Problem 1
Given the dataset in problem1.csv
A. Calculate the Mean, Variance, Skewness and Kurtosis of the data

- Mean: 0.05019795790476916
- Variance: 0.010332476407479581
- Skewness: 0.1206257259522635
- Kurtosis: 0.23006981317028652

B. Given a choice between a Normal Distribution and a T-Distribution, which one would you choose to model the data? Why?

- The sample size is 1000 and larger than 30. I'll choose the normal distribution method. The sampling distribution of the mean is approximately the mean for the population if the sample size is large enough.
- The Shapiro-Wilk test and Q-Q plot could check if the data follows a normal distribution. If the test statistic is close to 1 then it could tell that the data follows a normal distribution.
- I'll use the normal distribution method if the population standard deviation is known. If it is unknown, then I'll use t-distribution.
- Overall, with a statistically significant sample size and Shapiro-Wilk test statistics which is close to 1, I'll choose to use normal distribution.

C. Fit both distributions and prove or disprove your choice in B using methods presented in Class.

	Moment	Original Data	Normal Fit	T_Distribution Fit
0	Mean	0.050198	0.049238	0.049706
1	Variance	0.010332	0.010321	0.010860
2	Skewness	0.120445	0.009246	-0.014636
3	Kurtosis	0.222927	-0.267684	0.328657

Both distributions approximate the variance and mean well. The original data for skewness is 0.120445. The normal distribution is close to the original data and positive. The negative skewness generated from the t-distribution is incorrect. For Kurtosis, t-distribution is close to the original data. The data doesn't disprove my choice in B. I still think t-distribution is a better way in this case.

Problem 2
Given the data in problem2.csv
A. Calculate the pairwise covariance matrix of the data.

	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

B. Is the Matrix at least positive semi-definite? Why?

- No, at least one of its eigenvalues is negative. There are two negative eigenvalues (-0.31024286 and -0.13323183). A matrix is positive semi-definite if all of its eigenvalues are non-negative.

C. If not, find the nearest positive semi-definite matrix using Higham's method and the near-psd method of Rebenato and Jackel.

	x1	x2	x3	x4	x5
x1	1.615133	1.441960	0.897144	1.780426	1.433794
x2	1.441960	1.346968	0.585086	1.554552	1.211409
x3	0.897144	0.585086	1.298916	1.115956	1.076692
x4	1.780426	1.554552	1.115956	1.983165	1.621373
x5	1.433794	1.211409	1.076692	1.621373	1.404936

D. Calculate the covariance matrix using only overlapping data.

- I have selected the rows without missing values.

	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	0.441360	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

E. Compare the results of the covariance matrices in C and D. Explain the differences.
Note: the generating process is a covariance matrix with 1 on the diagonals and 0.99 Elsewhere.

- The psd covariance matrix is higher than the covariance using overlapping data
- On the diagonal elements, there is a large difference. The variance is higher by using psd methods than overlapping-data methods.
- The overlapping-data methods may underestimate correlation since off-diagonal elements are higher.
- The psd methods ensure the covariance remains positive semi-definite.

Problem 3
Given the data in problem3.csv

A. Fit a multivariate normal to the data.

```
[0.04600157 0.09991502] [[0.0101622  0.00492354]
 [0.00492354 0.02028441]]
```

Mean: 0.04600157, 0.09991502

Covariance:

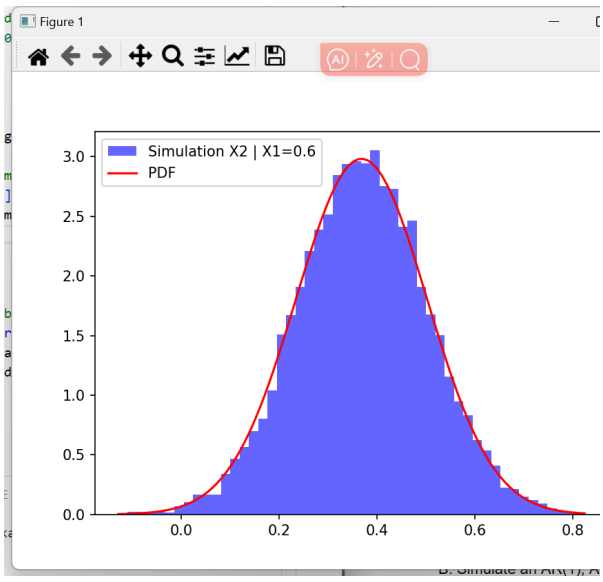
0.0101622	0.00492354
0.00492354	0.02028441

