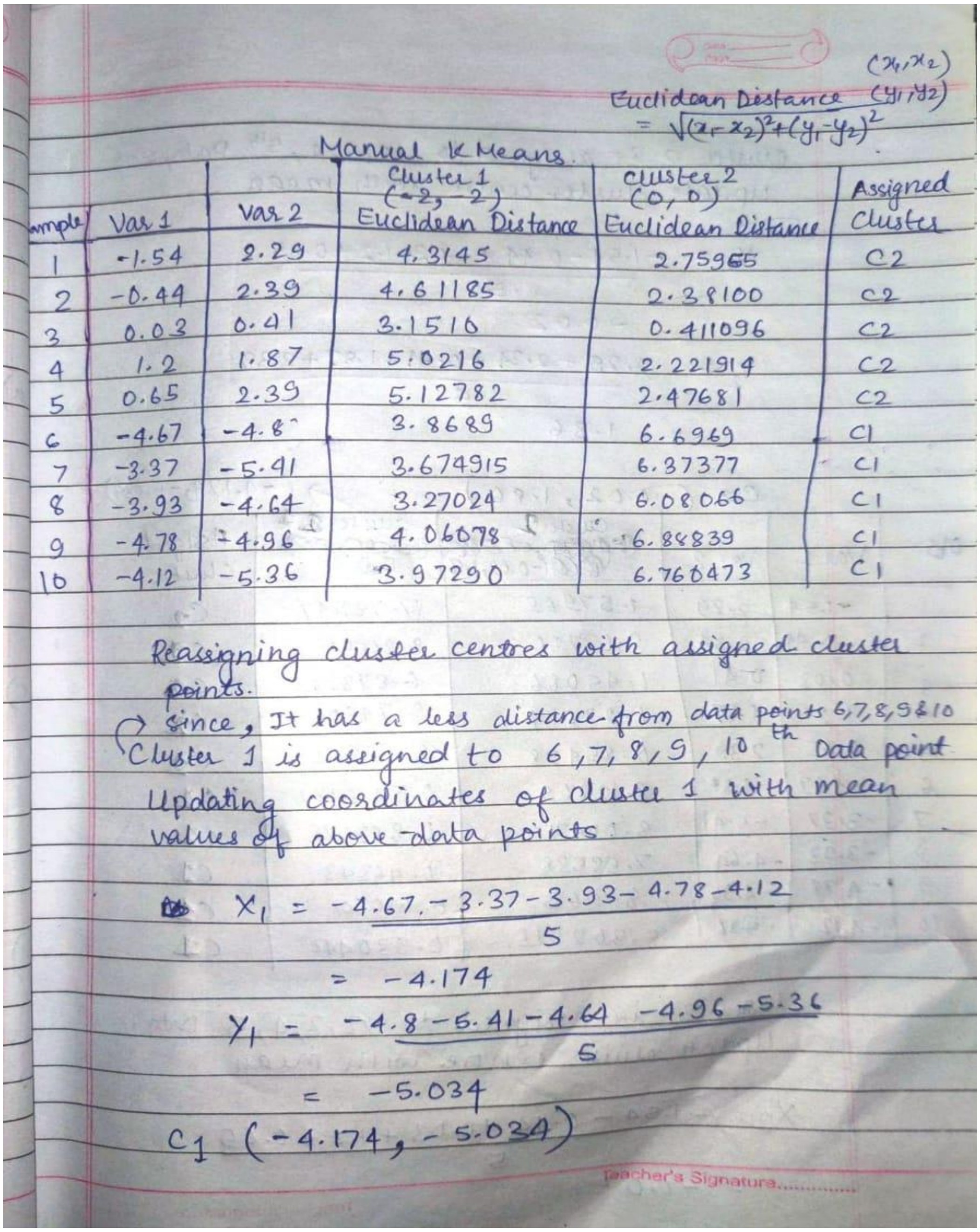
CLL788 Assignment 4

Rishav Kumar Rajak 2018CH70302

1)



(x_1, x_2)

Euclidean Distance (y_1, y_2)

$= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Manual K Means.

