CS215 Assignment 4

Aquib Nawaz(190050023),Rajesh Dasari(190050030),
Paavan Kumar(190050051)

11 November 2020

Question 1

a)

X_1,X_2 are uniformly distributed ,
hence $P(X_1 = a) = \frac{1}{2}$ for $-1 < a < 1$ and $P(X_2 = b) = \frac{1}{2}$ for $-1 < b < 1$.
Therefore , we have $P(X = (a, b)) = P(X_1 = a, X_2 = b)$
$P(X = (a, b)) = P(X_1 = a)P(X_2 = b) = \frac{1}{4}$
Now to calculate the probability that the point lies in a circle of radius 1 , we integrate the probability $p_X(a, b)$ over the area of the circle.
let the area of circle be A

Hence, $p = \int\int_{x,y \in A} p_X(x, y)dxdy$
$p = \int\int_{x,y \in A} \frac{1}{4}dxdy$
$p = \frac{1}{4}\int\int_{x,y \in A} dxdy$
$p = \frac{1}{4}A$
Therefore $\boxed{p = \frac{\pi}{4}}$

b)

Suppose N is the sample size of the random variable generated for computing the value of π.
And Y be the number of points that actually lie inside the circle when we plot the values of $X = (X_1, X_2)$ taking X_1 on the x-axis and X_2 on y-axis
Then by the formula determined above $p = \frac{\pi}{4}$
We also know that actual probability $p = \frac{Y}{N}$
Hence equating both we get the value of π to be $\frac{4Y}{N}$

c)

The code for this part is attached with the name $q1.m$ in the Code folder
Our code doesnt handle this case, since we generate a N smaple and the count the values that lie in the region bounded by the circle
Generating a random sample of size 10^9 usually takes a lot of time in MATLAB, so this is not a very good idea , so instead we can use the bernoullli confidence interval and estimate the value

of π using the formula mentioned in part D

$$\hat{p} \pm z\sqrt{\frac{(1-\hat{p})\hat{p}}{M}} \tag{1}$$

Given the value of M which is the sample size , we can calculate the value of z and hence estimate the confidence interval and we can cut the value of π estimated to a small range with more than 99% confidence interval

d)

Lets take each trail to be a Bernoulli trial with parameter $p = \pi/4$, Here if the the chosen point in a trial lies in the circle then it will be considered success else failure.
Now we know for a binomial distribution p is estimated as

$$\hat{p} \pm z\sqrt{\frac{(1-\hat{p})\hat{p}}{M}} \tag{2}$$

Where z for 95% confidence interval is 1.96 and M is size of data points.
Also error in π is 4× error in $\hat{p}$ We can find M from here which turns out to be around 100000 using $\pi = 3.141592$ and $\hat{p} = \pi/4$.

Question 2

a)

A general multivariate gaussian variable is given by $X = AW + \mu$
Its coefficient matrix $C = AA^T$. So to generate random sample of given distribution we have to first find A.
Now covariance C is a symmetric positive definite matrix so it can be written as QDQ^T.
Where Q is an orthogonal matrix whose columns are the eigenvectors of C, and D is a diagonal matrix whose entries are the eigenvalues of C.
Hence $C = QDQ^T = (Q\sqrt{D})(\sqrt{D}Q^T) = (Q\sqrt{D})(Q\sqrt{D})^T$ as entries in D are positive.
Hence one possible value of A is $Q\sqrt{D}$.
Q and D can be obtained by eig() function and W can be obtained by randn() function.

The code for this question is attached as $q2.m$

Question 3

a)

Principal component analysis can be used to find a linear fit for scatter plots
First we determine the mean of the two varaibles X,Y (name it as m_1, m_2)
By PCA, the linear fit must pass through the (m_1, m_2)
Now we determine the variance and co-variance of the random variables X,Y (name them v_1, v_2, c
Now we transform the X,Y to X-m_1,Y-m_2(actually this is not necessary)
Then assume a unit vector v,$[\cos(\theta), \sin(\theta)]'$ along the direction of the resultant linear fit, $slope = \tan(\theta)$

Now by PCA , we have to maximise the value of $v'Cv$ where v' is the transpose of vector v
Here C is the covariance matrix $[[v_1 c][c v_2]]$
Hence expanding $v'Cv$ we get an expression E

$E = v_1 \cos(\theta)^2 + v_2 \sin(\theta)^2 + 2c\cos(\theta)\sin(\theta)$
Differentiating w.r.t θ, we get

