An introduction to the use of hidden Markov models for stock return analysis

Chun Yu Hong[*], Yannik Pitcan[†]

December 4, 2015

Abstract

We construct two HMMs to model the stock returns for every 10-day period. Our first model uses the Baum-Welch algorithm for inference about volatility, which regards volatility as hidden states and uses a mean zero Gaussian distribution as the emission probability for the stock returns. Our second model uses a spectral algorithm to perform stock returns forecasting. We analyze the tradeoffs of these two implementations as well.

1 Introduction

Hidden Markov models (HMMs) are known for their applications to speech processing and pattern recognition. They are attractive models for discrete time series analysis because of their simple structures. It is therefore not surprising that there has been research on the applications of HMMs to finance.

Hassan and Nath (2005) use HMM to forecast the price of airline stocks. The goal is to predict the closing price on the next day based on the opening price, the closing price, the highest price and the lowest price today. The performance of the HMM is similar to that of artificial neural networks (ANN).

O et al. (2004) propose a three-level hierarchical HMM to model the dynamics of the stock prices. The first level consists of the hidden states that describe the trend of the stocks: strong bear, weak bear, random walk, weak bull and strong bull. The second level consists of the hidden states responsible for the components of a Gaussian mixture. The third level consists of the outputs: the relative closing prices, defined as the percent change in closing price relative to the previous closing price.

Since many of these HMM models for stock returns focus on forecasting, we decide to introduce a very simple HMM for performing inference about volatility changes. The idea of using HMM for volaility analysis is not new: there are a few existing papers on HMM-GARCH (generalized autoregressive conditional heteroskedasticity) models for volatility forecasts (for example, Zhuang and Chan, 2004; Rossi and Gallo, 2005). However, these models are often too complex to be interpreted properly. Our model allows simple and natural interpretations yet provides important insight into the heteroskedastic nature of

[*]Department of Statistics, UC Berkeley: jcyhong@berkeley.edu
[†]Department of Statistics, UC Berkeley: pitcany@berkeley.edu

stock returns. It provides a solid foundation for further study of the volatility changes.

Spectral learning methods have recently gained excitement in the machine learning community. Hsu et al. (2012) introduces a spectral algorithm for learning HMMs. The usual EM approach for learning HMM tends to suffer from local optima issues and is often considered as computationally hard. The algorithm proposed by Hsu et al. can be proven to be efficient and correct. In our paper, we use the spectral algorithm to perform stock returns forecasting.

2 Volatility Analysis

Let $\{P_t\}_{t=1}^T$ is the price of a certain stock from $t = 1$ to $t = T$. The geometric brownian motion (GBM) model for stock prices suggests that

$$dP_t = \mu P_t dt + \sigma P_t dW_t,$$

where $\{W_t\}$ is a standard Brownian motion, and μ and σ are unknown constants governing the drift and the volatility of the prices, respectively. Rearranging the term and setting $dt = 1$, $dP_t = P_{t+1} - P_t$, and $dW_t = W_{t+1} - W_t$, we obtain the following (approximate) relationship:

$$\frac{P_{t+1} - P_t}{P_t} = \mu + \sigma(W_{t+1} - W_t).$$

If this relationship is indeed true, the price returns $\frac{P_{t+1}-P_t}{P_t}$ should follow a white noise process[1] plus a certain constant. Based on the histograms and ACF plot shown in Figures 3 and 4, it is tempting to conclude that white noise processes are appropriate models. However, we suspect that the volatility of stock returns is non-constant. This motivates our model for volatility analysis.

2.1 Model Specification

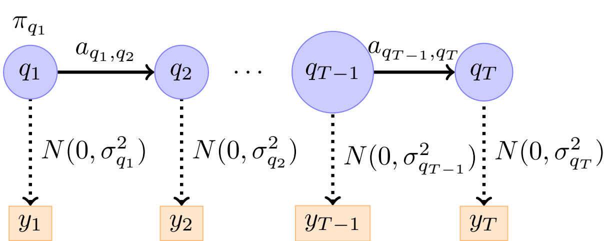

Figure 1: An HMM for centered returns. Each of the hidden states takes values in $\{1, ..., M\}$. Each of the emission probabilities is based on a mean zero Normal distribution with variance depending on the corresponding hidden state q_t.

