Lecture 24: Covariance

Anup Rao

May 29, 2019

We have seen a few statistics that are useful to understand real valued random variables. We can measure the expected value (a.k.a. mean) and the variance (or closely related, the standard deviation). We can measure all of the moments. All of this information is captured in the moment generating function of the random variable.

Today, we discuss random variables that take values in higher dimensional space. What are the statistics that capture their distribution? To keep things simple, let us start by talking about 2-dimensional points; points in the plane. Suppose $A = (X_1, Y_1)$ is a random point sampled uniformly at random from the 4 points shown in Figure 1, and $B = (X_2, Y_2)$ is a random point sampled uniformly at random from the 16 points shown in Figure 2. The distribution of X_1, X_2, Y_1 and Y_2 are all identical. Each of them is uniformly random in the set $\{1, 2, 3, 4\}$. But the distributions of A and B are quite far from each other.

Both A and B are random points in the plane. How are they different? One way they are different is that in A the two coordinates are very *correlated*—knowing that X_1 is larger than its expectation tells you that Y_1 is also larger than its expectation. In B the two coordinates are independent—knowing X_2 gives you no information about Y_2.

Now, what about the point $C = (X_3, Y_3)$ which is sampled uniformly at random from the points shown in Figure 3? X_3 and Y_3 are certainly not independent—for example, knowing that $X_3 = 2$ tells you that Y_3 cannot be 3—but X_3 and Y_3 certainly do not determine each other. If X_3 is larger than its expectation, would you expect Y_3 to also exceed its expectation?

What we are after is a general quantitative measure to help us answer this kind of question. This is exactly what *covariance* gives us.

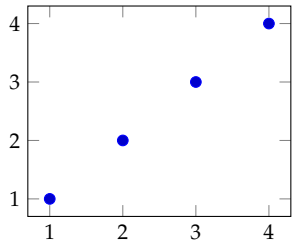

Figure 1: X_1, Y_1 determine each other.

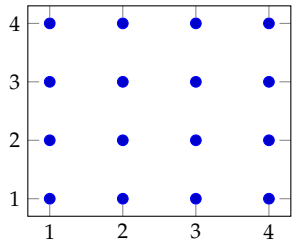

Figure 2: X_2, Y_2 are independent.

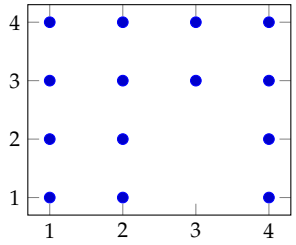

Figure 3: X_3, Y_3 are correlated, but to what degree?

Covariance

The covariance of two random variables X, Y is defined to be

$$\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Immediately, from the definition, you see that this generalizes the concept of variance, since

$$\mathrm{Cov}[X, X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathrm{Var}[X].$$

Recall that for variance, we showed that $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. There is a similar formula that holds for covariance.

Fact 1. $\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Proof. By linearity of expectation:

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

□

Intuitively, if X and Y were truly independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ and the covariance would be 0. A positive covariance means that X, Y are likely to be high at the same time, and a negative covariance means that when one is high, the other is likely to be low.

Returning to our examples from the previous section, we see that

$$\begin{aligned}\text{Cov}[X_1, Y_1] &= 1.25, \\ \text{Cov}[X_2, Y_2] &= 0, \\ \text{Cov}[X_3, Y_3] &= 6.5 - (34/14)(37/14) = 0.0816.\end{aligned}$$

So X_3, Y_3 are slightly positively correlated.

It is important to understand that just because two variables have 0 covariance *does not mean* that they are independent. For example, consider a uniformly random point $D = (X_4, Y_4)$ sampled from the points shown in Figure 4. You can check that $\text{Cov}[X_4, Y_4] = 0$, even though these two variables are not independent.

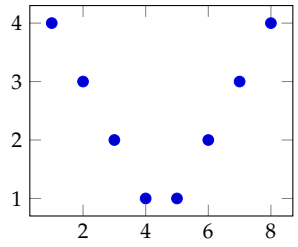

Figure 4: The two coordinates have 0 covariance.

