lecture 4 2020

- Covariance and Correlation
- Tests for correlation
- Revised notes on PDF and CDF
- Introduction to harmonic analysis

Recap Variance

Variance is a measure of the spread of a dataset

$$\sigma^2 = \frac{1}{M} \sum_{i=1}^{M} (x_i - \mu)^2$$

Aside:
$\bar{x}$: sample mean
μ: population mean

Most of the time you don't know the population mean, so you have to work with the sample mean.

Covariance & correlation

Covariance is a measure of the strength of correlation between 2 (or more) sets of variables:

$$\mathrm{cov}(X,Y) = \sum_i^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$\mathrm{cov}(X,Y) = \langle XY \rangle - \langle X \rangle \langle Y \rangle$$

For uncorrelated (i.e. independent) variates:

$$\mathrm{cov}(X,Y) = 0$$

Covariance & correlation

Sign of the covariance is determined by the joint signs of the residuals of *x* and *y*.

$$\mathrm{cov}(X,Y) = \sum_i^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$$

$x - \overline{x}$	$y - \overline{y}$	
$+$	$+$	$+$
$-$	$-$	$+$
$+$	$-$	$-$
$-$	$+$	$-$

What does the sign of the covariance mean?

Covariance & correlation

The covariance of two variables is simply a number. The covariance between more than two variables is generally presented as a (symmetric) covariance matrix.

In order to calculate the correlation between two data sequences it is necessary that each possess the same number of terms.

The correlation also assumes that the sequences are ordered in such a way that each observation in one sequence is associated with the corresponding observation in the other sequence: either they were made at the same time or the same place.

A generalisation of the correlation considers shifting the one sequence with respect to the other before generating the correlation - we'll come back to this later.

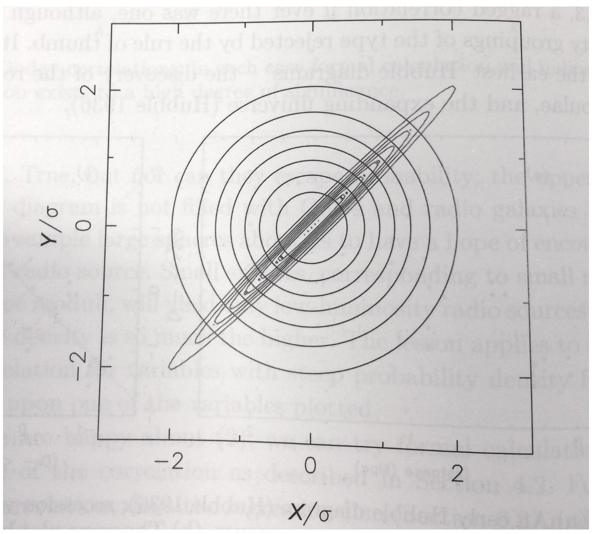
Correlation

Correlation coefficient:

$$\rho = \frac{\mathrm{cov}[x, y]}{\sigma_x \sigma_y}$$

can be estimated by *r*

(Pearson product-moment correlation coefficient)



