TSDA Lecture 6

For all work on Fourier series, see `vanderplas_lombscargle.pdf` in Reading folder on cloudcape.

In this lecture:

1. Nyquist frequency & FFT
2. What's a window function?
3. Uneven sampling
4. Lomb Scargle periodogram
5. Choosing a frequency range

Nyquist frequency

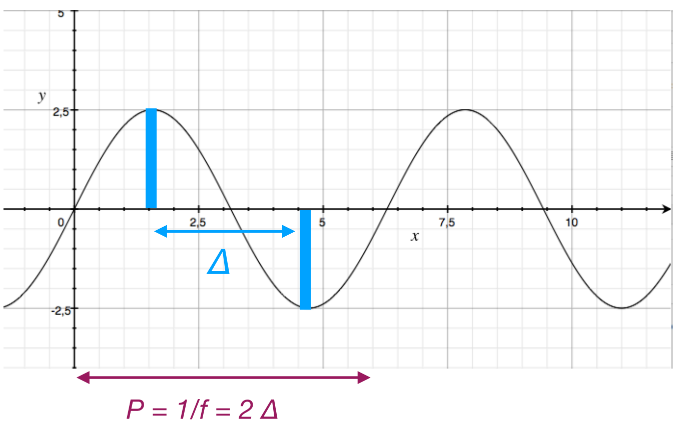
$$f_c \equiv \frac{1}{2\Delta}$$

where Δ is the sampling interval.

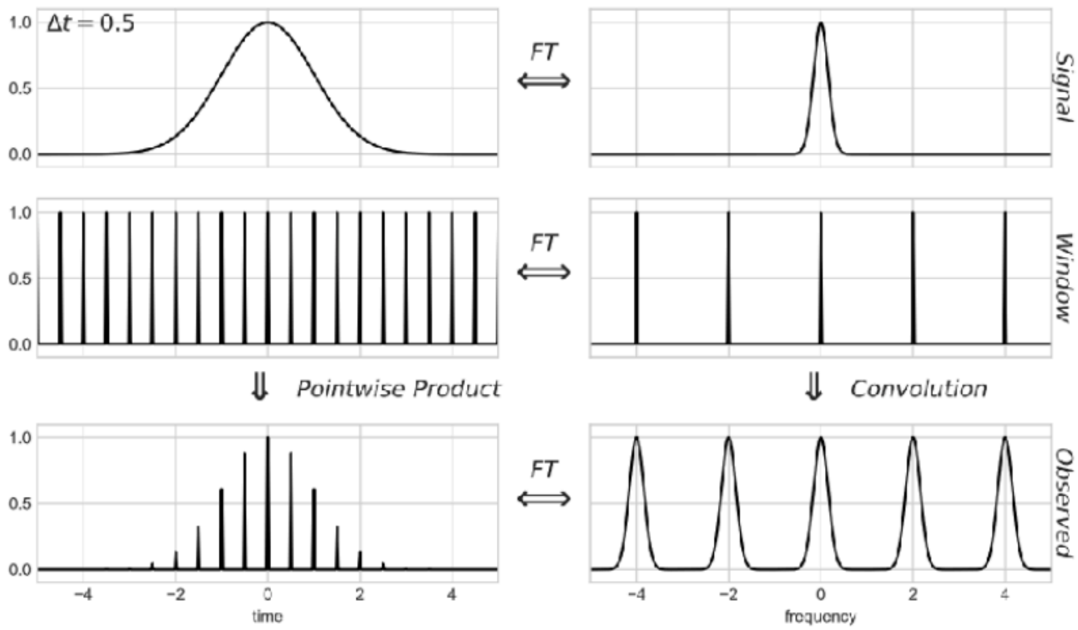
If a sine wave of frequency f_c is sampled at its positive peak value, the next sample in Δ will be at the trough.

Critical sampling of a sine wave is 2 points per cycle.

Undersampling of the sinusoid allows there to be a lower-frequency alias, which is a different function that produces the same set of samples.



DFT - sampling



Nyquist good news: Sampling theorem

- If a continuous function *h(t)*, which is sampled at an interval *Δ,* happens to be bandwidth limited to frequencies $f < f_c$, then *h(t)* is completely determined by its samples h_n.
- This is the **sampling theorem**.
- Think about a signal passed through an amplifier with a known frequency response.

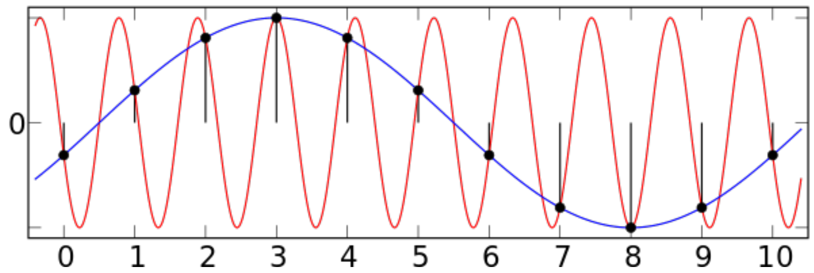
Nyquist bad news: Aliasing

- If you sample a continuous function that is **not** bandwidth limited to less than the Nyquist frequency, then power at frequencies outside the range $-f_c < f < f_c$ is spuriously moved into that range.

