Hidden Markov Model for Stock Trading

Nguyet Nguyen

Replication by Grace Beyoko and Arina Agaronyan

Contents

1 Introduction

Stock price prediction is a question that has always sparked interest in the finance industry. Many try to predict stock prices for personal profit, but being able to predict future stock prices allows us to gain a deeper understanding of the economic cycles as well as predict potential crashes. Indeed, stock prices are influenced by a wide range of factors, such as geopolitical events, economic cycles, the general trust level in the market, etc. Therefore, by predicting stock prices accurately, we can gain insight into the future state of the economy, helping to identify risks and opportunities within different sectors and markets.

This study attempts to replicate the findings of the 'Hidden Markov Model for Stock Trading' paper by Nguyet Nguyen, published in 2018. The goal of the original paper is to predict the monthly Standard & Poor's 500 (S&P 500) closing prices using historical price data from 1950 to 2016. In the following sections, the source and type of the data will be presented, along with descriptive statistics. Then, the methods and models used to estimate the stock prices will be outlined, along with performance results. The results will be explained and interpreted. Subsequently, our replication study will be presented and discussed. Finally, our the implications and limitations will be considered.

2 Methodology

2.1 Data Collection

The data source of the original paper is the Yahoo Finance website. The author used monthly price data (closing price, open price, high price, low price) of the S&P 500 between January 1950 and November 2016. The summary of the data is presented below, with corrected period captions.

Price	Min	Max	Mean	Std.
Open	16.66	2173.15	504.06	582.20
High	17.09	2213.35	520.46	599.28
Low	16.66	2147.56	486.91	562.96
Close	17.05	2213.35	506.73	584.98

Table 1: Summary of S&P 500 index monthly prices from 1 January 1950 to 30 November 2016.

2.2 HMM Foundation

The author used the Hidden Markov Model (HMM), a statistical model frequently used in machine learning, to estimate the S&P 500 stock prices. The HMM model is based on the idea that the data follow an underlying process made of states, and the transition from one state to another depends solely on the current state. There are 3 main parameters of this model. First, the transition matrix, which gives the probability of transitioning from one state to another.

$$A = (a_{ij}), \quad a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i), \quad i, j = 1, 2, \ldots, N.$$

Next, the observation probability matrix, which presents the probability of observing a certain output given that we are in a specific state. Here, we assume that the probabilities follow a Gaussian distribution.

$$b_i(O_t) = N(O_t = v_k, \mu_i, \sigma_i)$$

Finally, we have the vector of initial probability which gives us the probability of being in a given state at time t=1.

$$p = (\pi), \quad \pi = P(q_1 = S_i), \quad i = 1, 2, \ldots, N$$

Thus, we are able to model the HMM parameters:

$$\lambda \equiv \{A, \mu, \sigma, p\}$$

where μ and σ are vectors of means and variances of the Gaussian distributions, respectively.

Prior to starting the prediction, the probability of observations $P(O \mid \lambda)$ must be calculated, "best fit" hidden states of observations must be found, and the parameters needed to be estimated and calibrated. This is done algorithmically with the in-sample data – here being the historical monthly S&P monthly prices from January 1950 to October 2006. Then, applying the Baum-Welch algorithm allows for identification of the parameters that locally maximise $P(O \mid \lambda)$.

2.3 Model Selection

For the calibration, the number of hidden states must be selected. Using a rolling window of 120 months (10 years), the parameters are calibrated for 5 different hidden states (from 2 to 6). Beginning the calibration from the period of December 1996 to November 2006, and moving the window up by a month each iteration 120 times, the information criteria is calculated for each model:

$$\begin{aligned}
\text{AIC} &= -2\ln(L) + 2k, \\
\text{BIC} &= -2\ln(L) + k\ln(M), \\
\text{HQC} &= -2\ln(L) + k\ln(\ln(M)), \\
\text{CAIC} &= -2\ln(L) + k(\ln(M) + 1),
\end{aligned}$$

where L is the likelihood function for the model (assuming that all the observations are independent), M is the number of observations, and k is the number of estimated parameters in the model – this is formulated as $k = N^2 + 2N - 1$, where N is numbers of states used in the HMM.