Our objective is to study the price return process $\{r_t\}_{t=2}^T$, where r_t is the rate of returns:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}.$$

[1] A time series $\{\epsilon_t\}_{t=1}^T$ is a white noise process if $\epsilon_1, ..., \epsilon_T$ are i.i.d. $N(0, \sigma^2)$.

For simplicity, each of the rate of returns is computed over a 10-day period (each unit time is equal to 10 days: $t-(t-1)=10$ trading days).

We propose the following model for r_t:

$$r_t = \mu + y_t, \qquad t = 1, ...T,$$

where $\mu \in \mathbb{R}$ is an unknown parameter and $\{y_t\}_{t=1}^T$ follows a hidden Markov process specified by Figure 1:

1. number of states: $M \geq 2$
2. number of observations: T
3. hidden states: $\{q_t\}_{t=1}^T$
4. observations: $\{y_t\}_{t=1}^T$, where $y_t|q_t \sim N(0, \sigma_{q_t}^2)$
5. probability distribution of q_1: $\pi = \{\pi_1, ..., \pi_M\}$, where $\sum_{i=1}^M \pi_i = 1$, $\pi_i \geq 0$ for all $i = 1, ..., M$, and $\pi_i = \mathbb{P}(q_1 = i)$.
6. transition matrix: $A \in [0,1]^{M\times M}$ with $A = (a_{ij})$ and $a_{ij} = \mathbb{P}(q_{t+1} = j|q_t = i)$ (we assume that $\{q_t\}_{t=1}^T$ is a homogeneous Markov chain)

.

The q_t's represents the "volatility stages" in which the stock is undergoing. To see why, note that q_t has a direct relationship with the variance of y_t, which serves as a natural proxy for the volatility of the stock returns. There are three interesting quantities in this model:

1. $\sigma_1^2, ..., \sigma_M^2$: Assume that the σ_i^2's reflect the volatility of a particular stock. Are these values small or big? Are they of similar magnitudes? How are they vary for different stocks?
2. M: How many volatility stages does a particular stock possess? Small-cap stocks tend to be more volatile than large-cap stocks, so do small-cap stocks have more volatility stages, or is this volatility behavior already captured by the magnitudes of the σ_i's?
3. A: The ith diagonal element of A, a_{ii}, measures the "stickiness" of the ith volatility stage. If a_{ii} is close to 1, it means that the stock is likely to be "stuck" in the ith volatility stage once it enters the ith stage.

To estimate μ, a natural estimator is the method of moments estimator, the arithmetic average:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^T r_t =: \bar{r}.$$

To estimate the stochastic component y_t, we use the centered return, $\tilde{y}_t := r_t - \bar{r}$.

We estimate all the model parameter using the Baum-Welch algorithm. To find the most likely sequence of the hidden states, we use the Viterbi algorithm based on the estimated model parameters obtained from Baum-Welch.

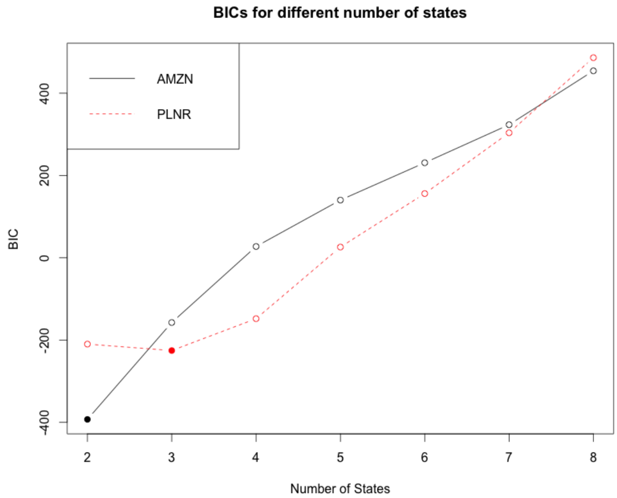

Figure 2: A plot of the BICs versus different numbers of states for AMZN and PLNR. Each of the filled points indicates the lowest BIC for the corresponding stock.

