Introduction to Probability

Lecture 7: Independence, Covariance and Correlation

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Independence of Random Variables

This definition covers the discrete and continuous case!

Definition of Independence

Two random variables X and Y are independent if for all values a, b:

$$\mathbf{P}[X \leq a, Y \leq b] = \mathbf{P}[X \leq a] \cdot \mathbf{P}[Y \leq b].$$

For two discrete random variables, an equivalent definition is:

$$\mathbf{P}[X = a, Y = b] = \mathbf{P}[X = a] \cdot \mathbf{P}[Y = b].$$

This is useless for continuous random variables.

Remark

Using the joint probability distribution, the above is equivalent to for all a, b,

$$F(a, b) = F_X(a) \cdot F_Y(b).$$

All these definitions extend in the natural way to more than two variables!

Factorisation

Factorisation

The definition of independence of X and Y implies the following factorisation formula: for any "suitable" sets A and B,

$$\mathbf{P}[X \in A, Y \in B] = \mathbf{P}[X \in A] \cdot \mathbf{P}[Y \in B]$$

For continuous distributions one obtains by differentiating both sides in the formula for the joint distribution:

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$

Example

Let X and Y be two independent variables. Let $I = (a,b]$ be any interval and define $U := \mathbf{1}_{X \in I}$ and $V := \mathbf{1}_{Y \in I}$. Prove U and V are independent.

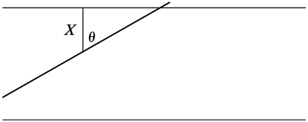
Answer

$$\begin{aligned}\mathbf{P}[U = 0, V = 1] &= \mathbf{P}\left[X \in I^c, Y \in I\right] \\ &= \mathbf{P}[X \notin I]\,\mathbf{P}[Y \in I] = \mathbf{P}[U = 0]\,\mathbf{P}[V = 1].\end{aligned}$$

Verification for other combination of values is similar ($U = 1$, $V = 0$ and $U = 0$, $V = 0$, etc).

Buffon's Needle Problem (1/2)

Georges-Louis Leclerc de Buffon 1707–1788 (Source Wikipedia)



Source: Ross, Probability 8th ed.

- A table is ruled with equidistant, parallel lines a distance D apart.
- A needle of length L is thrown randomly on the table.
- What is the probability that the needle will intersect one of the two lines?

Let X be the distance of the middle point of the needle to the closest parallel line. Needle intersects a line if hypotenuse of the triangle is less than $L/2$, i.e.,

$$\frac{X}{\cos(\theta)} < \frac{L}{2} \qquad \Leftrightarrow \qquad X < \frac{L}{2}\cos(\theta).$$

We assume that $X \in [0, D/2]$ and $\theta \in [0, \pi/2]$ are independent and uniform.

Can be thought of as: 1. Sample the middle point of needle, 2. Sample the angle.

Buffon's Needle Problem (2/2)

Let us compute the probability that the line intersects:

$$
\begin{aligned}
\mathbf{P}\left[X < \frac{L}{2}\cdot\cos(\theta)\right] &= \iint\limits_{x<(L/2)\cos y} f_{X,\theta}(x,y)\,dx\,dy \\
&= \iint\limits_{x<(L/2)\cos y} f_X(x)f_\theta(y)\,dx\,dy \\
&= \frac{4}{\pi D}\int_0^{\pi/2}\int_0^{L/2\cos(y)} dxdy \\
&= \frac{4}{\pi D}\int_0^{\pi/2}\frac{L}{2}\cos(y)dy \\
&= \frac{2L}{\pi D}.
\end{aligned}
$$

This gives us a method to estimate π!

Covariance

Definition of Covariance

Let X and Y be two random variables. The covariance is defined as:

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

Interpretation:

- If $\mathbf{Cov}[X, Y] > 0$ and X has a realisation larger (smaller) than $\mathbf{E}[X]$, then Y will likely have a realisation larger (smaller) than $\mathbf{E}[Y]$.
- If $\mathbf{Cov}[X, Y] < 0$, then it is the other way around.

