Will Neuner
Midterm Fintech 545
10/20/2025

1 Explain the difference in thinking between data modeling for risk analysis vs data modeling for forecasting

When it comes to data modeling, forecasting and risk analysis are completely separate tasks that utilize very different mindsets. While the data being worked with is often the same, the way of thinking about that data is vastly different.

When modeling data for forecasting, we are often concerned mostly with what the expected value of some variable will be. Methods like linear regression and other machine learning models are mostly concerned with finding the best “fit” for our data. That is, creating a model that can predict an outcome with the best accuracy or lowest loss possible. While there may be some outliers occasionally, we are generally happy with a model if it is able to capture and predict a target variable somewhat closely.

However, in the world of risk analysis, those outliers are exactly the thing we must be worried about. While forecasting is concerned with what the expected value of a stock may be, risk analysts are more concerned about the distribution of values the stock could move to. This risk analysis comes in the form of modeling variance (risk) and visualizing what worst case scenarios may be. Once an analyst has an idea of what their 5%, 1%, 0.1%, etc worst case scenario may be, they can set up their assets such that their institution can be prepared in the event that those situations occur.

In short, while forecasters are busy attempting to predict what will happen, risk analysts concern themselves with understanding everything that could happen.

2. Using problem 2

a. Calculate the Mean, Variance, Skewness and Kurtosis of the data (8)

Mean: -0.0003457749550813146
Variance (unbiased): 0.000485
Skew (unbiased): 0.11425266
Kurtosis (unbiased, excess): 0.96838872

b. Given a choice between a normal distribution and a t-distribution, which one would you choose to model the data and why based on part a alone?

I would choose to model the data with a t-distribution, as given that we only have a sample of the data, we do not know the population standard deviation of our variable. The t-distribution incorporates fatter tails to account for this uncertainty in the second moment of our distribution.

c. Fit both distributions and prove or disprove your choice in b)

Normal distribution

- Mean: -0.0003457749550813146
- Standard Deviation: 0.02203232224742731

T distribution

- Mean: -0.000477589421055936
- Standard Deviation: 0.019398966622270863
- Degrees of freedom: 8.85936027362565

After fitting the two models, we can see that the standard deviation for the normal distribution is much higher than for the T distribution. This revelation indicates that our sample distribution has the fat tails mentioned in part b. If a normal distribution were the correct choice for this data, then our t-distribution should have a much higher number of degrees of freedom, indicating that it has converged to a normal distribution. This is not the case, so we can be confident that a t-distribution is a better model for the data

3. Using problem2.csv and your fitted models. These are returns of a stock

a. Calculate the VaR (5% alpha) as distance from 0 for both models (8)

   VaR for normal: 0.03623994515884444
   VaR for t: 0.035624875050607886

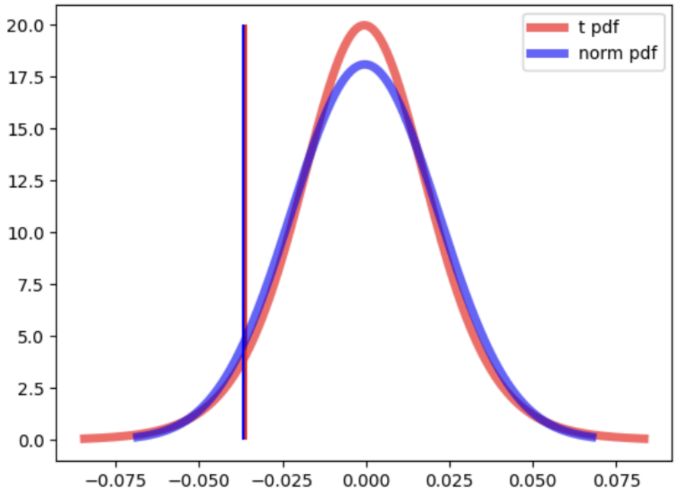
b. Calculate the ES (5% alpha) as the distance from 0 for both models (8)

   ES for normal: 0.045792128233980406
   ES for t: 0.048230230954124355

c. Discuss the results. What do you notice? Why is that? (4)

The VaR for the normal distribution was higher than for the t distribution. However, the ES for the t distribution was higher than for the normal distribution. This seems odd at first, as the two risk metrics each declare that a different distribution is riskier. However, when thinking about the definition of VaR and ES, we can see why this phenomena occurs. The t-distribution in this case has a smaller standard deviation with fatter tails, while the normal distribution has a larger standard deviation with slimmer tails. This means that the VaR for the normal distribution should be higher, as the 5% quantile point will reflect a greater loss for the normal distribution than for the t-distribution. However, since the t-distribution has more probability mass in its tails, once we condition on our loss being worse than a 5% day, the expected value of our loss is greater for our t-distribution than for our normal distribution.

