Joint Probability Distributions (cont.)

Joint Probability Density Functions

The joint probability density function for the continuous random variables X and Y, denotes as $f_{XY}(x,y)$, satisfies the following properties:

(1) $f_{XY}(x,y) \geq 0$ for all x, y

(2) $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,dy = 1$$

(3) $$P\big((X,Y) \subset R\big) = \iint_R f_{XY}(x,y)\,dx\,dy \qquad (5\text{−}2)$$

An example of a joint PDF

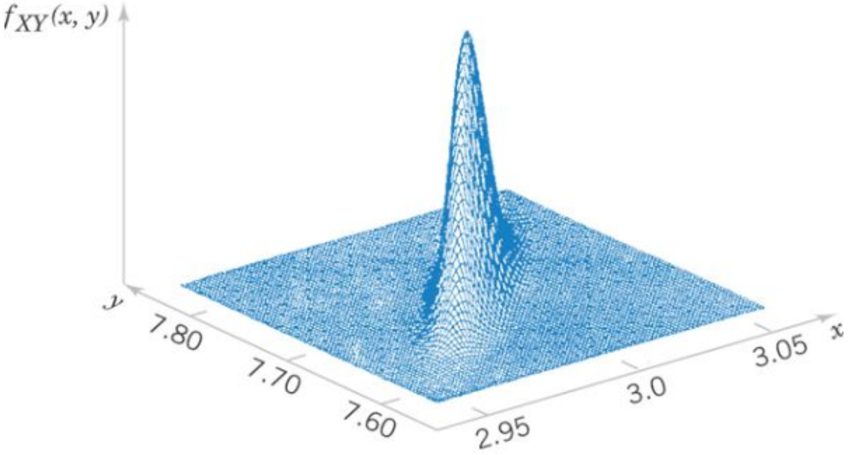


Figure 5-3 Joint probability density function for the continuous random variables *X* and *Y* of expression levels of two different genes. Note the asymmetric, narrow ridge shape of the PDF – indicating that small values in the *X* dimension are more likely to occur when small values in the *Y* dimension occur.

Marginal PDFs

- Rather than summing a discrete joint PMF, we integrate a continuous joint PDF.
- The marginal PDFs are used to make probability statements about one variable.
- If the joint probability density function of random variables *X* and *Y* is *fXY*(*x*,*y*), the marginal probability density functions of *X* and *Y* are:

$$f_X(x) = \int_y f_{XY}(x,y)\,dy$$
$$f_Y(y) = \int_x f_{XY}(x,y)\,dx \qquad (5\text{–}3)$$

$$f_X(x) = \sum_y f_{XY}(x,y)$$
$$f_Y(y) = \sum_x f_{XY}(x,y)$$

Conditional PDFs

Given continuous random variables X and Y with
joint probability density function $f_{XY}(x,y)$,
the conditional probability densiy function of Y given X=x is

$$f_{Y|x}(y) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{XY}(x,y)}{\int_y f_{XY}(x,y)\,dy} \text{ if } f_X(x) > 0 \qquad (5-4)$$

which satifies the following properties:

(1) $f_{Y|x}(y) \geq 0$

(2) $\int f_{Y|x}(y)dy = 1$

Conditional Probabilities

Conditional probability distributions can be developed for multiple random variables by extension of the ideas used for two random variables.

Suppose p = 5 and we wish to find the distribution of X1, X2 and X3 conditional on X4=x4 and X5=x5.

$$f_{X_1X_2X_3|x_4x_5}(x_1,x_2,x_3) = \frac{f_{X_1X_2X_3X_4X_5}(x_1,x_2,x_3,x_4,x_5)}{f_{X_4X_5}(x_4,x_5)}$$

for $f_{X_4X_5}(x_4,x_5) > 0$.

Independence for Continuous Random Variables

For random variables X and Y, if any one of the following properties is true, the others are also true. Then X and Y are independent.

(1) $f_{XY}(x, y) = f_X(x) \cdot f_Y(y)$

(2) $f_{Y|x}(y) = f_Y(y)$ for all x and y with $f_X(x) > 0$

(3) $f_{X|y}(y) = f_X(x)$ for all x and y with $f_Y(y) > 0$

(4) $\mathrm{P}(X \subset A, Y \subset B) = P(X \subset A) \cdot P(Y \subset B)$ for any sets A and B in the range of X and Y, respectively. (5−7)

Covariance and Correlation

Covariance Defined

Covariance is a number quantifying the average ***linear*** dependence between two random variables.

