Fundamentals in Time Domain Astronomy: Grouping to Reveal Structure

G. Bélanger

2014 October 29 (draft)

This first set of two one and a half hour lectures was prepared for the morning of Thursday October 30, 2014 at the first ESAC Data Analysis and Statistics Workshop, which was held over the course of five days, from Monday to Friday, October 27 to 31. They address many if not most fundamental aspects of the analysis of time domain signals in astronomy by posing a single question—how do we group data to reveal their structure—and working through to the answer with attention to the details relating to the statistical concepts involved.

Introductory Remarks

Much of what is done in astronomy and astrophysics falls in one of three domains: imaging, spectral, and timing. What we are going to talk about is timing. And what I mean here by timing is everything that has to do with measurements of anything at all as a function of time.

As far as how the data is presented, there are basically only two possibilities: we either get individual time-tagged events,[1] or we get measurements of what we can refer to generally as the "intensity" (whatever the actual quantity may be), but which is estimated by as an average over a specific time interval.[2]

An example of the first is an X-ray or γ-ray event list, and an example of the second is a time series of energy density measurements based on 30 second snapshots with a near-infrared camera. The event file contains at the very least a list of the times at which the events were detected/recorded.[3] The energy density measurements extracted by PSF fitting at the position of the source of interest in the 30 second snapshots, on the other hand, have the start and end times of the frame, and the measured energy flux.

In these two lectures, we will ask some questions and look at some data. We will ask very basic questions, and we will look at our data very closely. I don't use "basic" to mean simplistic; I use it to me mean fundamental and therefore essential. In the same way want to clarify right from the start that the use of the word "fundamental" in the title, does not mean "easy". In fact, the material we'll cover is *not* so easy because it is quite technical. Rather, "fundamental" here means most essential and important.

We have three hours together, which is not so much, but it is enough to look at quite a few things. And I will ask a lot of ques-

[1] These are ***unbinned*** because readout frequency is higher than event rate.

[2] These are ***binned*** at the instrumental level because the signal is accumulated for a particular time before readout.

[3] And usually also their coordinates on the detector plane, their energy, and other things like quality flags, etc.

tions, rhetorical questions to stimulate you thinking about the issue at hand. Nobody needs to answer. *You* don't need to answer. Do not feel like you're being put on the spot: you are not! Just think about the question and the issue we are considering. If an answer pops up, then great: just say it. In general, I'll provide all the answers to my own questions in the discussion that follows them. Once again, please, don't stress about it. It's just a way to engage you and keep you on your toes. Let's start now.

Distribution of Waiting Times

WHAT IS THE FIRST THING we can do with a list of event arrival times? We are only interested in their temporal characteristics, and not with their energy or spatial distribution. The very first thing I do is to take a look at the ensemble of times intervals between events: the interarrival or waiting times.

Why? Because, firstly, this is the most basic feature of their distribution in time: how much time passes between events? What is the statistical distribution of the interval between detected events? And secondly, it is something that can be done immediately without requiring any kind of manipulation of the data.

What do we gather from it? Well, if we were in a physics lab looking at the photons recorded by a detector encapsulating a radioactively decaying source, then we would expect the waiting times to be distributed precisely as an exponential distribution with a mean given by the inverse of the detection rate.[4]

The exponential distribution is characteristic of a memory-less process, a process in which each event takes place at a given average rate, *independently* of the previous and the next. The exponential always peaks at zero, and extends to positive values. The speed with which the density drops towards zero (the x-axis) is determined by the mean of the distribution.[5]

What else do we look for? If the process is constant, or rather, if the process gives rise to an apparently constant event rate, then this means that all waiting times will belong to the same parent population and be distributed according to a single, well-defined distribution. We can check that by overlaying on our distribution of waiting times the exponential density function. It is also a good way to identify anomalies, of course: peaks, bumps or breaks in the distribution, for example. We'll see some of that later.

What if the distribution is exponential in appearance, but doesn't agree with the analytical density function? It has a longer tail, for example. This immediately tells us that the event rate is *not* constant,

[4] If the decay process yields 1 photon per second, then the average waiting time is 1 second between photons. If it produces 4 photons per second, then the average waiting time is 1/4 of a second.

[5] We also speak of the decay constant, which is the inverse of the mean: a mean of 2 translates into a decay constant of 1/2. In the application we are considering of photon detection, the mean of the distribution is the average waiting time, and the decay constant, therefore, equals the detection rate.

and that it is, therefore, variable. It can be just slightly or strongly variable, but if the distribution of waiting times is not a pure exponential, then it means that it can be thought of as a mixture of different exponentials from the different "rate-states". We'll see that later as well, but let's not go too fast.

How about a simpler question first: how do we group the data?

Structure and *Resolution*

YOU KNOW THAT making a light curve is the very first thing an astronomer does to look at their data, sometimes even in near-real time during the course of an observation. Have you ever asked yourself "what bin size should I use to make this light curve?" I'm sure you have. So, what did you choose as the bin time: a millisecond, a second, 60 or 100 seconds, 3600 seconds? And why?

This is the question that will drive us in our investigations throughout these two lectures: ***Is there an ideal—an optimal—timescale for grouping data?*** If so, what is it and how do we find it?

Naturally, it depends what we want to do with the light curve. So, let's first restrict ourselves to the following goal: *We want to group the data into a time series that will reveal the most about the variability structure of these data.* Personally, I think that whether you have phrased it in this way or not, this is generally what we all want to see when we look at a light curve as a means to getting a sense of what's going on in the source we're observing.

This grouping to reveal ***structure*** applies in the same way to event data as it does to instrumentally binned data, and the implementation is just slightly different in some of its details. We will look at even data first, which is simpler because we can bin the events as we wish without any concerns about uncertainties and the like. This is not quite the case when we are working with binned data that we want to resample, as we will see later on.

How do we decide, on a *quantitative* basis, what time scale will best reveal the structure in our data? Let's think of the two extreme cases: having only one bin, or having an infinite number of bins. Will we see structure in our light curve? No. What will be the mean deviation along the length of the time series? Nil or very close to it. This gives us an indication, firstly, that the answer is somewhere between one and an infinite number of bins (obviously), but also of how we can go about choosing or picking out that ideal scale for looking at the data.

Another point that we cannot overlook is that whether we are working with event or binned data, one, and sometimes, *the* most important issue, is ***resolution***. Therefore, whatever we may be interested in investigating in the data, it is crucial to retain the best possible resolution.[6]

Hence, we have *two* essential concerns in what relates to grouping the data: We want to reveal the variability structures in the most effective way, and we want, *at the same time*, to retain the highest possible resolution. For this, we must turn to statistics.

Underlying Statistics: The Distribution of the Data

LET'S COME BACK to the extreme of an infinite, or in practical terms, a very large number of bins. Would we be able to see any structure by looking at this extremely sparsely distributed time series where each bin would have at most a single event, and where most bins would be empty? Clearly not. Now, do the thought experiment of looping through the data, from the first event to the last, over and over again, grouping it with a slowly increasing bin size. This will in effect, squeeze the data together, gradually taking out the empty bins and clustering more events together. We'll begin to see the data grow vertically, so to speak, growing out of the one dimensional line of mostly zeros into taller, increasingly better defined structures.

Processes in physics appear to be statistical in nature.[7] In terms of measuring such processes, this means that there is always going to be scatter around the mean. The amount of scatter, in absolute terms, the spread of the measurements around the mean, depends both on the process and the characteristics of the measurements, but it does *not* change, and it does *not* depend on the number of measurements we make. What *does* change, however, and what *does* depend on the number of measurements, is 1) the precision with which we can determine the most likely value of the measured quantity, 2) the spread of the set of measurements around that value, and 3) the shape of the distribution of measurements.

Hence, the first relevant question we should ask is this: What is the statistical nature of the data? Or, in other words: ***how are the data distributed?*** Of all the basic questions we should ask when working with data, this is the *most* fundamental, because it is the answer to this question that tells us what are the relevant statistics. In the case we are currently considering, we are detecting and counting individual events, and therefore, are working with Poisson statistics.

[6] For many, maybe most analyses, like spectral modelling in the energy or power spectral domains, for instance. we can, *and should*, aim to work with the highest resolution available.

Each instrument has its particular characteristics, like its timing resolution and its number of energy channels, frequency or wavelength bands, and we should use all the information that is available to us without throwing any of it away.

Throwing information away is, in effect, what we do when we group energy channels together: artificially degrading the energy resolution of our instrument. This is never necessary when using the correct statistics and appropriate statistical analysis methods.

[7] I used "appear" because making the stronger statement that physical processes *are* statistical in nature could lead into much lengthier discussions that would clearly be more philosophical than scientifically pragmatic in flavour, independently of whether it is even possible to bring such discussions to a conclusive end. Even though I personally find such investigations very interesting, it is not the purpose of these lectures, which is indeed pragmatic.

Statistical Uncertainty: Statistical Fluctuations

I WILL NOT DISCUSS generalities of Poisson statistics, because you are probably already familiar with this by now. The crucial element I want to use is that the Poisson probability density is a single parameter function whose value defines both the mean and the variance.[8] In addition, I want to highlight and clarify the distinction between what I will refer to as *homogeneous* and *non-homogeneous* Poisson processes.

A homogeneous Poisson process is, very simply, one for which the value of the parameter is constant. A non-homogeneous Poisson process is one for which the value of the parameter changes. What is important to recognise in this context is that *any* process which is intrinsically variable *and* which is measured as individual events—no matter what the nature of the process and the single or multiple causes of the variability—can be considered as a non-homogenous Poisson process, and thus treated and understood as such.[9]

For a homogeneous Poisson process, the statistical fluctuations expected in the measurements, which translates directly to the inherent statistical uncertainty, is defined by the variance that is also the mean. Hence, for a mean number of events per bin of n, the variance about the mean will be n, and the statistical uncertainty, expressed as the standard deviation, will therefore be given by $\sqrt{n}$.

