CS 4980/6980: Data Science © Spring 2018

Lecture 17: Estimating and Manipulating Covariance

Instructor: Daniel L. Pimentel-Alarcón Scribed by: Paul Trimor and Jeevitha Meyyappan

This is preliminary work and has not been reviewed by instructor. If you have comments about typos, errors, notation inconsistencies, etc., please email the scribers.

17.1 Mean, Variance, and Covariance

In the previous lecture, we learned that the mean, variance, and covariance of random variable(s) can be expressed using Expected value as follows:

Remark 17.1. The mean, variance, and covariance of random variable(s)

$$mean(x) = \mu_{\mathbf{x}} = \mathbb{E}[x] \tag{17.1}$$

$$var(x) = \sigma^2_{\mathbf{x}} = \mathbb{E}[(x - \mu_x)^2] \tag{17.2}$$

$$cov(x, y) = \sigma^2_{\mathbf{x},\mathbf{y}} = \mathbb{E}[(x - \mu_x)(y - \mu_y)] \tag{17.3}$$

We also learned that the boundry of the covariance can be implied given the following rules:

1. If $x \perp y$, then $cov(x, y) = 0$
2. If $x = y$, then $cov(x, y) = var(x)$
3. Therefore $0 \leq cov(x, y) \leq max(var(x), var(y))$

However, we only know the boundry. How can we estimate the covariance of x and y if $x \not\perp y$ and $x \neq y$?

The best way to conceptualize this problem is to think of the mean, variance, and covariance of a random variable(s) as a value that is outside the scope of human induction. Determining these values requires knowledge of all possible outcomes of the random variable(s). In other words, only God knows.

17.2 The Issue

The issue is that we are mere mortals with a limited scope of understanding. The best we can do is approximate the mean, variance, and covariance by taking samples of data.

Given a random variable x, a sample can be represented as follows.

$$\mathbf{x} = \left[\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, ...\mathrm{x}_n\right]$$

Where n is finite.

We can assume that the more samples we take, the closer our approximation converges to the Expected value of our random variable.

$$\frac{1}{N}\sum_{i=1}^{N} f(\mathrm{x}_i) \longrightarrow \mathbb{E}[f(x)]$$

Using the sample of a random variable and the assumption above, we can explicitly represent our pervious equations as follows:

Remark 17.2. The mean, variance, and covariance of a sample

$$mean(\mathbf{x}) = \widehat{\mu}_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{x}_i \tag{17.4}$$

$$var(\mathbf{x}) = \widehat{\sigma}^2_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} (\mathrm{x}_i - \widehat{\mu}_{\mathrm{x}})^2 \tag{17.5}$$

$$cov(\mathbf{x}, \mathbf{y}) = \widehat{\sigma}^2_{\mathbf{x},\mathbf{y}} = \frac{1}{N}\sum_{i=1}^{N} (\mathrm{x}_i - \hat{\mu}_{\mathrm{x}})(\mathrm{y}_i - \widehat{\mu}_{\mathbf{y}}) \tag{17.6}$$

Note: We use the ˆ symbol to denote estimations:

These three equations is how we can go back and forth between the abstract idea of random variables to practical equations for which we can apply real data to. This is also how we conquer our goal. But now that we have these equations, how can we play God and generate numbers that correspond to a mean, variance, and covariance that we choose. The following homework exerice will explore this.

17.3 Homework - Playing God

For this homework, we must generate pairs of scalar values xs and ys that are correlated according to the covariance that we have determined.

In mathematical terms, how can I generate pairs $(\mathrm{x}_1, \mathrm{y}_1), (\mathrm{x}_2, \mathrm{y}_2), ..., (\mathrm{x}_n, \mathrm{y}_n)$ such that:

$$\mu_{\mathbf{x}} = 2$$
$$\mu_{\mathbf{y}} = 4$$
$$\sigma^2_{\mathbf{x}} = 5$$
$$\sigma^2_{\mathbf{y}} = 7$$

But most important,

$$\sigma^2_{\mathbf{x},\mathbf{y}} = 3$$

Note: These values have been chosen arbirarily. Remember, we are playing God.

Manipulating the mean and the variance is easy. Assuming $randn(1, N)$ is a function that generates a random matrix where (1,N) is the size of matrix. Here the random matrices $\mathbf{X}$ and $\mathbf{Y}$ are

$$\mathbf{X} = randn(1, N)$$
$$\mathbf{Y} = randn(1, N)$$

,

To manipulate the mean and the variable of a variable, we just need to manipulate the normal distribution equation, $N(0, 1)$ where 0 is the mean and 1 is the variance.

$$\mathbf{X} = randn(1, 1) \longleftarrow x \sim N(0, 1)$$

To manipulate the mean, we simply shift our value from the mean by addition or subtraction.

