Week 9 Lecture: Covariance and Correlation

Yilin

DigiPen

COVARIANCE

Covariance

Goal

Develop a meaningful way to quantify the relationship between random variables X and Y.

Definition

We say an random variable. X is **centered** if $\overline{X} = 0$. If X is any random variable., then we center it by making a new random variable $X - \overline{X}$.

Idea

Let X and Y be centered r.v.s. Measure the expectation of the random variable XY.

- If both X and Y are (jointly) often “large” then $E(XY)$ should be ”large”.
- If X and Y (jointly) often have the same sign (both positive or negative) then $E(XY)$ should be positive.
- If X and Y (jointly) often have opposite signs then $E(XY)$ should be negative.
- If X and Y are often different with not much of a joint relation, then $E(XY)$ should be small.

Example: Relation of X and Y

Consider the joint distribution given by

Joint Probability	$Y = 1$	$Y = 2$	$Y = 3$	X Marginal
$X = 0$	0.02	0.08	0.01	0.11
$X = 1$	0.03	0.21	0.04	0.28
$X = 2$	0.16	0.11	0.34	0.61
Y Marginal	0.21	0.40	0.39	

1. Find the distribution of the random variable XY. (Wont use this for a few more slides)
2. Center X and Y.
3. Find the distribution of the random variable $(X - \overline{X})(Y - \overline{Y})$.
4. Compute $E\left((X - \overline{X})(Y - \overline{Y})\right)$.

1. Distribution of XY

- XY takes values in products of 0,1,2 with 1,2,3: $\{0, 1, 2, 3, 4, 6\}$.
- Use joint probabilities of X and Y to compute each probability.

$$
\begin{aligned}
P(XY = 0) &= P(X = 0) && = 0.11 \\
P(XY = 1) &= P(X = 1, Y = 1) && = 0.03 \\
P(XY = 2) &= P(X = 1, Y = 2) + \Pr(X = 2, Y = 1) && = 0.37 \\
P(XY = 3) &= P(X = 1, Y = 3) && = 0.04 \\
P(XY = 4) &= P(X = 2, Y = 2) && = 0.11 \\
P(XY = 6) &= P(X = 2, Y = 3) && = 0.34
\end{aligned}
$$

2. Centering X and Y

Compute

$$\overline{X} = 0(0.11) + 1(0.28) + 2(0.61) = 1.5,$$

and

$$\overline{Y} = 1(0.21) + 2(0.4) + 3(0.39) = 2.18.$$

- The r.v. $X - \overline{X}$ takes values in $\{-1.5, -0.5, 0.5\}$
- $Y - \overline{Y}$ takes values in $\{-1.18, -0.18, 0.82\}$.

Joint Probability	$Y - \overline{Y} = -1.18$	$Y - \overline{Y} = -0.18$	$Y - \overline{Y} = 0.82$
$X - \overline{X} = -1.5$	0.02	0.08	0.01
$X - \overline{X} = -0.5$	0.03	0.21	0.04
$X - \overline{X} = 0.5$	0.16	0.11	0.34

3. Distribution of $(X - \overline{X})(Y - \overline{Y})$

The r.v. $(X - \overline{X})(Y - \overline{Y})$ takes values in products of numbers in $\{-1.5, -0.5, 0.5\}$ and $\{-1.18, -0.18, 0.82\}$, that is

$$
\begin{aligned}
P((X - \overline{X})(Y - \overline{Y}) &= (-1.5)(-1.18) = 1.77) && = 0.02 \\
P((X - \overline{X})(Y - \overline{Y}) &= (-1.5)(-0.18) = 0.27) && = 0.08 \\
P((X - \overline{X})(Y - \overline{Y}) &= (-1.5)(0.82) = -1.23) && = 0.01 \\
P((X - \overline{X})(Y - \overline{Y}) &= (-0.5)(-1.18) = 0.59) && = 0.03 \\
P((X - \overline{X})(Y - \overline{Y}) &= (-0.5)(-0.18) = 0.09) && = 0.21 \\
P((X - \overline{X})(Y - \overline{Y}) &= (-0.5)(0.82) = -0.41) && = 0.04 \\
P((X - \overline{X})(Y - \overline{Y}) &= (0.5)(-1.18) = -0.59) && = 0.16 \\
P((X - \overline{X})(Y - \overline{Y}) &= (0.5)(-0.18) = -0.09) && = 0.11 \\
P((X - \overline{X})(Y - \overline{Y}) &= (0.5)(0.82) = 0.41) && = 0.34
\end{aligned}
$$

