Review of degrees of freedom / time indep.

Sample size impacts the variance of sampling distribution means. Persistence impacts the variance of the sampling distribution since it leads to an overestimate of the sample size

Effective sample size (t-stat)

$$N^* = N\left(\frac{1-\rho_1}{1+\rho_1}\right)$$

ρ_1 = lag-1 autocorrelation
N = orig. # samples
N^* = effective samples.

Assumes a first order Markov process

For white noise $N^* = N$, then as r increases N^* decreases.

Leith (1973) proposed:

$$N^* = N\left(\frac{\Delta t}{2 \cdot T}\right)$$

T = e-folding time scale of autocorrelation function.

Factor of 2 included because any given point in a red noise time series can be predicted by points before and after that point.

As before, as the the redness of time series goes up, N^* decreases.

For variance and covariance, Bretherton et al. (1999) suggest

$$N^* = N\left(\frac{1-\rho_1^2}{1+\rho_1^2}\right)$$

Recall we encountered something along these lines when estimating N^* for EOF significance testing.

Whatever you may use, should have an idea of how it affects your dof.

Auto correlation	0.15	0.3	0.5	0.7	0.9
Leith (1973) $\frac{N^*}{N}$	1	0.6	0.35	0.18	0.05
Bretherton (1999) $\frac{N^*}{N}$	1	0.83	0.6	0.34	0.1

The latter one can yield 2x the dof!

Harmonic analysis

Basic idea: Interpret a time or space series as a summation of contributions from harmonic functions - each with a unique temporal or spatial scale.

Recall the normal equations for the least squares best fit of 'y' to predictors 'x'

$$y = a_0 + a_1 x_1 + a_2 x_2 \cdots a_N x_N$$

where $\bar{y} = a_0 + a_1 \bar{x} + a_2 \bar{x}_2 \cdots a_N \bar{x}_N$

and:

$$\overline{x_1' y'} = a_1 \overline{x_1'^2} + a_2 \overline{x_1' x_2'} + a_3 \overline{x_1' x_3'} \cdots$$

$$\overline{x_2' y'} = a_1 \overline{x_1' x_2'} + a_2 \overline{x_2'^2} + a_3 \overline{x_2' x_3'} \cdots$$

In the case of harmonic analysis, the functions to be fit (i.e. the x's), or basis functions, are of the form:

$$\cos\left(2\pi K \frac{t}{T}\right) \quad \text{and} \quad \sin\left(2\pi K \frac{t}{T}\right)$$

$$K = 1, 2, 3, \cdots$$

Hence y can be written as:

$$y = a_0 + \sum_{K=1}^{N/2} A_K \cos\left(2\pi K \frac{t}{T}\right) + \sum_{K=1}^{N/2} B_K \sin\left(2\pi K \frac{t}{T}\right)$$

$K = 1, 2, 3 \ldots$

T = total time length of the record
N = # of grid points or time steps
A_K & B_K = regression coefficients for each predictor.

Notes

- Each predictor is a harmonic function with frequency K/T, and hence fits into the interval $[0, T]$ 'K' times

- if $K = 1$, the frequency is $1/T$ and hence predictor completes 2π radians in $[0, T]$

- if $K = N/2$, the frequency is the highest that can be resolved, the predictor completes π radians in one timestep
  → Nyquist frequency

- for a given K, fitting K cosine and K sine waves with variable amplitudes A & B into $[0, T]$. Equivalent to fitting K cosine waves (w/ no sine waves) with variable amp & phase

The amplitude of each predictor is found by solving the normal equations (like in regression)

Discrete Fourier transform

In the special case of evenly spaced data points (e.g. a constant time interval)

Consider domain $[0, T]$ where 0 and T coincide with data points $i = 1$ and $i = N+1$ where N is an even number.

1) The functions assume the form

$$\cos\left(2\pi K \frac{i\Delta t}{T}\right) \quad \text{and} \quad \sin\left(2\pi K \frac{i\Delta t}{T}\right)$$

where Δt is the spacing between grid points.

2) The average of each sine / cosine function on the interval $[0, T]$ is zero. Hence $a_0 = \overline{y}$

3) The harmonic functions are mutually orthogonal on $[0, T]$ because the correlation between sine and cosine functions is zero. Therefore, the off-diagonal elements in covariance matrix = 0.

Hence, the normal equations reduce to a set of equations of the form:

$$a_K = \frac{\overline{x'_K y'}}{\overline{x'^2_K}}$$

4) the functions (the K's) each have variance $\overline{x'^2_K} = \frac{1}{2}$ except for $A_{N/2}$ and $B_{N/2}$, of which the former takes on a value of $\pm$ unity at each data point, and ~~the~~ the latter a value of zero at each grid point.

Hence, the variance of the predictor with amplitude $A_{N/2}$, $B_{N/2}$ is one and zero, respectively.

