DS-GA 1002 Lecture notes 4 Fall 2016

Expectation

1 Introduction

The mean, variance and covariance allow us to describe the behavior of random variables very succinctly, without having to specify their distribution. The mean is the value around which the distribution of a random variable is centered. The variance quantifies the extent to which a random variable fluctuates around the mean. The covariance of two random variables is a measure of whether they tend to deviate from their means in a similar way. In order to define these concepts rigorously, we begin by introducing the expectation operator.

2 Expectation operator

The expectation operator maps a function of a random variable or of several random variables to an average weighted by the corresponding pmf or pdf.

Definition 2.1 (Expectation for discrete random variables)**.** *Let X be a discrete random variable with range R. The expected value of a function $g(X)$, $g : \mathbb{R} \to \mathbb{R}$, of X is*

$$\mathrm{E}\left(g\left(X\right)\right) = \sum_{x \in R} g\left(x\right) p_X\left(x\right). \tag{1}$$

Similarly, if X, Y are both discrete random variables with ranges R_X and R_Y then the expected value of a function $g(X, Y)$, $g : \mathbb{R}^2 \to \mathbb{R}$, of X and Y is

$$\mathrm{E}\left(g\left(X, Y\right)\right) = \sum_{x \in R_X} \sum_{x \in R_Y} g\left(x, y\right) p_{X,Y}\left(x, y\right). \tag{2}$$

If $\vec{X}$ is an n-dimensional discrete random vector, the expected value of a function $g\left(\vec{X}\right)$, $g : \mathbb{R}^n \to \mathbb{R}$, of $\vec{X}$ is

$$\mathrm{E}\left(g\left(\vec{X}\right)\right) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} g\left(\vec{x}\right) p_{\vec{X}}\left(\vec{x}\right). \tag{3}$$

Definition 2.2 (Expectation for continuous random variables)**.** *Let X be a continuous random variable. The expected value of a function $g(X)$, $g : \mathbb{R} \to \mathbb{R}$, of X is*

$$\mathrm{E}\left(g\left(X\right)\right) = \int_{x=-\infty}^{\infty} g\left(x\right) f_X\left(x\right)\, dx. \tag{4}$$

Similarly, if X, Y are both continuous random variables then the expected value of a function $g(X,Y)$, $g : \mathbb{R}^2 \to \mathbb{R}$, of X and Y is

$$\mathrm{E}\left(g\left(X,Y\right)\right) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} g\left(x,y\right) f_{X,Y}\left(x,y\right)\, dx\, dy. \tag{5}$$

If $\vec{X}$ is an n-dimensional random vector, the expected value of a function $g(X)$, $g : \mathbb{R}^n \to \mathbb{R}$, of $\vec{X}$ is

$$\mathrm{E}\left(g\left(\vec{X}\right)\right) = \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{\infty} \cdots \int_{x_n=-\infty}^{\infty} g\left(\vec{x}\right) f_{\vec{X}}\left(\vec{x}\right)\, dx_1\, dx_2 \ldots\, dx_n \tag{6}$$

In the case of quantities that depend on both continuous and discrete random variables, the product of the marginal and conditional distributions plays the role of the joint pdf or pmf.

Definition 2.3 (Expectation with respect to continuous and discrete random variables)**.** *If C is a continuous random variable and D a discrete random variable with range R_D defined on the same probability space, the expected value of a function $g(C,D)$ of C and D is*

$$\begin{aligned}\mathrm{E}\left(g\left(C,D\right)\right) &= \int_{c=-\infty}^{\infty} \sum_{d \in R_D} g\left(c,d\right) f_C\left(c\right) p_{D|C}\left(d|c\right)\, dc \qquad (7)\\ &= \sum_{d \in R_D} \int_{c=-\infty}^{\infty} g\left(c,d\right) p_D\left(d\right) f_{C|D}\left(c|d\right)\, dc. \qquad (8)\end{aligned}$$

The expected value of a certain quantity may be infinite or not even exist if the corresponding sum or integral tends towards infinity or has an undefined value. This is illustrated by Examples 2.4 and 3.2 below.

Example 2.4 (St Petersburg paradox)**.** A casino offers you the following game. You will flip an unbiased coin until it lands on heads and the casino will pay you 2^k dollars where k is the number of flips. How much are you willing to pay in order to play?

Let us compute the expected gain. If the flips are independent, the total number of flips X is a geometric random variable, so $p_X(k) = 1/2^k$. The gain is 2^X which means that

$$\mathrm{E}\left(\mathrm{Gain}\right) = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \infty. \tag{9}$$

The expected gain is infinite, but since you only get to play once, the amount of money that you are willing to pay is probably bounded. This is known as the St Petersburg paradox.

A fundamental property of the expectation operator is that it is linear.

Theorem 2.5 (Linearity of expectation)**.** *For any constant $a \in \mathbb{R}$, any function $g : \mathbb{R} \to \mathbb{R}$ and any continuous or discrete random variable X*

$$\mathrm{E}\left(a\, g\left(X\right)\right) = a\, \mathrm{E}\left(g\left(X\right)\right). \tag{10}$$

For any constants $a, b \in \mathbb{R}$, any functions $g_1, g_2 : \mathbb{R}^n \to \mathbb{R}$ and any continuous or discrete random variables X and Y

$$\mathrm{E}\left(a\, g_1\left(X, Y\right) + b\, g_2\left(X, Y\right)\right) = a\, \mathrm{E}\left(g_1\left(X, Y\right)\right) + b\, \mathrm{E}\left(g_2\left(X, Y\right)\right). \tag{11}$$

Proof. The theorem follows immediately from the linearity of sums and integrals. □

Linearity of expectation makes it very easy to compute the expectation of linear functions of random variables.

Example 2.6 (Coffee beans (continued from Ex. 3.19 in Lecture Notes 3))**.** Let us compute the expected total amount of beans that can be bought. C is uniform in $[0, 1]$, so $\mathrm{E}\left(C\right) = 1/2$. V is uniform in $[0, 2]$, so $\mathrm{E}\left(V\right) = 1$. By linearity of expectation

$$\begin{aligned} \mathrm{E}\left(C + V\right) &= \mathrm{E}\left(C\right) + \mathrm{E}\left(V\right) && (12) \\ &= 1.5 \text{ tons.} && (13) \end{aligned}$$

Note that this holds even if the two quantities are *not* independent.

If two random variables are independent, then the expectation of the product factors into a product of expectations.

Theorem 2.7 (Expectation of functions of independent random variables)**.** *If X, Y are independent random variables defined on the same probability space, and $g, h : \mathbb{R} \to \mathbb{R}$ are univariate real-valued functions, then*

$$\mathrm{E}\left(g\left(X\right) h\left(Y\right)\right) = \mathrm{E}\left(g\left(X\right)\right) \mathrm{E}\left(h\left(Y\right)\right). \tag{14}$$

Proof. We prove the result for continuous random variables, but the proof for discrete random variables is essentially the same.

$$\begin{aligned}
\mathrm{E}\left(g\left(X\right)h\left(Y\right)\right) &= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} g\left(x\right)h\left(y\right)f_{X,Y}\left(x,y\right)\,\mathrm{d}x\,\mathrm{d}y && (15)\\
&= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} g\left(x\right)h\left(y\right)f_{X}\left(x\right)f_{Y}\left(y\right)\,\mathrm{d}x\,\mathrm{d}y \quad \text{by independence} && (16)\\
&= \mathrm{E}\left(g\left(X\right)\right)\mathrm{E}\left(h\left(Y\right)\right). && (17)
\end{aligned}$$

□

3 Mean and variance

3.1 Mean

The mean of a random variable is equal to its expected value.

Definition 3.1 (Mean)**.** *The mean or first moment of X is the expected value of X:* $\mathrm{E}\left(X\right)$*.*

Table 1 lists the means of some important random variables. The derivations can be found in Section A of the appendix. As illustrated by Figure 3, the mean is the center of mass of the pmf or the pdf of the corresponding random variable.

If the distribution of a random variable is very *heavy tailed*, which means that the probability of the random variable taking large values decays slowly, its mean may be infinite. This is the case of the random variable representing the gain in Example 2.4. The following example shows that the mean may not exist if the value of the corresponding sum or integral is not well defined.