2.2 Results

We first study Amazon (AMZN), a stock which is considered as a "mega-cap" stock (a stock with market capitalization over \$200 billion). We study its adjusted close prices from January 3, 2007 to November 24, 2015, and choose the number of states M based on BICs (Bayesian Information Criterion), which is defined as

$$BIC = -2 \log \text{likelihood} + p \log n,$$

where p is the number of estimated parameters in the model and n is the number of observations. For the proposed HMM, $n = T$ and the number of estimated parameters is computed as follows:

$$\begin{aligned} p = &\# \text{ of parameters estimated for } \pi + \# \text{ of parameters estimated for } A \\ &+ \# \text{ of parameters estimated for } \sigma_1^2, ..., \sigma_M^2 \\ &= (M-1) + M(M-1) + M \\ &= M^2 + M - 1. \end{aligned}$$

We would like to choose a model with the smallest BICs. Since the HMM model with $M = 2$ has the smallest BICs, we continue our analysis with $M = 2$ (Figure 2).

As shown in Figure 3, the centered returns for AMZN resembles a bell-shaped distribution, consistent with the choice of mean zero Gaussian emission probability. The ACF plot shows that the autocorrelations are very weak, which is consistent with the HMM having no arrows among the observations y_t themselves. The bottom right panel shows that plot of the volatility stages (VSs) based on the most likely sequence of hidden states. VS 1 corresponds to $\sigma_1^2 = 0.003758$ ($\sigma_1 = 0.061$) and VS 2 corresponds to $\sigma_2^2 = 0.027972$ ($\sigma_2 = 0.167$). It seems

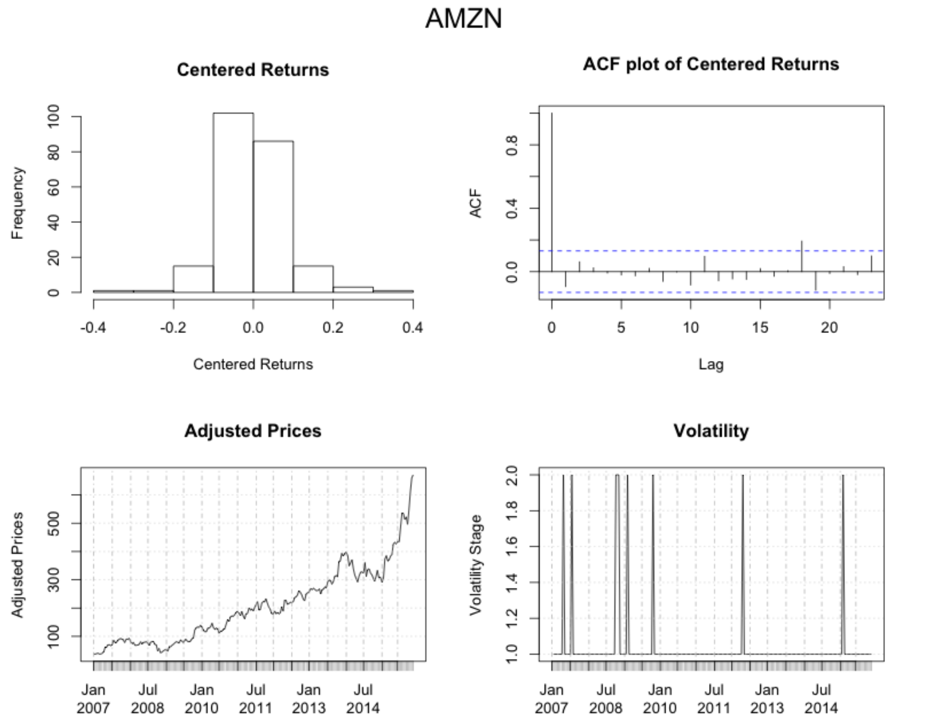

Figure 3: A 2×2 panel plot of the stock AMZN. Top left: A histogram of the centered returns for AMZN. Top right: An ACF (autocorrelation function) plot of the centered returns. Bottom left: A line plot of the adjusted price time series. Bottom right: A plot of the volatility stages.

that VS2 (high volatility) is ephemeral relative to VS1 (low volatility), as the volatility plot stays flat in VS 1 most of the time, and it has occasional abrupt jump to VS 2 and immediate return to VS 1. We can also deduce this behavior based on the estimated transition matrix $\hat{A}$:

$$\hat{A} = \begin{bmatrix} 0.917 & 0.083 \\ 0.654 & 0.346 \end{bmatrix}.$$

Note that the estimated probability of going from VS 2 to VS 1 $\hat{a}_{21}$ is greater than 60%, indicating there is a high probability of returning to VS 1 if the stock AMZN is currently in VS 2, based on the HMM model.

