On Different Notions of Correlation, and their Relation to Covariance

Mike Wilensky

March 2021

1 Introduction

The purpose of this memo is to investigate the difference between covariance, and what I and perhaps others think of as correlation. These terms are often used interchangeably, which I think leads to confusion in some instances when one member of the discussion means something other than another member, without explicit acknowledgement that the term ”correlation” is being overloaded.

First, the covariance of two random variables, X and Y, denoted $C(X,Y)$ has a precise mathematical definition, which is

$$C(X,Y) = E[XY] - E[X]E[Y], \tag{1}$$

where $E[\cdot]$ denotes an ensemble expectation. This is a centralized moment in that the means are subtracted off in the expectation. Usually, the “correlation coefficient” is defined as

$$\rho(X,Y) = \frac{C(X,Y)}{\sigma_X \sigma_Y}, \tag{2}$$

where σ_X and σ_Y are the standard deviations of X and Y. This choice of definition makes covariance and correlation practically indistinguishable.

On the other hand, I tend to think of correlation as the following statement: X and Y are strongly correlated if, given X, the value of Y is known with little uncertainty (and vice-versa). Translated into probability theory, this is saying that the conditional distributions, $P(Y|X)$ and $P(X|Y)$ (related via Bayes' Theorem), are strongly concentrated[1] for all values of X and Y. This is different than common mathematical definitions of correlation, which often is just a normalized covariance, but it is a way of defining it that I feel is connoted by many who use the term. Below, I use simulations, some analytic math, and some thought experiments to explore these related, but different concepts.

[1] relative to some measure the analyst cares about

2 Correlation and Covariance are Similar

Figure 1 shows an example of correlated random variables. In this example, X is drawn from a standard normal distribution (mean 0, variance 1). Then, we define Y by

$$Y = X + \mathcal{N}(0, 0.1). \tag{3}$$

This is to say that Y is X with added independent gaussian noise of significantly smaller standard deviation than X. In this case, Y and X are clearly correlated in the sense of the definition given in the introduction. For each X, $P(Y|X)$ is a normal distribution (approximately) centered on the given X, with standard deviation of (approximately) 0.1, which is much more concentrated than the variations in X itself. This is the same as saying that if X is known, then Y can be inferred with little uncertainty. X and Y are strongly correlated.

We can calculate the covariance of X and Y by explicitly plugging equation 3 into equation 1. We obtain

$$C(X, Y) = E[X^2 + NX] - E[X]E[X + N]. \tag{4}$$

Using that expectations are linear, and that the sum or product of two independent zero-mean random variables is also zero-mean, we see that

$$C(X, Y) = \mathrm{Var}[X] = 1. \tag{5}$$

Using the values for our random variables, we have a correlation coefficient of approximately 0.995.

$$\rho(X, Y) \approx 0.995 \tag{6}$$

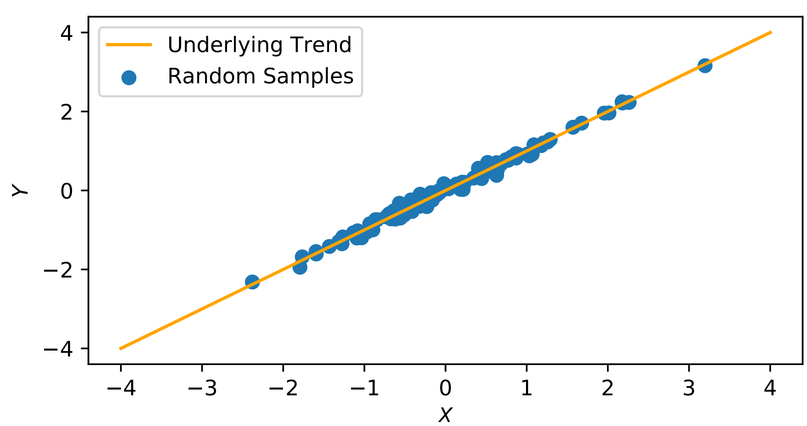

Figure 1: Scatter plot of samples from the relation $Y = X + N$ where X is a standard normal random variable and N is a normal random variable with a standard deviation of 0.1. The correlation between X and Y is clear, which makes sense since Y is simply X with a small noise term.

Indeed, these variables are strongly correlated in the conditional distribution sense and also the covariance sense.

3 Correlation and Covariance are Different

Now, consider the following highly idealized thought experiment. An ensemble of identical massless springs with identical massive blocks attached to each one, are all located on separate space stations in free-fall. Each spring-block system is attached to a wall, and is undergoing tiny, perfectly synchronized sinusoidal oscillations over a frictionless track due to an initial perturbation. A stationary, floating, high-speed camera records the motion of each block. Meanwhile, in the biology lab on each station, a troop of angry baboons pounds on the walls, causing tiny vibrations to propagate through the station. This perturbs the sinusoidal oscillations that the floating camera would otherwise observe by small additive gaussian noise, say $\mathcal{N}(0, 0.1)$. To be explicit, let us say the block position for any spring in the ensemble, $Y(t)$ is given by

$$Y(t) = \cos(\omega t) + \mathcal{N}(0, 0.1) \tag{7}$$

where $\mathcal{N}$ is drawn independently for each time *and* spring. Figure 2 shows an example period of the perturbed oscillation observed by the camera.

