Assignment 2 : CS 215

Saikiran 200050023
Krishna Kamal 200050142

Report for Question 2

Instructions for running the code is given at the end

Question 2.1:

Generate N points (with N taking the values 10, 102 , 103 , 104 , 105) from a multivariate 2D Gaussian probability density function with mean μ = [1, 2]’ and a covariance matrix C with the first row as [1.6250, −1.9486] and the second row as [−1.9486, 3.8750].

X = A*W + μ

- Here C is symmetric and PSD with all positive eigenvalues so there exists an orthogonal matrix Q and diagonal matrix D such that C = QDQ’ where D is composed of eigenvalues of C and Q with eigenvectors of C.

Now coming to generation of points:

- First calculated A using C and eigen function.

```
r = np.sqrt(evalue) #eigenvalues
A = r*(evect)#eigen vector matrix Q
```

- Generated W consisting of two randomly generated values using np.random.normal.
- Then calculated the X with the given μ.

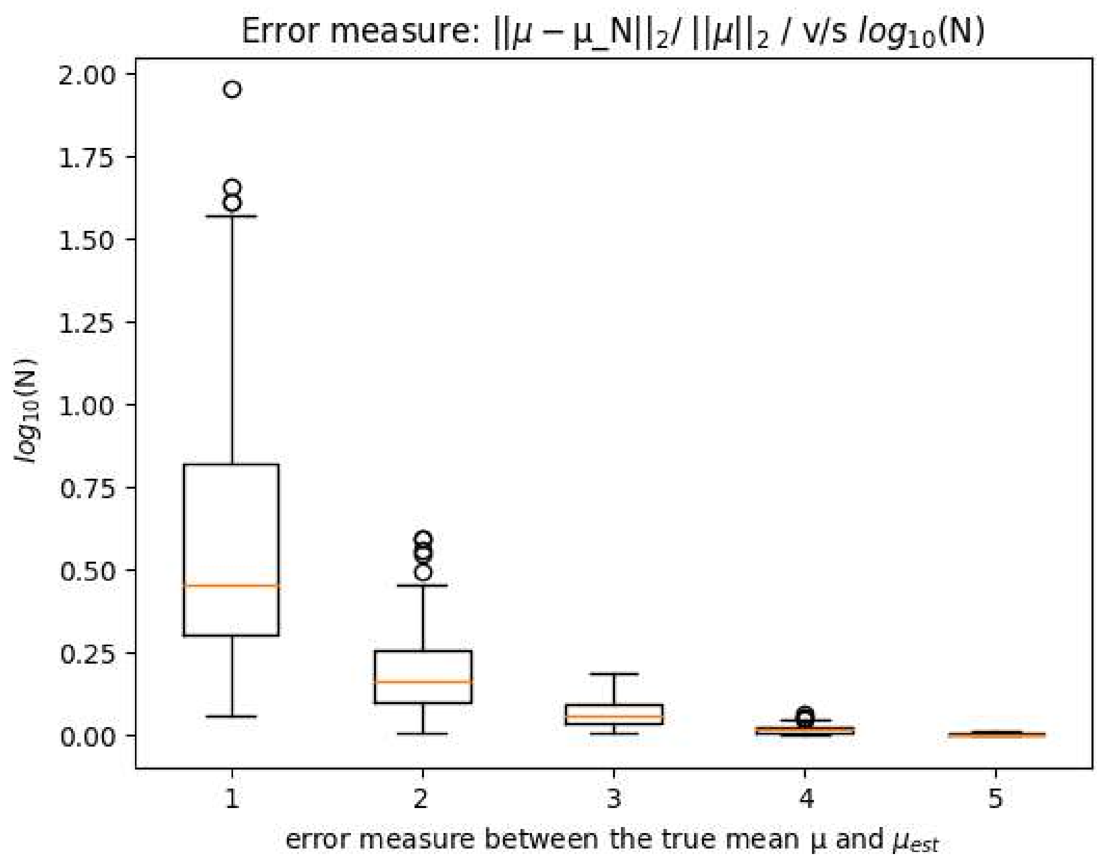
```
        p = np.random.normal()
        q = np.random.normal()
        W = np.array([[p],[q]])
        X = matrixsum(np.dot(A,W) , mean)
```

- Here matrix A is QS,hence the covariance matrix is AA’ which is also equal to QSS’Q’ but here S = S’and S^2 = D
- So AA’ = QDQ’ , this matches perfectly with the given matrix C and also the generated mean would also be equal to μ.
- Hence it is justified that we have the same Covariance matrix as μ is already fixed.