5) From #3, the solutions for each predictor's amplitude:

$$a_K = \frac{\overline{x'_K y'}}{\overline{x'^2_K}} \quad \text{reduce to.}$$

$$A_K = \frac{2}{N} \sum_{i=1}^{N} y_i \cos\left(2\pi K \frac{i\Delta t}{T}\right)$$

$$B_K = \frac{2}{N} \sum_{i=1}^{N} y_i \sin\left(2\pi K \frac{i\Delta t}{T}\right)$$

$$K = 1, N/2 - 1$$

$$A_{N/2} = \frac{1}{N} \sum_{i=1}^{N} y_i \cos\left(\pi N \frac{\Delta t}{T}\right)$$

$$B_{N/2} = 0.$$

The function y can now be written as a sum of sines and cosines:

$$y(t) = \bar{y} + \sum_{K=1}^{N/2-1} A_K \cos\left(2\pi K \frac{t}{T}\right) + \sum_{K=1}^{N/2-1} B_K \sin\left(2\pi K \frac{t}{T}\right) + A_{N/2} \cos\left(\pi N \frac{t}{T}\right)$$

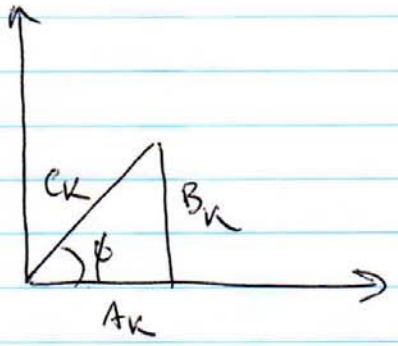
Can also write this in amplitude/phase form:

$$y(t) = \bar{y} + \sum_{K=1}^{N/2-1} C_K \cos\left(2\pi K \frac{(t-t_K)}{T}\right) + A_{N/2} \cos\left(\pi N \frac{t}{T}\right)$$

where $C_K^2 = A_K^2 + B_K^2$

$$t_K = \frac{T}{2\pi K} \tan^{-1}\left(\frac{B_K}{A_K}\right)$$

In geometric terms:



$$\phi = \tan^{-1}\left(\frac{B_K}{A_K}\right)$$

Fraction of variance explained for each K:

$$r^2(y, x_K) = \frac{\left(\overline{x_K' y'}\right)^2}{\overline{x_K'^2}\ \overline{y'^2}}$$

Using the expressions for A_K and B_K, and the fact that $\overline{x_K'^2} = \frac{1}{2}$

$r^2(y, x_K)$ is $\frac{A_K^2}{2\overline{y'^2}}$ for the cosine functions

$r^2(y, x_K)$ is $\frac{B_K^2}{2\overline{y'^2}}$ for the sine functions:

" " $\frac{A_{N/2}^2}{\overline{y'^2}}$ for $N/2$

Notes on expressing power spectrum coefficients.

Can also express $y(t)$ with complex notation:

$$y(t) = \bar{y} + \sum_{K=1}^{N/2-1} H_K e^{i\left(2\pi K \frac{t}{T}\right)}$$

Recall $e^{i\theta} = \cos\theta + i \sin\theta$

Real part of $H_K = A_K$
Imaginary part $H_K = B_K$

The term with the "K" is more compactly expressed in terms of angular frequency (ω)

$$\omega = \frac{2\pi K}{T} \quad \text{or} \quad \frac{2\pi K}{N}$$

* Be aware of these notational conventions in the Hartmann and Wilks references — or you may find yourself confused!

The Power Spectrum

The plot of $c_k^2/2$ vs. k is called the power spectrum of $y(t)$

Frequency spectrum if t is time
Wavenumber spectrum if t is space

The 'line spectrum' (i.e. computing the Fourier transform using the entire length of record) is fine if you have infinite record.

For a finite spectrum, the line spectrum has 3 main drawbacks.

1) Integral values of k do not have any special relationship to population being sampled. Simply chosen by the length of data record, which is usually just what is available.

2) Individual spectral lines have very few degrees of freedom (~ 2 dof)

N data points determine: mean, $N/2$ amplitudes, $N/2 - 1$ phases
(Recall dof relates to samples − # of estimated parameters)

3) Most geophysical data are not truly periodic, but only quasi-periodic, so they're better represented by spectral bands rather than individual spectral lines.

Just about any thing in climate that has the word 'oscillation' in it behaves this way!

e.g. ENSO → 3-7 years
PDO → 20-30 years

Given these drawbacks, most commonly used is the continuous power spectrum, where variance of y is given per unit frequency!

$$\overline{y'^2} = \int_0^{K^*} \Phi(k)\, dk$$

$\Phi(k)$ = continuous power spectrum
K^* = Nyquist frequency
One cycle per $2\Delta t$
Highest frequency that can be resolved.

Can also write as $\Phi(\omega)$, which is commonly done.

Then for a spectral band: