right panels show that while f_{min} and f_{max} are identical for AstroML and EBellm implementations, they are more extreme for AstroPy (bigger f_{max}, smaller f_{min}). For all N the length of baseline ΔT is unchanged, i.e. we sample more points from the same time interval. Note that we plot $\omega = 2\pi f$.

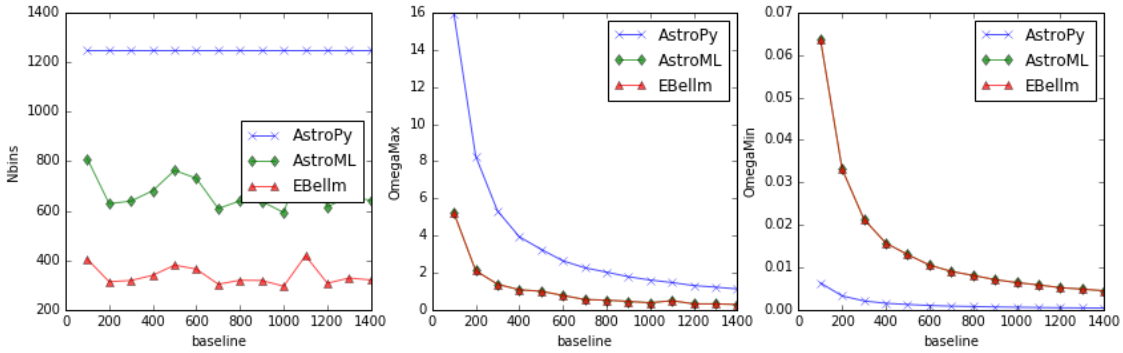

Figure 16. We simulate time series with fixed $N = 100$ observations, sampled at times randomly chosen from between 0 and baseline ΔT, varying ΔT from 100 to 1500 in intervals of 100. On the left panel, we see that the number of bins for $AstroPy$ remains constant, since it is proportional to $0.5 Sy_F N$, while for AstroPy and EBellm the number of bins is approximately constant, because as ΔT increases, so does $median(\Delta t)$, with fixed $N = 100$. For this reason, in the middle panel for AstroML and EBellm, $f_{max} \propto 1/median(\Delta t) \propto 1/\Delta T$. The right panel shows f_{min} decreasing with an increase of baseline as $1/\Delta T$ for AstroML and EBellm, and as $1/(10\Delta T)$ for AstroPy.

APPENDIX B: WEIGHTED MEAN AND STANDARD DEVIATION

Here we consider the difference between calculating unweighted and weighted versions of mean and standard deviation. For instance, consider a set of points x_i with errors e_i . If we assume that each point x_i has an equal probability, so that we are calculating unweighted mean and standard deviation. Then the mean is :

$$\mu(x_i) = \frac{\sum x_i}{N} \tag{B1}$$

and the standard deviation :

$$\sigma(x_i) = \sqrt{\frac{1}{N}\sum (x_i - \mu)^2} \tag{B2}$$

where if the mean μ is calculated based on the data, we loose one degree of freedom, so that $1/N$ becomes $1/(N-1)$. This is called Bessel's correction, but for large N does not significantly change the result.

Now, if instead of having equal probabilities each point x_i has a probability p_i (often called weight w_i), then the mean is called weighted mean :

$$\mu_w(x_i) = \frac{\sum p_i x_i}{\sum p_i} \tag{B3}$$

where we divided by the sum of probabilities to ensure that they are normalized. If they are properly normalized, then $\sum p_i = 1$. If all values are equally likely, then $p_i = 1$, so that $\sum_{i=0}^{N}(p_i) = N$, and we recover Eq. B1.

Similarly, for standard deviation we have

$$\sigma_w(x_i) = \sqrt{\frac{\sum p_i (x_i - \mu_w)^2}{\sum p_i}} \tag{B4}$$

where we divide by the sum of probabilities (weights), to ensure that the probability is normalized (especially if we are using weights). Note, that here again if all weights (probabilities) are equally likely (but not necessarily normalized),

then $p_i = 1$, and $\sum p_i = N$, and we recover Eq. B2. Thus, for all weights equal, $\sigma_w = \sigma$, and $\mu_w = \mu$, as we would expect.

REFERENCES

Ivezić Ž., Connolly A. J., VanderPlas J. T., Gray A., 2014, Statistics, Data Mining, and Machine Learning in Astronomy

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