B. Given that fit, what is the distribution of X2 given X1=0.6. Use the 2 methods described in class

- The first method is conditional distribution formula
  - μ1 = 0.0460, μ2 = 0.0999
  - σ11 = 0.01016, σ22 = 0.02028, σ12 = 0.00492
  - E[X2|X1 = x1] = μ2 + σ12 / σ11 (x1 - μ1) =0.3683
  - Var(X2|X1) = σ22 - σ12 ^ 2 / σ11 = 0.0179
- The second method is Regression Interpretation
  - X2 = β0 + β1X1 + e
  - β1 = Cov(X1, X2) / Var(X1) = 0.00492 / 0.01016 = 0.4842
  - β0 = μ2−β1μ1 = 0.0999 - (0.4842 * 0.0460) = 0.0776
  - E[X2|X1 = x1] = β0+β1x1 = 0.0776 + (0.4842 * 0.6) = 0.3681
  - Var(X2|X1) = σX2 ^ 2 (1-p^2) = σ12 / σX1 * σX2 = 0.00492 / rad0.01016 * rad0.02028 = 0.3432
  - Var(X2|X1) = 0.02028 * (1 - 0.03432^2) = 0.0179

C. Given the properties of the Cholesky Root, create a simulation that proves your distribution of X2 | X1=0.6 is correct.