The covariance between the random variables X and Y, denoted as $\text{cov}(X, Y)$ or σ_{XY} is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$$

Montgomery, Runger 5th edition Eq. (5–14)

The units of σ_{XY} are the units of X times the units of Y.

Unlike the range of the variance, covariance can be negative: $-\infty < \sigma_{XY} < \infty$.

Covariance Defined

Covariance is a number quantifying the average ***linear*** dependence between two random variables.

The covariance between the random variables X and Y, denoted as $\text{cov}(X, Y)$ or σ_{XY} is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y$$

Montgomery, Runger 5th edition Eq. (5–14)

The units of σ_{XY} are the units of X times the units of Y.

Unlike the range of the variance, covariance can be negative: $-\infty < \sigma_{XY} < \infty$.

Notes About Covariance

Covariance measures dependencies between variables

If X and Y are independent, Cov(X,Y) = 0

$-\infty < \mathrm{Cov}(X,Y) < \infty$: So covariances can be negative

Covariances And PMFs

y = number of times city name is stated	x = number of bars of signal strength		
	1	2	3
1	0.01	0.02	0.25
2	0.02	0.03	0.20
3	0.02	0.10	0.05
4	0.15	0.10	0.05

Do you think the covariance will be:

A. Positive
B. Negative
C. Zero

Covariances And PMFs

y = number of times city name is stated	x = number of bars of signal strength		
	1	2	3
1	0.01	0.02	0.25
2	0.02	0.03	0.20
3	0.02	0.10	0.05
4	0.15	0.10	0.05

By inspection, note that the larger probabilities occur as X and Y move in opposite directions. This indicates a negative covariance.

Covariances and Scatter Plots

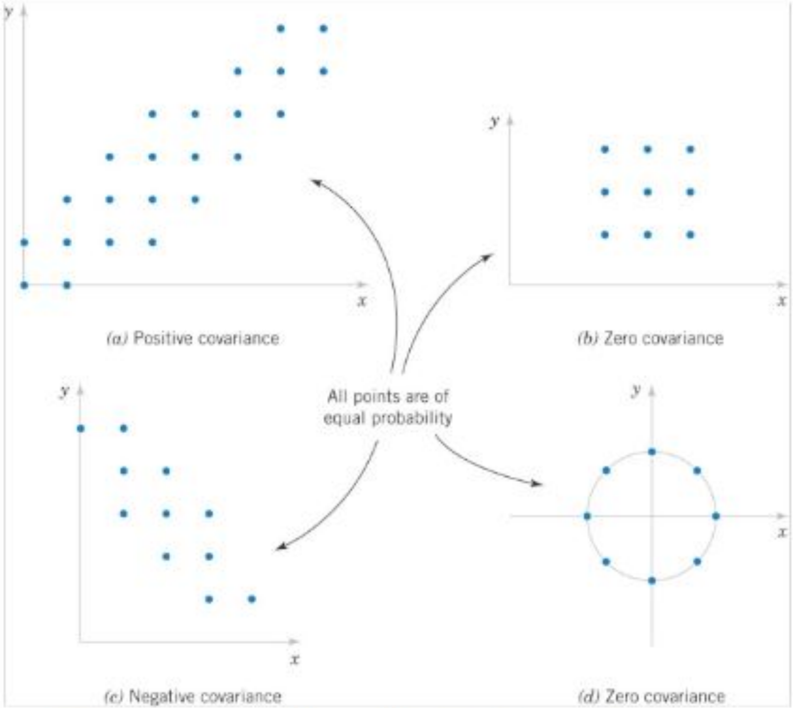


Figure 5-13 Joint probability distributions and the sign of cov(*X*, *Y*). Note that covariance is a measure of linear relationship. Variables with non-zero covariance are correlated.