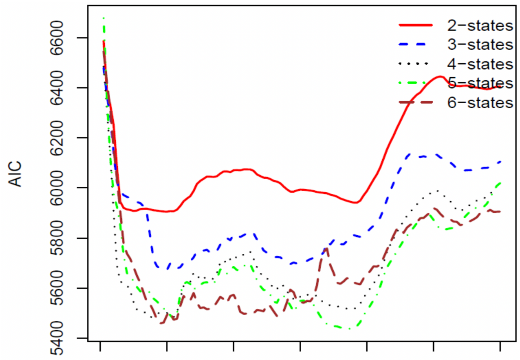
Initial parameters for the first calibration are chosen as such:

$$\begin{aligned}
&A = (a_{ij}), \quad a_{ij} = N_1, \\
&\rho = (1, 0, \ldots, 0), \\
&\mu_i = \mu(O) + Z, \quad Z \sim \mathcal{N}(0, 1), \\
&\sigma_i = \sigma^{(O)},
\end{aligned}$$

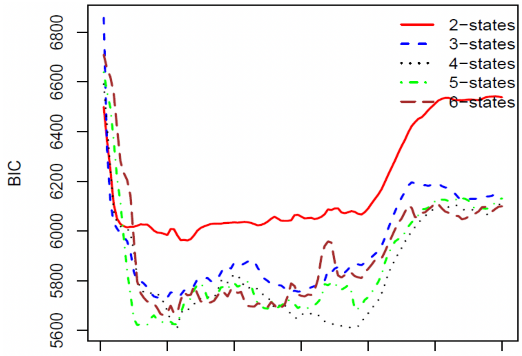
where $i, j = 1, ..., N$ and $\mathcal{N}(0, 1)$ is the standard normal distribution, and O represents the historical observed data of the open, low, high, and close prices for the fixed length $T = 120$ of 10 months. In this paper, rather than testing HMM with a single observation sequence, these multiple independent variables are considered instead.

$$O = \left\{O^{(1)}, O^{(2)}, O^{(3)}, O^{(4)},\, t = 1, 2, \ldots, T\right\}.$$

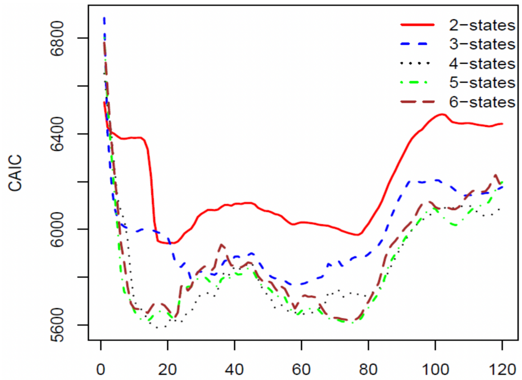
Visually guided by these plots of these 120 calibrations below, the author concluded that the best model – that is, the one with the consistently lower criterion value – was the Gaussian Hidden Markov Model with 4 states.



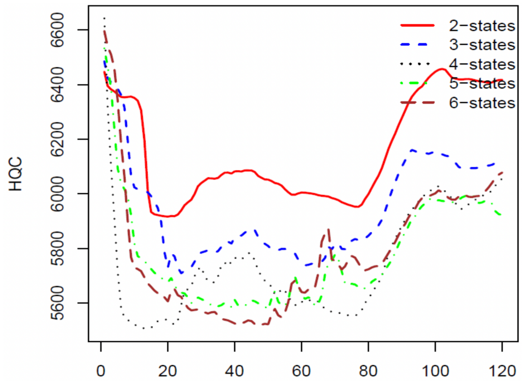
(a) AIC



(b) BIC



(c) HQC



(d) CAIC

Figure 1: Summary of model selection criteria: AIC, BIC, HQC, and CAIC.