Example 3.2 (Cauchy random variable)**.** The pdf of the Cauchy random variable, which is shown in Figure 1, is given by

$$f_X(x) = \frac{1}{\pi(1+x^2)}. \tag{18}$$

By the definition of expected value,

$$\mathrm{E}(X) = \int_{-\infty}^{\infty}\frac{x}{\pi(1+x^2)}\,\mathrm{d}x = \int_{0}^{\infty}\frac{x}{\pi(1+x^2)}\,\mathrm{d}x - \int_{0}^{\infty}\frac{x}{\pi(1+x^2)}\,\mathrm{d}x. \tag{19}$$

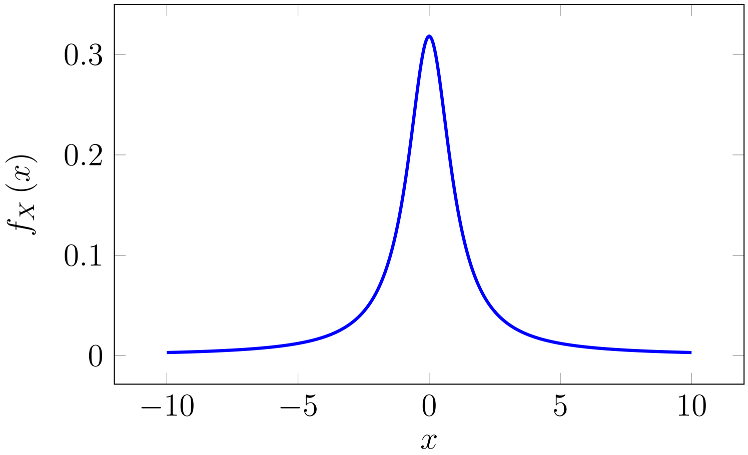


Figure 1: Probability density function of a Cauchy random variable.

Now, by the change of variables $t = x^2$,

$$\int_0^\infty \frac{x}{\pi(1+x^2)}\,\mathrm{d}x = \int_0^\infty \frac{1}{2\pi(1+t)}\mathrm{d}t = \lim_{t\to\infty} \frac{\log(1+t)}{2\pi} = \infty, \tag{20}$$

so $\mathrm{E}(X)$ does not exist, as it is the difference of two limits that tend to infinity.

The mean of a random vector is defined as the vector formed by the means of its components.

Definition 3.3 (Mean of a random vector)**.** *The mean of a random vector $\vec{X}$ is*

$$\mathrm{E}\left(\vec{X}\right) := \begin{bmatrix} \mathrm{E}\left(X_1\right) \\ \mathrm{E}\left(X_2\right) \\ \cdots \\ \mathrm{E}\left(X_n\right) \end{bmatrix}. \tag{21}$$

As in the univariate case, the mean can be interpreted as the value around which the distribution of the random vector is centered.

It follows immediately from the linearity of the expectation operator in one dimension that the mean operator is linear.

Theorem 3.4 (Mean of linear transformation of a random vector)**.** *For any random vector $\vec{X}$ of dimension n, any matrix $A \in \mathbb{R}^{m\times n}$ and $\vec{b} \in \mathbb{R}^m$*

$$\mathrm{E}\left(A\vec{X} + \vec{b}\right) = A\,\mathrm{E}\left(\vec{X}\right) + \vec{b}. \tag{22}$$

Proof.

$$\mathrm{E}\left(A\vec{X}+\vec{b}\right) = \begin{bmatrix} \mathrm{E}\left(\sum_{i=1}^{n} A_{1i}X_i + b_1\right) \\ \mathrm{E}\left(\sum_{i=1}^{n} A_{2i}X_i + b_2\right) \\ \cdots \\ \mathrm{E}\left(\sum_{i=1}^{n} A_{mi}X_i + b_n\right) \end{bmatrix} \tag{23}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} A_{1i}\mathrm{E}\left(X_i\right) + b_1 \\ \sum_{i=1}^{n} A_{2i}\mathrm{E}\left(X_i\right) + b_2 \\ \cdots \\ \sum_{i=1}^{n} A_{mi}\mathrm{E}\left(X_i\right) + b_n \end{bmatrix} \quad \text{by linearity of expectation} \tag{24}$$

$$= A\,\mathrm{E}\left(\vec{X}\right) + \vec{b}. \tag{25}$$

□

3.2 Median

The mean is often interpreted as representing a *typical* value taken by the random variable. However, the probability of a random variable being equal to its mean may be zero! For instance, a Bernoulli random variable cannot equal 0.5. In addition, the mean can be severely distorted by a small subset of extreme values, as illustrated by Example 3.6 below. The median is an alternative characterization of a *typical* value taken by the random variable, which is designed to be more robust to such situations. It is defined as the midpoint of the pmf or pdf of the random variable. If the random variable is continuous, the probability that it is either larger or smaller than the median is equal to 1/2.

Definition 3.5 (Median)**.** *The median of a discrete random variable X is a number m such that*

$$P\left(X \leq m\right) \geq \frac{1}{2} \quad \textit{and} \quad P\left(X \geq m\right) \geq \frac{1}{2}. \tag{26}$$

The median of a continuous random variable X is a number m such that

$$F_X\left(m\right) = \int_{-\infty}^{m} f_X\left(x\right)\, dx = \frac{1}{2}. \tag{27}$$

The following example illustrates the robustness of the median to the presence of a small subset of extreme values with nonzero probability.

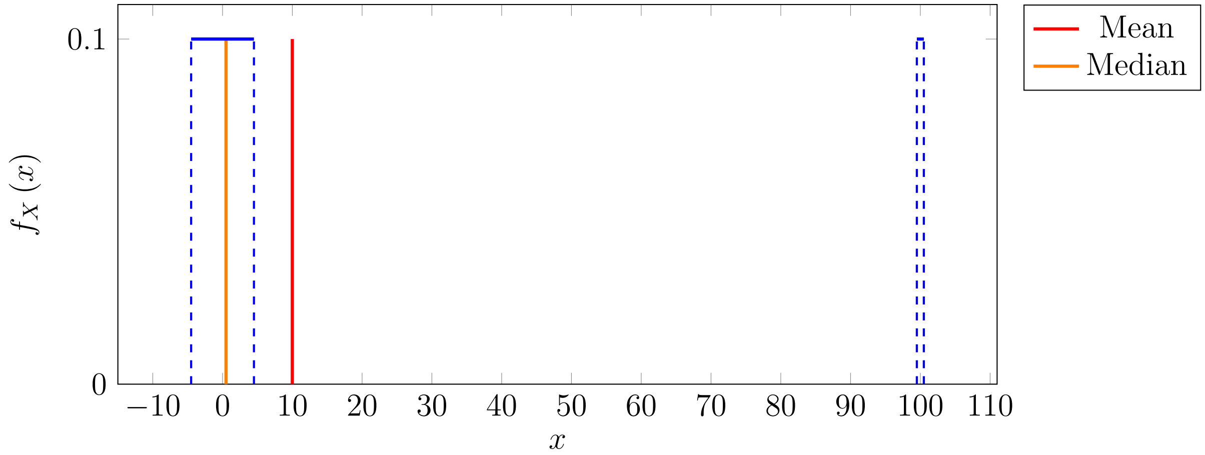


Figure 2: Uniform pdf in $[-4.5, 4.5] \cup [99.5, 100.5]$. The mean is 10 and the median is 0.5.

Example 3.6 (Mean vs median)**.** Consider a uniform random variable X with support $[-4.5, 4.5] \cup [99.5, 100.5]$. The mean of X equals

$$\begin{aligned} \mathrm{E}(X) &= \int_{x=-4.5}^{4.5} x f_X(x) \, \mathrm{d}x + \int_{x=99.5}^{100.5} x f_X(x) \, \mathrm{d}x && (28) \\ &= \frac{1}{10} \frac{100.5^2 - 99.5^2}{2} && (29) \\ &= 10. && (30) \end{aligned}$$

The cdf of X between -4.5 and 4.5 is equal to

$$\begin{aligned} F_X(m) &= \int_{-4.5}^{m} f_X(x) \, \mathrm{d}x && (31) \\ &= \frac{m + 4.5}{10}. && (32) \end{aligned}$$

Setting this equal to 1/2 allows to compute the median which is equal to 0.5. Figure 2 shows the pdf of X and the location of the median and the mean. The median provides a more realistic measure of the center of the distribution.

3.3 Variance and standard deviation

The expected value of the square of a random variable is sometimes used to quantify the *energy* of the random variable.

Random variable	**Parameters**	**Mean**	**Variance**
Bernoulli	p	p	$p\,(1-p)$
Geometric	p	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial	n, p	np	$np\,(1-p)$
Poisson	λ	λ	λ
Uniform	a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gaussian	μ, σ	μ	σ^2

Table 1: Means and variance of common random variables, derived in Section A of the appendix.

Definition 3.7 (Second moment)**.** *The mean square or second moment of a random variable X is the expected value of X^2:* $\mathrm{E}\left(X^2\right)$.

The mean square of the difference between the random variable and its mean is called the variance of the random value. It quantifies the variation of the random variable around its mean. The square root of this quantity is the standard deviation of the random variable.