To make sure this is perfectly clear, it means that if we had an infinite time series of events detected at a constant rate of ν from a non-variable process, and we grouped these events in bins of width dt, we would have on average $n = \nu dt$ events per bin. If we now took the number of events in each bin and made a frequency histogram, putting in each bar the number of times we find in a bin of the time series zero events, one event, two events, three events, and so on, the histogram would trace the Poisson distribution function with a mean of n, variance of n, and standard deviation of $\sqrt{n}$.

With this in mind, how would you estimate the magnitude of the statistical fluctuations we might expect in our measurements of a *non-homogeneous* process? Let's do another thought experiment and imagine we can tune variability. Let's set the mean count rate to ν and keep it constant. This implies that the total number of events for an arbitrary observation duration T will always be exactly $N = \nu T$. It also implies that the average number of events per bin of width dt will always be $\langle n \rangle = \nu dt$.

Let's tune up and down variability and consider what we would see. First, turn it down to the minimum, no variability, and ask yourself these questions: Is the shape and spread of the distribution of events in each bin dependent on something other than the statistical

[8] As is also the case for the exponential distribution, which is, in other ways, entirely different.

[9] The motivation or philosophical basis for this, is that the act of measuring the process by detecting, collecting and counting individual, discrete events, imprints onto the process the signature of Poisson statistics. This is independent of the process, its nature, and what causes it to be variable.

nature of the process? Do the instrument characteristics depend on the variability of the process being observed?

Now, turn up the variability gradually, so that you see larger and better defined structures in the time series. Does the intrinsic variability have anything to do with the measurement uncertainty? Is there any reason to believe that there should be more or less *statistical* fluctuations in the measurements as the variability of the process increases or decreases while the mean rate and bin size remain constant? And is there any reason to believe that the *statistical* fluctuations should ever be greater or lesser than those associated with the first, non-variable process observed in our experiment, a homogeneous Poisson process, of exactly the same mean rate and bin size?

The answer to all these questions is no: The properties of the intrinsic variability of a source have nothing to do with the statistical measurement uncertainty associated with the detection process. They are distinct and independent. And the fact that we convolve one with the other when measuring the physical process as a discrete number of measurements with our instrument, does not imply that statistical fluctuations should increase or decrease with greater or lesser variability.

On the contrary, it shows us that the most reasonable estimate we can make of the magnitude of statistical fluctuations for *any* Poisson process is given by the variance (for a given bin size) of the homogeneous process with mean rate equal to that observed.

The Frequency Domain of Fourier Space

LET US COME BACK once more to our motivating concern of grouping the data to best reveal structure while retaining the highest resolution. Translating these two requirements into a statistical statement, we could term the question as follows: ***What is the time scale at which the magnitude of the statistical fluctuations are equal to the fluctuations due to the intrinsic variability?***

Answering this question will yield what we're looking for. But how do we estimate, *quantitatively*, the magnitude of the fluctuations, not for the statistical fluctuations of the Poisson part of the process that can be estimated simply using the mean event rate, but for the intrinsic variability, the non-homogeneous part, so to speak, of the observed process? For this, we must again turn to statistics, but we must also move to a different space, a different domain: the frequency domain of Fourier space.

Transforming the time domain signal we are working with to its representation in Fourier space as a frequency domain signal by constructing a periodogram, allows us to estimate the amount of "power", the amount of activity, so to speak, at each frequency that is accessible in the signal.[10] The frequencies that can be tested are defined by the duration of the data set, on the low frequency end, and by the spacing in time between consecutive measurements, on the high frequency end. For event arrival times, the highest frequency that can be tested is related to the minimum time between two consecutive events, whereas for binned data it is related to the sampling rate (bin size).

The highest testable frequency is referred to as the Nyquist frequency, and is given by half the sampling frequency. The lowest frequency is given by the inverse of the time spanned by the data (the total duration). The inverse of the duration also defines the distance between independent frequencies, the step between frequencies, that we call an Independent Fourier Spacing or IFS for short.

The simplest and most intuitive way to think of the information conveyed by a periodogram is to imagine that you place a sinusoidal wave over your time series, and the closer it is to the shape of the time series, the more "power" you get for that particular frequency of the wave. Start at the lowest frequency (the longest wavelength) and go through all the testable frequencies, doing the same thing each time. Each result is shown on the periodogram at the height of the point at the corresponding frequency.

Another way to understand the operation of making the periodogram is to consider the list of arrival times, scattered along the length of the observation timeline, and, taking the first test frequency, think of it as a period instead. Now, calculate to which point in the phase of this period between 0 and 1 each arrival time corresponds. This is done by dividing the time by the period, and dropping the integer part of the result, keeping only the decimals.

Hence, from a list of arrival times, we have constructed a list of phases for the period we just used. We can now, for each phase, multiply it by 2π to get radians, take its sine and cosine, square each one, and sum them. Doing the same for each phase, and then summing all these terms, and then dividing the result by the total number of phases in the list (and multiplying by 2), we get the value of the Fourier power for *that* period in *this* data set. We repeat this procedure for each test period and construct the periodogram.[11]

I have been able to delay using equations until now, but this kind of discussion is easier to see in mathematical terms. What we did for each testable frequency f, is to map each arrival time t_i, to its phase ϕ_i, within the periodic cycle that corresponds to that frequency

[10] If we consider the time series to be a degraded signal which is a single, finite length realisation of the actual signal at the source of the physical process, then the periodogram is, in exact analogy, a single, degraded estimate of the actual power spectrum at the source over the bounds of testable frequencies. The underlying assumption in this is that the nature of the process—stationary or non stationary—does *not* change. This implies that the physical system being observed is assumed to remain in the same general state, because a state change would imply a change in the characteristics of the emission, and also a change in the power spectrum.

[11] This particular periodogram is called the Rayleigh periodogram. It is the simplest and, at the same time, the most powerful for detecting sinusoidal signals.

$(p = 1/f)$, and calculate the Rayleigh statistic:

$$R^2 = 2N(C^2 + S^2) \tag{1}$$

where C and S are defined as:

$$C = \frac{1}{N}\sum_{i=1}^{N} \cos\phi_i \quad \text{and} \quad S = \frac{1}{N}\sum_{i=1}^{N} \sin\phi_i. \tag{2}$$

And in terms of the statistics of the periodogram, first, the expectation value of the functions $\cos\phi$ and $\sin\phi$ is zero: $\langle\cos\phi\rangle = \langle\sin\phi\rangle = 0$. Therefore, so are those of C and S. Second, the variances of $\cos\phi$ and $\sin\phi$ both equal one half: $\mathrm{V}[\cos\phi] = \mathrm{V}[\sin\phi] = 1/2$. Therefore, those of C and S are a factor of N times smaller: $\mathrm{V}[C] = \mathrm{V}[S] = 1/2N$. Finally, since $\mathrm{V}[mX] = m^2\mathrm{V}[X]$, where m is a numerical constant, the scaled variables $c = \sqrt{2N} \cdot C$ and $s = \sqrt{2N} \cdot S$ have a variance of one: $\mathrm{V}[\sqrt{2N} \cdot C] = \mathrm{V}[\sqrt{2N} \cdot S] = 2N \cdot \mathrm{V}[C] = 1$.

Note however, that the phases are uniformly distributed between 0 and 2π, and the sine and cosine are distributed between -1 and 1 with their characteristic, symmetric U-shaped distribution with minimum at 0 and rising slowly toward the edges where it peaks very sharply. It is the summing and averaging of several identically distributed values that yields the two normal variables C and S, and standard normal c and s.

This implies that

$$R^2 = c^2 + s^2 = 2NC^2 + 2NS^2 = 2N(C^2 + S^2) \tag{3}$$

is the sum of the squares of two standard normal variables. Squaring a standard normal yields a χ^2 variable with one degree of freedom (dof). Summing χ^2 variables yields another χ^2 variable with a number of dof that is the sum of the dof of the variables being summed (this is illustrated in Figure 1). Therefore, the power being the sum of two χ^2_1 variables is χ^2_2 distributed with a mean and standard deviation of two (or variance of four). This is convenient due to the simplicity of the purely exponential χ^2_2 density function:

$$\chi^2_2(x) = \frac{1}{2}e^{-x/2}. \tag{4}$$

The caveat here is that this is only true if the power estimates at different frequencies are independent, which is only true for *non-variable* processes. The Fourier transform of a constant is also a constant. Consequently, the time series of such non-variable processes yield a globally flat periodogram with equal power at all frequencies, but with the statistical fluctuations characteristic of the χ^2_2 with its mean and standard deviation of two.

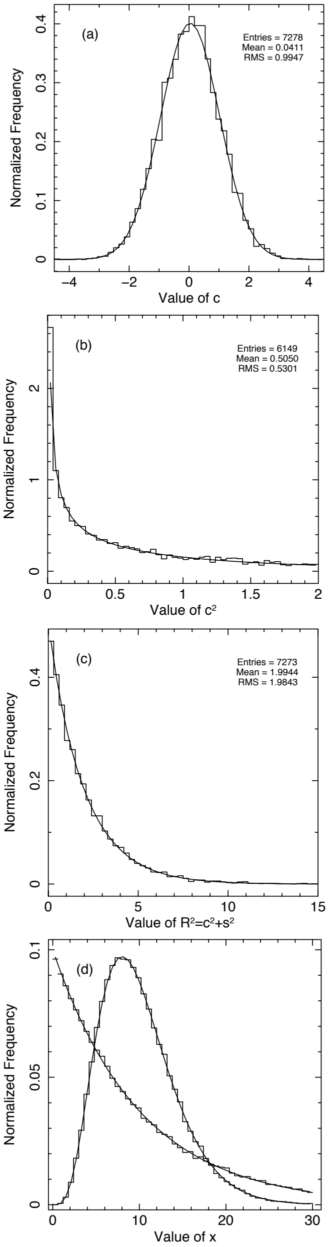

Figure 1: Illustration of the relationship between standard normal, χ^2 and exponential variables using the normalised frequency distributions and the analytical density functions: In panel (a) we see the variable $c = \sqrt{2N}C$ (standard normal); in panel (b) we see its square, $c^2 = 2NC$ (χ^2_1); and in panel (c) we see Rayleigh statistic, $R^2 = c^2 + s^2$ (χ^2_2: an exponential with $\tau = 2$). Panel (d) illustrates the difference between summing five χ^2_2 variables (χ^2_{10}), and scaling by five a χ^2_2 (exponential with $\tau = 10$).

