ETF & Stock Market Prices Prediction with HMM

Alexander Gofman and Tal Kaspani

Technion – Israel Institute of Technology

Introduction

The stock market is a network that provides a platform for almost all major economic transactions in the world.
Unfortunately, predicting how the stock market will perform is a very difficult thing to do. There are numerous factors involved in the prediction varying from physical and physiological factors, rational behavior, political events and more.
In this project, we will try to use Hidden Markov Models to predict the stock market behavior while inspecting various stocks and ETFs using years of trading data and compare our results to common Time Series methods.

The Data and Translation

We used the trading days data of various ETFs & stocks from the year 2010 onwards with each trading day consisting of the opening price.
We then translated each trading day into a series of symbols, where each symbol represents a specific change:

Date	Open	High	Low	Close	Volume
2010-01-04	20.574	20.7050	20.5370	20.687	9504561
2010-01-05	20.715	20.7150	20.5370	20.659	22483600

$$O_t = (Open_t, High_t, Low_t, Close_t, Volume_t)$$

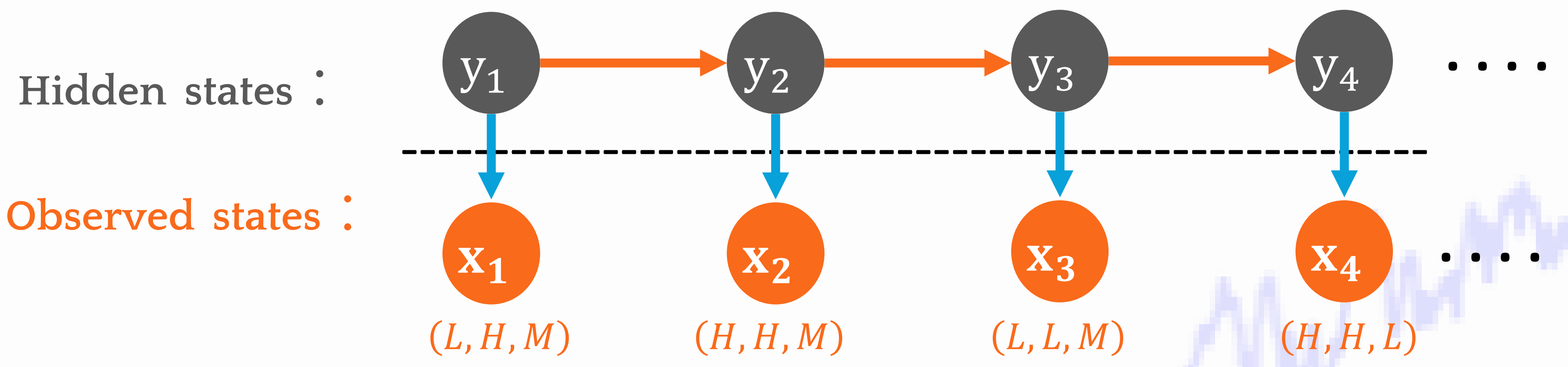
$$O_t = \left(\frac{close_t - open_t}{open_t}, \frac{high_t - low_t}{open_t}, \frac{open_t - close_{t-1}}{close_{t-1}}\right)$$

