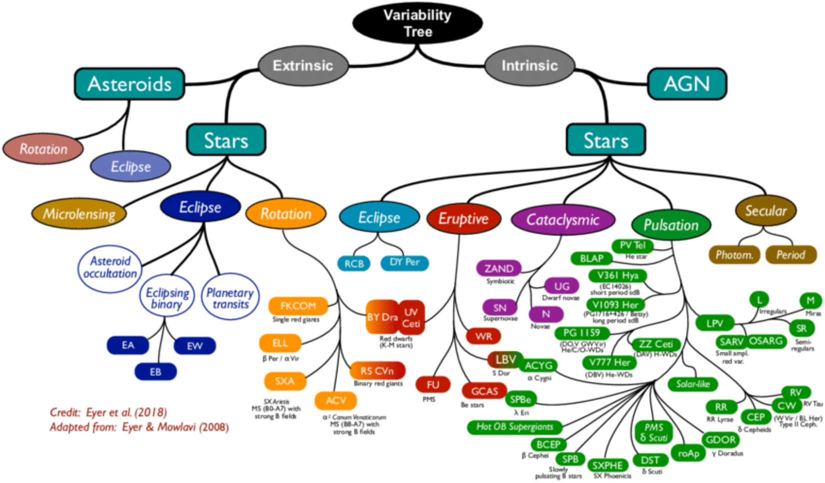
ASTR 3890 - Selected Topics: Data Science for Large Astronomical Surveys (Spring 2022)

Time Series Analysis: I

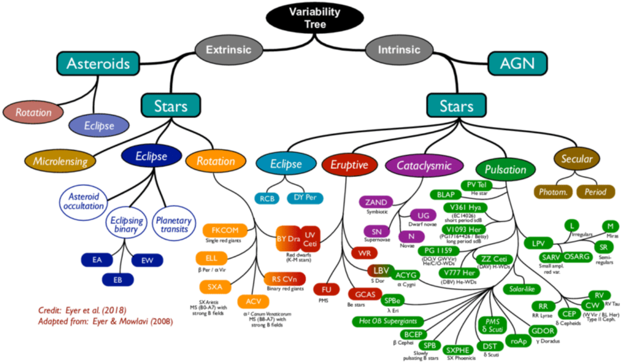
Dr. Nina Hernitschek
March 21, 2022

Motivation



Credit: Eyer et al. (2018)
Adapted from: Eyer & Mowlavi (2008)

Motivation



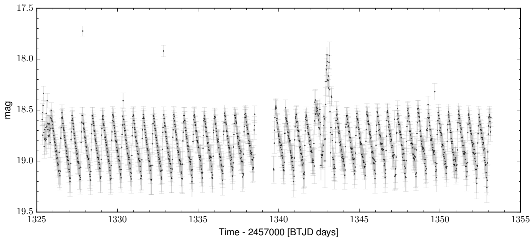
Credit: Eyer et al. (2018)
Adapted from: Eyer & Mowlavi (2008)

many astronomical sources vary - describe and classify astronomical sources by their variability

Astronomical Light Curves

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

RRab (RR Lyrae type ab):



Astronomical Light Curves

Motivation

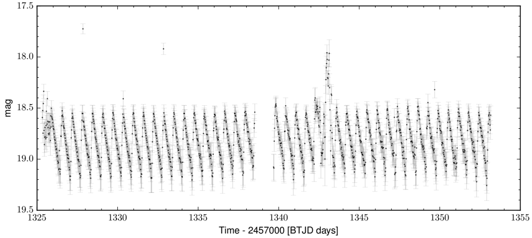
Intro: Time Series Analysis

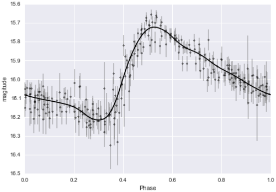
Parameter Estimation and Model Selection

Detecting Periodic Signals

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

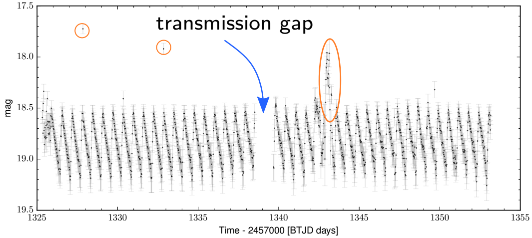
RRab (RR Lyrae type ab):