Stochastic Variability: Red Noise

What about randomly variable processes? These are the ones in which we are, in fact, interested. We want to find the ideal time scale to group our data so that when we look at them in the form of a time series we can see what they have to show us about the features of the emission with the optimal level of detail.

Variable processes are different in the sense that the very fact of variability in the data implies that there is some relation between the activity at different time scales. For anything other than coherent sinusoidal variability, which will appear as a single peak at the frequency of the modulation, what this means is that the activity of the system gives rise to multiple physical processes that vary on different timescales but are influenced by one another, and that, therefore, the power estimates at different frequencies are related to a certain extent, maybe tightly or maybe not, maybe across the entire power spectrum, maybe not, but the bottom line is that if there is some level of random variability there is correlation between the power at different frequencies.

What does this mean about the way the various power estimates are distributed then? Are we working with a different χ^2 or exponential variable, or some other kind? Let's look back at the thought experiment we did to determine the extent of expected statistical fluctuations while tuning up or down variability, and, in analogy to this, ask ourselves the following: does the shape or form of the underlying power spectrum as it is as the source of the emission have anything to do with the statistics of the periodogram? No, it doesn't. And for this reason—for *any* power spectral shape—the power estimates at each testable frequency are distributed as the basic χ^2_2 variable resulting from the sum of squared normal variates, scaled by the underlying power spectrum. The powers are therefore all exponential variables (as demonstrated in Figure 2) for which the mean is given by the best fit model of the periodogram.

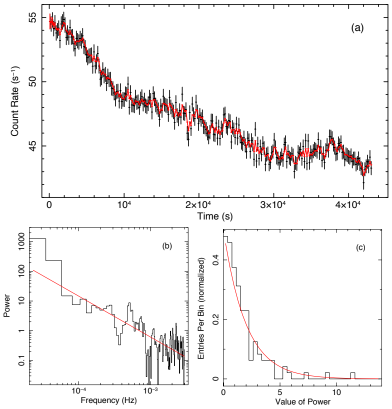

Figure 2: Illustration of periodogram powers of astrophysical red noise as scaled χ^2_2 (exponential) variables using an *XMM-Newton* observation of Mkn 421: Panel (a) shows the RGS time series in rates (0.3–2 keV with 85 s bins); panel (b) shows the periodogram with the best fit power-law model; and panel (c) is the distribution of de-trended periodogram powers overlaid with the analytical form of the χ^2_2 density function, the exponential density with mean of 2 (decay constant of 1/2): $f(x) = \frac{1}{2}e^{-x/2}$.

Using the language of the frequency domain just introduced, we can rephrase the question that we expressed in the terms "What is the time scale at which the magnitude of the *statistical fluctuations* are equal to the *fluctuations* due to the intrinsic variability?", as "What is the time scale at which the *Fourier power* associated with the statistical fluctuations is equal to the *power* of the intrinsic variability?".

To answer this question, recall our thought experiment in which we looped through the data grouping it using an extremely small bin size at first, and then gradually increasing the bin size, seeing the structure of the data begin to grow out of the flat, one-dimensional array of sparsely distributed measurements. Translating this to our analysis in Fourier space, which, in effect, does exactly this by giving us the power (the amount of activity) associated with each different time scale (or frequency) that can be tested.

For stochastic variability, which is almost always what we see and thus have to work with in astrophysics, the distribution of powers in the power spectrum follow a power-law that peaks at low frequencies and decays towards the higher frequencies. Such a power spectrum is most often referred to as "red noise", in analogy to the optical spectrum of visible light where the red wavelengths are the longest and thus have the lowest frequencies. The place where the power-law has a value of two is the place where there is equal power from the intrinsic variability as from the statistical fluctuations. Therefore, this is finally the answer we were seeking.

The Optimal Grouping Time Scale

MATHEMATICALLY, we express it in the following way: The power p, is given as a function of the frequency f, by

$$p = Nf^{-\alpha}, \tag{5}$$

where N is the normalisation and α is the power-law index. The frequency at which there is as much power from the inherent variability as from the statistical fluctuations, is where the power in the power-law equals the value of the constant level of the Poisson floor c.[12] Hence, it is found by solving $Nf^{-\alpha} = c$, and yields

$$f_{\mathrm{grp}} = \left(\frac{N}{c}\right)^{1/\alpha}. \tag{6}$$

The corresponding time scale is just the inverse, and thus equal to

$$dt_{\mathrm{grp}} = \left(\frac{c}{N}\right)^{1/\alpha}, \tag{7}$$

where we use the subscript "grp" for "grouping".

We've made quite a lot of progress in following this investigation, and we're almost there. I know I've been throwing a lot of information at you for quite a while, but I was wondering if anyone started wondering, in the last few minutes, how we know what value of the power-law index α to use in order to determine that frequency and time scale we are interested in. Where do we get α from? We have to fit the periodogram.

[12] In our case, using the Rayleigh normalisation yields a value of two for the ***Poisson floor***. It can, however, in general be any other value depending on the periodogram normalisation.

Fitting the Periodogram: The B-Statistic

WHO REMEMBERS what we said about the distribution of power estimates in the periodogram? How are the power values at a given frequency distributed? What kind of variable are we working with? We're working with exponential variables. So, how are we going to fit the periodogram? Who remembers how we go about constructing an optimal fit statistic based purely on probability theory and without having to make any kinds of approximations or assumptions? Let's go back to something familiar and ask the questions that should lead us to what we want to know.

Say you're *planning* to observe a Poisson process for which you expect a constant event rate of ν. At this stage, you can ask—however many times you want—what is the *probability* of detecting n events in the reference one second interval? The answer is given by plugging these numbers in the equation for the Poisson probability density function:[13]

$$f(n;\nu) = \frac{\nu^n e^{-\nu}}{n!}. \tag{8}$$

In the same way, you can ask what is the probability, if I make two independent measurements, of getting n_1 and n_2 events from the first and the second respectively? The answer is that we now have to multiply the probability P of the first with the probability of the second. Hence,

$$P(n_1, n_2; \nu) = \frac{\nu^{n_1} e^{-\nu}}{n_1!} \times \frac{\nu^{n_2} e^{-\nu}}{n_2!}. \tag{9}$$

Now, say you are actually running your experiment and measuring the number of events detected by your instrument every second. You make the first measurement and detect n_1 events. The question you can ask now, already at this early stage, but also at any other stage, what is the *likelihood* of having detected n_1 events when I was expecting ν? And the answer is exactly the same as the one for the first question we asked while planning our experiment about the probability of seeing n events in a single measurement.[14]

Let's now transpose our problem to the periodogram we have made of our data set. At this stage, because we have already collected the data and carried out the "experiment", we can only talk about likelihoods and not about probabilities anymore. So, we can ask: what is the likelihood of this value of power, p, that we see at frequency f, when we expect to find ρ? In this case, the answer is derived from the exponential probability density, and is given by

$$L(\rho|p) \propto f(p;\rho) = \frac{1}{\rho} e^{-p/\rho}. \tag{10}$$

[13] In the section that follows, as previously, I use Greek letters to denote the model expectation, and Latin letters to denote the measurements.

More specifically: ν expected for n observed events in a counting experiment, and ρ expected for p estimated power in the periodogram.

[14] The distinction is subtle but important: Before the experiment, we calculate probabilities; once we have made a measurement, we calculate likelihoods, and both are calculated directly from the probability density function of the variable describing the statistical nature of the measurements.

Furthermore, the probability makes a statement about the measurement given a specific value of the parameter, whereas the likelihood makes a statement about the parameter given the actual data, which is now fixed by the fact that it was measured.

Finally, the third crucial distinction is that a probability is normalisable such that the area of a probability density always equals unity, whereas the likelihood function is not. For this reason, only *relative* values and ratios of likelihood are meaningful.

Note that here, exactly as it was the case above, asking and answering this question can only be done in light of an expected outcome, and this means, having a model. The model for the homogeneous Poisson process considered above is as simple as can be: a single parameter model which is a constant. The model for the periodogram is somewhat more complex, although not much more: it is a power-law, which contains two parameters, the spectral index and the normalisation.[15]

The periodogram contains many independent frequencies over which the power is estimated. What is, then, the likelihood of having measured the powers we have compared to the powers we were expecting as defined by the power-law model? As above, it is given by the product of the individual likelihoods from each frequency:

$$L(\boldsymbol{\rho}) = \prod_i \frac{1}{\rho_i} e^{-p_i/\rho_i}, \tag{11}$$

where, unlike in the example of the homogeneous Poisson process we used above where the expected number of events is always the same, each frequency channel has a different measured power p_i, but also a *different* expected power ρ_i given by the model (hence the bold $\boldsymbol{\rho}$ to denote that it is a vector of values, one per frequency).

