TIME SERIES

GOAL

the goal of time series data is to analyze trends in values that occured at certain time intervals. This usually means that you want to predict future values or find the period of the data you have.

DEFINITIONS

Stationary: the time series is invariant to shift's in time, it is oscillating around some constant mean

Auto correlation: a measure of how related a time series is to itself at different lags

White noise: a time series that is random w/ constant mean & variance, it has no autocorrelation

Lag: how many samples of a difference from the one you're looking at. If measurements are taken daily a lag of 2 would be the measurement from 2 days ago.

Even: taken at evenly spaced times.

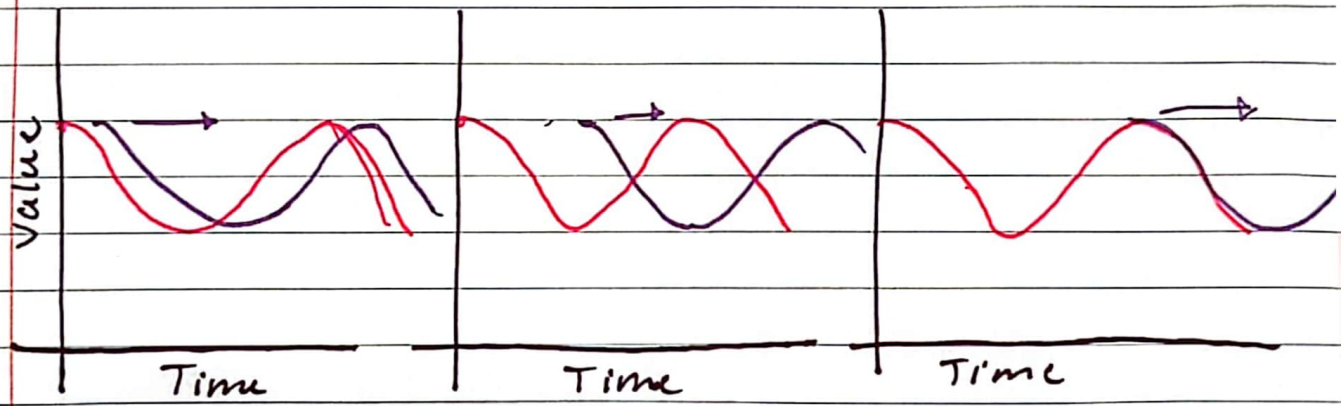
Uneven: taken at varying time intervals → many of the methods do not work for uneven data.

Methods

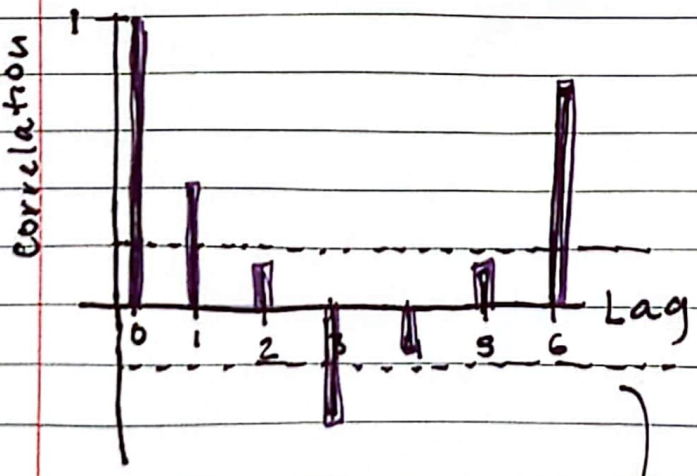
Autocorrelation Function (ACF)

used for evenly spaced data to see if there is some sort of period or time correlation

The correlation at different lags is calculated



The purple moves over the actually data at lags that are equal to the time spacing. This results in a graph that looks like:



we see the correlation from various lags, from this plot it appear the data may have a period of around 6.