$\rho = 0.01$ **near circular**

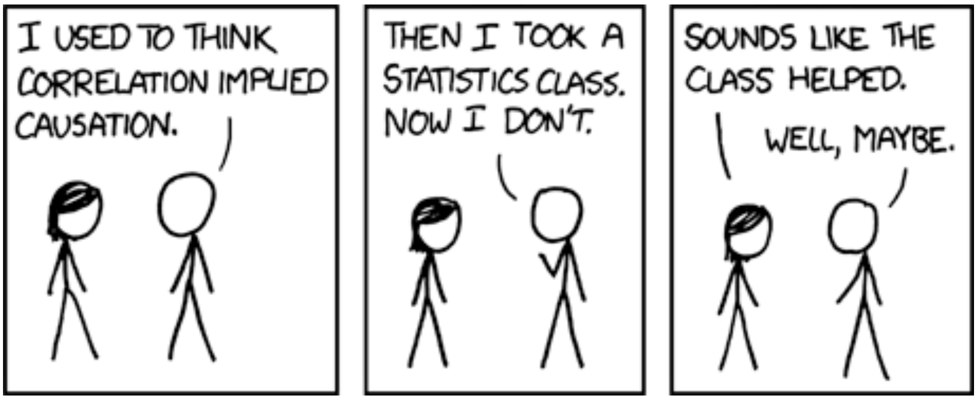
$\rho = 0.99$ **highly elliptical**

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 (Y_i - \bar{Y})^2}}$$

ρ↑

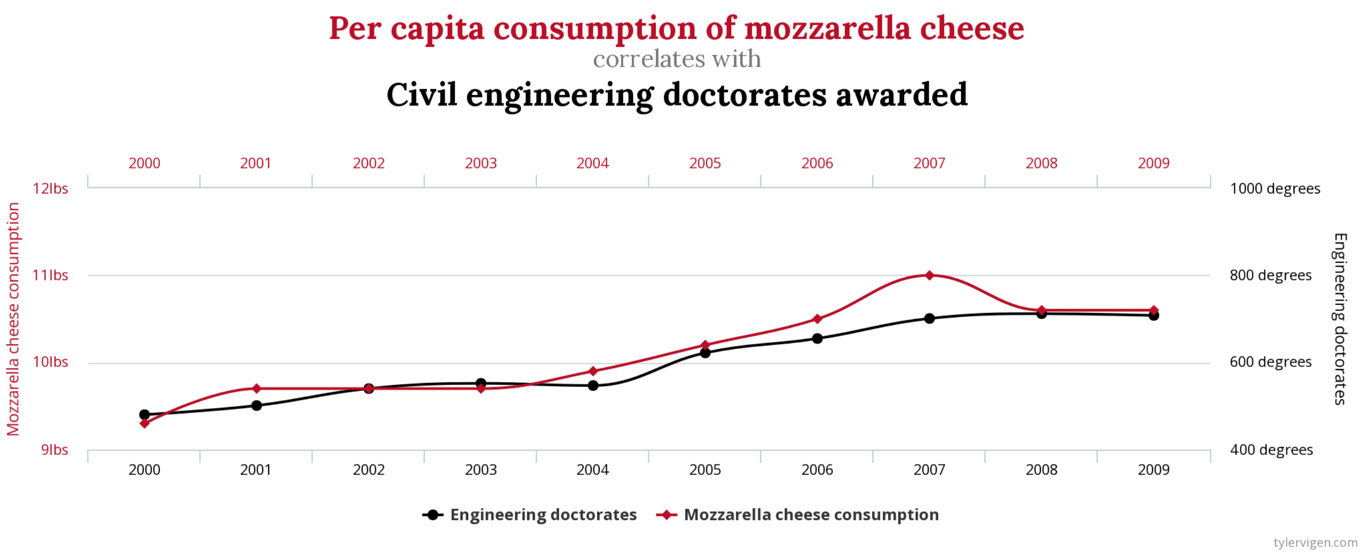
Correlation & causality

http://xkcd.com



A strong correlation between two variables does not imply causality.

Correlation & causality



$r = 0.9586$

http://www.tylervigen.com/spurious-correlations

Autocovariance & autocorrelation

Autocovariance is analogous to covariance. However, rather than calculating the covariance between x_i and y_i, the autocovariance is the covariance of x_i with another instance of itself, x_{i+k}, that has been shifted k samples relative to x_i.

The autocovariance at lag k is defined as

$$\gamma_k = \mathrm{cov}(x_i, x_{i+k})$$

Autocorrelation

The corresponding autocorrelation function at lag k is:

$$\rho_k = \frac{\gamma_k}{\gamma_0}$$

The autocorrelation function (ACF) is also known as the serial correlation. By symmetry

$$\rho_k = \rho_{-k}$$

The sample ACF is

$$r_k = \frac{1}{N} \sum_{i=1}^{N-k} (x_i - \overline{x})(x_{(i+k)} - \overline{x})$$