To make the calculations easier, and work with sums instead of a products, we can take the log of the expression for the join likelihood.[16] Moreover, because we are interested in actually constructing a fit statistic whose value we will minimise by iteratively adjusting the parameter values, we define the B Statistic as $-2\ln L$:

$$B = 2\sum_i (\ln \rho_i + p_i/\rho_i). \tag{12}$$

And with this we have all the elements we need to take the final step.

Application to Event Data

We will now apply what we have learnt this far by actually doing it. We will take a time series of a stochastically variable process,

1. look at the distribution of interarrival times,
2. construct the periodogram,
3. determine the best fit slope by model fitting using the B-stat,
4. identify the optimal grouping time scale,
5. bin the data and examine the result,
6. compare to a range of different shorter and longer bin times.

As we do this, we'll simultaneously look, at each step, at what we would expect from a homogeneous Poisson process: a white noise

[15] Technically, our complete model for the periodogram is a power-law plus a constant to account for the Poisson floor, which in the most general case would lead to three free parameters. However, the constant level is defined by the normalisation used to compute the periodogram.

[16] The exponential log-likelihood for a vector of model expectations, $\boldsymbol{\rho}$, and a vector of measurements, $\boldsymbol{p}$, is

$$\ln L(\boldsymbol{\rho}|\boldsymbol{p}) = -\sum_i (\ln \rho_i + p_i/\rho_i).$$

without variability structure. We will first work on event data, and then on a time series binned at the instrument level.

For the event data, the unbinned time series, we will use simulated observations that have the characteristics we desire, and that we know precisely (of course). The data sets have an observation duration of 10^4 s, and mean rate of $10\,\mathrm{s}^{-1}$. The power-law index α of the stochastically variable (red noise) process *before* drawing the events to make up the count rate is 2.5. For the binned time series, we use 10 hours of VLT and Keck near-infrared data.

Distribution of Interarrival Times

Our first step is to construct the frequency distribution (histogram), of the interarrival or waiting times. We therefore do this for our red noise event list and for the white noise as a means to compare them and train ourselves to become sensitive and attuned to differences that can sometimes be subtle. The histograms are shown in Figure 3.

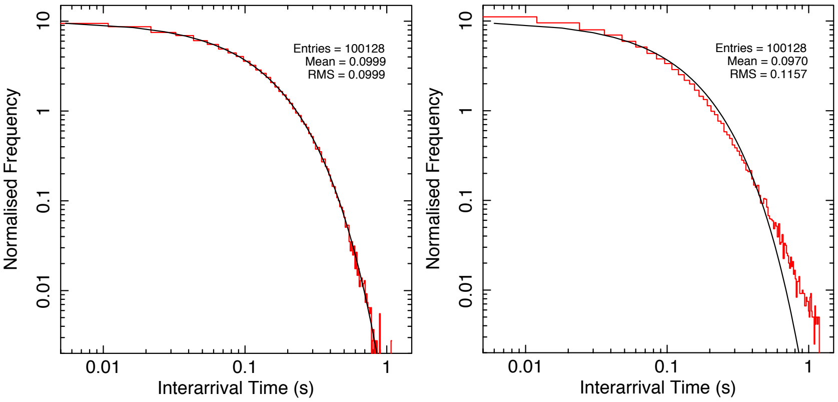

Figure 3: Normalised frequency distributions of interarrival times constructed on 100 bins in the range between 0 and 1.2, overlaid with the probability density function for a white noise process of equal count rate. The left panel shows the histogram for a white noise process, and the right panel shows the corresponding histogram for the variable process we are considering in our analysis. Using a log-log scale helps highlight the differences between the histogram and the density function.

Even just casting a glance at these histograms, we immediately see, on the one hand, the perfect agreement of the normalised frequency and probability density in the case of the white noise process we are using as reference shown in the left panel, and on the other, the evident departure from the expected density function seen in the variable process displayed in the right panel.

A departure from the expected density will generally be easy to detect, both by eye and using a goodness of fit statistic like Pearson's. Although not so useful for us now, quantifying the departure with goodness of fit would only be useful in applications where the data is being processed and categorised by a computer without a person to look at it, such as in machine learning applications. This is obviously that can be exploited for various purposes and in different circumstances, but we'll just leave it at that for now.

Computing the Periodogram

Next, we use the arrival times to calculate the periodogram using the Rayleigh statistic, exactly as we described earlier. To do this, we need to specify the range of frequencies to be tested by choosing the minimum and maximum frequencies as well as the sampling, which we will take as one frequency per IFS to keep things simple.[17]

Note that because we are using the arrival times directly, without having grouped them in any way, we need to define the frequency range over which to compute the periodogram. What should we pick for the minimum frequency? This one is easy: one over the duration, which is 10^{-4} Hz. What about for the maximum frequency? Think about it for a second. How can we go about guessing what a good maximum test frequency would be?

We said that the highest frequency we can test in a data set is the Nyquist frequency, which is twice as fast as the sampling rate. But we don't really need to go that far. In fact, it is extremely rare, from what I have seen, that it is useful to go as far as the Nyquist frequency in analysing astrophysical signals. The main reason why this is so is that the fluctuations in the power spectrum that arise from the statistical fluctuations in the time series completely dominate already at frequencies much lower than the Nyquist frequency. And so, going to higher frequencies in the periodogram simply means that it will take a *lot* longer to compute, and that the Poisson floor will extend along most of the length of the periodogram. Even extreme phenomena like kHz QPOs in black hole binaries are located at frequencies that are at least 3 orders of magnitude lower than the Nyquist frequency for instruments with $\sim \mu$s sampling rates.

Therefore, what should we use for the maximum frequency? Should the answer depend on the mean count rate? Well, that makes sense because the higher the count rate, the shorter the time between events. So we can easily go as high as the count rate in Hz. But then again, we are not, at this stage, interested in the high frequency part of the spectrum, because we want to estimate the power-law index, and this is constrained to the low frequency part of the periodogram. So, how about a maximum frequency corresponding to a time scale of 10 s, and therefore, to 0.1 Hz? Let's try that.

Figure 4 shows the Rayleigh periodograms of our two event lists between 10^{-4} and 0.1 Hz, with sampling at the independent frequencies only.[18] As expected for the white noise and discussed earlier, we indeed see that the periodogram is globally flat, with equal power at all frequencies, and sitting at the expected constant level of two. Whereas the red noise process in which there are variability structures on different time scales which are connected to one another,

[17] This is an aspect of the Rayleigh periodogram that we didn't discuss to keep the presentation on the essential elements. In fact, this is indeed an important detail pertaining to periodogram analysis: the ability to "oversample" the periodogram, but it involves some technical details that I don't want to cover at this stage to avoid muddling the basic points.

[18] How many test frequencies will that make? How many IFSs are there between 10^{-4} and 0.1 Hz?

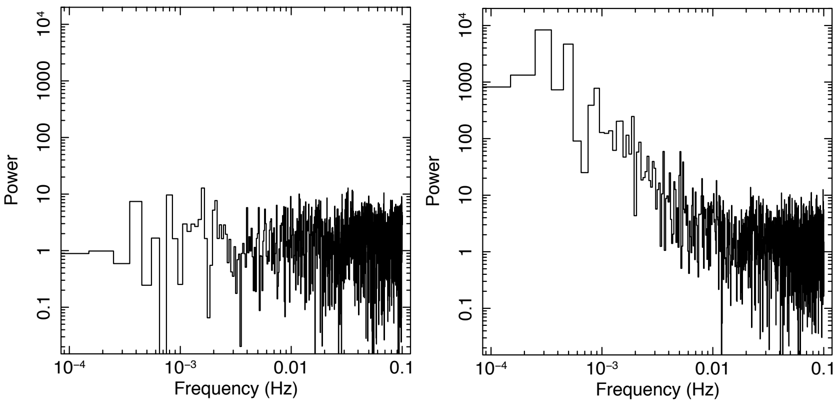

Figure 4: Rayleigh periodograms of the event data for the homogenous Poisson process (left) and the stochastically variable or red noise process, computed on the independent frequencies in the range displayed on the graphs.

leads to a power-law spectrum where the power estimates in neighbouring frequency channels are correlated that we see in the shape of the periodogram, at least up to a certain point.

And *that* point is the point of everything we've done: the time scale at which there is an equal amount of power from the inherent variability as from the fluctuations contributing to the periodogram; the time scale at which we can display the time series to reveal the structures within the data while keeping the highest possible resolution and see as much detail as we can; the time scale that is the ideal, the optimal, time scale for grouping the data. We now just need to fit the periodogram, get the best estimate of the power-law index, and determine precisely what that point, that frequency, that timescale actually is.

Fitting the periodogram

Fitting first requires defining the model, which we already know: it is a power-law plus a constant component for the Poisson floor. The model function is therefore the same as Equation (5) with the addition of the constant c, such that

$$p(x) = Nx^{-\alpha} + c, \tag{13}$$

using N for the normalisation and α for the power-law index, as before, but x instead of f for the frequency to keep it perfectly clear that it is the independent variable in the problem.

The procedure of fitting is one of varying the values of the parameters in little steps in order to find the model that best fits the data, which means minimising the different between them. For the periodogram, the optimal measure of the difference between the model and the data is the B statistic because it is constructed using the joint likelihood function defined for a collection of exponential variables with different means. Probability theory ensures us that we cannot do better than this,[19] because the shape of the probability density is automatically and seamlessly incorporated into the procedure and therefore also in the results obtained from it.

