MANIPAL SCHOOL OF INFORMATION SCIENCES
MANIPAL
(A constituent unit of MAHE, Manipal)

AML5102 | Applied Machine Learning | PCA Problem Set

1. Consider the direction $\mathbf{u} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$. Using the image template below where the samples in the data matrix

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix}$$

are shown,

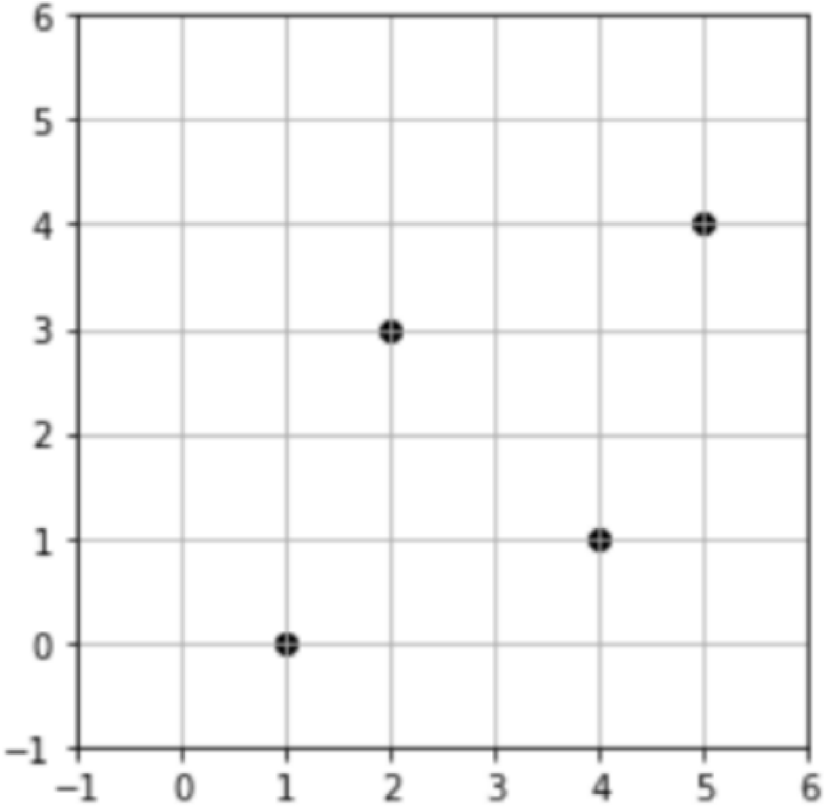
   (a) identify the samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}$, and $\mathbf{x}^{(4)}$ in the image (counting starts from 1);

   (b) clearly draw the direction $\mathbf{u}$;

   (c) clearly show the projections of all samples in the data matrix $\mathbf{X}$ onto $\mathbf{u}$;

   (d) identify which two samples are nearest and farthest from each other after the projections.

   (e) Calculate the variance of the samples after the projections.


6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6

2. At the beginning of the 20th century, one researcher obtained measurements on seven physical characteristics for each of 3000 convicted male criminals. The characteristics he measured are:

   X_1: length of head from front to back (in cm.)

   X_2: head breadth (in cm.)

   X_3: face breadth (in cm.)

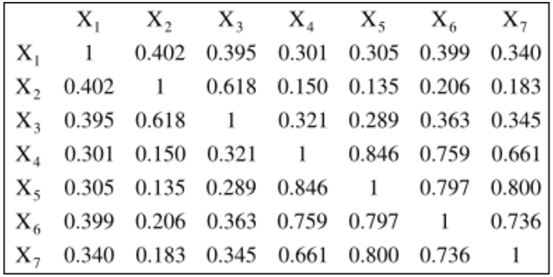
   X_4: length of left forefinger (in cm.)

   X_5: length of left forearm (in cm.)

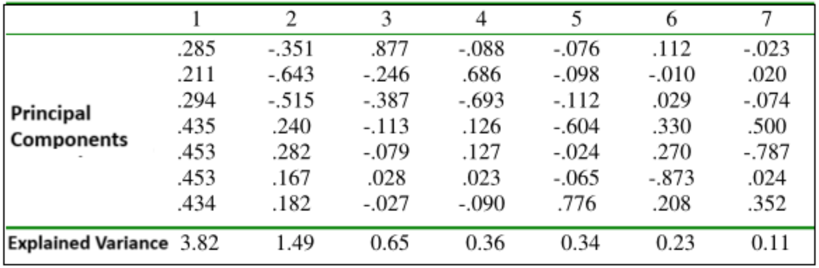
   X_6: length of left foot (in cm.)

   X_7: height (in inches)

   The sample correlation matrix, eigenvalues, and eigenvectors of the sample correlation matrix are shown below:



	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1	0.402	0.395	0.301	0.305	0.399	0.340
X_2	0.402	1	0.618	0.150	0.135	0.206	0.183
X_3	0.395	0.618	1	0.321	0.289	0.363	0.345
X_4	0.301	0.150	0.321	1	0.846	0.759	0.661
X_5	0.305	0.135	0.289	0.846	1	0.797	0.800
X_6	0.399	0.206	0.363	0.759	0.797	1	0.736
X_7	0.340	0.183	0.345	0.661	0.800	0.736	1



	1	2	3	4	5	6	7
Principal Components	.285	-.351	.877	-.088	-.076	.112	-.023
	.211	-.643	-.246	.686	-.098	-.010	.020
	.294	-.515	-.387	-.693	-.112	.029	-.074
	.435	.240	-.113	.126	-.604	.330	.500
	.453	.282	-.079	.127	-.024	.270	-.787
	.453	.167	.028	.023	-.065	-.873	.024
	.434	.182	-.027	-.090	.776	.208	.352
Explained Variance	3.82	1.49	0.65	0.36	0.34	0.23	0.11

(a) What is the shape of the data matrix?

(b) Length of the left forearm has the highest correlation with which other feature?

(c) What proportion of variance is explained by the first principal component?

(d) How many minimum principal components are needed to explain more than 90% of the variance in the data?

(e) Which two features are identically loaded for calculating the 1st principal component score?

(f) Which principal component assigns the greatest weight (in magnitude) to head breadth?

(g) In plain English, interpret what the PC-3 score of a sample captures in reference to the data.