Project 1

Beth Tian

February 8, 2025

1 Problem 1

A

Mean: 0.0502
Variance: 0.0103
Skewness: 0.1204
Kurtosis: 0.2229

B

I would like to choose normal distribution. Given the question C, here I just give two reasons related to the data features and visualization.

1. According to the information about X, we can know that the amount of observations in X is 1000. This means X is a relative large sample. The standard deviation of X is 0.1 and there is no extreme high or low values.

2. The visualization of X shows that X does not have fat tail or very extreme values. The data also exhibits the symmetric and bell-shaped distribution.

C

The fitting result:
The results of normal distribution: $\mu = 0.05$, $\sigma = 0.10$
The result of t distribution: df=28.71, loc=0.05, scale=0.10

The fitting graph is:

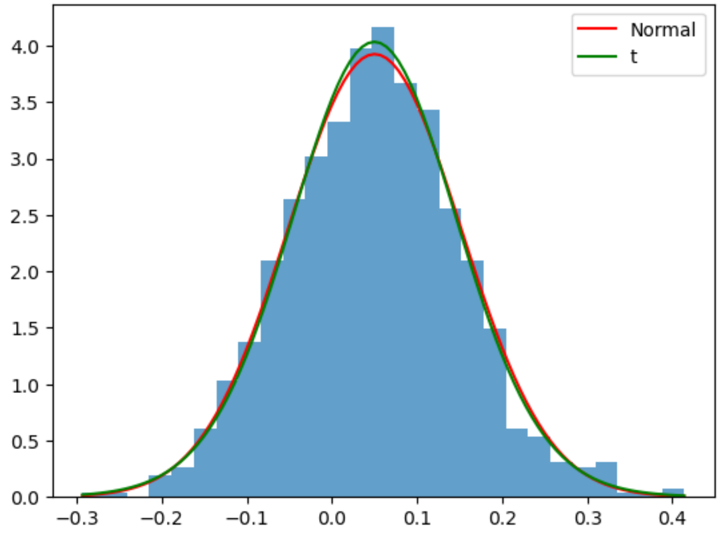


Figure 1: The fiiting visualization

Comparison:

1. Comparison through the **goodness of fit**:
Normal: $R^2 = 0.9935$, Adj $R^2 = 0.9934$
t: $R^2 = 0.9932$, Adj $R^2 = 0.9930$
From result we can know that the R^2 and Adj R^2 of normal distribution are both higher than the t distribution. This means accoring to the R^2 and Adj R^2, normal distribution fits the data better.

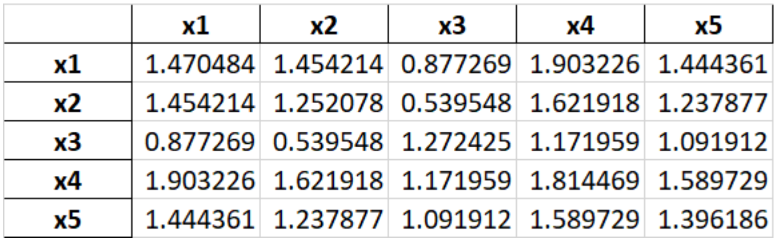
2. Comparison through **AIC/BIC**:
Normal distribution: AIC_{norm}=-1731.59, BIC_{norm}=-1721.77
t distribution: AIC_t=-1731.42, BIC_t=-1716.70
From the result of AIC and BIC of both distributions, we can find that both AIC and BIC of norm distribution are lower than that of t distribution. This means that the fitting result of norm distribution is better than t distribution.

Conclusion: Hence, the comparisons through goodness of fit and AICBIC indicate that the norm distribution fits data better.

2 Problem 2

A

The pairwise covariance matrix is:



	x1	x2	x3	x4	x5
x1	1.470484	1.454214	0.877269	1.903226	1.444361
x2	1.454214	1.252078	0.539548	1.621918	1.237877
x3	0.877269	0.539548	1.272425	1.171959	1.091912
x4	1.903226	1.621918	1.171959	1.814469	1.589729
x5	1.444361	1.237877	1.091912	1.589729	1.396186

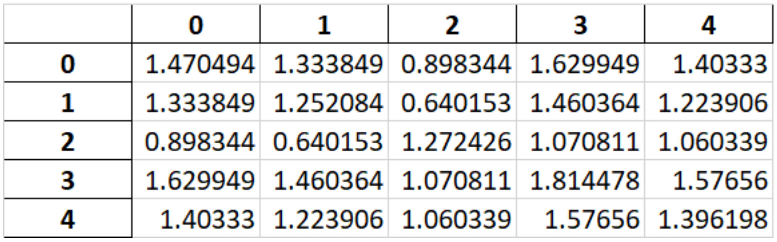
Figure 2: The pairwise covariance matrix

B

The matrix is not positive semi-definite, because at least one value in eigen vectors is less than 0. According to the code, the eigen vectors is array([6.78670573, 0.83443367, -0.31024286, 0.02797828, -0.13323183]). Because there are some eigen values of the covariance matrix less than 0, we cannot prove the matrix is a semi-definite matrix.

C

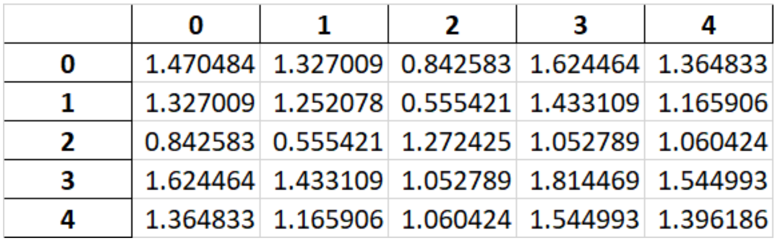
Higham's Method:



	0	1	2	3	4
0	1.470494	1.333849	0.898344	1.629949	1.40333
1	1.333849	1.252084	0.640153	1.460364	1.223906
2	0.898344	0.640153	1.272426	1.070811	1.060339
3	1.629949	1.460364	1.070811	1.814478	1.57656
4	1.40333	1.223906	1.060339	1.57656	1.396198