$(v_2 - v_1)\sin(2\theta) + 2c\cos(2\theta)) = 0 \implies \tan(2\theta) = \frac{2c}{v_1 - v_2}$ and Hence $\boxed{\theta = \frac{1}{2}\arctan\frac{2c}{v_1 - v_2}}$

b)

The code for this part is attached with the name $q3b.m$
The scatter plot overlaid with plot for pca is attached below

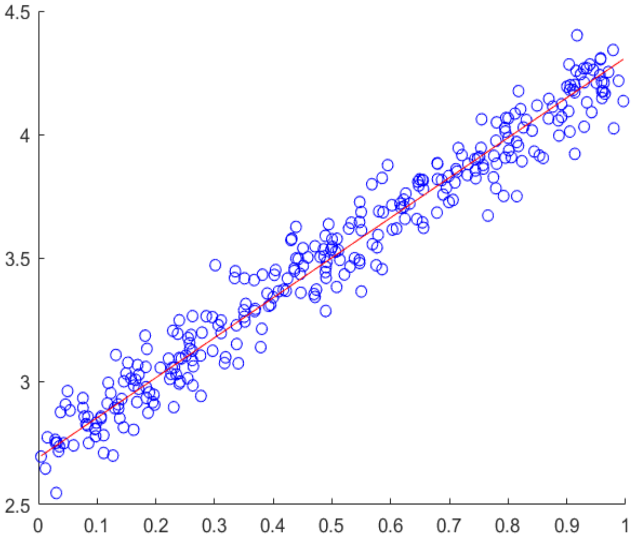

Figure 1: points2D Set1

c)

The code for this part is attached with the name $q3c.m$
The scatter plot overlaid with plot for pca is attached below
Since the covariance of the set of vaues of X and Y is approximately 0,the value of theta determined from the relation in part a) comes out to be zero, this is a pretty bad approximation of the graph since we are assuming the graph to be linear where in practical its not zero , for example the set 2 given has approx zero correlation and hence the outcome of the result is quite unsuitable.

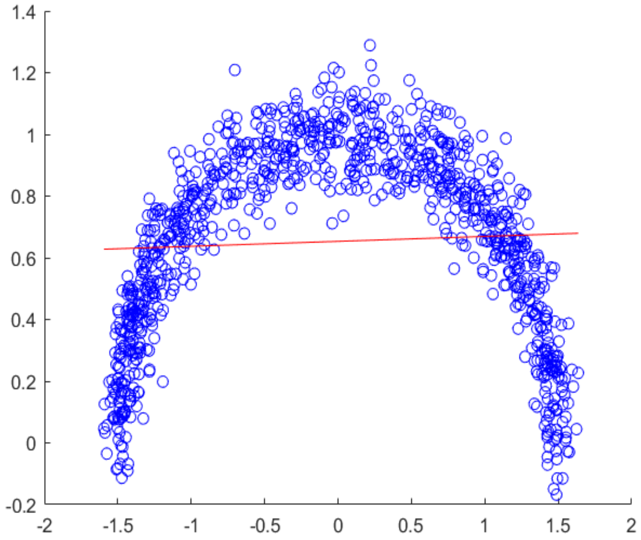

Figure 2: points2D Set2

The quality of approximation use for this part is very poor since we actually don't know that the data is scattered around a line , which is the assumption that we made in the pca part to estimate the direction of maximal variance, But from the scatter plot we can actually say that the data is distributed along a semicircular arc,and only one dimension is enough to find the direction of maximal variance it may be a curved direction depending on the distance from origin

Question 4

MATLAB code for this question is attached as Q4.m

On plotting eigen values for different digits we found that the number of significant eigen values is around 100 which is far less than 784.
It means that value of only few pixels varies much among all the images of same digit.
It is because some style or orientation of digit may change but its basic skeleton remain same.

On plotting three images for each digit we found that the mode of variation showing images are tilted around its mean. So people deviates in writing.

For digit 1 in third image it is tilted more rightward while first image is straight which means some people write it mostly straight while other more toward tilted write.