Also we can see the error to be diminishing as N approaches to larger values as in the next subparts of the question.

Question 2.2:

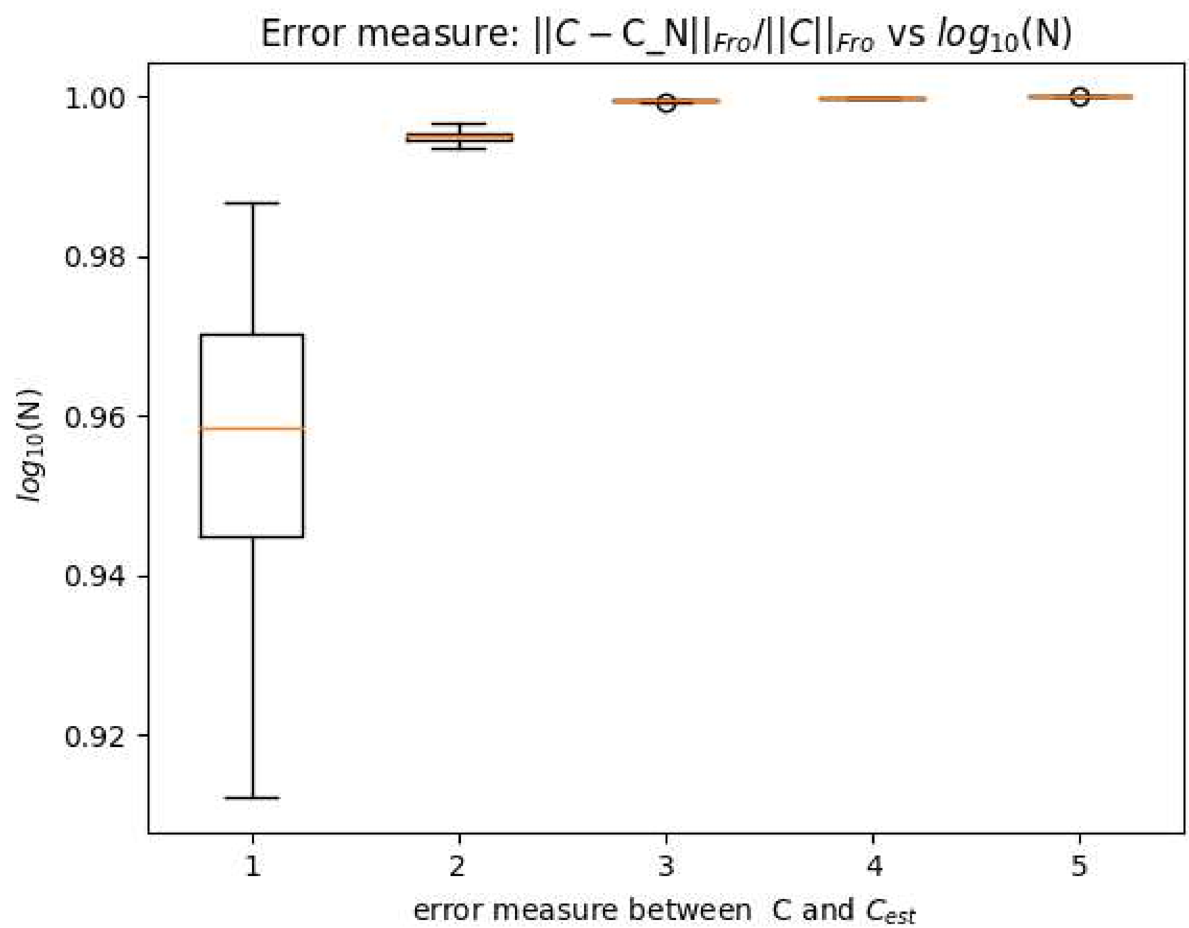
For each value of N, repeat the experiment 100 times, and plot a boxplot of the error between the true mean μ and the ML estimate μ_N (which depends on N), where the error measure is || μ − μ_N ||2 / ||μ||2. Use a logarithmic scale on the horizontal axis, i.e., log10 N.



As we can see in the graph as said above the difference between the estimate and true mean gets reduced as N tends to move as 10,100,1000,10000,100000.

Question 2.3:

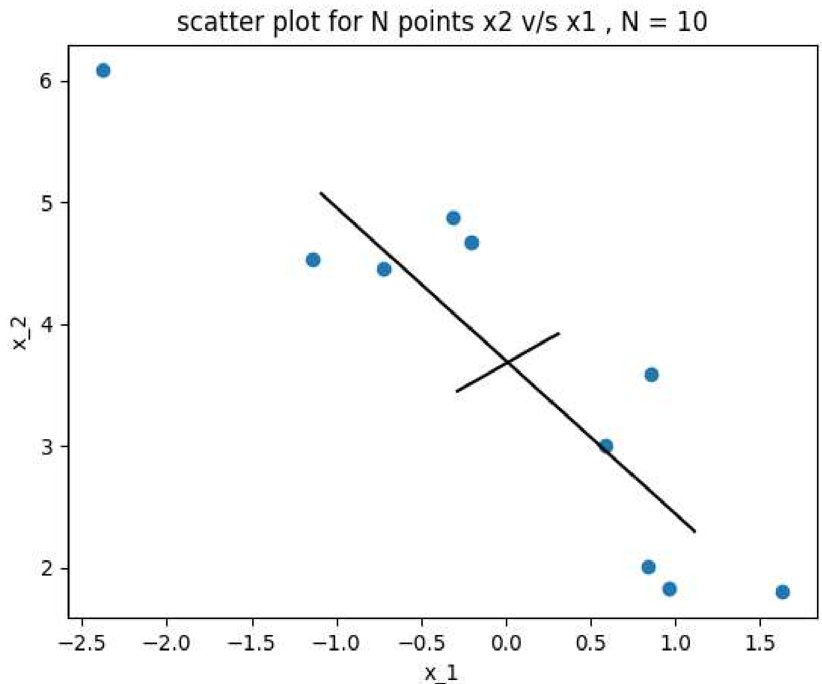
For each value of N, repeat the experiment 100 times, and plot a boxplot of the error between the true covariance C and the ML estimate CbN (which depends on N), where the error measure is || C − C_N ||Fro / || C ||Fro. Use a logarithmic scale on the horizontal axis, i.e., log10 N.