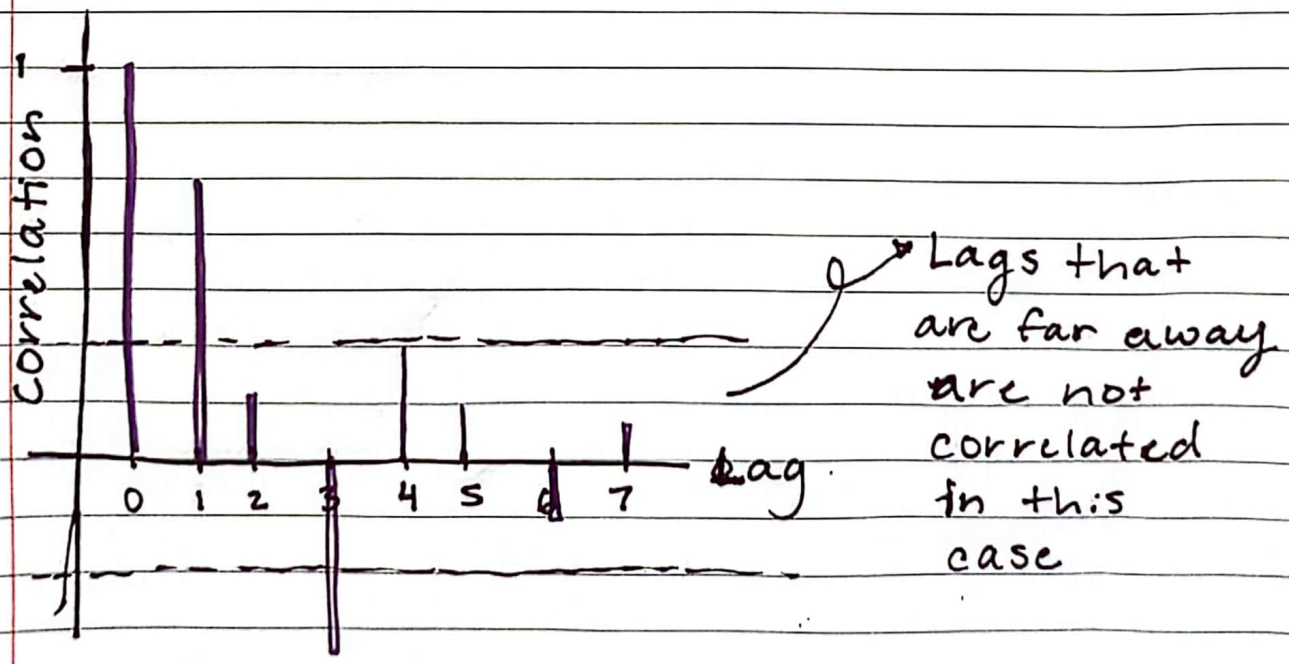
A lot of times there a significance lines

Partial Autocorrelation Function (PACF)

this is similar to ACF, but it gives the auto correlation at a value of k by removing the effects of shorter Ks. This can help to better identify the underlying structure → the direct relationship between lags.

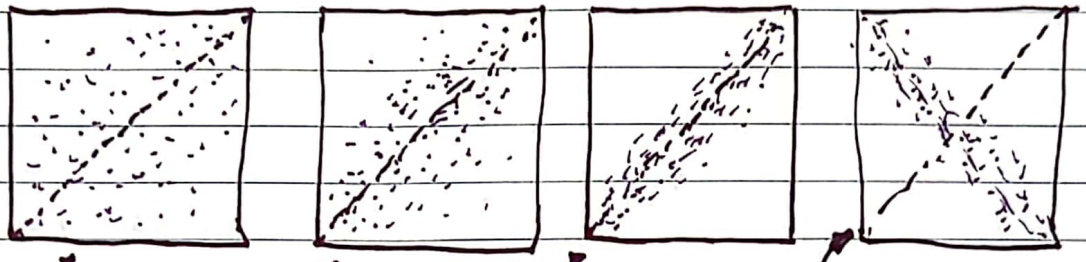
This is done by fitting a regressive model that accounts for each lag.