Question 5

Code for this question is attached as q5.m
The code written by us for this part is slightly slower than usual time

Question 6

Code for this Question is attached as q6.m

a)

The plot for mean and first four eigen vectors have been attached in the *results* folder

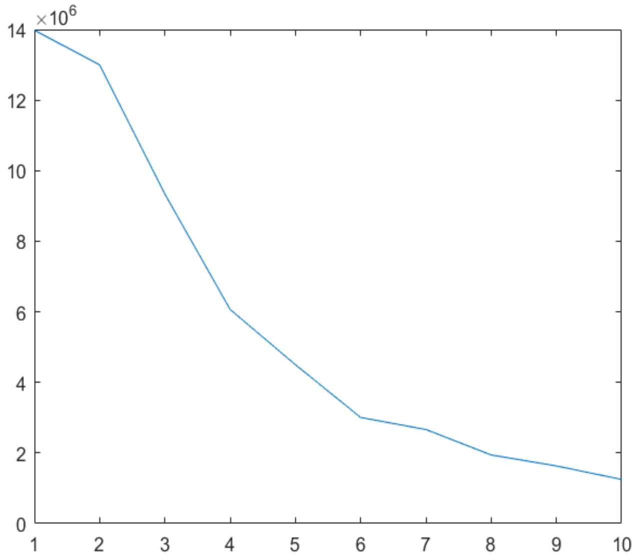

Figure 3: first 10 eigen values

b)

Let $sz = size(I)$ Let the eigen vectors be $\vec{V_1}, \vec{V_2}, \vec{V_3}, \vec{V_4}$ found out from the first part.
let I' be the closest representation of the image I in terms of the mean $\vec{\mu}$ and the eigen vectors

$I' = a_0\vec{\mu} + \sum_1^4 a_i\vec{V_i}$

Our job is to minimise $I - I'$
Therefore to minimise the magnitude of $T = I - I'$

Let $T[j] = I[j] - a_0\vec{\mu}[j] - \sum_0^4 a_i\vec{V_i}[j]$

Therefore to minimise the value of $F(a_0, a_1, a_2, a_3, a_4) = \sum_1^{sz} T[j] \times T[j]$

Differentaiting w.r.t $a_i, i \neq 0$, we get

$\frac{\partial F}{\partial a_i} = \sum_1^{sz} 2 \times T[j] \times \frac{\partial T[j]}{\partial a_i}$

We Know for $i \neq 0, \frac{\partial T[j]}{\partial a_i} = -\vec{V_i}[j]$

Therefore ,$\frac{\partial F}{\partial a_i} = \sum_1^{sz} 2 \times T[j] \times -\vec{V_i}[j]$

Hence $\frac{\partial F}{2\partial a_i} = -I \cdot \vec{V_i} + a_0\vec{V_i} \cdot \vec{\mu} + \sum_{j=1}^4 a_j\vec{V_i} \cdot \vec{V_j} = -I \cdot \vec{V_i} + a_0\vec{V_i} \cdot \vec{\mu} + a_i\vec{V_i} \cdot \vec{V_i} + \sum_{j \neq i} a_j\vec{V_i} \cdot \vec{V_j}$

Since these are eigen vectors of a covariance matrix, $\vec{V_i}\cdot\vec{V_j} = 0$ for $j \neq i$ and $\vec{V_i}\cdot\vec{V_i} = 1$

Therefore, $\frac{\partial F}{2\partial a_i} = -I\cdot\vec{V_i} + a_0\vec{V_i}\cdot\vec{\mu} + a_i$

Equating to zero , we get $\frac{\partial F}{2\partial a_i} = -I\cdot\vec{V_i} + a_0\vec{V_i}\cdot\vec{\mu} + a_i = 0 \implies \boxed{a_i = I\cdot\vec{V_i} - a_0\vec{V_i}\cdot\vec{\mu} \qquad ,i \neq 0}$

Now similarly differentiating w.r.t a_0, we get

$\frac{\partial F}{\partial a_0} = \sum_1^{sz} 2\times T[j]\times\frac{\partial T[j]}{\partial a_0}$

We Know for $\frac{\partial T[j]}{\partial a_0} = -\vec{\mu}[j]$

Therefore ,$\frac{\partial F}{\partial a_i} = \sum_1^{sz} 2\times T[j]\times -\vec{\mu}[j]$

Hence, $\frac{\partial F}{2\partial a_0} = -I\cdot\vec{\mu} + a_0\vec{\mu}\cdot\vec{\mu} + \sum_{i=1}^4 a_i\vec{\mu}\cdot\vec{V_i} = -I\cdot\vec{\mu} + a_0\vec{\mu}\cdot\vec{\mu} + \sum_{i=1}^4(I\cdot\vec{V_i} - a_0\vec{V_i}\cdot\vec{\mu})\vec{\mu}\cdot\vec{V_i}$