To make comparison with AMZN, we fit our HMM model to Planar Systems, Inc. (PLNR), a small-cap stock with market capitalization around \$150 million, over the same time period. Based on the BICs, we find that the optimal number of states is 3 (Figure 2). The three volatility stages corresponds to $\sigma_1^2 = 0.005697$ ($\sigma_1 = 0.075$), $\sigma_2^2 = 0.046706$ ($\sigma_2 = 0.216$), and $\sigma_3^2 = 0.271467$ ($\sigma_3 = 0.521$). The estimated transition matrix $\hat{A}$ for PLNR is

$$\hat{A} = \begin{bmatrix} 0.939 & 1.36 \times 10^{-2} & 4.76 \times 10^{-2} \\ 1.60 \times 10^{-31} & 0.946 & 5.36 \times 10^{-2} \\ 1.00 & 5.78 \times 10^{-10} & 2.00 \times 10^{-29} \end{bmatrix}.$$

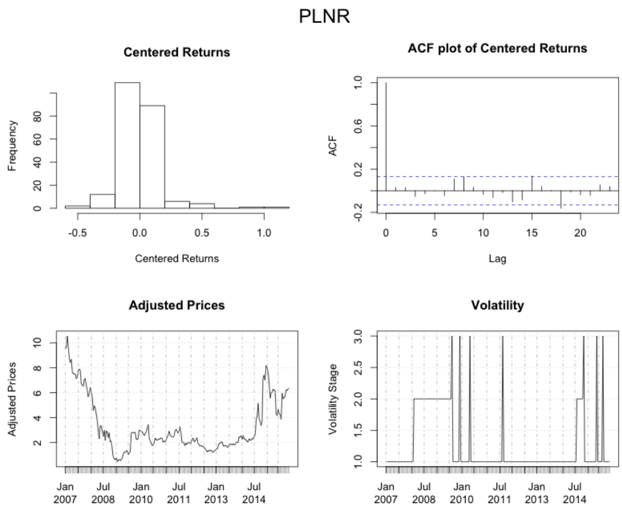

Figure 4: A 2×2 panel plot of the stock PLNR. Note that the histogram of the centered returns is approximated bell-shaped and the ACF plot shows that almost all of the autocorrelations are insignificant.

The entry $\hat{a}_{31} = 1$ indicates that we are almost sure that the stock will return to VS1 (low volatility) after it enters VS3 (very high volatility). Both $\hat{a}_{11}$ and $\hat{a}_{22}$ are very close to 1, which indicates that the lower volatility regimes are "sticky".

With the specific examples of AMZN and PLNR in mind, we can turn to a more general analysis of large-cap to mega-cap stocks and small-cap stocks. Table 1 shows the 10 large-cap to mega-cap stocks and the 10 small-cap stocks selected for the analysis.

Figure 5 shows that σ_i's versus the volatility stages for the 20 stocks. The numbers of VSs are chosen such that the BICs are minimized. We find that most of the selected stocks, regardless of whether the stocks are small-cap or large-cap, have only 2 volatility stages (high and low) in our model. To our surprise, quite a few small-cap stocks and a few large-cap stocks possess similar $\hat{\sigma}_1$ and $\hat{\sigma}_2$, as shown in the substantial overlap among red and blue lines.

We remark that the model is best suited for the purpose of volatility analysis and not for the purpose of forecasting. We use the proposed HMM model to forecast the returns from 2015-07-20 to 2015-11-24. Figure 6 shows that the predicted returns does not capture the behavior of the actual returns very well.

