Lecture 4: Correlation, covariance, and random walks

Recap

Lecture 3 examined transformation from one probability density function, joint probability density functions, and conditional probability.

When we think about joint probability and conditional probability, it's natural to think about how quantities correlate, and we started discussing this with definitions of covariance of x and y:

$$C_{xy} = \langle x'y' \rangle, \tag{1}$$

and correlation

$$\rho_{xy} = \frac{\langle x'y' \rangle}{\sqrt{\langle x'^2 \rangle \langle y'^2 \rangle}}. \tag{2}$$

We started outlining a linear model that links x and y in the form:

$$\hat{y}' = \alpha x', \tag{3}$$

where α is a constant chosen to make $\hat{y}$ approximate y.

In Lecture 3, we defined the the mean-square error (mse):

$$\epsilon = \langle (\hat{y} - y)^2 \rangle = \alpha^2 \langle x^2 \rangle - 2\alpha \langle xy \rangle + \langle y^2 \rangle. \tag{4}$$

We found the best α to minimize the mse by differentiating with respect to α, setting the result equal to zero, and solving for α. Because $\epsilon \to \infty$ as $\alpha \to \pm\infty$, the result is a minimum.

$$\frac{\partial \epsilon}{\partial \alpha} = 2\alpha \langle x^2 \rangle - 2\langle xy \rangle = 0. \tag{5}$$

Thus:

$$\alpha = \frac{\langle xy \rangle}{\langle x^2 \rangle} \tag{6}$$

The term α is a regression coefficient, and it assumes a fully linear relationship between x and y.

We then plugged α into the equation for the mse to find the misfit:

$$\epsilon = \langle y^2 \rangle \left(1 - \frac{\langle xy \rangle^2}{\langle x^2 \rangle \langle y^2 \rangle} \right) \tag{7}$$

$$= \langle y^2 \rangle \left(1 - \rho_{xy}^2 \right) \tag{8}$$

Fraction of variance explained

If we take the results for our linear model, we can now ask what fraction of the variance in y (that is $\langle y^2 \rangle$) is explained by our model $\hat{y}$ (with variance $\langle \hat{y}^2 \rangle$).

$$\text{fraction of variance explained} = \frac{\langle \hat{y}^2 \rangle}{\langle y^2 \rangle} \tag{9}$$

$$= \frac{\alpha^2 \langle x^2 \rangle}{\langle y^2 \rangle} \tag{10}$$

$$= \frac{\langle xy \rangle^2}{\langle x^2 \rangle^2} \frac{\langle x^2 \rangle}{\langle y^2 \rangle} \tag{11}$$

$$= \frac{\langle xy \rangle^2}{\langle x^2 \rangle \langle y^2 \rangle} \tag{12}$$

$$= \rho_{xy}^2. \tag{13}$$

Thus the correlation squared is a measure of the skill of our linear model.

We can compute the correlation coefficient fairly efficiently. In Matlab, for random noise, we have

```
N=100000;
x=randn(N,1); y=randn(N,1);
corrcoef(x,y)
```

and in python the syntax is similar:

```
N=100000
x=np.random.normal(size=N)
y=np.random.normal(size=N)
np.corrcoef(x,y)
```

Correlation, joint pdfs, and rotation

When we looked at a joint pdf in class, we considered a case when x and y were dependent on each other, for example by defining x to be a normally distributed random number, and y to be a different normally distributed random number added to x, as in:

```
N=100000;
x=randn(N,1); y=randn(N,1)+x;

histogram2(x,y,'Normalization','pdf','DisplayStyle','tile'); colorbar
xlabel('x','FontSize',14)
ylabel('y','FontSize',14)
h=gca;
set(h,'FontSize',14)
```

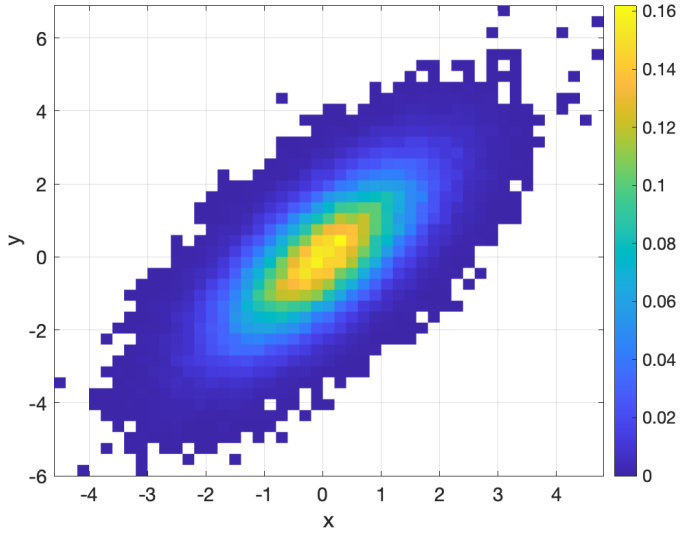
or in python (with a bonus variant):

```
N=1000000
x=np.random.normal(size=N)
y=np.random.normal(size=N)+x
z=np.random.normal(size=N)+x

plt.figure(figsize = (10,4))
plt.subplot(1,2,1)
H, xedges, yedges = np.histogram2d(x,y,[30,30],density=True)
plt.pcolormesh(xedges,yedges,H.T)
plt.colorbar()

plt.subplot(1,2,2)
H, xedges, yedges = np.histogram2d(y,z,[30,30],density=True)
plt.pcolormesh(xedges,yedges,H.T)
plt.colorbar()
```

