2. Covariance and Correlation

170B Probability Theory, Puck Rombach

Last updated: October 5, 2016

Bertsekas & Tsitsiklis: Section 4.2.

Assumed knowledge: Expectation and variance.

We introduce the covariance of two random variables:

\[ \text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) = \mathbb{E}(XY) - \mathbb{E}X \cdot \mathbb{E}Y. \]

Let’s check this very carefully.

\[
\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) \\
= \mathbb{E}(XY - X\mathbb{E}(Y) - \mathbb{E}(X)Y + \mathbb{E}(X)\mathbb{E}(Y)) \\
= \mathbb{E}(XY) - \mathbb{E}(X\mathbb{E}(Y)) - \mathbb{E}(\mathbb{E}(X)Y) + \mathbb{E}(\mathbb{E}(X)\mathbb{E}(Y)) \\
= \mathbb{E}(XY) - \mathbb{E}(Y)\mathbb{E}(X) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) \\
= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\]

The important observation is that \( \mathbb{E}(X\mathbb{E}(Y)) = \mathbb{E}(Y)\mathbb{E}(X) \), because \( \mathbb{E}(Y) \) is just a constant, and \( \mathbb{E}(aX) = a\mathbb{E}(X) \).

We show the following properties of the covariance:

- \( \text{cov}(X, X) = \text{var}(X) \)
- \( \text{cov}(X, aY + b) = a \cdot \text{cov}(X, Y) \)
- \( \text{cov}(X, Y + Z) = \text{cov}(X, Y) + \text{cov}(X, Z) \)
- \( X, Y \) independent \( \Rightarrow \) \( \text{cov}(X, Y) = 0 \) (but the converse is not true).

Variance of the sum of random variables

Remember that if random variables \( X_1, \ldots, X_n \) are independent, then

\[
\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}(X_i).
\]
We generalize this to any set of random variables $X_1, \ldots, X_n$:

$$\text{var} \left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \text{var}(X_i) + \sum_{\{i,j\mid i \neq j\}} \text{cov}(X_i, X_j).$$

I suppose we can also express this as $\sum_{i=1}^{n} \sum_{j=1}^{n} \text{cov}(X_i, X_j)$, but that is maybe less useful.

**Example 2.1.** Please work through example 4.15 on page 221 in the book. I try not to read straight from the book to you in lectures, so here is an alternative and slightly more complex version of the problem. We have $2^n$ people and $2^n$ hats. Hats are either red or blue, and indistinguishable otherwise. People either own a blue or a red hat, place them all in a box, and then each pick out a hat again uniformly at random. Someone “picks their own hat” if they pick a hat of the same color as the original. Let $X_i$ be the indicator random variable that takes value 1 if person $i$ picks their own (color) hat. We are interested in the expectation and variance of $X = X_1 + \ldots + X_{2n}$, the number of people with their own color hats. We have $X_i$ is Bernoulli with probability $1/2$, so

$$\mathbb{E}(X_i) = \frac{1}{2}, \quad \text{var}(X_i) = \frac{1}{4}.$$ 

By linearity of expectation, we have

$$\mathbb{E}(X) = \mathbb{E} \left( \sum_{i=1}^{2n} X_i \right) = \frac{1}{2} \cdot 2n = n.$$ 

To find the variance, we need

$$\text{var}(X) = \text{var} \left( \sum_{i=1}^{2n} X_i \right) = \sum_{i=1}^{2n} \text{var}(X_i) + \sum_{\{i,j\mid i \neq j\}} \text{cov}(X_i, X_j).$$

We split the pairs $i \neq j$ into two cases: either person $i$ and $j$ started out with the same color, or they started out with different colors.

- **Same color.** We have that if $X_i = 1$, then person $i$ has their own color hat. This means that for person $j$, there are $2n - 1$ hats left in the box, with $n - 1$ of those the right color. So,

  $$\mathbb{E}(X_iX_j) = \mathbb{P}(X_i = 1, X_j = 1) = \mathbb{P}(X_i = 1)\mathbb{P}(X_j = 1|X_i = 1) = \frac{1}{2} \cdot \frac{n - 1}{2n - 1}.$$ 

  There are $n(n - 1)$ ordered pairs $(i, j)$ for each color, giving $2n(n - 1)$ ordered pairs in total.

- **Different color.** We have that if $X_i = 1$, then person $i$ has their own color hat. This means that for person $j$, there are $2n - 1$ hats left in the box, with $n$ of those the right color. So,

  $$\mathbb{E}(X_iX_j) = \mathbb{P}(X_i = 1, X_j = 1) = \mathbb{P}(X_i = 1)\mathbb{P}(X_j = 1|X_i = 1) = \frac{1}{2} \cdot \frac{n}{2n - 1}.$$ 

  There are $n^2$ ordered pairs $(i, j)$ where $i$ is red and $j$ blue, so there are $2n^2$ ordered pairs in total.
Putting this all together, we get
\[ \text{var}(X) = 2n \cdot \frac{1}{4} + 2n(n - 1) \cdot \left( \frac{1}{2} \cdot \frac{n - 1}{2n - 1} - \frac{1}{4} \right) + 2n^2 \cdot \left( \frac{1}{2} \cdot \frac{n}{2n - 1} - \frac{1}{4} \right) = \frac{n^2}{2n - 1}. \]

We define the **correlation coefficient** between two variables to be

\[ \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}, \]

The correlation coefficient gives us a normalized measure of how much two random variables co-vary. This is shown by the following lemmas.