Alternative Formula

Using the linearity of expectation rule, one has the equivalent definition:

$$\mathbf{Cov}[X, Y] = \mathbf{E}[X \cdot Y] - \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

- Note that $\mathbf{Cov}[X, X] = \mathbf{V}[X]$.
- Two variables X, Y with $\mathbf{Cov}[X, Y] > 0$ are positively correlated.
- Two variables X, Y with $\mathbf{Cov}[X, Y] < 0$ are negatively correlated.
- Two variables X, Y with $\mathbf{Cov}[X, Y] = 0$ are uncorrelated.

Illustration of 3 Cases for Cov $[X, Y]$

500 outcomes of randomly generated pairs of RVs (X, Y) with different joint distributions

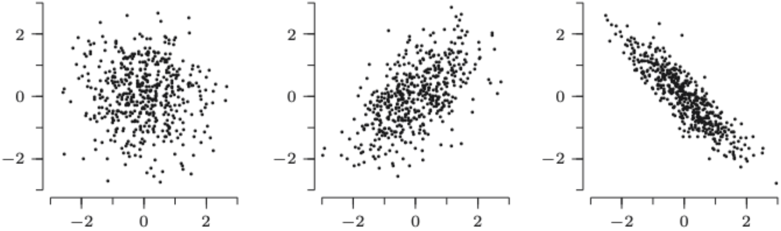


Fig. 10.1. Some scatterplots.
Source: Textbook by Dekking

1. What is the covariance (positive, negative, neutral)?
   - Left: set of sampled points has a circular shape, so uncorrelated.
   - Middle: looks like ellipsoids with $y = x$ as main axis, so positively correlated.
   - Right: looks like ellipsoids with $y = -x$ as main axis, so negatively correlated.
2. Where is the covariance the largest (in magnitude)?
   - Right: points more closely concentrated, hence correlation is largest.

Independence implies Uncorrelated

Example

Let X and Y be two independent random variables. Then X and Y are uncorrelated, i.e., $\mathbf{Cov}[X, Y] = 0$.

Answer

We give a proof for the discrete case:

$$\begin{aligned}
\mathbf{E}[X \cdot Y] &= \sum_i \sum_j a_i \cdot b_j \cdot \mathbf{P}[X = a_i, Y = b_j] \\
&= \sum_i \sum_j a_i \cdot b_j \cdot \mathbf{P}[X = a_i] \cdot \mathbf{P}[Y = b_j] \\
&= \left(\sum_i a_i \cdot \mathbf{P}[X = a_i] \right) \cdot \left(\sum_j b_j \cdot \mathbf{P}[Y = b_j] \right) \\
&= \mathbf{E}[X] \cdot \mathbf{E}[Y].
\end{aligned}$$

Uncorrelated may not imply Independence

Example

Find a (simple) example of two random variables X and Y which are uncorrelated but dependent.

Answer

- Let X be uniformly sampled from $\{-1, 0, +1\}$ and $Y := \mathbf{1}_{X=0}$.

$\Rightarrow$ $X \cdot Y = 0$ (for all outcomes), and thus

$$\mathbf{E}[X \cdot Y] = 0.$$

- Further, $\mathbf{E}[X] = 0$ (and $\mathbf{E}[Y] = 1/3$), and hence:

$$\mathbf{Cov}[X, Y] = \mathbf{E}[X \cdot Y] - \mathbf{E}[X] \cdot \mathbf{E}[Y] = 0.$$

- On the other hand, $\mathbf{P}[X = 0] = 1/3$ and $\mathbf{P}[Y = 0] = 2/3$, and thus

$$1 = \mathbf{P}[X \cdot Y = 0] > \mathbf{P}[X = 0] \cdot \mathbf{P}[Y = 0] = 2/9.$$

Variance of Sums and Covariances

Variance of Sum Formula

- For any two random variables X, Y,

$$\mathbf{V}[X+Y] = \mathbf{V}[X] + \mathbf{V}[Y] + 2 \cdot \mathbf{Cov}[X, Y].$$

- Hence if X and Y are uncorrelated variables,

$$\mathbf{V}[X+Y] = \mathbf{V}[X] + \mathbf{V}[Y].$$

Generalisation of the case where X and Y are even independent!