4. Expectation of $(X-\overline{X})(Y-\overline{Y})$

Now that we know the distribution of $(X-\overline{X})(Y-\overline{Y})$, we can compute its mean.

$$
\begin{aligned}
E\Big((X-\overline{X})(Y-\overline{Y})\Big) &= (1.77)(0.02) + (0.27)(0.08) \\
&\quad + (-1.23)(0.01) + (0.59)(0.03) \\
&\quad + (0.09)(0.21) + (-0.41)(0.04) \\
&\quad + (-0.59)(0.16) + (-0.09)(0.11) \\
&\quad + (0.41)(0.34) \\
&= 0.1
\end{aligned}
$$

This is called the **covariance** of X and Y.

Definition

Definition

The covariance of r.v.s X and Y is the expectation of the product of the centered versions of X and Y:

$$\mathrm{Cov}(X, Y) = E\left((X - \overline{X})(Y - \overline{Y})\right)$$

- If $\mathrm{Cov}(X, Y) > 0$ we say X and Y are **positively correlated**.
- If $\mathrm{Cov}(X, Y) < 0$ we say X and Y are **negatively correlated**.
- If $\mathrm{Cov}(X, Y) = 0$ we say X and Y are **uncorrelated** (This is different from independence).

Some Intuition

If X and Y are similar,

- Note:
  $\mathrm{Cov}(X, X) = E\left((X - \overline{X})(X - \overline{X})\right) = E\left((X - \overline{X})^2\right) = \mathrm{var}(X)$.
- If X is similar to Y then $\mathrm{Cov}(X, Y) = E\left((X - \overline{X})(Y - \overline{Y})\right)$ is close to $\mathrm{var}(X)$ and $\mathrm{var}(Y)$.

If X and Y are not similar,

- $(X - \overline{X})(Y - \overline{Y})$ should not be "much more" likely to be positive than negative, i.e. the terms of $\mathrm{Cov}(X, Y)$ will likely balance each other out.
- Thus the sum $\mathrm{Cov}(X, Y)$ will be smaller if X and Y are different.

Previous example: $\mathrm{Cov}(X, Y) = 0.1$ so X and Y are related but not by much.

Example 2: Covariance of two biased coins.

Consider X and Y distributed via

	$Y=0$	$Y=1$	P_X
$X=0$	0.12	0.28	
$X=1$	0.18	0.42	
P_Y			

To compute $\text{Cov}(X, Y)$, we need to find

$$\overline{X} = 0(0.4) + 1(0.6) = 0.6, \qquad \overline{Y} = 0(0.3) + 1(0.7) = 0.7.$$

Note: X and Y are independent. (Why?)

Example 2 Cont'd

Rewrite the joint probabilities for the centered r.v.s:

Joint Probability	$Y - \overline{Y} = -0.7$	$Y - \overline{Y} = 0.3$	$P_{X-\overline{X}}$
$X - \overline{X} = -0.6$	0.12	0.28	0.4
$X - \overline{X} = 0.4$	0.18	0.42	0.6
$P_{Y-\overline{Y}}$	0.3	0.7	

Note: $X - \overline{X}$ and $Y - \overline{Y}$ are independent.

$(X - \overline{X})(Y - \overline{Y})$ takes values in

$(-0.7)(-0.6) = 0.42$	$(0.3)(-0.6) = -0.18$
$(-0.7)(0.4) = -0.28$	$(0.3)(0.4) = 0.12$

Example 2: Covariance

Thus

$$\Pr\Big((X-\overline{X})(Y-\overline{Y})=0.42\Big)=0.12$$
$$\Pr\Big((X-\overline{X})(Y-\overline{Y})=-.18\Big)=0.28$$
$$\Pr\Big((X-\overline{X})(Y-\overline{Y})=-.28\Big)=0.18$$
$$\Pr\Big((X-\overline{X})(Y-\overline{Y})=0.12\Big)=0.42,$$

and

$$\begin{aligned}\mathrm{Cov}(X,Y)&=E\Big((X-\overline{X})(Y-\overline{Y})\Big)\\&=0.42(0.12)-0.18(0.28)-0.28(0.18)+0.12(0.42)\\&=0\end{aligned}$$

Question: Does the independence of X and Y play a role?