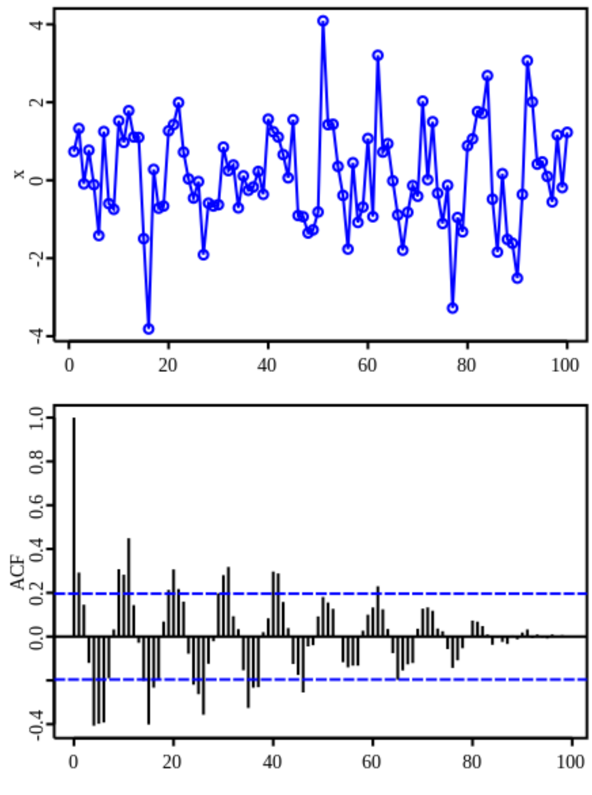
Autocovariance & autocorrelation

A correlogram is a plot of ρ_k versus k.

The analysis of a correlogram may reveal some characteristics of a time series:

- **random** if $r_k \simeq 0$ for all $k \neq 0$;
- **short-term correlation** if r_1 is relatively large but the subsequent coefficients tend rapidly to zero;
- **alternating** if values in x_i alternate on either side of the mean then the correlogram alternates around zero
- **trend** if the x_i include a trend then the correlogram will only reach zero after a relatively large number of terms;
- **seasonal cycle** if x_i includes a seasonal cycle then the same cycle will be reflected in the corrrelogram.

Autocovariance & autocorrelation



Probability Density Function (PDF)

- *p(x) dx* is the probability that an event has a value between *x* and *x + dx.*

Cumulative Distribution Function (CDF)

- The *cumulative distribution function* (CDF) at *x* is given by the integral of the *probability density function* (PDF) up to *x*.