3 Results

3.1 Prediction

Consequently, the prediction follows a clear process. First, the likelihood of a given window in the in-sample training set is calculated, then compared with previous periods month by month, until a data set with a similar likelihood is identified:

$$O^* = \left\{O^{(1)}, O^{(2)}, O^{(3)}, O^{(4)},\, t = T^* - D + 1, T^* - D, \ldots, T^*\right\}$$

such that $P(O^*|\lambda) \approx P(O|\lambda)$, where the training window $D = 120$

Following this, the difference between the previous and next months define the predicted prices as such:

$$O_T^{(4)+1} = O_T^{(4)} + \left(O_{T^*+1}^{(4)} - O_{T^*}^{(4)}\right) \times \operatorname{sign}\left(P(O \mid \lambda) - P(O^* \mid \lambda)\right).$$

With the parameters calibrated, the in-sample period is then trained on the historical patterns of the dataset. The model is thereafter applied on the out-of-sample data – the historical period of prices from November 2006 to November 2016.

The figure below contains the results of the HMM prediction against the real Close prices of that period, which is visibly similar in both magnitude and trend. Indeed, the model is able to capture the price changes around the 2008-2009 crash.

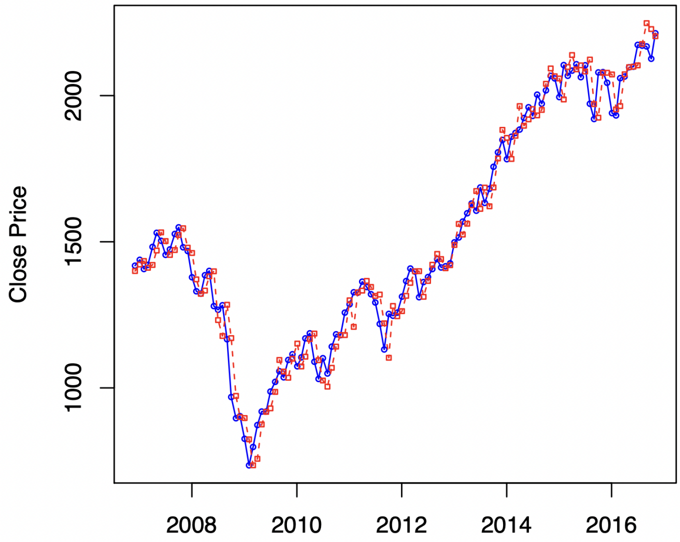


Figure 2: Predicted S&P 500 monthly prices from November 2006 to November 2016 using four-state HMM.

3.2 Performance

As stock prices can also be calculated based on their predicted returns and vice versa, in order to effectively evaluate the HMM predicted prices, the author compared the performance of HMM against historical average return model (HAR) for both predicted prices and predicted return.

The error estimators used are the absolute percentage error (APE), the average absolute error (AAE), the average relative percentage error (ARPE) and the root-mean-square error (RMSE). Included in the Table 2 is also a measure of efficiency (Eff.), which is calculated by subtracting the HMM/HAR ratio of the error estimator from 1; when positive, this indicates that the HMM surpasses the HAR.

Error Estimators		**APE**	**AAE**	**ARPE**	**RMSE**
Predicted Return	HMM	0.3325	0.0245	0.0303	2.4371
	HAR	0.1708	0.0249	0.0308	2.4752
	Eff.	**−0.9467**	**0.0161**	**0.0161**	**0.0154**
Predicted Price	HMM	0.0242	43.5080	54.8606	0.0240
	HAR	0.0246	44.2091	55.4853	0.02440
	Eff.	**0.0163**	**0.0159**	**0.0113**	**0.0164**

Table 2: Error and efficiency of S&P 500's predicted prices and predicted returns using the HMM versus HAR model

Based on these values, while the HAR has a lower APE for predicted stock returns, it can be concluded that the HMM outperforms the HAR in most cases.

