2.

a. We are given a covariance matrix;

$$C = \begin{bmatrix} 1.6250 & -1.9486 \\ -1.9486 & 3.8750 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

We already know,

if $X = AZ + \mu$

where X has μ as mean and AA^T as covariance

$$\Rightarrow \quad C = AA^T$$

we have to find solutions for A.

Many people are attempting to find a triangular matrix A. But there is more better method !!! using the fact that C is POSITIVE DEFINITE < eigen values must be non-negative>.

Let

$$AA^T = C$$

Let $C = V Z V^T$

Then $A = V\sqrt{Z}$ is a solution where $\sqrt{Z}$ is diagonal matrix with entries as square root of eigen values.

$$\boxed{A^T = (\sqrt{Z})^T V^T = \sqrt{Z}\, V^T}$$

We used eig method in numpy to get V and Z.

Then used general math to find A.

After finding A computed X using A, μ and the random value array of length N we got using randn function.

After finding the X array we have computed the mean and covariance for X.

It can be observed that as N value increases the mean and covariance will be getting closer to the true mean and covariance we got which makes sense.

b.

CODE ⇒ q2_b.py   IMAGE ⇒ q2_b.png

For each N, we have to repeat the experiement 100 times draw a boxplot of the error between the true mean μ and ML estimate $\hat{\mu}_N$, where the error is $\|\mu - \hat{\mu}_N\|_L / \|\mu\|_L$ as a function of $\log_{10} N$.

For each $n = 10^{d}$ err [d][k] ↳ This axis would signify 100 values [errors in each case].

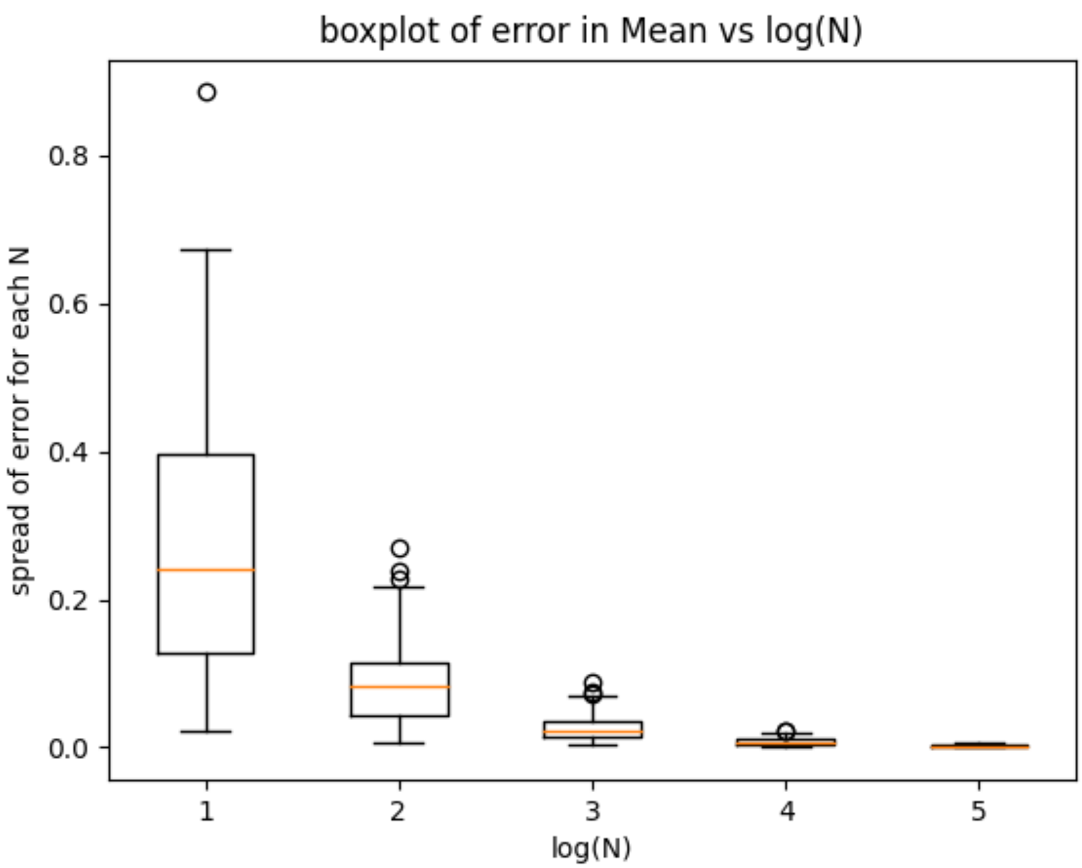
$\|\mu - \hat{\mu}_N\|_L / \|\mu\|_L$ is the measure of error.

