CMSC5724: Exercise List 8

Problem 1. Consider the following 2D dataset:

What is the variance of the projections of the 6 points onto the line $y = x$?

Answer. The projections are as shown below:

The projection of, for example, $a$ is $a'$, which we will regard as a 1D value, equal to the negated distance $-3\sqrt{2}$ from $a'$ to the origin (it is negated because it is below the y-axis). Similarly, the projections $b', c', d', e'$ and $f'$ can be regarded as 1D values: $-2\sqrt{2}, -\sqrt{2}/2, \sqrt{2}, 3\sqrt{2}/2$ and $3\sqrt{2}$, respectively. The mean of the six 1D values is 0; hence, their variance is:

$$\frac{1}{6}((-3\sqrt{2} - 0)^2 + (-2\sqrt{2} - 0)^2 + (-\sqrt{2}/2 - 0)^2 + (\sqrt{2} - 0)^2 + (3\sqrt{2}/2 - 0)^2 + (3\sqrt{2} - 0)^2) = (18 + 8 + 0.5 + 2 + 4.5 + 18)/6$$

$$= 51/6$$

Problem 2. What is the co-variance matrix of the dataset in Problem 1?
Answer. The co-variance matrix is $A = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}$, where $\sigma_{xx}$ ($\sigma_{yy}$) is the variance along the x- (y-) dimension, and $\sigma_{xy}$ ($= \sigma_{yx}$) is the covariance of the x- and y-dimensions. We calculate:

\[
\sigma_{xx} = \frac{((-4)^2 + (-3)^2 + (-1)^2 + (2)^2 + (2)^2 + (4)^2)/6}{16} = 25/3
\]

\[
\sigma_{yy} = \frac{((-2)^2 + (-1)^2 + (0)^2 + (0)^2 + (1)^2 + (2)^2)/6}{4 + 1 + 0 + 1 + 4} = 5/3
\]

\[
\sigma_{xy} = \frac{((-4)(-2) + (-3)(-1) + (-1)0 + 2 \cdot 0 + 2 \cdot 1 + 4 \cdot 2)/6}{8 + 3 + 0 + 2 + 8} = 3.5
\]

Hence, $A = \begin{bmatrix} 25/3 & 3.5 \\ 3.5 & 5/3 \end{bmatrix}$.

Problem 3. Use PCA to find the line passing the origin on which the projections of the points in Problem 1 have the greatest variance.

Answer. We first compute the eigenvectors $v$ of the covariance matrix $A$. Specifically, there should be a non-zero value $\lambda$ such that:

\[
Av = \lambda v \iff \begin{bmatrix} 25/3 & 3.5 \\ 3.5 & 5/3 \end{bmatrix} v = \lambda v \iff \begin{bmatrix} 25/3 - \lambda & 3.5 \\ 3.5 & 5/3 - \lambda \end{bmatrix} v = 0
\]

Setting the determinant of $\begin{bmatrix} 25/3 - \lambda & 3.5 \\ 3.5 & 5/3 - \lambda \end{bmatrix}$ to 0 gives:

\[
(25/3 - \lambda)(5/3 - \lambda) = 3.5^2
\]

which has two solutions (a.k.a. eigenvalues):

\[
\lambda_1 = 59/6 \\
\lambda_2 = 1/6
\]

The line that we are looking for (i.e., the one capturing the most variance of the projections) is given by an eigenvector $v = [x, y]$ corresponding to the larger eigenvalue $\lambda_1$. To find $v$, we fit $\lambda_1 = 59/6$ into Equation 1 (after simplification):

\[
\begin{bmatrix} -1.5 & 3.5 \\ 3.5 & -49/6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0
\]

namely:

\[-1.5x + 3.5y = 0
\]

It suffices to find any solution of $x$ and $y$, e.g.:

\[
x = 3.5 \\
y = 1.5
\]

The line we are looking for is therefore the one passing origin and point (3.5, 1.5), as shown below:
Problem 4. Consider that we use FASTMAP to reduce the 2D dataset in Problem 1 to a set of 1D values. Recall that this algorithm essentially finds a line to project all the points on. What is this line?

Answer. The algorithm first picks an arbitrary point $p_1$, e.g., $p_1 = c$. Then, it chooses another point $p_2$ to be the one in the dataset farthest from $p_1$, in our case $p_2 = f$. Next, it re-selects $p_1$ as the point in the dataset farthest from $p_2$, in our case $p_1 = a$. Finally, it determines the line to be the one passing $a$ and $f$.

You can verify yourself that the result remains the same if $p_1$ was initialized to be another point.