Category	Stocks
Large-cap/Mega-cap	Amazon.com, Inc. (AMZN), Apple Inc. (AAPL), Microsoft Corporation (MSFT), Exxon Mobil Corporation (XOM), AT&T Inc. (T), Johnson & Johnson (JNJ) Wells Fargo & Co (WFC), General Electric (GE), Chevron (CVX), Wal-Mart Stores, Inc. (WMT)
Small-cap	Planar Systems, Inc. (PLNR), Harsco Corporation (HSC), Trex Company, Inc. (TREX), Abercrombie & Fitch Co. (ANF), Alaska Air Group, Inc. (ALK), United Bancorp, Inc. (UBCP) Citizens Holding Company (CIZN), Kentucky First Federal Bancorp (KFFB) Dominion Resources Black Warrior Trust (DOM), CSP Inc. (CSPI)

Table 1: 10 Large-mega cap stocks and 10 small-cap stocks.

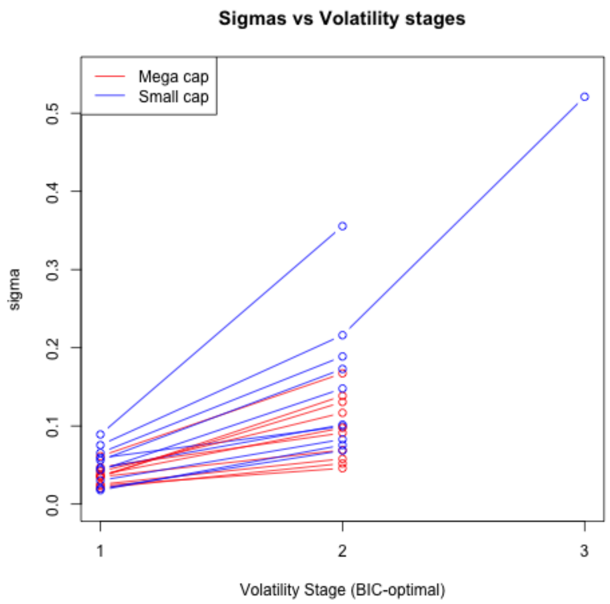

Figure 5: A plot of σ_i's versus the volatility stages for various large to mega cap stocks and small-cap stocks. For each stock, the numbers of volatility stages is chosen to minimize the BIC. The plot shows that most of the stocks have only two volatility stages.

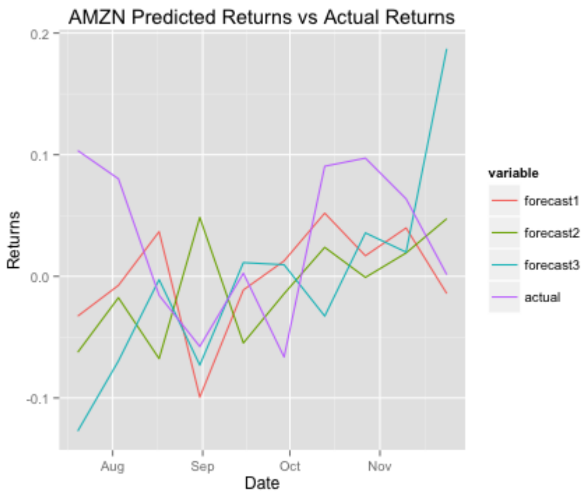

Figure 6: Predicted returns versus actual returns for AMZN from 2015-07-20 to 2015-11-24.

3 A Spectral HMM for stock returns forecasting

Here, we implement an alternative HMM parameter estimation method using the method of moments. In the HMM setting, we use a spectral decomposition technique to carry out this approach. This technique is fairly new, emerging recently within the past five years.

In particular, we analyze a spectral algorithm proposed by [Hsu et al. 2012] and assess how well it predicts stock returns. This method serves as a dual to the first model as it relies solely on observable quantities and therefore can be used for time series forecasting. To our knowledge, this is the first time spectral HMM methods have been used for predicting stock returns, but they have been applied to other datasets in the past.

3.1 Model Specification

The model details covered here are taken from the paper *A Spectral Algorithm for Learning Hidden Markov Models* [Hsu et al. 2012].

In this setting, we assume our transition and emission matrices are invertible and our initial state distribution is strictly positive. Although we don't work with these explicitly in the spectral setting, those conditions are necessary for HMMs to have a parametrization depending only on observable quantities.

For clarity, let $[n] = \{1, \dots, n\}$ be the set of observations and x_t represent the observation at timestep t.

We define our HMM in the following way [Hsu et al. 2012]:

1. $[P_1]_i = Pr[x_1 = i]$
2. $[P_{2,1}]_{ij} = Pr[x_2 = i, x_1 = j]$
3. $[P_{3,x,1}]_{ij} = Pr[x_3 = i, x_2 = x, x_1 = 1] \quad \forall x \in [n]$.

$P_1 \in \mathbb{R}^n$ is a vector while $P_{2,1} \in \mathbb{R}^{n \times n}$ and $P_{3,x,1} \in \mathbb{R}^{n \times n} \quad \forall x \in [n]$ are matrices. We can treat these three quantities as moment matrices for the first, second, and third moments.

The algorithm we implement takes in the number of hidden states and the sample size as parameters.

Algorithm LearnHMM(m,N) [Hsu et al. 2012]:
Inputs: m-number of states, N-sample size
Returns: HMM model parametrized by $\{\hat{b_1}, \hat{b_\infty}, \hat{B}_x \, \forall x \in [n]\}$

1. Sample N observation triples (x_1, x_2, x_3) from the HMM to form empirical estimates $\hat{P}_1, \hat{P}_{2,1}, \hat{P}_{3,x,1} \, \forall x \in [n]$ of $P_1, P_{2,1}, P_{3,x,1} \, \forall x \in [n]$.

2. Compute the SVD of $\hat{P}_{2,1}$, and let $\hat{U}$ be the matrix of left singular vectors corresponding to the m largest singular values.

3. Compute model parameters $\hat{b_1}, \hat{b_\infty}, \hat{B}_x \, \forall x \in [n]$.

   (a) $\hat{b_1} = \hat{U}^T \hat{P}_1$,

   (b) $\hat{b_\infty} = (\hat{P}_{2,1}^T \hat{U})^+ \hat{P}_1$,

   (c) $\hat{B}_x = \hat{U}^T \hat{P}_{3,x,1} (\hat{U}^T \hat{P}_{2,1})^+$

The model parameters can be used to predict the probability of a sequence of observations and the conditional probability of x_t given $x_1, \ldots, x_{t-1}$ as follows.

$$\hat{p}(x_1, \ldots, x_t) = \hat{b}_\infty^T \hat{B}_{x_t} \cdots \hat{B}_{x_1} \hat{b_1}$$

$$\hat{p}(x_t | x_1, \ldots, x_{t-1}) = \frac{\hat{b}_\infty^T \hat{B}_{x_t} \hat{b}_t}{\sum_x \hat{b}_\infty^T \hat{B}_x \hat{b}_t}.$$

$$\hat{b}_{t+1} = \frac{\hat{B}_{x_t} \hat{b}_t}{\hat{b}_\infty^T \hat{B}_{x_t} \hat{b}_t}$$

3.2 Experiment Design

We use the LearnHMM algorithm to forecast returns on the S&P 500 index returns for a five year period from 11-23-2010 to 11-23-2015 with each time step being a 10-day interval. Training sets were created from the first 40, 50, 60, and 80 timesteps (or 400, 500, 600, and 800 days) and then used to predict the rest of the index returns.

We transform our observation space to a discrete setting because otherwise the spectral algorithm will not apply–the observation transition matrices only make sense in a discrete setting. We do this via binning — identifying each data point with a quantile corresponding to the observation state x_t. For example, if our price change at the second timestep is in the third quantile, we say $x_2 = 3$. We used 6 observation states (quantiles). The idea was that we could rank our returns by 'very high', 'high', 'average/slightly above average', 'average/slightly below average', 'low', and 'very low'.

