COMP527
Data Mining and Visualisation
Problem Set 3

Danushka Bollegala

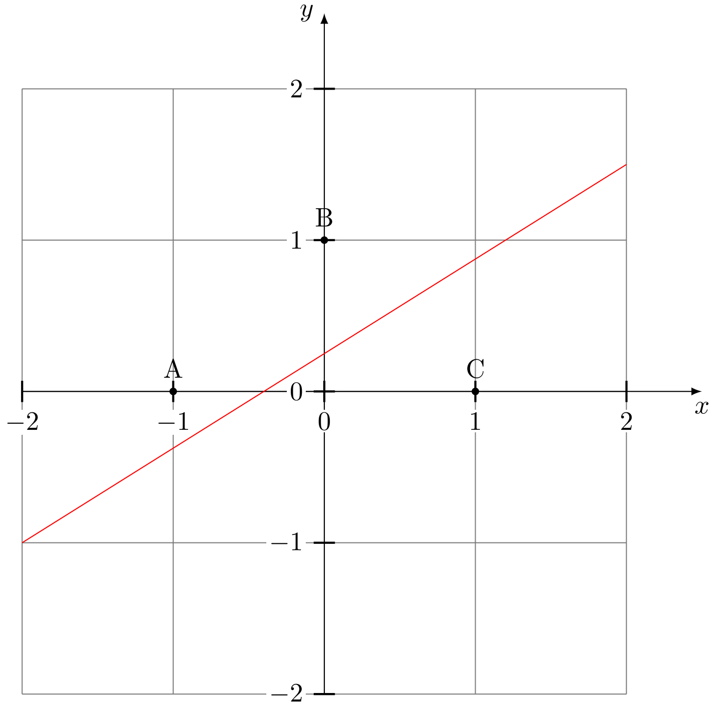

Figure 1: Projecting three points A, B, C onto the line $y = mx + c$.

Question Consider the problem of projecting a two-dimensional dataset consisting of three points $A = (-1, 0), B = (0, 1)$, and $C = (1, 0)$ onto the one-dimensional line given by $y = mx + c$. The dataset and the line is shown in Figure 1. Answer the following questions.

A. Compute the co-ordinates of the base of the perpendicular from point (α, β) to line $y = mx + c$. **(10 marks)**

Assume the base of the coordinates to be (p, q). Then it must satisfy the following equations as it is on the line $y = mx + c$ and the line section connecting base to (α, β) is perpendicular (hence gradient of $-1/m$).

$$q = mp + c \tag{1}$$

$$\frac{q - \beta}{p - \alpha} = -\frac{1}{m} \tag{2}$$

Solving these equations we get,

$$p = \frac{\alpha + m\beta - mc}{1 + m^2}$$

and

$$q = \frac{m\alpha + m^2\beta + c}{1 + m^2}.$$

B. Compute the perpendicular distance to the line $y = mx + c$ from point (α, β). **(10 marks)**

The distance d is given by,

$$d^2 = (p - \alpha)^2 + (q - \beta)^2.$$

By substituting for (p, q) from the previous question we get,

$$d = \frac{|\alpha m - \beta + c|}{\sqrt{1 + m^2}}$$

.

C. Show that if $y = mx + c$ is a solution to the one dimensional PCA projection, then $y = mx + c'$ is also a solution. Here, $c \neq c'$. **(10 marks)**

Any line parallel to $y = mx + c$ will have the same distances between the corresponding pairs of projected points. Therefore, the variance of the distances between projected points be independent of c. Because the variance does not change, its maximiser is also unaffected by c. Therefore, any parallel line of $y = mx+c$ is also a solution. In particular, we can set $c = 0$ for the rest of the questions to simplify the calculations and find the value of m that maximises variance of the projected points.

Alternative approach (maximising the squared pairwise distance between the projected points): *As noted in the lecture, a third alternative for obtaining PCA is to maximise the pairwise distance of the projected points given as follows:*

$$d^2 = (A'B')^2 + (A'C')^2 + (B'C')^2 \tag{3}$$

$$= \frac{(1+m)^2 + (m+m^2)^2 + (m-1)^2 + (m^2-m)^2 + 4 + 4m^2}{(1+m^2)^2} \tag{4}$$

$$= 2 + \frac{4}{m^2+1} \tag{5}$$

Therefore, d^2 is maximised by $m = 0$.

