Further Topics on Random Variables: Covariance and Correlation

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

Reference:

- D. P. Bertsekas, J. N. Tsitsiklis, *Introduction to Probability* , Sections4.2

Covariance (1/3)

- The covariance of two random variables X and Y is defined by

$$\operatorname{cov}(X,Y) = \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right]$$

  – An alternative formula is

$$\operatorname{cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$$

- A positive or negative covariance indicates that the values of $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ tends to have the same or opposite sign, respectively
- A few other properties

$$\operatorname{cov}(X,X) = \operatorname{var}(X)$$

$$\operatorname{cov}(X, aY + b) = a \operatorname{cov}(X,Y)$$

$$\operatorname{cov}(X, Y + Z) = \operatorname{cov}(X,Y) + \operatorname{cov}(X,Z)$$

Covariance (2/3)

- Note that if X and Y are independent

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$$

  - Therefore

$$\operatorname{cov}(X, Y) = 0$$

- Thus, if X and Y are independent, they are also uncorrelated
  - However, the converse is generally not true! (See Example 4.13)

Covariance (3/3)

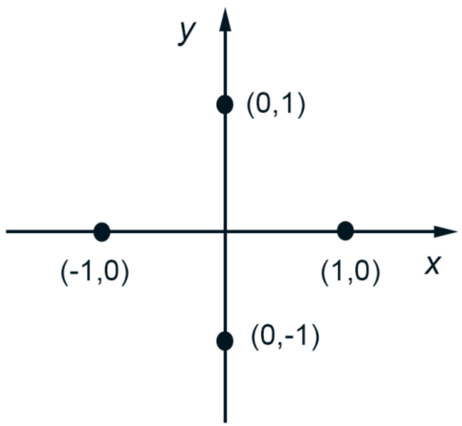
- **Example 4.13.** The pair of random variables (*X, Y*) takes the values (1, 0), (0, 1), (−1, 0), and (0,−1), each with probability ¼ Thus, the marginal pmfs of *X* and *Y* are symmetric around 0, and **E**[*X*] = **E**[*Y*] = 0
  - Furthermore, for all possible value pairs (*x, y*), either *x* or *y* is equal to 0, which implies that *XY* = 0 and **E**[*XY*] = 0*.* Therefore, cov(*X, Y*) = **E**[(*X* − **E**[*X*])(*Y* − **E**[*Y*])] = **E**[*XY*] = 0, and *X* and *Y* are uncorrelated
  - However, X and Y are not independent since, for example, a nonzero value of X fixes the value of Y to zero

$$P(X=0)=\frac{1}{2}$$

$$P(X=1)=P(X=-1)=\frac{1}{4}$$

$$P(Y=0)=\frac{1}{2}$$

$$P(Y=1)=P(Y=-1)=\frac{1}{4}$$



For example :

$$P(X=1,Y=1)=\frac{1}{4}$$

$$\neq P(X=1)P(Y=1)=\frac{1}{16}$$

Correlation (1/3)

- Also denoted as "Correlation Coefficient"
- The correlation coefficient of two random variables X and Y is defined as

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$$

  – It can be shown that (see the end-of-chapter problems)

$$-1 \leq \rho \leq 1$$

Note that
the sign of ρ only depends on $\operatorname{cov}(X,Y)$

  - $\rho > 0$: positively correlated
  - $\rho < 0$: negatively correlated
  - $\rho = 0$: uncorrelated $(\Rightarrow \operatorname{cov}(X,Y) = 0)$

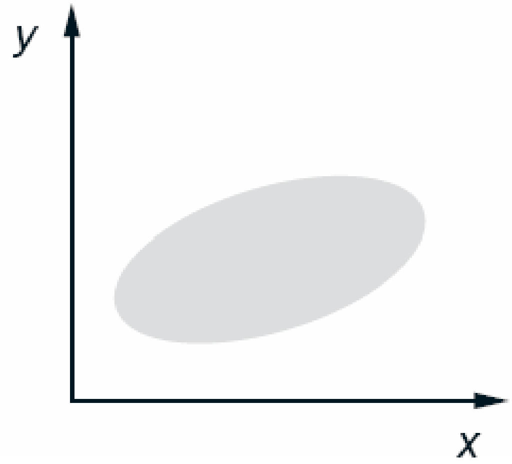
Correlation (2/3)