Intuition: Covariance and Linear Regression

Given this last example, one might wonder what the covariance is really measuring. It turns out that there is a nice geometric interpretation of covariance. In a sentence—$\text{Cov}[X, Y]$ is related to the slope in the best linear explanation of the dependence between X and Y.

To understand this, let us look at some examples. Suppose X is a random variable, and $Y = aX + b$, for some constants a, b. Then by linearity of expectation:

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[X(aX + b)] - \mathbb{E}[X]\mathbb{E}[aX + b] \\ &= \mathbb{E}\left[aX^2 + bX\right] - a\,\mathbb{E}[X]^2 + b\,\mathbb{E}[X] \\ &= \mathbb{E}\left[aX^2 + bX\right] - a\,\mathbb{E}[X]^2 + b\,\mathbb{E}[X] \\ &= a(\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2) = a \cdot \text{Var}[X].\end{aligned}$$

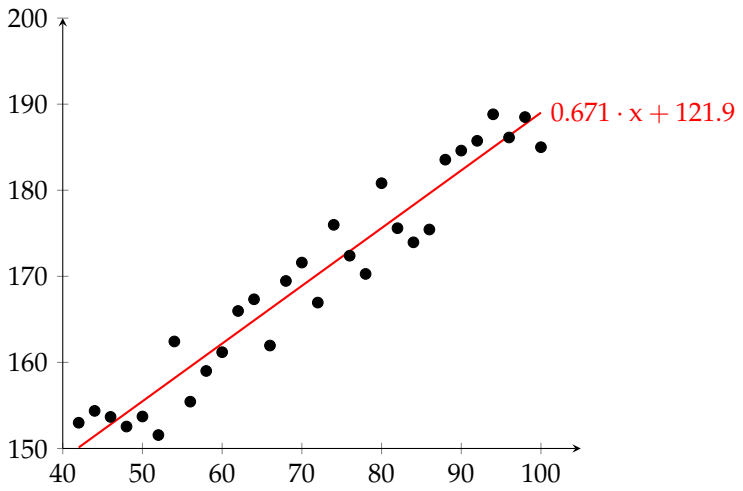

Figure 5: This data was generated randomly near a line.

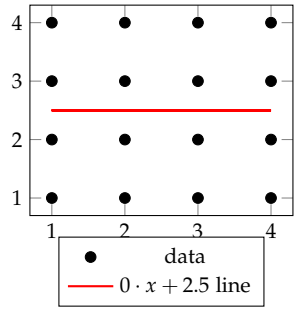

Figure 6: This is the best line approximating the data in Figure 2.

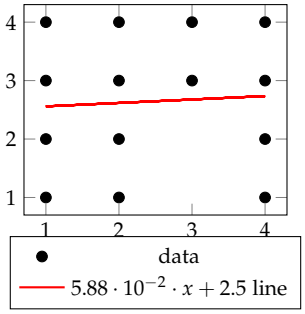

Figure 7: This is the best line approximating the data in Figure 3.

So, in this particular case, we have $a = \frac{\text{Cov}[X,Y]}{\text{Var}[X]}$. In fact, this is not a coincidence.

Suppose now that X, Y are arbitrary real valued random variables. We would like to find the numbers a, b such that $Y \approx aX + b$. Our measure of the error in this approximation will be the expected value of the square of the distance: $\mathbb{E}\left[(Y - aX - b)^2\right]$. What are the a, b that minimize this expected error?

To simplify the problem, suppose that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ and $\text{Var}[X] = 1$. Observe that making the means 0 and variances 1 corresponds to shifting and scaling the points in our examples. If we invert the process, we obtain the best fit line for the original data.