**Lemma 2.2.** *Schwarz Inequality for expectations:*

\[ \mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2). \]

**Proof.** This is shown in the book on page 250, but here is a slightly different way of doing it. We have

\[ \mathbb{E}((aX \pm bY)^2) \geq 0, \quad \forall a, b \in \mathbb{R}, \]

because a square is never negative. Then,

\[ \mathbb{E}((aX + bY)^2) = a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(XY) \geq 0. \]

Here’s the trick: we said this was true for any real \( a, b \), so we may as well set \( a^2 = \mathbb{E}(Y^2) \) and \( b^2 = \mathbb{E}(X^2) \). This yields:

\[ -2a^2b^2 \leq 2ab\mathbb{E}(XY) \leq 2a^2b^2. \]

Divide by \( 2ab \) to get our desired result:

\[ -\sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)} \leq \mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}. \]

\[ \square \]

**Lemma 2.3.** *The correlation coefficient is normalized in the sense that*

\[ -1 \leq \rho(X, Y) \leq 1, \]

*and \( \rho = 1 \ (or \ -1) \ if \ there \ is \ a \ positive \ (or \ negative) \ constant \ c \ such \ that \ Y - \mathbb{E}Y = c(X - \mathbb{E}X).)*

**Proof.** The result \(-1 \leq \rho(X, Y) \leq 1\) follows immediately from lemma 2.2. If \( Y - \mathbb{E}Y = c(X - \mathbb{E}X) \), then \( \rho(X, Y) = \frac{c}{|c|} = \pm 1 \). (The converse is proven on page 251 in the book.) \( \square \)

As we saw in the last lecture, the covariance of two random variables has some nice linear properties, such as \( \text{cov}(X, aY + b) = a \cdot \text{cov}(X, Y) \) and \( \text{cov}(X, Y + Z) = \text{cov}(X, Y) + \text{cov}(X, Z) \). We can write these properties very generally, as follows. (The lemma follows from the two properties just mentioned.)
Lemma 2.4. If we can write two random variables as linear combinations of other random variables, i.e.

\[ X = \sum_i a_i X_i, \quad Y = \sum_j b_j Y_j \]

then we can write the covariance as

\[
\text{cov}(X, Y) = \text{cov} \left( \sum_i a_i X_i, \sum_j b_j Y_j \right) \\
= a_1 b_1 \cdot \text{cov}(X_1, Y_1) + a_2 b_1 \cdot \text{cov}(X_2, Y_1) + \ldots \\
= \sum_i \sum_j a_i b_j \cdot \text{cov}(X_i, Y_j).
\]

This is a very powerful tool, as shown in the following example.

Example 2.5. Please read Example 4.14 on page 220. We consider a generalization of this example. I’ll describe two settings producing the same distribution of two random variables \( X \) and \( Y \).

- Roll a fair \( n \)-sided die, which produces values \( 1, \ldots, n \) uniformly at random, (a coin if \( n = 2 \)) \( k \) times. Let \( X \) be the number of outcomes 1, and \( Y \) the number of outcomes 2.

- Place \( k \) balls in \( n \) buckets, by placing each ball in one of the buckets uniformly at random, independently of the other balls. Let \( X \) be the number of balls in bucket 1, and \( Y \) the number in bucket 2.

What is \( \rho(X, Y) \)?

If you try to find \( \mathbb{E}(XY) \) directly you will get stuck pretty quickly. Evaluating this would involve a double sum over all the pairs of integers with a given product, which is pretty awful. However, we can use indicator variables again, and express \( X \) and \( Y \) as sums. This way, we can “filter out” the dependence between \( X \) and \( Y \), which get more and more independent as \( n \) grows. Let \( X_i \) be the indicator random variable that ball \( i \) is in bucket 1, and \( Y_j \) the indicator random variable that ball \( j \) is in bucket 2. Now

\[ X = \sum_i X_i, \quad Y = \sum_j Y_j \]

and we can use lemma 2.4, giving:

\[
\text{cov}(X, Y) = \sum_i \sum_j \text{cov}(X_i, Y_j).
\]

We have that \( X_i \) and \( Y_i \) are all Bernoulli with probability \( 1/n \):

\[
\mathbb{E}(X_i) = \mathbb{E}(Y_j) = \frac{1}{n}, \quad \text{var}(X_i) = \text{var}(Y_j) = \frac{n - 1}{n}.
\]

If \( i \neq j \), then \( X_i \) and \( Y_j \) are independent (why?). If \( i = j \), then

\[
\text{cov}(X_i, Y_j) = \mathbb{E}(X_i Y_j) - \mathbb{E}(X_i)\mathbb{E}(Y_j) = 0 - \frac{1}{n^2} = -\frac{1}{n^2},
\]

\[ 170B \text{ Probability Theory, Rombach} \]

Lecture Notes
because $P(X_i = 1, Y_j = 1) = 0$. Therefore,

$$\text{cov}(X, Y) = \sum_i \sum_j \text{cov}(X_i, Y_j) = \sum_{(i,j) \neq j} \text{cov}(X_i, Y_j) + \sum_i \text{cov}(X_i, Y_i) = 0 + \sum_{i=1}^{k} \left( -\frac{1}{n^2} \right) = -\frac{k}{n^2}.$$ 

Since $X$ and $Y$ are binomial with parameters $k$ and $1/n$, we have

$$\text{var}(X) = \text{var}(Y) = \frac{k}{n} \left( 1 - \frac{1}{n} \right) = \frac{k(n - 1)}{n^2}.$$ 

Finally, we find

$$\rho(X, Y) = \frac{-\frac{k}{n^2}}{\frac{k(n-1)}{n^2}} = -\frac{1}{n - 1}.$$ 

Note how this does indeed yield $\rho = -1$ when $n = 2$ (ex. 4.14 in the book).