If the position of the block on a spring is known at some time, t, then its position one period later is well-constrained by this information as long as the period is known. A similar statement holds for points separated by half a

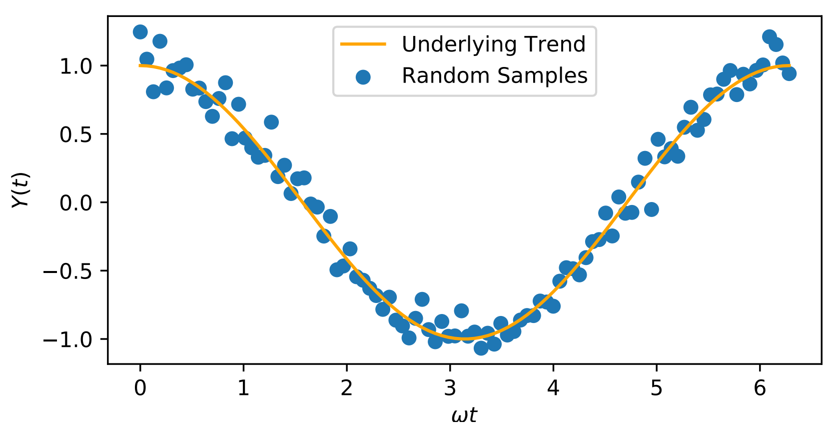

Figure 2: Realization of one oscillation period for a thermodynamic spring. Values of Y separated by one period are well-correlated, while values of Y separated by half a period are anticorrelated. Interestingly, the covariance of any two points is 0.

period. In fact, if the amplitude, frequency, and phase of the oscillation can be constrained by the data, which is true when the additive noise is small relative to the amplitude of oscillation and the oscillation is sampled appropriately, a reasonable guess for any point in the series can be inferred given the value at any other point. In this sense, the random variables $Y(t)$ and $Y(t')$ are well correlated. The conditional distribution of $Y(t')$ is gaussian, where the mean is determined by phase-shifting the appropriate wave-form by $\omega(t'-t)$ (adding this phase to its argument), and the same standard deviation as $Y(t)$.

Now, the prescription for covariance is that we calculate an *ensemble* expectation, which is only equal to a time expectation when the process is ergodic.[2] Now let us suppose, in addition to our assumption of synchornized oscillation, that we also have perfectly synchronized sampling of the spring-block waveforms. This is to say that t is not a random variable in the ensemble sense, and thus neither is $\cos(\omega t)$. This means that

$$C\big(Y(t), Y(t')\big) = E[\cos(\omega t)\cos(\omega t') + \mathcal{N}(t)\cos(\omega t') + \mathcal{N}(t')\cos(\omega t) + \mathcal{N}(t)\mathcal{N}'(t')] \\ - E[\cos(\omega t) + \mathcal{N}(t)]E[\cos(\omega t') + \mathcal{N}(t')] \tag{8}$$

Now, since $\mathcal{N}$ is drawn independently for each time and spring, and is zero-mean, all the terms with $\mathcal{N}$ vanish. Since $\cos(\omega t)$ is constant over the ensemble, and the ensemble average of a constant is just that constant, we have

$$C\big(Y(t), Y(t')\big) = \cos(\omega t)\cos(\omega t') - \cos(\omega t)\cos(\omega t') = 0. \tag{9}$$

This might be shocking to one that considers covariance and correlation equivalent. Earlier we reasoned that, given the appropriate information, one could see that certain points were well-correlated. However, here we see that any two points have precisely zero covariance. I simulated this, and found the estimated covariance to be of order 10^{-5}.

4 Discussion

As with any mathematical endeavor, this has really just been an exercise in adhering to definitions. The definition for correlation that I have used here is in disagreement with definitions in the literature, namely that correlation is a statement about conditional distributions rather than about covariance. It is interesting that sometimes these definitions are at odds and sometimes they are not. To reconcile this disagreement, I offer the following intuition.

When calculating the covariance, the means are subtracted off. This means only a relationship between fluctuations about the means remain. Thus, covariance is a measure of how the fluctuations in two random variables track one another, regardless of the values they are fluctuating about. Since the noise term in the sinusoidal example above was drawn independently for each time and spring in the example above, this resulted in zero covariance.

[2] Something I hardly understand and should not get into here.

On the other hand, the definition of correlation I have listed here answers the question, "If I know the value of one random variable, can I make a good guess about the value of the other variable?" This question cares not only about whether fluctuations track each other, but whether the mean value of the variables track each other in some meaningful way as well. In fact, with this definition of correlation, the fluctuations can fail to track and yet the variables can be well-correlated, as was the case in our spring ensemble.

One final definition of correlation that I have not yet touched upon is the following, which may be familiar to radio astronomers,[3] called the cross-correlation function. The cross-correlation function of two deterministic signals $X(t)$ and $Y(t)$ is commonly defined by

$$R(\tau) = \int_{-\infty}^{+\infty} dt\, X^*(t) Y(t+\tau). \tag{10}$$

In principle one could compute this function for a deterministic signal that is corrupted by noise, and the result is not totally meaningless, particularly if the signal and noise are zero-mean, and the noise is wide-sense stationary, i.e. that the mean and variance do not change in time, or over the integration period should it be finite. In this case, the resulting function seems to agree with the conditional distribution notion of correlation for the sinusoidal example above, as long as the integration period is made finite.

Hopefully this document helps shed some light on the importance of definitional clarity in discussions about mathematical interpretation of data.

[3] :)