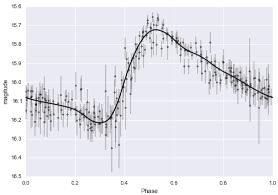




example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

RRab (RR Lyrae type ab):

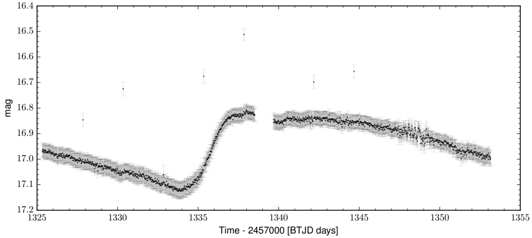




Astronomical Light Curves

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

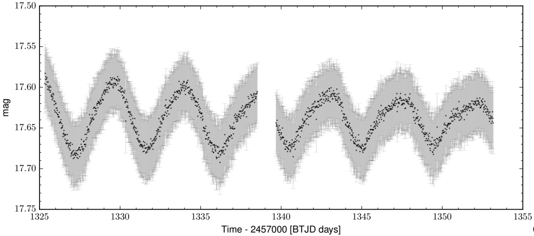
Cepheid

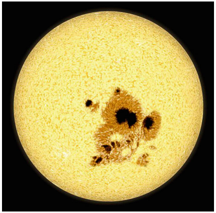


Astronomical Light Curves

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

rotational variable star:



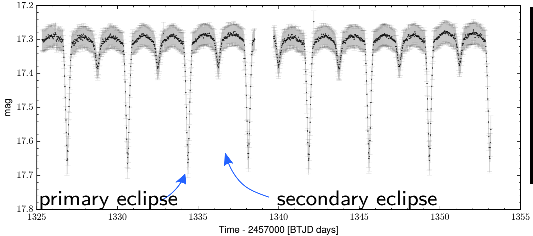


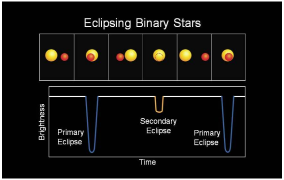
credit: Observer's Guide to Variable Stars, M. Griffiths

Astronomical Light Curves

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

eclipsing binary star:





credit: Wikimedia, NASA

When the smaller star partially blocks the larger star, a primary eclipse occurs, and a secondary eclipse occurs when the smaller star is occulted.

Other Astronomical Time Series Data

Motivation

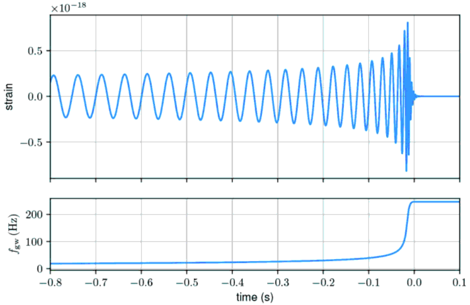
Intro: Time Series Analysis

Parameter Estimation and Model Selection

Detecting Periodic Signals

Light curves show variability from electromagnetic sources.
In addition: gravitational-wave variability as time series data

Gravitational Wave Signal:



Typical GW signal of a compact binary coalescence. The GW strain (above) and the GW frequency (below) are plotted as function of the time before merging. credit: Vallisneri et al. (2015).

Time Series Data

A time series is a sequence of random variables $\{\mathbf{X}_t\}_{t=1,2,\cdots}$.

Thus, a time series is a **series of data points ordered in time**. The time of observations provides a source of additional information to be analyzed.

Time Series Data

A time series is a sequence of random variables $\{\mathbf{X}_t\}_{t=1,2,\cdots}$.

Thus, a time series is a **series of data points ordered in time**. The time of observations provides a source of additional information to be analyzed.

Since there may be an infinite number of random variables, we consider **multivariate distributions of random vectors**, that is, of finite subsets of the sequence $\{\mathbf{X}_t\}_{t=1,2,\cdots}$.

Time Series Data

A time series is a sequence of random variables $\{\mathbf{X}_t\}_{t=1,2,\cdots}$.

Thus, a time series is a **series of data points ordered in time**. The time of observations provides a source of additional information to be analyzed.