Computing Covariance

We can compute the covariance in a simpler fashion:

Proposition

$Cov(X, Y) = E(XY) - E(X)E(Y) = E(XY) - \overline{X} \cdot \overline{Y}.$

Proof.

$$\begin{aligned}
E\big((X - \overline{X})(Y - \overline{Y})\big) &= E(XY - \overline{X}Y - \overline{Y}X + \overline{X} \cdot \overline{Y}) && (1)\\
&= E(XY) + E(-\overline{X}Y) + E(-\overline{Y}X) + E(\overline{X} \cdot \overline{Y}) && (2)\\
&= E(XY) - \overline{X}E(Y) - \overline{Y}E(X) + \overline{X} \cdot \overline{Y} && (3)\\
&= E(XY) - \overline{X} \cdot \overline{Y} - \overline{X} \cdot \overline{Y} + \overline{X} \cdot \overline{Y} && (4)\\
&= E(XY) - \overline{X} \cdot \overline{Y} && (5)
\end{aligned}$$

$\square$

Covariance of Independent Random Variables

Theorem

If X and Y are independent, then $E(XY) = E(X)E(Y)$, so $Cov(X, Y) = 0$.

Proof.

$$
\begin{aligned}
\overbrace{\sum_i \sum_j x_i y_j \Pr(X = x_i, Y = y_j)}^{E(XY)} &= \sum_i \sum_j x_i y_j \overbrace{\Pr(X = x_i) \Pr(Y = y_j)}^{\text{Independence}} \\
&= \sum_i x_i \Pr(X = x_i) \sum_j y_j \Pr(Y = y_j) \\
&= \left(\sum_i x_i \Pr(X = x_i) \right) \left(\sum_j y_j \Pr(Y = y_j) \right) \\
&= E(X)E(Y)
\end{aligned}
$$

Example

A fair die is independently rolled two times. X = number on the first roll, Y = number on the second roll. Find Cov(X, Y).

Solution.

- X and Y come from two independent experiments $\Rightarrow$ independent.
- Cov$(X, Y) = E(XY) - E(X)E(Y) = 0$ by the theorem.

Concrete Formula for Covariance

- $\mu_X = E(X), \mu_Y = E(Y) \Rightarrow \text{Cov}(X, Y) = E(XY) - E(X)E(Y)$.
- If X and Y are discrete with joint PMF $p(x, y)$, then

$$\text{Cov}(X, Y) = \sum_{x,y} xyp(x, y) - E(X)E(Y)$$

- If X and Y are continuous with joint PDF $f(x, y)$, then

$$\text{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dydx - E(X)E(Y)$$

Take-away

- If X and Y are positively correlated, then they are "similar" in some sense.
- If X and Y are negatively correlated then they are "opposite" in some sense.
- If X and Y are uncorrelated then they are... uncorrelated. Not related to each other in joint distribution (but still possibly dependent!)
- The covariance is a way to measure the closeness of X and Y by measuring when $X - \overline{X}$ and $Y - \overline{Y}$ are close or far away.

Example 3: Dependent but Uncorrelated r.v.s

- Let X take values in $-1, 0, 1$ with probabilities $0.25, 0.5, 0.25$.
- Let $Y = X^2$. Then Y takes values in 0 and 1 each with probability 0.5.
- It is clear that X and Y are dependent.
- We see $E(X) = 0$ and $E(Y) = E(X^2) = 0.5$.
- $XY = X^3$ takes values in $-1, 0, 1$ each with respective probability $0.25, 0.5, 0.25$.

Then

$$E(XY) = -1(0.25) + 0(0.5) + 1(0.25) = 0,$$

and

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0 - 0(0.5) = 0.$$

So X and X^2 are uncorrelated but are most certainly **dependent**.

Example 4

Flip a fair coin 3 times. $X = \#$ heads in the first 2 flips, $Y = \#$ heads on the last 2 flips.

(a) Find the joint PMF $p(x, y)$ of X and Y

(b) Compute $\text{Cov}(X, Y)$.

Solution.

- Sample space

$$\Omega = \{\text{HHH,HHT,HTH,THH,HTT,THT,TTH,TTT}\}$$

- Possible values of X, Y are $0, 1, 2$.
- Need to compute

$$p(x, y) = P(X = x, Y = 0) \text{ for } x, y \in \{0, 1, 2\}.$$

Example 4: Joint PMF of X and Y

$$p(x, y) = P(X = x, Y = y)$$

- $p(0, 0) \Rightarrow$ all tosses are tails $\Rightarrow p(0, 0) = P(\text{TTT}) = 1/8$.
- $p(1, 1) \Rightarrow$ 1 head in the first two tosses and 1 head in the last two tosses $\Rightarrow p(1, 1) = P(\text{THT,HTH}) = 2/8$.
- The rest of values are computed similarly.