→ Gen_data function is called 100_times to generate data 100 times.

OBSERVATIONS:

1. Exactly like law of large numbers for some estimate we got error becoming smaller as more data is taken, similarly here also the error for the mean estimate goes on decreasing as N is increasing exponentially.

- We can see from the above plot that as $\log_{10} N$ increases the spread of error from true mean decreases. This can be clearly seen from the box-plot graphs.
- This is like the what we saw in the case of law of large numbers in uni-variate case.



BOX-PLOT GRAPH

c. CODE ⇒ q2-c.py IMAGE ⇒ q2-c.png [Box-plot].

For each N, we have to repeat the experiment 100 times draw a boxplot of the error between the true variance C and ML estimate $\hat{C}_N$, where the error is $||C - \hat{C}_N||_{Fro} / ||C||_{Fro}$ as a function of $\log_{10} N$.

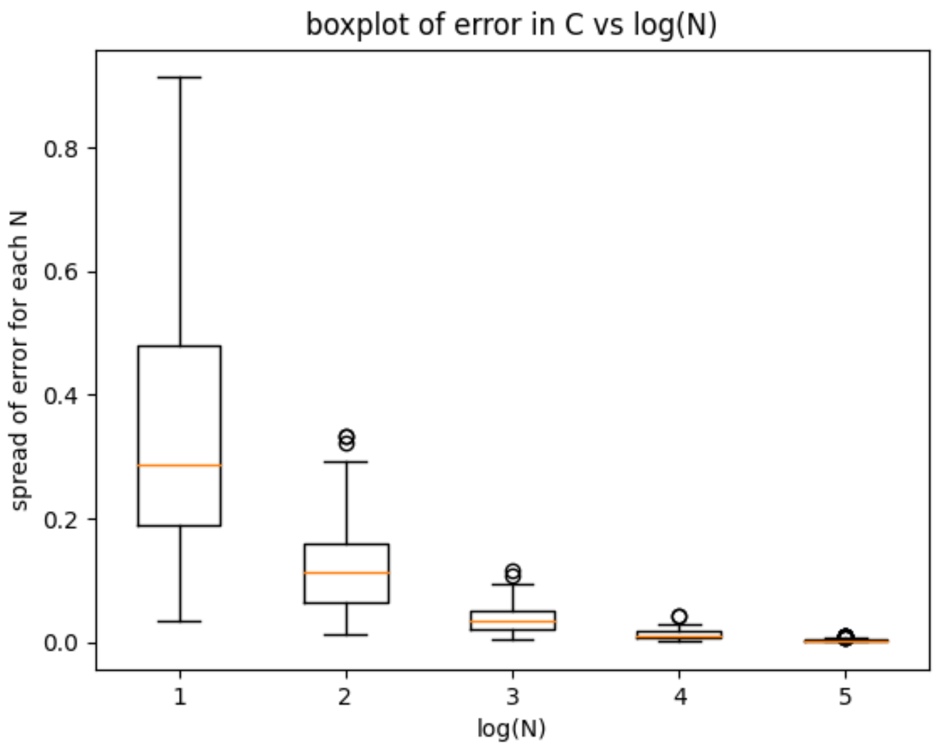
Definition:- $||X||_{Fro} = \sum (a_{ij})^2$

For each $n = 10^d$ err [C][K] ↳ This axis would signify 100 values [errors in each case].

$||C - \hat{C}_N||_{Fro} / ||C||_{Fro}$ is the measure of error.

→ Gen-data function is called 100-times to generate data 100 times.

OBSERVATIONS:



1. Exactly like law of large numbers for some estimate we got error becoming smaller as more data is taken, similarly here also the error for the mean estimate goes on decreasing as N is increasing exponentially.

- We can see from the above plot that as $\log_{10} N$ increases the spread of error from true covariance decreases. This can be clearly seen from the box-plot graphs.

- This is like the what we saw in the case of Law of large numbers in uni-variate case.

