QUESTION 5

Imagine we extend the study on Informatics students to have 4 variables, with the following summary statistics.

Variable	Mean	Sample standard deviation
Drink (Tea -1, Coffee - 9)	5	2
Language (Haskell - 1, Java - 9)	6	2
Platform (Teams - 1, Collaborate - 9)	3	1
Data science (Hates - 1, Loves - 9)	7	2

We ensure that all of the data is standardised and run PCA on the data to give the following loadings of the first two principal components:

Variable	PC1 loadings	PC2 loadings
Drink	0.5	0.5
Language	0.5	-0.5
Platform	-0.5	0.5
Data Science	-0.5	-0.5

A student has the following characteristics:

- Drink: 7
- Language: 6
- Platform: 1
- Data Science: 9

Indicate the coordinates of their principal component scores in the plot.

If we convert this problem into the notation used in the lecture notes, we get the following:

Data vector: $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 7 \\ 6 \\ 1 \\ 9 \end{pmatrix}$ (Drink) (Language) (Platform) (D.S.)

Sample means $\begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \\ 3 \\ 7 \end{pmatrix}$

Sample standard deviations

$$\begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 1 \\ 2 \end{pmatrix}$$

Principal component loadings:

PC 1:

$$\underline{p}_1 = \begin{pmatrix} p_{11} \\ p_{21} \\ p_{31} \\ p_{41} \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix}$$

PC 2:

$$\underline{p}_2 = \begin{pmatrix} p_{12} \\ p_{22} \\ p_{32} \\ p_{44} \end{pmatrix} = \begin{pmatrix} 0.5 \\ -0.5 \\ 0.5 \\ -0.5 \end{pmatrix}$$

First standardise the data vector using the formula

$$z_i = \frac{x_i - \bar{x}_i}{s_i}$$

$$\Rightarrow \quad \underline{z} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix} = \begin{pmatrix} (7-5)/2 \\ (6-6)/2 \\ (1-3)/1 \\ (9-7)/2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -2 \\ 1 \end{pmatrix}$$

Now use the formulae in the lectures to compute the principal component scores:

$$t_1 = p_{11} z_1 + p_{21} z_2 + p_{31} z_3 + p_{41} z_4$$

$$= 0.5 \times 1 + 0.5 \times 0 - 0.5 \times (-2) - 0.5 \times 1$$

$$= 0.5 + 0 + 1 - 0.5$$

$$t_1 = 1$$

$$t_2 = p_{12} z_1 + p_{22} z_2 + p_{32} z_3 + p_{42} z_4$$

$$= 0.5 \times 1 - 0.5 \times 0 + 0.5 \times (-2) - 0.5 \times 1$$

$$= 0.5 - 0 - 1 - 0.5$$

$$t_2 = -1$$

Hence the coordinates of the first two PC scores are $(t_1, t_2) = (1, -1)$.

Note: in the lecture notes we refer to the data point $\underline{x}_i$ rather than $\underline{x}$ (without the subscript i). Here we have dropped the i index throughout.

QUESTION 6

This question uses the same data and principal component loadings as the question above. We are given the following principal component scores for a student:

- PC1 = 2
- PC2 = -1

What is the value of the student's preference for data science that we can reconstruct from these scores?

Here we assume that PC3 and PC4 are both equal to zero. We therefore have the principal component scores:

$$t_1 = 2$$
$$t_2 = -1$$
$$t_3 = 0$$
$$t_4 = 0$$

We convert from PC scores to standardised data using the following formula:

$$z_j = p_{j1} t_1 + p_{j2} t_2 + p_{j3} t_3 + p_{j4} t_4$$

We want the preference for data science, i.e. x_4

$$z_4 = p_{41} t_1 + p_{42} t_2 + p_{43} t_3 + p_{44} t_4$$

$$= -0.5 \times 2 - 0.5 \times (-1) + p_{43} \times 0 + p_{44} \times 0$$
$$= -1 + 0.5$$
$$= -0.5$$

Now convert back to the original variables:

$$x_4 = \bar{x}_4 + s_4 z_4 = 7 + 2 \times (-0.5) = 6$$

$\Rightarrow$ Student's preference for data science is 6.

Note on notation for PCA

We have used component-wise notation for these solutions. In matrix notation we define the principal component matrix as

$$P = (\underline{p}_1 \quad \underline{p}_2 \quad \underline{p}_3 \quad \underline{p}_4)$$

The component scores are computed using the transpose of the principal component matrix:

$$\underline{t} = P^T \underline{z}$$

to compute the standardised data from the scores, we invert the equation:

$$\underline{z} = P \underline{t}$$

Note that because P is an orthogonal matrix that $P^{-1} = P^T$

In question 6, P contained only two columns:

$$\tilde{P} = (\underline{p}_1 \quad \underline{p}_2)$$

Thus $\underline{z} = t_1 \underline{p}_1 + t_2 \underline{p}_2$