- For any random variables $X_1, X_2, \ldots, X_n$:

$$\mathbf{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{V}[X_i] + 2 \cdot \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbf{Cov}[X_i, X_j].$$

Computing Variances of Sums of Uncorrelated Variables

Example

Recall the example where $X \in \{-1, 0, +1\}$ uniformly and $Y := \mathbf{1}_{X=0}$. Compute $\mathbf{V}[X+Y]$.

Answer

- We first compute $\mathbf{V}[X]$:

$$\mathbf{V}[X] = \frac{1}{3} \cdot (-1)^2 + \frac{1}{3} \cdot 0^2 + \frac{1}{3} \cdot 1^2 = \frac{2}{3}.$$

- Now for $\mathbf{V}[Y]$:

$$\mathbf{V}[Y] = \frac{1}{3} \cdot (1 - \frac{1}{3})^2 + \frac{2}{3}(0 - \frac{1}{3})^2$$
$$= \frac{2}{9}.$$

$\Rightarrow$ Hence:

$$\mathbf{V}[X+Y] = \mathbf{V}[X] + \mathbf{V}[Y] + 2 \cdot \mathbf{Cov}[X, Y]$$
$$= \frac{2}{3} + \frac{2}{9} + 0 = \frac{8}{9}.$$

Correlation Coefficient: Normalising the Covariance

The definition of covariance is not scaling invariant:

- If X increases by a factor of α, then $\mathbf{Cov}[X, Y]$ increases by a factor of α.

$\Rightarrow$ Even if X and Y both increase by α, then $\mathbf{Cov}[X, Y]$ will change. (Exercise: It changes by?)

Correlation Coefficient

Let X and Y be two random variables. The correlation coefficient $\rho(X, Y)$ is defined as:

$$\rho(X, Y) = \frac{\mathbf{Cov}[X, Y]}{\sqrt{\mathbf{V}[X] \cdot \mathbf{V}[Y]}}.$$

If $\mathbf{V}[X] = 0$ or $\mathbf{V}[Y] = 0$, then it is defined as 0.

Properties:

1. The correlation coefficient is scaling-invariant, i.e., $\rho(X, Y) = \rho(\alpha \cdot X, \beta \cdot Y)$ for any $\alpha, \beta > 0$.
2. For any two random variables X, Y, $\rho(X, Y) \in [-1, 1]$.

Range of the Correlation Coefficient

Example

Verify that the correlation coefficients' range satisfies $\rho(X, Y) \in [-1, 1]$.

Answer

- We will only prove $\rho(X, Y) \geq -1$ (the other direction follows in analogous way).
- Let σ_x^2 and σ_y^2 denote the variances of X and Y, and σ_x and σ_y their standard deviations.
- Then:

$$
\begin{aligned}
0 &\leq \mathbf{V}\left[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right] \\
&= \mathbf{V}\left[\frac{X}{\sigma_X}\right] + \mathbf{V}\left[\frac{Y}{\sigma_Y}\right] + 2\,\mathbf{Cov}\left[\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right] \\
&= \frac{\mathbf{V}[X]}{\mathbf{V}[X]} + \frac{\mathbf{V}[Y]}{\mathbf{V}[Y]} + 2 \cdot \frac{\mathbf{Cov}[X, Y]}{\sigma_X \cdot \sigma_X} \\
&= 2 \cdot (1 + \rho(X, Y)).
\end{aligned}
$$