- This is called **aliasing**

Nyquist bad news: Aliasing

- Consider a real signal (red) of frequency 9 cycles per 10 seconds, i.e. $f_{red} = 0.9$.
- You are sampling at $\Delta = 1s$, i.e. $f_s = 1$.
- The real frequency (f_{red}) is above the Nyquist frequency ($f_c = 0.5$) so you don't detect it
- But the real period can beat with the sampling period to give a frequency below the Nyquist frequency $f_{blue} = f_s - f_{red}$
- So power from a frequency outside your critical range has "leaked" or "aliased" into your range.



Fast Fourier Transform

- Algorithm that does the transform of N points in a time proportion to $N \log N$, rather than the N^2 timing of a brute force implementation. Cooley & Tukey 1965

- Probably one of the most well used transforms on the planet.

https://rob-bell.net/2009/06/a-beginners-guide-to-big-o-notation/

Power Spectral Density & Periodogram

- The *power spectral density* (or power spectrum) is given by:

$$\mathcal{P}_g \equiv |\mathcal{F}\{g\}|^2$$

where $F(g) = \hat{g}$ (see earlier in course notes) is the Fourier transform of g (a function of t)

- The *classical periodogram* is defined by Schuster (1898) as:

$$P_S(f) = \frac{1}{N}\left|\sum_{n=1}^{N} g_n e^{-2\pi i f t_n}\right|^2$$

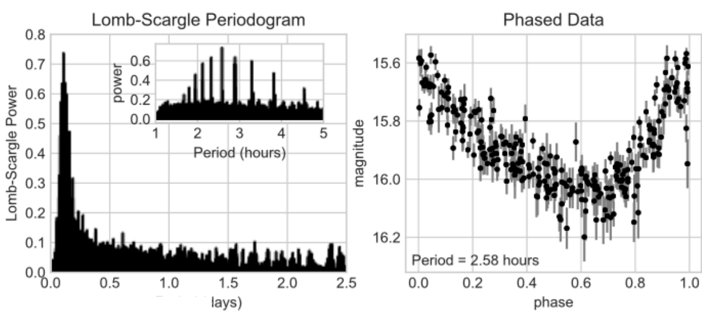
while having a similar form as above, it's defined in the context of the discrete Fourier transform.

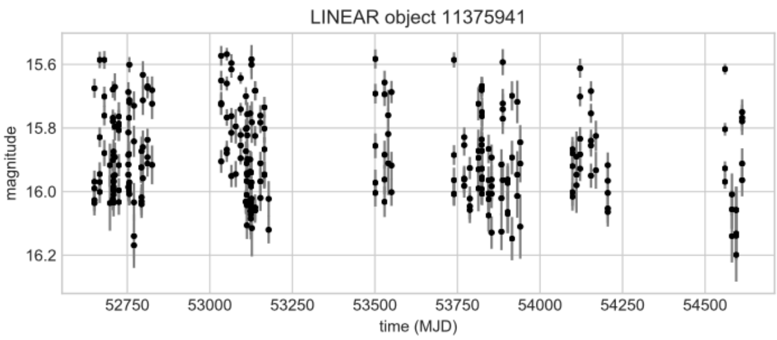
Power Spectral Density & Periodogram

- While astronomers often use the terms *power spectrum* and *periodogram* interchangeably, we should be aware that the periodogram is an estimate of the power spectrum.
- This is because we're estimating the power spectrum (which is for a continuous underlying function) by discrete samples of this continuous function.
- Also, there are a few other quirks related to periodograms…

