COMP527
Data Mining and Visualisation
Problem Set 3

Danushka Bollegala

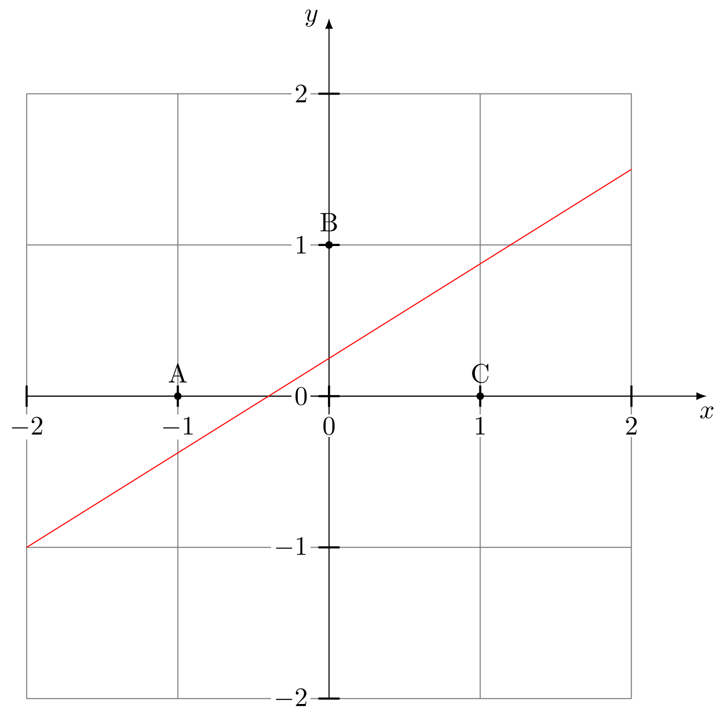

Figure 1: Projecting three points A, B, C onto the line $y = mx + c$.

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 variations between the corresponding pairs of projected points. Therefore, any parallel line would be a solution. Therefore, we can set $c = 0$ for the rest of the question and find the solution for m. Then any line $y = mx + c$ for the found m will be a solution for any $c \in \mathbb{R}$.

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'. Then the variation v is given by,

$$v = A'B'^2 + A'B'^2 + B'C'^2$$

Substituting for the base points

$$A' = \left(\frac{-1 - mc}{1 + m^2}, \frac{-m + c}{1 + m^2}\right), \tag{3}$$

$$B' = \left(\frac{1 - mc}{1 + m^2}, \frac{m^2 + c}{1 + m^2}\right), \tag{4}$$

$$C' = \left(\frac{1 - mc}{1 + m^2}, \frac{m + c}{1 + m^2}\right). \tag{5}$$

we get, $v = \frac{2m^2+6}{m^2+1} = 2 + \frac{4}{m^2}$ Therefore, v is maximised when $m = 0$, giving the line $y = c$.

E. Find m such that the projection error is minimised. **(20 marks)**

$$d^2 = AA'^2 + BB'^2 + CC'^2 \tag{6}$$

$$= \frac{m^2 + 1 + m^2}{m^2 + 1} \tag{7}$$

$$= 2 - \frac{1}{m^2 + 1}. \tag{8}$$

Note that we have set $c = 0$ following question 3. Therefore, d is minimised by setting $m = 0$, giving the solution $y = c$.

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,

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

$$(\boldsymbol{x}_2 - \boldsymbol{\mu})(\boldsymbol{x}_2 - \boldsymbol{\mu})^\top = \begin{bmatrix} 1 & -1/3 \\ -1/3 & 1/9 \end{bmatrix}, \tag{10}$$

$$(\boldsymbol{x}_3 - \boldsymbol{\mu})(\boldsymbol{x}_3 - \boldsymbol{\mu})^\top = \begin{bmatrix} 0 & 0 \\ 0 & 4/9 \end{bmatrix}. \tag{11}$$

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

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

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

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

$$\mathbf{S}\boldsymbol{\theta} = \lambda\boldsymbol{\theta} \tag{12}$$

$$|\mathbf{S} - \lambda\mathbf{I}| = 0 \tag{13}$$

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* $y = c$*.*