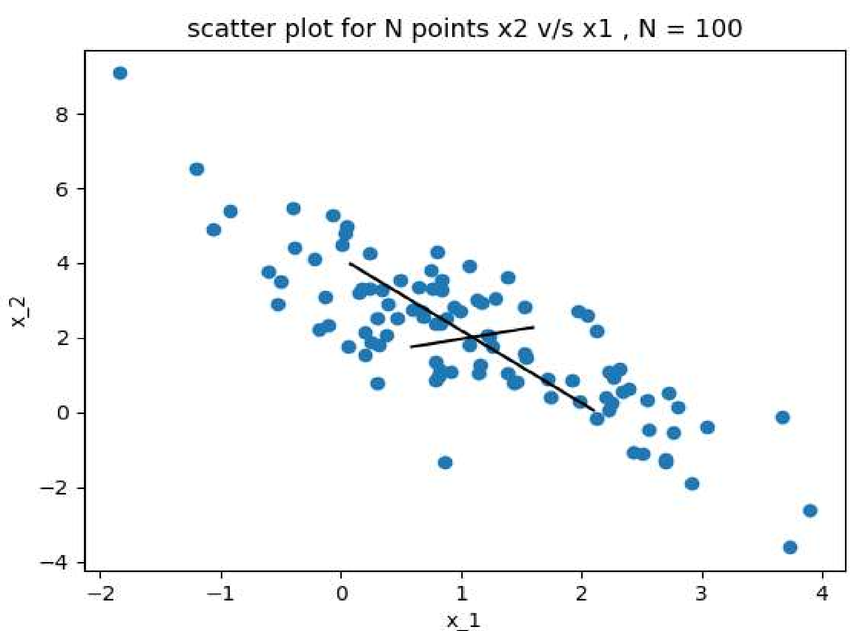


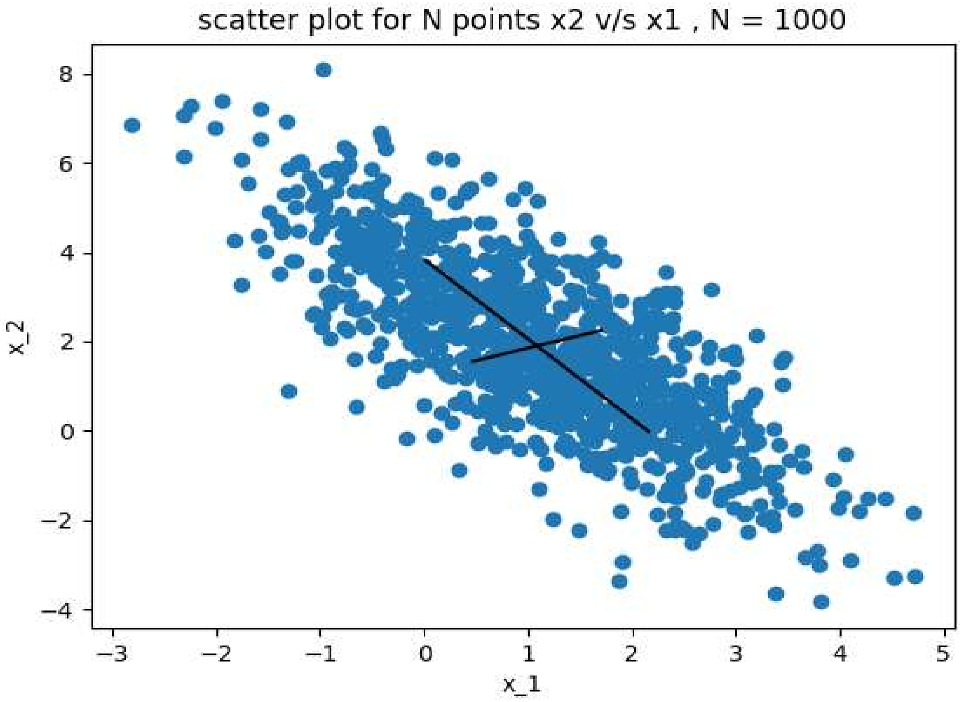
As we can see in the graph as said above the difference between the estimate and true mean gets reduced as N tends to move as 10,100,1000,10000,100000.

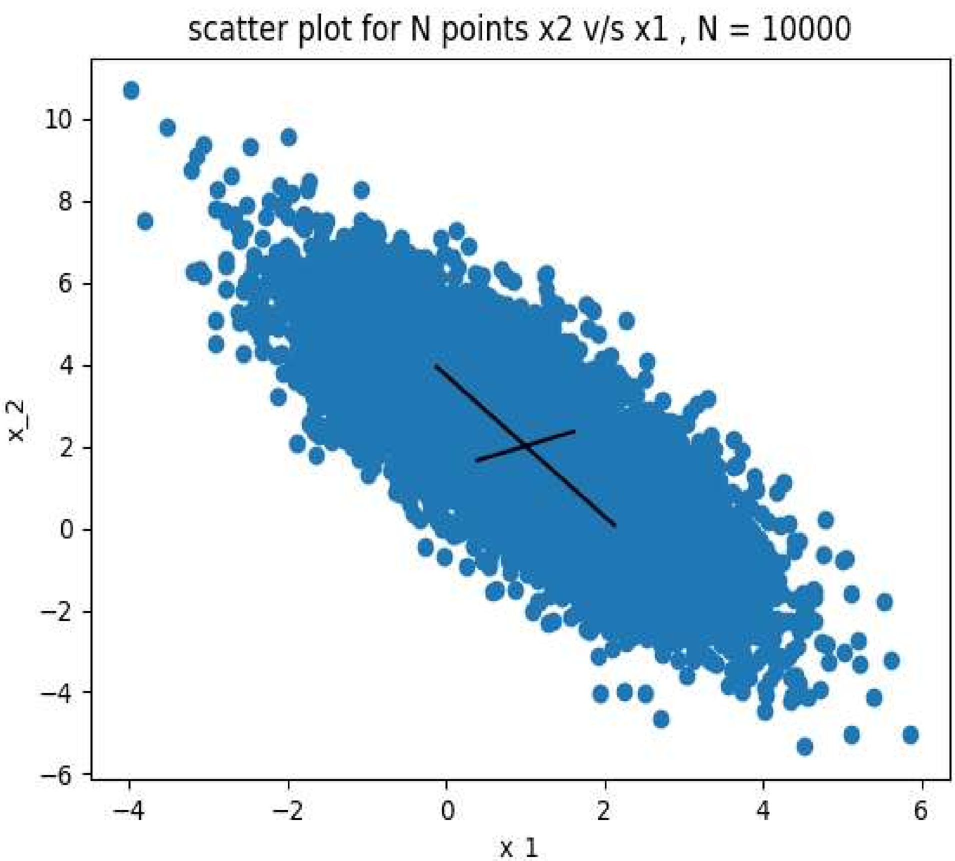
Question 2.4:

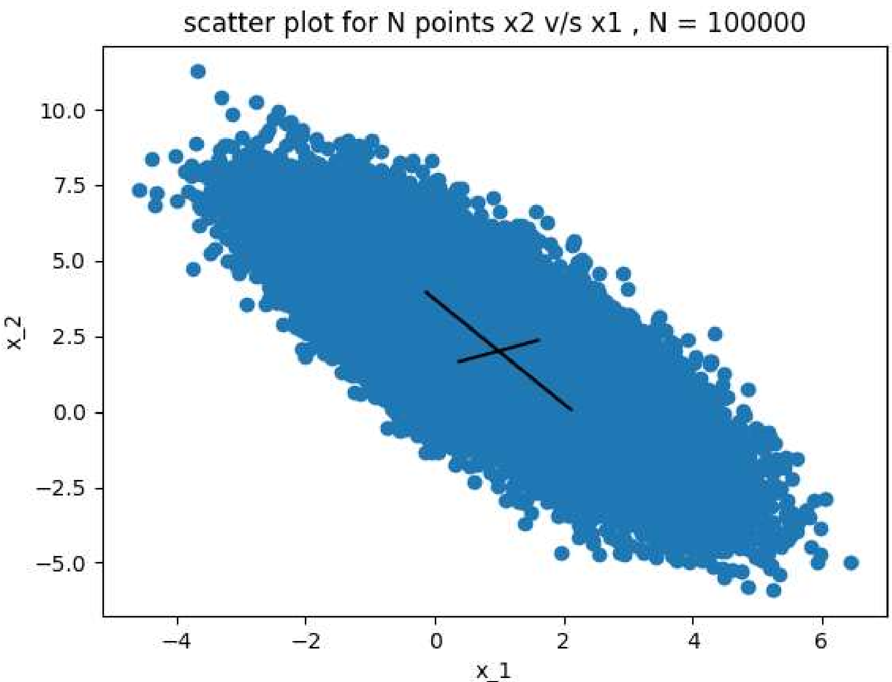
For each value of N, for a single data sample, within a single figure, plot the 2D scatter plot of the generated data and show the principal modes of variation of the data by plotting a line starting at the empirical mean and going a distance equal to the empirical eigenvalue along a direction given by the empirical eigen-vector.






scatter plot for N points x2 v/s x1 , N = 1000
x_2
x_1


scatter plot for N points x2 v/s x1 , N = 10000
x_2
x_1



- Didn't use a for loop for change of N as to store the data of the error into a separate array so that boxplot can work properly.
- As N gets to large values such as 10^4 it takes time to generate picture and the same with the box plot.

Instructions to run the code :-

Please move to the Q2 directory and

- python3 ./code/q2.py will run the python script q2.py, the program will plot and show the plots one by one (exit the current graph to view the next one)