$$PACF(2) = \frac{ACF(2) - ACF(1)^2}{1 - ACF(1)^2}$$



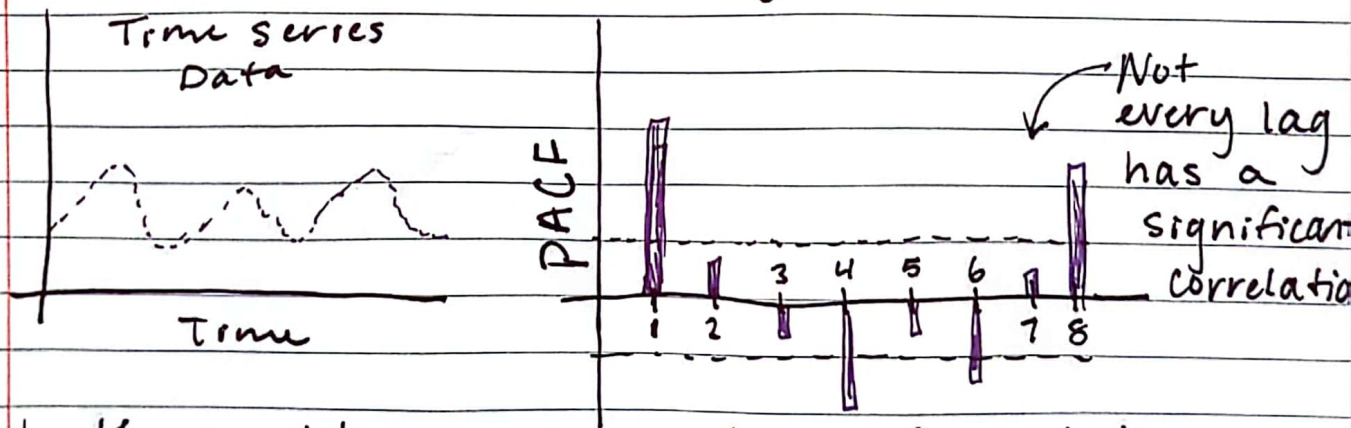
Lag K Scatter Plot

these plots plot the values of different lags → X_t vs X_{t+K}. This helps to visually determine patterns.



Autoregressive Modal (AR)

Tries to predict future values based on past values (still for evenly spaced)



In the modul we only want significant lags so it will look something like:

$$m_t = \beta_0 + \beta_1 m_{t-1} + \beta_4 m_{t-4} + \beta_6 m_{t-6} + \beta_8 m_{t-8}$$

Moving Average Model (MA)

Instead of basing your guess on the actual past values, MA bases it on the error. There's some value you're trying to predict, but at each time you're probably off by a little bit.

$$X_i = \varepsilon_i + \beta_1 \varepsilon_{i-1} + \ldots + \beta_q \varepsilon_{i-q}$$

These are the errors at different lags

So if you're really far off a one lag, you'll want to change your answer more the next time.

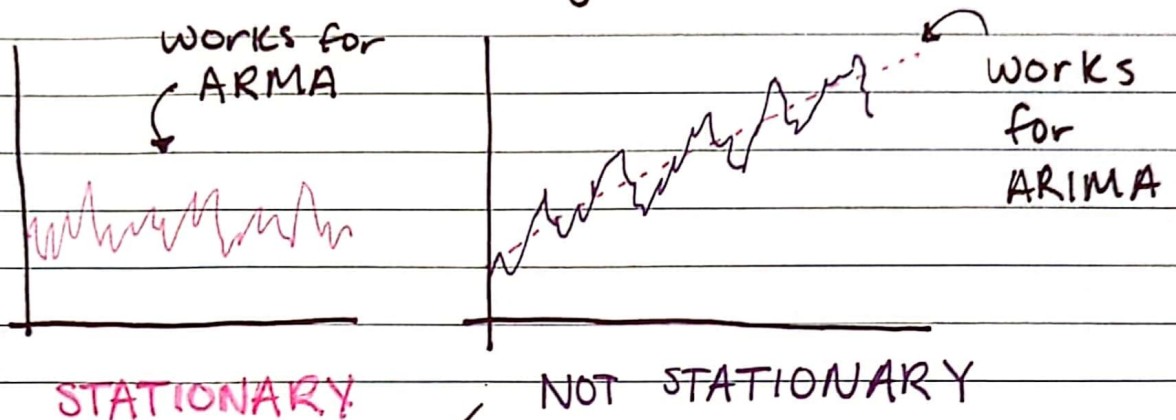
ARMA (Autoregressive - Moving Average)

This model combines AR & MA into one. It uses both the past values & the errors to make future predictions.

