18.5.3

Mean, Variance, and Covariance

Now that you convinced Martha that feature extraction is the way to go, she needs some background on why this works in case questions come up during her presentation on how she can "magically" combine these features in a meaningful way. To start, you dust off your stats knowledge and refresh your memory on mean, variance, and covariance. These will be the building blocks used for PCA.

There is a mathematical way to use feature extraction, but first let's review some stats concepts.

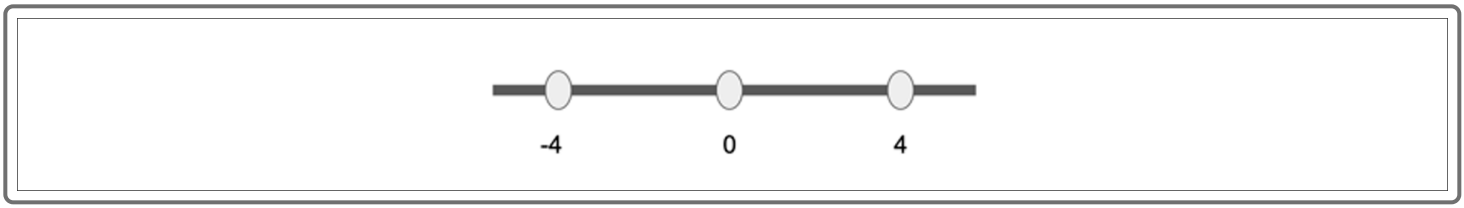
Mean

Recall that the **mean** is the sum of a group of numbers divided by the total amount of numbers. For example, we start with points 2, 3, and 7. First, we add up all the numbers: 2 + 3 + 7 = 12. Then we divide the result by the total amount of points, which is 3. So, 12 / 3 = 4, so the mean of those three points is 4.

Variance

Variance is the square distance from each point from the center, added together, and divided by the total number of points. The variance measures the spread of a set of numbers. The center of the points may look familiar, and it should, because it is the mean of all the points. Variance, in other words, is a measure of how far apart the data points are from the mean.

Look at the following points on a line:



Using 0 as the center point, the distances are -4 from the center, 0 from the center (the center point is still a point), and 4 from the center.

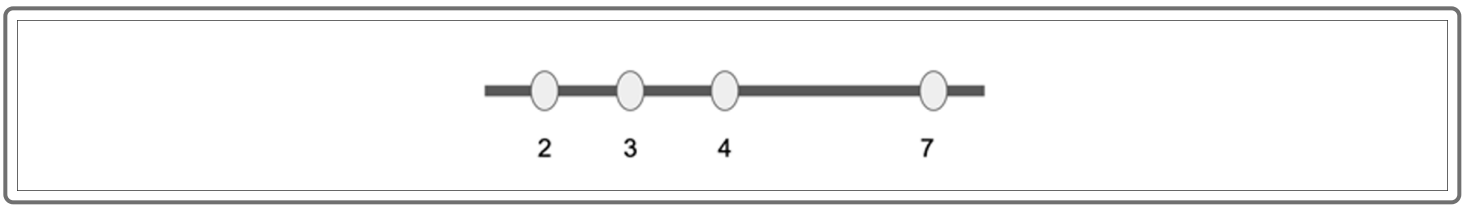
The sum of squared distances would be (-4)^2 + (0)^2 + (4)^2 = 16 + 0

- 16 = 32. We use squared distance so they are all positive.

Divide by the total number of points, which is 3. The variance of this dataset would be 32/3, or 10⅔.

Normally, there won't be an even distribution of points around the center. The points 2, 3, and 7 from the previous example don't have a clear center.

This is where the mean comes into play. The center of the line is set to the mean, which we found to be 4. Here is what the points look like on a line:



The distance from 4 to 2 is -2, the distance from 4 to 3 is -1, the distance from 4 to 4 is 0, and the distance from 4 to 7 is 3.

Add up the squares of each distance: (-2)^2 + (-1)^2 + (0)^2 + (3)^2 = 4 + 1 + 0 + 9 = 14.