Lomb-Scargle periodogram

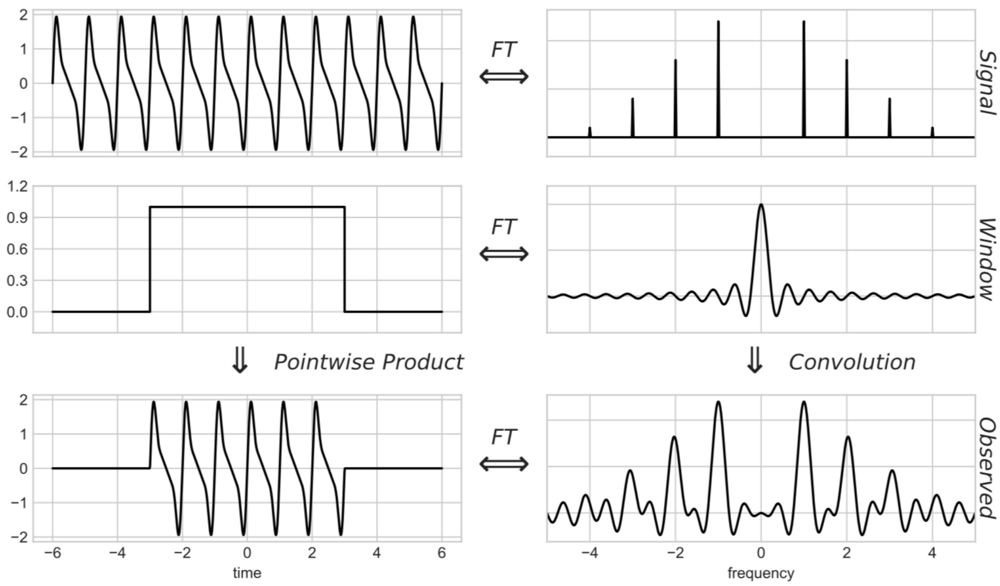
- Widely used in astronomy
- Can deal with uneven sampling





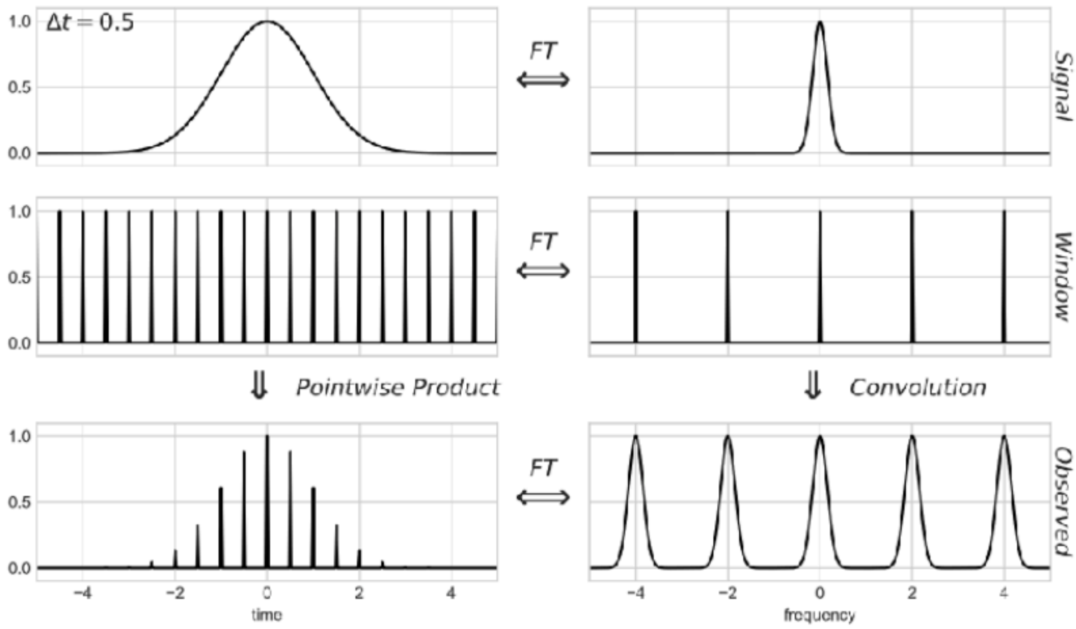
Window function

Remember the convolution theorem:



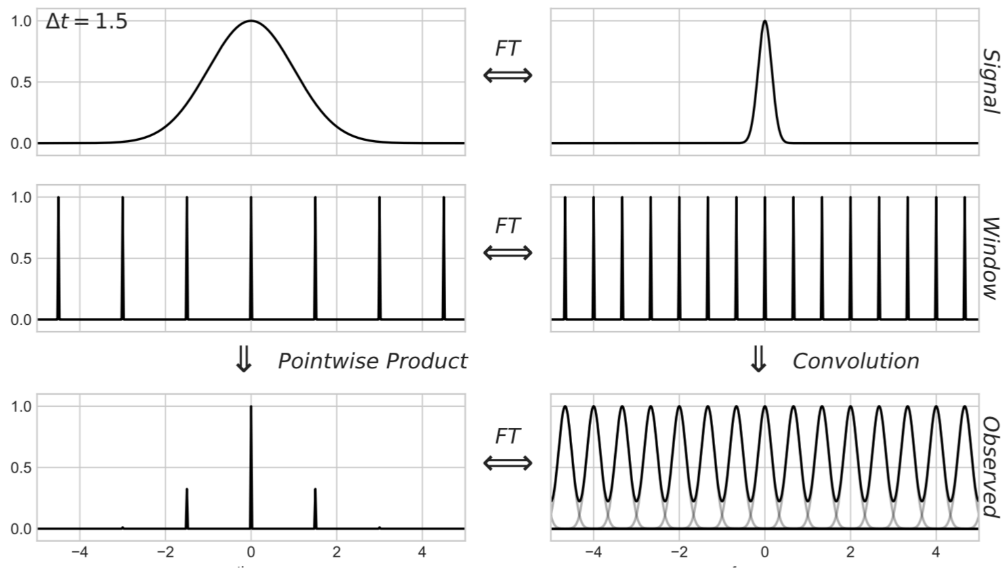
Window function

- Observed signal: $g_{\mathrm{obs}}(t) = g(t)W(t)$



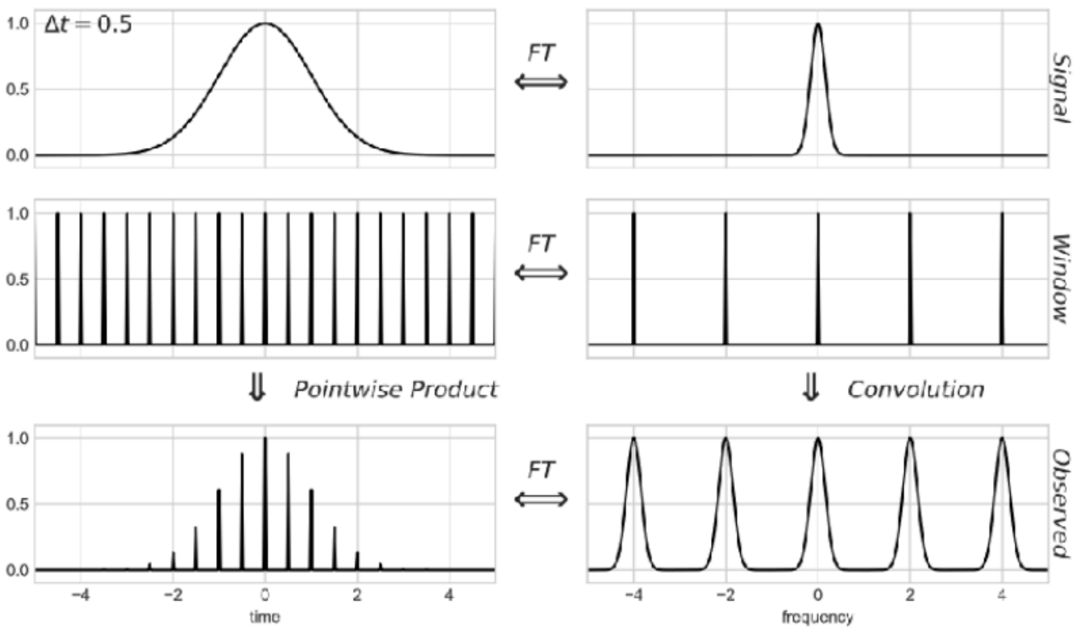
Window function

- Another way to think about the Nyquist limit:
- Same as figure above, but sampling is 1.5 d instead of 0.5

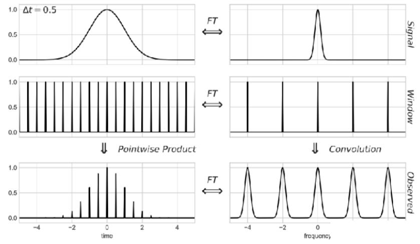


Window function

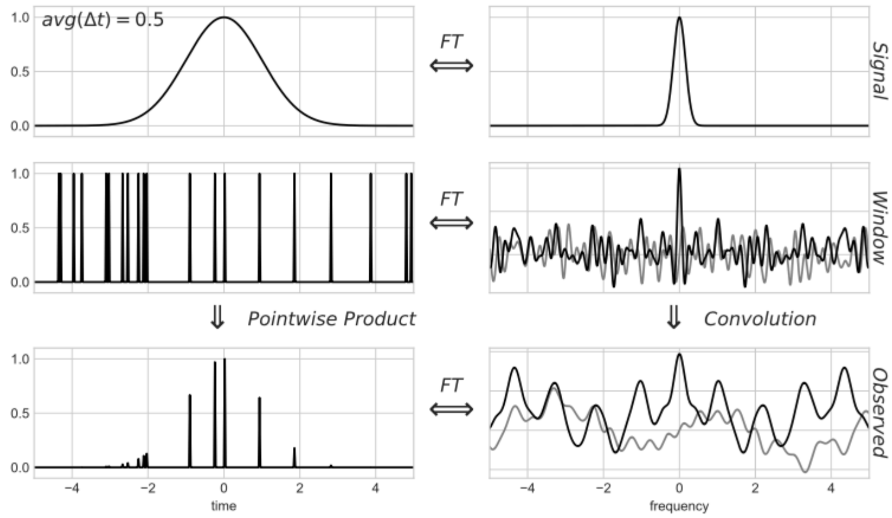
- Observed signal: $g_{obs}(t) = g(t)W(t)$



Uneven sampling



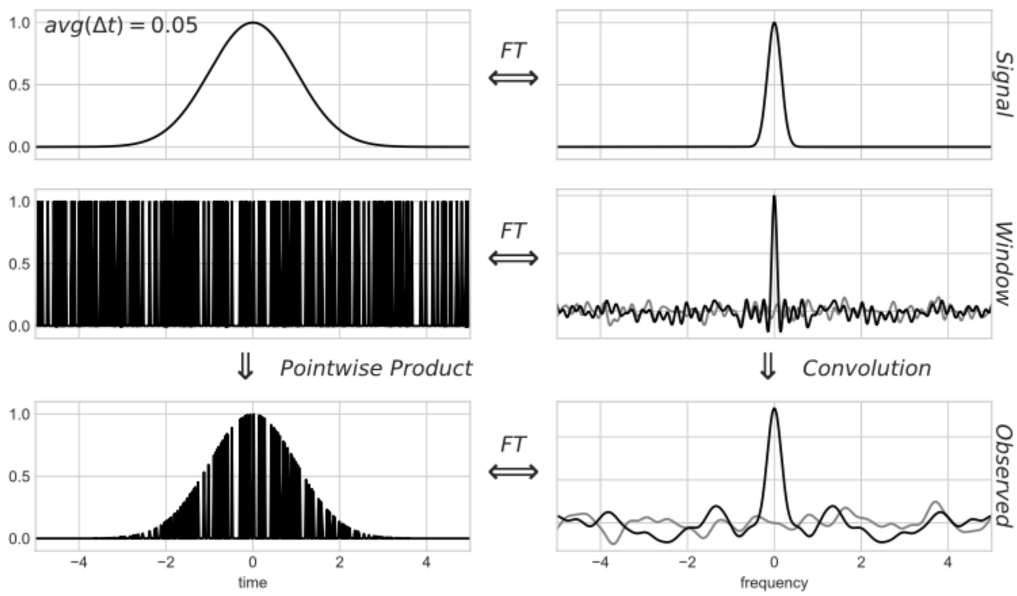
- FT of sampling looks like random noise
- No exact aliasing of the true signal



Nyquist frequency undefined

Uneven sampling

- Denser sampling helps, but doesn't get rid of the noise in the FT.



Lomb-Scargle periodogram

$$P_{LS}(f) = \frac{1}{2}\left\{ \left(\sum_n g_n \cos(2\pi f[t_n - \tau])\right)^2 \Big/ \sum_n \cos^2(2\pi f[t_n - \tau]) + \left(\sum_n g_n \sin(2\pi f[t_n - \tau])\right)^2 \Big/ \sum_n \sin^2(2\pi f[t_n - \tau]) \right\}$$

$$\tau = \frac{1}{4\pi f} \tan^{-1}\left(\frac{\sum_n \sin(4\pi f t_n)}{\sum_n \cos(4\pi f t_n)}\right)$$

where *tau* is specified for each *f* to ensure time-shift invariance

1. Periodogram reduces to classical form in the case of equally spaced observations

2. Periodogram's statistics are analytically computable

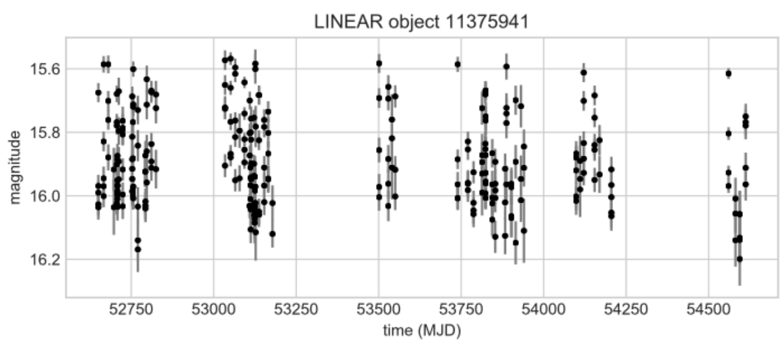
3. Periodogram is insensitive to global time-shifts in data

Lomb-Scargle periodogram

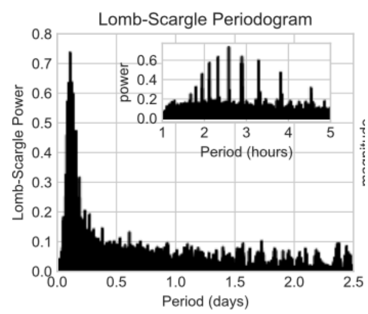
- Does the Nyquist limit exist in unevenly sampled data?

- Not in the same sense as for the DFT - it is a direct result of the even sampling of a Dirac comb.

- Some kinds of aliasing can still occur though, so best to be careful and use simulations.



Mean sampling 7 days



Period recovered 2.58 hours!