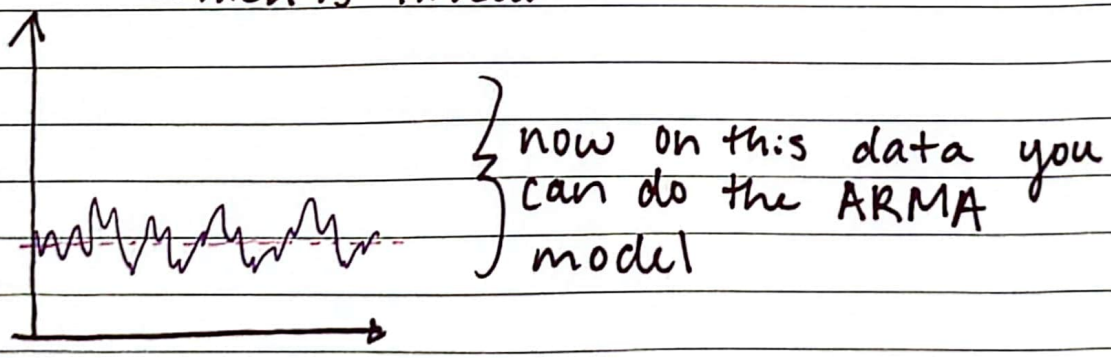
$$X_i = \underbrace{\alpha_1 X_{i-1} + \ldots + \alpha_p X_{i-p}}_{\text{Autoregressive}} + \underbrace{\varepsilon + \beta_1 \varepsilon_{i-1} + \ldots + \beta_q X_{i-q}}_{\text{Moving Average}}$$

ARIMA Model (Autoregressive integrated moving average)

Still used for predicting the future, with evenly spaced data, but this time the data does not have to be stationary.



This data is tranformed by take $X_{t+1} - X_t$ to get the "average" change which is linear



Three parameters

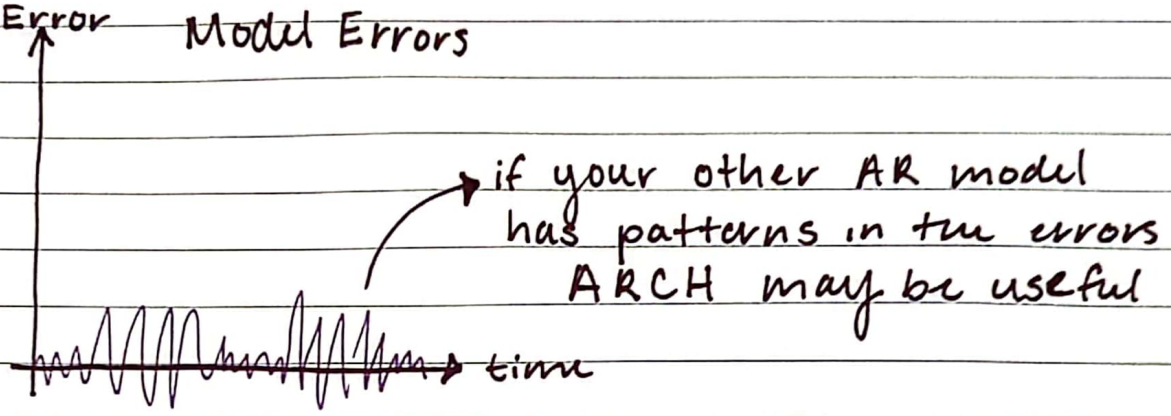
p: autoregressive order

d: differencing to account for the change in the mean

q: order for MA model

ARCH Model (autoregressive conditional heteroscedastic)

This model is used to better understand errors in the other AR models



For ARCH (1):

$$Var(\varepsilon_t) = \sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}$$

so your variance relates to the variance of previous day

$$\varepsilon_t = w_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2}$$

So, essentially this just models your error.

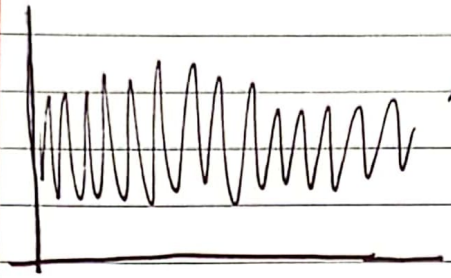
Ljung-Box Statistic

H_0: the time series in random or white noise
H_1: there is serial correlation

$$Q(m) = n(n+2) \sum_{j=1}^{m} \frac{r_j^2}{n-j}$$

↳ m's are the different lags

Power Spectral Density



→ for a stationary process

$$S_x(f) = \lim_{T\to\infty} \frac{E\left[|F_{x_T}(f)|^2\right]}{2T}$$

↑ mean squared power

$$\bar{X}^2 = \int_{-\infty}^{\infty} S_x(f)\, df$$

$$X_T(t) = \begin{cases} X(t) & |t| \le T \le \infty \\ 0 & |t| > T \end{cases}$$