mple	Var 1	Var 2	Cluster 1 (-2, -2) Euclidean Distance	cluster 2 (0, 0) Euclidean Distance	Assigned Clusters
1	-1.54	2.29	4.3145	2.75965	C2
2	-0.44	2.39	4.61185	2.38100	C2
3	0.03	0.41	3.1510	0.411096	C2
4	1.2	1.87	5.0216	2.221914	C2
5	0.65	2.39	5.12782	2.47681	C2
6	-4.67	-4.8	3.8689	6.6969	C1
7	-3.37	-5.41	3.674915	6.37377	C1
8	-3.93	-4.64	3.27024	6.08066	C1
9	-4.78	-4.96	4.06078	6.88839	C1
10	-4.12	-5.36	3.97290	6.760473	C1

Reassigning cluster centres with assigned cluster points.

Since, It has a less distance from data points 6,7,8,9&10 Cluster 1 is assigned to 6, 7, 8, 9, 10th Data point

Updating coordinates of cluster 1 with mean values of above data points.

$$X_1 = \frac{-4.67 - 3.37 - 3.93 - 4.78 - 4.12}{5}$$

$$= -4.174$$

$$Y_1 = \frac{-4.8 - 5.41 - 4.64 - 4.96 - 5.36}{5}$$

$$= -5.034$$

$C_1\ (-4.174, -5.034)$

Teacher's Signature..............

Cluster 2 is assigned to 1, 2, 3, 4, 5th Datapoint
Update cluster centre with mean

$$X_2 = \frac{-1.54 - 0.44 + 0.03 + 1.2 + 0.65}{5}$$

$$= -0.02$$

$$Y_2 = \frac{2.29 + 2.34 + 0.41 + 1.87 + 2.39}{5}$$

$$= 1.86$$

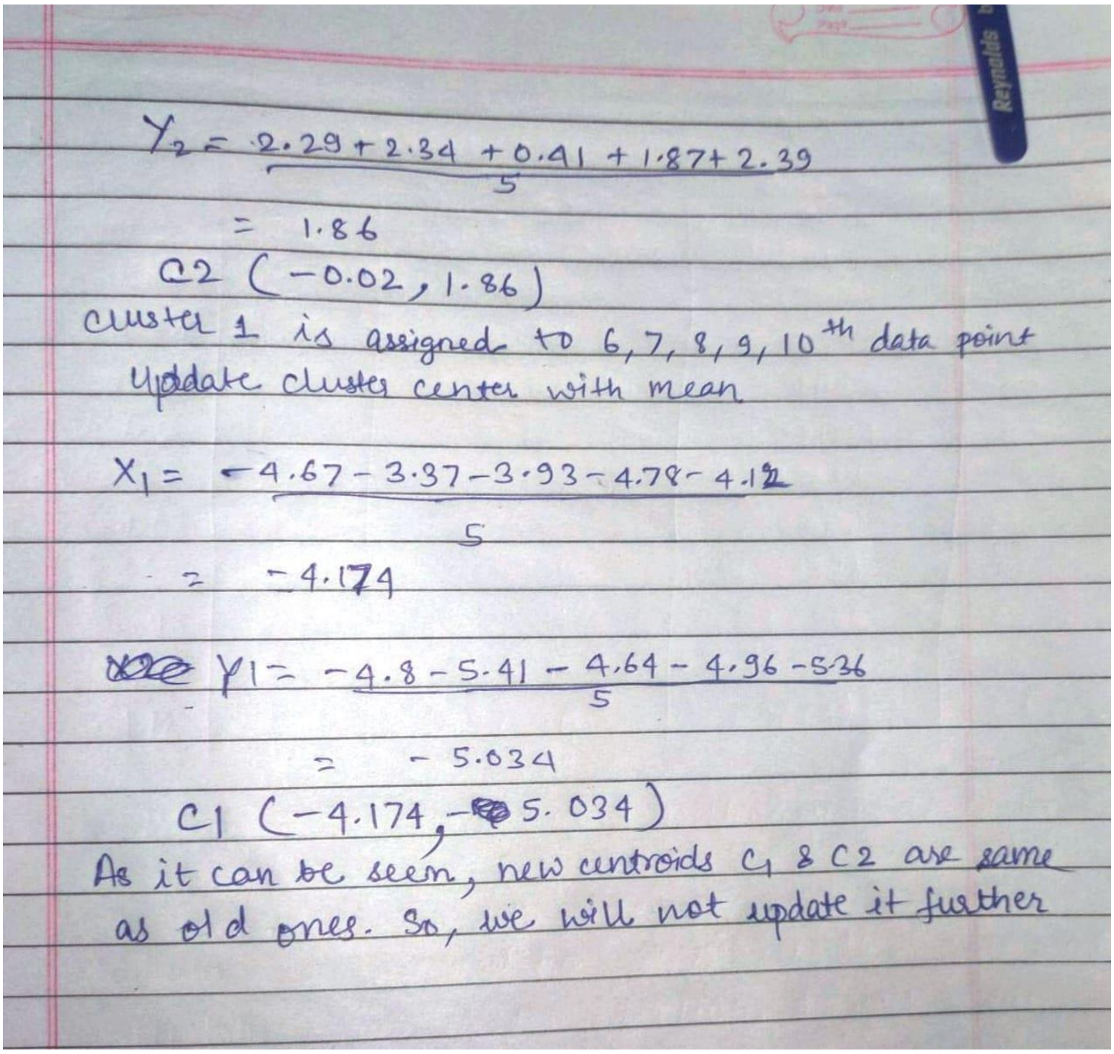
C2 (−0.02, 1.86) ↗ (−4.174, −5.034)