Definition 3.8 (Variance and standard deviation)**.** *The variance of X is the mean square deviation from the mean*

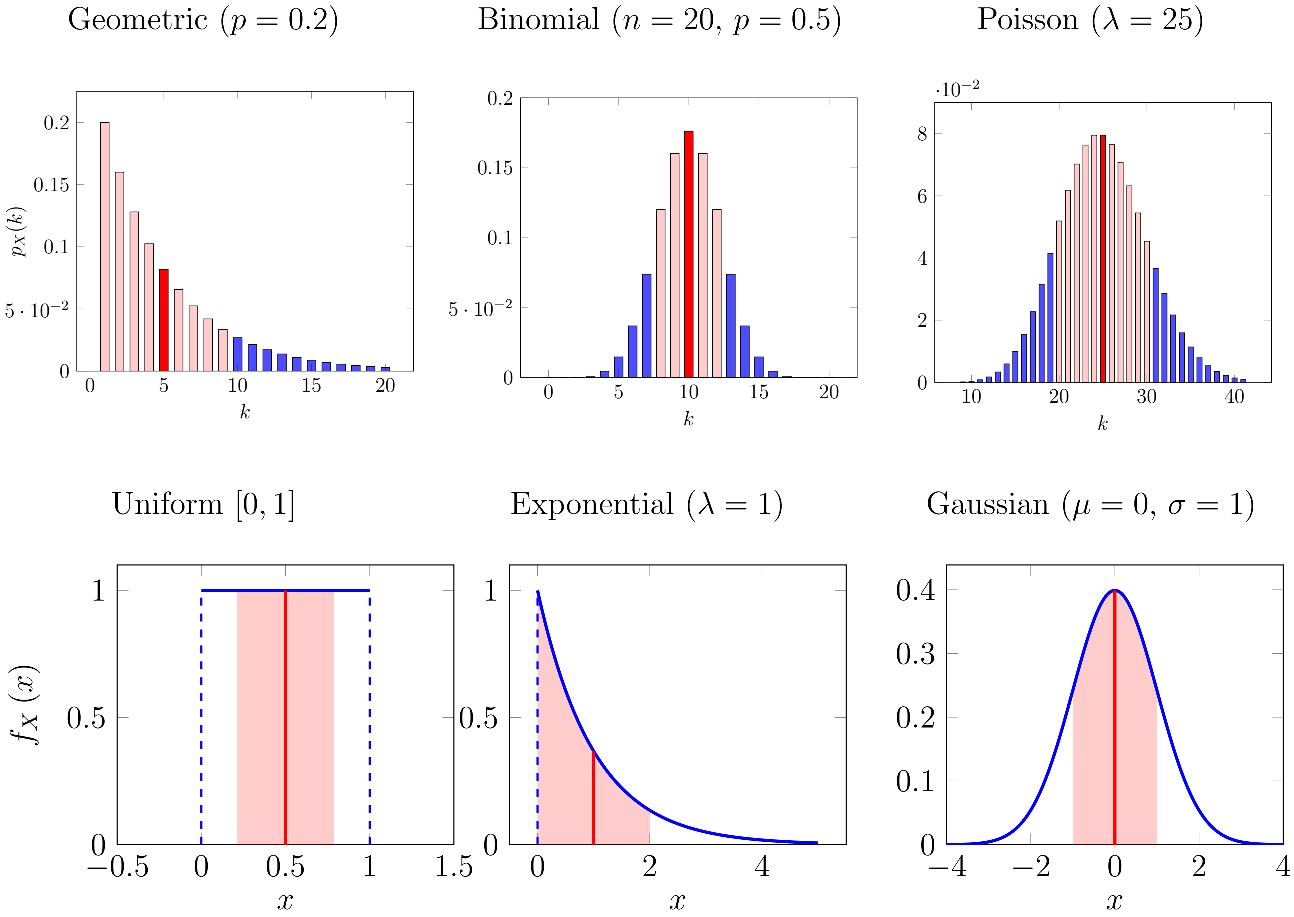
$$\begin{aligned} \mathrm{Var}\,(X) &:= \mathrm{E}\left((X-\mathrm{E}\,(X))^2\right) && (33)\\ &= \mathrm{E}\left(X^2\right)-\mathrm{E}^2\,(X)\,. && (34)\end{aligned}$$

The standard deviation σ_X of X is

$$\sigma_X := \sqrt{\mathrm{Var}\,(X)}. \tag{35}$$

We have compiled the variances of some important random variables in Table 1. The derivations can be found in Section A of the appendix. In Figure 3 we plot the pmfs and pdfs of these random variables and display the range of values that fall within one standard deviation of the mean.

The variance operator is not linear, but it is straightforward to determine the variance of a linear function of a random variable.


Geometric ($p = 0.2$)

Binomial ($n = 20$, $p = 0.5$)

Poisson ($\lambda = 25$)

Uniform $[0, 1]$

Exponential ($\lambda = 1$)

Gaussian ($\mu = 0$, $\sigma = 1$)

Figure 3: Pmfs of discrete random variables (top row) and pdfs of continuous random variables (bottom row). The mean of the random variable is marked in red. Values that are within one standard deviation of the mean are marked in pink.

Lemma 3.9 (Variance of linear functions)**.** *For any constants a and b*

$$\operatorname{Var}(a\,X+b)=a^2\operatorname{Var}(X)\,. \tag{36}$$

Proof.

$$\begin{aligned}\operatorname{Var}(a\,X+b)&=\operatorname{E}\left((a\,X+b-\operatorname{E}(a\,X+b))^2\right) && (37)\\&=\operatorname{E}\left((a\,X+b-a\operatorname{E}(X)-b)^2\right) && (38)\\&=a^2\operatorname{E}\left((X-\operatorname{E}(X))^2\right) && (39)\\&=a^2\operatorname{Var}(X)\,. && (40)\end{aligned}$$

□

This result makes sense: If we change the center of the random variable by adding a constant, then the variance is not affected because the variance only measures the deviation from the mean. If we multiply a random variable by a constant, the standard deviation is scaled by the same factor.

3.4 Bounding probabilities using the mean and variance

In this section we introduce two inequalities that allow to characterize the behavior of a random valuable to some extent just from knowing its mean and variance. The first is the Markov inequality, which quantifies the intuitive idea that if a random variable is nonnegative and small then the probability that it takes large values must be low.

Theorem 3.10 (Markov's inequality)**.** *Let X be a nonnegative random variable. For any positive constant $a>0$,*

$$\operatorname{P}(X\geq a)\leq\frac{\operatorname{E}(X)}{a}. \tag{41}$$

Proof. Consider the indicator variable $1_{X\geq a}$. We have

$$X-a\,1_{X\geq a}\geq 0. \tag{42}$$

In particular its expectation is non negative (as it is the sum or integral of a non-negative quantity over the positive real line). By linearity of expectation and the fact that $1_{X\geq a}$ is a Bernoulli random variable with expectation $\operatorname{P}(X\geq a)$ we have

$$\operatorname{E}(X)\geq a\operatorname{E}(1_{X\geq a})=a\operatorname{P}(X\geq a)\,. \tag{43}$$

□

Example 3.11 (Age of students)**.** You hear that the mean age of NYU students is 20 years, but you know quite a few students that are older than 30. You decide to apply Markov's inequality to bound the fraction of students above 30 by modeling age as a nonnegative random variable A.

$$\mathrm{P}(A \geq 30) \leq \frac{\mathrm{E}\,(A)}{30} = \frac{2}{3}. \tag{44}$$

At most two thirds of the students are over 30.

As you can see from Example 3.14, Markov's inequality can be rather loose. The reason is that it barely uses any information about the distribution of the random variable.

Chebyshev's inequality controls the deviation of the random variable from its mean. Intuitively, if the variance (and hence the standard deviation) is small, then the probability that the random variable is far from its mean must be low.

Theorem 3.12 (Chebyshev's inequality)**.** *For any positive constant $a > 0$ and any random variable X with bounded variance,*

$$\mathrm{P}\left(|X - \mathrm{E}\,(X)| \geq a\right) \leq \frac{\mathrm{Var}\,(X)}{a^2}. \tag{45}$$

Proof. Applying Markov's inequality to the random variable $Y = (X - \mathrm{E}\,(X))^2$ yields the result. □

An interesting corollary to Chebyshev's inequality shows that if the variance of a random variable is zero, then the random variable is a constant or, to be precise, the probability that it deviates from its mean is zero.

Corollary 3.13. *If* $\mathrm{Var}\,(X) = 0$ *then* $\mathrm{P}\left(X \neq \mathrm{E}\,(X)\right) = 0$.