The fitting procedure using the B statistic as the similarity metric yields for the normalisation, index and constant the values $\hat{N}_B = 8.123 \times 10^{-5}$, $\hat{\alpha}_B = 2.169$ and $\hat{c} = 1.970$ ($B = 3933.84$), which were used to compute the best fit model shown in red in Figure 5 overlaid on the periodogram constructed in the previous step.[20]

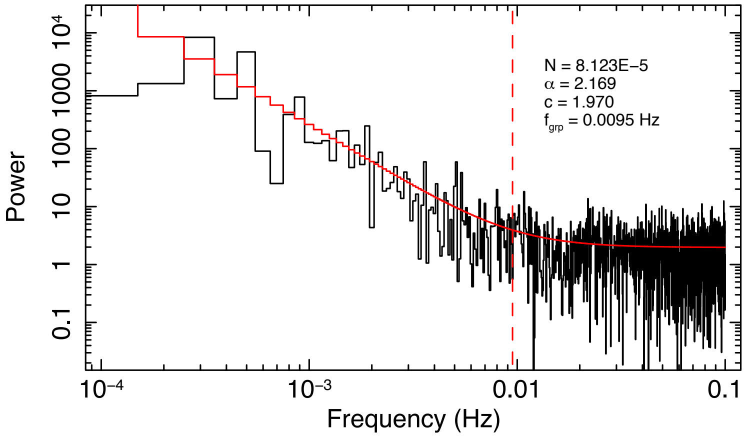

Figure 5: The red noise periodogram shown in Figure 4 on which is overlaid the best fit model of a power-law plus a constant obtained by model fitting using the B stat. The best fit values of the parameters are $\hat{N}_B = 3.7851 \times 10^{-4}$ and $\hat{\alpha}_B = 1.8940$, with c fixed at 2.

Optimal Grouping

Substituting into Equation (7) the best fit values of N, α and c: respectively $\hat{N}$, $\hat{\alpha}$ and $\hat{c}$, gives us the time scale whose value has been our primary motivation throughout this investigation.[21] Thus,

$$dt_{\text{grp}} = \left(\frac{\hat{c}}{\hat{N}}\right)^{1/\hat{\alpha}} = \left(\frac{1.970}{8.123 \times 10^{-5}}\right)^{1/2.169} = 105.218 \text{ s.} \qquad (14)$$

We are finally ready to take a look at our time series for the first time.[22] We will see it a moment. In addition, in order to visually assess, at least partially, the differences that can be seen from using different time scales on the same time series, and from this, get a more intuitive sense that an optimal time scale, must, in fact, exist, and must, in fact, be uniquely defined for every unique time series that has even the smallest level of random variability, and that for the data we have used it is indeed this time scale of 105 s that we have determined in this investigation, we will display the same data with a few different time scales, both shorter and longer than 105 s.

Examining the seven time series displayed in Figure 6, three below and three above the one binned to 105 s, we see exactly what we could have expected: Below the optimal grouping time scale, the

[19] This point *cannot* be overemphasised.

[20] To help the minimisation converge quickly, it is easy to use as initial parameter estimates those that we can compute analytically using a least squares estimator for a power-law function, restricting the fit in the pure power-law part at low frequencies. Doing this yields $\hat{N}_{LS} = 4.285 \times 10^{-4}$ and $\hat{\alpha}_{LS} = 1.817$. Uncertainties for each parameter can obviously be computed, but we leave these details out of the discussion for sake of simplicity.

[21] Substituting $\hat{N}$ and $\hat{\alpha}$ into Equation (6) to get f_{grp} yields 0.0095 Hz, shown in Figure 5 as the vertical dashed line. It is the frequency where the modelled power equals exactly 4.

[22] After all this, are you not just dying to finally see what it looks like: what these data actually *look* like?

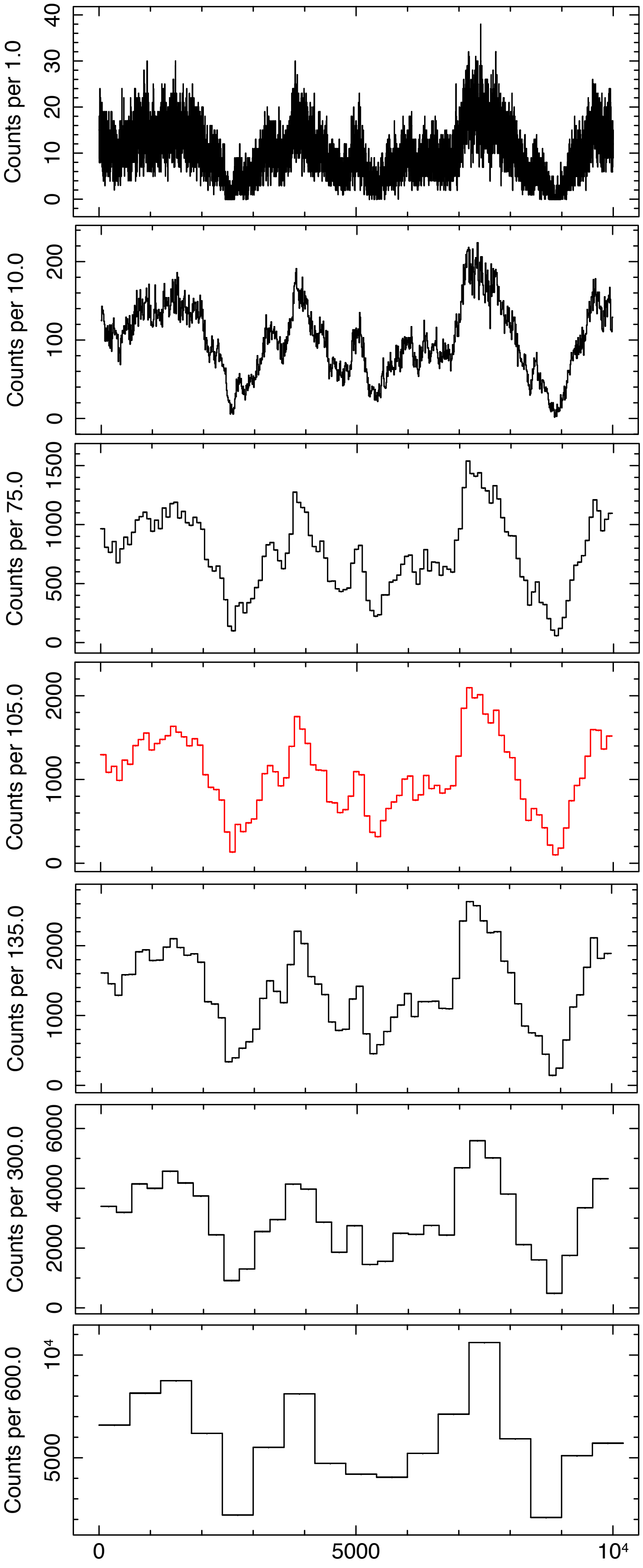

Figure 6: Time series of the red noise process under analysis constructed by grouping the data using different time scales. All are display on exactly the same time axis between -500 and 10500 s, and therefore the axis values are suppressed except for the bottom most panel for clarity.

The purpose is to illustrate the difference between these and the one derived from our analysis, and found in Equation (14). The time series grouped using the optimal time scale $dt_{grp} = 105$ s is shown in red. The other time scales used are, from the top: 1, 10 and 75 s below dt_{grp}; and then 135, 300 and 600 s above.

statistical fluctuations dominate the appearance of the time series either completely, as is the case for the first with a bin size of 1 s, or affecting mostly the fine details of the structure, as demonstrated also by the second time series binned to 10 s; above that timescale, the time series with coarser resolution continue to retain the information of the larger (300 s) to largest structures (600 s), and, in fact, more prominently, but lose all of the fine features of the shorter time scale variability.

It is quite instructive to compare the three central time series (75, 105 and 135 s) to witness the transition between just below—where we see a little more statistical fluctuation from one bin to another, to just above what was identified as the ideal timescale—where we are already losing some of the details of the fine features that are apparent in the 105 s time series.

We have completed what we set out to do. We have, however, only worked with event data up to this point. To show both a strong similarities in approach, but also highlight the differences, we will now look at a near-infrared time series binned at the instrumental level, for which we therefore do not have information about individual photons as we do with event data.

Application to Binned Data

We will perform the same steps as we did in our application to event data with three minor but important differences, all arising from the fact that the data are already grouped and thus have either never had or lost the information about individual photon arrival times. The first is that it is not possible to construct the histogram of interarrival times; instead, we will look at the distribution of time between measurements, which is determined by the length of the integration and dead time between snapshots. The second is that the periodogram must be computed differently. And the third is that the procedure for resampling a grouped time series to a different time scale brings about complications that do not arise with event data.[23] We will spend enough time on these so that hopefully, everything will be clear. Let's get to it.

Distribution of Time Between Measurements

Instead of looking at interarrival times, we look at time between measurements. The first way of looking at these, is to make a time series where we can see on the horizontal axis when the gap occurred and how long it lasted, and display again on the vertical axis its

[23] These modification in the procedure pertain to steps 1 (interarrival times), 2 (periodogram), 5 and 6 (resampling to different time scales).

duration. This allows us to see immediately where and how long the gaps are, and also identify irregularities or systematic differences in different sections of the data set, as we will see in our example.

