How Strongly Correlated are Two Quantities?

Having spent much of the previous two lectures warning about the dangers of assuming uncorrelated uncertainties, we will now address the issue of correlations (albeit without incorporating the possibility of measurement uncertainties). Correlations abound in astronomy, with some of the most famous (such as the quantities plotted on a Hertzsprung-Russell diagram) becoming the basis for whole fields of study.

Our data set for this lecture comes from Stella, White, and Rosner 1986 (ApJ, 308, 669). We extract from their Table 1 the rotation frequencies and the maximum X-ray luminosities for the accreting neutron stars in supergiant X-ray binaries:

$f_{\rm rot}$(Hz)	L_x(max)(erg s^{-1})
0.00143	1×10^{37}
0.00189	4×10^{36}
0.00353	6×10^{36}
0.00387	3×10^{36}
0.20833	8×10^{37}
1.42857	6×10^{38}

We plot these data in Figure 1, in log-log space.

We'd like to know how strongly these two observables are correlated. To do that, we need to introduce the concept of a covariance matrix (also sometimes called a variance-covariance matrix).

Covariance matrix

Suppose that we measure several properties each of a number of objects. We'd like to know whether those properties are correlated, and if so how strongly and whether they are correlated or anticorrelated. The standard way to do this is to use a covariance matrix. To define this, recall that the variance of some number of measurements of a variable x is $\langle (x-\mu)^2 \rangle$, where μ is the arithmetic mean of x over the measurements and the angle brackets denote an average over the measurements. If for each object you measure n quantities $x_1, x_2, \ldots x_n$, then the covariance matrix is

$$\Sigma = \begin{pmatrix} \langle (x_1-\mu_1)(x_1-\mu_1) \rangle & \langle (x_1-\mu_1)(x_2-\mu_2) \rangle & \cdots & \langle (x_1-\mu_1)(x_n-\mu_n) \rangle \\ \langle (x_2-\mu_2)(x_1-\mu_1) \rangle & \langle (x_2-\mu_2)(x_2-\mu_2) \rangle & \cdots & \langle (x_2-\mu_2)(x_n-\mu_n) \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle (x_n-\mu_n)(x_1-\mu_1) \rangle & \langle (x_n-\mu_n)(x_2-\mu_2) \rangle & \cdots & \langle (x_n-\mu_n)(x_n-\mu_n) \rangle \end{pmatrix}$$