- It can be shown that $\rho = 1 \left(\text{or } \rho = -1\right)$ if and only if there exists a positive (or negative, respectively) constant c such that

$$Y - \mathbf{E}[Y] = c\left(X - \mathbf{E}[X]\right)$$

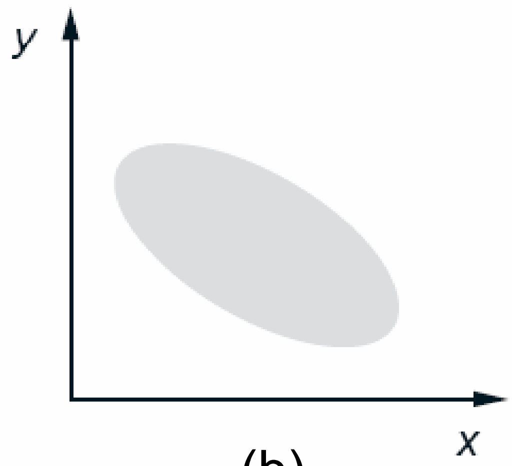
Correlation (3/3)

- **Figure 4.11:** Examples of positively (a) and negatively (b) correlated random variables



(a)

$$\operatorname{cov}(X, Y) > 0$$



(b)

$$\operatorname{cov}(X, Y) < 0$$

An Example

- Consider n independent tosses of a coin with probability of a head to p. Let X and Y be the numbers of heads and tails, respectively, and let us look at the correlation coefficient of X and Y.

$$X + Y = n$$

$$\Rightarrow \mathbf{E}[X] + \mathbf{E}[Y] = n$$

$$\Rightarrow X - \mathbf{E}[X] = -(Y - \mathbf{E}[Y])$$

$$\begin{aligned}
\operatorname{cov}(X, Y) &= \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] \\
&= -\mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \\
&= \text{-var}(X)
\end{aligned}$$

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} = \frac{\text{- var}(X)}{\sqrt{\operatorname{var}(X)\operatorname{var}(X)}} = -1$$

Variance of the Sum of Random Variables

- If $X_1, X_2, \ldots, X_n$ are random variables with finite variance, we have

$$\operatorname{var}(X_1 + X_2) = \operatorname{var}(X_1) + \operatorname{var}(X_2) + 2\operatorname{cov}(X_1, X_2)$$

  – More generally,

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

    - See the textbook for the proof of the above formula and see also Example 4.15 for the illustration of this formula

An Example

- Example 4.15. Consider the hat problem discussed in Section 2.5, where n people throw their hats in a box and then pick a hat at random. Let us find the variance of X , the number of people who pick their own hat.

$$X = X_1 + X_2 + \cdots + X_n$$

(Note that all X_i are Bernoulli with parameter $p = \mathbf{P}(X_i = 1) = \frac{1}{n}$;

X_i are not independen t of each other!)

$$\mathbf{E}[X_i] = \frac{1}{n}; \operatorname{var}(X_i) = \frac{1}{n}\left(1 - \frac{1}{n}\right)$$

?

For $i \neq j$, we have

$$\operatorname{cov}(X_i, X_j) = \boxed{\mathbf{E}[X_i X_j]} - \mathbf{E}[X_i]\mathbf{E}[X_j] = \boxed{\mathbf{P}(X_i = 1 \text{ and } X_j = 1)} - \mathbf{E}[X_i]\mathbf{E}[X_j]$$

$$= \mathbf{P}(X_i = 1)\mathbf{P}(X_j = 1 \mid X_i = 1) - \frac{1}{n^2} = \frac{1}{n} \cdot \frac{1}{n-1} - \frac{1}{n^2} = \frac{1}{n^2(n-1)}$$

Therefore,

$$\operatorname{var}(X) = \operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{var}(X_i) - \sum_{\{(i,j) \mid i \neq j\}} \operatorname{cov}(X_i, X_j)$$

$$= n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right) + n(n-1)\frac{1}{n^2(n-1)} = 1$$