PDF and CDF

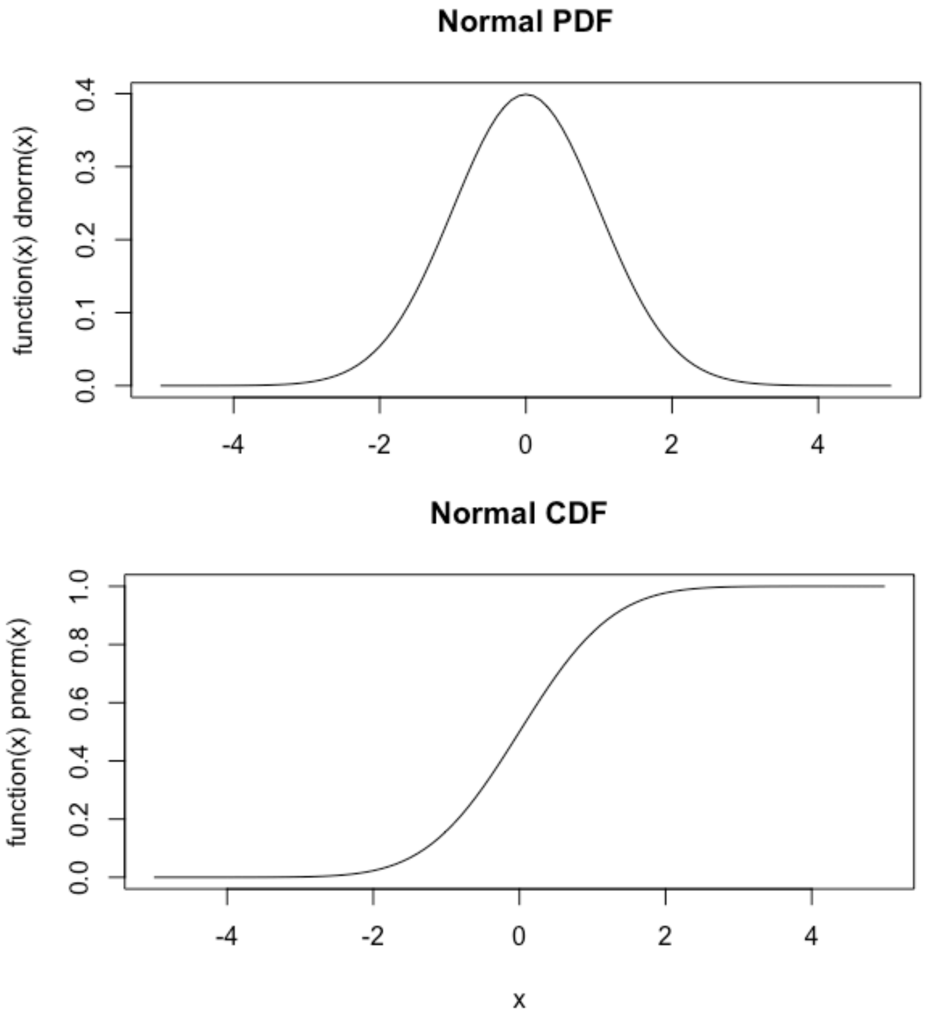
Example: Uniform Distribution

Suppose that values of x are distributed uniformly between 1 and 5. The PDF is then:

$$p(x) = \begin{cases} 0 & x < 1 \\ 1/4 & 1 \leq x \leq 4 \\ 0 & x > 4 \end{cases}$$

The corresponding CDF is then:

$$P(x) = \begin{cases} 0 & x < 1 \\ x/4 & 1 \leq x \leq 4 \\ 1 & x > 4 \end{cases}$$



PDF and CDF

- The PDF is the standard bell shaped curve which peaks at x = 0 and tapers to zero as $x \to \pm\infty$.
- The CDF is a curve which starts at zero for $x \to -\infty$ and flips to unity as $x \to \infty$.
- For any particular X, the value of the CDF corresponds to the area under the PDF between $x = -\infty$ and $x = X$ (the shaded region under the PDF).
- As $X \to -\infty$ one has $P(X) \to 0$ and conversely as $X \to \infty$ one has $P(X) \to 1$.

- **NB**: the CDF is found by **integrating** the PDF and the PDF is the derivative of the CDF
-

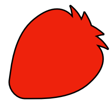
Fourier series and harmonic analysis

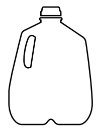
Why harmonic analysis?

- Removal of seasonal trends from data
- Orbits of stars
- Pulsations in stellar atmospheres
- Rotation of neutron stars
- Correlating data streams from radio telescopes

The Fourier Transform of a smoothie gives the recipe.

1 x 

8 x 

1/8 x 

Metaphor: Ian Heywood

Joseph Fourier

Born 1768, Auxerre, France

Orphaned at age 9. Educated by Benedictine Order

Scientific advisor on Napoleon’s Egyptian expedition 1798.

Part of his work on theory of heat laid foundation for Fourier theory

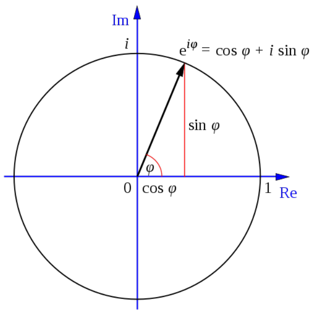
Suggested the greenhouse effect (Earth warmer than it should be)

Fourier theory

Any signal may be represented as the sum of sines and cosines:

$$f(t) = \int_{-\infty}^{+\infty} F(\omega) e^{i\omega t} d\omega$$

where the function *F* representing phased amplitudes of the sinusoidal components of *f* is known as the Fourier transform (FT).



Remember:

$$e^{i\omega t} = \cos \omega t + i \sin \omega t$$

A Fourier series

Lucas Barbosa - Wikipedia

Continuous Fourier transform

Fourier transform of a continuous signal *g(t)*

$$\hat{g}(f) \equiv \int_{-\infty}^{+\infty} g(t) e^{-2\pi i f t} dt$$

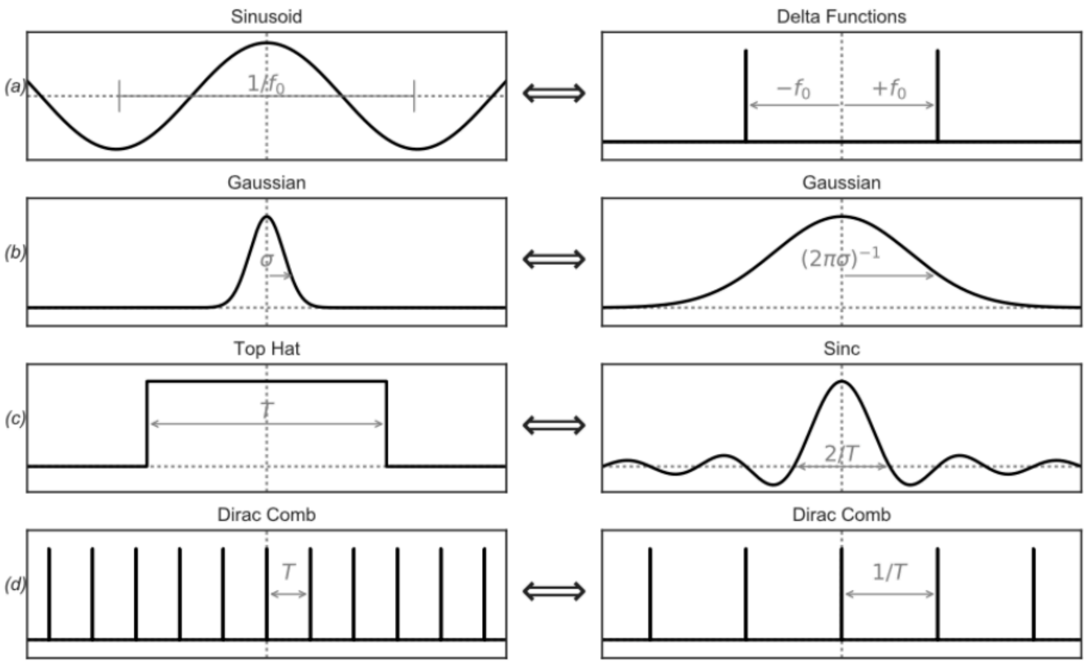
Inverse relationship

$$g(t) \equiv \int_{-\infty}^{+\infty} \hat{g}(f) e^{2\pi i f t} df$$

For convenience: define the Fourier transform operator $\mathcal{F}$:

$$\begin{aligned} \mathcal{F}\{g\} &= \hat{g} \\ \mathcal{F}^{-1}\{\hat{g}\} &= g \end{aligned}$$

Useful Fourier Pairs



Time domain **Frequency domain**

Convolution theorem

FT: convert convolutions into point-wise products

Convolution of two functions, *f(t)* and *g(t)* is defined:

$$h(t) = f \otimes g = \int_{-\infty}^{+\infty} f(\tau) g(t-\tau) d\tau$$

The FT of the convolution, *h(t)* is given by the product of the FTs of *f(t)* and *g(t)*:

$$\hat{h}(\omega) = \hat{f}(\omega) \cdot \hat{g}(\omega)$$

Convolution

1. Express each function in terms of a dummy variable τ.
2. Reflect one of the functions: $g(\tau) \rightarrow g(-\tau)$.
3. Add a time-offset, *t*, which allows $g(t-\tau)$ to slide along the τ-axis.
4. Start *t* at −∞ and slide it all the way to +∞. Wherever the two functions intersect, find the integral of their product. In other words, compute a sliding, weighted-sum of function $f(\tau)$, where the weighting function is $g(-\tau)$.

The resulting waveform (not shown here) is the convolution of functions *f* and *g*.

If *f(t)* is a unit impulse, the result of this process is simply *g(t)*. Formally:

$$\int_{-\infty}^{\infty} \delta(\tau) g(t-\tau)\, d\tau = g(t)$$

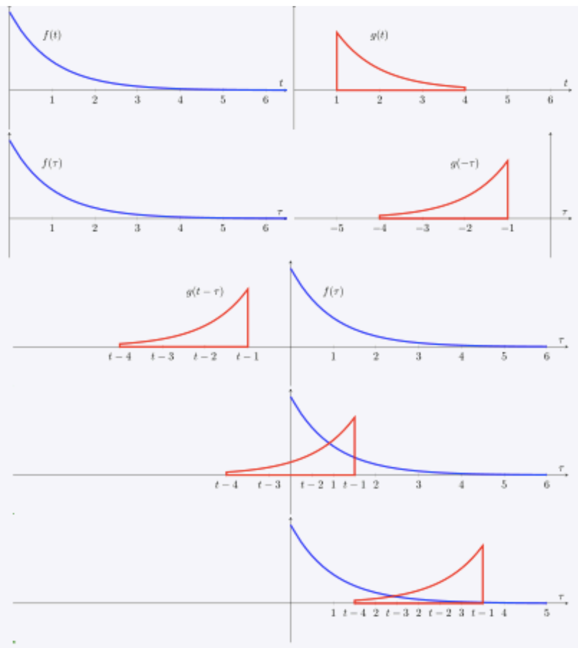


Image convolution examples

A convolution is very useful for signal processing in general. There is a lot of complex mathematical theory available for convolutions. For digital image processing, you don't have to understand all of that. You can use a simple matrix as an image convolution kernel and do some interesting things!

Series: Convolutions:

1. Basics of convolutions
2. **Image convolution examples**

Simple box blur

Here's a first and simplest. This convolution kernel has an averaging effect. So you end up with a slight blur. The image convolution kernel is:

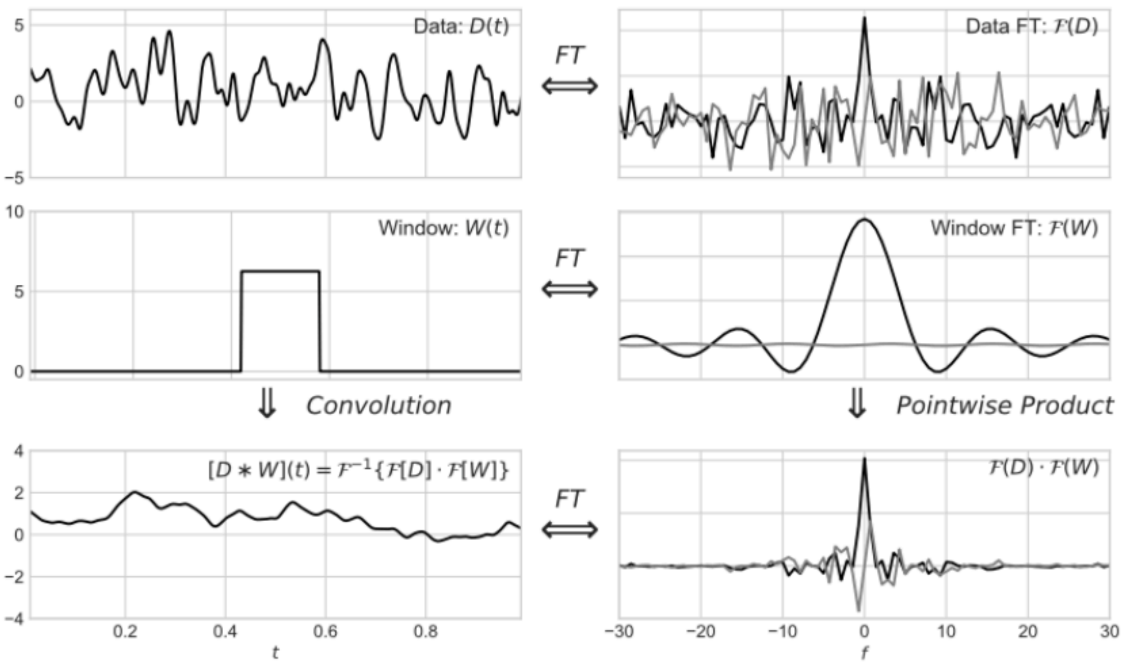
1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

Note that the sum of all elements of this matrix is 1.0. This is important. If the sum is not exactly one, the resultant image will be brighter or darker.