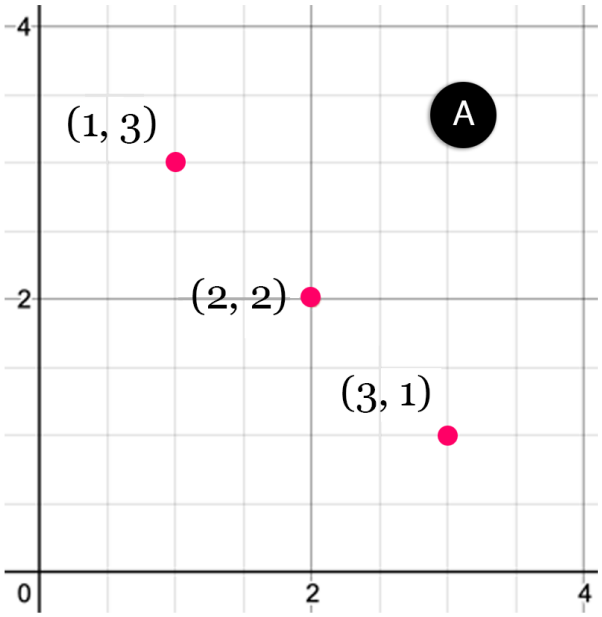
Finally, divide the distances by the total number of points: 14 / 3. The variance equals 14/3, or 4⅔.

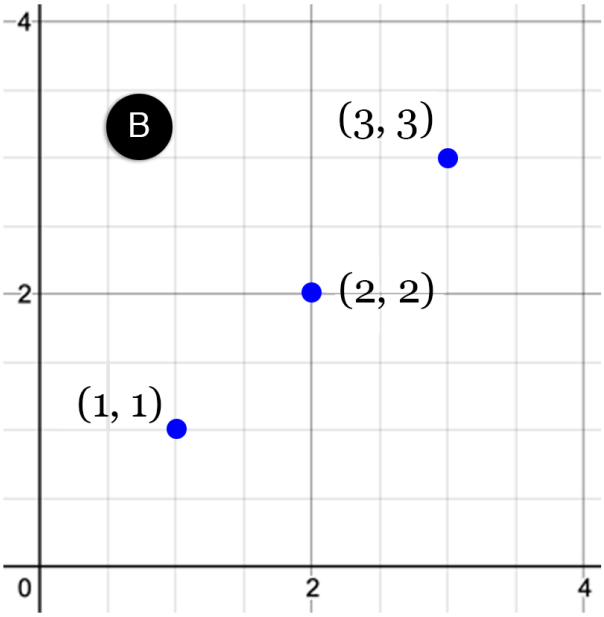
NOTE

These examples showed points on the x-axis, and thus, form the x variance. The same process applies to elements on the y-axis, forming the y variance.

Covariance

Before defining what covariance is, look at the following two plots:





These two plots clearly are very different. Each has the same center, with different points on the left and the right, one sloping negatively and the other sloping positively.

Let's find the x and y variance for each line.

For graph A:

- The center point is (2, 2).
- The distances for the points are the distance from (2, 2).

- Point (1, 3) is a distance of -1 away on the x-axis and 1 on the y-axis.
- Point (3, 1) is a distance of 1 away on the x-axis and -1 on the y-axis.
- x variance = (-1)^2 + 0^2 + (1)^2 = 2 / 3
- y variance = (1)^2 + 0^2 + (-1)^2 = 2 / 3