For forecasting, we determine $\hat{x}_{t+1}$ by taking the maximizing value of the conditional probability given the past observations. In succinct form,

$$\hat{x}_{t+1} = \max_{x_{t+1}} \hat{p}(x_{t+1} | x_1, \ldots, x_t).$$

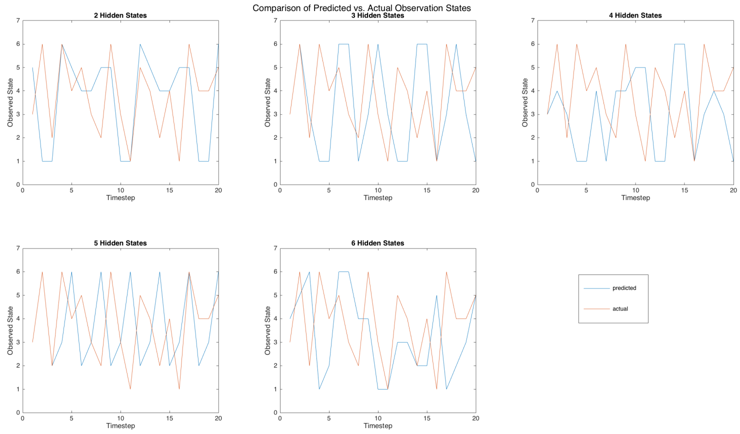

Figure 7: Accuracy of forecasting dependent on number of hidden states when training on 400 days (40 timesteps).

Then using this, we can calculate b_{t+1} and keep alternating between the two.

3.3 Results

Noticeably, the spectral HMM fared poorly for predicting returns in the S&P 500. This section will discuss the algorithm's empirical performance in more detail.

First, the forecasts, regardless of the number of hidden states and training dataset size used, would cycle through sets of states.

The plots shown in Figures 7 and 8 focus on a window of 20 timesteps corresponding to days 700 to 899.

The mean squared errors between our predicted observations and actual observations were computed for the different training sets and numbers of hidden states. We could have used our baseline prediction of 3 or 4 throughout and outperformed the predictions from LearnHMM as seen in our tables 2 and 4. It was interesting that the MSE was lower for smaller numbers of hidden states. When there were only two hidden states, the mean square error was at its lowest. When training on 600 days (60 timesteps), our LearnHMM for two hidden states actually outperformed our baseline with respect to the mean square error as seen in 5.

However, the cycling through states is not the only flaw with the LearnHMM algorithm. We saw while calculating likelihoods (joint probabilities) for different sets of observations that there were negative values. The table 6 provides more detail. On further inspection, there are also stability issues stemming from the SVD calculation of $\hat{P}_{2,1}$. As the sample size increases, the SVD does not converge. This was why our joint probability fluctuated wildly with respect to the sample size–the SVD was extremely sensitive to small perturbations in $\hat{P}_{2,1}$.

Originally, we thought there was a bug in our code until we saw these stability issues were discussed in a recent paper *A Sober Look at Spectral Learning* [Zhao and Poupart, 2014].

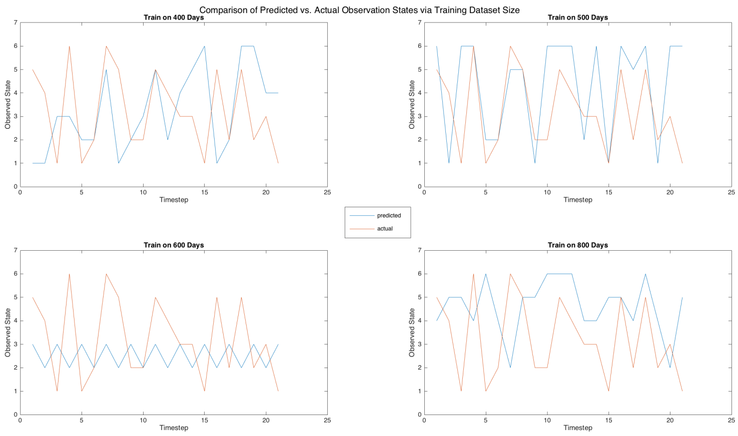

Figure 8: Accuracy of forecasting dependent on training dataset.

Number of hidden states	MSE
2	5.4588
3	5.7647
4	5.6235
5	6.2118
6	6.7647

Table 2: Our mean squared error between predicted and actual observations varying by the number of hidden states hyperparameter used in our LearnHMM algorithm. Here we train on 400 days (40 timesteps).

