Correlation Coefficients

A correlation coefficient (R) is a helpful way to estimate how connected two variables are to each other. If you've ever wondered how many extra boxes of tissue you buy when you get a bad cold and run a fever (is it one box per added degree?), you've employed correlation. A positive correlation occurs when one variable increases as another does, a negative correlation occurs when one variable increases as another decreases, and no correlation occurs when as one variable increases there is no discernible trend in another. Values near to unity (+1, or -1) are the strongest, and values near zero (between -0.5 and $+0.5$) indicate weak correlations.

In Figure 1, we observe the appearance of two variables with differing levels of correlation. We have deliberately left off any definition of the quantities being measured, so that you will focus on the trends that you see in the data. We can see that there is a high degree of correlation whether y changes linearly or has a more complex relationship to x, as long as the two variables change together (they increase, or decrease, together). Correlation is strong when if you know the behavior of x, you can predict the behavior of y accurately, and correlation is weak when they have little or no connection.

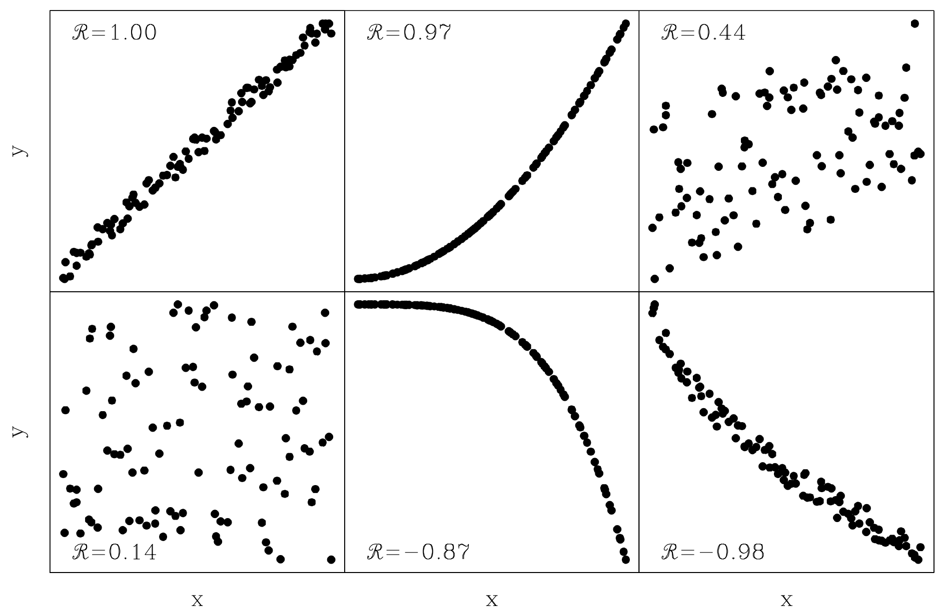

Figure 1: These six panels show the change in the correlation coefficient (R) from extremely positive (upper left) to weak (third and fourth panels) to extremely negative (lower right). For $-0.5 < R < 0.5$ there is no strong relationship between x and y. The values of x increase from left to right, and those of y increase from bottom to top, in each panel.

Correlation can be a very powerful tool in astronomy. Let's examine a particular usage, determining how far away a distant galaxy lies by comparing the patterns in the light which its stars emit with those found in objects at rest. Because of the expansion of the Universe, light emitted from distant galaxies appears at longer wavelengths (it seems redder) than it would in a laboratory here on Earth. The farther the galaxy lies from us, the larger the shift in its spectral features. Because we know the wavelengths at which certain spectral feature should appear for an unmoving object, we can calculate the shift, and thus determine the distances to galaxies which lie far, far away from us.

Figure 2, illustrates how to take two spectra, and shift one in wavelength along the x-axis until it lines up with the other. In each panel, compare the offset between the two lines with the value of the correlation coefficient.

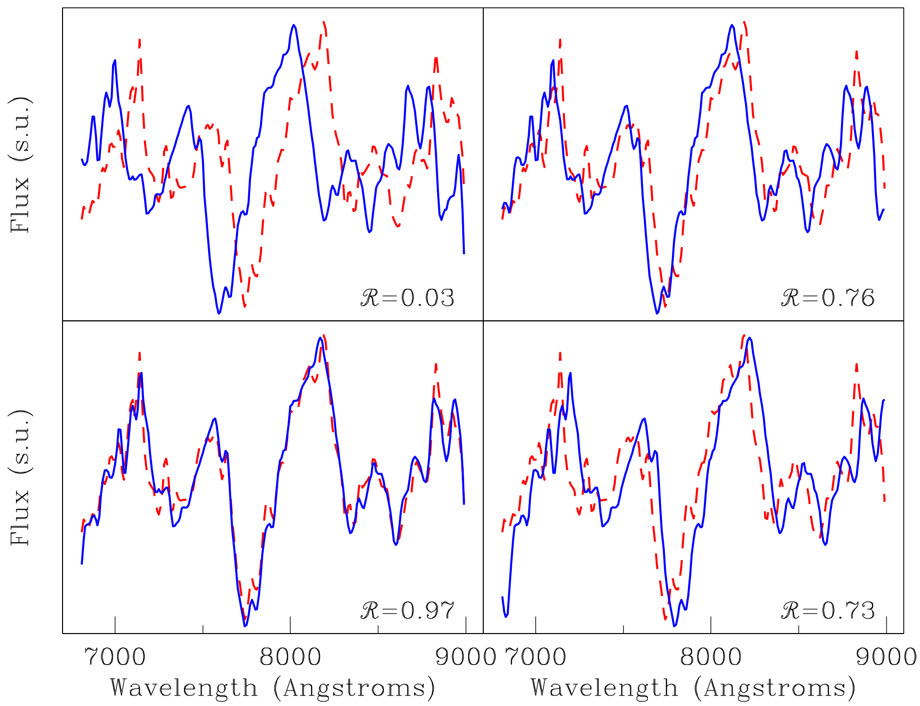

Figure 2: These four panels show the change in correlation coefficient (R) as we shift a solid blue line representing a galaxy spectrum to the right or left until it matches up with a red dashed line representing the position of the peaks and valleys as they would appear in an object at rest. The larger the shift in wavelength, the larger the relative velocity between the Earth and this galaxy. The correlation coefficient quickly drops to below 0.80 as the galaxy signal shifts the position of the peaks slightly off-center, and drops to around zero once the features cease to line up at all. When the two spectra overlay each other, once can see the alignment and also observe small variations in signal strength between the two lines.

In which panel of Figure 2 do the two spectra line up best with each other? Does this panel also have the highest correlation coefficient? (Yes / No)