Figure 3: The psd matrix from Higham Method

Rebenato and Jackel Method

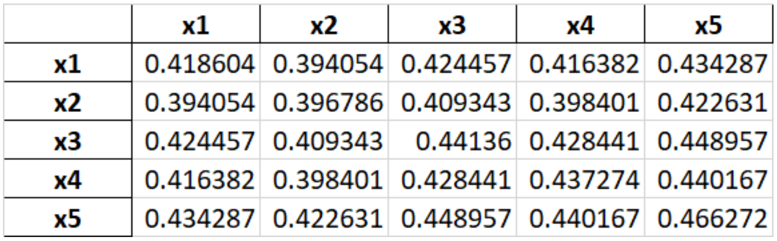


	0	1	2	3	4
0	1.470484	1.327009	0.842583	1.624464	1.364833
1	1.327009	1.252078	0.555421	1.433109	1.165906
2	0.842583	0.555421	1.272425	1.052789	1.060424
3	1.624464	1.433109	1.052789	1.814469	1.544993
4	1.364833	1.165906	1.060424	1.544993	1.396186

Figure 4: The psd matrix from Rebenato and Jackel Method

D

The covariance matrix of the overlapping data is:



	x1	x2	x3	x4	x5
x1	0.418604	0.394054	0.424457	0.416382	0.434287
x2	0.394054	0.396786	0.409343	0.398401	0.422631
x3	0.424457	0.409343	0.44136	0.428441	0.448957
x4	0.416382	0.398401	0.428441	0.437274	0.440167
x5	0.434287	0.422631	0.448957	0.440167	0.466272

Figure 5: The covariance matrix of overlapping data

E

The values in two psd matrices is similar to each other, ranging 1.0 from 1.8. The PSD matrices show greater variation in their diagonal elements. The values vary with each other. The The vaues in overlapping covariance matrix is relative small, only ranging 0.39 from 0.47. It has relatively close diagonal elements. The values are more similar within the matrix.

The overlapping data only capture relationships with the overlapping window, lack lots of information. This lead to the underestimation for the true covariance values.

3 Problem 3

A

The mean of x1 in the multivariate normal distribution is 0.046
The mean of x2 in the multivariate normal distribution is 0.0999
The covariance matrix in the multivariate normal distribution is:

	x_1	x_2
x_1	0.0101622	0.00492354
x_2	0.00492354	0.02028441

The fitting graph of multivariate normal distribution is:

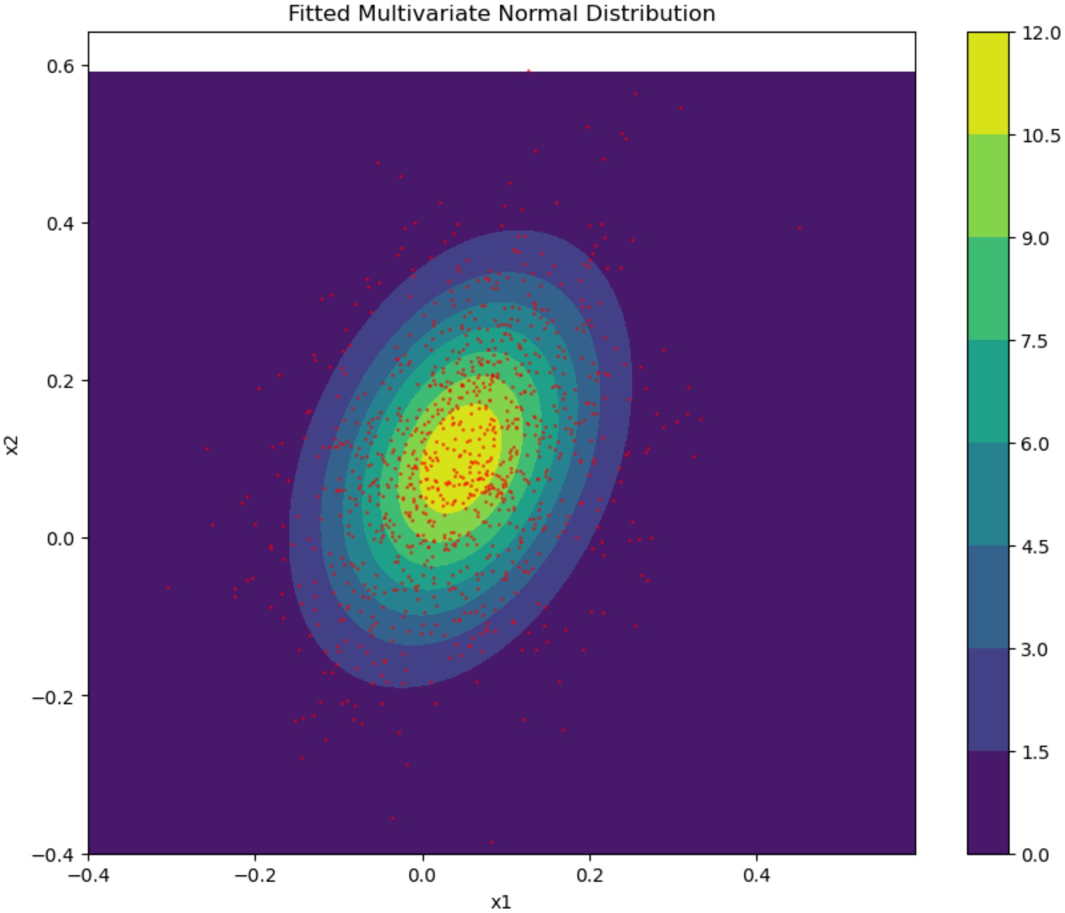


Figure 6: Visualization of multivariate normal distribution

B

1. Conditional probability distribution

The conditional expected value is 0.3683
The conditional standard deviation is 0.0179

The distribution visualization is:

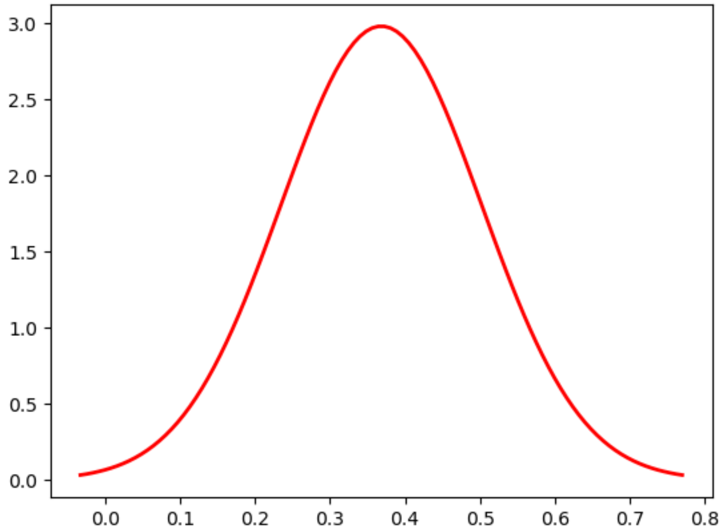
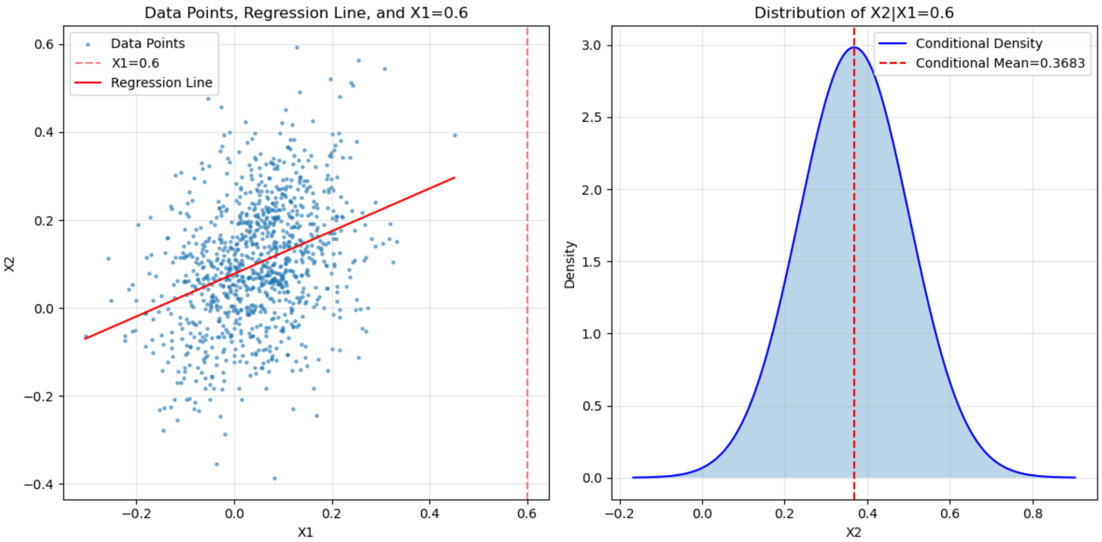


Figure 7: Visualization of conditional probability distribution

2. OLS

When X1 = 0.6:
The conditional expectation is E[X2—X1=0.6] = 0.3683
The conditional standard deviation is = 0.0179

The fitting visualization is:


Data Points, Regression Line, and X1=0.6

Distribution of X2|X1=0.6

Figure 8: Visualization of OLS fitting result

C

According to the code result:
The results from the Cholesky Root:
Theoretical Mean: 0.3683
Theoretical Variance: 0.0179

The results from the OLS:
The conditional expectation is E[X2—X1=0.6] = 0.3683
The conditional standard deviation is = 0.0179

The results from the conditional probability method:
The conditional expected value is 0.3683
The conditional standard deviation is 0.0179

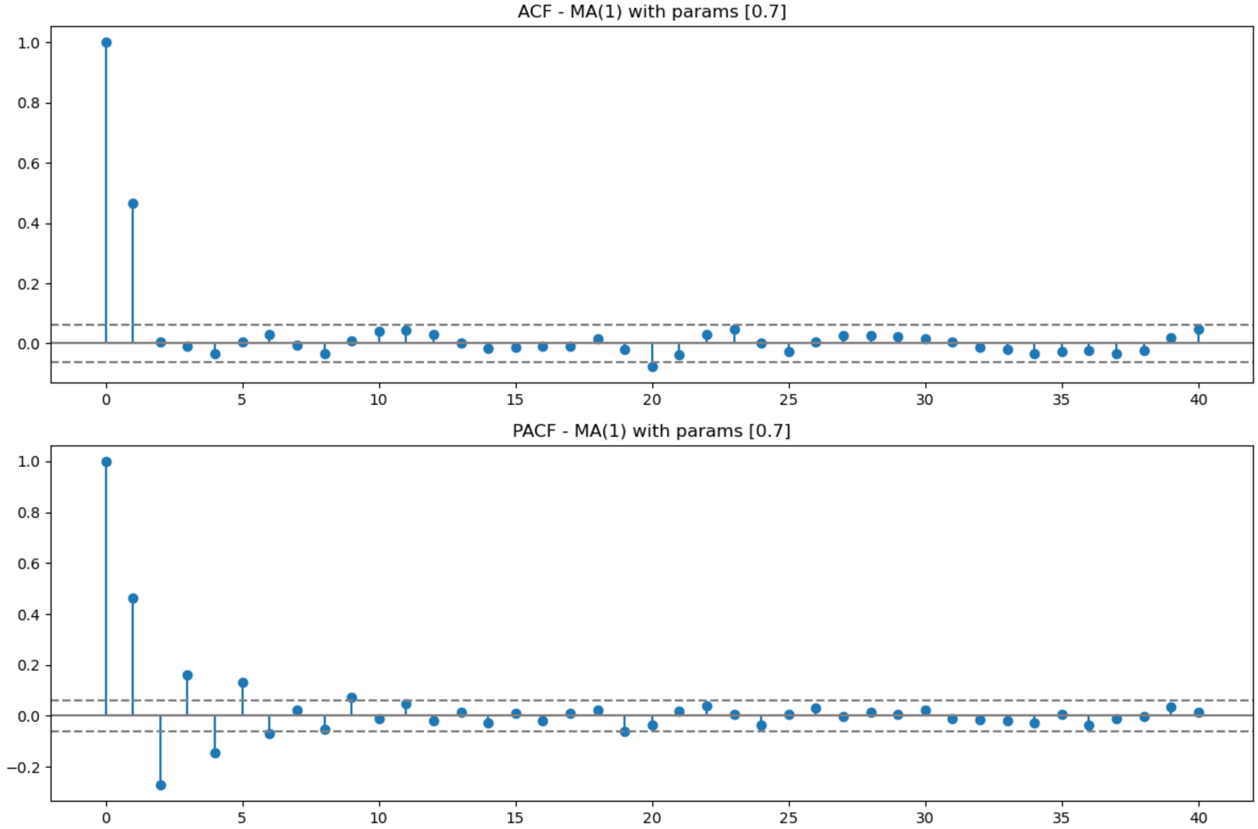
Hence, the distribution can be proved to be correct.