Since there may be an infinite number of random variables, we consider **multivariate distributions of random vectors**, that is, of finite subsets of the sequence $\{\mathbf{X}_t\}_{t=1,2,\cdots}$.

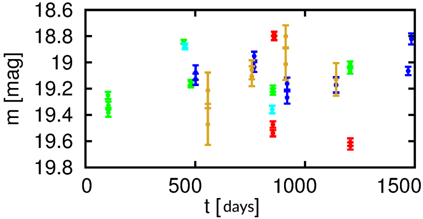
A **time series model** for the observed data $\{x_t\}$ is defined to be a specification of all of the joint distributions of the random vectors $\mathbf{X} = (X_1, \cdots, X_n)^{\mathrm{T}}$, $n = 1, 2, \cdots$ of which $\{x_t\}$ are possible realizations, that is, at all of these probabilities

$$P(X_1 \leq x_1, \cdots, X_n \leq x_n),\ -\infty < x_1, \cdots, x_n < \infty, \\ n = 1, 2, \cdots .$$

Time Series Data

Astronomical time series are typically assumed to be generated at irregularly spaced interval of time (**irregular time series**).

example: light curves from multi-band surveys



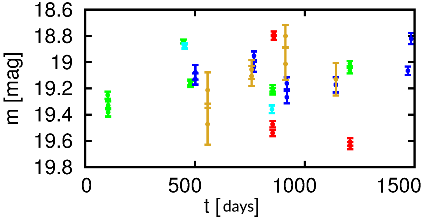
Time Series Data

A time series is a **series of data points ordered in time**. The time of observations provides a source of additional information to be analyzed.

Astronomical time series are typically assumed to be generated at irregularly spaced interval of time (**irregular time series**).

Time series can have one or more variables that change over time. If there is only one variable varying over time, we call it **univariate time series**. If there is more than one variable it is called **multivariate time series**.

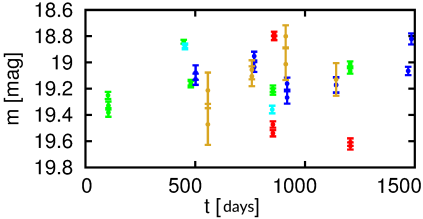
example: light curves from multi-band surveys



Characteristics of Astronomical Time Series Data

Astronomical time series data in general is:

- irregularly sampled
- multivariate
- not sampled to fully characterize the variability process
- not an independent random variable in their y values: often $y_{i+1} = f(y_i)$

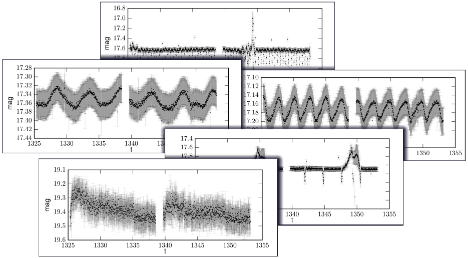


Goals of Time Series Analysis

Time series analysis extracts meaningful statistics and other characteristics of the dataset in order to understand it.

The main tasks of time series analysis are:

- **characterize** the temporal correlation between different values of y, including its significance
  example: classification of variable sources



- **forecast** (predict) future values of y
  example: transient detection, e.g. early supernovae detection

Goals of Time Series Analysis

When dealing with time series data, the first question we ask is ***Does the time series vary over some timescale?*** (if not, there is no point doing time series analysis)

Variability does not mean necessarily periodicity. Stochastic processes are variable over some timescale, but are distinctly aperiodic through the inherent randomness.

Goals of Time Series Analysis

When dealing with time series data, the first question we ask is ***Does the time series vary over some timescale?*** (if not, there is no point doing time series analysis)

Variability does not mean necessarily periodicity. Stochastic processes are variable over some timescale, but are distinctly aperiodic through the inherent randomness.

If we find that a source is variable (almost all astronomical sources are), then time-series analysis has two main goals:

1. Characterize the temporal correlation between different values of y (i.e., characterize the light curve), e.g. by learning the parameters for a model.
2. Predict future values of y.

Detecting Variability

For known and Gaussian uncertainties, we can compute χ^2 and the corresponding p values for variation in a signal.
For a sinusoidal variable signal $A \sin(\omega t)$, with homoescedastic measurement uncertainties, the data model would be

$$y(t) = A \sin(\omega t) + \epsilon$$

where $\epsilon \sim N(0, \sigma)$. The overall data variance is then $V = \sigma^2 + A^2/2$.