Proof. Take any $\epsilon > 0$, by Chebyshev's inequality

$$\mathrm{P}\left(|X - \mathrm{E}\,(X)| \geq \epsilon\right) \leq \frac{\mathrm{Var}\,(X)}{\epsilon^2} = 0. \tag{46}$$

□

Example 3.14 (Age of students (continued))**.** You are not very satisfied with your bound on the number of students above 30. You find out that the standard deviation of student age is actually just 3 years. Applying Chebyshev's inequality, this implies that

$$\begin{aligned} \mathrm{P}(A \geq 30) &\leq \mathrm{P}\left(\left|A - \mathrm{E}\left(A\right)\right| \geq 10\right) && (47) \\ &\leq \frac{\mathrm{Var}\left(A\right)}{100} = \frac{9}{100}. && (48) \end{aligned}$$

So actually at least 91% of the students are under 30 (and above 10).

---

4 Covariance

4.1 Covariance for two random variables

The covariance of two random variables describes their joint behavior. It is the expected value of the product between the difference of the random variables and their respective means. Intuitively, it measures to what extent the random variables fluctuate together.

Definition 4.1 (Covariance)**.** *The covariance of X and Y is*

$$\begin{aligned} \mathrm{Cov}\left(X, Y\right) &:= \mathrm{E}\left(\left(X - \mathrm{E}\left(X\right)\right)\left(Y - \mathrm{E}\left(Y\right)\right)\right) && (49) \\ &= \mathrm{E}\left(XY\right) - \mathrm{E}\left(X\right)\mathrm{E}\left(Y\right). && (50) \end{aligned}$$

If $\mathrm{Cov}\left(X, Y\right) = 0$*, X and Y are* ***uncorrelated****.*

Figure 4 shows samples from bivariate Gaussian distributions with different covariances. If the covariance is zero, then the joint pdf has a spherical form. If the covariance is positive and large, then the joint pdf becomes skewed so that the two variables tend to have similar values. If the covariance is large and negative, then the two variables will tend to have similar values with opposite sign.

The variance of the sum of two random variables can be expressed in terms of their individual variances and their covariance. As a result, their fluctuations reinforce each other if the covariance is positive and cancel each other if it is negative.

Theorem 4.2 (Variance of the sum of two random variables)**.**

$$\mathrm{Var}\left(X + Y\right) = \mathrm{Var}\left(X\right) + \mathrm{Var}\left(Y\right) + 2\,\mathrm{Cov}\left(X, Y\right). \qquad (51)$$

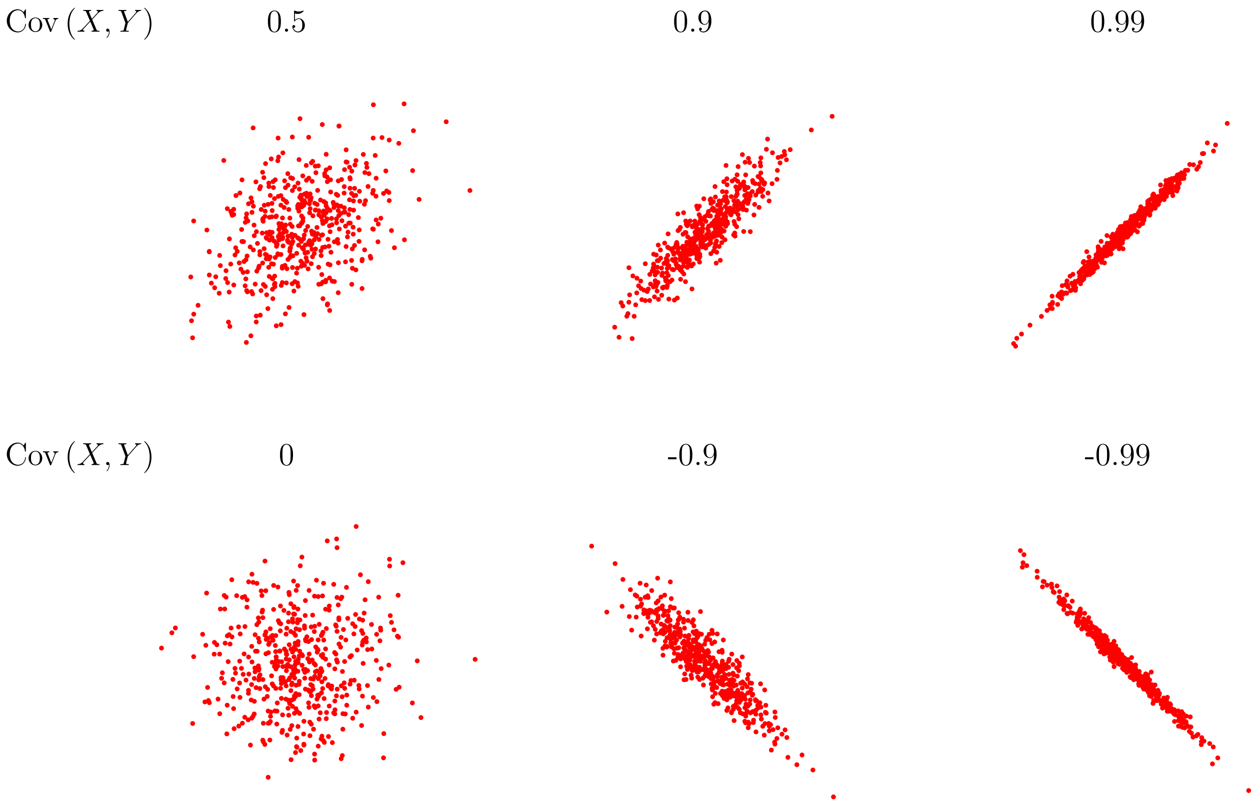


Figure 4: Samples from 2D Gaussian vectors (X, Y), where X and Y are standard Gaussian random variables with zero mean and unit variance, for different values of the covariance between X and Y.

Proof.

$$\begin{aligned}
\operatorname{Var}(X+Y) &= \mathrm{E}\left((X+Y-\mathrm{E}(X+Y))^2\right) && (52)\\
&= \mathrm{E}\left((X-\mathrm{E}(X))^2\right)+\mathrm{E}\left((Y-\mathrm{E}(Y))^2\right)+2\mathrm{E}\left((X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\right)\\
&= \operatorname{Var}(X)+\operatorname{Var}(Y)+2\operatorname{Cov}(X,Y). && (53)
\end{aligned}$$

□

An immediate consequence is that if two random variables are uncorrelated, then the variance of their sum equals the sum of their variances.

Corollary 4.3. *If X and Y are uncorrelated, then*

$$\operatorname{Var}(X+Y)=\operatorname{Var}(X)+\operatorname{Var}(Y). \tag{54}$$

The following lemma and example show that independence implies uncorrelation, but uncorrelation does not always imply independence.

Lemma 4.4 (Independence implies uncorrelation)**.** *If two random variables are independent, then they are uncorrelated.*

Proof. By Theorem 2.7, if X and Y are independent

$$\operatorname{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\,\mathrm{E}(Y)=\mathrm{E}(X)\,\mathrm{E}(Y)-\mathrm{E}(X)\,\mathrm{E}(Y)=0. \tag{55}$$

□

Example 4.5 (Uncorrelation does not imply independence)**.** Let X and Y be two independent Bernoulli random variables with parameter $1/2$. Consider the random variables

$$\begin{aligned}
U &= X+Y, && (56)\\
V &= X-Y. && (57)
\end{aligned}$$

Note that

$$\begin{aligned}
p_U(0) &= \mathrm{P}(X=0,Y=0)=\frac{1}{4}, && (58)\\
p_V(0) &= \mathrm{P}(X=1,Y=1)+\mathrm{P}(X=0,Y=0)=\frac{1}{2}, && (59)\\
p_{U,V}(0,0) &= \mathrm{P}(X=0,Y=0)=\frac{1}{4}\neq p_U(0)\,p_V(0)=\frac{1}{8}, && (60)
\end{aligned}$$

so U and V are not independent. However, they are uncorrelated as

$$\begin{aligned}
\operatorname{Cov}(U,V) &= \mathrm{E}(UV) - \mathrm{E}(U)\,\mathrm{E}(V) && (61)\\
&= \mathrm{E}((X+Y)(X-Y)) - \mathrm{E}(X+Y)\,\mathrm{E}(X-Y) && (62)\\
&= \mathrm{E}\left(X^2\right) - \mathrm{E}\left(Y^2\right) - \mathrm{E}^2(X) + \mathrm{E}^2(Y) = 0. && (63)
\end{aligned}$$

The final equality holds because X and Y have the same distribution.

4.2 Correlation coefficient

The covariance does not take into account the magnitude of the variances of the random variables involved. The Pearson correlation coefficient is obtained by normalizing the covariance using the standard deviations of both variables.

