Week 6 Homework

Math 510 Christina Morgenstern

---

1. Compute the covariance matrix for $X = 4, 2, -1, 1, 5$ and $Y = 2, 3, 3, 5, 6$.

$$Cov(X,Y) = \begin{bmatrix} cov(X,X) & cov(X,Y) \\ cov(Y,X) & cov(Y,Y) \end{bmatrix}$$

$$\bar{X} = \frac{4+2+(-1)+1+5}{5} = \frac{11}{5} = 2.2$$

$$\bar{Y} = \frac{2+3+3+5+6}{5} = \frac{19}{5} = 3.8$$

$$cov(X,X) = \sigma^2(X) = \frac{3.24+0.04+10.24+1.44+7.84}{4} = \frac{22.8}{4} = 5.7$$

$$cov(Y,Y) = \sigma^2(Y) = \frac{3.24+0.64+0.64+1.44+4.84}{4} = \frac{10.8}{4} = 2.7$$

$$cov(X,Y) = \frac{(4-2.2)(2-3.8)+(2-2.2)(3-3.8)+(-1-2.2)(3-3.8)+(1-2.2)(5-3.8)+(5-2.2)(6-3.8)}{4} = \frac{-3.24+0.16+2.56-1.44+6.16}{4} = \frac{4.2}{4} = 1.05$$

$$Cov(X,Y) = \begin{bmatrix} 5.7 & 1.05 \\ 1.05 & 2.7 \end{bmatrix}$$

2. Give an example of a real world data set that you would expect to have a low standard deviation (relative to the size of the values). What about one with a high standard deviation? Explain why these examples fit.

The standard deviation represents a number, that tells us how spread out the data of a group are from the mean value. A low standard deviation means, that most of the values are close to the average. A high standard deviation means, that the values are spread apart.

Considering an example from sports e.g. soccer. The teams that are ranked at the top perform better throughout the season, win many games and show more consistency. Thus, the standard deviation for the winner team is low.
Teams with a higher standard deviation are less predictable and have good games and bad games ending up somewhere in the middle.

Another example is the amount of salary in a company. The salary might show great variability from very low salaries for interns and very high salaries for CEOs. The standard deviation for this pool of data is high because the data is spread apart.
On the other hand, if you consider people with the same tasks and similar positions within the company, like a trainee, then the variance between the salaries is low because you expect them to earn similar amounts of money. The standard deviation in this case is low.

3. Give an example of a real world dataset that you would expect to have a covariance near 1. What about one with a covariance near -1? What about one with a covariance near zero? Explain why each dataset fits.

In general, the covariance is a measure of how much two random variables vary together.

A covariance of -1 means a strong negative relation (if *X* goes up, *Y* goes down): An example is the predator-prey-interaction. If the number of predators increases, the number of prey decreases because there are more predators that hunt down the prey. Vice versa if the population of predators decreases because there are not enough prey the population of prey increases because it can recover.

A covariance of 1 means a strong positive relation (if *X* goes up, *Y* goes up): An example would be greenhouse gas emissions and atmospheric temperature. If the emissions of greenhouse gases rise, the atmospheric temperature rises as well.

If the covariance is near zero, then it´s likely that the two variables are independent. However, there might be still a non-linear relationship between the two. An example would be eating pizza and speaking Italian. The amount of pizzas I am eating will unlikely have an impact (positive or negative) on my ability speaking Italian.

4. Perform a principal component analysis (PCA) on the following data. Interpret the answers - what do they mean? (Feel free to use technology to compute your solution.)

Let 2.0, 1.8, 1.5, 1.2, 1.0, 0.8, 0.5, 0.2 be the distance (in miles) from a train station and let 19, 21, 30, 54, 61, 82, 83, 102 be the average income (in thousands of dollars) for a person living at each distance away from the train station.

Table 1. Data organized in table. Calculation of average values.

	Distance (in miles)	**Average income (in thousands of dollars)**
	2.0	19
	1.8	21
	1.5	30
	1.2	54
	1.0	61
	0.8	82
	0.5	83
	0.2	102
Average	**1.125**	**56.5**

Calculation of covariance matrix:

$cov(X,X) = \sigma^2(X) = \frac{2.735}{7} = 0.39$

$cov(Y,Y) = \sigma^2(Y) = \frac{6818}{7} = 974$

$Cov(X,Y) = \begin{bmatrix} 0.39 & -19.2 \\ -19.2 & 974 \end{bmatrix}$

Calculation of eigenvectors and eigenvalues (using www.symbolab.com; The values for the eigenvectors and eigenvalues are rounded to two decimal places)

Eigenvectors for $\begin{bmatrix} 0.39 & -19.2 \\ -19.2 & 974 \end{bmatrix}$:

We get an eigenvector $\begin{bmatrix} -0.02 \\ 1 \end{bmatrix}$ with an eigenvalue $\lambda = 974.38$

We get an eigenvector $\begin{bmatrix} 50.73 \\ 1 \end{bmatrix}$ with an eigenvalue $\lambda = 0.01$

From looking at the data, I can see, that there is a relationship between the distance a person lives from the train station and their income. The further away a person lives from the train station, the lower her income is.

I tried different scenarios with the eigenvectors to come up with values that correspond to my hypothesis.

I have chosen the first eigenvector as the principal component. I can multiply it with 10, because each multiple of the eigenvector is an eigenvector as well. I have chosen to multiply with 10, because it reflects the distance values more realistically.

$$\begin{bmatrix} -0.02 \\ 1 \end{bmatrix} * 10 = \begin{bmatrix} -0.2 \\ 10 \end{bmatrix}$$

$$\begin{bmatrix} -0.2 \\ 10 \end{bmatrix} corresponds\ to \begin{bmatrix} distance\ in\ miles \\ average\ income\ (in\ thousands\ of\ \$) \end{bmatrix}$$

My interpretation of the PCA using the eigenvector $\begin{bmatrix} -0.2 \\ 10 \end{bmatrix}$:
For every 0.2 miles, that a person lives from the train station, their salary is cut for about 10 thousand USD.