---

d. for each N

Steps :

1. Generate a data sample [In the code all the required data is actually generated all at once.]

2. Plot the 2D-scatter plot using the data generated from the above step.

3. Using eig function in numpy find the eigen values and eigen vectors for the covariance matrix C.

4. Using the above eigenvalue and eigenvector draw the line showing the principle mode of variation.

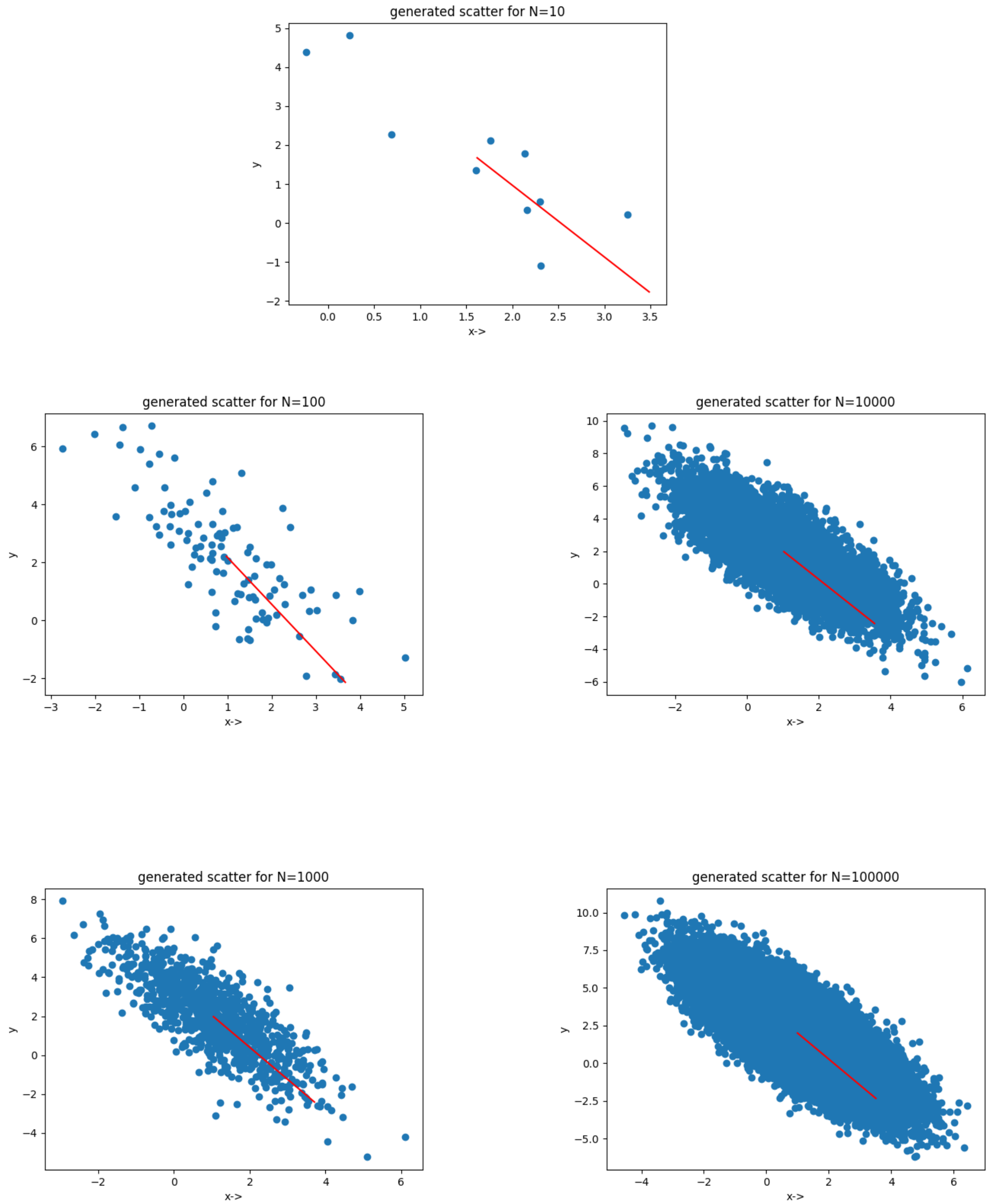
→ Principle modes of variation for drawing this here we have to pick the largest eigenvalue and the corresponding eigenvector.

We do this because eigenvalue is directly related to variance hence large eigen value ⇒ large variance ⇒ Most of the data will be spread farthest along that direction.

∴ From the plot's we can also observe that most the data is maximum spread along that direction of line we have drawn.

Also as N is increasing we can clearly see the hyper-elliptic shape of the distribution and also our line staying in the direction of major axis of the ellipse (2D-case).
Which is what we expected theoritically also.

→ The following are the scatter-plots for all N varying from 10 to 10^5.



Scatter Plots for N = 10,100,1000,10000,100000