CS215 ASSIGNMENT 2

Utkarsh Varun

Question 2

Contents

1 Introduction

I and my partner had agree to use Python programming platform for this question.
And for plotting graphs, we use pyplot function from matplotlib package.
Also, i used numpy mathematical package for coding.

I had implement the problem statement given in assignment in 3 different files.
I had also mention some comments in the code for better understanding.

Those 3 files are as follows:-
1) q2_b.py -> For plotting error in mean calculation for 5 different N values in boxplot.
2) q2_c.py -> For plotting error in covariance calculation for 5 different N values in boxplot.
3) q2_d.py -> For plotting scatter plot of generated data and modes of variation.

When you run first q2_b.py and q2_c.py file, the boxplot showing error in mean calculation for 5 different N values will be saved in the file named mean_error.png and cov_error.png respectively in the same directory.

Upon running q2_d.py, a combined plot of scatter plot of generated data and line plot of modes of variation will be saved in the file named q2_d.png.

I had also submit all plots in the "results" directory and all 3 python files in "code" directory.

2 Method for generating sample points from 2D Gaussian

Firstly, i had find the eigenvalues and eigenvectors of the given covariance matrix.

Then define two matrix :-
(1) A diagonal matrix L with diagonal element as square-root of eigenvalues.
(2) Matrix A defined as the resultant matrix of the multiplication of two matrix (eigenvector) and the above defined matrix L.
Why i had defined these 2 matrix, i had tried to proof myself in image uploaded in next page.

Thenafter, i had simply generate (N*2) random values as per the standard normal distribution.

Then, use the concept studied in the class
(X = (A * W) + Mean)
to get the desired random 2D gaussian distribution points with given mean and covariance.

The above variable X will consists of random (N*2) 2D random Gaussian points according to the mean and covariance given in Question.
If we consider high value of N, then our empirical data will conserves to the true data values/conditions.

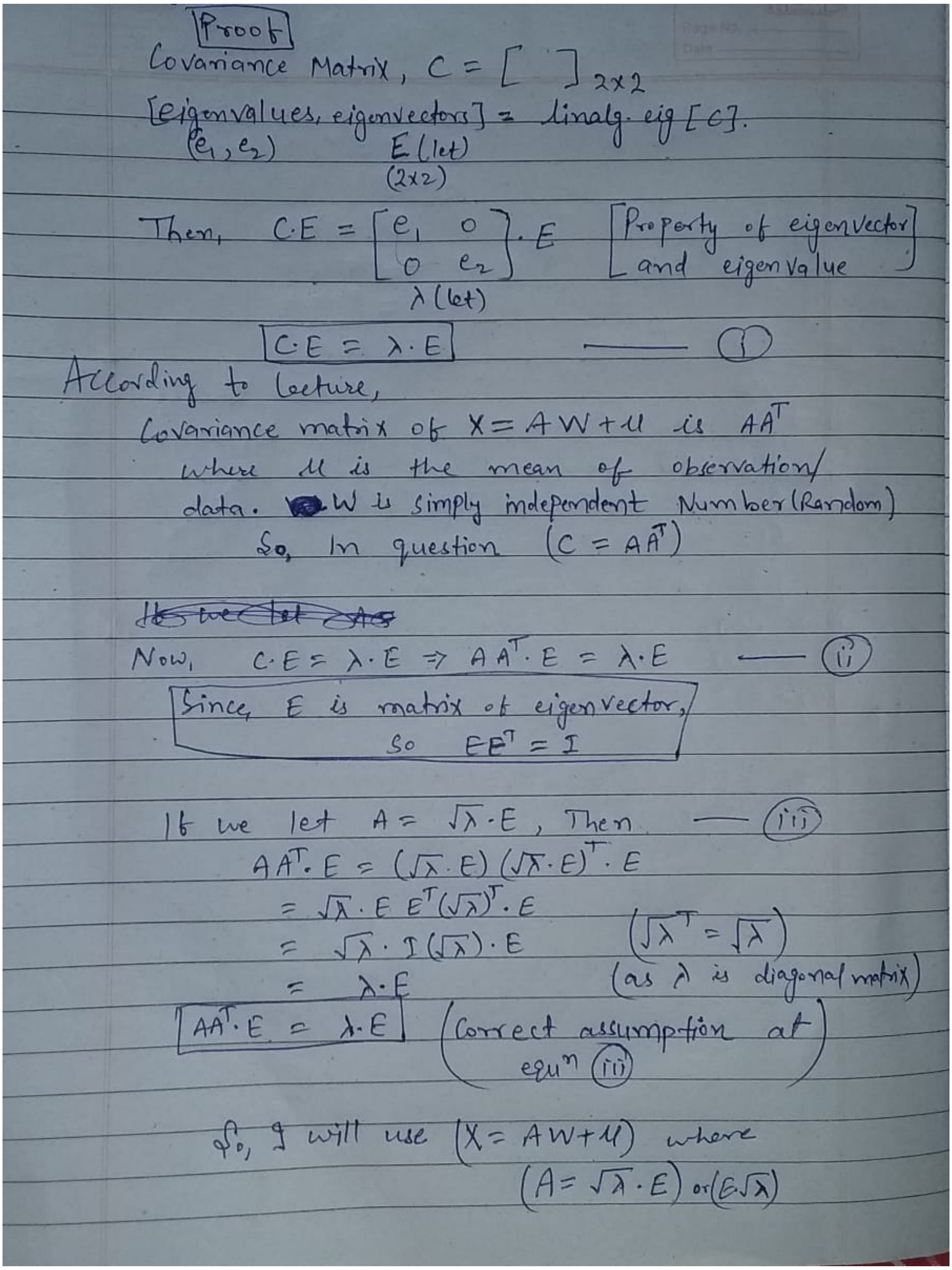
2.1 Maximum-Likelihod(ML) estimates of mean and covariance matrix