Correlation

Correlation is normalized covariance

This is also known as Pearson correlation coefficient

$\rho_{XY} = \sigma_{XY} / \sigma_X \sigma_Y$

is the covariance normalized to be $-1 \le \rho_{XY} \le 1$

Spearman Correlation

Pearson correlation tests for linear relationship between X and Y

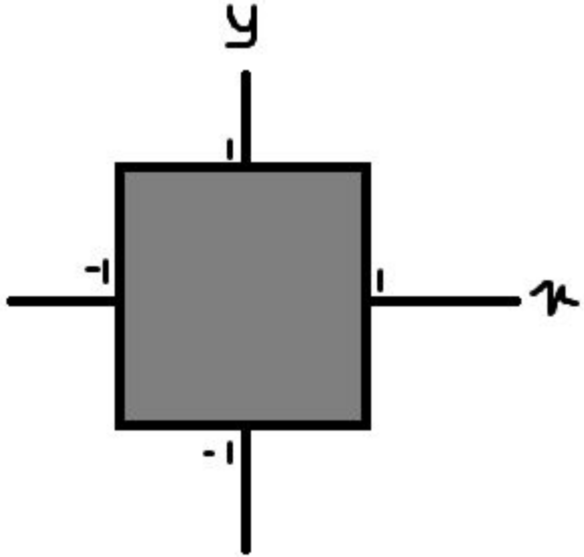
Spearman correlation tests for any monotonic relationship between X and Y

Calculate ranks (1 to n), rX(i) and rY(i) of variables in both samples. Calculate Pearson correlation between ranks: Spearman(X,Y) = Pearson(rX, rY)

Ties: convert to fractions, e.g. tie for 6s and 7s place both get 6.5. This can lead to artefacts.

Example 1

Uniform distribution in the square $-1<X<1,-1<Y<1$



Example 1

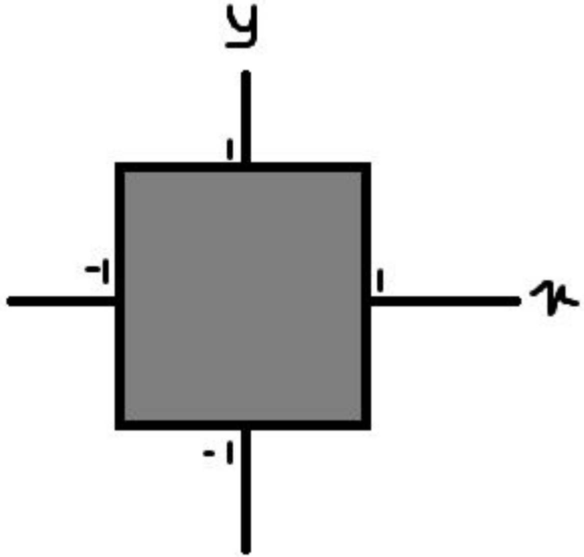
Uniform distribution in the square -1<X<1,-1<Y<1

$$f_{XY}(x,y) = c \text{ if } -1 < x < 1, -1 < y < 1$$

$$1 = \int_{\text{square}} f_{XY}(x,y)\, dxdy$$

$$1 = c * \text{area} = 4c$$

$$c = 1/4$$



Example 1

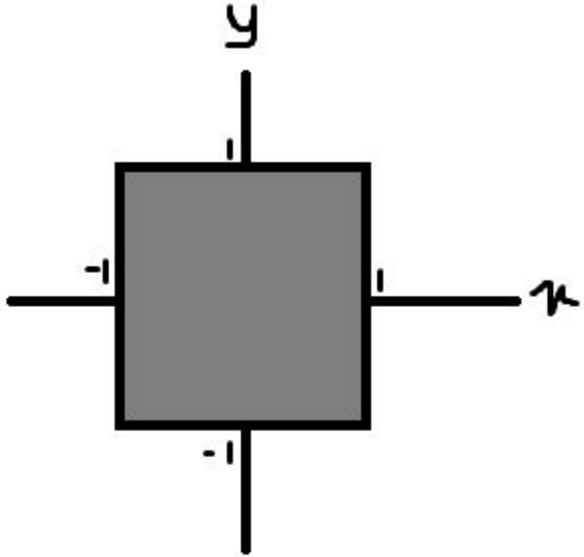
Are X and Y independent?

