Interpreting Autocorrelation

In discrete form, the autocorrelation C is:

$$C(n) = \frac{\sum_{i=1}^{N-n}(x_i - \overline{x})(x_{i+n} - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{1}$$

for time separations $\Delta t = n\delta t$, where δt is the separation between observations. In essence, we sum $N - n$ elements in the numerator but N elements in the denominator, so as n increases, the value of C will converge to zero. Thus this is called a biased estimator. The Matlab function "xcov" will compute this; in order to carry out the computation in (1), use the option "coeff". Matlab leaves open the debate about how best to normalize the autocorrelation. Matlab defaults to a biased estimator, and this makes sense. Since the autocorrelation at large n is based on fewer estimates, its uncertainty is greater. The biased estimator keeps the uncertain values at large lag from governing our analysis. Matlab will compute an unbiased estimator using the option "unbiased", but in this case it does not be default normalize the zero lag autovariance to be one. (You can normalize by hand if necessary, although Matlab does not have a mode to do it automatically.) The autocorrelation becomes highly variable for large time lags because $N - n$ is small, so few data points are averaged, and the uncertainty in the autocorrelation scales like $1/\sqrt{N-n}$ as n increases. Usually we use a biased estimate of the autocorrelation, but regardless of which one you choose, you should keep in mind that the autocorrelation will be unreliable at large lags.

The useful information in the autocorrelation is concentrated near zero lag. The autocorrelation is 1 at zero lag, and it drops to zero (or below) with increasing time lags.

How do we interpret the autocorrelation? Let's consider three possible scenarios:

(a) A time series of temperatures are independent and completely random. Effectively they are like output from a random number generator, so there is no relationship between measurements at t and $t + \delta t$. In this case, we would expect the autocorrelation to be 1 at $t = 0$ but to drop to zero for all other lags.

(b) Temperatures are completely controlled by an identical sinusoidal seasonal cycle every year, so that January of year 1 is identical to January of year 2, and so forth. Then we'd expect the autocorrelation for $n = 12$ to be 1, like the autocorrelation for $n = 0$. With six months separation, at $n = 6$ or $n = 18$, we would expect temperatures to be completely anticorrelated, with an autocorrelation of -1.

(c) Temperatures vary slowly over months or years, but are not cyclical in any predictable way. In this case, we would expect the autocorrelation to taper gradually from 1 to 0 with increasing lags. The time-scale at which the autocorrelation tapers, the "decorrelation scale" will be a useful measure of the variability in the system.

Real temperature observations often show a combination of effects from scenario (b) and scenario (c).

How do we determine the decorrelation scale for our measurements? A simple measure is to look for the first zero crossing of the autocorrelation function. Once the autocorrelation drops to zero, the data are no longer correlated. Any further positive wiggles are then interpreted as noise.

The zero crossing is by no means the only way to estimate a decorrelation scale. We might alternatively estimate a decorrelation time scale based on twice the time required for C to drop to $1/2$. More formally we can compute it by integrating the autocorrelation function as:

$$\tau = n_d\delta t = \sum_{n=-M}^{M} \frac{N - |n|}{N} C_{\text{unbiased}}(n)\delta t = \sum_{n=-N}^{N} C_{\text{biased}}(n)\delta t. \tag{2}$$

In practice the results drop to zero if you sum over all possible values of C; instead one can vary N in the summation and look for a maximum in the decorrelation time scale τ. Normally the formal estimate for τ does not differ much from the estimate based on the first zero crossing. In this example, the largest plausible decorrelation scale would be 9.3 months as indicate in Figure 1.

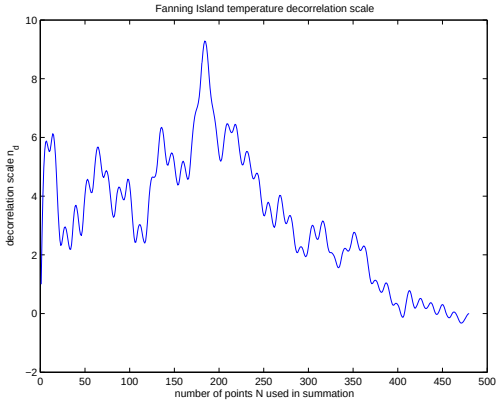

Figure 1: Decorrelation scale n_d for Fanning Island temperature record, computed as a function of N, the number of points over which $C_{\text{biased}}(n)$ is summed.

Degrees of Freedom

The decorrelation scale tells us about typical scales of variability, so it can inform us about the physical characteristics of our system. It is also useful for deciding how many degrees of freedom we have in a set of measurements.

Consider a case where we measure temperature every millisecond, but our thermometer has a slow response time and takes 1000 milliseconds = 1 second to respond to changes. As a result, the 1000 recorded observations each second are all effectively the same. The autocorrelation function will show that it takes about a second for measurements to decorrelate, so $\tau = n_d \delta t = 1000$ ms. After 10 minutes of observation, we would have 600,000 observations which might lead us to believe that our estimate of the mean temperature should have a very samll error bar ($\sigma/\sqrt{N} = \sigma/775$). In reality, our data would represent only 600 independent observations. The effective number of degrees of freedom $N_{\text{eff}} = N/n_d$, so our estimated uncertainty about the mean should be $\sigma/\sqrt{N_{\text{eff}}} = \sigma/24.5$.

The effective number of degrees of freedom should also be considered when we compute the significance of correlation coefficients. Thus the threshold value used to determine whether a correlation coefficient is statistically significant should be

$$r_{sig} = \text{erf}^{-1}(s)\sqrt{\frac{2}{N_{\text{eff}}}}. \tag{3}$$

In the case of our Fanning Island time series, the total number of data points was $N = 480$, corresponding to 40 years of monthly data, but this analysis suggests that we have only about 2 degrees of freedom per year, or $N_{\text{eff}} = 80$. This translates into an increase in the 95% significance level, r_{sig}, from 0.09 for $N = 480$ to 0.22 for $N_{\text{eff}} = 80$. In other words, if we think that consecutive observations are strongly correlated then we have fewer degrees of freedom. If we were sure that this slow variation were an important part of the signal that we are studying, we wouldn't worry about determining our effective number of degrees of freedom. But perhaps our slowly varying signal is just the result of a random process that changes the ocean every six months. Our signal might appear correlated with another slowly varying random signal, but for no real physical reason.