Detecting Variability

For known and Gaussian uncertainties, we can compute χ^2 and the corresponding p values for variation in a signal.
For a sinusoidal variable signal $A \sin(\omega t)$, with homoescedastic measurement uncertainties, the data model would be

$$y(t) = A \sin(\omega t) + \epsilon$$

where $\epsilon \sim N(0, \sigma)$. The overall data variance is then $V = \sigma^2 + A^2/2$.

If $A = 0$ (no variability, with $\overline{y} = 0$):

- $\chi^2_{\rm dof} = N^{-1} \sum_j (y_j/\sigma)^2 \sim V/\sigma^2$
- $\chi^2_{\rm dof}$ has expectation value of 1 and std dev of $\sqrt{2/N}$

Detecting Variability

If $|A| > 0$ (variability):

- $\chi^2_{\rm dof}$ will be larger than 1.
- probability that $\chi^2_{\rm dof} > 1 + 3\sqrt{2/N}$ is about 1 in 1000 (i.e., $> 3\sigma$ above 1, where 3σ is 0.997).

Detecting Variability

If $|A| > 0$ (variability):

- $\chi^2_{\rm dof}$ will be larger than 1.
- probability that $\chi^2_{\rm dof} > 1 + 3\sqrt{2/N}$ is about 1 in 1000 (i.e., $> 3\sigma$ above 1, where 3σ is 0.997).

If this false-positive rate (1 in a 1000) is acceptable (because even without variability 1 in 1000 will be above this threshold) then the minimum detectable amplitude is $A > 2.9\sigma/N^{1/4}$ (from $V/\sigma^2 = 1 + 3\sqrt{2/N}$, so that $A^2/2\sigma^2 = 3\sqrt{2/N}$).

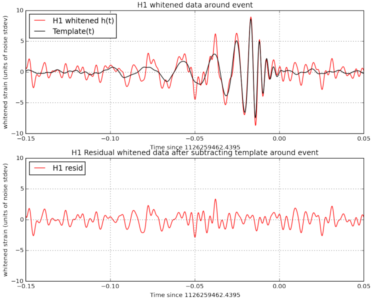
Depending on how big your sample is, you may want to choose a higher threshold. E.g., for 1 million non-variable stars, this criterion would identify 100 as variable.

1. For $N = 100$ data points (not 100 objects), the minimum detectable amplitude is $A_{\rm min} = 0.92\sigma$
2. For $N = 1000$, $A_{\rm min} = 0.52\sigma$

Detecting Variability

We do this under the assumption of the null hypothesis of no variability. If instead we have a model, we can perform a **matched filter analysis** by correlating a known template with an unknown signal to detect the presence of the template in the unknown signal

example: gravitational wave event GW150914



credit: https://www.gw-openscience.org/tutorials/

Parameter Estimation and Model Selection

We can fit a model to N data points (t_i, y_i):

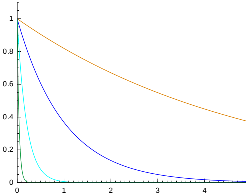
$$y_i(t_i) = \sum_{m=1}^{M} \beta_m T_m(t_i|\theta_m) + \epsilon_i,$$

with (not necessarily periodic) basis functions T_m, t_i with arbitraray sampling, and model parameters θ_m.
Common deterministic models include a **sine wave**

$$T(t) = \sin(\omega t)$$

and a **decaying burst** (exponential decay)

$$T(t) = \exp(-\alpha t),$$



with parameters to be estimated from the data.

Parameter Estimation and Model Selection

Determining whether a variable model is favored over a non-variable model is the same as previously in frequentist and Bayesian model selection. In a Bayesian sense, we can use the tools we know like the AIC, BIC, or Bayesian odds ratio.

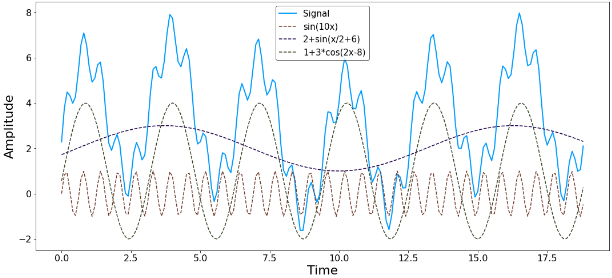
Parameter Estimation and Model Selection

Determining whether a variable model is favored over a non-variable model is the same as previously in frequentist and Bayesian model selection. In a Bayesian sense, we can use the tools we know like the AIC, BIC, or Bayesian odds ratio.