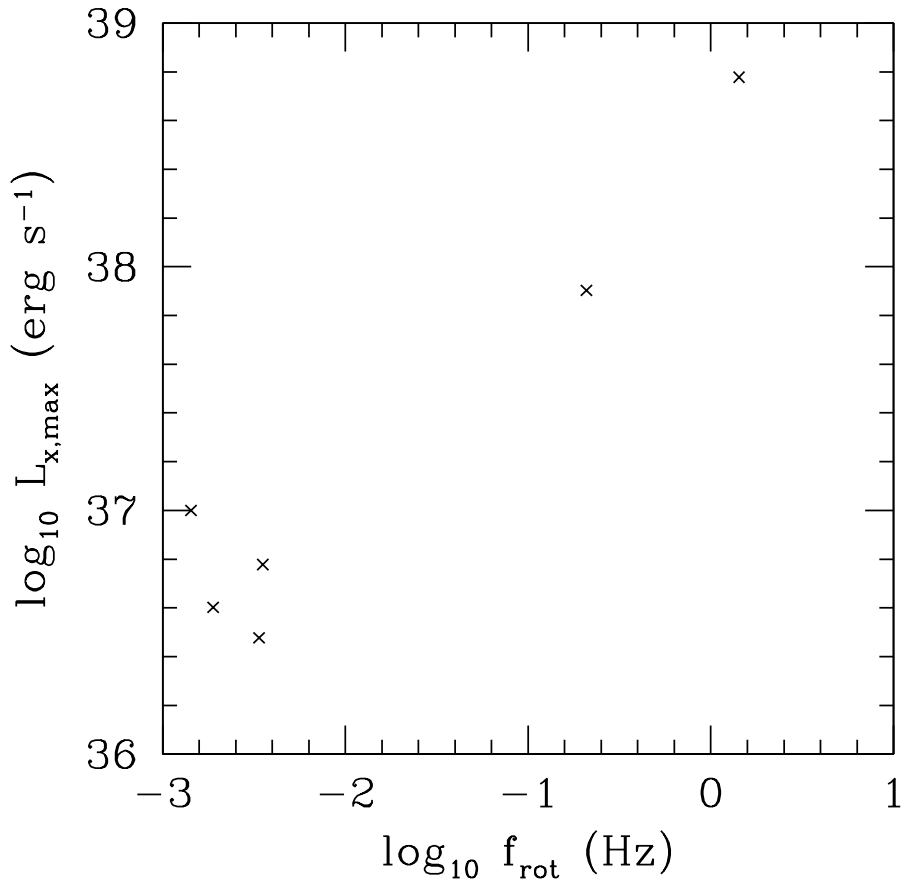

Fig. 1.— Log_{10} of the neutron star rotation frequency versus $\log_{10}$ of the maximum X-ray luminosity for six supergiant X-ray binaries. Original data from Table 1 of Stella, White, and Rosner 1986 (ApJ, 308, 669).

You can see that the matrix is symmetric (the ij component equals the ji component for any i and j). If we designate the standard deviations of the variables by $\sigma_1, \sigma_2, \ldots, \sigma_n$, then we can write the covariance matrix in the form

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{pmatrix}$$

Here ρ_{ij} is Pearson's correlation coefficient between variables x_i and x_j; clearly $\rho_{ij} = \rho_{ji}$. ρ_{ij} can have any value between -1 (a perfect anticorrelation) to $+1$ (a perfect correlation), and $\rho_{ij} = 0$ means that the two variables are not correlated at all.

An aside: principal component analysis.—Believe it or not, our setup so far has allowed us to go much of the way toward principal component analysis, which is a technique often used to make some sense out of balls of data. Correlation coefficients only test linear correlations between *pairs* of parameters. But there are plenty of cases in which multiple parameters might align with each other in some fashion. For example, this is the basis for various "fundamental planes" in different areas of study (e.g., the relation between the effective radius, the average surface brightness, and the central velocity dispersion of elliptical galaxies).

You'd like to be able to extract linear combinations of the measured parameters that have significant correlations with each other. A simple way to do this is to construct the covariance matrix as above, and then find the eigenvalues and eigenvectors of the matrix (I use "jacobi.c" from Numerical Recipes, which works for real symmetric matrices such as the covariance matrix). The eigenvectors give the linear combinations of your original parameters corresponding to the eigenvalues. The largest eigenvalue indicates the longest axis, i.e., the direction in which your quantities are most correlated, your second-largest eigenvalue indicates the direction with the second-largest correlation, and so on.

For example, for our data set above, we have two quantities and therefore two eigenvalue/eigenvector combinations. The larger eigenvalue is 1.94 and the associated eigenvector is (0.815065,0.57937); note that the first coefficient is for $\log_{10} f_{\rm rot}$(Hz) and the second is for $\log_{10} L_x$(max)(erg s^{-1}), so this points in the lower-left to upper-right direction that is obviously the primary correlation direction. The smaller eigenvalue is 0.042 and the associated eigenvector is (-0.57937,0.815065). Note that the eigenvectors are orthonormal due to the algorithm used in jacobi.c. For more complicated data sets, with many measured quantities, principal component analysis can help you find linear combinations of parameters that are potentially significantly correlated.

Now back to correlation coefficients. This sounds easy! In our case we only have two

variables, so we just plug the numbers into the tables and compute ρ_{12} between the frequency and the maximum luminosity (not between the log frequency and the log luminosity). We find that $\rho_{12} = 0.9997$.

What?!? That's ridiculous; do the data *look* like they form a perfect straight line? Have we performed the calculation incorrectly?

No, but we have run into a problem with blind use of the correlation coefficient. What's going on can be seen more clearly in Figure 2: two of the points have much higher luminosity and rotation frequency than the others, and the line between those two points happens to more or less pass through the origin. The remaining four points are pretty well scattered, but because all of those stars have low luminosities and rotation frequencies, they look like a point.

Now it's your turn: calculate the correlation coefficient between the frequency and maximum X-ray luminosity for Be X-ray binaries (the data set on the website is also taken from Stella, White, and Rosner 1986). What do you get? What do you get if you remove a point or two? Do you draw any particular conclusions?