Definition 4.6 (Pearson correlation coefficient)**.** *The Pearson correlation coefficient of two random variables X and Y is*

$$\rho_{X,Y} := \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{64}$$

The correlation coefficient between X and Y is equal to the covariance between X/σ_X and Y/σ_Y. Figure 5 compares samples of bivariate Gaussian random variables that have the same correlation coefficient, but different covariance and vice versa.

Although it might not be immediately obvious, the magnitude of the correlation coefficient is bounded by one because the covariance of two random variables cannot exceed the product of their standard deviations. A useful interpretation of the correlation coefficient is that it quantifies to what extent X and Y are linearly related. In fact, if it is equal to 1 or -1 then one of the variables is a linear function of the other! All of this follows from the notorious Cauchy-Schwarz inequality. The proof is in Section C of the appendix

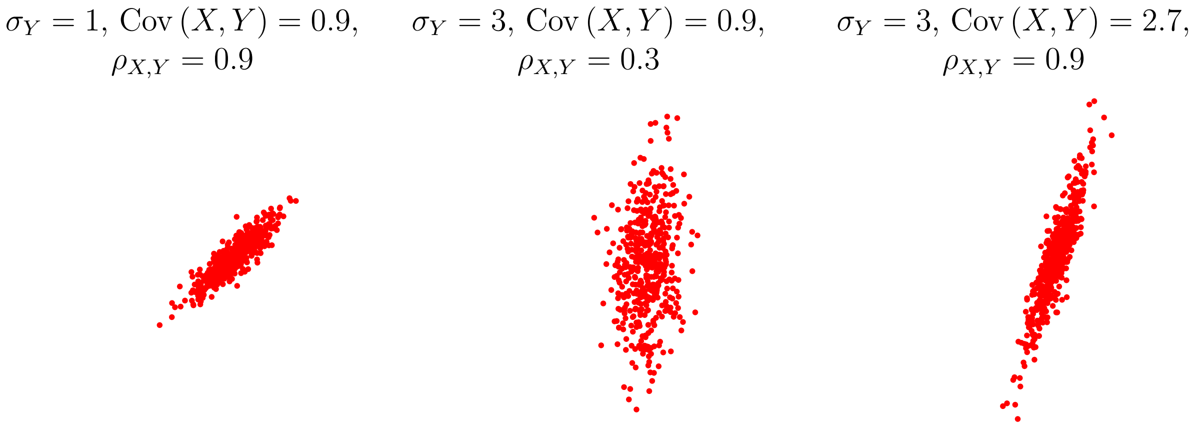
Theorem 4.7 (Cauchy-Schwarz inequality)**.** *For any random variables X and Y defined on the same probability space*

$$|\mathrm{E}(XY)| \leq \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}. \tag{65}$$

Assume $\mathrm{E}(X^2) \neq 0$*,*

$$\mathrm{E}(XY) = \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)} \iff Y = \sqrt{\frac{\mathrm{E}(Y^2)}{\mathrm{E}(X^2)}}X, \tag{66}$$

$$\mathrm{E}(XY) = -\sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)} \iff Y = -\sqrt{\frac{\mathrm{E}(Y^2)}{\mathrm{E}(X^2)}}X. \tag{67}$$



$\sigma_Y = 1$, $\mathrm{Cov}\left(X,Y\right) = 0.9$, $\rho_{X,Y} = 0.9$

$\sigma_Y = 3$, $\mathrm{Cov}\left(X,Y\right) = 0.9$, $\rho_{X,Y} = 0.3$

$\sigma_Y = 3$, $\mathrm{Cov}\left(X,Y\right) = 2.7$, $\rho_{X,Y} = 0.9$

Figure 5: Samples from 2D Gaussian vectors (X, Y), where X is a standard Gaussian random variables with zero mean and unit variance, for different values of the standard deviation σ_Y of Y (which is mean zero) and of the covariance between X and Y.

Corollary 4.8. *For any random variables X and Y,*

$$\mathrm{Cov}\left(X,Y\right) \le \sigma_X \sigma_Y. \tag{68}$$

Equivalently, the Pearson correlation coefficient satisfies

$$|\rho_{X,Y}| \le 1, \tag{69}$$

with equality if and only if there is a linear relationship between X and Y

$$|\rho_{X,Y}| = 1 \iff Y = c\,X + d. \tag{70}$$

where

$$c := \begin{cases} \frac{\sigma_Y}{\sigma_X} & \text{if } \rho_{X,Y} = 1, \\ -\frac{\sigma_Y}{\sigma_X} & \text{if } \rho_{X,Y} = -1, \end{cases} \qquad d := \mathrm{E}\left(Y\right) - c\,\mathrm{E}\left(X\right). \tag{71}$$

Proof. Let

$$U := X - \mathrm{E}\left(X\right), \tag{72}$$

$$V := Y - \mathrm{E}\left(Y\right). \tag{73}$$

From the definition of the variance and the correlation coefficient,

$$\mathrm{E}\left(U^2\right) = \mathrm{Var}\left(X\right), \tag{74}$$

$$\mathrm{E}\left(V^2\right) = \mathrm{Var}\left(Y\right) \tag{75}$$

$$\rho_{X,Y} = \frac{\mathrm{E}\left(UV\right)}{\sqrt{\mathrm{E}\left(U^2\right)\mathrm{E}\left(V^2\right)}}. \tag{76}$$

The result now follows from applying Theorem 4.7 to U and V. □

4.3 Covariance matrix of a random vector

The covariance matrix of a random vector captures the interaction between the components of the vector. It contains the variance of each component in the diagonal and the covariances between different components in the off diagonals.

Definition 4.9. *The covariance matrix of a random vector $\vec{X}$ is defined as*

$$\Sigma_{\vec{X}} := \begin{bmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_n, X_2) & \operatorname{Cov}(X_n, X_2) & \cdots & \operatorname{Var}(X_n) \end{bmatrix} \tag{77}$$

$$= \operatorname{E}\left(\vec{X}\vec{X}^T\right) - \operatorname{E}\left(\vec{X}\right)\operatorname{E}\left(\vec{X}\right)^T. \tag{78}$$

Note that if all the entries of a vector are uncorrelated, then its covariance matrix is diagonal.

From Theorem 3.4 we obtain a simple expression for the covariance matrix of the linear transformation of a random vector.

Theorem 4.10 (Covariance matrix after a linear transformation)**.** *Let $\vec{X}$ be a random vector of dimension n with covariance matrix Σ. For any matrix $A \in \mathbb{R}^{m\times n}$ and $\vec{b} \in \mathbb{R}^m$,*

$$\Sigma_{A\vec{X}+\vec{b}} = A\Sigma_{\vec{X}}A^T. \tag{79}$$

Proof.

$$\Sigma_{A\vec{X}+\vec{b}} = \operatorname{E}\left(\left(A\vec{X}+\vec{b}\right)\left(A\vec{X}+\vec{b}\right)^T\right) - \operatorname{E}\left(A\vec{X}+\vec{b}\right)\operatorname{E}\left(A\vec{X}+\vec{b}\right)^T \tag{80}$$

$$\begin{aligned} &= A\operatorname{E}\left(\vec{X}\vec{X}^T\right)A^T + \vec{b}\operatorname{E}\left(\vec{X}\right)^T A^T + A\operatorname{E}\left(\vec{X}\right)\vec{b}^T + \vec{b}\vec{b}^T \\ &\quad - A\operatorname{E}\left(\vec{X}\right)\operatorname{E}\left(\vec{X}\right)^T A^T - A\operatorname{E}\left(\vec{X}\right)\vec{b}^T - \vec{b}\operatorname{E}\left(\vec{X}\right)^T A^T - \vec{b}\vec{b}^T \end{aligned} \tag{81}$$

$$= A\left(\operatorname{E}\left(\vec{X}\vec{X}^T\right) - \operatorname{E}\left(\vec{X}\right)\operatorname{E}\left(\vec{X}\right)^T\right)A^T \tag{82}$$

$$= A\Sigma_{\vec{X}}A^T. \tag{83}$$

□

An immediate corollary of this result is that we can easily decode the variance of the random variable *in any direction* from the covariance matrix. Formally, the variance of the random variable in the direction of a unit vector $\vec{u}$ is equal to the variance of its projection onto $\vec{u}$.

