Computational Methods in Statistics (STAT 8670)
Spring 2020

MOHAMMED SHOEBUDDIN HABEEB
002521737

Stock market prediction for three major Semiconductor Manufacturing stocks using Hidden Markov Model (HMM)

''A growing economy consists of prices falling, not rising" - Kel Kelly

INTRODUCTION:

The stock market is consistent in only one aspect- it is always changing. Stock prices change every day as a result of market forces. There is a change in share price because of supply and demand. According to the supply and demand, the stock price either moves up or undergoes a fall. Stock markets normally reflect the business cycle of the economy: when the economy grows, the stock market typically reflects this economic growth in an upward trend in prices. In contrast, when the economy slows, stock prices tend to be more mixed. Markets may take time to form bottoms or make tops, sometimes of two years or more. This makes it difficult to determine when the market hits a top or a bottom A Hidden Markov Model is a tool that allows us to represent the probability distribution over a sequence of observations. The HMM assumed that an observation at time t was generated by a process whose state is hidden from an observer.

In this paper, the trend analysis of the stock market is found using Hidden Markov Model by considering the one day difference in close value for a particular period.This HMM is used to forecast the trend of the stock markets for three major semiconductor manufacturing companies, Lam Research. (LRCX), KLA Tencor Corp. (KLAC), and ASML Holding. (ASML). For a given observation sequence, the hidden sequence of states and their corresponding probability values are found. The probability values of π gives the trend percentage of the stock prices. Decision makers make decisions in case of uncertainty. The proposed approach gives a platform for decision makers to make decisions on the basis of the percentage of probability values obtained from the steady state probability distribution.

DATA:

The data consists of 3 .csv files (KLAC,ASML,LRCX) having information of Open ,High ,Low ,Close,Volume,Adjusted for a aprticular month between November 2014 to November 2019. With the HMM we are interested in modeling the difference between the close value and the opening value of the current day. There are two values for observations, "*I*" for price increase and "*D*" for price decrease respectively. Moreover, if the closing value of a month is greater than the closing value of previous month then we conisder the hidden state observation to be 'I' whereas otherwise 'D'.

An HMM, defined as λ=(P, E, π), consists of following elements →

S : *set of hidden states*
O : *set of hidden observations*
P : *state transition probability*
E : *observation emission probability*
π : *prior probability*

There are three states the majority of stock market analysts are interested in, “Low”, “Moderate” and “High”. These states dictate stock values and are often arduous for the investor to find and interpret. The three states are estimated by the observations related to hidden states.

DATA PREPARATION FOR PARAMETERS: For sake of calculation and ease of understanding I declare any current stock value will be “High” if it is greater than [0.5*mean(Data) + sd(Data)] and similarly any value will be “Low” if it is less than [0.5*mean(Data) – sd(Data)] and all the values in the middle will be “Medium”.

I will be manually calculating intial probability distribution, Transition Probability Matrix(TPM) and Emission Probability Matrix(EPM) for Hidden states{High,Medium,Low} and Observation states{“Increase” and “Decrease”}. For initial distribution I have taken all the available data from start of the company till November 2014 as the ‘Dataset’ and calculated initial distribution by seeing how many [High, Medium, Low] are present.

ASML	LRCX	KLAC
initial_distribution	initial_distribution	initial_distribution
[0.27004 0.48101 0.24894]	[0.2561308 0.4959128 0.2479564]	[0.2975610 0.4853659 0.2170732]

Now I want to consolidate the results for Hidden and Observed states in a dataframe ‘Data_Final’ with updated Dataset from November 2014 to November 2019 where Hidden are encoded as [“H”, “M”, “L”] and observed are encoded as [1,2] for [“I”, “D”].

Example :

	Hidden	Visible
1	M	2
2	M	2
3	M	1
4	L	2

Now for TPM I will calculate all the possible 9 combinations of hidden states and then accumulate to get probabilities for each position. Similarly for EPM I will calculate all the six possible combinations and aggregate them to calculate the matrix.

KLAC	LRCX	ASML
Transition Probability Matrix[TPM]:		
[,1] [,2] [,3]	[,1] [,2] [,3]	[,1] [,2] [,3]
[1,] 0.357 0.357 0.286	[1,] 0.20 0.333 0.4667	[1,] 0.20 0.467 0.33333
[2,] 0.133 0.600 0.267	[2,] 0.17 0.621 0.2069	[2,] 0.31 0.346 0.34620
[3,] 0.333 0.400 0.267	[3,] 0.467 0.333 0.200	[3,] 0.22 0.556 0.22220

Emission Probability Matrix[EPM]:

```
      [,1]   [,2]                 [,1]   [,2]                 [,1] [,2]
[1,] 1.0000 0.0000   [1,] 1.0000 0.0000   [1,]  1.0  0.0
[2,] 0.6098 0.3902   [2,] 0.6098 0.3902   [2,]  0.5  0.5
[3,] 0.0000 1.0000   [3,] 0.0000 1.0000   [3,]  0.0  1.0
```

With these parameters we have λ=(P, E, π) ready for each of the stock.

BAUM WELCH ALGORITHM:

Baum-Welch is an iterative procedure for estimating θ^*(best possible λ). It works by maximizing a proxy to the log-likelihood, and updating the current model to be closer to the optimal model. Each iteration of Baum-Welch is guaranteed to increase the log-likelihood of the data. But of course, convergence to the optimal solution is not guaranteed.
I have defined Baum-Welch Function which expects inputs [Observed sequence, TPM, EPM, π_0 and the iterations]. This recursively is calculated until convergence.

1. Compute $Q(\theta, \theta^s) = \sum_{z \in \mathcal{Z}} \log\left[P(\mathcal{X}, z; \theta)\right] P(z|\mathcal{X}; \theta^s)$.

2. Set $\theta^{s+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^s)$.

$$\pi_i^{(s+1)} = \frac{1}{D} \sum_{d=1}^{D} P(z_1^{(d)} = i | X^{(d)}; \theta^s)$$

$$A_{ij}^{(s+1)} = \frac{\sum_{d=1}^{D} \sum_{t=2}^{T} P(z_{t-1}^{(d)} = i, z_t^{(d)} = j | X^{(d)}; \theta^s)}{\sum_{d=1}^{D} \sum_{t=2}^{T} P(z_{t-1}^{(d)} = i | X^{(d)}; \theta^s)}$$

$$B_i^{(s+1)}(j) = \frac{\sum_{d=1}^{D} \sum_{t=1}^{T} P(z_t^{(d)} = i | X^{(d)}; \theta^s) I(x_t^{(d)} = j)}{\sum_{d=1}^{D} \sum_{t=1}^{T} P(z_t^{(d)} = i | X^{(d)}; \theta^s)}$$

Here θ= (π,A,B), π_i=P(z_1=i), A_{ij}=P(z_{t+1}=j|z_t= 1), andB_i(j) =P(x_t=j|z_t=i).

KLAC

```
$a
     [,1]      [,2]        [,3]
[1,] 6.08e-01 1.072e-01 2.83e-01
[2,] 6.17e-29 2.47e-24 1.00e+00
[3,] 7.61e-01 2.38e-01 1.13e-80

$b
       [,1]     [,2]
[1,] 1.000000 0.00000
[2,] 0.634492 0.365508
[3,] 0.000000 1.000000

$initial_distribution
[1] 0.2975610 0.4853659 0.2170732
0.2489451
```

LRCX

```
$a
     [,1]     [,2]     [,3]
[1,] 7.00e-01 0.300 1.84e-19
[2,] 7.18e-01 0.187 9.40e-02
[3,] 7.60e-73 0.999 6.58e-06

$b
          [,1] [,2]
[1,] 1.000000e+00    0
[2,] 7.388986e-12    1
[3,] 0.000000e+00    1

$initial_distribution
[1] 0.2561308 0.4959128 0.2479564
```

ASML

```
$a
     [,1]     [,2]     [,3]
[1,] 1.12e-89 3.97e-01 0.602
[2,] 8.98e-14 9.09e-01 0.091
[3,] 7.49e-01 1.19e-09 0.250

$b
       [,1]      [,2]
[1,] 1.0000000 0.0000000
[2,] 0.6124166 0.3875834
[3,] 0.0000000 1.0000000

$initial_distribution
[1] 0.27004 0.48101
```

These are the updated parameters from the Baum Welch Algorithm which will help Maximize P(O|λ). For convergence I have varied the iterations from 50 to 500 and I believe our values converge at around 500.

VITEBI ALGORITHM:

The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states—called the Viterbi path—that results in a sequence of observed events.Viterbi algorithm features similar steps to the forward algorithm: Initialization, Recursion and Termination, but also the Backtracking step to find the sequence of hidden states.

I have defined a Viterbi function which takes [Observations, TPM, EPM and initial distribution] which will give us a sequence of 60 most probable hidden sequence.

The accuracy for each of three stocks from the predefined sequence of Hidden state is :

	KLAC	LRCX	ASML
Initial parameters:	46.67%	33.33%	45%
Baum-Welch :	50%	58.33%	55%

Example Output from Viterbi:

```
[1] "L" "M" "H" "M" "H" "H" "M" "L" "M" "M" "H" "H" "H" "M" "H" "H" "M" "H" "H" "H" "H" "H"
"H"
[24] "H" "H" "H" "H" "H" "H" "H" "M" "H" "H" "H" "H" "M" "M" "H" "H" "H" "M" "H" "M" "H" "M"
"L"
[47] "M" "H" "M" "H" "H" "H" "H" "M" "H" "H" "H" "H" "H" "M"
```

FORWARD/BACKWARD ALGORITHM:

The Forward–Backward algorithm is the conventional, recursive, efficient way to evaluate a Hidden Markov Model, that is, to compute the probability of an observation sequence given the model. This probability can be used to classify observation sequences in recognition applications.

I have designed a function which takes in Observations,TPM,EPM and initial distribution and returns a Forwards Table. We can then calculate $P(O|\ \lambda)$ by adding up all the 3 elements pertaining to the 60^{th} value. We also calculate all the $P(O|\ \lambda)$ using Backward table by adding up all the 3 values in the first column of Backward table. We can then verify our to see that Baum-Welch maximize P(O| λ)

$P(O|\ \lambda)$	KLAC	LRCX	ASML
BAUM -WELCH PARAMETERS:			
FORWARD:	1.484745e-16	1.598551e-16	7.549831e-18
BACKWARD:	1.439399e-16	6.261543e-17	7.106226e-18
INITIAL PARAMETERS:			
FORWARD:	3.429274e-18	1.011832e-18	8.293648e-19
BACKWARD:	3.429274e-18	1.78416e-18	8.293648e-19
DIFFERENCE INCREASE(MAXIMIZED):			
(FORWARD)	+42%	+156%	+8.1%

STATISTICAL ANALYSIS:

The first Chart on the left shown below represents the LRCX's price chart from Nov 2014 through Nov 2019. Along with the volume traded. The second chart series on the right shows the **Bollinger Band chart**, % Bollinger change, Volume Traded and Moving Average Convergence Divergence. The moving average is important to understanding LAM Research(LRCX)'s technical charts. It smooths out daily price fluctuations by averaging stock prices and is effective in identifying potential trends.

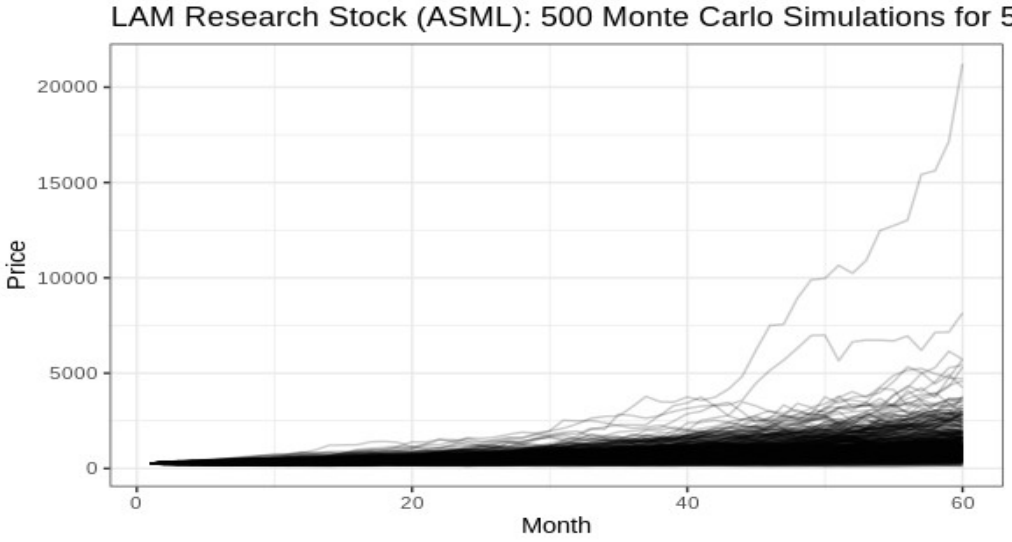
The Bollinger Band chart plots **two** standard deviations away from the moving average and is used to measure the stock's **volatility**. The Volume chart shows how its stocks are traded on the daily. The Moving Average Convergence Divergence gives technical analysts buy/sell signals. The rule of thumb is: **If it falls below the line, it is time to sell. If it rises above the line, it is experiencing an upward momentum.**

Apart from that to get a better idea of how our stocks will perform I am going to predict prices using Monte-Carlo for stable prediction. Monte Carlo methods basically refer to class of algorithms which use Randomness to give an estimate. These prices are simulated for accuracy. I have done 500 Monte Carlo simulations using Random Exponential Distribution.

LAM RESEARCH CORPORATION (LRCX):

Price Chart **Bollinger Band chart**

ONTE CARLO SIMULATION FOR 5 YEARS



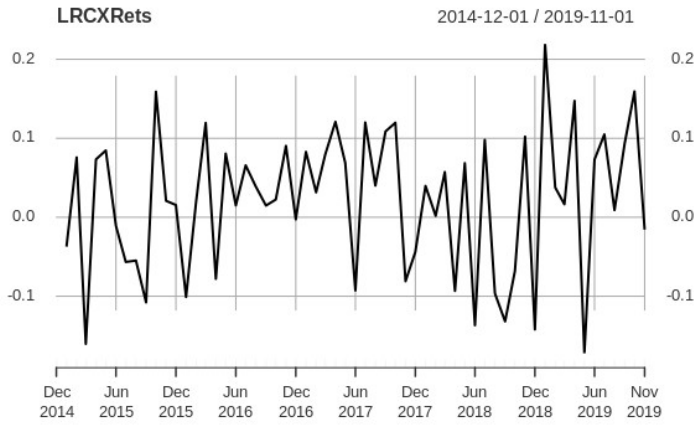
PRICE PREDICTION

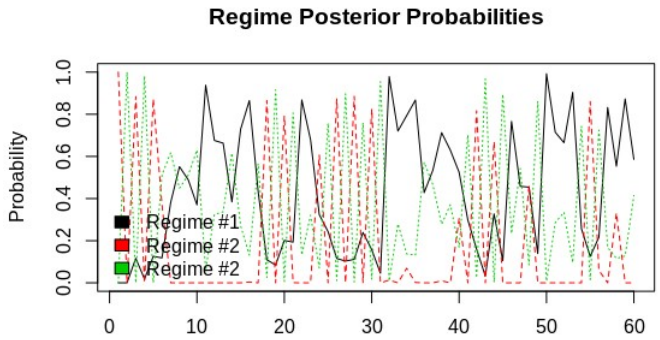
5%	50%	95%	99.5%
302.95	760.16	2002.500	3243.9956

From these simulations the prices can go as low as 302 and as high as 3243.9

VOLATILITY AND REGIME DETECTION:

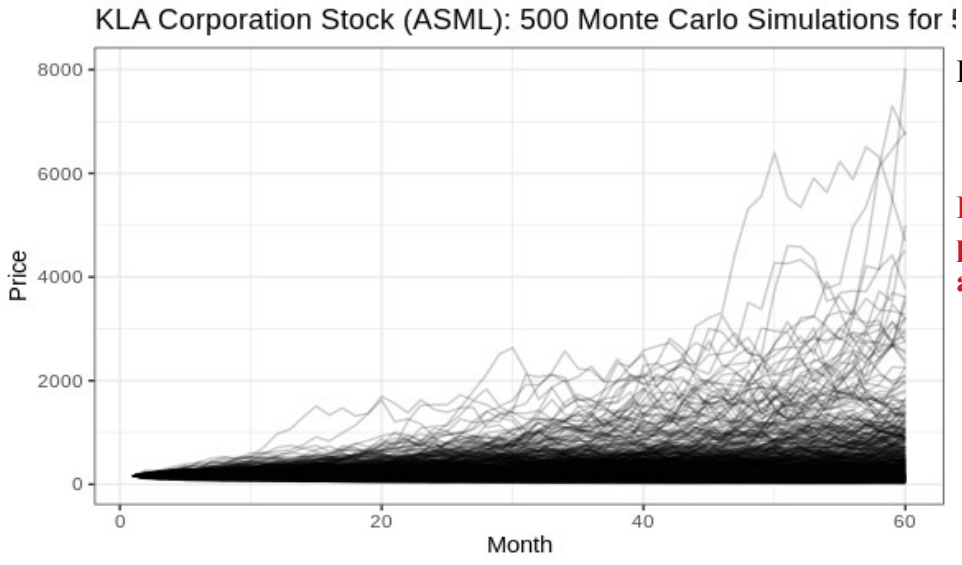
I will next plot the time series data for checking volatility of the stock and try to predict regime which could be Bullish or Bearish. The chart on the left shows the volatility the higher the inflection points the more volatile it was during that time.





KLA CORPORATION (KLAC):

MONTE CARLO SIMULATION FOR 5 YEARS

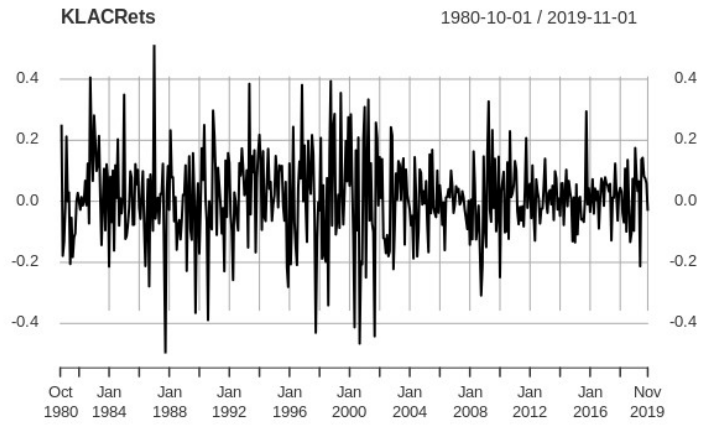


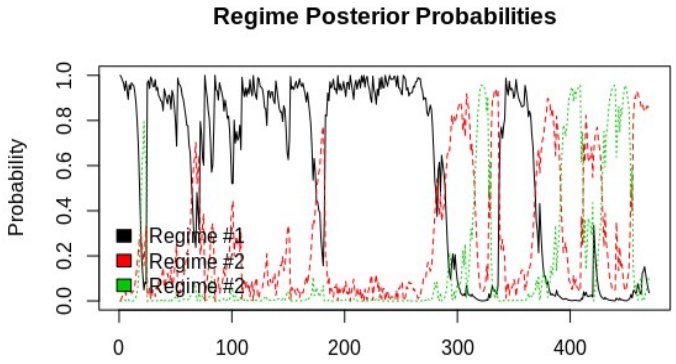
PRICE PREDICTION

5%	50%	95%	99.5%
50.22	299.37	1440.30	3891.731

From these simulations the prices can go as low as 50.22 and as high as 3891.7

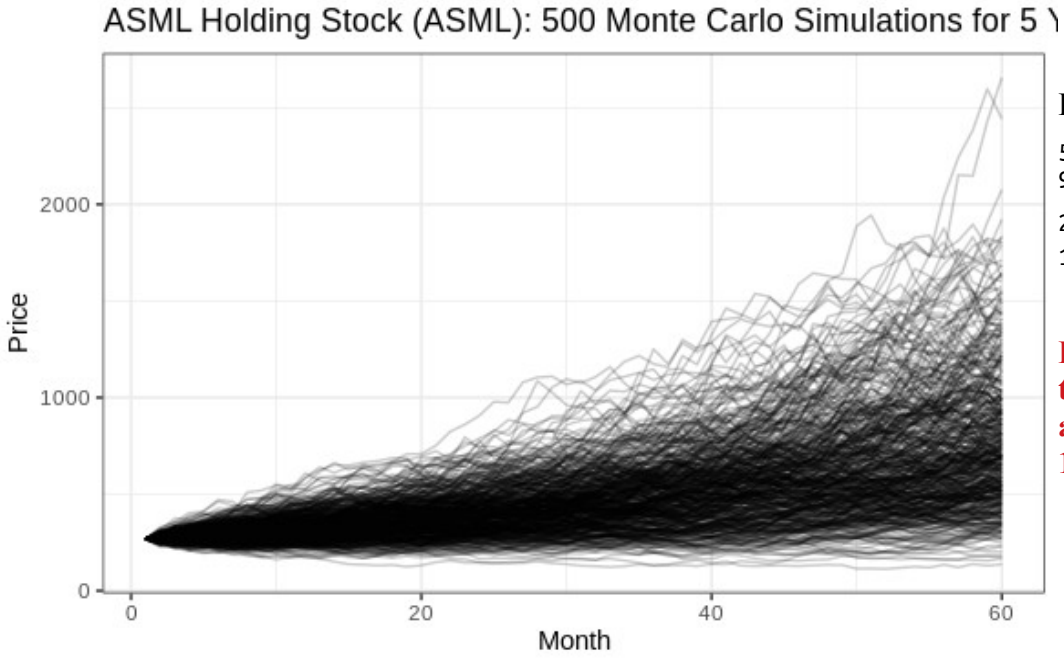
VOLATILITY AND REGIME DETECTION:





ASML HOLDING (ASML):

MONTE CARLO SIMULATION FOR 5 YEARS

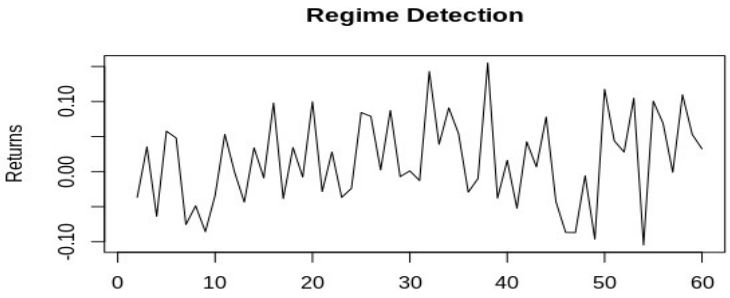


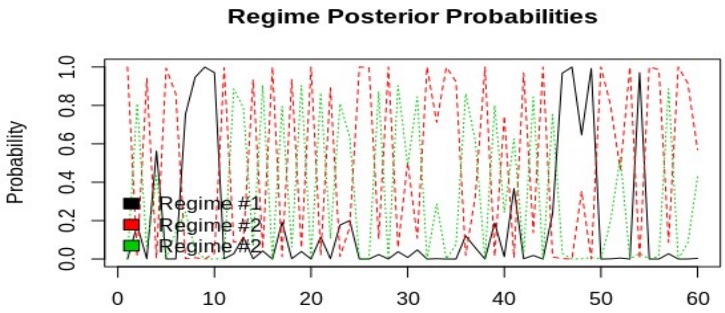
PRICE PREDICTION:

5%	50%	95%	99.5%
292.55	573.35	1149.98	1678.6919

From these simulations the prices can go as low as 292.5 and as high as 1678.7

VOLATILITY AND REGIME DETECTION:





From the above graphical inferences about each stock we get measures for comparison.

1) From Bollinger Band chart we see how many times have the expected values crossed Moving Average Convergence Divergence (Red Dotted) curve and what is the current trend.

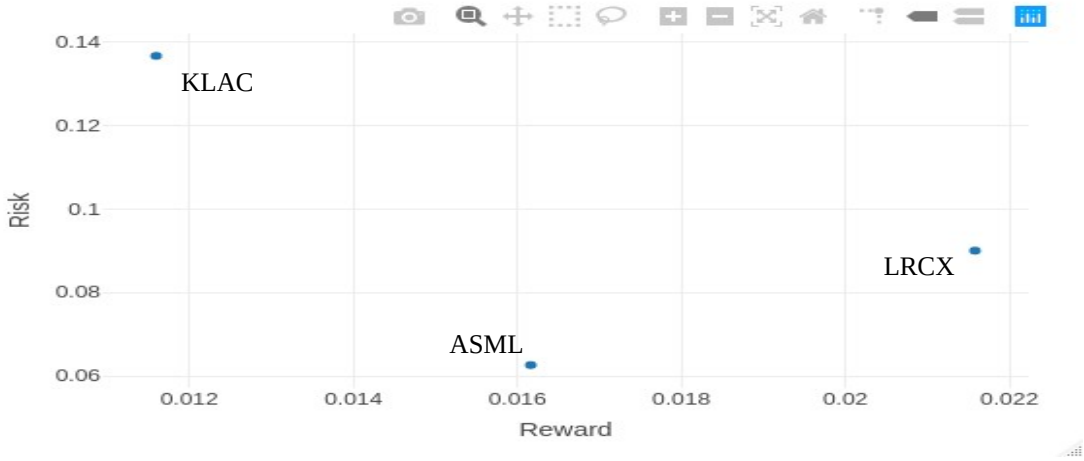
2)From Monte Carlo we see the limits of our investment

3) From Volatility and Regime detection we understand how volatile has the stock been historically and what are its regime reactions to depressions.

	Bollinger Band Chart	Monte Carlo Simulation	Regime and Volatility
LRCX	Rarely crosses MACD downawards and is currently on upward trend	Density of simulations is high around $400-$900	Volatility between [-0.2,0.2] and bullish tendencies in good markets
KLAC	Never crossed MACD and doesnt deviate much but currently on a high trend	Density of simulations is high around $150-$400	Volatility between [-0.45,0.45] and bearish tendencies to depressions
ASML	Historically high performing stock with a current high trend	Density of simulations is high around $400-$800	Volatility between [-0.1,0.15] and rare downward spiral

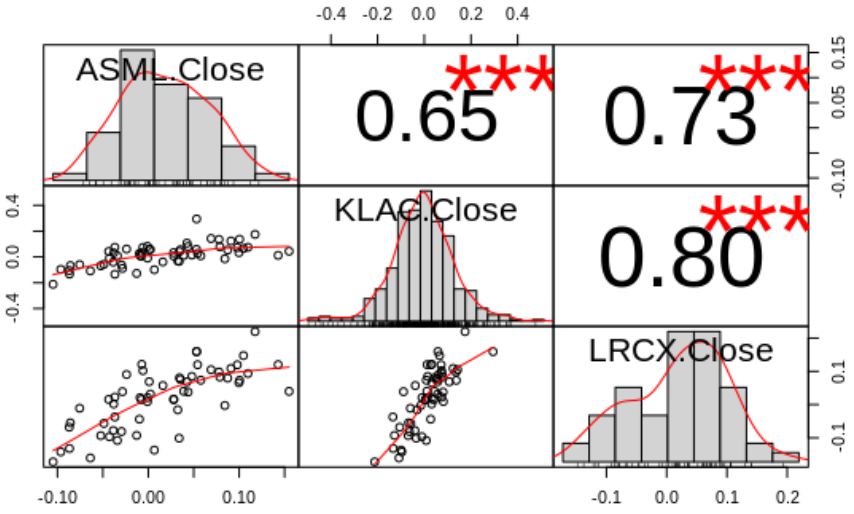
COMPARISON:

I compared the risk/return rate of each stock. I took the mean of log return and standard deviation of log return. The mean is assumed as the consistent rate of return while standard deviation is the risk that comes with purchasing the stock. I used plotly, an interactive visualization tool, to illustrate my findings.



KLA(KLAC) stock has the highest risk and the lowest return. ASML is the least risky with good returns but LRCX has the best returns provided the extra risk. If you are risk-inclined, Lam Research(LRCX) is a good investment as it has high risk and high returns. But if you are risk-adverse, like me, ASML Holdings(ASML) is the best choice.

A popular investing principle is to diversify your investments: do not put all your eggs in one basket. When purchasing stocks you should try to purchase stocks that share a small correlation because you want to maximize the total rate of return. From the below correlation plot ASML and LRCX have a lower correlation and better returns so I would recommend buying those 2 stocks.



CONCLUSION:

From the current stock data available I have found the best possible hidden states which give the maximum likelihood for each stock and then I would propose that Lam Research and ASML Holdings are good investment options for the future.