Since x and y are correlated, their joint pdf is tilted. We can represent this as an ellipse that is tilted by angle θ relative to the standard x–y coordinate system. The major axis of the ellipse is along the rotated coordinate $\hat{x}$, and the minor axis is along the rotated coordinate $\hat{y}$.



We can use the covariance to determine the principal axes. Given two random variables, the goal is to find a coordinate rotation that results in two variables that are uncorrelated. Such a rotation might be desirable if the two variables are eastward and northward current near a coast of arbitrary orientation. The rotated axes give the directions of maximum and minimum fluctuation. Define the complex random variable z as:

$$z = x + iy \tag{14}$$

The variable $\hat{z}$ in a coordinate system rotated by an angle θ is given by:

$$\hat{z} = ze^{-i\theta} \tag{15}$$

The variables $\hat{x}$ and $\hat{y}$ in the rotated coordinate system are found by considering the real and imaginary parts of (15).

$$\hat{x} = x\cos\theta + y\sin\theta \tag{16}$$
$$\hat{y} = -x\sin\theta + y\cos\theta \tag{17}$$

Then the covariance of $\hat{x}$ and $\hat{y}$ is

$$\langle\hat{x}\hat{y}\rangle = \left(\langle y^2\rangle - \langle x^2\rangle\right)\cos\theta\sin\theta + \langle xy\rangle\left(\cos^2\theta - \sin^2\theta\right) \tag{18}$$

We want to find the angle θ that makes the covariance be zero. Setting (18) to zero, and using a couple of trigonometric identities yields

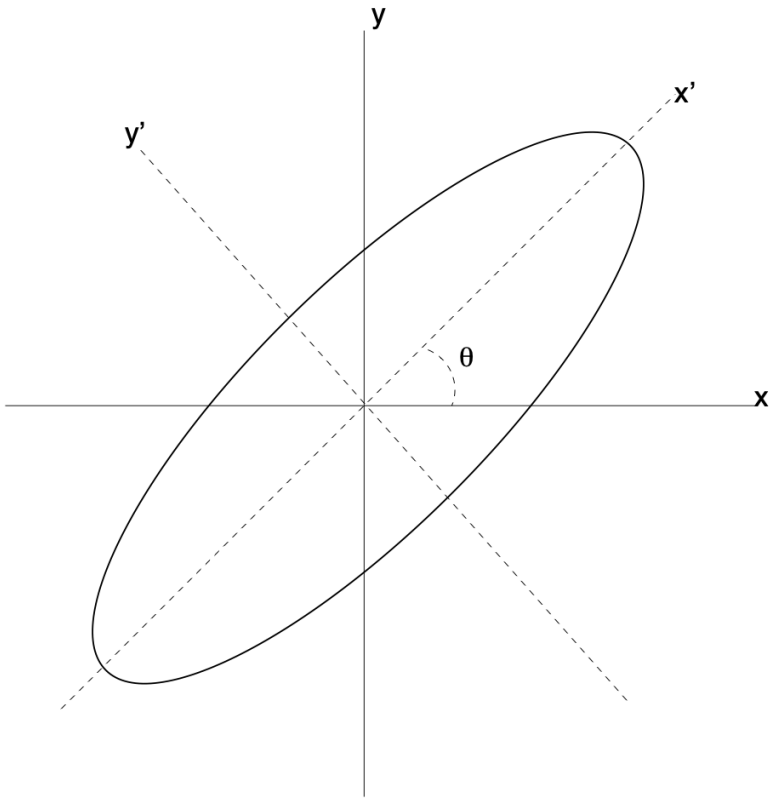
$$\left(\langle y^2\rangle - \langle x^2\rangle\right)\frac{1}{2}\sin(2\theta) + \langle xy\rangle\cos(2\theta) = 0. \tag{19}$$