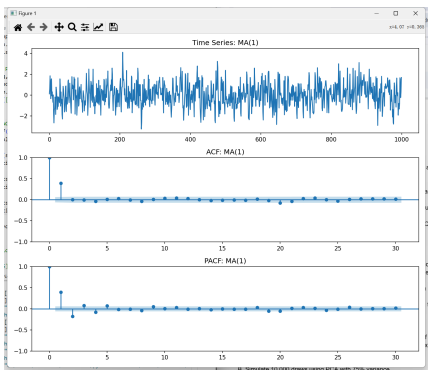
- Shown in the Code
-

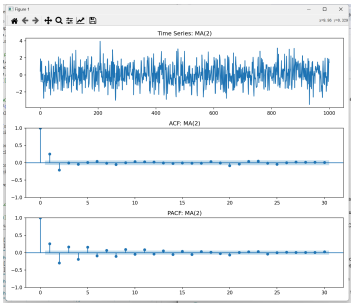


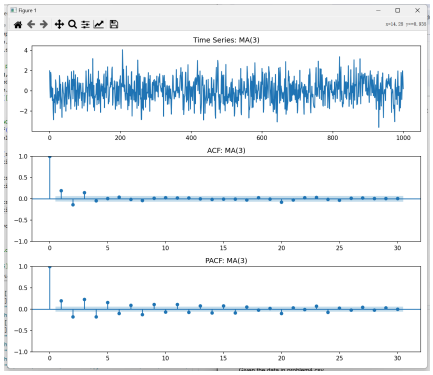
Problem 4

Given the data in problem4.csv

A. Simulate an MA(1), MA(2), and MA(3) process and graph the ACF and PACF of each. What do you notice?

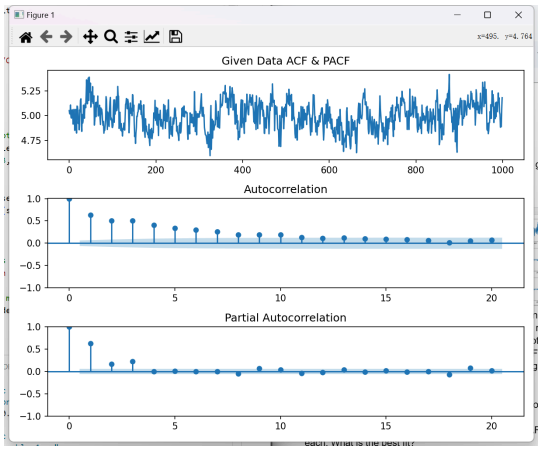


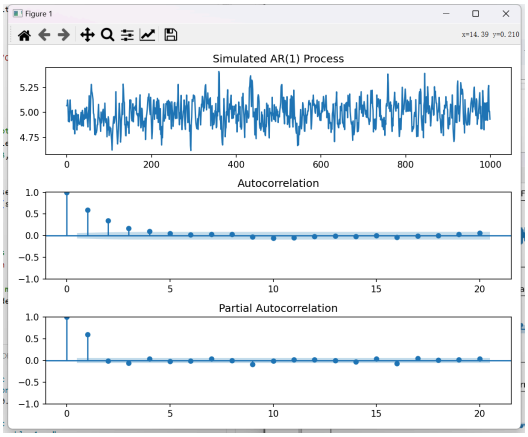


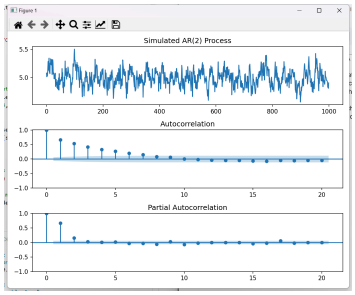


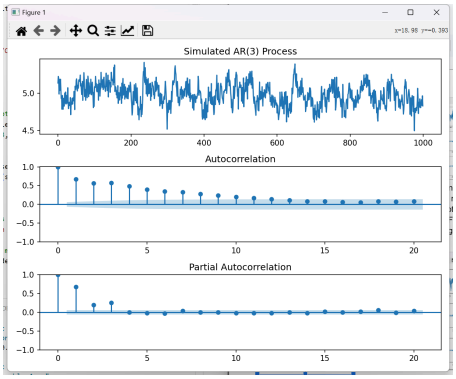
- MA(1): ACF shows spike at lag 1; PACF does not show spike
- MA(2): ACF shows spike at lag 1 and 2; PACF remains bounded
- MA(3): ACF shows correlations; PACF does not show spike
- ACF shows large values only up to lag q; PACF shows a more gradual decline

B. Simulate an AR(1), AR(2), and AR(3) process and graph the ACF and PACF of each. What do you notice?









- AR(1): ACF shows an decay; PACF shows autoregressive structure
- AR(2): ACF shows an oscillatory decay; PACF cuts off after lag 2

- AR(3): CF shows a complex circumstance: PACF cuts off after lag 3

C. Examine the data in problem4.csv. What AR/MA process would you use to model the data? Why?

- The best model would be AR(3) and MA(0) since it shows an autoregressive model with three lagged terms

D. Fit the model of your choice in C along with other AR/MA models. Compare the AICc of each. What is the best fit?

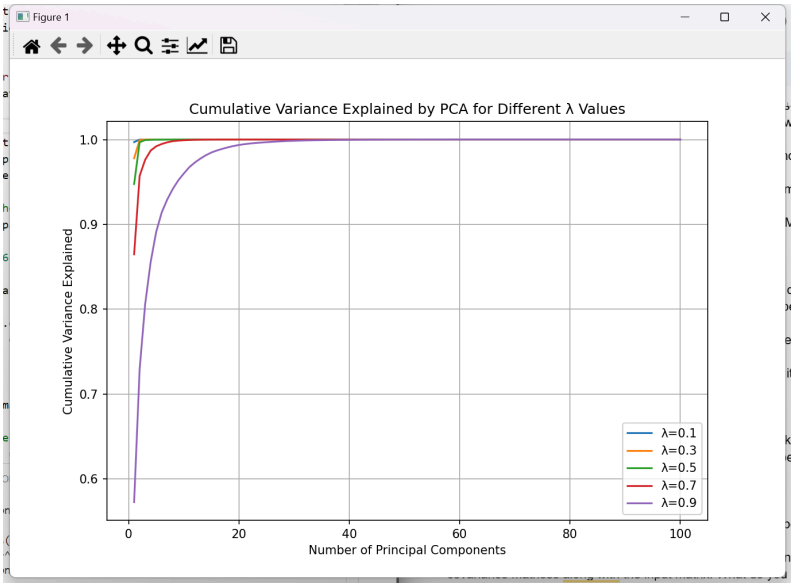
- Lowest AICs value is -1746.22. It best fits AR(3) and MA(0)

Problem 5

Given the stock return data in DailyReturns.csv.

A. Create a routine for calculating an exponentially weighted covariance matrix. If you have a package that calculates it for you, verify it produces the expected results from the testdata folder.

B. Vary λ. Use PCA and plot the cumulative variance explained of λ in (0,1) by each eigenvalue for each λ chosen.



C. What does this tell us about the values of λ and the effect it has on the covariance Matrix?

- Lower lamda make the covariance matrix more reactive to the market. The pCA curve for lower lamda prone to increase more gradually. However, higher lamda put more equal weight on past and present data. It makes the matrix more stable and smoother.

Problem 6

Implement a multivariate normal simulation using the Cholesky root of a covariance matrix. Implement a multivariate normal simulation using PCA with percent explained as an input. Using the covariance matrix found in problem6.csv

A. Simulate 10,000 draws using the Cholesky Root method.

B. Simulate 10,000 draws using PCA with 75% variance

C. Take the covariance of each simulation. Compare the Frobenius norm of these matrices to the original covariance matrix. What do you notice?
D. Compare the cumulative variance explained by each eigenvalue of the 2 simulated covariance matrices along with the input matrix. What do you notice?
E. Compare the time it took to run both simulations.
F. Discuss the tradeoffs between the two methods.