$X \setminus Y$	0	1	2	$p(x)$
0	$1/8$	$1/8$	0	$1/4$
1	$1/8$	$2/8$	$1/8$	$1/2$
2	0	$1/8$	$1/8$	$1/4$
$p(y)$	$1/4$	$1/2$	$1/4$	1

Example 4: $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$

- $E(X) = \sum_{x=0}^{2} xp(x) = 1 \cdot 1/2 + 2 \cdot 1/4 = 1.$
- $E(Y) = \sum_{y=0}^{2} yp(y) = 1 \cdot 1/2 + 2 \cdot 1/4 = 1.$
- $E(XY) = \sum_{x,y=0}^{2} xyp(x, y) = 1 \cdot \frac{2}{8} + 2 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + 4 \cdot \frac{1}{8} = \frac{5}{4}$
- $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{5}{4} - 1 = \frac{1}{4}.$

Covariance and variance

Lemma 2. Let X be a random variable. Then

$$\mathrm{var}(X) = \mathrm{Cov}(X, X)$$

Proof. By Theorem 1,

$$\mathrm{Cov}(X, X) = E(X^2) - [E(X)]^2 = \mathrm{var}(X).$$

Properties of covariance

Theorem 2. Let X, Y, Z be jointly distributed random variables. Let a, b, c, d be real numbers. Then the following hold.

(a) $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$

(b) $\mathrm{Cov}(a, X) = 0$.

(c) $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$.

Properties of covariance (continued)

(d) $\mathrm{Cov}(a + bX, Y) = b\mathrm{Cov}(X, Y)$.

(e) $\mathrm{Cov}(a + bX, c + dY) = bd\mathrm{Cov}(X, Y)$.

Correlation

Correlation

- Covariance has the potential to be very large if $E(XY)$ is large (negative or positive).
- We can scale $\mathrm{Cov}(X, Y)$ to give us a number in the range of $[-1, 1]$ that gives us a sort of "score" of how correlated X and Y are.
- This is called the **correlation**.

Definition

Definition

Let X and Y be r.v.s with standard deviation σ_X and σ_Y respectively. The **correlation** of X and Y is called $\rho_{X,Y}$ (Greek letter 'rho') and is computed

$$\rho_{X,Y} := \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- If $\rho_{X,Y}$ is close to 1 we say there is a **strong positive correlation** between X and Y, etc.
- If $\rho_{X,Y} = 1$ we say that X and Y have perfect correlation.
- If $\rho_{X,Y} = -1$ we say that X and Y have perfect negative correlation.
- If $\rho_{X,Y} = 0$ then $\text{Cov}(X, Y) = 0$, which means that X and Y are uncorrelated.

Why is $-1 \leq \rho_{X,Y} \leq 1$?

Theorem

For r.v.s X and Y, we have

$$-1 \leq \rho_{X,Y} \leq 1.$$

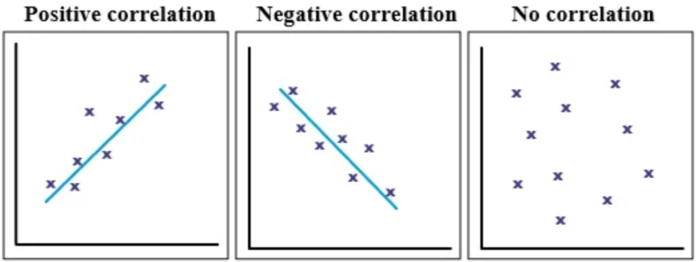
Proof Outline.

This follows from the **Cauchy-Schwarz Inequality** (do not need to know any of this). Assume X and Y are centered, then $[E(XY)]^2 \leq E(X^2)E(Y^2)$. Dividing and taking a square root, we get

$$-1 \leq \frac{E(XY)}{\sqrt{E(X^2)E(Y^2)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \leq 1$$

$\square$

What correlation says?