Lomb Scargle Periodogram

The non-uniform Nyquist limit

Eyer & Bartholdi (1999), Koen (2006)

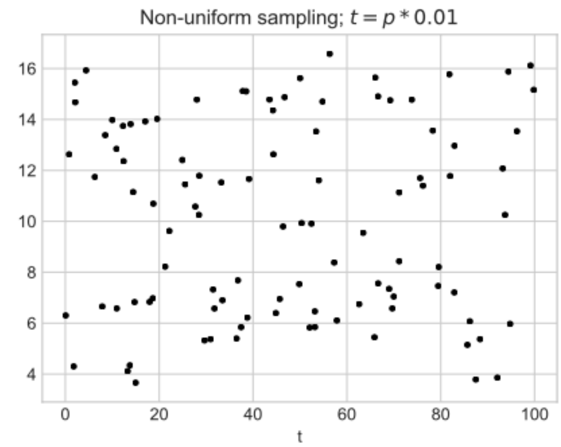
$$\Delta t_i = t_i - t_0 = n_i p$$

Choose *p* to the largest value that allows each time interval to be written as an exact integer multiple of this factor *p*.

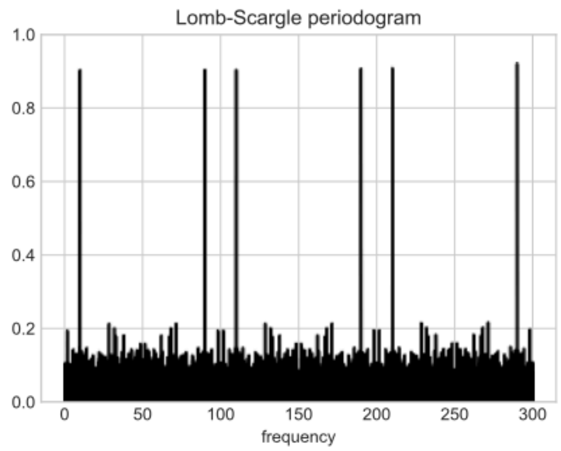
Lomb Scargle Periodogram

The non-uniform Nyquist limit

Then Nyquist frequency is $f_{\mathrm{Ny}} = \dfrac{1}{2p}$

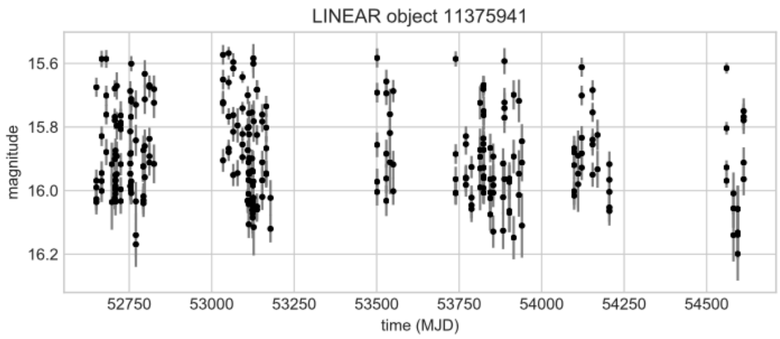


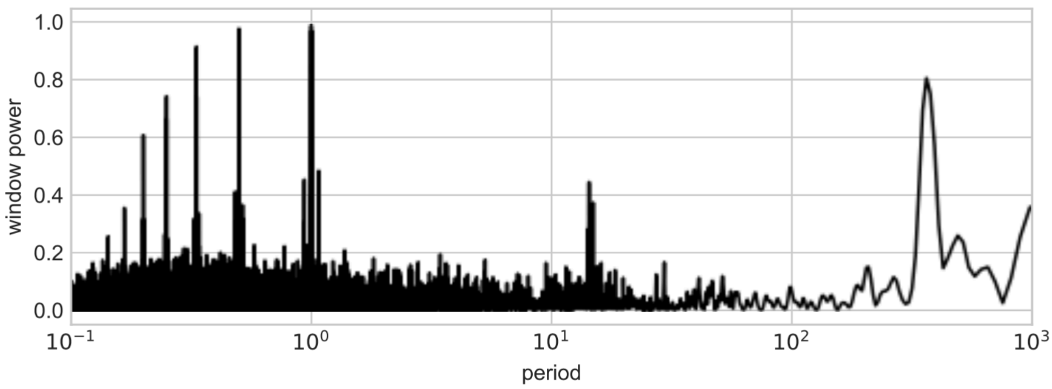
$p = 0.01$



$f_{\mathrm{Ny}} = 50$

Lomb Scargle window power



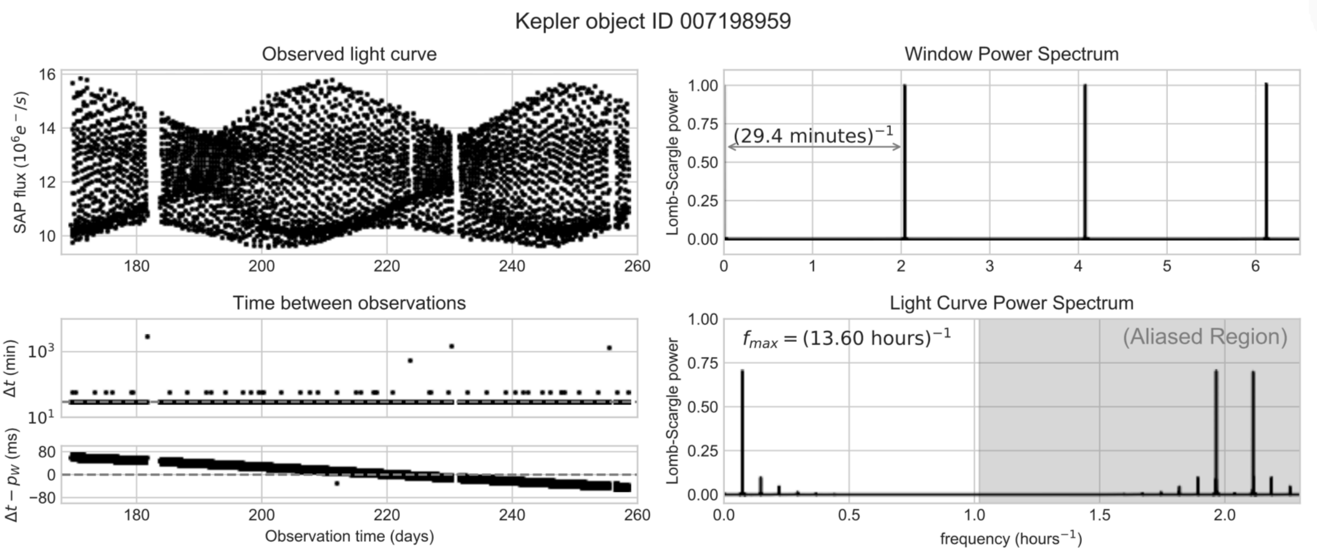