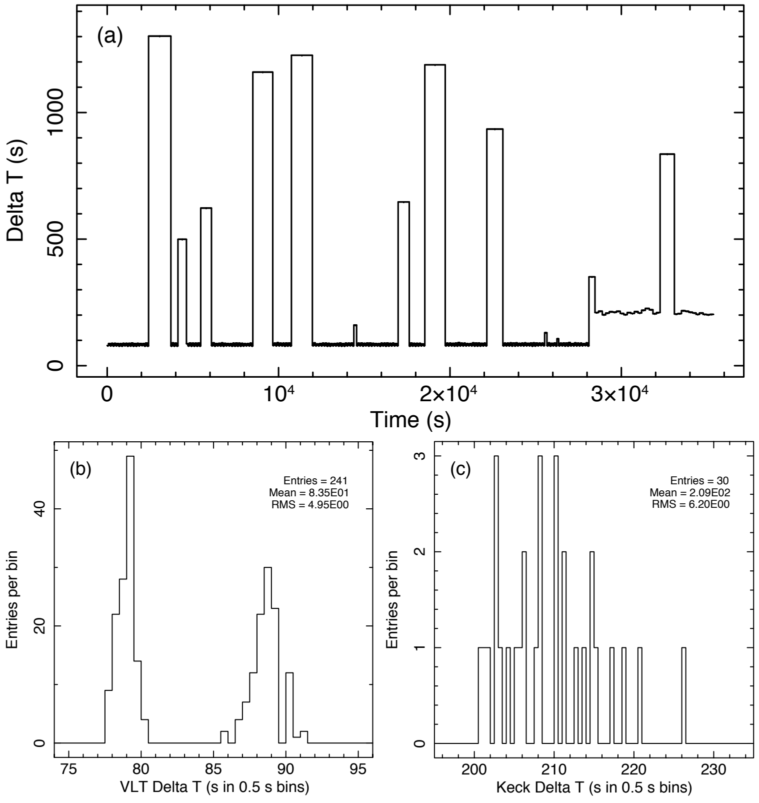

Figure 7: Time series and histograms of time between measurements for the combined data set and the individual parts (VLT and Keck). The number of entries in the histograms indicates that, in terms of the number of measurements, the VLT observation accounts for 89% of the data and the Keck for the remaining 11%.

Figure 7 shows in its top panel the time series of time between measurements. Besides the most obvious long gaps that stand out very clearly, we can also see that the baseline sits just below 100 s for most of the timeline, and jumps to about 200 s for the last part. The point of discontinuity is where the VLT observation ends and the Keck observation begins. The histograms below the time series show in detail how the time between actual measurements (excluding gaps) is distributed, and that, in fact, it is not a well-defined value as it is often reported to be in both instrumental and data analysis papers.[24]

Computing the Periodogram

To make the periodogram, instead of using the Rayleigh statistic in the form presented earlier, we must use a form that is appropriate for binned data. Here is the form of the statistic to use:

$$\mathcal{R}^2 = \frac{\left(\sum_{i=1}^{n} r_i \cos\phi_i\right)^2}{\sum_{i=1}^{n}(\sigma_i \cos\phi_i)^2} + \frac{\left(\sum_{i=1}^{n} r_i \sin\phi_i\right)^2}{\sum_{i=1}^{n}(\sigma_i \sin\phi_i)^2}, \tag{15}$$

[24] It is important to appreciate that depending on the level of detail we incorporate into the analysis of our data, these details of the duration of each snapshot, and the differences between them can be important to consider. But even if they are not, there is no excuse to be unaware of such fundamental characteristics of the data we are working with. It is therefore imperative that we look at these details at the very start of the analysis process.

where r_i is the rate and σ_i is the error for the measurement in bin i, whereas ϕ_i is the phase as before corresponding to the time in the centre of the bin, and the summing is performed on the ensemble of n measurements. Each sine and cosine of the phases is scaled by the value of the rate to transfer the information of the distribution of intensity as a function of time, and the numerators are normalised by the associated measure of variance.

But wait! Where does the measure of error for each bin come from, and how do we get that? For this data set, the rates are given by PSF fitting at the position of the source of interest in the individual images of each snapshot. This cannot give us an error on that measurement. So, how do we get an estimate of the error?

The way it was done in this case, and how it is often done in this type of observation, is maybe the most intuitive way of doing this: just pick at least one calibrator star in the field of view, a non-variable star with relatively bright and stable emission, extract its energy flux by the same PSF fitting procedure in the images, and construct a time series. This yields a companion time series made under the same instrumental conditions that we will assume to adequately represent the measurement uncertainty because it is expected to be constant, and therefore, the trends, magnitude of fluctuations, and the distribution of measurements will together give us a good idea of the appropriate uncertainty to assign to our measurements of the source flux.

Will we derive from this analysis of the calibrator star individual uncertainties for each measurement? Or a single value of an average uncertainty that can be assigned to each measurement?[25] We will get an average uncertainty, and assign to each measurement *the same* uncertainty.[26]

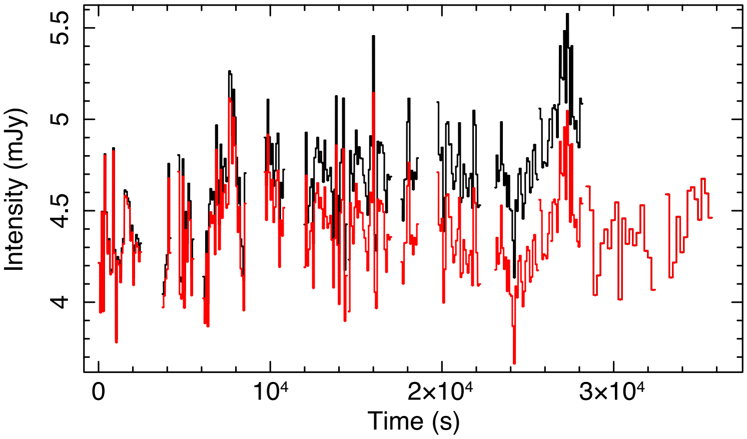

Figure 8: Time series of the flux measurements from the calibrator star that will be used to estimate the average measurement uncertainty to be assigned to the flux measurements of the source in which we are interested. The black line shows the measurements, and the red line shows the result of a liner detrending of the time series.

Figure 8 shows the time series of the calibrator. Looking at the

[25] Could we derive a measurement uncertainty for each frame?

[26] We could also, at this stage, modify this average value of uncertainty for each measurement by taking into account the differences in effective exposure time in each frame, by assigning slightly different weights: the longer the exposure, the more accurate the estimate, and vice versa. The effects in this case are minute, and thus negligible.

I use this opportunity to emphasise what good sense it makes to assign the same or practically the same uncertainty to each of the measurements in an observation performed by a given instrument over a relatively short period of time, in comparison to what is unfortunately the standard practice in high energy astronomy of assigning to each bin of grouped events an uncertainty that is derived from the square root of the number of events in that bin, as if the uncertainty in making the measurement was dependent on the number of events detected during a particular interval of time, and as if this uncertainty depended on whether the source was brighter or dimmer, or not from one moment to the next.

black line, we can see at first sight that there is an upward trend along the length of the VLT observation. This trend is also very clear, maybe even more so, in the asymmetry seen in the distribution of these flux measurements shown in Figure 9, also in black. However, removing this linear trend in the data yields a very nice and symmetric normal distribution of fluxes, whose standard deviation (labelled RMS) we use as our estimate of the average measurement uncertainty.

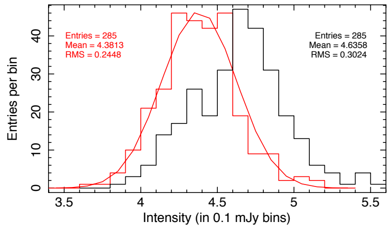

Figure 9: Histogram of the flux measurements from the calibrator star. As in Figure 8, the black line is used for the measurements, and the red for the detrended fluxes.

Do we have everything we need to compute the periodogram? We have the measurements, we have identified a simple linear trend that we have removed, and we have a good estimate of the average measurement uncertainty for the observation. Therefore, we do have everything we need to compute the periodogram, and this is what we show in Figure 10. But there are two periodograms that are similar, especially at frequencies around 10^{-3} Hz, but the on at the top extends a little more towards the lower frequencies, whereas the bottom one extends quite a bit more towards higher frequencies. Any guesses as to how this was done? Think about it for a few seconds, and here's a hint: what is it that defines the frequency range that can be tested and the IFS step size?

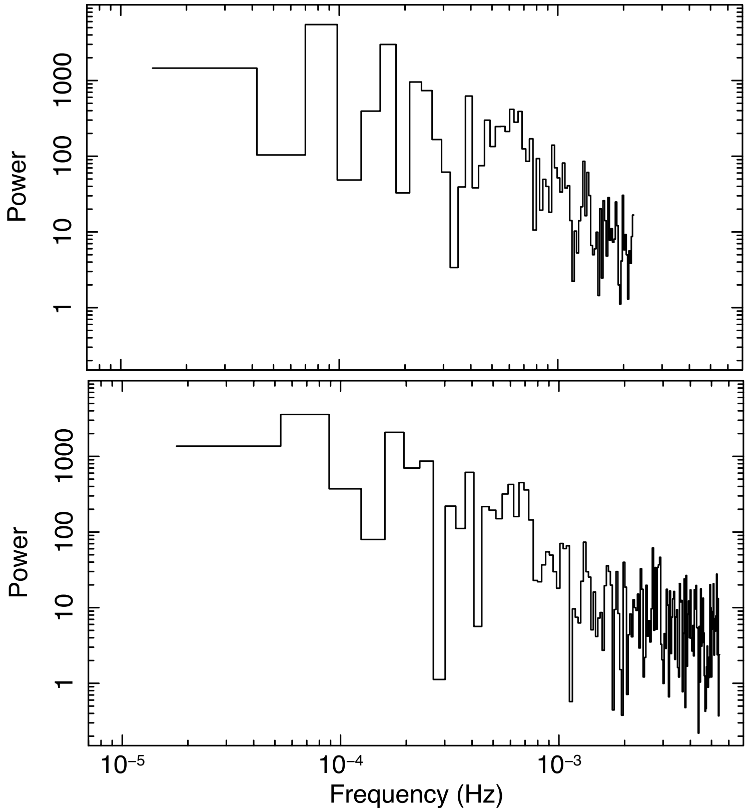

Figure 10: Modified Rayleigh periodograms of the combined VLT and Keck data set (top) computed at the independent frequencies between 2.80×10^{-5} and 2.21×10^{-3} Hz; and of the VLT data alone computed between 3.55×10^{-5} and 5.43×10^{-3} Hz.