Once the model parameters, θ_m have been determined, we can apply supervised or unsupervised **classification methods** to gain further insight (lecture 10 - 13).

Fourier Analysis

Fourier analysis plays a **major role** in the analysis of time series data. In Fourier analysis, general **functions are approximated by integrals or sums of trigonometric functions**.



For periodic functions, such as periodic light curves in astronomy, often a relatively small number of terms (less than 10) suffices to reach an approximation precision level similar to the measurement precision.

Fourier Analysis

The **Fourier transform (FT)** $H(f)$ of function $h(t)$ is defined as

$$H(f) = \int_{-\infty}^{\infty} h(t) \exp(-i2\pi ft) dt$$

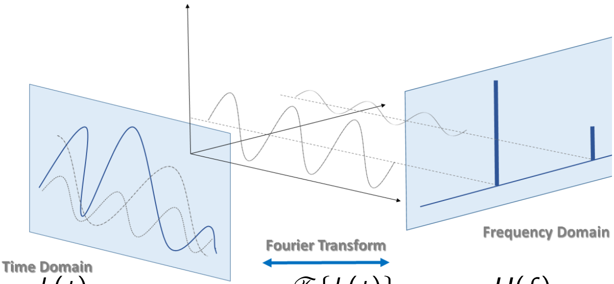
with inverse transformation

$$h(t) = \int_{-\infty}^{\infty} H(f) \exp(-i2\pi ft) df$$

where t is time and f is frequency (for time in seconds, the unit for frequency is hertz, or Hz).

Fourier Analysis

In other words, FT transforms a periodic function in **Time Domain** to a function in **Frequency Domain**:



$h(t)$ $\mathscr{F}\{h(t)\}$ $H(f)$

Fourier Analysis

For a real function $h(t)$, $H(f)$ is in general a **complex function**.

special case:
When $h(t)$ is an even function such that $h(-t) = h(t)$, $H(f)$ is real and even as well.

example:
The Fourier transform of a pdf of a zero-mean Gaussian $\mathcal{N}(0, \sigma)$ in time domain is a Gaussian $H(f) = \exp(-2\pi^2\sigma^2 f^2)$ in the frequency domain.

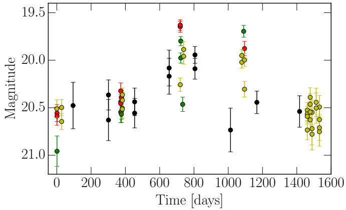
Detecting Periodic Signals

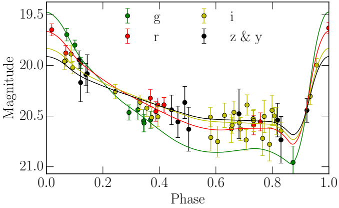
many objects/ systems have periodic signals: e.g., pulsars, RR-Lyrae, Cepheids, eclipsing binaries

For a periodic signal, if the period is known

- we can write $y(t+P) = y(t)$, where P is the period.
- we can create a **phased light curve** that plots the data as function of phase: $\phi = \frac{t}{P} - \mathrm{int}\left(\frac{t}{P}\right)$ with $\mathrm{int}(x)$ being the integer part of x.

phase folding

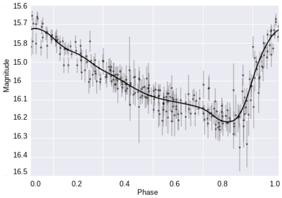




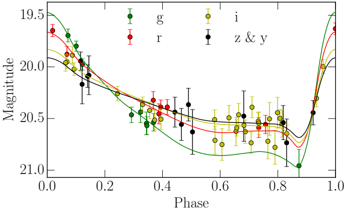
Detecting Periodic Signals

for well-sampled, high-cadence data: easy, standard methods can be applied

for sparse, low-cadence data: harder, specialized methods like template fitting necessary



vs.



measure the period and amplitude in the face of both noisy and incomplete data

Detecting Periodic Signals - An Approach

Let's take the case where the **data are drawn from a single sinusoidal signal**

$$y(t) = a\sin(\omega t) + b\cos(\omega t),$$

where $A = (a^2 + b^2)^{1/2}$ and $\phi = \tan^{-1}(b/a)$
and determine whether or not the data are indeed consistent with periodic variability and, if so, what is the period.