Case 1: Here ($\mathbf{X}$+2) is distributed as Gaussian N(2,1),

$$(\mathbf{X} + 2) \sim N(0 + 2, 1)$$
$$(\mathbf{X} + 2) \sim N(2, 1)$$

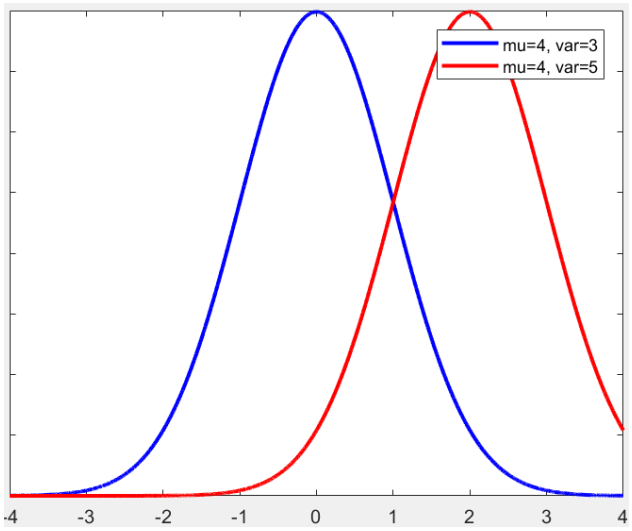

figure 1 *shows that the* μ_x *shifted by* 2.

To manipulate the variance, we just stretch or squeeze the mean through multiplication. Continuing from our previous case:

Case 2: Here $\sqrt{5}$($\mathbf{X}$+2) is distributed as Gaussian N(2,5),

$$\sqrt{5}(\mathbf{X} + 2) \sim N(2, 5)$$

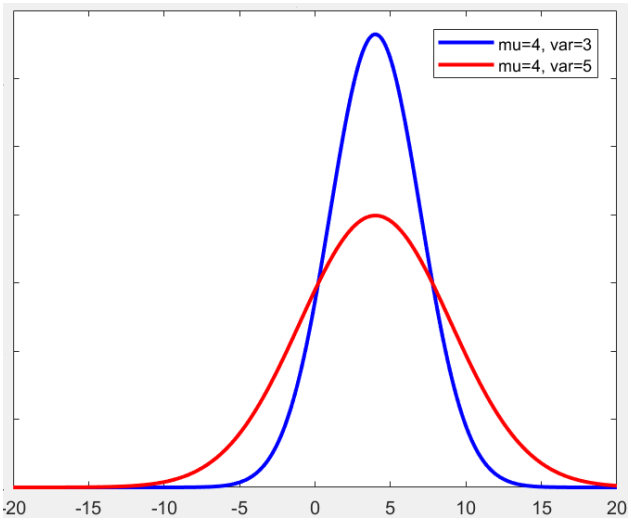

figure 2 shows that the μ_x shifted by 4, with variance 5.

Case 3: Here $\sqrt{7}(\mathbf{Y}+4)$ is distributed as Gaussian N(2,7),

$$\sqrt{7}(\mathbf{Y}+4) \sim N(4,7)$$

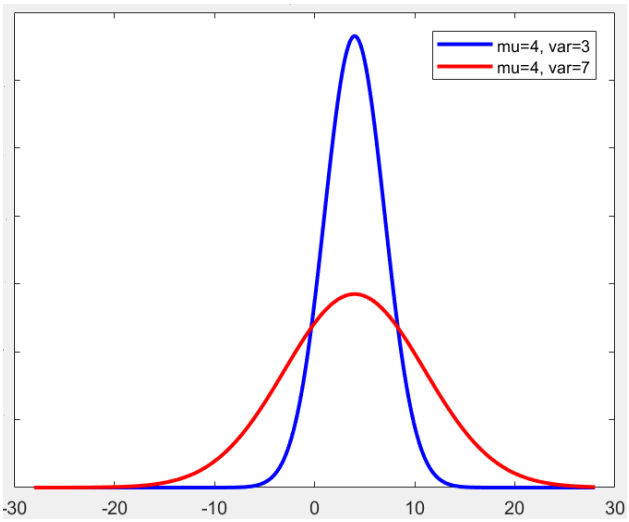

figure 3 shows that the μ_x shifted by 4, with variance 7.

Let us formulize the tricks in these cases as follows:

Tricks to Transform Random Variables

Trick 1. if x has mean μ_x, then $x+k$ has a mean $\mu_x + k$

Trick 2. if x has variance σ^2, then cx has variance $c^2\sigma_x^2$

However, this is not enough. We want to assign the covariance between two variables. To do this we must encapsulate the two variables of interest in a vector.

Lets encapsulate our two random variables x and y into a vector Z

$$Z = \begin{bmatrix} x \\ y \end{bmatrix}$$

If we extend our ideas from the first part, we have these rules:

Tricks to transform Random Vectors

Trick 3. if Z has mean μ_z, then $Z + K$ has a mean $\mu_z + K$

Trick 4. if Z has a cavariance matix C, then Az has a covariance ACA^T

The Mean μ_z of the vector Z is,

$$\mu_z = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$$

So the Covariance matrix can be deduced as,

$$C = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$$

The next question then is, how do we come up with matrix A? inorder to give us our desired C.

17.4 Conclusion

Random variables describe the probability of all possible outcomes of a given event. Random variables are divine, in the sense that one must be omnipotent to know the true mean, variance, and covariance of a random variable. However, we are humans, not Gods, the best we can do is to estimate. Inorder to estimate the mean, variance, and covariance of random variables, one must take samples from these variables. This takes us out of the realm of the divine into the realm of the physical.

We also learned how to play God and manipulate the mean, variance, and covariance of random variables. Manipulating the mean and variance is just a matter of arithmetic. Therefore, to relate two random variables, we must use a vector.