The top periodogram was computed from the combined data set. Because it covers a longer duration, we can test slightly lower frequencies, and the step size of the IFS is smaller. However, because the sampling of the Keck data is longer (binning is coarser), and because we cannot test frequencies higher than the sampling frequency (Nyquist limit), we have to use the longest sampling time of 226 s (Figure 7 panel (c)) that translates to a maximum of 2.21×10^{-3} Hz.

The bottom periodogram was computed from the VLT data alone. In this case, the duration is a little shorter, but the maximum sampling time is 92 s (Figure 7 panel (b)), which in this case translates to a maximum frequency of 5.43×10^{-3} Hz, and therefore, as you can see, quite a few more testable frequencies in the upper range.

Looking at the higher frequency end of these periodograms, the most important feature for us now in trying to identify the ideal grouping time scale, is that in there is no clear flattening from reaching the Poisson floor: definitely not in the periodogram of the combined data set, and maybe not as clearly from the VLT data, but still obviously well above a power level of two where sits the power from statistical fluctuations.

Another important question we must ask at this stage is how much of the structure in the periodogram is due to the sampling function, to how data were sampled.[27] In the previous example we didn't even have to think about this because we had continuous sampling without interruptions and, most crucially, without the application of an inherent grouping of the data during acquisition, which is the case here. Hence, this is the next thing we have to look at: what would the *expected* periodogram of a white noise process look like if we applied to it the sampling function of our near-infrared data? And what would the periodograms of the data look like if we actually took away this power due to the sampling function?

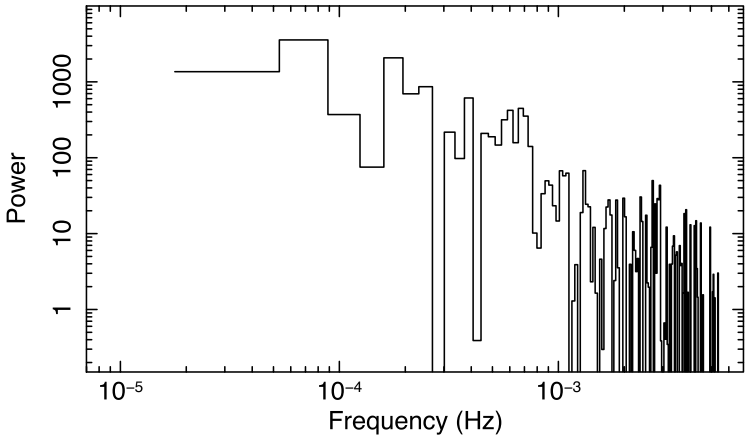

Figure 11: Periodogram of the VLT data as shown in Figure 10, but from which was subtracted the power attributable to the sampling function.

[27] And don't believe for one second that the Lomb-Scargle periodogram is not affected by the sampling function: it is! And in fact, it is virtually indistinguishable in shape from the modified Rayleigh periodogram we used here.

Looking at the periodograms in Figure 12 of the sampling functions for the combined data set and for the VLT data alone, we first notice that they are not perfectly flat at the level of two as would be expected from a uniform sampling of white noise, and we notice that they are different, especially at the lowest and highest frequencies, but nonetheless similar in the range around 10^{-3} Hz. At low frequencies the combined data set, the coarser binning and sampling gaps causes a rise in the power that is not present in the VLT sampling.

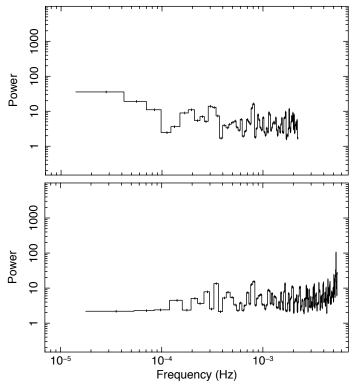

Figure 12: Periodograms of the respective sampling functions for the combined VLT and Keck data set (top), and of the VLT data alone. The test frequencies and axes are identical to those of Figure 10. These were derived from the average of 50 simulated white noise data sets binned as the data.

On the other hand, the finer VLT sampling and gap structure causes a rise in power at the highest frequencies. These effects are expected, of course, but the important point is that it is possible to calculate precisely what is the contribution to the power estimates in the periodogram that is due to the way the data are sampled, and that this contribution should not be neglected or ignored when working in Fourier space.

The adjusted data periodogram shown in Figure 11 for the VLT data that contains quite a bit more of the high frequency information in which we are interested in order to identify the inflection point from which begins the Poisson floor, does not give such an indication because the power-law does not break and flatten. What does this tell us, besides the fact that there is no point in fitting the periodogram to find the grouping time scale? It tells us that the sampling is too low for the signal to noise of the data, which tells us that we have lost information about the variability structure contained in these data by grouping them how it was done: the integration time was too long.

Therefore, the best we can do for viewing this time series is to use the instrumental binning with as much resolution as is available. The time series in full detail is shown in Figure 13: The VLT data were detrended based on the time series of the calibrator star (Figure 8), and the size of the error bar on the intensity was derived from the distribution of calibrator flux measurements (Figure 9).

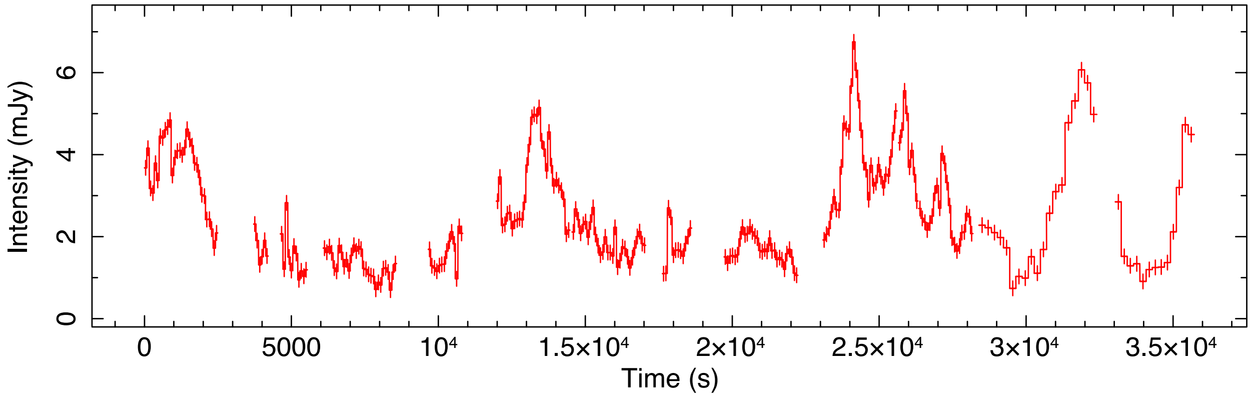

Figure 13: Time series of the combined VLT and Keck data sets displayed using the maximum (instrumental) resolution.

Resampling Binned Data

It turned out that in this case we do not need to resample,[28] because the data were already grouped at the instrumental level on a time scale that is longer than what we defined earlier as the optimal grouping time scale to reveal structure. But what if we needed to, or simply wanted to to see what it looks like on different time scales: how would we do it? How do *you* do this in your own work? Let's state the problem explicitly, and at the same time emphasise that this is *not* an easy problem, as you will see in a moment.

We have an ensemble of measurements as a function of time. Each of these measurements has an associated uncertainty. We cannot assume that these uncertainties are either the same for each measurement, or that they are in any way proportional to the measured value. We don't know how they were derived. Therefore, we have to treat each measurement and each uncertainty as they are.

We want to resample the time series of already binned measurements in a way that *preserves* intensity, *preserves* the statistical character of the data, and *preserves* the structures and trends in the data as much as possible, and without the need for complicated modelling of the data in order to do it. The key concern is conservation.

The method involves calculating the intensity for each part of the bin based on the local trends, and then introducing noise (Poisson, normal or otherwise) to preserve variance, while ensuring that the intensity is always preserved. The introduction of random fluctuations to the intensity in a bin, naturally induces "diffusion" of the intensity to the neighbouring bins. This is inevitable. Nonetheless, using a mechanism to preserve the trends on either side of the bin works to minimise the diffusion.

The diagram in Figure 14 helps illustrate the method. It shows a detailed view of three adjacent bins with different intensities and uncertainties. The centre of the central bin is labelled x_i, and the centre of the bins preceding and following it are respectively labelled x_{i-1} and x_{i+1}.

Similarly, the rates or intensities corresponding to these three bins are labelled r_{i-1}, r_i and r_{i+1}, from left to right. And since we are considering the splitting of the central bin into two uneven parts, we label the uncertainty associated with its rate simply as σ, without a subscript.

Further quantities that we define are the width of the central bin, δx, (also without subscript), and the widths of the two parts of the bin resulting from the split: δ_- on the negative, left side, and δ_+ on the positive, right side. The slopes of the lines connecting r_{i-1} to r_i, and r_i to r_{i+1}, are labelled m_- and m_+, respectively.

[28] The word "resample" is used because we are modifying the sampling in a very different sense than when we group bins together or "rebin" without having to split bins apart, which is what is done when we resample.