Detecting Periodic Signals - An Approach

Let's take the case where the **data are drawn from a single sinusoidal signal**

$$y(t) = a\sin(\omega t) + b\cos(\omega t),$$

where $A = (a^2 + b^2)^{1/2}$ and $\phi = \tan^{-1}(b/a)$
and determine whether or not the data are indeed consistent with periodic variability and, if so, what is the period.

Assuming constant uncertainties on the data, the **likelihood for this model** becomes

$$\begin{aligned} L &\equiv p(t, y|\omega, a, b, \sigma) \\ &= \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-[y_j - a\sin(\omega t_j) - b\cos(\omega t_j)]}{2\sigma^2}\right), \end{aligned}$$

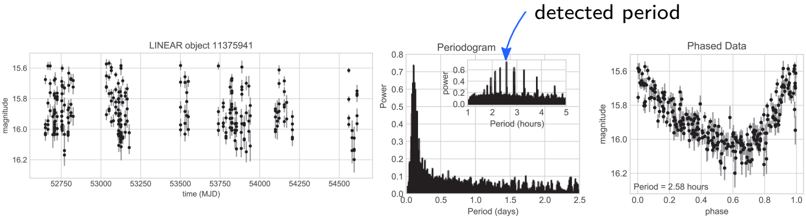
where y_i is the measurement (e.g., the brightness of a star) taken at time t_i.

Detecting Periodic Signals - An Approach

The **posterior** can be simplified to

$$p(\omega|\{t,y\},\sigma) \propto \sigma^{-N} \exp\left(\frac{-NQ}{2\sigma^2}\right) \propto \exp\left(\frac{P(\omega)}{\sigma^2}\right)$$

where $P(\omega)$ is the **periodogram**, which is a plot of the *power* in the time series at each possible period (as illustrated below):



left panel: observed light curve from LINEAR object ID 11375941
middle panel: periodogram computed from the light curve
right panel: light curve folded over the detected 2.58 hr period
credit: VanderPlas (2018)

Detecting Periodic Signals - The Periodogram

The periodogram is defined as

$$P(\omega) = \frac{1}{N}\left[\left(\sum_{j=1}^{N} y_j \sin(\omega t_j)\right)^2 + \left(\sum_{j=1}^{N} y_j \cos(\omega t_j)\right)^2\right]$$

The **best value** ω is given by

$$\chi^2(\omega) = \chi_0^2 \left[1 - \frac{2}{N\,V}\,P(\omega)\right],$$

where $P(\omega)$ is the periodogram, V the variance of the data y, and χ_0^2 is the χ^2 for the null-hypothesis model $y(t) = \text{const}$:

$$\chi_0^2 = \frac{1}{\sigma^2}\sum_{j=1}^{N} y_j^2 = \frac{N\,V}{\sigma^2}$$

Detecting Periodic Signals - The Periodogram

We can renormalize the periodogram, defining the **Lomb-Scargle periodogram** as

$$P_{\mathrm{LS}}(\omega) = \frac{2}{NV} P(\omega),$$

where $0 \leq P_{\mathrm{LS}}(\omega) \leq 1$.

With this renormalization, the ratio of $\chi^2(\omega)$ (for the periodic model) relative to χ_0^2 (for the pure noise model) is

$$\frac{\chi^2(\omega)}{\chi_0^2} = 1 - P_{LS}(\omega).$$

Detecting Periodic Signals - The Periodogram

How to determine if our source is variable or not:

- compute Lomb-Scargle periodogram $P_{\mathrm{LS}}(\omega)$
- model the odds ratio for our variability model vs. a no-variability model.

If our variability model is correct, then the **peak** of $P(\omega)$ (found by grid search) gives the best period ω.

⇨ The Lomb-Scargle periodogram (Lomb 1976; Scargle 1982) is the **standard method** to search for periodicity in unevenly-sampled time-series data.

Break & Questions

afterwards we continue with `lecture_8.ipynb` from the `github` repository