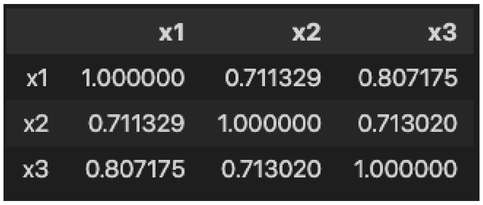
This plot shows the pdfs of our normal and t distributions. As was described, the normal distribution's VaR, shown by the blue vertical line, is just barely larger in magnitude than the t distribution VaR. However, it is clear to see that the t distribution has much fatter tails that lead to the greater ES that we see in the earlier parts.



While this initially seems strange, it actually highlights the key differences between normal and t distributions.

4. Using problem4.csv

a. Calculate the exponentially weighted correlation matrix with lambda = 0.94

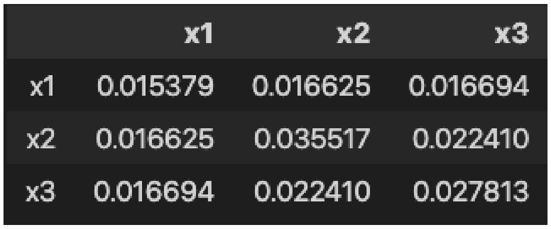


	x1	x2	x3
x1	1.000000	0.711329	0.807175
x2	0.711329	1.000000	0.713020
x3	0.807175	0.713020	1.000000

b. Calculate the exponentially weighted variances with lambda=0.97

- X1 = 0.01537881
- X2 = 0.03551743
- X3 = 0.02781346

c. Combine A and B to form a covariance matrix



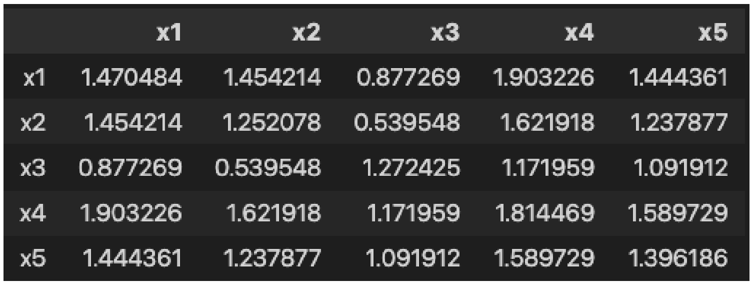
	x1	x2	x3
x1	0.015379	0.016625	0.016694
x2	0.016625	0.035517	0.022410
x3	0.016694	0.022410	0.027813

d. Why would you do something like this in practice?

This method is helpful for separating how the variance and correlation weights are computed. When just computing the exponentially weighted covariance, you are locked into one lambda value for your exponential weight decay. However, computing covariance this way allows you to select two lambda values, one for correlation and one for variance. This means that if, for example, you think that older historical values have a greater effect on your data's current variances than for its correlations, you can change your lambda values accordingly for your two computations in an attempt to better model your data.

5. (30 pts) Using the data in problem5.csv. These data contain missing values

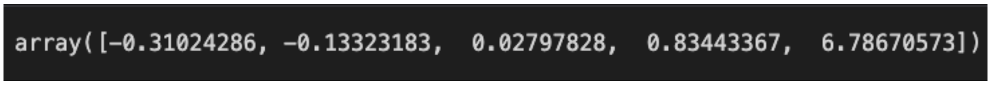
a. Calculate the pairwise covariance of the data



	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 your matrix Positive Definite, Positive Semi-definite, or Non Definite?

The eigenvalues of this covariance matrix are

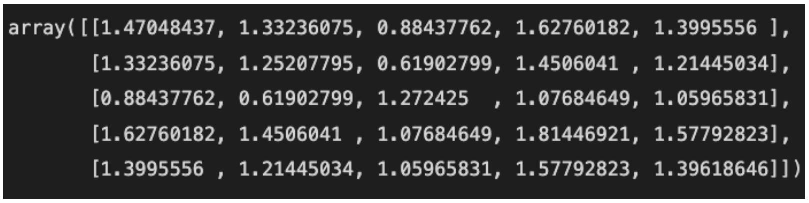


```
array([-0.31024286, -0.13323183,  0.02797828,  0.83443367,  6.78670573])
```

Since some of these eigenvalues are negative, our matrix is Non-Definite.

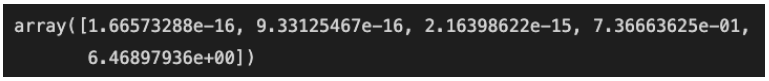
c. If the matrix is non definite, use Higham's method to fix the matrix

Higham corrected covariance matrix



```
array([[1.47048437, 1.33236075, 0.88437762, 1.62760182, 1.3995556 ],
       [1.33236075, 1.25207795, 0.61902799, 1.4506041 , 1.21445034],
       [0.88437762, 0.61902799, 1.272425  , 1.07684649, 1.05965831],
       [1.62760182, 1.4506041 , 1.07684649, 1.81446921, 1.57792823],
       [1.3995556 , 1.21445034, 1.05965831, 1.57792823, 1.39618646]])
```

For confirmation, the new eigenvalues of this corrected matrix are



```
array([1.66573288e-16, 9.33125467e-16, 2.16398622e-15, 7.36663625e-01,
       6.46897936e+00])
```

Since all of these eigenvalues are positive, the Higham method has done its job and given us a Positive semi definite matrix to work with.

d. For each principal component, list the variance explained and the cumulative variance explained, sorted from largest to smallest variance explained.

PC number	Principal component	Variance Explained	Cumulative Variance Explained
1	6.46897936e+00	8.97765734e-01	0.8977657335513296
2	7.36663625e-01	1.02234266e-01	0.9999999999999997

3	2.16398622e-15	3.00318268e-16	0.9999999999999999
4	9.33125467e-16	1.29499264e-16	1.0
5	1.66573288e-16	2.31170609e-17	1.0

6. Using problem6.csv. These data are prices of 3 stocks. You own 100 shares of each stock. Using arithmetic returns:

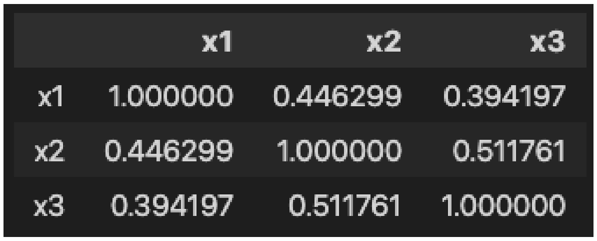
a) De-mean the return series so that the mean of each is 0. Fit a Student T model for each stock. Report the fit values.
Stock1:
- Mu = -0.0004786520929827172
- Sigma = 0.012907549628639717
- Nu = 4.729830909131829

Stock2:
- Mu= -4.350220543514708e-05
- Sigma= 0.009058419601620012
- Nu = 6.766945042505089

Stock 3:
- Mu= 7.48934264331786e-05
- Sigma= 0.01706272432180566
- Nu= 39.864383658541556

b) Simulate the system using a Gaussian Copula. Report the correlation matrix you used in the copula

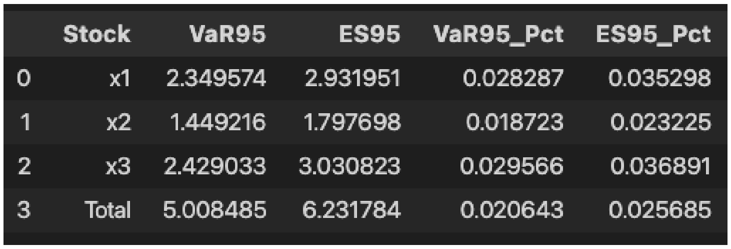
Correlation matrix computed from quantile vectors using the Spearman correlation



	x1	x2	x3
x1	1.000000	0.446299	0.394197
x2	0.446299	1.000000	0.511761
x3	0.394197	0.511761	1.000000

c) What is the VaR and ES at the 5% alpha level for each stock expressed in $?

After running a simulation for 1 million trials, the VaR and ES at the 5% alpha level (in $) for each stock is expressed in the VaR95 and ES95 columns respectively, with the percentage loss also represented in the other columns.



	Stock	VaR95	ES95	VaR95_Pct	ES95_Pct
0	x1	2.349574	2.931951	0.028287	0.035298
1	x2	1.449216	1.797698	0.018723	0.023225
2	x3	2.429033	3.030823	0.029566	0.036891
3	Total	5.008485	6.231784	0.020643	0.025685

d) What is the VaR and ES at the 5% alpha level for the total portfolio expressed in $?
See part c’s answer above, as the dataframe also has a row for the total portfolio results.