4 Problem 4

A

I set the parameters of MA(1) as [0.7], the parameters of MA(2) as [0.7, 0.4], the parameters of MA(3) as [0.7, 0.4, 0.2].

The pictures of ACF and PACF of MA(1), MA(2), MA(3) process can be shown as following:


ACF - MA(1) with params [0.7]

PACF - MA(1) with params [0.7]

Figure 9: ACF and PACF of MA(1)

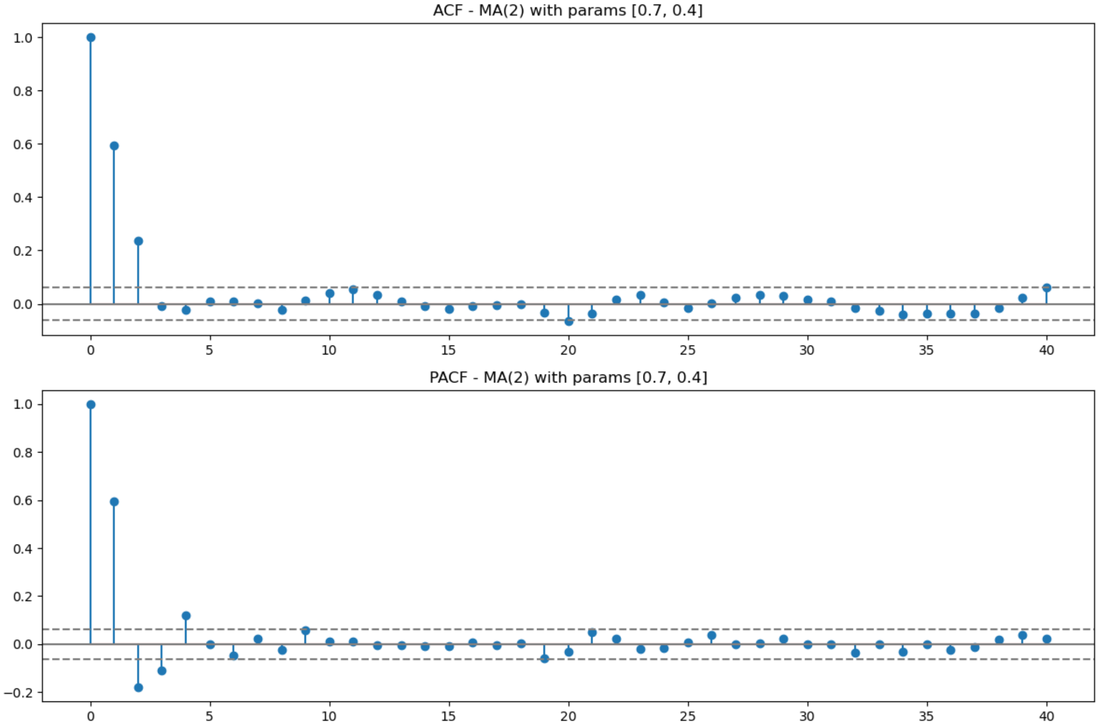


Figure 10: ACF and PACF of MA(2)