3.3 Stock Trading

Another way to compare the 4-state HMM to the HAR model is through a stock trading simulation, wherein if the next predicted monthly stock return is positive, then the stock is bought, and sold if the prediction is of a negative return. Assuming that the trading cost is \$7.00 per trade action (this is based on the trading fee on the U.S. market at the time), and with each trade set to buy or sell 100 shares of the S&P over different trading periods set out in Table 3:

Investment = share price $*$ 100,
Earnings = Investment + (remaining shares $*$ price),
Cost = total buy and total sell trades $*$ \$7.00,
Profit = % return after trading costs.

The difference between the models here is that the HMM model will do nothing if the next moth return prediction is increasing, and the Buy & Hold model simply simulates and investor buying 100 shares at the beginning and holding them until the end of the trading period.

Simply by looking at the simulated profits, it is clear that the 4-state HMM model consistently exceeds the HAR and Buy & Hold model profits, except for the single instance where the Buy & Hold model yields a slightly higher profit in the 100 month trading period. The occasional identical HAR and Buy & Hold profit values arise due to the HAR predicted returns being positive.

This simulation highlights a limitation of the HAR model wherein, due to its predictions depending entirely on the mean of the historical data, it is not sensitive to point changes in the price. Thus, an advantage of the HMM is that it captures this point change credibly, making it quite suitable for stock price forecasting compared to the HAR.

Trading Period	Model	Investment (\$)	Earning (\$)	Cost (\$)	Profit (%)
40 months	HMM	168,155	65,172	168	38.66
(7/2013	HAR	168,573	52,762	14	31.29
–11/2016)	Buy & Hold	168,573	52,762	14	31.29
60 months	HMM	124,696	95,205	378	76.05
(12/2011	HAR	131,241	90,094	14	68.64
–11/2016)	Buy & Hold	124,696	96,639	14	77.49
80 months	HMM	116,943	113,971	392	97.12
(4/2010	HAR	116,943	104,392	14	89.26
–11/2016)	Buy & Hold	116,943	104,392	14	89.26
100 months	HMM	126,738	94282	616	73.91
(8/2008	MAR	126,738	73,010	84	57.54
–11/2016)	Buy & Hold	126,738	94,597	14	74.63
120 months	HMM	141,830	100,614	770	70.39
(11/2006	HAR	140,063	81,272	14	58.15
–11/2016)	Buy & Hold	140,063	81,272	14	58.15

Table 3: S&P 500 trading using the HMM, HAR, and Buy & Hold models for different time periods

4 Replication Study

4.1 Data Collection

The data source for the present study is the yfinance Python library. We used the same monthly price data as the author (closing price, open price, high price, low price) of the S&P 500 between January 1950 and November 2024. Unfortunately, we have some irregularities in our data. For instance, for the full period before December 29th, 1961, the same value was reported for Open, Low, High, and Close prices. Then, between the 29th December 1961 and 20th April 1982, the Open prices were set to 0. We chose to encode these 0 instances as missing values to avoid biasing our results. These irregularities are explained by the fact that these data were obtained through estimation and backtracking of historical prices.

It is also important to note that we were unable to obtain the CSV files used by the author. Yahoo Finance no longer provides the option to download CSV files from its free version. While the yfinance Python library allows full access to daily data, the monthly data are only available from 1985 onward. Consequently, we calculated and aggregated the monthly data from the daily data. Due to these combined inconsistencies in data collection, our dataset differs from the one used by the author. As such, we chose to conduct our replication study over a more recent period, thus our data spans from 1st January 1950 to 30th November 2024.

Our full clean dataset consists of 3,596 data points. We have 899 rows – each representing the months between January 1950 and November 2024 – and 4 columns: open, high, low and close prices. The average closing price is $1,106 while the average opening price is $1,098. We also observe a maximum high price of $6,044 reached in November 2024 and min low price of $17, in January 1950. We also observe a clear overall upward trend in prices, i.e all the prices have increased significantly over time.