Correlation in Example 1

Let X and Y be as in example 1:

- $\overline{X} = 1.5$
- $\overline{Y} = 2.18$
- $P(X = 0) = 0.11, \quad P(X = 1) = 0.28, \quad P(X = 2) = 0.61$
- $P(Y = 1) = 0.21, \quad P(Y = 2) = 0.40, \quad P(Y = 3) = 0.39$

Then

$$E(X^2) = 0(.11) + 1(.28) + 4(.61) = 2.72,$$

and

$$E(Y^2) = 1(0.21) + 4(0.4) + 9(0.39) = 5.32.$$

Thus the variances are

$$\sigma_X^2 = E(X^2) - (\overline{X})^2 = 2.72 - (1.5)^2 = 0.47,$$

and

$$\sigma_Y^2 = E(Y^2) - \overline{Y}^2 = 5.32 - 2.18^2 = 0.5676.$$

Example 1 Correlation Cont'd

Then

- Recall $\mathrm{Cov}(X, Y) = 0.1$
- $\sigma_X = \sqrt{0.47}$
- $\sigma_Y = \sqrt{0.5676}$

Thus

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{0.1}{\sqrt{0.47}\sqrt{0.5676}} \approx \frac{0.1}{0.5165} \approx 0.1936$$

Example 4: Correlation measures linearity

Let X be an r.v. with distribution

$$P(X = 0) = 0.5, \quad P(X = 1) = 0.4, \quad P(X = 2) = 0.1.$$

Define $Y := 2X + 1$. Then

$$P(Y = 1) = 0.5, \quad P(Y = 3) = 0.4, \quad P(Y = 5) = 0.1.$$

The relation between X and Y is **linear**.

Question

What is the joint distribution of X and Y?

Joint Probability	$Y = 1$	$Y = 3$	$Y = 5$
$X = 0$	0.5	0	0
$X = 1$	0	0.4	0
$X = 2$	0	0	0.1

Example 4: Covariance

To find $\mathrm{Cov}(X, Y)$ we see that XY lies in $\{0, 3, 10\}$.

$$P(XY = 0) = 0.5, \qquad P(XY = 3) = 0.4, \qquad P(XY = 10) = 0.1,$$

and

$$E(XY) = 3(0.4) + 10(0.1) = 2.2.$$

The means are

$$E(X) = 1(0.4) + 2(0.1) = 0.6,$$

and

$$E(Y) = 1(0.5) + 3(0.4) + 5(0.1) = 2.2 = 2(0.6) + 1.$$

So

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 2.2 - (0.6)(2.2) = (0.4)(2.2) = 0.88.$$

Example 4: Correlation

The variances can be computed by first finding

$$E(X^2) = 1(0.4) + 4(0.1) = 0.8,$$

and

$$E(Y^2) = 1(0.5) + 9(0.4) + 25(0.1) = 6.6.$$

Thus

$$\sigma_X^2 = 0.8 - (0.6)^2 = 0.44, \qquad \sigma_Y^2 = 6.6 - 2.2^2 = 1.76.$$

Finally the correlation is

$$\rho_{X,Y} = \frac{0.88}{\sqrt{0.44}\sqrt{1.76}} = \frac{0.88}{0.88} = 1.$$

Take-away: Linearly related r.v.s, $Y = aX + b$ are perfectly correlated.

Theorem

Theorem

Let $Y = aX + b$. If $a > 0$, then X and Y are perfectly correlated and if $a < 0$ then X and Y are perfectly negatively correlated.

Proof.

If X is centered then $E(Y) = E(aX + b) = b$

Theorem

Theorem

Let $Y = aX + b$. If $a > 0$, then X and Y are perfectly correlated and if $a < 0$ then X and Y are perfectly negatively correlated.

Proof.

If X is centered then $E(Y) = E(aX + b) = b$so $Y - \overline{Y} = aX$. Thus

$$\mathrm{Cov}(X, Y) = E(X(aX)) - E(X)(E(aX)) = aE(X^2) = a\sigma^2.$$

Also if $\mathrm{var}(X) = \sigma^2$, then $\mathrm{var}(Y) = \mathrm{var}(aX) = E(a^2X^2) = a^2\sigma^2$. Thus

$$\rho_{X,Y} = \frac{a\sigma^2}{\sigma \cdot \pm a\sigma} = \begin{cases} 1 & : a > 0 \\ -1 & : a < 0 \end{cases}.$$

$\square$

Example

Roll a coin three times.
X = # heads in the first roll, Y total # heads.
Find the correlation of X and Y.

$X \setminus Y$	0	1	2	3	$p_X(x)$
0	1/8	2/8	1/8	0	4/8
1	0	1/8	2/8	1/8	4/8
$p_Y(y)$	1/8	3/8	3/8	1/8	**1**

Solution

Solution