We can compute a, b using calculus. Using linearity of expectation and the above assumptions, we have

$$\begin{aligned}
&\mathbb{E}\left[(Y - aX - b)^2\right] \\
&= \mathbb{E}\left[Y^2 + a^2X^2 + b^2 - 2aYX - 2bY + 2abX\right] \\
&= \mathbb{E}\left[Y^2\right] + a^2\,\mathbb{E}\left[X^2\right] + b^2 - 2a\,\mathbb{E}[YX] - 2b\,\mathbb{E}[Y] + 2ab\,\mathbb{E}[X] \\
&= \text{Var}[Y] + a^2 + b^2 - 2a\,\mathbb{E}[XY].
\end{aligned}$$

To compute the best value of b, we take the derivative of this expression with respect to b, and set the derivative to be 0. This gives

$$0 = \frac{d\,\mathbb{E}\left[(Y - aX - b)^2\right]}{db} = 2b,$$

so $b = 0$. To compute the best value of a, we set the derivative to be 0:

$$0 = \frac{d\,\mathbb{E}\left[(Y - aX - b)^2\right]}{da} = 2a - 2\,\mathbb{E}[XY],$$

giving that $a = \mathbb{E}[XY] = \text{Cov}[X, Y]$.

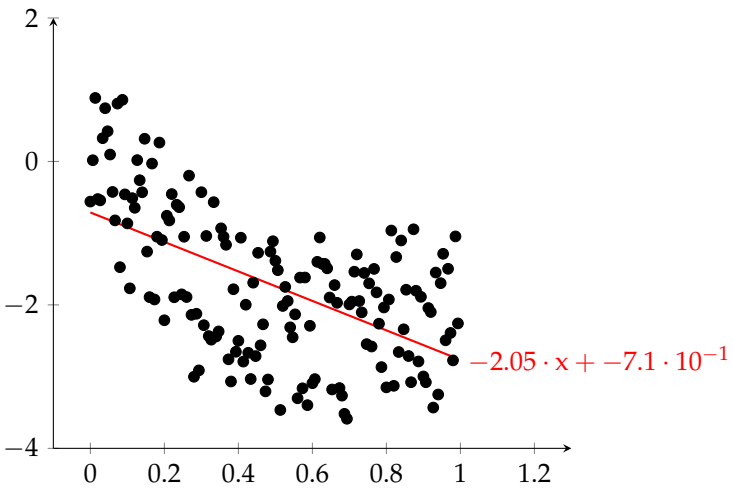

Figure 8: Another example of linear regression.

So, if the means of X, Y are 0 and the variance of X is 1, then the best fit line is

$$y = \text{Cov}[X, Y] \cdot x.$$

To compute the best fit line in general, set $X' = \frac{X - \mathbb{E}[X]}{\sigma(X)}$ and $Y' = Y - \mathbb{E}[Y]$. The best fit line for X', Y' is then

$$y = \text{Cov}\left[X', Y'\right] \cdot x = \frac{\text{Cov}[X, Y]}{\sigma(X)} \cdot x$$

since X', Y' satisfy the conditions above. Now we can recover the best fit line for X, Y by rescaling and shifting this line. Scaling the line by $\sigma(X)$ in the horizontal direction gives the line

$$y = \frac{1}{\sigma(X)} \cdot \frac{\text{Cov}[X, Y]}{\sigma(X)} \cdot x = \frac{\text{Cov}[X, Y]}{\text{Var}[X]} \cdot x.$$

Finally, shifting the line by $\mathbb{E}[X], \mathbb{E}[Y]$ gives the line:

$$y = \frac{\text{Cov}[X, Y]}{\text{Var}[X]} \cdot x - \mathbb{E}[X] \cdot \frac{\text{Cov}[X, Y]}{\text{Var}[X]} + \mathbb{E}[Y].$$

This is the solution to linear regression in general.

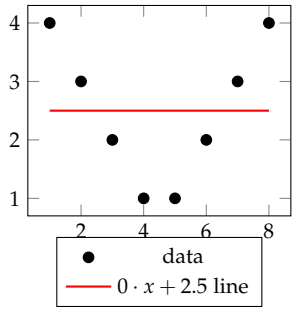

Figure 9: This is the best line approximating the data in Figure 4.