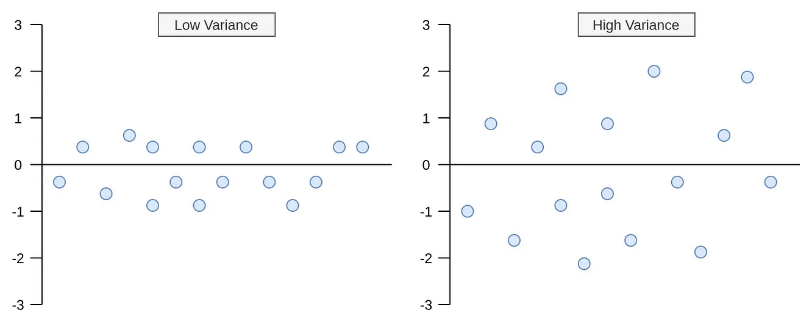
COMS20011 - Data-Driven Computer Science

Problem Sheet 2 - covariances, eigenvectors, eigenvalues, normalisation

February 2024

1. Sketch a simple scatterplot to show low variance in an example data set, and then another plot to show high variance in another dataset.

   Answer:
   Since we are plotting variance, and not covariance, then the amount of variance is only in one dimension. An example set of plots are these:



2. You are given this simple 2D dataset:

   X = (1 2 3 4 5)
   Y = (2 4 6 8 10)

   Work out the covariance matrix of this dataset (by hand!).

   Answer:
   The covariance can be calculated as follows. The mean of X and Y can be calculated as:

$$\bar{X} = \frac{1+2+3+4+5}{5} = 3$$
$$\bar{Y} = \frac{2+4+6+8+10}{5} = 6$$

   The deviation of X and Y from their mean can be calculated as:

$$X_i - \bar{X} = [1,2,3,4,5] - 3 = [-2,-1,0,1,2]$$
$$Y_i - \bar{Y} = [2,4,6,8,10] - 6 = [-4,-2,0,2,4]$$

   Next, calculate the variances of X and Y:

$$
\begin{aligned}
\textit{Var}(X) &= \frac{\sum_{i=1}^{5}(X_i - \bar{X})^2}{5-1} \\
&= \frac{(-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2}{5-1} \\
&= \frac{4+1+0+1+4}{4} \\
&= 2.5 \\
\textit{Var}(Y) &= \frac{\sum_{i=1}^{5}(Y_i - \bar{Y})^2}{5-1} \\
&= \frac{(-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (4)^2}{5-1} \\
&= \frac{16+4+0+4+16}{4} \\
&= 10
\end{aligned}
$$

Now, calculate the covariance:

$$
\begin{aligned}
\textit{Cov}(X,Y) &= \frac{\sum_{i=1}^{5}(X_i - \bar{X})(Y_i - \bar{Y})}{5-1} \\
&= \frac{(-2)(-4) + (-1)(-2) + (0)(0) + (1)(2) + (2)(4)}{5-1} \\
&= \frac{8+2+0+2+8}{4} \\
&= 5
\end{aligned}
$$

The covariance matrix can finally be states as:

$\begin{bmatrix} \textit{Var}(X) & \textit{Cov}(X,Y) \\ \textit{Cov}(X,Y) & \textit{Var}(Y) \end{bmatrix} = \begin{bmatrix} 2.5 & 5 \\ 5 & 10 \end{bmatrix}$

3. Find the eignevalues and eigenvectors of:

$$A = \begin{bmatrix} 1 & 6 \\ 5 & 2 \end{bmatrix}$$

Answer:
To solve for the eigenvalues and eigenvectors of a covariance matrix, we need to find the roots of the characteristic equation.

Using $det(A - \lambda I) = 0$*, where I is the 2x2 identity matrix:*

$|A - \lambda I| = \begin{bmatrix} 1-\lambda & 6 \\ 5 & 2-\lambda \end{bmatrix} = (1-\lambda)(2-\lambda) - (5 \cdot 6) = 0$

Thus:

$(1-\lambda)(2-\lambda) - 30 = 0 \;\Rightarrow\; \lambda^2 - 3\lambda - 28 = 0 \Rightarrow (\lambda - 7)(\lambda + 4) = 0$

Solving for λ*, then the eigenvalues are* $\lambda_1 = 7$ *and* $\lambda_2 = -4$

For $\lambda_1 = 7$*, using* $Av = \lambda v$*, the eigenvector v is any non-zero scalar multiple of* $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

For $\lambda = -4$, using $Av = \lambda v$, the eigenvector v is any non-zero scalar multiple of $\begin{bmatrix} 1 \\ -5/6 \end{bmatrix}$.

4. Calculate the normalised versions of the following vectors:

   (a) $V = (2, 1, 5)$
   (b) $W = (5, -3, 8, 2)$
   (c) $Z = (2, 2, 2, 2)$

   Answer:
   The normalised vector is a vector in the same direction but with norm (length) 1. It is obtained by dividing each element of the vector by its magnitude.

   (a) The magnitude or length of vector V is $\sqrt{(2^2 + 1^2 + 5^2)} = \sqrt{30}$. The normalised vector of V would then be
   $V' = (2/\sqrt{30}, 1/\sqrt{30}, 5/\sqrt{30})$ *or* $V' = (0.3651, 0.1826, 0.9129)$

   (b) The magnitude or length of vector W is $\sqrt{(5^2 + (-3)^2 + 8^2 + 2^2)} = \sqrt{102}$. The normalised vector of W would then be
   $W' = (5/\sqrt{102}, -3/\sqrt{102}, 8/\sqrt{102}, 2/\sqrt{102})$.

   (c) The magnitude or length of vector Z is $\sqrt{(2^2 + 2^2 + 2^2 + 2^2)} = \sqrt{16} = 4$. The normalised vector of W would then be
   $Z' = (0.5, 0.5, 0.5, 0.5)$.

5. After a class test, the teacher computes the mean score to be 60 and the standard deviation of the test scores to be 10. If Jack's test score was 48 and Jill's was 75, what would be their individual z-score?

   Answer:

   The z-score standardisation formula is $x' = \frac{(x-\mu)}{\sigma}$. Here x is the test score.

   For Jack, the z-score for a test score of 48 would be (48 - 60) / 10 = -1.2. This means that his test score of 48 is 1.2 standard deviations below the mean.

   For Jill, the z-score for a test score of 75 would be (75 - 60) / 10 = 1.5. This means that her test score of 75 is 1.5 standard deviations above the mean.

6. After measuring the height of all the players in its football team, a club finds that the minimum height is 155cm, while the maximum height is 197cm (probably the goalkeeper!).

   The team manager wants to rescale these heights to lie between 0 and 1.

   (a) What is the rescaled height of the central-defender in the team who is 189cm tall?

   (b) Argentinian player Messi is 169cm tall. What is his rescaled height in this range?

   (c) What would be Messi's rescaled height, if the new range was to be between 1 and 10?

   Answer:
   For a height of 189cm, the rescaled value would be:
   $\frac{(189-155)}{(197-155)} = 0.809$.

(b) For Messi's height of 169cm, the rescaled value would be:
$\frac{(169-155)}{(197-155)} = 0.333$.

(c) The rescaling range is not always between 0 to 1. Let's say it's between A and B. Then one can rescale to the 0 to 1 range at first and then multiply the result by the new range difference $(B-A)$ and add the starting value of the range to shift it into that range, hence

$x' = A + (B-A) \times \frac{x-min(x)}{max(x)-min(x)}$

So in this example, where the new scale range is 1 to 10, Messi's rescaled height is $1 + 9 \times 0.333 = 4$.

7. Find the eignevalues and eigenvectors of:

$$C = \begin{bmatrix} -2 & -4 & 2 \\ -2 & 1 & 2 \\ 4 & 2 & 5 \end{bmatrix}$$

Answer:

To solve for the eigenvalues and eigenvectors of a covariance matrix, we need to find the roots of the characteristic equation.

Using $\det(C - \lambda I) = 0$, where I is the 3x3 identity matrix:

$\left|C - \lambda I\right| = \left| \begin{bmatrix} -2 & -4 & 2 \\ -2 & 1 & 2 \\ 4 & 2 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right| = 0$

Expanding the determinant and solving for λ:

$$\begin{vmatrix} -2-\lambda & -4 & 2 \\ -2 & 1-\lambda & 2 \\ 4 & 2 & 5-\lambda \end{vmatrix} = 0 \quad \Rightarrow \quad \lambda^3 - 4\lambda^2 - 27\lambda + 90 = 0$$

Simplifying this to:

$(\lambda - 3)(\lambda^2 - \lambda - 30) = 0,$ *and then* $(\lambda - 3)(\lambda + 5)(\lambda - 6) = 0,$

we get the eigenvalues as:

$$\lambda_1 = 3,\ \lambda_2 = -5,\ \lambda_3 = 6$$

Use $Cv = \lambda v$ to work out the eigenvectors. You should get the following corresponding eigenvectors:

$\begin{bmatrix} 1 \\ \frac{-3}{2} \\ \frac{-1}{2} \end{bmatrix}, \quad \begin{bmatrix} 1 \\ \frac{-1}{2} \\ \frac{1}{2} \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 6 \\ 16 \end{bmatrix}.$