Let's see if fXY(x,y) = fX(x) * fY(y)

$$f_X(x) = \int_{-inf}^{inf} f_{XY}(x,y)\,dy$$
$$= \int_{-1}^{1} \frac{1}{4}\,dy = \frac{1}{2} \text{ if } -1 < x < 1$$

Same for fY(y). So,

$$f_X(x) * f_Y(y) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$$
$$= f_{XY}(x,y)$$



Independent!

Example 2

Are X and Y uniformly distributed in the disc $x^2+y^2 \leq 1$ independent?

$$f_{XY}(x,y) = \frac{1}{\text{Area}} = \frac{1}{\pi}$$

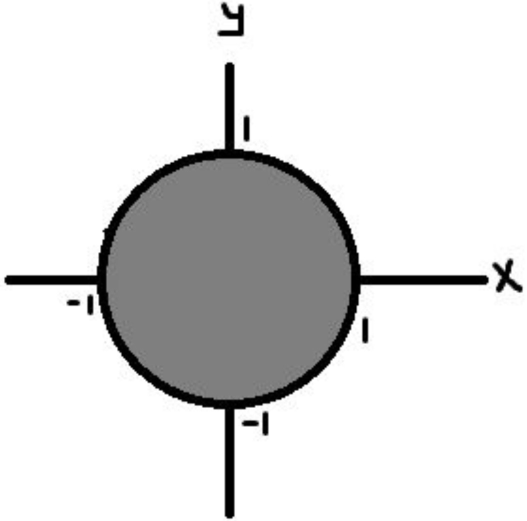
Let's see if fXY(x,y) = fX(x) * fY(y)

$$f_X(x) = \int_{-inf}^{inf} f_{XY}(x,y)\, dy$$

$$= \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\, dy = \frac{2 * \sqrt{1-x^2}}{\pi}$$

Same for fY(y). So,

$$(y) * f_Y(y) = \frac{2}{\pi}\sqrt{1-x^2} * \frac{2}{\pi}\sqrt{1-y^2}$$
$$\neq f_{XY}(x,y)$$



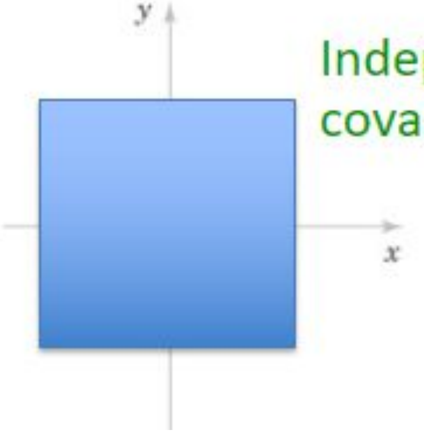
Not independent!

Independence Implies σ=ρ = 0 but <u>not vice versa</u>

- If *X* and *Y* are independent random variables,

  $$\sigma_{XY} = \rho_{XY} = 0 \qquad (5\text{-}17)$$

- $\rho_{XY} = 0$ is necessary, but not a sufficient condition for independence.