Frequency grid

- For DFT this is obvious: N points between $-f_c$ and f_c
- For Lomb-Scargle there are choices to be made: *frequency limits* and *grid spacing*
- f_{low} = either $1/T_{\{data\ length\}}$, or 0
- f_{high} = pseudo-Nyquist, window f^n, prior knowledge?

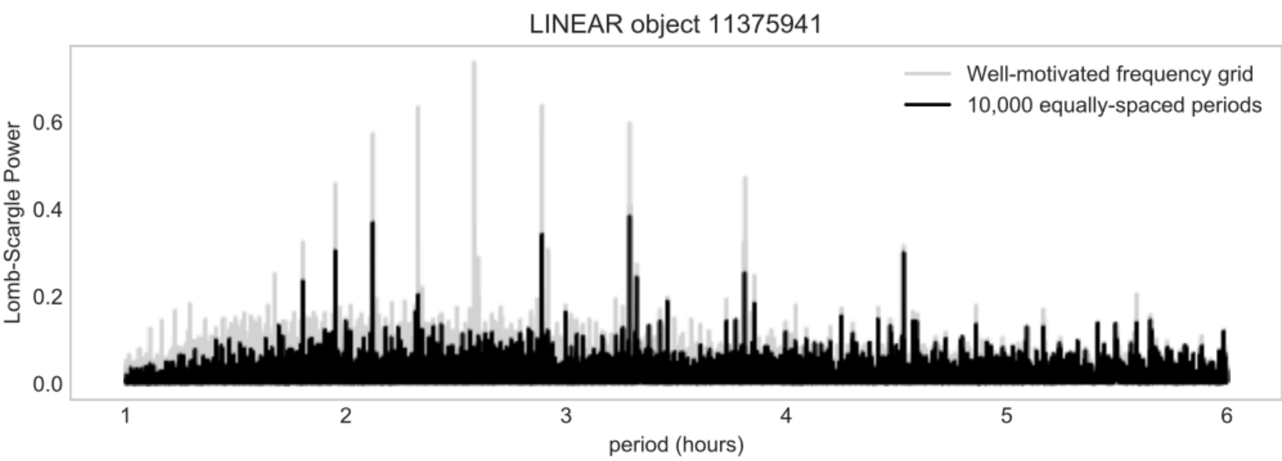
Frequency grid

Upper and lower limits



Frequency grid

How finely do we sample frequencies between f_{low} and f_{high}?



200 000 equally spaced frequencies

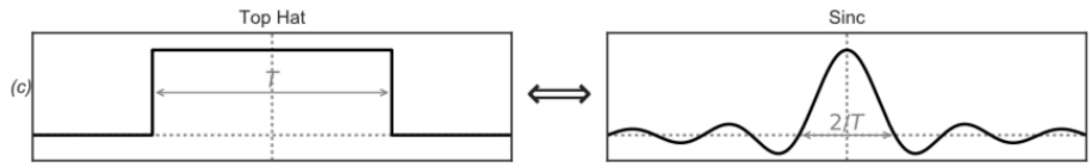
10 000 equally spaced periods

Frequency (not period) grid

How finely do we sample frequencies between f_{low} and f_{high}?