~~Obs~~	Var 1	Var 2	Cluster 2 ED (−0.02, 1.86)	Cluster 1 ED	Assigned Cluster
1	−1.54	2.29	1.57965	7.783247	C_2
2	−0.44	2.34	0.63780	8.26550	C2
3	0.03	0.41	1.45086	6.87828	C2
4	1.2	1.87	1.22004	8.74900	C2
5	0.65	2.39	0.85428	8.85362	C2
6	−4.67	−4.8	8.12269	0.548427	C1
7	−3.37	−5.41	8.00471	0.887576	C1
8	−3.93	−4.64	7.58538	0.46343	C1
9	−4.78	−4.96	8.31685	0.610501	C1
10	−4.12	−5.36	8.302815	0.330442	C1

Cluster 2 is assigned to 1, 2, 3, 4, 5 Data.
Update cluster centre with mean

$$X_2 = \frac{-1.54 - 0.44 + 0.03 + 1.2 + 0.65}{5}$$

= −0.02



$$Y_2 = \frac{2.29 + 2.34 + 0.41 + 1.87 + 2.39}{5}$$

$= 1.86$

C2 $(-0.02, 1.86)$

cluster 1 is assigned to 6, 7, 8, 9, 10th data point

update cluster center with mean

$$X_1 = \frac{-4.67 - 3.37 - 3.93 - 4.78 - 4.12}{5}$$

$= -4.174$

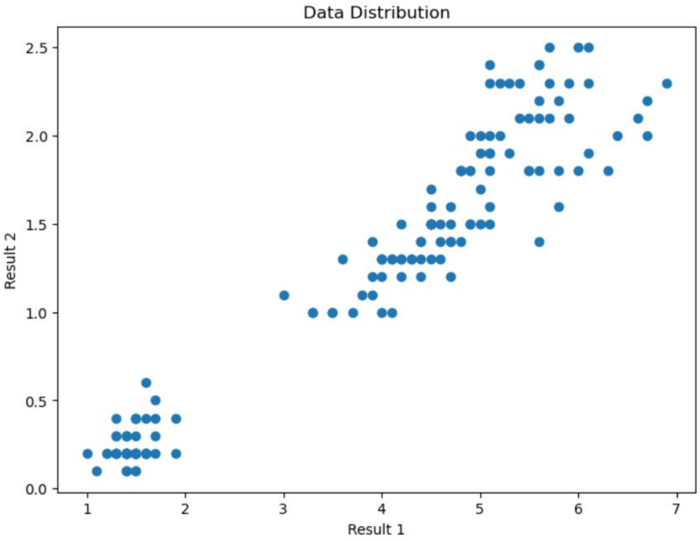
$$Y1 = \frac{-4.8 - 5.41 - 4.64 - 4.96 - 5.36}{5}$$