Here's a blur that I got on an image:

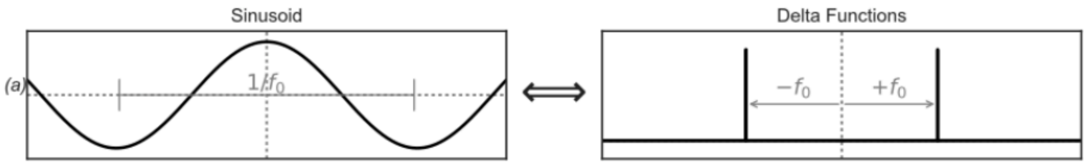
A simple blur done with convolutions

Convolution example



Properties of Fourier transforms

- The FT of a sine wave is a delta function in the frequency domain (searching for periodicities):



- The FT of $f \otimes g$, the cross-correlation or convolution of functions f and g, is $F \times G$. Important for spectroscopy, detection, image smoothing, etc.

Properties of Fourier transforms

- Shift theorem: transform of $f(t + \tau)$ is just the transform of f, times a simple exponential $e^{-i\omega t}$ - useful for redshifts

- The FT of a Gaussian is another Gaussian

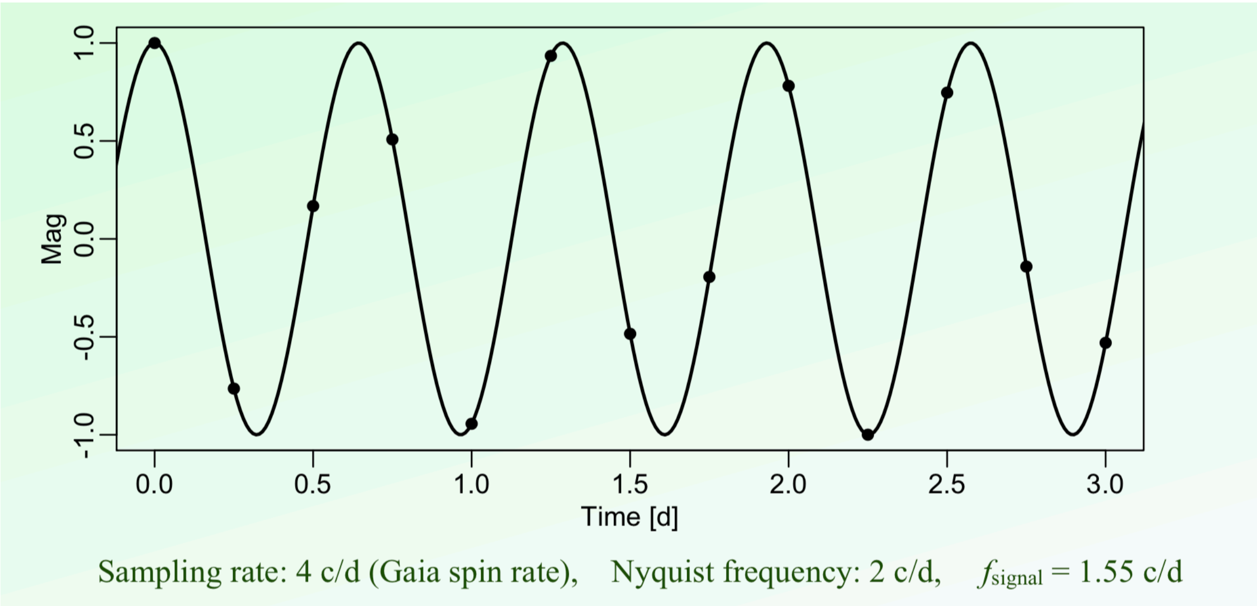
- Power spectrum $|F(\omega)|^2$ and the autocorrelation function $\int f(\tau) f(t+\tau) d\tau$ are Fourier pairs.

Remember: $$r_k = \frac{1}{N} \sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x}))$$