Choose grid spacing smaller than the expected widths of the periodogram peaks.

Data viewed through rectangular window of length T has sinc-shaped peaks of width ~1/T

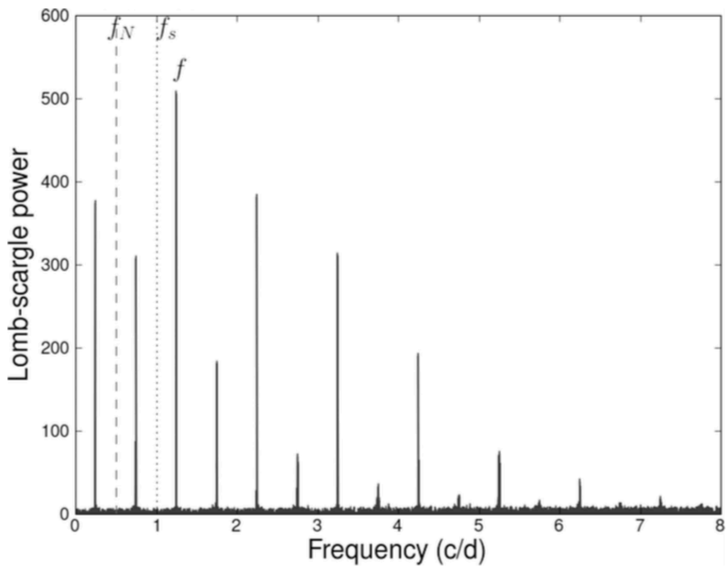


So, oversample the peak, so you don't miss it, e.g.

$$\Delta f = \frac{1}{n_0 T}$$

where n_0 is generally between 5 and 10.

Aliasing and window functions



f_s = sampling frequency

f = injected frequency (1.25 c/d)

f_N = badly estimated "Nyquist"

Bird, Coe & McBride 2012

False Alarm probability

What is the chance that a dataset with **no signal** would lead to a peak in the periodogram of the magnitude we found?

$$P_{\text{single}} = 1 - \exp(-Z)$$

where *Z = P(f_0)* is the periodogram value at f_0. Here P_{single} is the probability of getting this power at a single frequency f_0.

What is more useful, is the distribution of the highest peak in the periodogram. This can be quantified as the false alarm probability (FAP):

$$\text{FAP}(z) \approx 1 - [P_{\text{single}}(z)]^{N_{\text{eff}}}$$

where N_{eff} is the **effective number of independent frequencies**. If the expected peak width in the periodogram is $\delta f = 1/T$ and you're probing peaks in the range 0 to f_{max}, then $N_{\text{eff}} = f_{\text{max}}T$

False Alarm probability

Another way of estimating the false alarm probability is based on theory for extremes in stochastic processes (Baluev 2008):

$$\text{FAP}(z) \approx 1 - P_{\text{single}}(z)e^{-\tau(z)}$$

$$\tau(z) \approx W(1-z)^{(N-4)/2}\sqrt{z}$$

$$W = f_{\text{max}}\sqrt{4\pi\,\text{var}(t)}$$

Here W is an effective width of the observing window in units of the maximum frequency probed. The Baluev estimate provides an upper limit for the FAP where the window function does not have much time structure.

Significance = 1 - FAP

Remember the false alarm probability estimates the probability of observing a peak at some power in the periodogram in the presence of only Gaussian noise, and in the absence of a period.

This means that the estimates in the slides are not valid where in case where the noise has a frequency dependence, e.g red noise or pink noise (also called shot noise).

Significance through bootstrapping

1. Generate *N* light curves with the same time structure as original light curve, but with scrambled *y*-values.

2. Calculate periodogram for each light curve.

3. Find maximum power for each periodogram.

4. From these N value of the maximum power, you can calculate significance levels.

5. Plot on your periodogram

6. How big does *N* need to be?

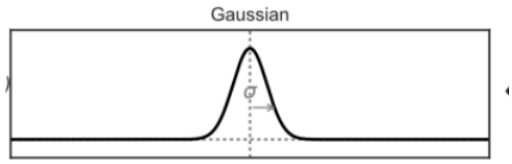
Frequency precision

Once you've found the peak period in your data, how do you report it?

Ideally by something like: **9.5 ± 0.5 days**

How do you get the error bar on the period?

The precision with which a peak's frequency can be determined is directly related to the width of the peak (half-width @ half-max):



$f_{\mathrm{peak}} \pm \sigma f$

While this is symmetric in frequency, it is not necessarily symmetric in period.

Frequency precision

This is a very simplistic estimate. You could do a better job by using simulations.

How?