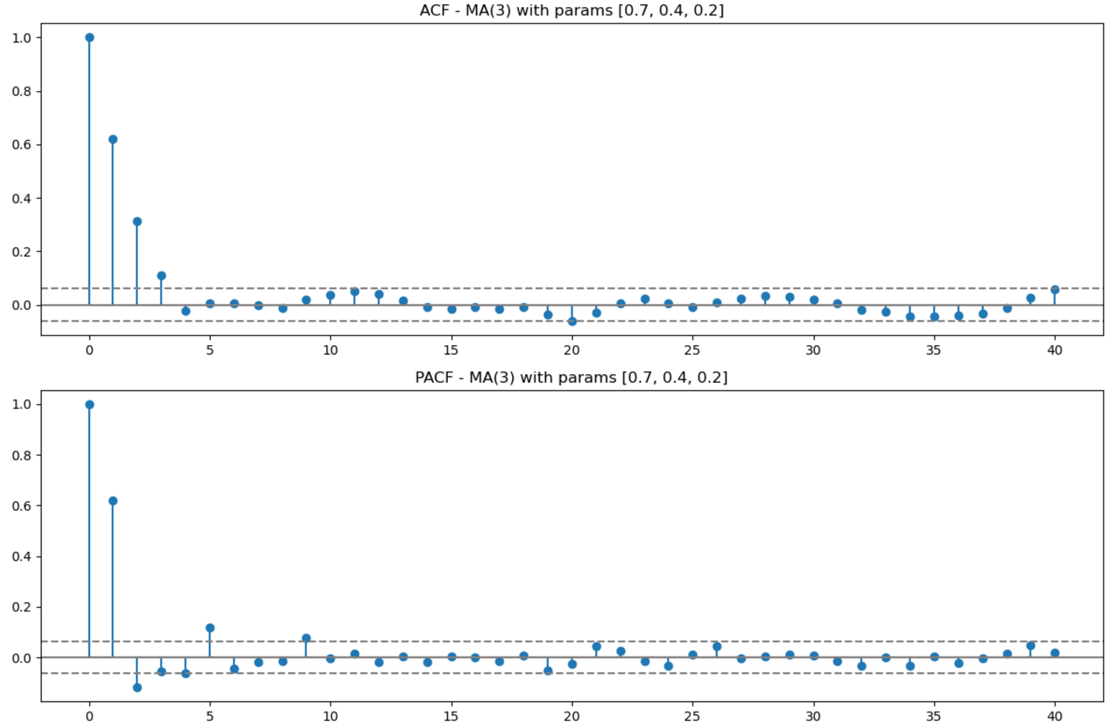
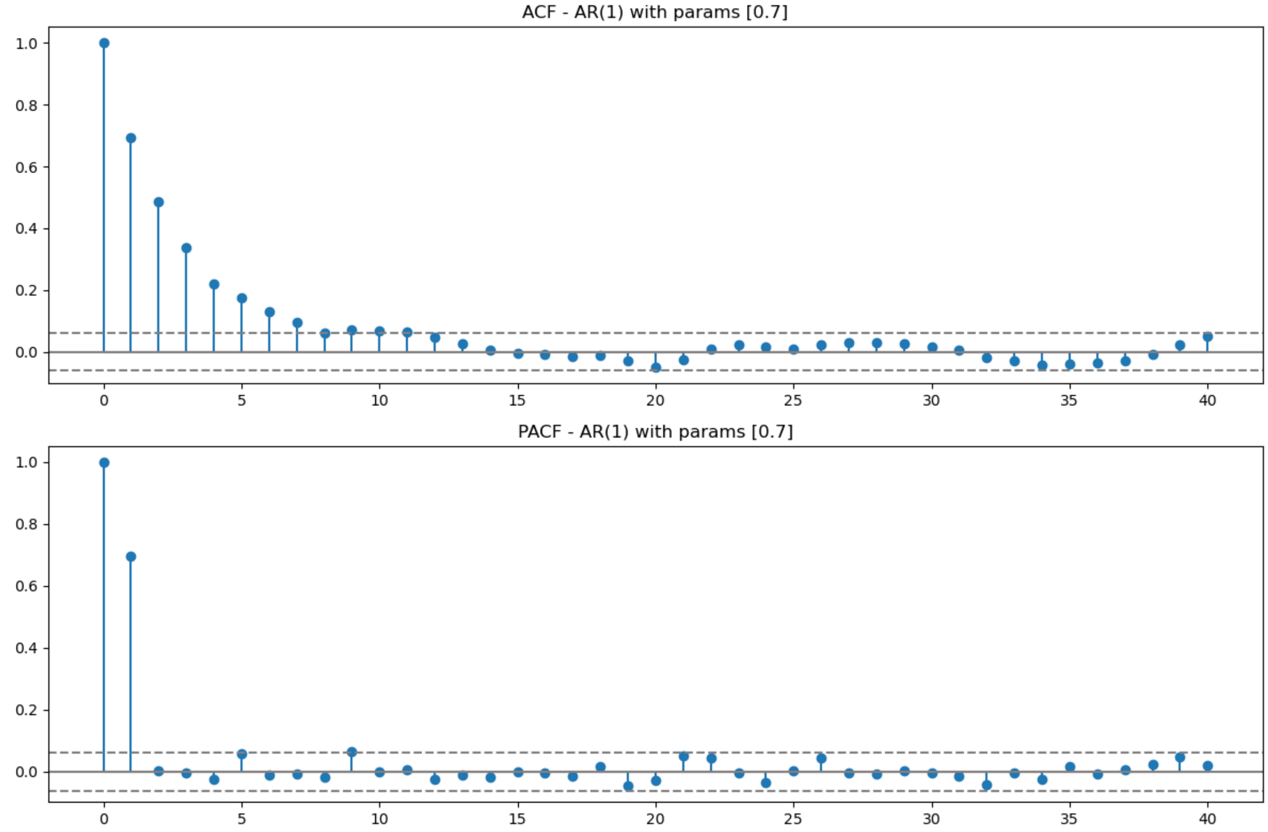


Figure 11: ACF and PACF of MA(3)

B

I set the parameters of AR(1) as [0.7], the parameters of AR(2) as [0.7, 0.4], the parameters of AR(3) as [0.7, 0.4, 0.2].

The pictures of ACF and PACF of AR(1), AR(2), AR(3) process can be shown as following:


ACF - AR(1) with params [0.7]

PACF - AR(1) with params [0.7]

Figure 12: ACF and PACF of AR(1)