$= -5.034$

C1 $(-4.174, -5.034)$

As it can be seen, new centroids C_1 & C2 are same as old ones. So, we will not update it further

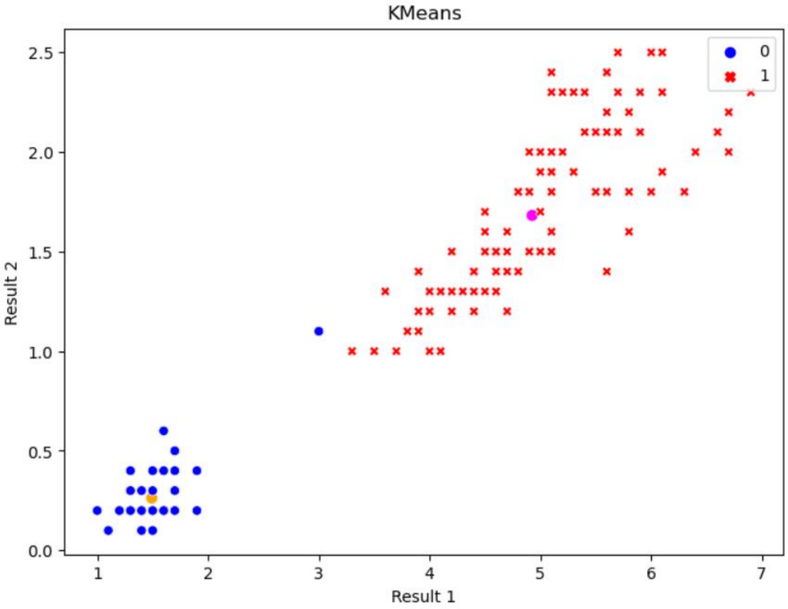
2 a)



From the above plot, we can see that the data points are grouped into two distinct spaces in the 2D coordinate system. One group is close to the origin and has low values for both the features (Result 1 and Result 2). Other group can be seen towards the alternate end, where the values of both the features range from an intermediate value to a high value.