Our analysis leads us to realize the first problem with the Pearson's correlation coefficient:

The coefficient is highly sensitive to a large range in values

We can sharpen our understanding of the problem by removing the 0.20833 Hz point from our data set and recalculating the correlation coefficient. Now instead of a mere 0.9997, the correlation coefficient becomes 0.99994(!). Thus the huge range of values, and in particular the large span between the largest-luminosity (and largest-frequency) point and all the other points, has the effect of collapsing the entire data set into two points: the high point and all the rest. In astronomy, particularly with relatively small data sets, it is fairly common that the biggest of the set (whatever "biggest" would mean in that context) is much bigger than the next biggest. Thus if you just throw the data into your code to compute the correlation coefficient, you could incorrectly convince yourself that you have a strong correlation.

Just for fun, I decided to pursue this by generating a synthetic data set in which one point was at $(x, y) = (1000, 1000)$ and the rest had x and y drawn randomly and independently from 0 to 1. Even when there were 99 random points, the formal correlation coefficient was 0.97. With no other context, you would thus conclude that x and y are strongly correlated. But if you eliminate the single high point, they aren't correlated at all. This leads to the second problem:

It is not easy to define what the correlation coefficient means

Suppose that you have a correlation coefficient of 0.3. Or 0.97. Or 0.9997. Does this

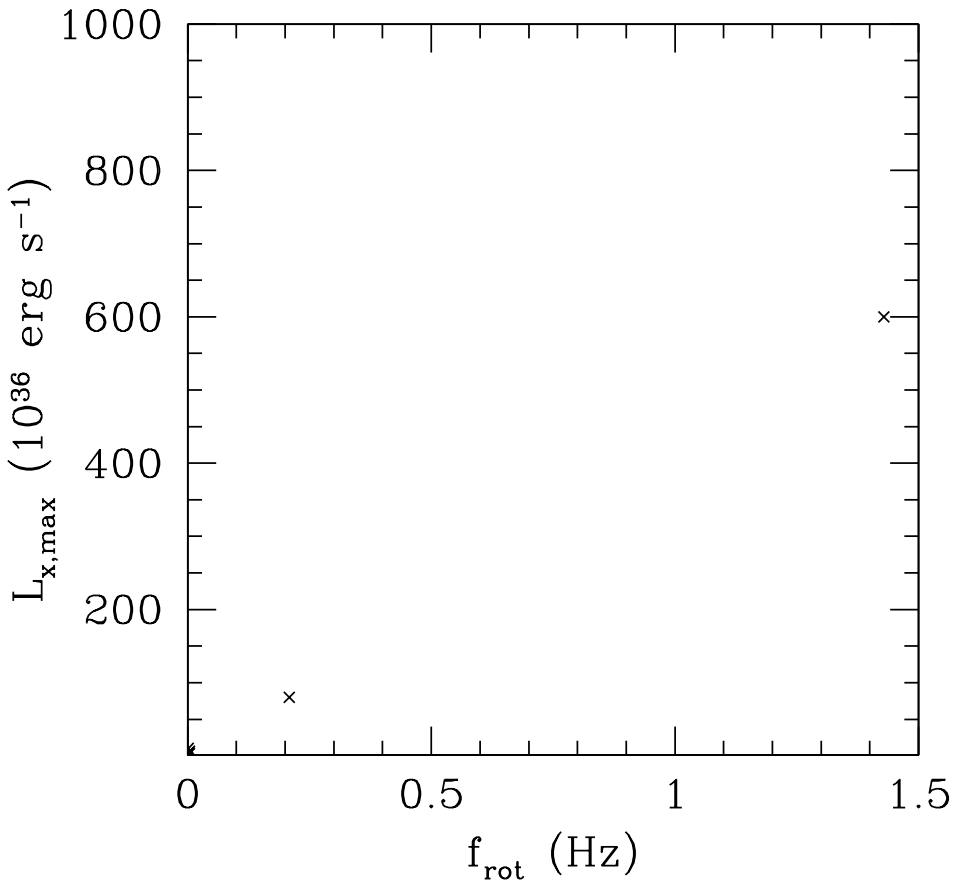

Fig. 2.— Neutron star rotation frequency versus the maximum X-ray luminosity for six supergiant X-ray binaries. These are the same data as we plotted in Figure 1. Now however, what strikes the eye are two points in a line that points back to the cluster of four points at lower luminosity. A standard calculation of the correlation coefficient returns a value of $\rho = 0.9997$; does this make sense, give the log-log plot in Figure 1?

imply that the quantities are strongly related to each other? As we've seen, that's not easy to tell. Even if you think that you have shown that two quantities really are strongly correlated (say, for example, that you've done 100 measurements and there are no big outliers), you have to keep in mind that correlation does not necessarily imply causation. For example, did you know that there is a strong tendency for elementary school students with larger feet to spell better? It's actually true!