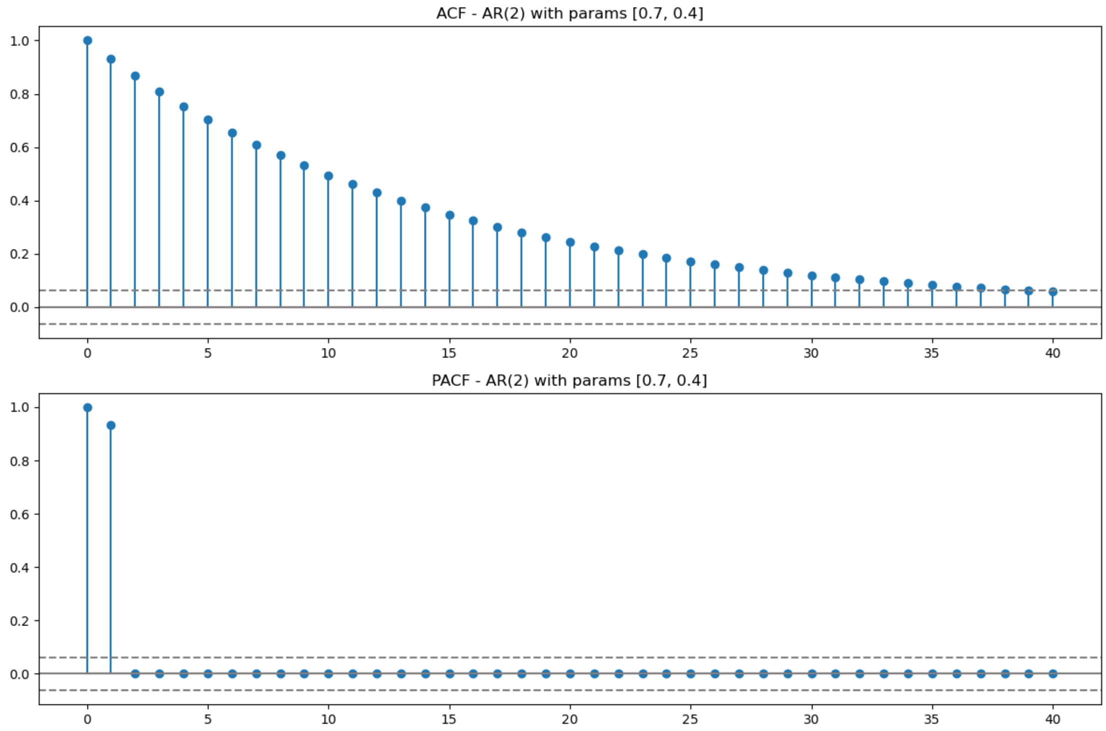


Figure 13: ACF and PACF of AR(2)

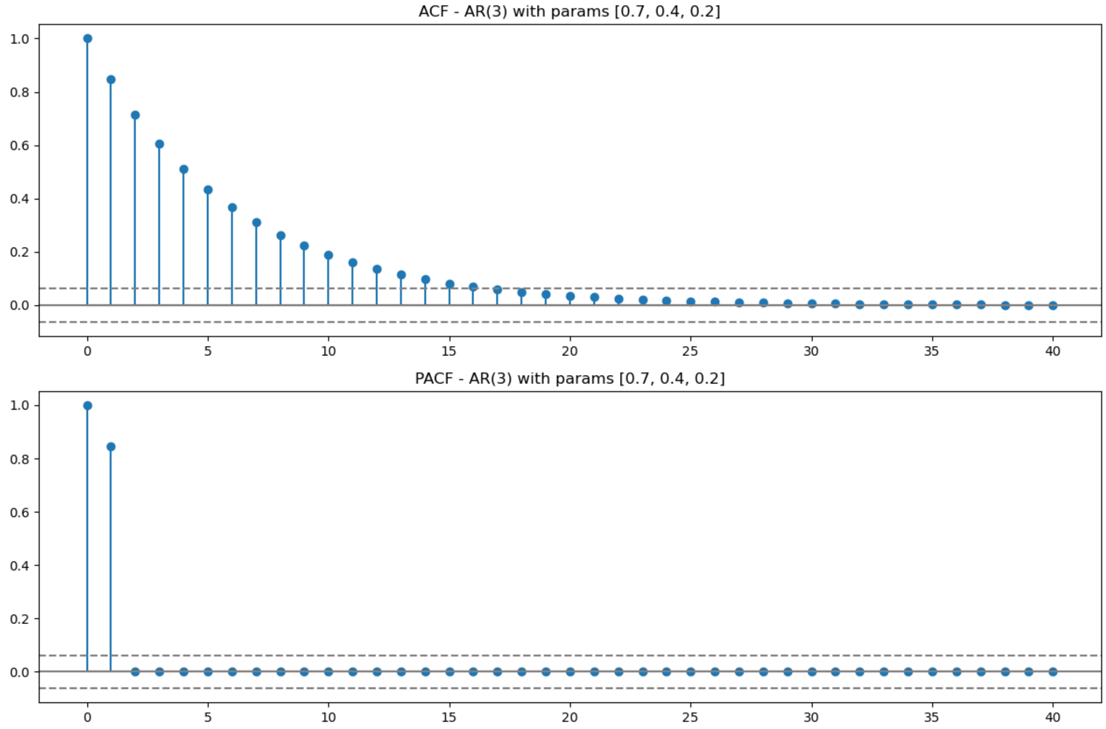


Figure 14: ACF and PACF of AR(3)

C

The ACF and PACF of the data is:

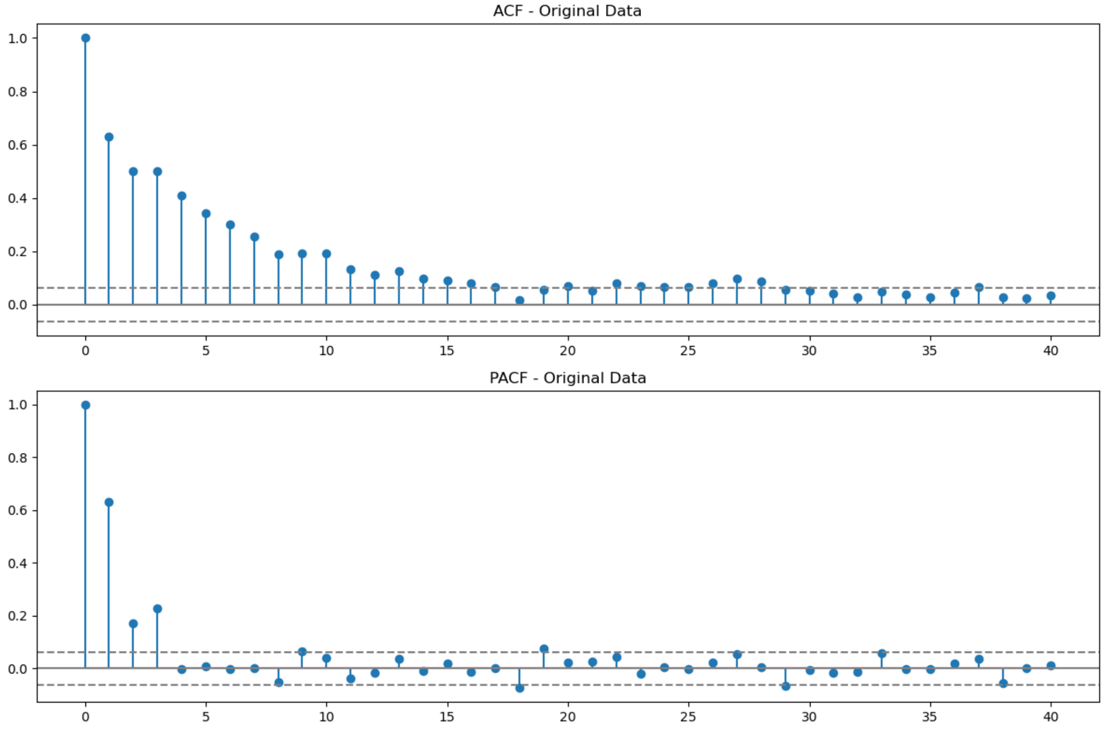


Figure 15: ACF and PACF of the data

According to the pictures of ACF and PACF, the ACF of the data is tailing off, similar to the ACF of simulation AR process. Hence, I will use the AR model.

Because the PACF of the data shows significant spikes up to lag 4, I would like to use the AR(4) model.

D

Because the spikes are not significant since 5, I tried to fit the AR model from AR(1) to AR(6).
The AICc of all models are:
AR(1): AICc = -1669.07
AR(2): AICc = -1696.05
AR(3): AICc = -1746.22
AR(4): AICc = -1744.22
AR(5): AICc = -1742.25
AR(6): AICc = -1740.23
MA(1): AICc = -1508.90
MA(2): AICc = -1559.21
MA(3): AICc = -1645.07

MA(4): AICc = -1677.50
MA(5): AICc = -1703.14

From the results we can find that the optimal model is AR(3), with lowest AICc -1744.22

I also fitted the MA model from MA(1) to MA(5). The AICc results shows that the AR(3) model fits the data best, with the lowest AICc.

The fitting result is a little bit different from my prediction. I guess the reason may be that the forth parameter does not provide sufficient explanatory power, but increases the increased model complexity. Hence, the AICc gives more punishment to the complexity than the increment to the contribution.

5 Problem 5

A

The routine has been shown in the code

B

I would like to use two sets of λ. One set of λ has large range, from 0.1 to 1, with interval 0.1. The other set of λ has small range, from 0.8 to 1, with interval 0.04. We have already known that the optimal λ is 0.97 or 0.94, so the large λ may be more practical in business or daily analysis. The goal of two sets is to reflect the relationship in a general large range and a specific practical small range.

The cumulative variance visualization of different λ is:

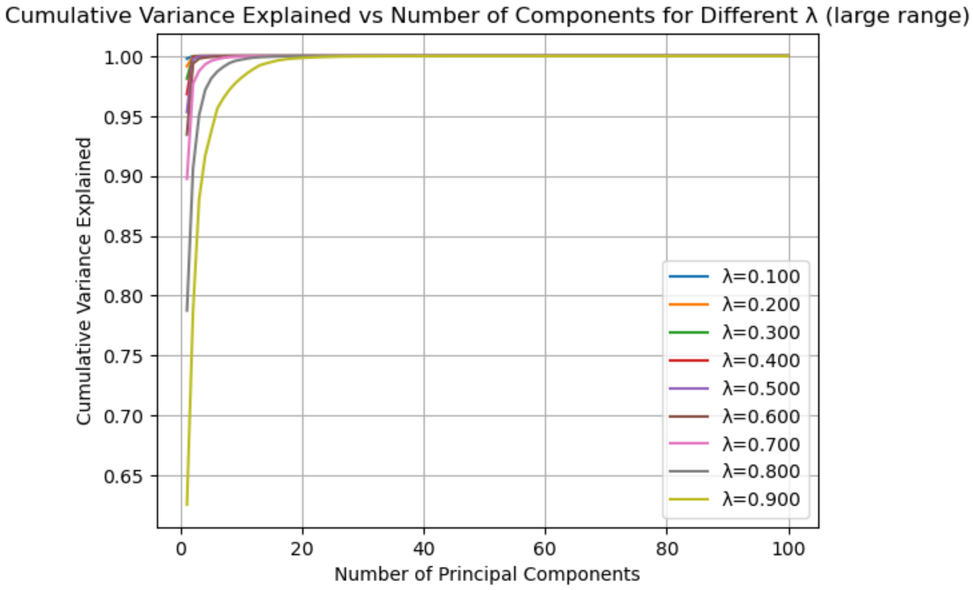


Figure 16: Cumulative Variance explained ny large range λ

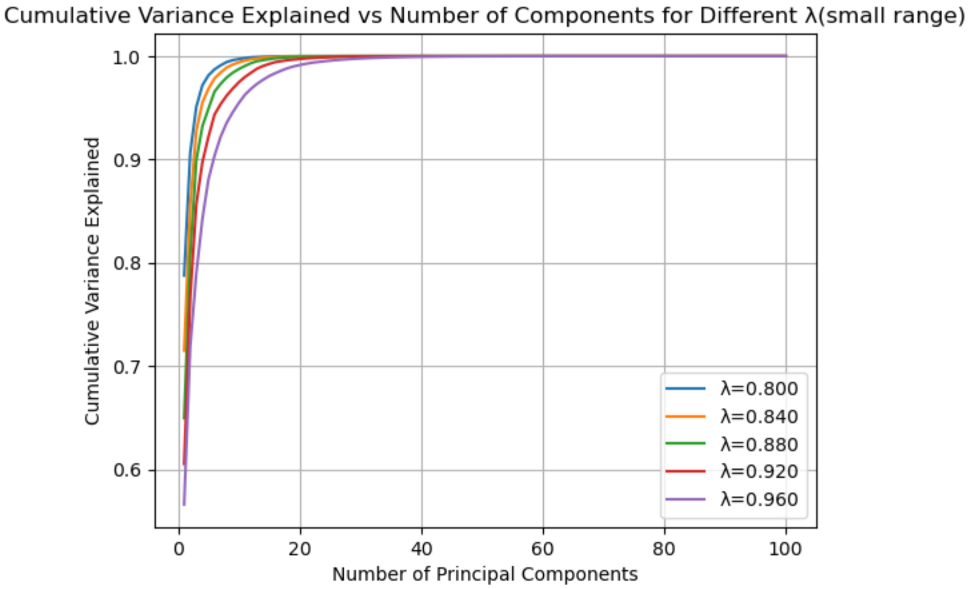


Figure 17: Cumulative Variance explained ny small range λ

C

The higher the λ is, more principle components needed to explain the variance. The smaller λ means the weight of recent data is higher, leading to more concentrated data.