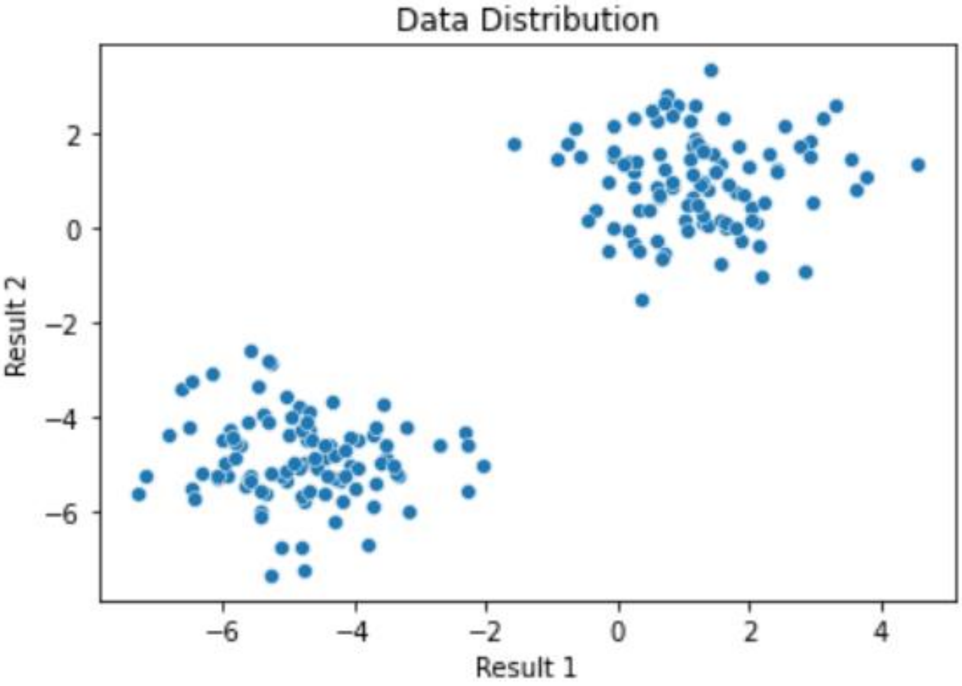
2) b) KMeans

Cluster Centres: ([1.49215686, 0.2627451], [4.92525253, 1.68181818])



Here I applied the K-Means algorithm to find out the two clusters. These clusters contain datapoints, whose characteristics are similar and by that, it can be understood that, between two clusters the degree of similarity is less and that is why there are two clusters to group them in the first place. The K Means algorithm groups the data points on the basis minimum Euclidean distance between them. We are required to make 2 clusters. So, the input parameters to this algorithm is the value of K and the dataset (which needs to be clustered). Now for every cluster, we have a cluster centre, popularly known as a centroid. I randomly initialise all the K centroids. Then for each datapoint in the given set, I first find the nearest centroid and assign it to the clusters. Now for each of these K clusters, we recompute the centroids by calculating the mean of all the points present in the respective clusters and continue this process until the cluster centres remain unchanged, at this point, we can say that the optimisation algorithm has converged.

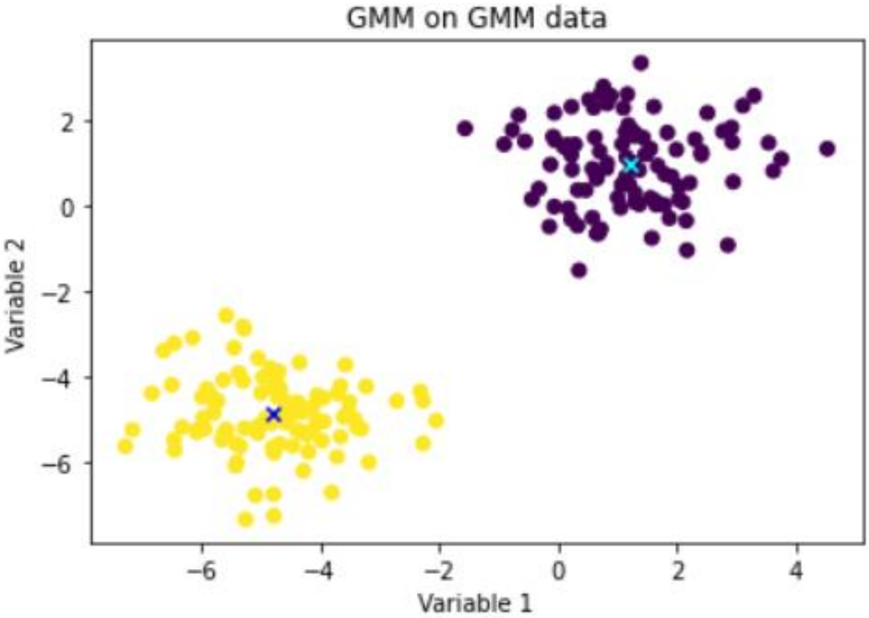
2 c) From the below plot, we can observe that there is grouping of certain datapoints. The first group has low values of both its features (Result 1 & Result 2) and the second group is at an alternate end in the plot with high values of both the features. This indicates that there are probably 2 clusters of datapoints present in the provided dataset



2 d) Means of two distributions: [[1.19721315, 1.00893775] [-4.81815103 -4.87158388]]