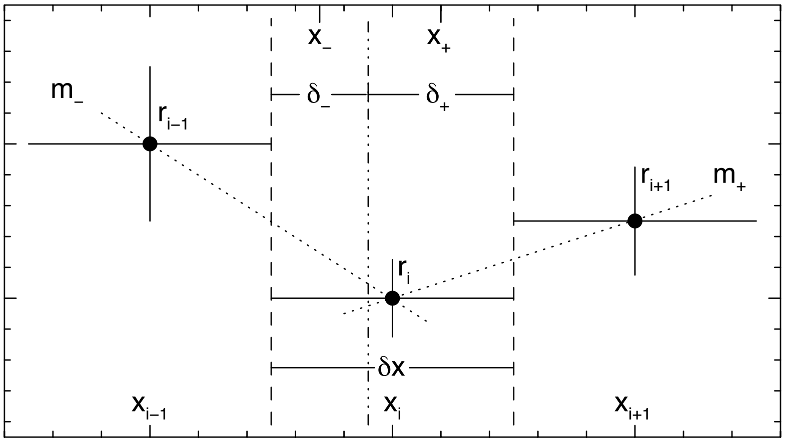

Figure 14: Diagram illustrating the splitting of the central bin into two new bins, and the variables that we use in the mathematical description of the process.

Our aim is to derive all the relevant quantities for the two new bins, including their respective centres: x_- and x_+, rates: r_- and r_+, and uncertainties: σ_- and σ_+.

Splitting and Resampling

The central bin is split into two parts that in general will be unequal. We refer to these as bin δ_-, on the negative side, and bin δ_+, on the positive side, based on their respective widths.

To preserve the trends in the data while keeping to the simplest possible model, we determine the rates that we will assign to each part of the bin using a linear model based on the adjacent bin heights in relation to that of the central bin.

The Negative Side — For bin δ_-, the linear model for the rate is

$$r_-(x) = m_- x + b_- \tag{16}$$

where the slope m_- is given by

$$m_- = \frac{r_i - r_{i-1}}{x_i - x_{i-1}} \tag{17}$$

and since the line passes through r_i, the ordinate is

$$b_- = r_i - m_- x_i \tag{18}$$

Substituting Eq. 18 into Eq. 16 yeilds

$$r_-(x) = r_i + m_-(x_- - x_i) \tag{19}$$

where x_-, the centre of the bin δ_-, is given by

$$x_- = x_i - \frac{1}{2}(\delta x - \delta_-) \tag{20}$$

The Positive Side — For bin δ_+, the linear model is

$$r_+(x) = m_+ x + b_+ \tag{21}$$

and since

$$m_+ = \frac{r_{i+1} - r_i}{x_{i+1} - x_i} \quad \text{and} \quad b_+ = r_i - m_+ x_i \tag{22}$$

we find in exact analogy to bin δ_- that

$$r_+(x) = r_i + m_+(x_+ - x_i) \tag{23}$$

where x_+, the centre of the bin δ_+, is

$$x_+ = x_i + \frac{1}{2}(\delta x - \delta_+) \tag{24}$$

The New Rates — Expressing the rates for bins δ_- and δ_+ in terms of known quantities we find that:
rate r_- is given by

$$\begin{aligned} r_-(x) &= r_i + \left(\frac{r_i - r_{i-1}}{x_i - x_{i-1}}\right)[x_i - \frac{1}{2}(\delta x - \delta_-) - x_i] \\ &= r_i - \frac{1}{2}(\delta x - \delta_-)\left(\frac{r_i - r_{i-1}}{x_i - x_{i-1}}\right) \end{aligned} \tag{25}$$

and rate r_+ is given by

$$\begin{aligned} r_+(x) &= r_i + \left(\frac{r_{i+1} - r_i}{x_{i+1} - x_i}\right)[x_i + \frac{1}{2}(\delta x - \delta_+) - x_i] \\ &= r_i + \frac{1}{2}(\delta x - \delta_+)\left(\frac{r_{i+1} - r_i}{x_{i+1} - x_i}\right) \end{aligned} \tag{26}$$

The Uncertainty on the New Rates — If the original binned time series has error bars, then in order derive the uncertainties associated with the rates for the two new bins, σ_- and σ_+, we require that the equation for the weighted mean holds true. This implies that

$$r_i = \frac{r_-/\sigma_-^2 + r_+/\sigma_+^2}{1/\sigma_-^2 + 1/\sigma_+^2} \quad \text{and} \quad \sigma_i^2 = \frac{1}{1/\sigma_-^2 + 1/\sigma_+^2} \tag{27}$$

But we also need to 'distribute' the uncertainty from the central bin correctly between the two new bins. And so we define two more variables, the weighting factors k_- and k_+, such that

$$\sigma_-^2 \equiv \frac{\sigma^2}{k_-} \quad \text{and} \quad \sigma_+^2 \equiv \frac{\sigma^2}{k_+} \tag{28}$$

Since

$$\sigma_i^2 = \frac{1}{k_-/\sigma_-^2 + k_+/\sigma_+^2} = \frac{\sigma^2}{k_- + k_+} \tag{29}$$

we find that $k_- + k_+ = 1$. Therefore, the natural solution is to use the bin fractions as the weighting factors:

$$k_- = \frac{\delta_-}{\delta x} \quad \text{and} \quad k_+ = \frac{\delta_+}{\delta x} \tag{30}$$

which naturally satisfies the condition that

$$\frac{\delta_-}{\delta x} + \frac{\delta_+}{\delta x} = 1 \tag{31}$$

Finally, substituting Eq. 30 into Eq. 28 yields

$$\sigma_-^2 = \sigma^2 \frac{\delta x}{\delta_-} \quad \text{and} \quad \sigma_+^2 = \sigma^2 \frac{\delta x}{\delta_+} \tag{32}$$

So that each part of the split central bin carries a portion of the original uncertainty that is inversely proportional to its width.

Summary of Results — Splitting the central bin into two yields: for bin δ_-

$$r_-(x) = r_i - \frac{1}{2}(\delta x - \delta_-)\left(\frac{r_i - r_{i-1}}{x_i - x_{i-1}}\right) \tag{33}$$

$$\sigma_-^2 = \sigma^2 \frac{\delta x}{\delta_-} \tag{34}$$

and for bin δ_+

$$r_+(x) = r_i + \frac{1}{2}(\delta x - \delta_+)\left(\frac{r_{i+1} - r_i}{x_{i+1} - x_i}\right) \tag{35}$$

$$\sigma_+^2 = \sigma^2 \frac{\delta x}{\delta_+} \tag{36}$$

To *preserve the trend* in the data is guaranteed by distributing the intensity according to Equations 33 and 35. These give the baseline intensities for the two new bins based on the linear models tracing the change in intensity from the adjacent bins on either side to the central bin being split. How well these trends are preserved depends on the magnitude of the difference in intensity: if the random noise component is of a magnitude comparable to the difference in intensity between the adjacent and the central bin, then the trend is lost. But this is as it should be since we are interested in preserving only those trends that are statistically significant.

To *preserve the variance* as much as is possible, we use the widest of the two new bins (that has the most reliable estimate of intensity), assign it an intensity proportional to its fraction of the original bin, and replace that by a pseudo-random number drawn from the appropriate distribution (maybe most often Poisson or normal). This ensures that the magnitude of statistical fluctuations in the original data remains the same on average.

Finally, to *preserve the intensity*, which is of the most fundamental importance, we assign to the adjacent new bin (the smaller of the two) the difference between the total intensity in the central bin before splitting, and the pseudo-random number drawn in the previous step. This guarantees that splitting a bin will never 'create' or 'destroy' intensity.[29]

Naturally, the logical thing to do at this point is to illustrate this resampling method, and evaluate how well it can preserve those quantities—trend, variance and intensity—that are most important to preserve. What I will do at this stage, though, is leave the most courageous and interested among you to do it yourself: code the resampling up, think of a way to evaluate, and see what you find when you do this. I will be very happy to discuss this with you in greater details.[30]

Concluding Remarks

As we started with "Introductory Remarks", it seems fitting to end with "Concluding Remarks". What we have seen and discussed together are all things related data analysis and statistics of time series, binned and unbinned, continuously sampled and with structured gaps, whose aspects we examined both in time and frequency space. We have looked at how different random variables are distributed differently, how the combination of certain kinds of variables can lead to different kinds, and how in the end and in the beginning, the most important element in our analysis is the use of the appropriate probability density function for the way in which the measurements are distributed as a means to construct the appropriate likelihood function that allows us to compare these data with whatever expectation or model we may have had about them.

I find it interesting, very interesting, that is order to answer such a simple question: "how should I group these data in order to reveal the structure within them?", we were brought to look at so many different things and so many different aspects of Time Domain Astronomy. I hope you found it interesting, engaging, instructive and, maybe most importantly, inspiring and motivating to explore the data further and always strive to improve on the methods and techniques you use.

And instead of going through the body of what was presented as is the traditional way to conclude, I will end by mentioning the three topics that I didn't discuss, but for which I would have loved to have time. The first is period searches, with all the details of constructing periodogram statistics, and comparing their behaviour in different

[29] The pseudo-random number drawn to define the number of events in the larger of the two new bins is not allowed to be greater than the intensity of the original central bin, as this would lead to a negative intensity for the adjacent new bin. If this occurs, we draw again until this requirement is satisfied.

[30] One way in which we could illustrate the method, would be to use an event list, and using a number of random grouping time scales (bin widths) from small to larger in comparison to the time span of the data, step through these in ascending order, group the events with this bin time to create our reference time series, and compare the result of this to the resampling of the previously binned time series.

circumstances, types of data and kinds of signals. The second is modelling stochastic or red noise, with the details relating to finding the means to estimate the properties of the emission at the source from the data at hand, as well as evaluating the probabilities or likelihoods of detecting a given strength of periodic modulation in a red noise background. And the third is the characterisation and subsequent categorisation or classification of data based on their time and frequency domain properties. I hope to have the chance to present and discuss these in a setting akin to this one, maybe next year if things go well with our administration and the faculty that funded this workshop. Thank you for listening.