Independent
covariance=0

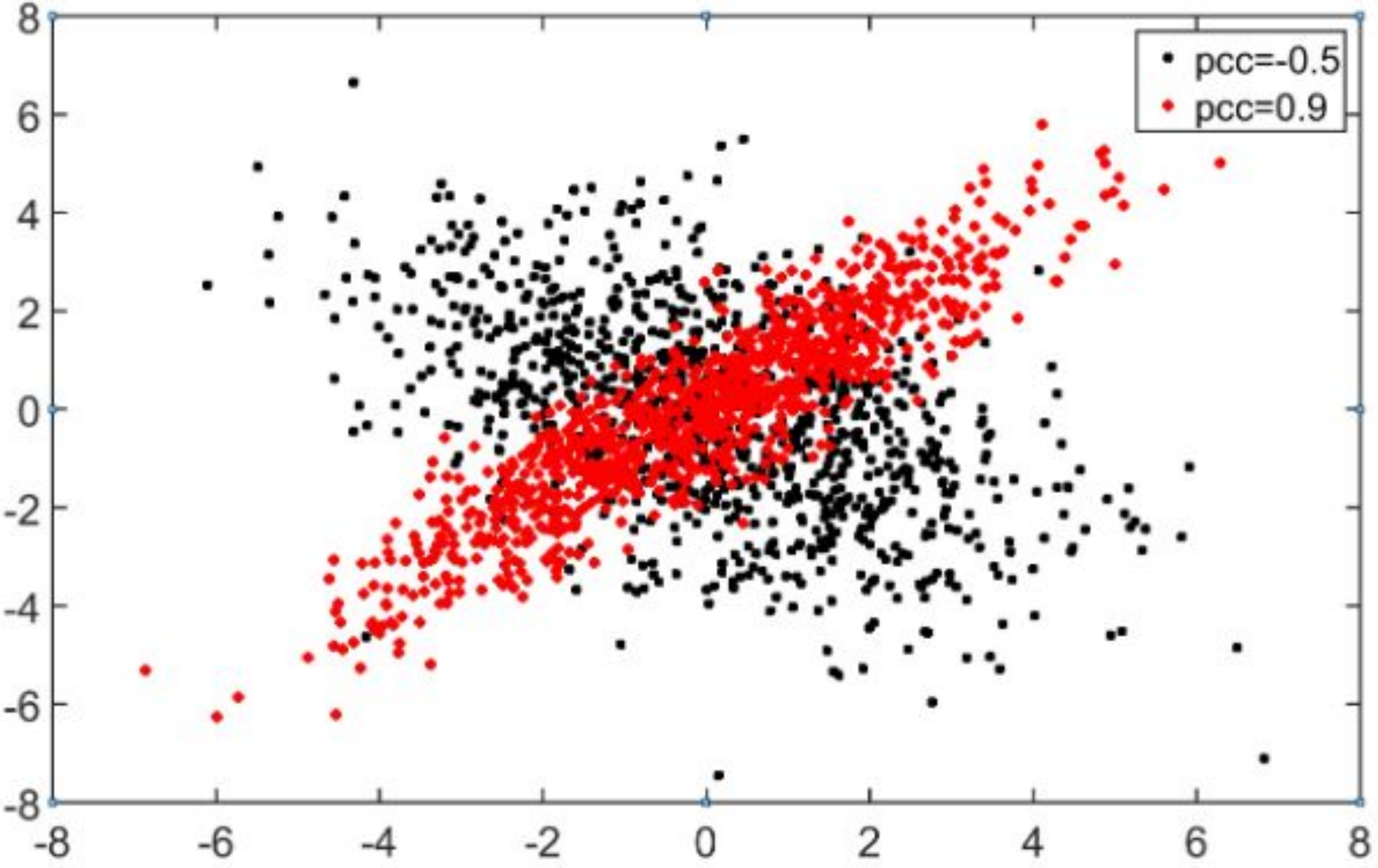
NOT independent
covariance=0

Matlab Exercise: Covariance/Correlation

- Generate a sample with Stats=100,000 of two Gaussian random variables r1 and r2 which have mean 0 and standard deviation 2 and are:
  - Uncorrelated
  - Correlated with correlation coefficient 0.9
  - Correlated with correlation coefficient -0.5
  - Trick: first make uncorrelated r1 and r2. Then make anew variable: r1mix=mix.*r2+(1-mix.^2)^0.5.*r1; where mix= corr. coeff.
- For each value of mix calculate covariance and correlation coefficient between r1mix and r2
- In each case make a scatter plot: plot(r1mix,r2,’k.’);

Matlab Exercise: Covariance/Correlation

```
1.  Stats=100000;
2.  r1=2.*randn(Stats,1);
3.  r2=2.*randn(Stats,1);
4.  disp('Covariance matrix='); disp(cov(r1,r2));
5.  disp('Correlation=');disp(corr(r1,r2));
6.  figure; plot(r1,r2,'k.');
7.  mix=0.9; %Mixes r2 to r1 but keeps same variance
8.  r1mix=mix.*r2+sqrt(1-mix.^2).*r1;
9.  disp('Covariance matrix='); disp(cov(r1mix,r2));
10. disp('Correlation=');disp(corr(r1mix,r2));
11. figure; plot(r1mix,r2,'k.');
12. mix=-0.5; %REDO LINES 8-11
```



pcc=-0.5
pcc=0.9
8
6
4
2
0
-2
-4
-6
-8
-8
-6
-4
-2
0
2
4
6
8