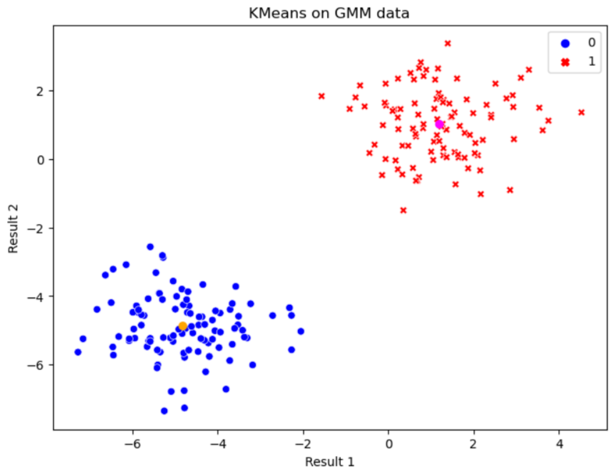
Covariance Matrix: array ([[1.19169015, 0.00910693], [0.00910693, 0.97049735]]),

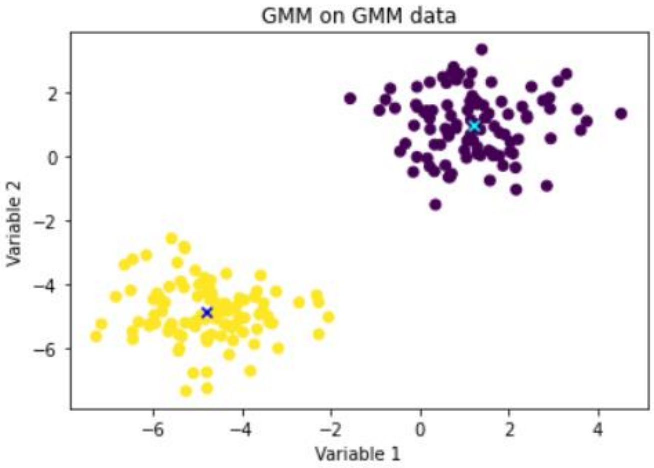
Array ([[1.16968737, -0.10528566], [-0.10528566, 0.80018909]])

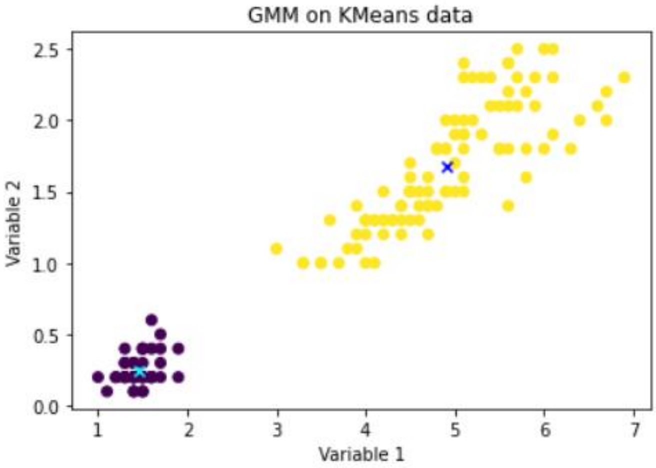


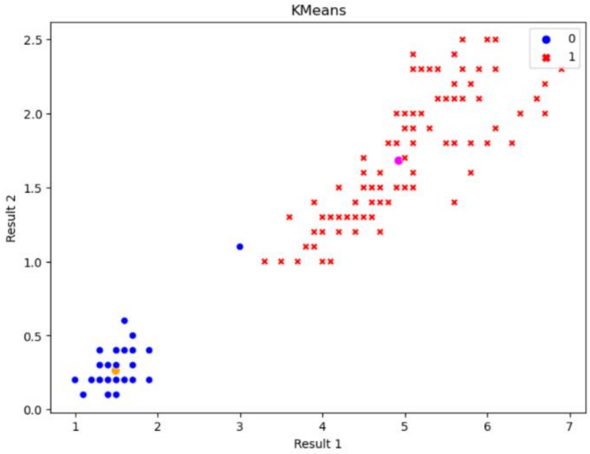
2 e) I applied KMeans and GMM on Data.xlsx (Data for Kmeans) as well as Data_GMM.xlsx (Data for GMM). As we can see from bottom right figure that one data point ~ (3,1.2) has been assigned cluster1

(blue). This seems to be a misclassification. But in the case of GMM this same point is assigned cluster 2(Distribution 2) because both methods use different approach to get clusters.