Thus

$$\tan(2\theta) = \frac{2\langle xy\rangle}{\langle x^2\rangle - \langle y^2\rangle} \tag{20}$$

so

$$\theta = \frac{1}{2}\tan^{-1}\left[\frac{2\langle xy\rangle}{\langle x^2\rangle - \langle y^2\rangle}\right]. \tag{21}$$



If we want to fully define the variance, we also want to find the major and minor axes of the rotated ellipse. If we define the major axis to be a and the minor axis to be b, then one way to address this (see Waterman and Lilly, JPO, 2015) is to note that the overall energy (specifically if our measured quantities are velocity) is

$$K = \frac{1}{2}(a^2 + b^2) = \frac{1}{2}\langle x^2 + y^2 \rangle \tag{22}$$

and does not depend on the rotation. The excess kinetic energy on the major axis relative to the minor axis is

$$\begin{aligned}
L = \frac{1}{2}(a^2 - b^2) &= \frac{1}{2}\left\langle (x\cos\theta + y\sin\theta)^2 - (-x\sin\theta + y\cos\theta)^2 \right\rangle & (23)\\
&= \frac{1}{2}\left\langle (x^2 - y^2)(\cos^2\theta - \sin^2\theta) + 4xy\cos\theta\ \sin\theta \right\rangle & (24)\\
&= \frac{1}{2}\left\langle (x^2 - y^2)\cos(2\theta) + 2xy\sin(2\theta) \right\rangle & (25)\\
&= \frac{1}{2}\frac{[\langle x^2 - y^2\rangle^2 + 4\langle xy\rangle^2]}{\sqrt{\langle x^2 - y^2\rangle^2 + 4\langle xy\rangle^2}} & (26)\\
&= \frac{1}{2}\sqrt{\langle x^2 - y^2\rangle^2 + 4\langle xy\rangle^2}. & (27)
\end{aligned}$$

From this we can solve for a and b by summing and differencing:

$$\begin{aligned}
a^2 &= \frac{1}{2}\left(\langle x^2 + y^2\rangle + \sqrt{\langle x^2 - y^2\rangle^2 + 4\langle xy\rangle^2}\right) & (28)\\
b^2 &= \frac{1}{2}\left(\langle x^2 + y^2\rangle - \sqrt{\langle x^2 - y^2\rangle^2 + 4\langle xy\rangle^2}\right). & (29)
\end{aligned}$$

All of this means that with the variance and covariance information, we can define the parameters necessary to characterize the variance ellipse.

Summing random variables

We've been looking at two random variables as more or less independent quantities, but now let's think about what happens when we sum two independent variables, so that $z = x + y$ (with the corresponding deterministic variables $t = r + s$). Conceptually, you can imagine that if you draw x from a uniform distribution between 0 and 1, and y from a uniform distribution between 0 and 1, the z will be between 0 and 2. There are more ways for z to be close to 1 than to be at the extremes, so the result will have a triangle distribution. The full derivation is in the appendix.

If we add enough variables together, the distribution will approach a central limit theorem. I'll skip the details of the derivation, which is in the notes from Dan Rudnick and Russ Davis and also in the material from SIOC 221A. But let's put this the concepts of covariance and summed random variables into practice by thinking about a random walk.

The random walk is central to understanding ocean mixing, and we can think about a random walk in relation to, for example, releasing dye or floats or drifters in the ocean. In a hypothetical situation in which we knew what every particle in the ocean was doing, we could imagine trying to track all of the motions. But in real world systems, we consider particles in statistical sense. Let's think about this as a random walk:

Random walk

The random walk is central to an understanding of ocean mixing. Consider a particle that is constrained to move along a straight line in a series of steps, each of uniform interval Δt, but with random velocity v_n. Then its position x_N after N such steps is:

$$x_N = x_0 + \Delta t \sum_{n=1}^{N} v_n, \tag{30}$$

where x_0 is the initial position. If we knew v_n for all steps, then the process is deterministic. Suppose we don't have this knowledge but we do know something about the statistics of v_n. First assume that the statistics are stationary, that is, they don't depend on n. We will take the average velocity to be zero:

$$\langle v_n \rangle = 0. \tag{31}$$

With no loss in generality, we take the initial position to be zero, so the average position is

$$\langle x_N \rangle = x_0 + \Delta t \sum_{n=1}^{N} \langle v_n \rangle = x_0 = 0. \tag{32}$$

Now let's compute the variance in position.

$$\langle x_N^2 \rangle = (\Delta t)^2 \sum_{n=1}^{N} \sum_{m=1}^{N} \langle v_n v_m \rangle. \tag{33}$$

Assuming that the statistics are stationary, that they don't depend on the step, the covariance of step distances in (33) is a function of difference $n - m$:

$$\langle x_N^2 \rangle = (\Delta t)^2 \sum_{n=1}^{N} \sum_{m=1}^{N} \langle C_{vv}(n - m) \rangle. \tag{34}$$