And finally, our last problem:

The Pearson correlation coefficient only tests linear correlations

Look at Figure 3. This is an unnecessarily detailed plot of $y = x^2$. Clearly, x and y are strongly correlated with each other. And yet, the Pearson's correlation coefficient between them is 0. The particular answer returned by a Pearson's correlation coefficient analysis only relates to linear correlations. You could, of course, look at this plot, guess that something like $y = x^2$ is involved, and try to correlate the square root of y with x. But in a more complex situation the choices might not be as obvious.

So what should you do?

It's not an easy question. The calculation of linear correlation coefficients is easy and relatively fast (as long as you don't have too many data points or variables). In the spirit of fast exploration I'd be happy to calculate the linear correlation between two quantities, but one might argue that simply plotting x versus y would achieve the same effect. Indeed, the plot might do better if your intent is to explore correlations, because your eye can pick up patterns that are more elaborate than mere lines. The risk in that case is that your eye will pick up patterns that aren't there; for our ancestors, seeing a leopard that isn't there had much less downside than not seeing a leopard that is there, which might help explain our tendency to impose patterns on data!

If we want to be rigorous Bayesians, we need to specify *in advance* what relation we're looking at. Let's think about how we would proceed.

A typical application of a correlation coefficient is to get a sense for whether two quantities are linearly related to each other. But this doesn't require anything special, from the Bayesian perspective. Suppose that we have two quantities, x and y, and we'd like to know whether there is a linear relation between them. Our two models are:

Model 1: x and y are not related, which is to say that if we use x as the independent variable, our hypothesis is that $y = y_0$, a constant, independent of x. Thus this model has as its single parameter y_0.

Model 2: x and y are linearly related, which is to say that $y = ax + b$ if we use x as the independent variable. Thus this model has two parameters: a and b.

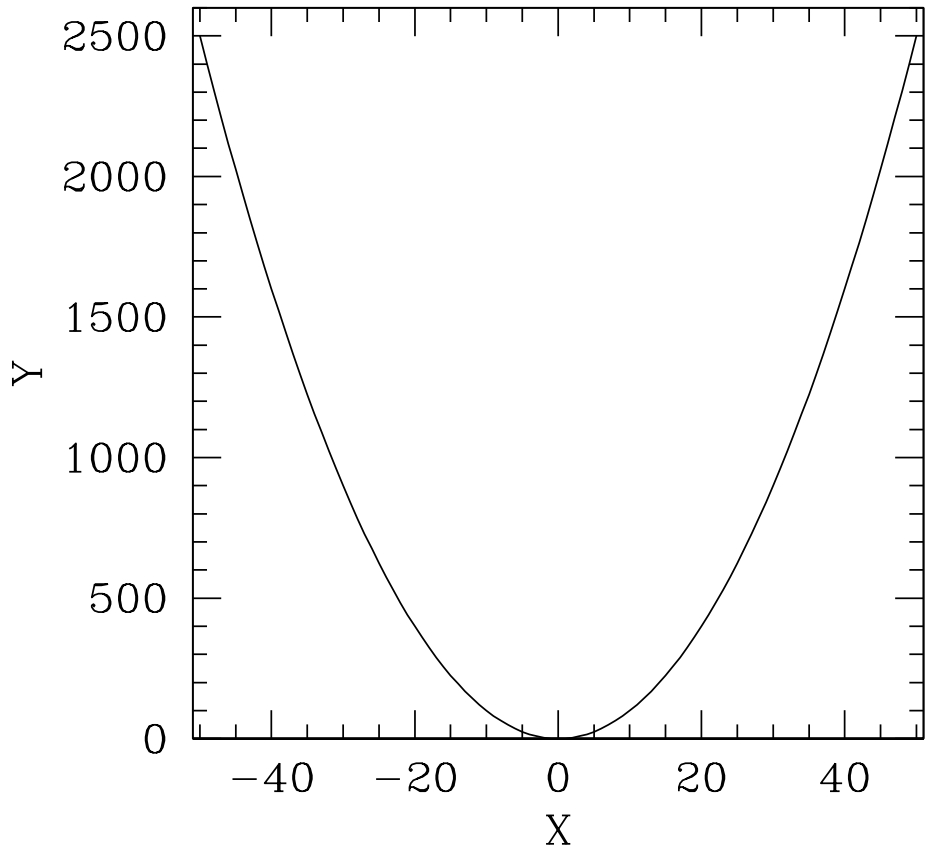

Fig. 3.— A plot of $y = x^2$ from $x = -50$ to $x = +50$. Although the correlation is obvious, the Pearson's correlation coefficient is zero, because it measures only linear correlations.