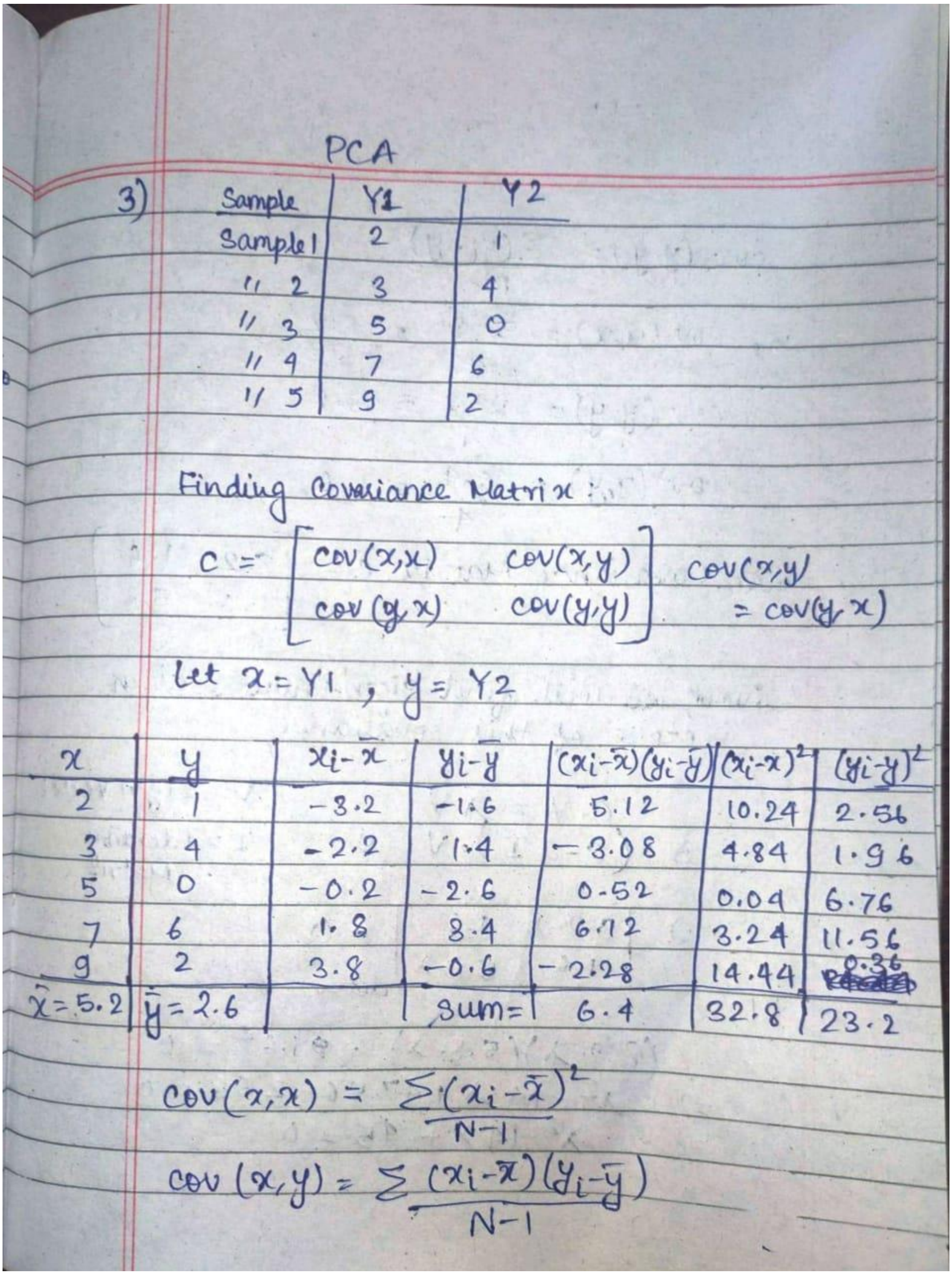








3)



PCA

3)