Corollary 4.11. *Let $\vec{u}$ be a unit vector,*

$$\operatorname{Var}\left(\vec{u}^T\vec{X}\right) = \vec{u}^T\Sigma_{\vec{X}}\vec{u}. \tag{84}$$

Consider the problem of finding the direction in which the random vector X has the largest variance. This boils down to finding the maximum of the quadratic form $\vec{u}^T\Sigma_{\vec{X}}\vec{u}$ over all unit-norm vectors $\vec{u}$. To analyze the properties of $\vec{u}^T\Sigma_{\vec{X}}\vec{u}$ we resort to linear algebra (check the additional notes for a review). Consider the eigendecomposition of the covariance matrix

$$\begin{aligned}\Sigma_{\vec{X}} &= U\Lambda U^T && (85)\\ &= \begin{bmatrix}\vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n\end{bmatrix}\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_n\end{bmatrix}\begin{bmatrix}\vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_n\end{bmatrix}^T, && (86)\end{aligned}$$

where X is n dimensional. By definition, $\Sigma_{\vec{X}}$, as all covariance matrices, is symmetric, so its eigenvectors u_1, u_2, …, u_n are orthogonal. Furthermore, the eigenvectors and eigenvalues have a very intuitive interpretation in terms of the quadratic form of interest.

Theorem 4.12. *For any symmetric matrix $A \in \mathbb{R}^n$ with normalized eigenvectors $u_1, u_2, \ldots, u_n$ and corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$*

$$\begin{aligned}\lambda_1 &= \max_{||u||_2=1} u^T A u, && (87)\\ u_1 &= \arg\max_{||u||_2=1} u^T A u, && (88)\\ \lambda_k &= \max_{||u||_2=1, u\perp u_1,\ldots,u_{k-1}} u^T A u, && (89)\\ u_k &= \arg\max_{||u||_2=1, u\perp u_1,\ldots,u_{k-1}} u^T A u. && (90)\end{aligned}$$

The maximum of $\vec{u}^T\Sigma_{\vec{X}}\vec{u}$ is equal to the largest eigenvalue λ_1 of $\Sigma_{\vec{X}}$ and is attained by the corresponding eigenvector $\vec{u_1}$. This means that $\vec{u_1}$ is the direction of maximum variance. Moreover, the eigenvector $\vec{u_2}$ corresponding to the second largest eigenvalue λ_2 is the direction of maximum variation that is orthogonal to $\vec{u_1}$. In general, the eigenvector $\vec{u_k}$ corresponding to the kth largest eigenvalue λ_k reveals the direction of maximum variation that is orthogonal to $\vec{u_1}$, $\vec{u_2}$, …, $\vec{u_{k-1}}$. Finally, $\vec{u_n}$ is the direction of minimum variance. Figure 6 illustrates this with an example, where $n = 2$. As we will see later in the course, principal component analysis– a popular method for unsupervised learning and dimensionality reduction– is based on this phenomenon.

4.4 Whitening

Whitening is a useful procedure for preprocessing data. Its goal is to transform samples from a random vector $\vec{X}$ linearly so that each component is uncorrelated and the variance

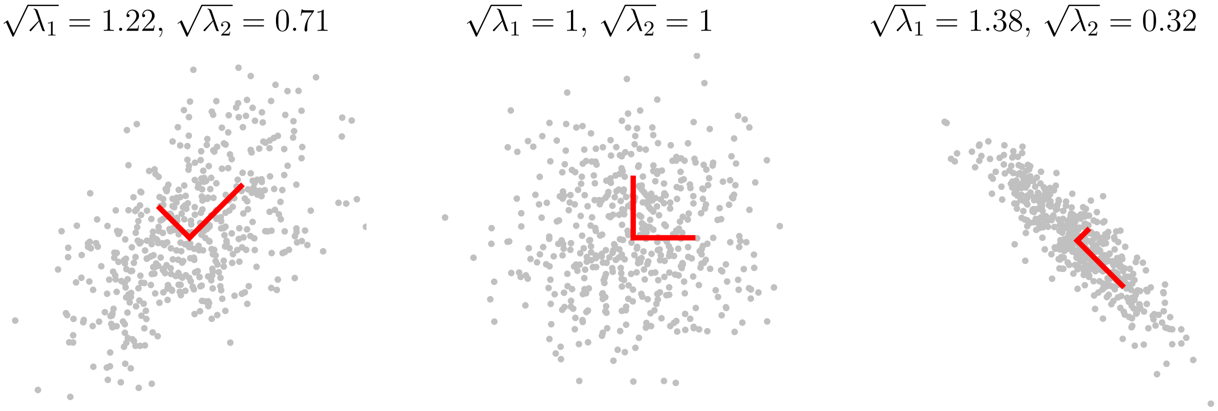


Figure 6: Samples from bivariate Gaussian random vectors with different covariance matrices are shown in gray. The eigenvectors of the covariance matrices are plotted in red. Each is scaled by the square roof of the corresponding eigenvalue λ_1 or λ_2.

of each component equals one. Equivalently, we aim to find a matrix A such that $A\vec{X}$ has a covariance matrix equal to the identity. As shown in the following lemma, this can be achieved using the eigendecomposition of the covariance matrix of $\vec{X}$.

Lemma 4.13 (Whitening)**.** *Let $\vec{X}$ be an n-dimensional random vector and let $\Sigma_X = U\Lambda U^T$ be the eigendecomposition of its covariance matrix Σ_X, which we assume to be full rank. Then all the entries of the random vector $\sqrt{\Lambda^{-1}}U^T\vec{X}$, where*

$$\sqrt{\Lambda^{-1}} := \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\lambda_2}} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \frac{1}{\sqrt{\lambda_n}} \end{bmatrix}, \tag{91}$$

are uncorrelated.

Proof. By Theorem 4.10, the covariance matrix of $\sqrt{\Lambda^{-1}}U^T\vec{X}$ equals

$$\begin{aligned} \Sigma_{\sqrt{\Lambda^{-1}}U^T\vec{X}} &= \sqrt{\Lambda^{-1}}U^T\Sigma_{\vec{X}}U\sqrt{\Lambda^{-1}} && (92) \\ &= \sqrt{\Lambda^{-1}}U^TU\Lambda U^TU\sqrt{\Lambda^{-1}} && (93) \\ &= \sqrt{\Lambda^{-1}}\Lambda\sqrt{\Lambda^{-1}} \quad \text{because } U^TU = I && (94) \\ &= I. && (95) \end{aligned}$$

□

This process is known as whitening because random vectors with uncorrelated entries are often referred to as white noise. Figure 7 shows the effect of the procedure on a Gaussian

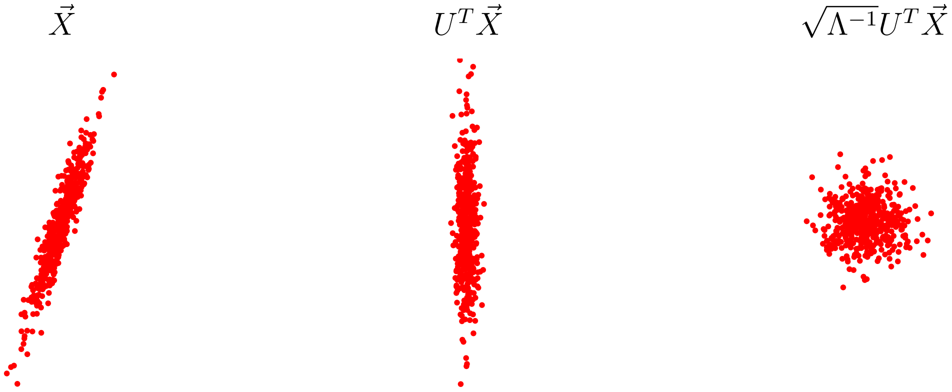


Figure 7: Samples from a bivariate Gaussian vector $\vec{X}$ with covariance matrix Σ_X (left). Samples from the transformed random variable $U^T\vec{X}$, where U is the matrix of eigenvectors of Σ_X (center). Samples from the whitened random variable $\sqrt{\Lambda^{-1}}U^T\vec{X}$, where Λ is the matrix of eigenvalues of Σ_X (right).

random vector. Applying U^T rotates the distribution to align it with the axes, making the entries uncorrelated. $\sqrt{\Lambda^{-1}}$ scales the variance of each axis to make them all equal to one.

4.5 Gaussian random vectors

You might have noticed that we have used mostly Gaussian vectors to visualize the different properties of the covariance operator. The reason is that Gaussian random vectors are completely determined by their mean vector and their covariance matrix. An important consequence, is that if the entries of a Gaussian random vector are uncorrelated then they are also mutually independent.