6 Problem 6

A

Because the covariance matrix is not positive semi-definite, I would like to use the near-psd method of Rebenato and Jackel to find th suitable matrix first. Then I will use the cholskey factorization to get the cholskey root.

The process is shown in the codes.

B

The process is shown in the codes.

C

The comparison Frobenius of Cholesky method is 0.021028
The comparison Frobenius of PCA is 0.082998

The Frobenius norm of PCA simulation (0.083) is higher than Choleskey simulation (0.021). This means that the covariance matrix of PCA simulation is much more different from that of Cholesky simulation. It might indicate that the data of PCA is much more different from original data. The reason here might be that the dimension reduction of PCA also lose some information from original data. This leads to the difference between them.

D

The cumulative variance explained of three covariance matrices are shown as following graph:

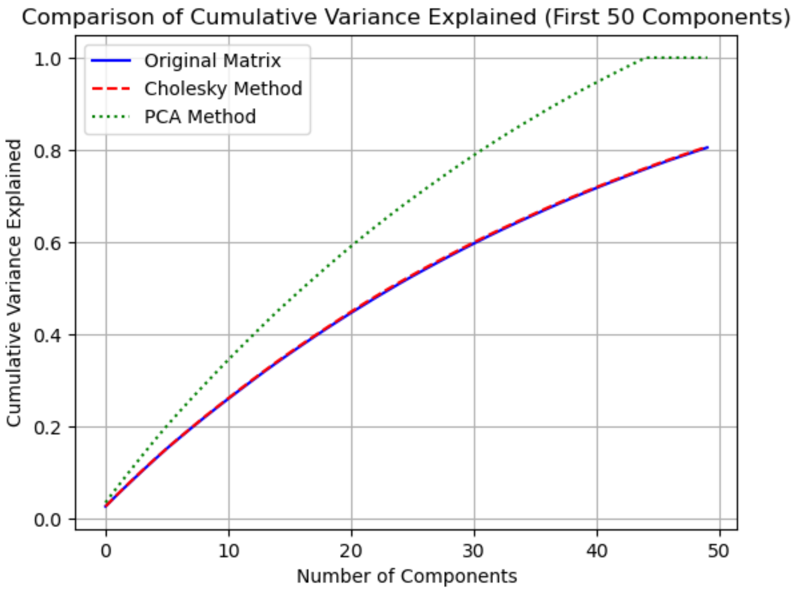


Figure 18: Cumulative Variance explained of three covariance matrices

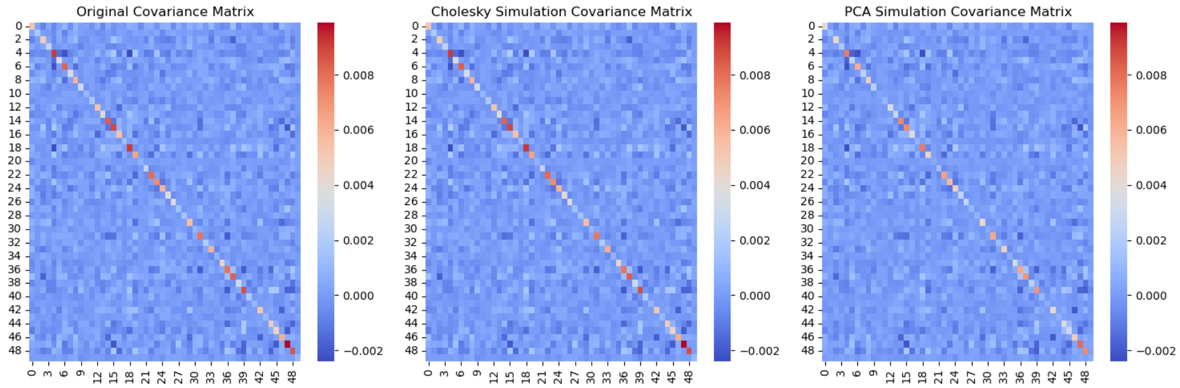


Figure 19: The heatmap comparison of three covariance matrices

The cholesky method keep the same covariance structures as the original data, because it overlaps the original matrix curve. The Cholesky better maintains the original data's complexity and correlation structure.

The PCA method shows the fatest growth in cumulative variance explained, reaching nearly 100% with about 40 components. Because PCA is used for the dimension reduction, it appears to concentrate the data variance more aggresively. For dimensionality reduction or data compression purposes, the PCA method might be more suitable as it explains more variance with fewer components

E

The runtime of PCA is 0.3297 seconds.
The runtime of Cholesky method is 0.5651 seconds.
The PCA method can process the simulation faster than the Cholesky method.

F

From the perspective of time, the PCA is faster than the Cholesky. This difference is much more obvious when the data set size is bigger. Moreover, the calculation of Cholesky roots and covariance are easily to crash the Python kernel, which means it may need more computility or memory to calculate.

From the perspective of Frobenius norm and cumulative variance explanation, the PCA and Cholesky has their own advantages.

The PCA can reduce the dimensions, making the date more concentrated and aggregated. The data structure of PCA simulation is more simple. It needs less components to ccontribute to higher variance explanation. However, the reduction of dimensions also leads to the loss of original information.This might increase the difference between the original data and simulation data.

The Cholesky can keep the data structure as the original data. The simulation data also similar to the origianl one. However, the data structure of it maybe more complex. It also needs more components to contribute to the variance.

Hence, when it comes to rapid data processing or preliminary data exploration, PCA offers a more efficient approach, providing a good balance between computational speed and variance explanation. For risk management or other purposes, the Cholesky method would be preferable as it better preserves the risk structure.