$$O_t = (fracChange, fracHigh, frachShift)$$

$$O_t = (L, L, M)$$

The Model

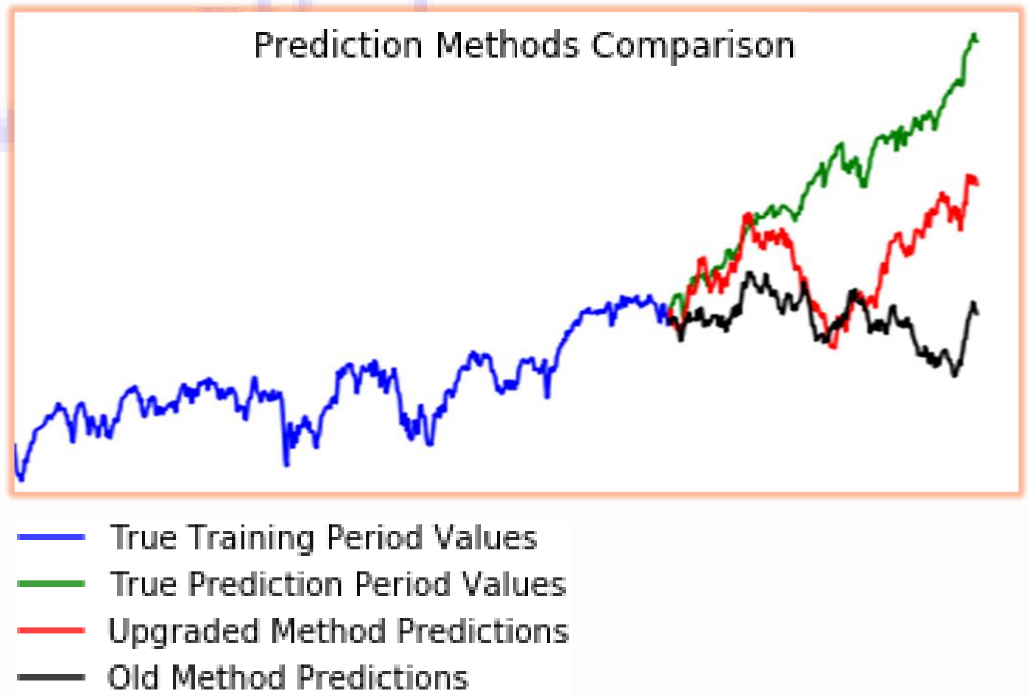
We used HMM to model the behavior of the stock market where the hidden states can be viewed as "*the state of the market*" and the observed states are the actual prices and changes in the value of the stock over time. In our case, the observed states are triplets/pairs of letters representing each trading day data.



The learning of the model was made by the Baum-Welch algorithm to achieve the appropriate transition matrices between the various states.
Once having the transition matrices, predictions were made by predicting the next most likely hidden state and from it, the most likely observed state. The predictions were in fact, a series of triplets that we later translated back to graphs using sampling methods and bootstrapping.

Advanced Part

In this part, we tried to predict the prices of the Technology Sector Fund. We did so by predicting the prices of it's composing stocks separately using our best model configuration and combining all results together while accounting for translation errors with linear regression:



Creative Part

In this part, we use the leading financial institutions' recommendations for promising stocks over the years to further improve our prediction method.
We did so by adding a new symbol for each instance where 'U' stands for Up and 'R' stands for Regular:

$$O_t = (L, L, M) \rightarrow O'_t = (L, L, M, \boldsymbol{R \backslash U})$$

After applying the transformation, we run our improved model and compared the results to our original model:

Method	MAPE	First Month	First Quarter
Original	143.3	Acc: **0.55** F1:**0.23**	Acc: **0.44** F1:**0.31**
Creative	136.1	Acc: **0.66** F1:**0.58**	Acc: **0.55** F1:**0.41**

Evaluation Measures

Our Basic evaluation measure is the Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|p_i - a_i|}{|a_i|} * 100\%$$

a_i - The actual stock value, p_i - Predicted stock value on day i and n is the number of days.
Furthermore, a more intuitive and useful measure we used is accuracy and F1 scores calculated based on the periodic changes of the stocks. We categorized behavior in the following ways:
"HARD SELL "– meaning stock value will decrease
"HARD BUY "– meaning stock value will increase
"HOLD" – meaning nothing significant will happen

Results

Model	MAPE
(Baseline) **Auto Regression**	**102.32**
(Baseline) **ARIMA**	**76.45**
HMM -4 hidden stats, 3 symbols, 4 years training data	**106.1**
HMM -2 hidden stats, 3 symbols, 6 years training data	**93.3**
HMM -2 hidden stats, 2 symbols, 6 years training data	**82.41**

Predicting the ETFs behavior, using our intuitive method our best model achieved:

Period	Accuracy	F1
First Month	**0.2**	**0.14**
First Quarter	**0.8**	**0.79**
First Year	**0.5**	**0.25**

Conclusions

- Predicting ETF prices is easier than predicting individual stocks.
- In contrast to other time series methods, old data is usable in HMM models (over 5 years).
- Real world insights are critical to the model's success.
- More features isn't necessarily better – a less robust representation might be more precise and prevents overfitting.