Lemma 4.14 (Uncorrelation implies mutual independence for Gaussian random vectors)**.** *If all the components of a Gaussian random vector $\vec{X}$ are uncorrelated, this implies that they are mutually independent.*

Proof. The parameter Σ of the joint pdf of a Gaussian random vector is its covariance matrix (one can verify this by applying the definition of covariance and integrating). If all the components are uncorrelated then

$$\Sigma_{\vec{X}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}, \tag{96}$$

where σ_i is the standard deviation of the ith component. Now, the inverse of this diagonal matrix is just

$$\Sigma_{\vec{X}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix}, \tag{97}$$

and its determinant is $|\Sigma| = \prod_{i=1}^n \sigma_i^2$ so that

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right) \tag{98}$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{(2\pi)}\sigma_i} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right) \tag{99}$$

$$= \prod_{i=1}^n f_{X_i}(x_i). \tag{100}$$

Since the joint pdf factors into a product of the marginals, the components are all mutually independent. □

5 Conditional expectation

The expectation of a function of two random variables X and Y conditioned on X taking a fixed value can be computed using the conditional pmf or pdf of Y given X.

$$\mathrm{E}(g(X,Y)|X=x) = \sum_{y\in R} g(x,y)\, p_{Y|X}(y|x), \tag{101}$$

if Y is discrete and has range R, whereas

$$\mathrm{E}(g(X,Y)|X=x) = \int_{y=-\infty}^{\infty} g(x,y)\, f_{Y|X}(y|x)\, \mathrm{d}y, \tag{102}$$

if Y is continuous.

Note that $\mathrm{E}(g(X,Y)|X=x)$ can actually be interpreted as a *function of x* since it maps every value of x to a real number. We can then define the conditional expectation of $g(X,Y)$ given X as follows.

Definition 5.1 (Conditional expectation)**.** *The conditional expectation of $g(X,Y)$ given X is*

$$\mathrm{E}(g(X,Y)|X) := h(X), \tag{103}$$

where

$$h(x) := \mathrm{E}\left(g(X,Y)\,|X = x\right). \tag{104}$$

Beware the confusing definition, the conditional expectation is actually a random variable!

Iterated expectation is a useful tool for computing expected values. The idea is that the expected value can be obtained as the expectation of the conditional expectation.

Theorem 5.2 (Iterated expectation)**.** *For any random variables X and Y and any function $g : \mathbb{R}^2 \to \mathbb{R}$*

$$\mathrm{E}\left(g(X,Y)\right) = \mathrm{E}\left(\mathrm{E}\left(g(X,Y)\,|X\right)\right). \tag{105}$$

Proof. We prove the result for continuous random variables, the proof for discrete random variables, and for quantities that depend on both continuous and discrete random variables, is almost identical. To make the explanation clearer, we define

$$\begin{aligned} h(x) &:= \mathrm{E}\left(g(X,Y)\,|X = x\right) && (106)\\ &= \int_{y=-\infty}^{\infty} g(x,y)\, f_{Y|X}(y|x)\, \mathrm{d}y. && (107)\end{aligned}$$

Now,

$$\begin{aligned} \mathrm{E}\left(\mathrm{E}\left(g(X,Y)\,|X\right)\right) &= \mathrm{E}\left(h(X)\right) && (108)\\ &= \int_{x=-\infty}^{\infty} h(x)\, f_X(x)\, \mathrm{d}x && (109)\\ &= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} f_X(x)\, f_{Y|X}(y|x)\, g(x,y)\, \mathrm{d}y\, \mathrm{d}x && (110)\\ &= \mathrm{E}\left(g(X,Y)\right). && (111)\end{aligned}$$

□

Iterated expectation allows to obtain the expectation of quantities that depend on several quantities very easily if we have access to the marginal and conditional distributions. We illustrate this with several examples taken from the previous lecture notes.

Example 5.3 (Desert (continued from Ex. 3.13 in Lecture Notes 3))**.** Let us compute the mean time at which the car breaks down, i.e. the mean of T. By iterated expectation

$$\begin{aligned}
\mathrm{E}\left(T\right) &= \mathrm{E}\left(\mathrm{E}\left(T|M,R\right)\right) && (112)\\
&= \mathrm{E}\left(\frac{1}{M+R}\right) \quad \text{because } T \text{ is exponential when conditioned on } M \text{ and } R && (113)\\
&= \int_0^1 \int_0^1 \frac{1}{m+r}\,\mathrm{d}m\,\mathrm{d}r && (114)\\
&= \int_0^1 \log\left(r+1\right) - \log\left(r\right)\,\mathrm{d}r && (115)\\
&= \log 4 = 1.39 \qquad \text{integrating by parts.} && (116)
\end{aligned}$$

Example 5.4 (Grizzlies in Yellowstone (continued from Ex. 4.3 in Lecture Notes 3))**.** Let us compute the mean weight of a bear in Yosemite. By iterated expectation

$$\begin{aligned}
\mathrm{E}\left(W\right) &= \mathrm{E}\left(\mathrm{E}\left(W|S\right)\right) && (117)\\
&= \frac{\mathrm{E}\left(W|S=1\right) + \mathrm{E}\left(W|S=1\right)}{2} && (118)\\
&= 170 \text{ kg.} && (119)
\end{aligned}$$

Example 5.5 (Bayesian coin flip (continued from Ex. 4.6 in Lecture Notes 3))**.** Let us compute the mean of the coin-flip outcome X. By iterated expectation

$$\begin{aligned}
\mathrm{E}\left(X\right) &= \mathrm{E}\left(\mathrm{E}\left(X|B\right)\right) && (120)\\
&= \mathrm{E}\left(B\right) \quad \text{because } X \text{ is Bernoulli when conditioned on } B && (121)\\
&= \int_0^1 2b^2\,\mathrm{d}b && (122)\\
&= \frac{2}{3}. && (123)
\end{aligned}$$

A Derivation of means and variances in Table 1

A.1 Bernoulli

$$\begin{aligned} \mathrm{E}\left(X\right) &= p_X\left(1\right) = p, && (124)\\ \mathrm{E}\left(X^2\right) &= p_X\left(1\right), && (125)\\ \mathrm{Var}\left(X\right) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right) = p\left(1-p\right). && (126) \end{aligned}$$

A.2 Geometric

To compute the mean of a geometric random variable, we apply Lemma B.3:

$$\begin{aligned} \mathrm{E}\left(X\right) &= \sum_{k=1}^{\infty} k\, p_X\left(k\right) && (127)\\ &= \sum_{k=1}^{\infty} k\, p\left(1-p\right)^{k-1} && (128)\\ &= \frac{p}{1-p}\sum_{k=1}^{\infty} k\left(1-p\right)^{k} && (129)\\ &= \frac{1}{p}. && (130) \end{aligned}$$

To compute the mean squared value we apply Lemma B.4:

$$\begin{aligned} \mathrm{E}\left(X^2\right) &= \sum_{k=1}^{\infty} k^2\, p_X\left(k\right) && (131)\\ &= \sum_{k=1}^{\infty} k^2\, p\left(1-p\right)^{k-1} && (132)\\ &= \frac{p}{1-p}\sum_{k=1}^{\infty} k^2\left(1-p\right)^{k} && (133)\\ &= \frac{2-p}{p^2}. && (134)\\ \mathrm{Var}\left(X\right) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right) = \frac{1-p}{p^2}. && (135) \end{aligned}$$

A.3 Binomial

By Lemma 5.8 in Lecture Notes 1, if we define n Bernoulli random variables with parameter p we can write a binomial random variable with parameters n and p as

$$X = \sum_{i=1}^{n} B_i, \tag{136}$$

where $B_1, B_2, \ldots$ are mutually independent Bernoulli random variables with parameter p. Since the mean of all the Bernoulli random variables is p, by linearity of expectation

$$\mathrm{E}\left(X\right) = \sum_{i=1}^{n} \mathrm{E}\left(B_i\right) = np. \tag{137}$$

Note that $\mathrm{E}\left(B_i^2\right) = p$ and $\mathrm{E}\left(B_i B_j\right) = p^2$ by independence, so

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \mathrm{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} B_i B_j\right) && (138)\\
&= \sum_{i=1}^{n} \mathrm{E}\left(B_i^2\right) + 2\sum_{i=1}^{n-1}\sum_{i=j+1}^{n} \mathrm{E}\left(B_i B_j\right) = np + n\left(n-1\right)p^2. && (139)\\
\mathrm{Var}\left(X\right) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right) = np\left(1-p\right). && (140)
\end{aligned}$$

A.4 Poisson

From calculus we have

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}, \tag{141}$$

which is the Taylor series expansion of the exponential function. Now we can establish that

$$\begin{aligned}
\mathrm{E}\left(X\right) &= \sum_{k=1}^{\infty} k p_X\left(k\right) && (142)\\
&= \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} && (143)\\
&= e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^{m+1}}{m!} && (144)\\
&= \lambda, && (145)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \sum_{k=1}^{\infty} k^2 p_X\left(k\right) && (146)\\
&= \sum_{k=1}^{\infty} \frac{k\lambda^k e^{-\lambda}}{(k-1)!} && (147)\\
&= e^{-\lambda}\left(\sum_{k=1}^{\infty} \frac{(k-1)\,\lambda^k}{(k-1)!} + \frac{k\lambda^k}{(k-1)!}\right) && (148)\\
&= e^{-\lambda}\left(\sum_{m=1}^{\infty} \frac{\lambda^{m+2}}{m!} + \sum_{m=1}^{\infty} \frac{\lambda^{m+1}}{m!}\right) && (149)\\
&= \lambda^2 + \lambda. && (150)\\
\mathrm{Var}\left(X\right) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right) = \lambda. && (151)
\end{aligned}$$