D. Find m such that the variance of the projected points on to the straight line is maximised. **(20 marks)**

Let us assume the projections of A, B, C onto $y = mx + c$ are given by A', B', C'. Let us denote the projections of A, B, C onto the line by A', B', C' given by:

$$A' = \left(\frac{-1}{1+m^2}, \frac{-m}{1+m^2}\right), \quad (6)$$

$$B' = \left(\frac{m}{1+m^2}, \frac{m^2}{1+m^2}\right), \quad (7)$$

$$C' = \left(\frac{1}{1+m^2}, \frac{m}{1+m^2}\right). \quad (8)$$

The mean of A, B, C is $\mu = (0, 1/3)$ and its projection on the line is $\mu' = \left(\frac{m/3}{m^2+1}, \frac{m^2/3}{m^2+1}\right)$. We can select any point on the line as the reference point for measuring distance. However, if we use μ' for this purpose it will simplify the calculations because the distance to μ' will be zero. Therefore, the variance v is given by

$$v = A'\mu'^2 + B'\mu'^2 + C'\mu'^2 \quad (9)$$

$$= \frac{(m+3)^2 + 4m^2 + (m-3)^2}{9(m^2+1)} \quad (10)$$

$$= \frac{2m^2+6}{3m^2+3} \quad (11)$$

$$= \frac{2}{3} + \frac{4}{3(m^2+1)} \quad (12)$$

Note that v achieves its maximum value of 2 when $m = 0$.

E. Find m such that the sum of squared projection errors is minimised. **(20 marks)**

The squared projection errors are respectively $AA'^2 = CC'^2 = \frac{m^2}{m^2+1}$ and $BB'^2 = \frac{1}{m^2+1}$. Therefore, the sum of squared projection errors is:

$$d^2 = AA'^2 + BB'^2 + CC'^2 \quad (13)$$

$$= \frac{m^2+1+m^2}{m^2+1} \quad (14)$$

$$= 2 - \frac{1}{m^2+1}. \quad (15)$$

Note that we have set $c = 0$ following question (C). Therefore, d is minimised by setting $m = 0$, giving the solution $y = c$. In this case, the minimum squared error is 1. (Note that although the values of m that maximises variance and minimises error are the same the actual optimal objective values are not equal.)

F. Compute the covariance matrix for this dataset. **(10 marks)**

Let $\boldsymbol{x}_1 = (-1, 0)^\top, \boldsymbol{x}_2 = (1, 0)^\top, \boldsymbol{x}_3 = (0, 1)^\top$*. Then their mean* $\boldsymbol{\mu} = (0, 1/3)^\top$*. Variances are computed as,*

$$(\boldsymbol{x}_1 - \boldsymbol{\mu})(\boldsymbol{x}_1 - \boldsymbol{\mu})^\top = \begin{bmatrix} 1 & 1/3 \\ 1/3 & 1/9 \end{bmatrix}, \quad (16)$$

$$(\boldsymbol{x}_2 - \boldsymbol{\mu})(\boldsymbol{x}_2 - \boldsymbol{\mu})^\top = \begin{bmatrix} 1 & -1/3 \\ -1/3 & 1/9 \end{bmatrix}, \quad (17)$$

$$(\boldsymbol{x}_3 - \boldsymbol{\mu})(\boldsymbol{x}_3 - \boldsymbol{\mu})^\top = \begin{bmatrix} 0 & 0 \\ 0 & 4/9 \end{bmatrix}. \quad (18)$$

Adding those three matrices and dividing by 3 *we get the covariance matrix* $\mathbf{S}$ *as follows:*

$$S = \begin{bmatrix} 2/3 & 0 \\ 0 & 2/9 \end{bmatrix}$$

G. Find the eigenvalues and eigenvectors of the covariance matrix. **(10 marks)**

Eigenvalue equation for $\mathbf{S}$ *is:*

$$\mathbf{S}\boldsymbol{\theta} = \lambda\boldsymbol{\theta} \quad (19)$$

$$|\mathbf{S} - \lambda\mathbf{I}| = 0 \quad (20)$$

From which we get $\lambda = 2, 2/3$*. The corresponding eigenvectors are respectively* $(1, 0)^\top$ *and* $(0, 1)^\top$*.*

H. Find the PCA projection using the eigenvalue decomposition of the covariance matrix. **(10 marks)**

For PCA we must select the eigenvector corresponding to the largest eigenvalue as it maximises the variance of the projected data points. Therefore, we select $(1, 0)^\top$*, which means that we select the* x*-axis, giving the solution* $y = c$*.*