Now we can consider some scenarios. If successive values of v_n are independent (as is the case with a random number generator), then the covariancle is the delta function:

$$C_{vv}(n-m) = \langle v^2 \rangle \delta_{nm}. \tag{35}$$

Then the variance in position becomes:

$$\langle x_N^2 \rangle = (\Delta t)^2 \langle v^2 \rangle \sum_{n=1}^{N} \sum_{m=1}^{N} \delta_{nm} = (\Delta t)^2 \langle v^2 \rangle N. \tag{36}$$

This means that the variance is proportional to the number of steps in the random walk. We could alternatively write:

$$\langle x_N^2 \rangle = (\Delta t) \langle v^2 \rangle t_N, \tag{37}$$

where $t_N = N\Delta t$ is the total elapsed time.

Since x_N is a sum, the central limit theorem applies, and its pdf approaches normal as N becomes large:

$$F_{x_N}(r) = \frac{1}{\sqrt{2\pi \langle x_N^2 \rangle}} \exp\left(\frac{-r^2}{2\langle x_N^2 \rangle} \right). \tag{38}$$

Appendix: Summing two random variables: Formalisms

Consider the sum for two random variables, $z = x + y$. The joint pdf can be written:

$$F_{xy}(r,s) = F_{xy}(r, t-r) \tag{39}$$

and the pdf of z is

$$F_z(t) = \int_{-\infty}^{\infty} F_{xy}(r, t-r)\, dr \tag{40}$$

As an example, suppose that x and y are independent and uniformly distributed between 0 and a. Then their pdfs are:

$$F_x(r) = \begin{cases} a^{-1} \text{ for } 0 \le r \le a \\ 0 \text{ elsewhere} \end{cases} \tag{41}$$

$$F_y(s) = \begin{cases} a^{-1} \text{ for } 0 \le s \le a \\ 0 \text{ elsewhere} \end{cases} \tag{42}$$

$$\tag{43}$$

So that the joint pdf is

$$F_{xy}(r,s) = \begin{cases} a^{-2} \text{ for } 0 \le r \le a, 0 \le s \le a \\ 0 \text{ elsewhere} \end{cases}. \tag{44}$$

Transforming from r, s to r, t space turns the square region of nonzero pdf to a parallelogram. Integration over the two triangular regions of the parallelogram yields:

$$F_z(t) = \begin{cases} \int_0^t a^{-2}\, dr = ta^{-2} \text{ for } 0 \le t \le a \\ \int_{t-a}^{a} a^{-2}\, dr = (2a-t)a^{-2} \text{ for } a \le t \le 2a \\ 0 \text{ elsewhere} \end{cases}. \tag{45}$$

So the distribution of the sum of two independent uniformly distributed variables is triangular with a clear mode. As the sum is over more variables, the result gets rounder and approaches a normal pdf. As the number of variables in the sum approaches infinity, the pdf approaches normal, as proven later in the Central Limit Theorem.

Statistical significance of a correlation

Once you compute a correlation, one perennial question is how to determine whether the correlation is large enough to indicate a real relationship between data. There are three ways that people think about this.

1. How big is the correlation? Sometimes investigators want to find a correlation that exceeds a fixed threshold (e.g. 0.5), maybe just because the threshold seems large enough to be useful. That might make sense intuitively, particularly since correlation reflects the fraction of variance explained. However, it's not really grounded in statistics.

2. An important consideration is whether the correlation is statistically different from zero. Matlab (corrcoef) and python (scipy.stats.pearsonr) will return a p-value, which is a measure of the likelihood that purely random data should produce a correlation as large as the observed correlation. These are often reported and can be useful measures.

3. An alternate approach is to determine a significance threshold. If the data were pure random noise, then correlation coefficients should be small but might sometimes stray to larger values. We can compute a threshold above which we'd expect to find fewer than 5% (or 2% or 1%) of correlation coefficients, in the case of random noise. *Numerical Recipes* has a nice discussion of this. The significance level is computed based on the inverse error function. In Matlab, it's:

   ```
   y=erfinv(sig_level).*sqrt(2 ./N);
   ```

   and in python:

   ```
   y=scipy.special.erfinv(sig_level)*np.sqrt(2/N)
   ```

   In either case, "sig_level" is typicaly 0.95, and N is the number of degrees of freedom—maybe the number of data points, or maybe a value that needs to be adjusted based on N_{eff} in Lecture 6.

4. Finally, sometimes it's nice to put an error bar on the correlation coefficients. Nominally this is

   $$\text{Standard error}(r) = \sqrt{\frac{1-r^2}{N-2}}, \tag{46}$$

   where r is the computed (empirical) correlation coefficient, and N is the number of (independent) data pairs.