Therefore, $\frac{\partial F}{2\partial a_0} = a_0(\vec{\mu}\cdot\vec{\mu} - \sum_{i=1}^4(\vec{\mu}\cdot\vec{V_i})^2) - (I\cdot\vec{\mu} - \sum_{i=1}^4((I\cdot\vec{V_i})(\vec{\mu}\cdot\vec{V_i}))$

Equating to zero, we get $\boxed{a_0 = \dfrac{I\cdot\vec{\mu} - \sum_{i=1}^4(I\cdot\vec{V_i})(\vec{\mu}\cdot\vec{V_i})}{\vec{\mu}\cdot\vec{\mu} - \sum_{i=1}^4(\vec{\mu}\cdot\vec{V_i})^2}}$

And Therefore, $a_i = I\cdot\vec{V_i} - \frac{I\cdot\vec{\mu} - \sum(I\cdot\vec{V_i})(\vec{\mu}\cdot\vec{V_i})}{\vec{\mu}\cdot\vec{\mu} - \sum(\vec{\mu}\cdot\vec{V_i})^2}\vec{V_i}\cdot\vec{\mu}, i \neq 0$ and $a_0 = \frac{I\cdot\vec{\mu} - \sum_{i=1}^4(I\cdot\vec{V_i})(\vec{\mu}\cdot\vec{V_i})}{\vec{\mu}\cdot\vec{\mu} - \sum_{i=1}^4(\vec{\mu}\cdot\vec{V_i})^2}$

The plots thus obtained are placed in the $results$ folder with the name $q\_i.png$ where each image corresponds to the ith fruit.
One such image is below as for image 16 where the fruit is an apple

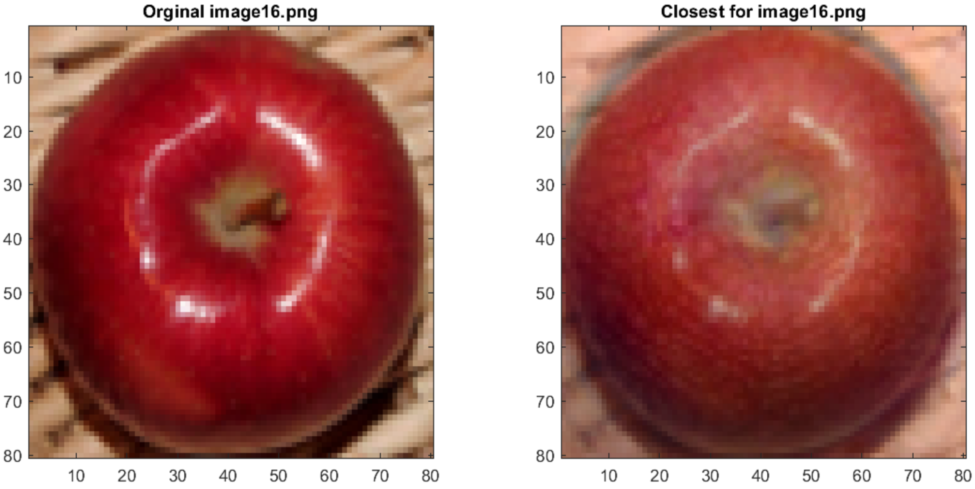

Figure 4: original and closest

This can be done in another way by considering the co-efficient of mean to be 1 and fitting the data onto the first four eigen vectors, called 4-D hyperplane fitting, we find the co-eff in the same as found out above the mean co-eff would be 1 which means $\boxed{a_i = (I - \vec{\mu})\cdot\vec{V_i} \qquad ,i \neq 0}$, from this we can find out the closest representation in the 4D hyperplae formed by the four eigen vectors c)

Let The eigen values corresponding to those four eigen vectors placed along a diagonal of a 4 by 4 matrix, name this matrix D can be computed using eigs function eficiently
With matrix as diagonal eigen valued matrix , we can find the matrix A for standard normal distribution given the eigen valued matrix as follows
$A = V'\sqrt{D}V$
Now we generate a random sample of size 19200,and using the relation X = AW + μ, we get a new fruit value with th egiven mean and eigen vectors as V W is a gaussian random variable that can be generated by using randn(19200,1) in matlab