I had simply use the MLE formula of mean and covariance.

$$\text{MLE\_mean} = \frac{sum(X)}{N}$$

$$\text{MLE\_cov} = \frac{(X - MLE\_mean) * (X - MLE\_mean)^T}{N-1}$$

Here, X is the total (N) 2D gaussian random variables.



$\boxed{\text{Proof}}$

Covariance Matrix, $C = [\quad]_{2\times 2}$

$\underset{(e_1, e_2)}{[\text{eigenvalues}}, \underset{\underset{(2\times 2)}{E\,(\text{let})}}{\text{eigenvectors}]} = \text{linalg.eig}[C].$

Then, $C\cdot E = \underset{\lambda\,(\text{let})}{\begin{bmatrix} e_1 & 0 \\ 0 & e_2 \end{bmatrix}} \cdot E$ [Property of eigenvector and eigenvalue]

$$\boxed{C\cdot E = \lambda\cdot E} \quad \text{—— (i)}$$

According to lecture,

Covariance matrix of $X = AW + \mu$ is AA^T where μ is the mean of observation/data. W is simply independent Number (Random)

So, In question $(C = AA^T)$

~~If we let A=~~

Now, $C\cdot E = \lambda\cdot E \Rightarrow AA^T\cdot E = \lambda\cdot E$ —— (ii)

$\boxed{\text{Since, } E \text{ is matrix of eigenvector, so } EE^T = I}$

If we let $A = \sqrt{\lambda}\cdot E$, Then —— (iii)

$$\begin{aligned} AA^T\cdot E &= (\sqrt{\lambda}\cdot E)(\sqrt{\lambda}\cdot E)^T\cdot E \\ &= \sqrt{\lambda}\cdot E\,E^T(\sqrt{\lambda})^T\cdot E \\ &= \sqrt{\lambda}\cdot I(\sqrt{\lambda})\cdot E \\ &= \lambda\cdot E \end{aligned}$$

$(\sqrt{\lambda}^T = \sqrt{\lambda})$ (as λ is diagonal matrix)

$\boxed{AA^T\cdot E = \lambda\cdot E}$ (Correct assumption at equ^n (iii))

So, I will use $(X = AW + \mu)$ where $(A = \sqrt{\lambda}\cdot E)$ or $(E\sqrt{\lambda})$

Figure 1: This is the image of my hand written proof of generating random sample points from 2D Gaussian.