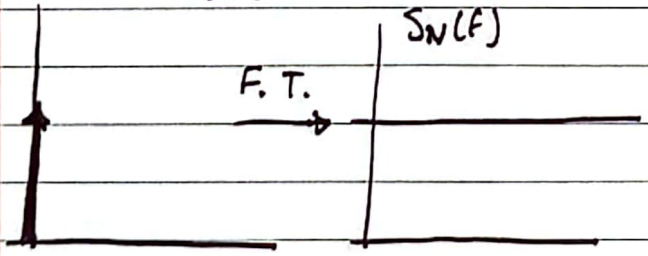
$$F_{x_T}(f) = \int_{-\infty}^{\infty} X_T(t)\, e^{-j2\pi ft}\, dt \qquad T < \infty$$

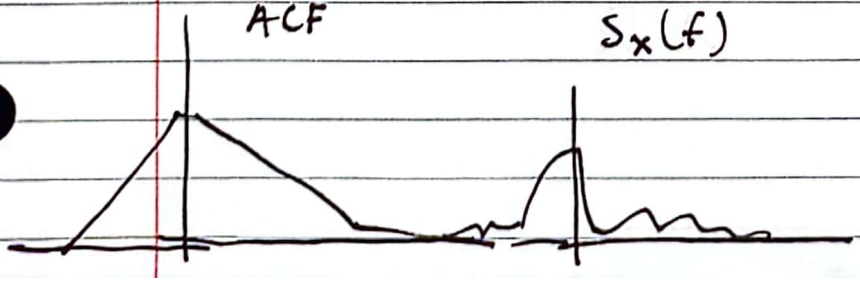
↳ take the fourier transform of

$$\int_{-\infty}^{\infty} x^2(t)\, dt = \int_{-\infty}^{\infty} |F_x(f)|^2\, df$$

$$S_x(f) = \underbrace{\text{F.T.}}_{\text{Fourier transform}}\left[\underbrace{R_x(\tau)}_{\text{auto correlation function}}\right]$$

ACF of Noise:





Periodogram

Used to identify dominant period of a ~~ob~~ time series, which have cyclical behavior.
This falls under spectral analysis. Even data.

Consider data w/ n points, represent the time series w/ sine & cosine

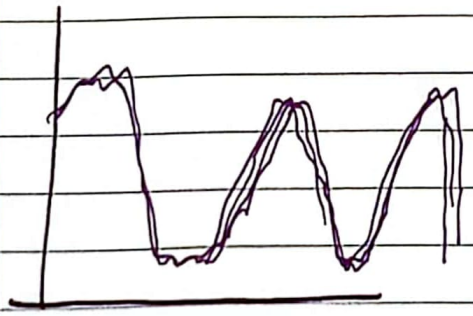
$$x_t = \sum_{j=1}^{n/2} \left[\beta_1 \left(\frac{j}{n}\right) \cos(2\pi \omega_j t) + \beta_2 \left(\frac{j}{n}\right) \sin(2\pi \omega_j t) \right.$$

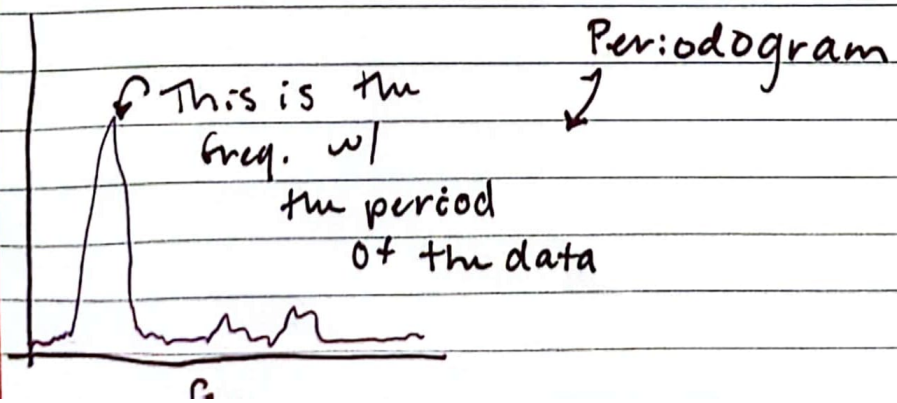
estimate β_1 & β_2

$$P\left(\frac{j}{n}\right) = \hat{\beta}_1^2 \left(\frac{j}{n}\right) + \hat{\beta}_2^2 \left(\frac{j}{n}\right)$$

↳ periodogram value at a specific frequency
↳ large P's indicate significant periods

So the data may look like:



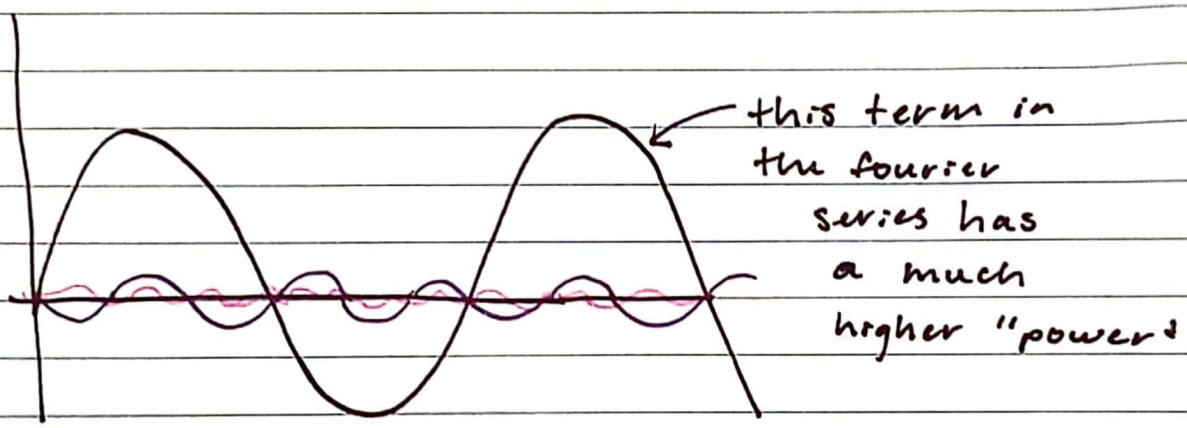


Lomb-Scargle Periodogram

Useful for finding the period of unevenly spaced time series data.

$$f(x) = \frac{1}{2} a_0 + \sum a_n \cos(nx) + \sum b_n \sin(nx)$$

the different "n's" have different power's



So you can keep track of the a_n & b_n for each freq. to see what freq. are "strong".

Now you want to take signals/data points at different times then recover the period

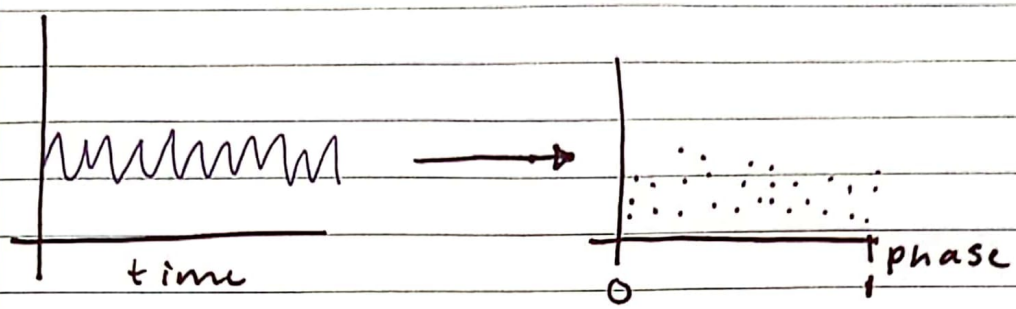
Could try fitting different curves to reconstruct signal.

LSP is very similar to a normal periodogram but it is modified to deal w/ uneven data.

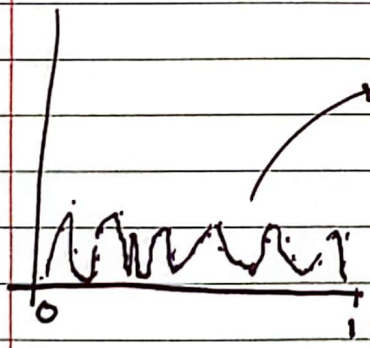
$$P_{LS}(\nu) = \frac{1}{2\sigma^2}\left[\frac{\left[\sum_{i=1}^{n} X_i \cos(2\pi\nu t_i)\right]^2}{\left[\sum \cos^2(2\pi\nu(t_i - \tau(\nu))\right]} + \frac{\left[\sum X_i \sin(2\pi\nu t_i)\right]^2}{\left[\sum \sin^2 2\pi\nu(t_i - \tau(\nu))\right]}\right]$$

Minimum String Length

Another way to find the period of the data, by using the folded modulo.



Each of the time values are converted into phase values. you have to cycle through different potential periods. So whatever period you're trying find the phase.



→ Then you want to find the distance needed to connect all the phase points
↳ the period w/ the minimum distance to connect the points is the estimated period.