A.5 Uniform

We apply the definition of expected value for continuous random variables to obtain

$$\begin{aligned}
\mathrm{E}\left(X\right) &= \int_{-\infty}^{\infty} x f_X\left(x\right)\mathrm{d}x = \int_a^b \frac{x}{b-a}\mathrm{d}x && (152)\\
&= \frac{b^2-a^2}{2\left(b-a\right)} = \frac{a+b}{2}. && (153)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \int_a^b \frac{x^2}{b-a}\mathrm{d}x && (154)\\
&= \frac{b^3-a^3}{3\left(b-a\right)} && (155)\\
&= \frac{a^2+ab+b^2}{3}. && (156)\\
\mathrm{Var}\left(X\right) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right) && (157)\\
&= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{\left(b-a\right)^2}{12}. && (158)
\end{aligned}$$

A.6 Exponential

Applying integration by parts,

$$\begin{aligned}
\mathrm{E}\left(X\right) &= \int_{-\infty}^{\infty} x f_X\left(x\right) \mathrm{d}x & (159)\\
&= \int_{0}^{\infty} x\lambda e^{-\lambda x} \mathrm{d}x & (160)\\
&= x e^{-\lambda x}]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} \mathrm{d}x & (161)\\
&= \frac{1}{\lambda}. & (162)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \int_0^{\infty} x^2 \lambda e^{-\lambda x} \mathrm{d}x & (163)\\
&= x^2 e^{-\lambda x}]_0^{\infty} + 2\int_0^{\infty} x e^{-\lambda x} \mathrm{d}x & (164)\\
&= \frac{2}{\lambda^2}. & (165)\\
\mathrm{Var}\left(X\right) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right) = \frac{1}{\lambda^2}. & (166)
\end{aligned}$$

A.7 Gaussian

We apply the change of variables $t = \left(x - \mu\right)/\sigma$.

$$\begin{aligned}
\mathrm{E}\left(X\right) &= \int_{-\infty}^{\infty} x f_X\left(x\right) \mathrm{d}x & (167)\\
&= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x & (168)\\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-\frac{t^2}{2}} \mathrm{d}t + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} \mathrm{d}t & (169)\\
&= \mu, & (170)
\end{aligned}$$

where the last step follows from the fact that the integral of a bounded odd function over a symmetric interval is zero.

Applying the change of variables $t = (x - \mu)/\sigma$ and integrating by parts, we obtain that

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x \qquad (171)\\
&= \int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x \qquad (172)\\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}}\,\mathrm{d}t + \frac{2\mu\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-\frac{t^2}{2}}\,\mathrm{d}t + \frac{\mu^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,\mathrm{d}t \qquad (173)\\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(t^2 e^{-\frac{t^2}{2}}]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,\mathrm{d}t \right) + \mu^2 \qquad (174)\\
&= \sigma^2 + \mu^2. \qquad (175)\\
\mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2(X) = \sigma^2. \qquad (176)
\end{aligned}$$

B Geometric series

Lemma B.1. *For any $\alpha \neq 0$ and any integers n_1 and n_2*

$$\sum_{k=n_1}^{n_2} \alpha^k = \frac{\alpha^{n_1} - \alpha^{n_2+1}}{1-\alpha}. \qquad (177)$$

Corollary B.2. *If $0 < \alpha < 1$*

$$\sum_{k=0}^{\infty} \alpha^k = \frac{\alpha}{1-\alpha}. \qquad (178)$$

Proof. We just multiply the sum by the factor $(1-\alpha)/(1-\alpha)$ which obviously equals one,

$$\begin{aligned}
\alpha^{n_1} + \alpha^{n_1+1} + \cdots + \alpha^{n_2-1} + \alpha^{n_2} &= \frac{1-\alpha}{1-\alpha}\left(\alpha^{n_1} + \alpha^{n_1+1} + \cdots + \alpha^{n_2-1} + \alpha^{n_2}\right) \qquad (179)\\
&= \frac{\alpha^{n_1} - \alpha^{n_1+1} + \alpha^{n_1+1} + \cdots - \alpha^{n_2} + \alpha^{n_2} - \alpha^{n_2+1}}{1-\alpha} \qquad (180)\\
&= \frac{\alpha^{n_1} - \alpha^{n_2+1}}{1-\alpha} \qquad (181)
\end{aligned}$$

□

Lemma B.3. *For $0 < \alpha < 1$*

$$\sum_{k=1}^{\infty} k\,\alpha^k = \frac{1}{(1-\alpha)^2}. \qquad (182)$$

Proof. By Corollary B.2,

$$\sum_{k=0}^{\infty} \alpha^k = \frac{1}{1-\alpha}. \tag{183}$$

Since the left limit converges, we can differentiate on both sides to obtain

$$\sum_{k=0}^{\infty} k\alpha^{k-1} = \frac{1}{(1-\alpha)^2}. \tag{184}$$

□

Lemma B.4. *For* $0 < \alpha < 1$

$$\sum_{k=1}^{\infty} k^2\, \alpha^k = \frac{\alpha\,(1+\alpha)}{(1-\alpha)^3}. \tag{185}$$

Proof. By Lemma B.3,

$$\sum_{k=1}^{\infty} k^2\alpha^k = \frac{\alpha\,(1+\alpha)}{(1-\alpha)^3}. \tag{186}$$

Since the left limit converges, we can differentiate on both sides to obtain

$$\sum_{k=1}^{\infty} k^2\alpha^{k-1} = \frac{1+\alpha}{(1-\alpha)^3}. \tag{187}$$

□

C Proof of Theorem 4.7

If $\mathrm{E}\,(X^2) = 0$ then $X = 0$ by Corollary 3.13 $X = 0$ with probability one, which implies $\mathrm{E}\,(XY) = 0$ and consequently that equality holds in (65). The same is true if $\mathrm{E}\,(Y^2) = 0$.

Now assume that $\mathrm{E}\,(X^2) \neq 0$ and $\mathrm{E}\,(Y^2) \neq 0$. Let us define the constants $a = \sqrt{\mathrm{E}\,(Y^2)}$ and $b = \sqrt{\mathrm{E}\,(X^2)}$. By linearity of expectation,

$$\begin{aligned}
\mathrm{E}\left((aX+bY)^2\right) &= a^2\mathrm{E}\left(X^2\right) + b^2\mathrm{E}\left(Y^2\right) + 2\,a\,b\,\mathrm{E}\,(XY) &(188)\\
&= 2\left(\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) + \sqrt{\mathrm{E}\,(X^2)\,\mathrm{E}\,(Y^2)}\mathrm{E}\,(XY)\right), &(189)\\
\mathrm{E}\left((aX-bY)^2\right) &= a^2\mathrm{E}\left(X^2\right) + b^2\mathrm{E}\left(Y^2\right) - 2\,a\,b\,\mathrm{E}\,(XY) &(190)\\
&= 2\left(\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) - \sqrt{\mathrm{E}\,(X^2)\,\mathrm{E}\,(Y^2)}\mathrm{E}\,(XY)\right). &(191)
\end{aligned}$$

The expectation of a non-negative quantity is nonzero because the integral or sum of a non-negative quantity is non negative. Consequently, the left-hand side of (188) and (190) is non-negative, so (189) and (191) are both non-negative, which implies (65).

Let us prove (67) by proving both implications.

($\Rightarrow$). Assume $\mathrm{E}\left(XY\right) = -\sqrt{\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right)}$. Then (189) equals zero, so

$$\mathrm{E}\left(\left(\sqrt{\mathrm{E}\left(X^2\right)}X + \sqrt{\mathrm{E}\left(X^2\right)}Y\right)^2\right) = 0, \tag{192}$$

which by Corollary 3.13 means that $\sqrt{\mathrm{E}\left(Y^2\right)}X = -\sqrt{\mathrm{E}\left(X^2\right)}Y$ with probability one.

($\Leftarrow$). Assume $Y = -\frac{\mathrm{E}\left(Y^2\right)}{\mathrm{E}\left(X^2\right)}X$. Then one can easily check that (189) equals zero, which implies $\mathrm{E}\left(XY\right) = -\sqrt{\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right)}$.

The proof of (66) is almost identical (using (188) instead of (189)).