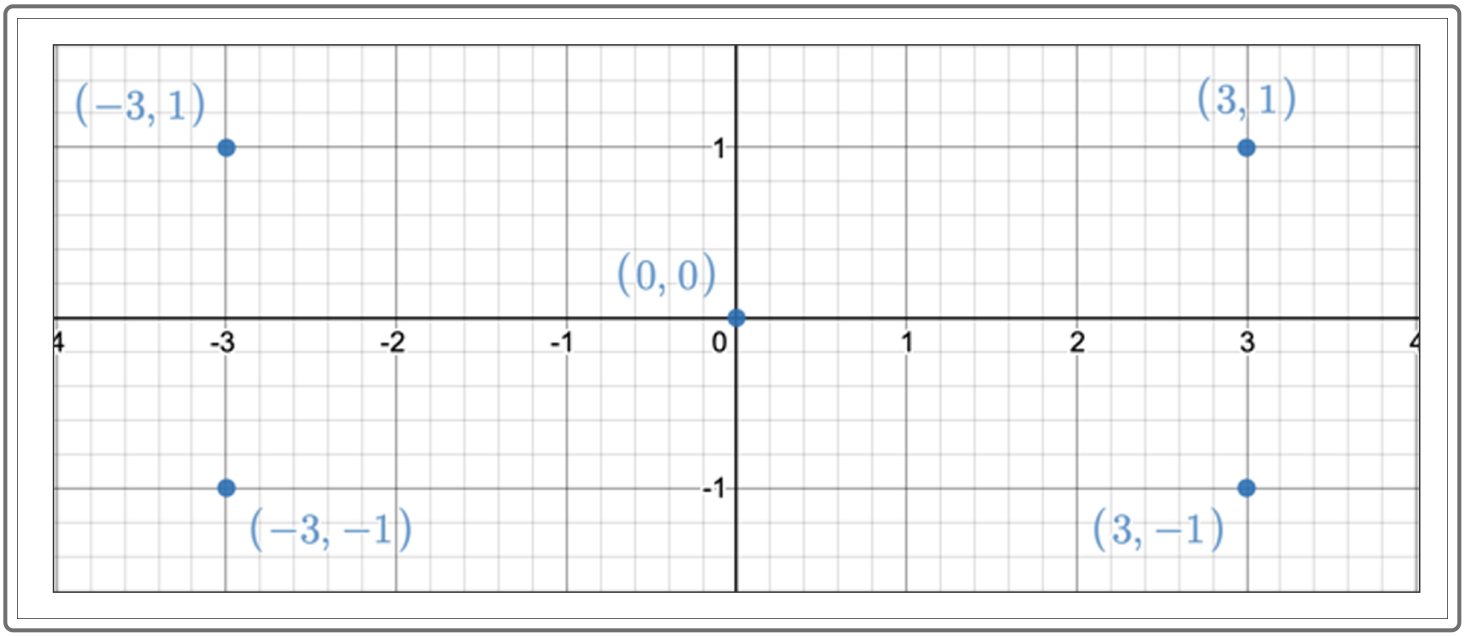
For graph B:

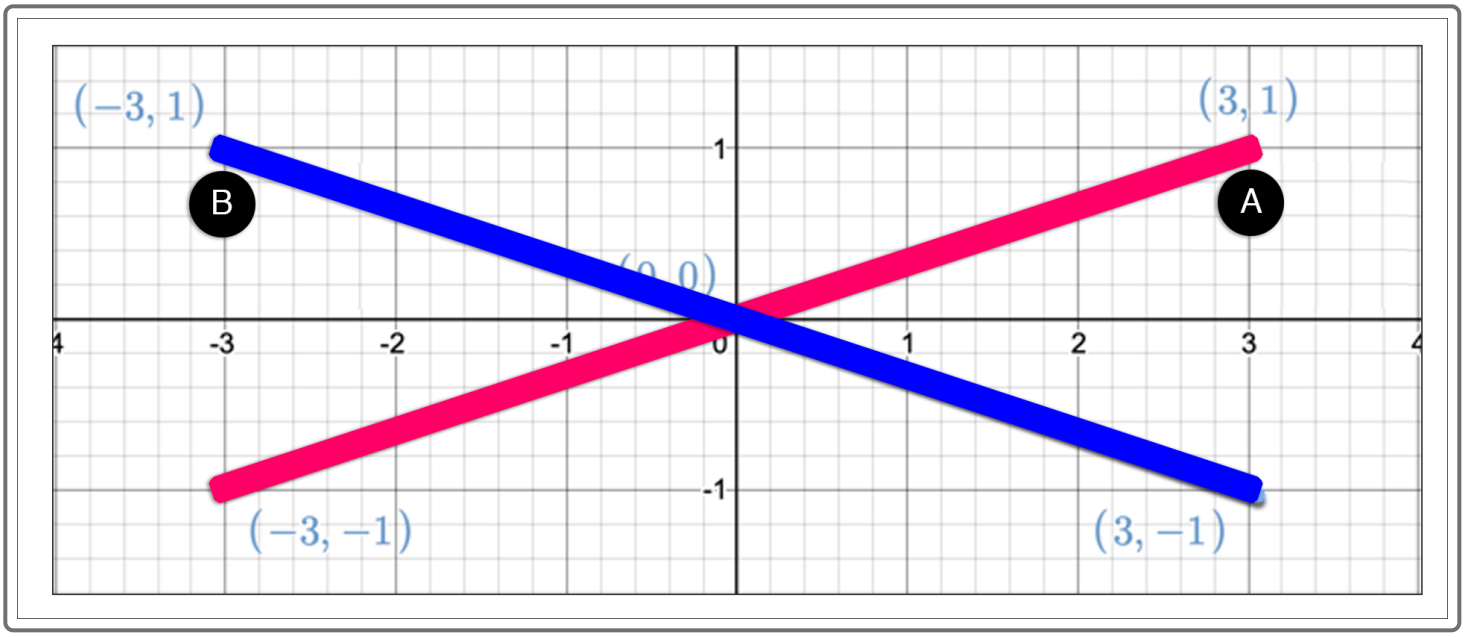
- The center point is (2, 2).
- The distances for the points are the distance from (2, 2).
- Point (1, 1) is a distance of -1 away on the x-axis and -1 on the y-axis.
- Point (3, 3) is a distance of 1 away on the x-axis and 1 on the y-axis.
- x variance = (-1)^2 + 0^2 + (1)^2 = 2 / 3
- y variance = (-1)^2 + 0^2 + (1)^2 = 2 / 3

Wait. Both of these variances are exactly the same; however, it is very obvious that these two graphs are totally different! How can we tell the difference?

This is where covariance comes into play. **Covariance** is a metric that allows us to tell these two different sets of points apart.

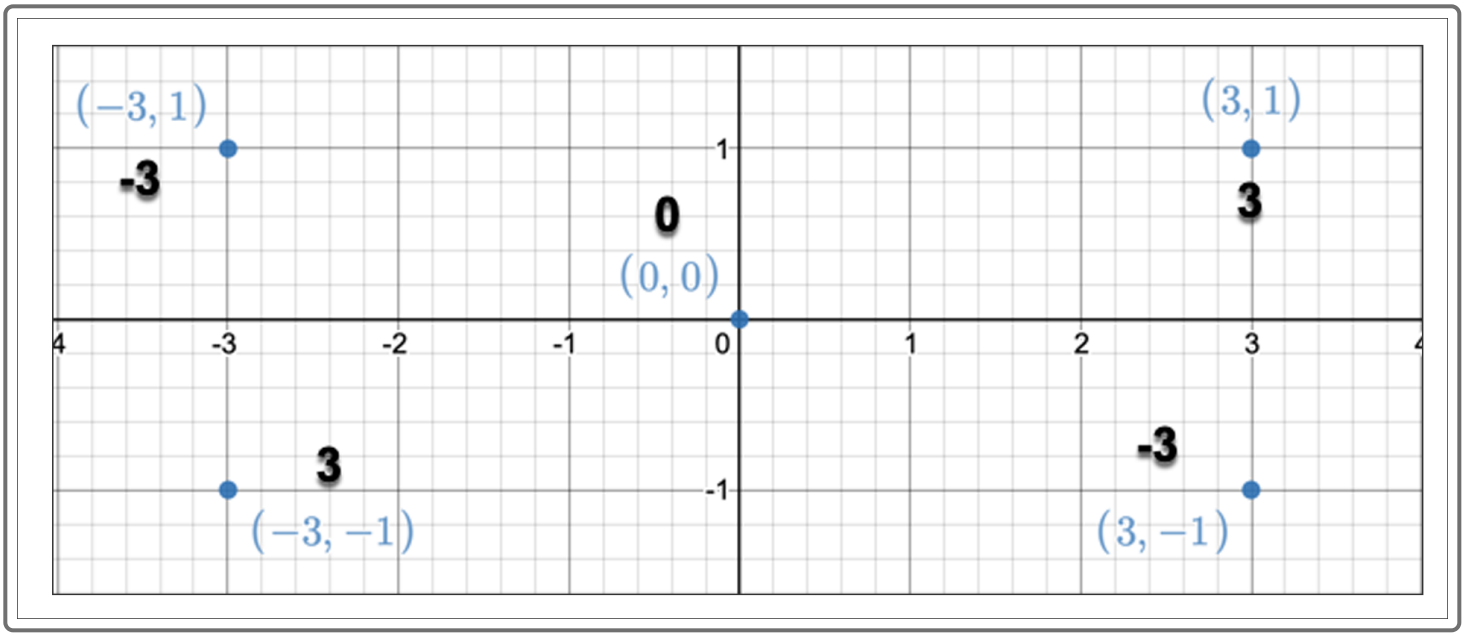
Let's look at the following examples:





How can we tell the difference between the points that lie along Line A versus the points that lie along Line B?

We can do this with the product of coordinates, which is the multiple of each of the two points:



Covariance is the sum of the product of coordinates divided by the number of points.

Covariance is used to determine the relationship between points.

The formula for covariance is as follows:

$$Cov\,(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\mathbb{N}}$$

What this equation is saying is that the covariance takes the sum of the product between each pair of coordinates and their difference from the mean divided by the total number of points. This may sound complicated but will make more sense once we look at an example.

Let's solve for the covariance of line A first which contains the points (-3, -1), (0, 0) and (3, 1).

First take the mean of the x coordinates in line A, -3 + 0 + 3 = 0 divided by 3 is zero. Then repeat for the y coordinates, -1 + 0 + 1 = 0 divided by 3 is also zero.

Then for each pair of coordinates find the difference between the point and their respective means.

X	**Y**	$(x_i - \bar{x})$	$(y_i - \bar{y})$
-3	-1	-3 - 0 = -3	-1 - 0 = -1
0	0	0 - 0 = 0	0 - 0 = 0
3	1	3 - 0 = 3	1 - 0 = 1

Now multiply the results of the coordinate pairs.

X	**Y**	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
-3	-1	-3 - 0 = -3	1 - 0 = -1	3
0	0	0 - 0 = 0	0 - 0 = 0	0
3	1	3 - 0 = 3	1 - 0 = 1	3

Finally add the product of all the coordinated paris and divide by the number of points to find the covariance.

3 + 0 + 3 = 6

Plug the results into the top part of the equation, and since we know there or 3 points, we plug that in for N to get.

$$Cov(x, y) = \frac{6}{3}$$

Reduce the equation.

$$Cov(x, y) = 2$$

The covariance for line A is 2.

Repeat the same process for line B would produce the following:

X	Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
-3	1	-3 - 0 = -3	1 - 0 = 1	-3
0	0	0 - 0 = 0	0 - 0 = 0	0
3	-1	3 - 0 = 3	1 - 0 = -1	-3

Add the product of all the coordinated pairs.

-3 + 0 + -3 = -6

Plug the results into the top part of the equation, and again we know there are 3 points, we plug that in for N to get.

$$Cov(x, y) = -\frac{6}{3}$$

Reduce the equation.

$$Cov(x, y) = -2$$

The covariance for line A is -2.

The covariance for Line A is 2 while the covariance for Line B is -2.

We can then say that Line A has a positive covariance (at 2) while Line B has a negative covariance (at -2). There is also a third type of covariance called **covariance zero**. This is when the points tend to form a horizontal line.

NOTE

Covariance is used to only describe the relationship between points, such as positive and negative as we just saw. You may recall another method for determining relationships is correlation. However, correlation is used to determine the strength of the relationship.