Sample	Y1	Y2
Sample1	2	1
〃 2	3	4
〃 3	5	0
〃 4	7	6
〃 5	9	2

Finding Covariance Matrix :

$$C = \begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix}$$

$cov(x,y) = cov(y,x)$

Let $x = Y1$, $y = Y2$

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-x)^2$	$(y_i-y)^2$
2	1	-3.2	-1.6	5.12	10.24	2.56
3	4	-2.2	1.4	-3.08	4.84	1.96
5	0	-0.2	-2.6	0.52	0.04	6.76
7	6	1.8	8.4	6.72	3.24	11.56
9	2	3.8	-0.6	-2.28	14.44	0.36
$\bar{x} = 5.2$	$\bar{y} = 2.6$		sum=	6.4	32.8	23.2

$$cov(x,x) = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

$$cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$\text{cov}(y,y) = \frac{\Sigma(y_i - y)^2}{N-1}$$

So, $\text{cov}(x,x) = \frac{32.8}{4} = 8.2$

$\text{cov}(y,y) = \frac{23.2}{4} = 5.8$

$\text{cov}(x,y) = \frac{6.4}{4} = 1.6$

Covariance Matrix $C = \begin{bmatrix} 8.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix}$

Now, we will find eigen-values & eigen vectors of this covariance

$C.V = \lambda V$

V= Eigen vector

$\Rightarrow (C - \lambda I).V = 0$

I = Identity Matrix

$\Rightarrow \begin{bmatrix} 8.2-\lambda & 1.6 \\ 1.6 & 5.8-\lambda \end{bmatrix}_{\text{deter}} = 0$

$(8.2-\lambda)(5.8-\lambda) - \cancel{2}1.6^2 = 0$

$\Rightarrow \lambda^2 - 14\lambda + 47.56 - 2.56 = 0$

$\Rightarrow \lambda^2 - 14\lambda + 45 = 0$

$$(\lambda-5)(\lambda-9)=0$$

$$\therefore \lambda = 5, 9.$$

Eigenvector corresponding to two root are:

$\lambda = 9$

$$(C-9I)\,V_1 = 0$$

$$\begin{pmatrix} -0.8 & 1.6 \\ 1.6 & -3.2 \end{pmatrix}\begin{pmatrix} v_1' \\ v_1'' \end{pmatrix} = 0$$

$$-0.8 v_1' + 1.6 v_1'' = 0$$

$$v_1' = 2 v_1''$$

$$V_1 = \begin{bmatrix} 1 \\ 1/2 \end{bmatrix}$$

$\lambda = 5$

$$(C-5I)\,V_2 = 0$$

$$\begin{pmatrix} 3.2 & 1.6 \\ 1.6 & 0.8 \end{pmatrix}\begin{pmatrix} v_2' \\ v_2'' \end{pmatrix} = 0$$

$$3.2 v_2' + 1.6 v_2'' = 0$$

$$v_2' = -\frac{1}{2} v_2'' = 0$$

$$V_2 = \begin{bmatrix} -1/2 \\ 1 \end{bmatrix}$$

So, Feature Matrix [Eigenvector matrix]

$$\phi = \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} 1 & -0.5 \\ 0.5 & 1 \end{bmatrix}$$

$\lambda_1 = 9$ $\quad$ $\lambda_2 = 5$

PC1 . PC2

↳ Highest Eigen Value

Variance explained by PC1 $= \frac{9}{9+5}$

$= 64.285\%$

Variance explained by PC2 $= \frac{5}{14}$

$= 35.714\%$

Transforming data along PC's.

Projection $(P) = X\phi$.

$$\begin{pmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{pmatrix}_{5\times 2} \begin{pmatrix} 1 & -1/2 \\ 1/2 & 1 \end{pmatrix}_{2\times 2}$$

$$\text{Projection} = \begin{pmatrix} 3/2 & 0 \\ 5 & -5/2 \\ 5 & -5/2 \\ 10 & 5/2 \\ 10 & -5/4 \end{pmatrix}$$

	PC1 scores	P2 scores
% Variance Explained =	64.285	35.714