Price	**Min**	**Max**	**Mean**	**Std.**
Open	16.66	5757.73	1098.44	1235.74
High	17.09	6044.17	1136.18	1279.12
Low	16.66	5696.51	1061.21	1195.06
Close	17.05	6032.38	1106.73	1248.43

Table 4: Summary of S&P 500 monthly prices from January 1950 to November 2024.

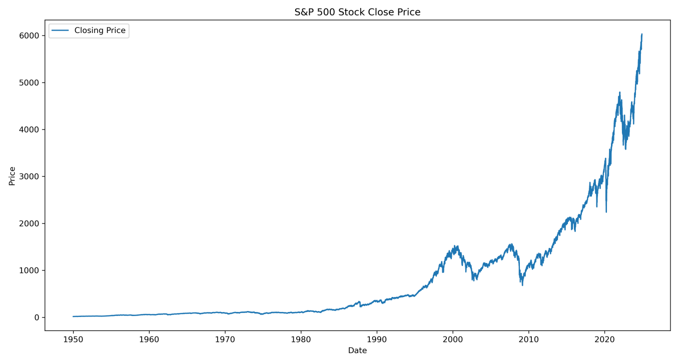
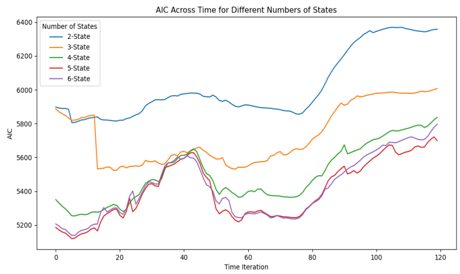


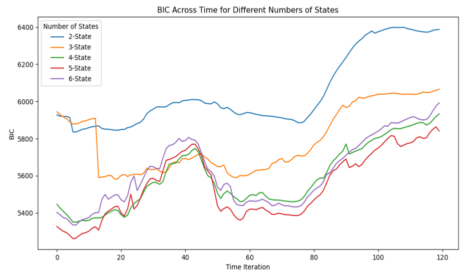
Figure 3: S&P 500 Stock Close Prices Trend from 1950 to 2024.

4.2 Model Selection

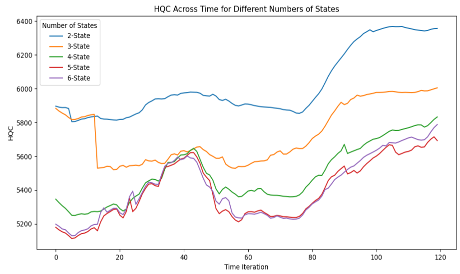
Similarly to the method used by Dr. Nguyen, we plotted the AIC, BIC, HQC and CAIC of 2 to 6 states for the period of December 1996 to November 2016 in order to determine the optimal number of hidden states that minimizes the Information Criteria. The visual analysis tends toward 5 hidden states.



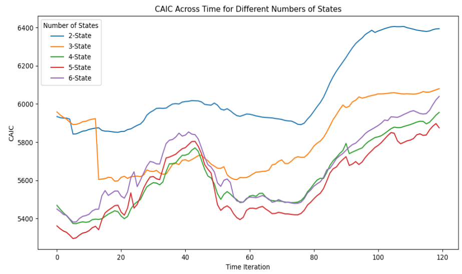
(a) AIC



(b) BIC



(c) HQC



(d) CAIC

Figure 4: Summary of model selection criteria: AIC, BIC, HQC, and CAIC.

In contrast, we decided to refrain from solely relying on the visual analysis, and so computed this using an algorithmic approach. By doing so, we found out that a model with 6 hidden states minimizes the Information Criteria. In fact, we conducted this study with both 5 and 6 hidden states to confirm this, and the model with 6 hidden states was more efficient.

4.3 Prediction

Figure 5 presents the results of the HMM price predictions against the actual observations for the ten year period between November 2014 and 2024. We are able to see that the match is very close. It is evident that the model is able to capture the overall trend of the price movements effectively, demonstrating the model's ability to follow both upward and downward trajectories.

However, discrepancies are observed in the magnitude, especially in more recent periods of following the 2020 crisis. These deviations suggest that while the model performs well in stable conditions, it struggles to fully account for extreme fluctuations, which may require further adjustments.

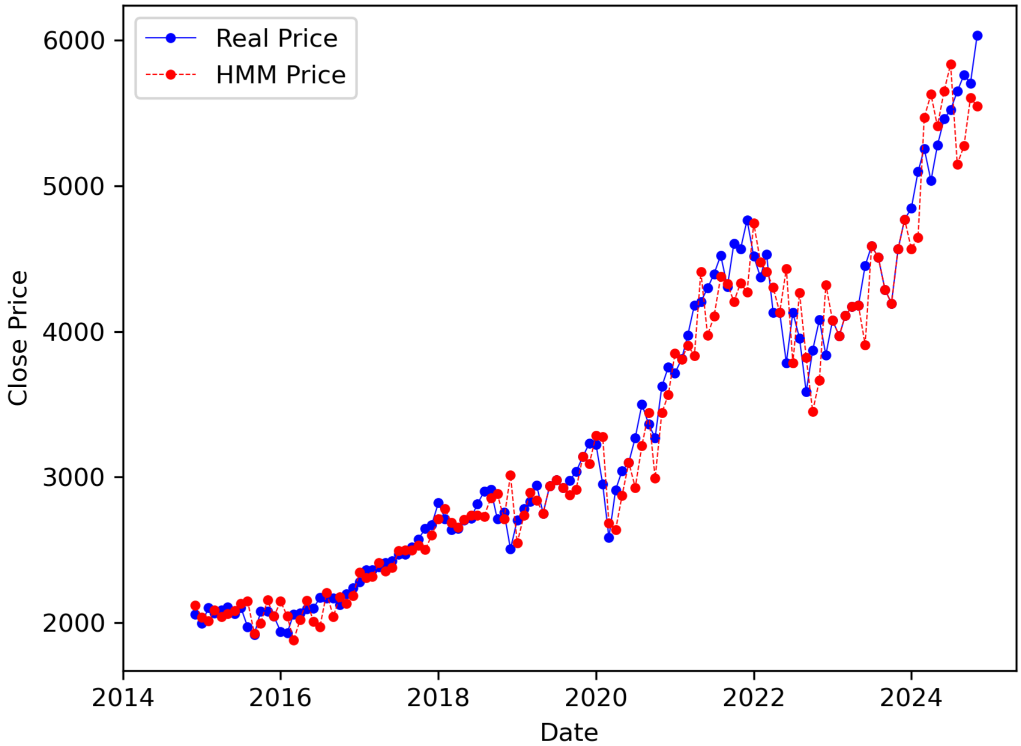


Figure 5: HMM Predicted Close Prices

4.4 Performance

Table 5 contains a summary of our performance validation measures. In comparison to Dr. Nguyen's error estimates, we replaced the APE with the MAPE for ease as well as included a separate R^2 score.

Here the Efficiency is significantly negative for all error estimators, and all error estimates are significantly higher for the HMM, clearly showing that for our prediction, the HAR surpasses the HMM model. Although, for both models, the AAE, ARPE, and especially the RMSE error estimates differ fundamentally to those in the original paper.

Model	**MAPE**	**AAE**	**ARPE**	**RMSE**	R^2
HMM	0.04	143.93	1.20	211.55	0.98
HAR	0.03	114.80	0.96	155.17	0.96
Efficiency	**-0.22**	**-0.25**	**-0.25**	**-0.36**	

Table 5: Error estimators of S&P 500's predicted prices using HMM versus HAR.

Additionally, in Figure 6 below, it is evident that the HAR predictions are more accurate to the real Close prices, with identical trend and magnitude, though with a slight lag.

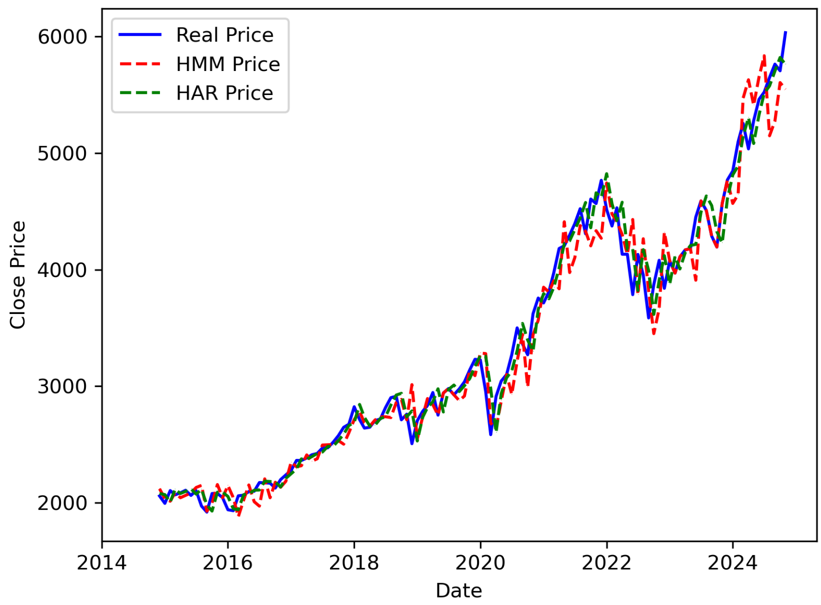


Figure 6: HMM vs HAR Predicted Close Prices.

4.5 Stock Trading

As stated previously, the HAR model is clearly superior for our replication. While the stock trading simulation displayed in Table 6 disproves this stance with HMM and Buy & Hold consistently yielding significantly higher profits compared to the HAR, the HMM model profits seem relatively unrealistic. With the cost consistently at \$7.00, this implies the HMM model investor is only trading once, and thus mimicking the Buy & Hold strategy, which is unexpected, given HMM's adaptive behavior. These inconsistencies may stem from potential inaccuracies in the replication code of the simulation itself, as well as an additional element of a changed US market trading fee from the assumed fixed cost of \$7.00 in 2006-2016.

Trading Period	Model	Investment (\$)	Earnings (\$)	Cost (\$)	Profit (%)
40 months	HMM	439,525	163,712	7	37.25
(7/2021	HAR	439,525	164,804	91	29.16
-11/2024)	Buy & Hold	439,525	163712	7	37.25
60 months	HMM	323,078	280,159	7	86.71
(12/2019	HAR	323,078	278,536	77	54.64
-11/2024)	Buy & Hold	323,078	280159	7	86.71
80 months	HMM	264,805	338,432	7	127.80
(4/2018	HAR	264,805	261,028	119	51.20
-11/2024)	Buy & Hold	264,805	338,432	7	127.80
100 months	HMM	197,218	406,019	7	205.87
(8/2015	HAR	197,218	330,433	119	64.82
-11/2024)	Buy & Hold	197,218	406,019	7	205.87
120 months	HMM	206,756	397,347	7	192.99
(11/2014	HAR	206,756	295,415	147	52.27
-11/2024)	Buy & Hold	206,756	396,481	7	191.76

Table 6: S&P 500 trading using the HMM, HAR, and Buy & Hold models for updated different time periods.

5 Extension

5.1 Limitations

After building our model, several limitations stand out. First, after analyzing our results, we observe that the predictions closely match the actual values up until 2020. After this point, the volatility of the predicted price increases, with more abrupt fluctuations. This can be explained by the fact that our model does not account for macroeconomic factors that significantly impact month-to-month variations. The prediction process is based entirely on past month-to-month differences from a time period with similar likelihood. This mechanism makes it difficult for the model to anticipate new patterns. For example, the effects of events like the COVID-19 pandemic may not be accurately captured, as these events are unprecedented in the historical data. In that light, incorporating an ARMA model with additional explanatory variables could potentially enhance the understanding of price movements and improve the forecasting accuracy.

Moreover, upon examining the data, we observed that the time series may not meet some of the key assumptions necessary for reliable forecasting. Specifically, it seems as if the time series is non-stationary and there is autocorrelation in the residuals. If these issues are present, they could compromise the accuracy of the predictions. A more thorough adapted time series analysis is needed to address these concerns.

5.2 Potential Improvements

As mentioned above, we felt the need to conduct a time series analysis to gain a deeper understanding. Through this analysis, we discovered that the time series was non-stationary and exhibited serial dependence in its residuals. This violates key assumptions of many statistical and machine learning models (including HMM models). Non-stationarity can result in unpredictable future patterns and model instability, while autocorrelation in residuals can lead to overfitting. Both issues can significantly increase forecast errors.

To address these issues, we transformed the time series into a stationary one by calculating the log returns. Furthermore, we addressed the serial dependence in the residuals by fitting a GARCH model, which allows us to model the error term as a function of past squared residuals and past variances, effectively capturing the volatility dynamics.

From this foundation, we attempted to build an HMM-GARCH model to predict S&P 500 stock returns. However, when fitting the GARCH(1,1) model for each hidden state, we observed that the GARCH component provided limited significance. The summary tables – see Figure 7 – revealed that, for most hidden states, the volatility was primarily driven by past volatility rather than new shocks. In other words, the GARCH element did not significantly enhance the model's performance, as the volatility appeared to be due to persistence rather than new information or unexpected events.

```
Regime 1 GARCH Summary:
                     Constant Mean - GARCH Model Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.000
Mean Model:             Constant Mean   Adj. R-squared:                  0.000
Vol Model:                      GARCH   Log-Likelihood:               -195.218
Distribution:                  Normal   AIC:                           398.436
Method:            Maximum Likelihood   BIC:                           407.863
                                        No. Observations:                   78
Date:                Wed, Jan 01 2025   Df Residuals:                       77
Time:                        22:27:30   Df Model:                            1
                               Mean Model
============================================================================
                 coef    std err          t      P>|t|  95.0% Conf. Int.
----------------------------------------------------------------------------
mu            -4.8324      0.339    -14.247  4.655e-46 [ -5.497, -4.168]
                             Volatility Model
=============================================================================
                 coef    std err          t      P>|t|       95.0% Conf. Int.
-----------------------------------------------------------------------------
omega          0.8846      1.021      0.867      0.386      [ -1.116,  2.885]
alpha[1]   2.3139e-12  2.358e-02  9.814e-11      1.000 [-4.621e-02,4.621e-02]
beta[1]        0.9026      0.135      6.685  2.311e-11      [  0.638,  1.167]
=============================================================================
```

Figure 7: GARCH(1,1) Summary Table for the 2nd Hidden State.

6 Conclusion

To conclude, in this study, we replicated and extended the analysis of the 2018 paper by Nguyet Nguyen, "Hidden Markov Model for Stock Trading". The Hidden Markov Models demonstrated strong reliability and the ability to capture the overall trends of stock price movements effectively. However, our extension also revealed some limitations. Specifically, the predictive performance declined significantly when the time period under consideration was extended up to November 2024. This period was marked by significant market volatility driven by various factors, highlighting the HMM's limitations, particularly during periods of heightened economic uncertainty.

Additionally, we explored several potential improvements, but none yielded significant enhancements in model performance. Future research could investigate more robust and advanced methods to address these challenges and improve forecasting accuracy in volatile conditions.

Overall, this study reinforces the potential of HMMs for financial modeling while acknowledging the complexities of stock price forecasting.