Discrete Fourier transforms

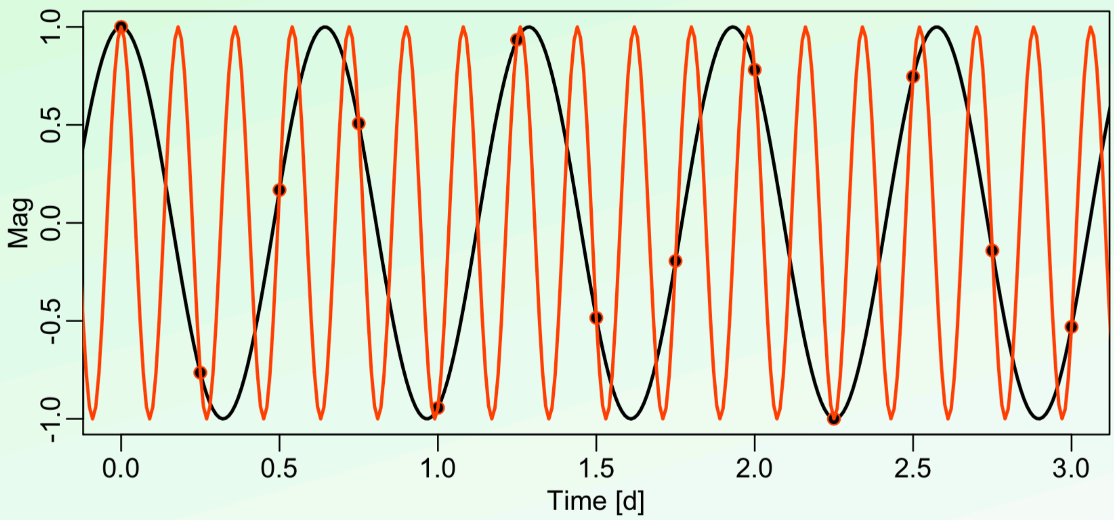
Take board notes from here.

Nyquist frequency & aliasing



Suveges (LSST data science)

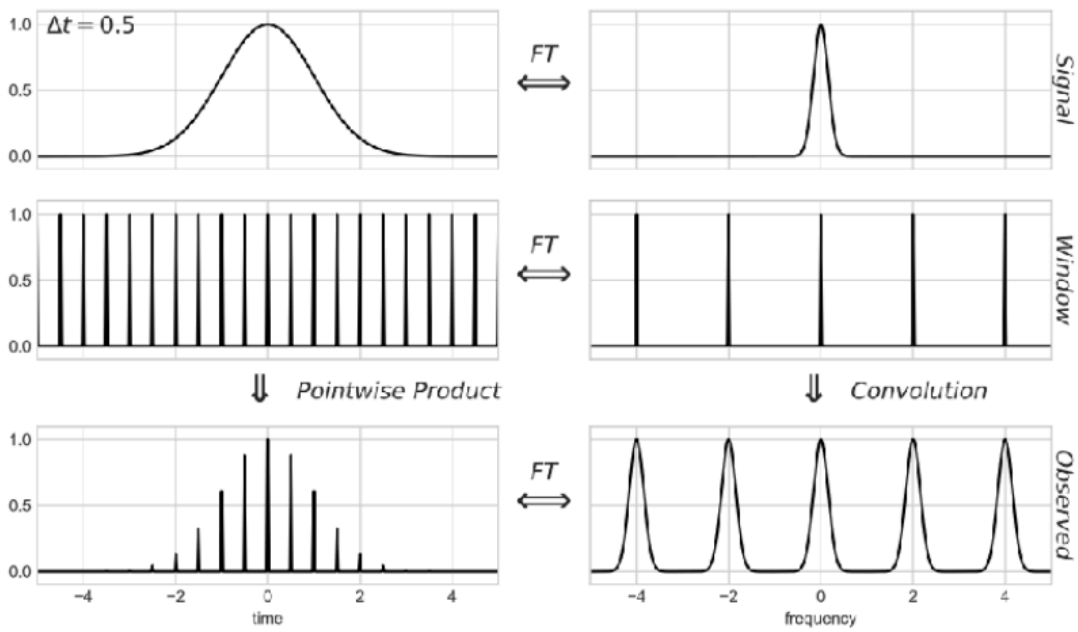
Nyquist frequency & aliasing



Sampling rate: 4 c/d (Gaia spin rate), Nyquist frequency: 2 c/d, $f_{\mathrm{signal}} = 1.55$ c/d
$f_{\mathrm{highfreq}} = 5.55$ c/d

Suveges (LSST data science)

DFT - sampling



Fast Fourier Transform

- Algorithm that does the transform of N points in a time proportion to $N \log N$, than than the N^2 timing of a brute force implementation. Cooley & Tukey 1965
- Probably one of the most well used transforms on the planet.

Exercise

Use direct numerical integration to do a numerical FT of a sine wave.

Compare the timing with an off-the-shelf FFT routine.

How many oscillations can you fit into your region of integration before the FFT accelerates away from the direct method?