Note that Model 1 is a special case of Model 2, in which $a = 0$. We can do model comparison as before, by computing the Bayes factor given the data. If we explicitly assume that prior to analyzing the data, we give equal probability to Model 1 and Model 2, then the final odds ratio is just the Bayes factor.

What would emerge from this analysis *is* specifically meaningful; it tells us the degree to which a linear relation between the two specified quantities is preferred to having those quantities unrelated. In addition, of course, we can easily put in more complex models (for example, quadratic relations).

But going back to the question of "what should you do?", there is an important point to make. Suppose that you don't know what you are looking for. You plot your data, and some pattern seems evident. Maybe it's a linear relation, maybe it's something else. Now, you have something specific to examine, so you want to know how significant it is. You therefore do a test for a linear correlation, or do a model comparison of models suggested by the data, or something else, and report the resulting significance.

Do you see the problem? If you proceed in this way, you are succumbing to *a posteriori* analysis. That is, you saw something interesting and then tried to figure out how likely that particular thing was. That's a statistical no-no.

In many cases this situation won't arise. You'll approach your data looking for something particular, and will do your analysis based on that. But in other cases, you're on a fishing expedition: maybe you have a brand-new survey that might reveal something unexpected. After all, such discoveries are part of what motivates us to do science! What should you do then?

If you're really disciplined, then one recommended approach is to wall off some small fraction of your data (say, 10%) and designate that as a playground in which you can do any analysis you want. This is, for example, a method adopted by the LIGO team. The point is that with that small fraction of your data, you don't worry about statistical trials or *a posteriori* analyses; plot it, look for trends, do lots of test analyses; have fun! After you have debugged your analysis codes, and after you have decided what you really want to do with your data, *then* you analyze the remaining 90%. *That* analysis is what would count.

Note that it is very important that *only* the analysis of the remaining 90% count in your final report. Why? Suppose that you thought you saw a strong trend in the 10% playground data. If you include that 10% data in your final analysis, then the trend will be there, at least a little bit. Thus the evidence for the strength of the trend is compromised.

Let me give a specific example that is not hypothetical, but to which I won't attach names. One exciting type of analysis these days is the analysis of quasar light curves to search for periodicities, with the implication that if there is real periodicity in a particular

object then the object might actually be a binary supermassive black hole. That would have lots of important implications.

Such searches are, however, difficult. Quasar light curves vary a lot, and thus finding periodicities in them is challenging. Various teams have analyzed large data sets and have claimed significant signals in some cases. Let's suppose that in a particular case, the light curves for a few hundred thousand quasars have been analyzed, of which a few tens are considered to be significant at the 10^{-5} level (defined somehow; such reports are often themselves incorrect). That's more than you might expect by chance, so if we assume that the analysis was carried out correctly, we have some hope that there might be a few binary supermassive black holes in the lot. However, given that the significance of any individual candidate isn't overwhelming, we'd like to be more confident in the best candidates.

A good way to do that is to accumulate more data. One group, which did that extra accumulation, then looked at their tens of candidates again, now with the whole data set. They argued that because they are looking at tens of candidates, rather than the initial few hundred thousand, their criterion for significance can change: rather than requiring 10^{-5} (defined somehow), they required only 10^{-2} because of the smaller number of candidates. They found that about half their original sample passed that test, so they announced that they had found very strong confirming evidence of the signals. Hooray!

But hold on. Their 10^{-2} estimate came from analyzing the *whole* data set, including the initial part that convinced them that an individual quasar was a good candidate for binarity. Thus adding new data took the overall significance from $\sim 10^{-5}$ to $\sim 10^{-2}$; that is, the new data strongly *reduced* the significance of the signal! As a result, their work actually showed that *none* of the initial candidates is strong.

If they had used the binary solution from the original data (phase, frequency, and amplitude) and then analyzed *only the new data* and found a significance of 10^{-2} that would have been different; the analysis would not have been biased by the original data. But that's not what they did.

In summary, correlation coefficients have many traps and deficiencies. Be careful!