Training Size	MSE
400	6.7647
500	6.4667
600	4.646
800	6.800

Table 3: Our mean squared error between predicted and actual observations varying by the training data size used for forecasting. Here we assume 6 hidden states and 6 observation states.

Predicted State	MSE
3	3.192
4	3.184

Table 4: Our mean squared error between predicted and actual observations given we constantly predicted an observed state of 3 or 4.

Number of hidden states	MSE
2	2.7846
3	5.4615
4	4.1231
5	5.9385
6	5.8000

Table 5: Our mean squared error between predicted and actual observations varying by the number of hidden states hyperparameter used in our LearnHMM algorithm. Here we train on 600 days (60 timesteps).

Sample Size	Probability
350	2.40446414856475e-45
360	-2.37288148310411e-44
370	7.11145514863548e-44
380	-2.27827302121708e-44
390	-5.78094561020260e-44
400	-2.22816732532757e-42
410	1.54518332772887e-43
420	2.63872954946976e-40
430	1.98182048452036e-39
440	-3.24626501478278e-40
450	-6.25464466480119e-39
460	5.27309019205740e-37
470	-9.22984754857112e-38
480	6.37774305996794e-39
490	1.67453020299202e-37
500	2.23366415402976e-37

Table 6: Probabilities of observation data occurring based on sample sizes given they are taken from the estimated HMM.

Negative joint probabilities for some of our observed sequences also caused our conditional probabilities $p(x_t|x_1, \dots, x_{t-1})$ to take on negative values sometimes as well. However, we still went ahead and still did forecasting for x_t outside the training set by a maximum likelihood approach as detailed in section 3.2, because we did not know of an alternative.

4 Conclusion

Based on our first HMM model, most stocks have two volatility stages (low volatility and high volatility). In our specific case study of AMZN and PLNR, we find that high-volatility stage has a tendency to be transient and low-volatility stage has a tendency to be stable. A potential explanation is that the sudden high-volatility stages result from important news or announcements about the stocks and they catch the attention of the traders, leading the traders to trade aggressively. Surprisingly, the large/mega-cap stocks and the small-cap stocks have similar volatility. One would normally expect that large-cap stocks have smaller volatility compared to the small-cap stocks. Finally, a potential extension of the HMM model is to introduce two layers of hidden states: momentum and volatility. This extension will allow for more complicated interactions between volume and volatility.

An approach using method of moments (spectral HMM) did not fare very well, even when changing on the number of states. Noticeably, the predictions were always periodic in nature, but with different patterns when changing the number of hidden states. Originally, when looking at a smaller time interval, we thought that these predictions can give a general idea of the movement ahead of time, but upon further inspection with larger intervals and seeing periodicity, we learned this wasn't the case. While calculating the predicted likelihood of observed stock returns, we discovered negative values. Also there were stability issues using the spectral algorithm as it didn't converge after a high number of samples. After noticing these issues, we conclude the spectral HMM algorithm is too unstable for use in application. This was also discussed in a paper (Zhao and Poupart, 2014) highlighting the flaws of spectral HMM, but we noticed this issue independently.

References

[1] R. Hassan and B. Nath. Stock Market Forecasting Using Hidden Markov Model: A New Approach. Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA'05). 2005.

[2] D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. Journal of Computer and System Sciences, 78(5):1460–1480, 2012b.

[3] J. O, J. W. Lee, S. Park, and B. Zhang. Stock Trading by Modelling Price Trend with Dynamic Bayesian Networks. DEAL 2004, LNCS 3177, pp. 794-799, 2004.

[4] H. Zhao, P. Poupart. A Sober Look at Spectral Learning. 2014.

[5] X. Zhuang and L. Chan. Volatility Forecasts in Financial Time Series with HMM-GARCH Models. IDEAL 2004, LNCS 3177, pp. 807-812, 2004

[6] A. Rossi and G. Gallo. Volatility Forecasts in Financial Time Series with HMM-GARCH Models. Journal of Empirical Finance. Volume 13. Issue 2. pp. 203-230. 2006.