3 q2_b.py and q2_c.py

3.1 Code Explanation

Implemented the error calculation of mean in file q2_b.py and error calculation of covariance matrix in q2_c.py.

Implemented the main logic of algorithm in these file to generate 2D gaussian random points and then calculating the MLE estimate of mean and covariance matrix.

I had run the program 100 times for each N value calculating the error and appending in the list mean_error and cov_error respectively.

I hadn't use norm function to calculate the norm value of matrix to get the error, just use the mathematical formula and do the calculation to get the error.

After getting the error(100 error count for each 100 time operation for each N's value), plot the boxplot graph for each N.

3.2 q2_b.py and q2_c.py Box Plot

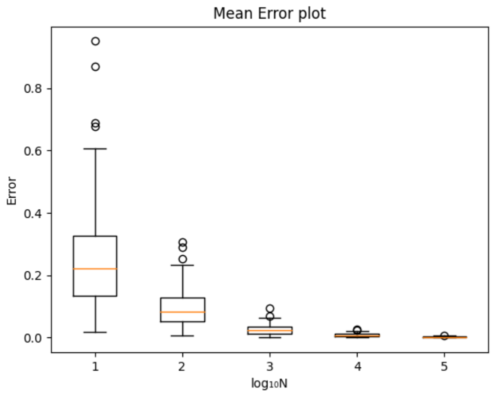


Figure 1: q2_b mean error plot

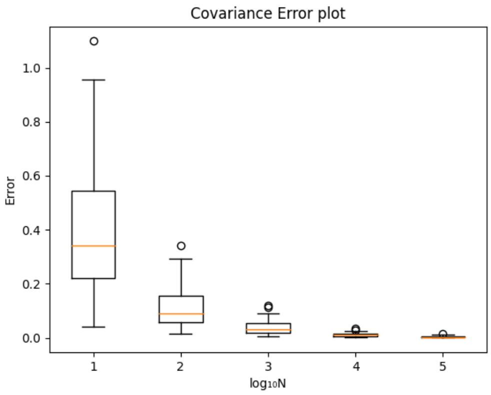


Figure 2: q2_c covariance error plot

4 q2_d.py

4.1 Code Explanation

firstly just generate random points for each N.

Then, using the concept of coordinate geometry, plot two lines which are starting from point P(x,y) = (empirical_mean[0], empirical_mean[1]) towards the two direction(of length equal to square-root of each eigenvalue) given by the eigenvectors corresponding to the respective eigenvalue.

4.2 q2_d.py Scatter plots and modes of variation

I had find the end points of both line by trigonometry algebra and then plot it.

I had also plot the scatter plot of random generated sample points along with the modes of variation represented by the above drawned line for each 5 N's value in the same plot.

4.2.1 q1_d.py Triangle 2D histogram plot

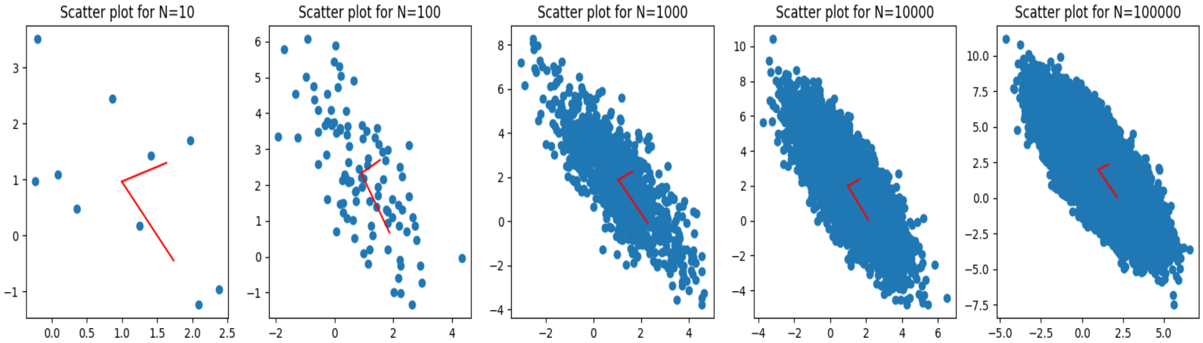


Figure 3: q2_d combined plots

Thanks